Aspects of Digital Libraries

Thesis for the Award of the Academic Degree of a Doctor of Technology at
Graz University of Technology

Josef Kolbitsch

Dr. Hermann Maurer (Thesis Supervisor)
Institute for Information Systems and Computer Media (IICM)
Graz University of Technology, Austria

Dr. Klaus Tochtermann (Thesis Reader)
Knowledge Management Institute
Graz University of Technology, Austria

Dr. Jennifer A. Lennon (Thesis Advisor)
Department of Computer Science
University of Auckland, New Zealand

Graz, Austria, January 2007.

Aspekte digitaler Bibliotheken

Dissertation in Telematik an der
Technischen Universität Graz

Dipl.-Ing. Josef Kolbitsch

o.Univ.-Prof. Dr.phil. Dr.hc.mult. Hermann Maurer (Betreuer und Begutachter)
Institut für Informationssysteme und Computer Medien (IICM)
Technische Universität Graz, Österreich

Univ.-Prof. Dr.rer.nat. Klaus Tochtermann (Begutachter)
Institut für Wissensmanagement
Technische Universität Graz, Österreich

Dr. Jennifer A. Lennon (Mitbetreuer)
Department of Computer Science
University of Auckland, Neuseeland

Diese Arbeit ist in englischer Sprache verfasst.

Graz, Österreich, im Januar 2007.

Preamble

This thesis is submitted in fulfilment of the requirements for the academic degree of a Doctor of Technology ("Dr. techn.") at Graz University of Technology, Austria.

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit – soweit nicht anders erwähnt – selbstständig und ohne fremde Hilfe oder unerlaubte Hilfsmittel verfasst, andere als die angegebenen Quellen nicht benutzt und die benutzten Quellen als solche kenntlich gemacht habe. Weiters versichere ich, dieses Dissertationsthema bisher weder im In- noch im Ausland in irgendeiner Form als Prüfungsarbeit zur Beurteilung vorgelegt zu haben.

Acknowledgements

Most PhD students wish for a visionary supervisor, a wise and experienced researcher. One of these great minds is Hermann Maurer. I have to thank him for giving me the exceptional opportunity of doing a PhD with him.

Imagine arriving at an airport at 5am, after a 35-hour flight, and there is someone to pick you up. This person had to get up at 3.30am. You are picked you up at the airport, and taken to a place where you can stay until you have found an apartment. It does not stop here. You are always welcome as a guest, there is ongoing support for you whenever you need it, … This person is Jennifer Lennon. Thank you Jennifer for all you have done for me, for your encouragement, your advice, and your truly honest comments. Thanks you for making me "jump over my own shadow". If my ideas are gemstones, you helped me polish them. You have introduced me to scientific writing and fuelled my desire to do research. I am deeply thankful for your friendship and hospitality.

My dearest thanks go to my parents for their support and confidence, and for never showing any doubts in my ventures—no matter how infeasible they might seem.

Thank you Amélie for your encouragement, support, time, and most importantly your patience during the busiest months.

I would also like to thank my friends, my "extended family", and all supporters of my work. Thank you Martin, for our frequent discussions.

My thanks also go to Edmund Haselwanter, Harald Krottmaier, Maria-Luise Lampl, Steven Mitter, and Christian Safran for their professional support and advice.

Thank you.

Abstract

Following profound transformations, the World Wide Web of the early 21st century has turned into a network allowing the user community to participate in the process of content production. As the driving force behind this change, several novel concepts, such as weblogs, wikis, legal file sharing services, and social networking services have emerged. These services constitute a new and unconventional type of large-scale library applications.

In this thesis, three particular aspects of these uncommon digital libraries are analysed: electronic encyclopaedias, metadata, and collaboration. The first part of the thesis presents a concept for an advanced community-based encyclopaedic environment. In this design, users from the community may author content while a quality assurance mechanism maintains a high standard for publications. In addition to this, enhanced functionality such as transclusions and adaptation are addressed. Furthermore, Wikipedia is examined with the intention of determining the influence of anonymous users on a collaborative environment.

An investigation of the use of Dublin Core metadata within academic and educational settings, conducted in this thesis, shows that the usage of this metadata standard is relatively infrequent. In contrast to this, tagging systems, a quasi-"competitor" of classic metadata systems, are very popular among users. A solution for the Vocabulary Problem, one of which is encountered in tagging systems, is presented in this work.

This thesis also introduces Kōrero, the concept of a platform capable of integrating collaborative technologies, including weblogs and wikis, on the basis of an underlying social network. Kōrero relies on a unified data model for its collaborative environment. The design of a structure that is compatible with Kōrero as well as existing systems, like weblogs, is also addressed.

Some concepts of this work are being implemented in project "Alexander"—a community-based system combining information from encyclopaedias and other high-quality resources. Other parts of this thesis were developed as prototypes and proof-of-concept implementations.

Keywords: Digital Libraries, Electronic Encyclopaedias, Collaborative Systems, Hypermedia, Metadata, Data Models, Weblogs, Wikis, Social Networks.

Categories (ACM 1998): H.3.5, H.3.7, H.5.1, H.5.3, H.5.4.

Kurzfassung

In den ersten Jahren des 21. Jahrhunderts hat sich das World Wide Web zu einem Netzwerk gewandelt, das es Benutzern erlaubt, selbst an der Produktion von Inhalten teilzunehmen. Die treibende Kraft hinter dieser Transformation sind neue Konzepte wie Weblogs, Wikis, legale Dateitauschbörsen und soziale Netzwerke. Diese Dienste stellen eine neue und unkonventionelle Art von Bibliotheksanwendungen dar.

In dieser Arbeit werden drei Aspekte solch ungewöhnlicher digitaler Bibliotheken beleuchtet: elektronische Enzyklopädien, Metadaten und Kollaboration. Im Hauptteil der Dissertation wird ein Konzept für eine fortschrittliche, die Benutzer-Community einbeziehende Enzyklopädie vorgestellt. Als Teil dieses Designs können Benutzer neue Inhalte erstellen, während Instanzen zur Qualitätssicherung in Kraft sind. Zudem werden in diesem Konzept erweiterte Funktionalitäten wie Transclusions und Adaptierung eingesetzt. Um die Rolle von anonymen Autoren in einer kollaborativen Umgebung besser verstehen zu können, wird Wikipedia, eine im Internet frei verfügbare Enzyklopädie, analysiert.

Eine im Rahmen dieser Dissertation durchgeführte Untersuchung der Verwendung von Dublin Core Metadaten in akademischen Einrichtungen zeigt, dass dieser Standard für die Beschreibung von Ressourcen nur relativ eingeschränkt verwendet wird. Tagging Systeme hingegen, eine Art „Konkurrenz“ für klassische Metadatensysteme, erfreuen sich bei Benutzern großer Beliebtheit. Für eines der Probleme in Tagging Systemen, das Vocabulary Problem, wird in dieser Arbeit ein Lösungsansatz vorgestellt.

Diese Arbeit stellt auch Kōrero vor, das Konzept einer Plattform, welche kollaborative Technologien wie Weblogs und Wikis auf Basis eines sozialen Netzwerkes integriert. Kōrero erfordert ein vereinheitlichtes Datenmodell für kollaborative Umgebungen. Das Design eines Modelles, das im Wesentlichen sowohl mit Kōrero als auch mit bereits existierenden Systemen kompatibel ist, wird erläutert.

Einige Konzepte dieser Arbeit werden im Rahmen von Projekt „Alexander“, einem System, das Enzyklopädien und weiteren qualitativ hochwertigen Quellen Inhalte entnimmt und die Benutzer-Community in deren Weiterentwicklung einbezieht, umgesetzt. Andere Teile der Dissertation wurden als Prototypen und Proof-of-Concept Implementierungen realisiert.

Schlüsselwörter: Digitale Bibliotheken, Elektronische Enzyklopädien, Kollaborative Systeme, Hypermedia, Metadaten, Datenmodelle, Weblogs, Wikis, Soziale Netzwerke.

Kategorien (ACM 1998): H.3.5, H.3.7, H.5.1, H.5.3, H.5.4.

Author's Note

Many people claim that computer science is short-lived, that it is probably the most "ephemeral" of all sciences. At times, this is hard to believe because, as computer scientists, we design and implement systems to be used for a potentially long time.

However, working on this thesis made me reconsider my attitude. When I started research for my PhD hardly anyone actually realised, for example, that not only a few new services were made available on the Web but that the entire Web was undergoing massive changes. Today, this view is commonly acknowledged and is considered "an old hat". When I did analyses on Wikipedia and concluded, for instance, that a completely editor-less encyclopaedia is probably not an optimal approach, I began designing an electronic encyclopaedia whose concept includes editor-like experts in order to ensure quality content. Only a few weeks before the final publication of this thesis one of the founders of Wikipedia announced a spin-off project that intents to implement an encyclopaedia very similar to our design …

Although these experiences can be interpreted as the thesis "going the right direction" they can, at the same time, be somewhat frustrating. In this context, my supervisor's advice to publish the individual chapters of this thesis as separate papers proved to be very appropriate. With such an approach, PhD students can make their current results instantly available even in short-lived areas.

I am glad that I could finish this thesis, I am fascinated about being on the leading edge, and I am happy to be involved in doing research. Most significantly, I have learnt one important lesson—computer science *is* short lived, and you have to be at least *two steps* ahead to be a good researcher in this field.

Contents

Introduction **1**

Chapter 1: The Transformation of the Web **5**
1.1 Introduction 5
1.1.1 Self-Organising Structures 6
1.1.2 Recent Developments on the Internet 6
1.2 Blogs 6
1.2.1 Types of Blogs and Applications 7
1.2.2 The Blogging Community 7
1.2.3 Advantages and Drawbacks of Blogs 7
1.3 Wikis 8
1.3.1 Technical Aspects of Wikis 8
1.3.2 Application Areas 8
1.3.3 Benefits and Shortcomings of Wikis 9
1.3.4 Counter Measures 9
1.4 Wikipedia 10
1.4.1 Advantages of Wikipedia 10
1.4.2 Drawbacks of the Approach 10
1.4.2.1 Quality and Authority 10
1.4.2.2 Background and Balance 10
1.4.2.3 Sensitive Information and Privacy 11
1.4.2.4 Wiki-Related Phenomena 11
1.4.2.5 Awareness of the Concept 11
1.4.2.6 Volume of Wikipedia 11
1.5 Wikinews 12
1.5.1 Goals of Wikinews 12
1.5.2 The Relationship between Wikinews and Wikipedia 12
1.5.3 Success of Wikinews 12
1.6 Podcasts 12
1.6.1 Use of Podcasts and Application Areas 13
1.6.2 Similar Technologies 13
1.7 File Sharing Tools 13
1.8 Social Networks 14
1.8.1 Large Communities 14
1.8.2 Technical Aspects 14
1.8.3 Use of Social Networks 15
1.8.4 Other Community-Based Networks 15
1.9 Summary and Outlook 16
1.9.1 Major Transformations 16

1.9.2 Opportunities and Future Trends 17
1.9.3 Challenges and Concerns 17
1.9.4 Technological Aspects 18
1.10 Conclusion 18

Part 1: Electronic Encyclopaedias 19

Chapter 2: Community-Building Around Encyclopaedic Knowledge 21
2.1 Introduction 21
2.1.1 Traditional Electronic Encyclopaedias 21
2.1.2 Brockhaus Digital Encyclopaedia 21
2.1.3 Wikipedia—A Radically Different Concept 22
2.1.4 Google Answers 22
2.2 Functionality Beyond Traditional Electronic Encyclopaedias 22
2.2.1 Annotations 22
2.2.2 Active Multimedia Documents 23
2.2.3 Links to External Resources 23
2.2.4 Private and Shared Workspaces 23
2.2.5 User Profiles and Adaptation 24
2.2.6 Trails 24
2.2.7 Quality Feedback and Review Mechanism 24
2.3 Community Building in Electronic Encyclopaedias 25
2.3.1 Purpose and Notion of a User Community 25
2.3.2 Foundations of a User Community 25
2.3.3 Quality Control 26
2.3.4 Active Knowledge Brokering 26
2.3.5 Social Aspects 27
2.3.6 Mixed and Split Content 27
2.3.7 Notification and Communication among Users 28
2.3.8 Discussion Forums 28
2.4 System Proposal 29
2.4.1 Overview 29
2.4.2 User Community 29
2.4.3 Basic Functionality 29
2.4.4 Authoring New Content, Linking, and Transclusions 30
2.4.5 Adaptation 30
2.4.6 Clustering and Knowledge Maps 30
2.4.7 Active Knowledge Brokering 30
2.4.8 Discussion Forums and Communication 31
2.5 Application Areas 31
2.5.1 Enhanced Encyclopaedic Environments 31
2.5.2 E-Learning 31
2.6 Summary and Outlook 31

Chapter 3: Adaptation of Content and Structure in Electronic Encyclopaedias 33
3.1 Introduction 33
3.2 User Adaptive Systems 33
3.2.1 Adaptation Technologies 34
3.2.1.1 Adaptation of User Interfaces 34
3.2.1.2 Content Adaptation 34
3.2.1.3 Structural Adaptation 35
3.2.2 Adaptation Mechanisms 35
3.2.2.1 Static Adaptation 35
3.2.2.2 Dynamic Adaptation 36
3.2.2.3 Flexible Dynamic Adaptation 36

3.3 Adaptation in Electronic Encyclopaedias 37
3.3.1 Replacement, Explanation, and Linking 37
3.3.2 Translation 38
3.3.3 Blacklisting 38
3.3.4 Whitelisting 39
3.3.5 Link Alteration 39
3.3.6 Navigation Support 39
3.3.7 Information Encountering 40
3.4 Technical Aspects and Proposed Implementation 40
3.4.1 Structure and Categorisation of Articles 40
3.4.2 Profile Matrices for Users 40
3.4.3 System Architecture 41
3.4.4 System Components 41
3.4.5 Source of Required Data 42
3.4.6 Performance Issues 42
3.5 Application Areas 42
3.6 Summary 42

Chapter 4: Transclusions in an HTML-Based Environment 45
4.1 Introduction 45
4.1.1 Background of Transclusions 45
4.1.2 Implications of Transclusions 46
4.1.3 Transclusions and HTML 46
4.2 Attempts to Implement Transclusions 46
4.2.1 Xanadu 46
4.2.2 Proposal to Amend the HTML Specification 47
4.2.3 Transclusions with IFrames and Embedded Objects 47
4.2.4 XML-Based Transclusions 47
4.2.5 Recent Projects Involving Transclusions 47
4.2.5.1 University of Nottingham 47
4.2.5.2 University of Bologna 48
4.2.5.3 IICM, Graz University of Technology 48
4.3 Implementation 48
4.3.1 Design Goals and Requirements 48
4.3.2 System Overview 48
4.3.3 System Architecture 49
4.3.4 Creating a Transclusion 49
4.3.5 Retrieving a Transclusion 50
4.4 Issues Encountered 51
4.4.1 Javascript Restrictions 51
4.4.2 Browser Specific Implementation 51
4.4.3 Modified Documents and Unavailable Resources 51
4.5 Discussion 52
4.5.1 Robustness 52
4.5.2 Aspects of the Design 53
4.5.3 Aspects of the Proxy Application 53
4.6 Conclusion 53

Chapter 5: Fine-Grained Transclusions of Multimedia Documents in HTML 55
5.1 Introduction 55
5.2 Fine-Grained Transclusions of Multimedia Documents 55
5.2.1 Drawings, Vector Graphics 56
5.2.2 Photos and Images 56
5.2.3 Video Content 56
5.2.4 Sound and Music 57
5.2.5 Compound Multimedia Documents 57

5.3 Approaches to an Implementation 57
5.3.1 Design Goals and Requirements 57
5.3.2 System Architecture 58
5.3.3 Extension Module for HTTP Servers 59
5.3.4 CGI Program 59
5.3.5 Plug-In Architecture 60
5.3.6 Prototype Implementation 60
5.4 Application Areas 61
5.4.1 News Providers 61
5.4.2 Galleries and Museums 61
5.4.3 Learning Environments 61
5.4.4 Movie Archives and Music Stores 62
5.5 Conclusion 62

Chapter 6: Anonymous Authorship in Wikipedia 63
6.1 Introduction 63
6.2 Previous Work 63
6.2.1 Wikipedia Statistics 64
6.2.2 Wikipedia and the Media 64
6.2.3 Social Phenomena in Wikipedia 64
6.2.4 Qualitative Comparison 64
6.2.5 Formality of Content 64
6.3 Anonymity in Information Systems 65
6.3.1 Varying Levels of Anonymity 65
6.3.1.1 Physical Identification 65
6.3.1.2 Strict Authentication 65
6.3.1.3 Lax Authentication 65
6.3.1.4 Pseudo Anonymity 65
6.3.1.5 True Anonymity 66
6.3.2 Impact of Anonymity 66
6.3.3 Anonymity in Wikipedia 66
6.4 Aim and Methodology 66
6.4.1 Selection of Articles 67
6.4.2 Classification and Categorisation of Articles 67
6.4.3 Tools 67
6.4.4 Data Analysis 68
6.4.5 Repeated Analysis 68
6.4.6 Methodological Concerns 69
6.4.6.1 Approach to the Analysis 69
6.4.6.2 Anonymous Users 69
6.5 Results and Interpretation 69
6.5.1 Anticipated Results 69
6.5.2 General Findings 69
6.5.2.1 Comparison of Data Sets 70
6.5.2.2 Skewness of Data 70
6.5.2.3 Edits per Article 70
6.5.2.4 Activity of Users 70
6.5.3 Analysis of Classes 71
6.5.4 Analysis of Categories 71
6.5.5 Anecdotal Note 72
6.6 Conclusion 72

Part 2: Metadata and Taxonomies 73

Chapter 7: The Use of HTML Encoded Dublin Core in Academic Web-Sites 75
7.1 Introduction 75
7.1.1 Core Elements 75
7.1.2 HTML-Encoded Dublin Core 76
7.2 Related Research 76
7.3 Aim and Hypotheses 76
7.3.1 Motivation and Aims 76
7.3.2 Hypotheses 77
7.3.2.1 Number and Proportion of Pages with DC Elements 77
7.3.2.2 Overhead Produced through DC 77
7.3.2.3 Creation of Pages with DC Elements 77
7.3.2.4 Syntactic Correctness of Pages with DC Elements 77
7.3.2.5 Most Frequently Used DC Elements 77
7.4 Methodology and Tools of this Study 78
7.4.1 Selection of Sites 78
7.4.2 Tools 78
7.4.2.1 Page Crawler 78
7.4.2.2 Tools for Data Extraction 78
7.4.2.3 Tools for Data Analysis 79
7.4.3 Methodological Concerns 79
7.5 Results and Interpretation 79
7.5.1 Use of DC Elements 79
7.5.2 Frequency of Simple DC Elements 80
7.5.3 Page Head Size 80
7.5.4 Page Generation 80
7.5.5 HTML Syntax Validity 81
7.6 Discussion 81
7.6.1 Additional Findings 81
7.7 Conclusion 82

Chapter 8: WordFlickr 83
8.1 Introduction 83
8.1.1 Tagging Systems 83
8.1.1.1 Structure of Tagging Systems 84
8.1.1.2 Benefits of Tagging Systems 84
8.1.1.3 Restrictions and Shortcomings 84
8.1.2 del.icio.us 85
8.1.3 Flickr 85
8.1.4 Motivation 86
8.2 Previous Research 86
8.2.1 Structural Analysis of Tagging Systems 87
8.2.2 Hierarchies in Tagging Systems 87
8.3 Enhancing Flickr Queries 87
8.3.1 WordFlickr 87
8.3.1.1 Concept 87
8.3.1.2 Implementation 87
8.3.1.3 Discussion of the Implementation 88
8.3.2 FlickrClustr 88
8.3.2.1 Concept 88
8.3.2.2 Implementation 88
8.3.2.3 Remark on the Implementation 88
8.4 Discussion 89
8.4.1 Concept and Implementation 89
8.4.2 Informal Comparison 89

8.4.2.1 General Findings 89
8.4.2.2 Tag Inconsistencies 89
8.4.2.3 Tag Usage and Interpretation 90
8.5 Conclusion 90

Part 3: Integration of Collaborative Concepts 91

Chapter 9: Kōrero: An Integrated Community-Based Platform for Collaboration 93
9.1 Introduction 93
9.1.1 Blended Systems 93
9.2 Concept 93
9.2.1 System Generated Information 94
9.2.2 Guided Information Discovery 94
9.2.3 Visualisation 95
9.3 The Proposed Platform 95
9.3.1 Core Technologies 96
9.3.2 Content Generating Components 96
9.3.3 Communication Facilities 96
9.4 Aspects of Organisation 96
9.4.1 Organisation of Content 96
9.4.2 Organisation of Users 97
9.5 Application Areas 97
9.5.1 Learner-Support Systems 97
9.5.2 Corporate and Organisational Environments 97
9.5.3 Special Interest Groups 98
9.6 Conclusion 98

Chapter 10: A Unified Structure for Current Collaborative Systems 99
10.1 Introduction 99
10.2 Characteristics of Collaborative Systems 99
10.2.1 Blogs 99
10.2.2 Wikis 100
10.2.3 Podcasts 100
10.2.4 Social Bookmarking and File Sharing 100
10.2.5 Social Networks 100
10.3 Proposed Structure 100
10.3.1 Entities 100
10.3.1.1 Content 100
10.3.1.2 Content Type 101
10.3.1.3 Metadata 101
10.3.1.4 Person and Group 101
10.3.1.5 Profile 101
10.3.1.6 Permission 101
10.3.2 Relations 101
10.3.2.1 Authorship 102
10.3.2.2 Content Link 102
10.3.2.3 Person Link 103
10.3.2.4 Metadata 103
10.3.2.5 Permission 103
10.4 Exemplary Applications 103
10.4.1 Blogs 104
10.4.2 Wikis 104
10.4.3 Podcasts 104
10.4.4 Tagging Systems 104
10.4.5 Social Network 105

10.5 Conclusion 105

Conclusion 107

Appendixes 109

Appendix A: The Growing Importance of E-Communities on the Web 111
A.1 Introduction 111
A.2 Community-Based Services on the Web 111
A.2.1 Blogs 112
A.2.1.1 Blog Styles 112
A.2.1.2 Technical Aspects 112
A.2.2 Wikis 113
A.2.2.1 Advantages and Drawbacks of Wikis 113
A.2.2.2 Wikipedia 114
A.2.3 File Sharing Services 114
A.2.3.1 Podcasting 115
A.2.4 Social Networks 115
A.2.5 Other Community-Based Applications 116
A.3 What Is Really New? What Is Different? 116
A.3.1 Blogs 116
A.3.2 Wikis 117
A.3.3 File Sharing 117
A.3.4 Social Networks 117
A.4 Impact of Recent Community-Based Developments 117
A.5 Future Advancements 118
A.5.1 Upcoming Technologies 119
A.6 Conclusion 119

Appendix B: Alexander: A Basic Functional Specification 121
B.0 Konventionen 121
B.1 Einleitung und Ziele 121
B.2 Allgemeine Definitionen 122
B.2.1 Benutzer 122
B.2.2 Artikel 122
B.2.3 Kategorien 122
B.2.4 Annotation 122
B.2.5 Hyperlink 122
B.3 Organisation 123
B.3.1 Benutzerrollen 123
B.3.2 Kategorien 123
B.4 Benutzerfunktionen 123
B.4.1 Artikel und Kategorien 123
B.4.1.1 Artikel suchen (M) 123
B.4.1.2 Durch Artikel und Kategorien browsen (M) 123
B.4.1.3 Artikel anzeigen (M) 124
B.4.1.4 Artikel, Annotationen, etc. verfassen (M/N) 124
B.4.1.5 Artikel ändern (M) 124
B.4.1.6 Artikel entfernen (M) 124
B.4.1.7 Artikel zertifizieren (N) 125
B.4.1.8 Artikel bewerten (M) 125
B.4.1.9 Artikel bookmarken (M) 125
B.4.1.10 Fragen an Artikel stellen (Active Documents) (M) 125
B.4.1.11 Artikel Kategorien zuordnen (M) 126
B.4.1.12 Kategorien anlegen (M) 126

B.4.1.13 Kategorien umbenennen (M) 126
B.4.1.14 Kategorien zusammenlegen (M) 126
B.4.1.15 Kategorien aufteilen (N) 126
B.4.1.16 Kategorien löschen (M) 127
B.4.1.17 Kategorien verschieben (M) 127
B.4.2 Diskussionen (M) 127
B.4.2.1 Beiträge lesen (M) 127
B.4.2.2 Beiträge verfassenen (M) 127
B.4.2.3 Beiträge bewerten (M) 127
B.4.2.4 Beiträge löschen (M) 128
B.4.3 Active Knowledge Brokering (M) 128
B.4.3.1 Fragen stellen (M) 128
B.4.3.2 Vorgeschlagener Content aus dem System (M) 128
B.4.3.3 Bereits vorhandene Antworten (Active Documents) (M) 128
B.4.3.4 Weiterleiten der Frage an Experten (M) 128
B.4.3.5 Antwort durch einen Experten (M) 129
B.4.4 Kommunikation (M) 129
B.4.4.1 Synchrone Kommunikation mit anderen Benutzern (M) 129
B.4.4.2 Asynchrone Kommunikation mit anderen Benutzern (M) 129
B.4.4.3 Externe Kommunikation (N) 129
B.4.4.4 Personalisierter Benachrichtigungsdienst (N) 129
B.4.5 Personalisierung (N/O) 130
B.4.5.1 Profile ausfüllen (N) 130
B.4.5.2 Suchergebnisse (N) 130
B.4.5.3 Anzeige von Artikeln (O) 130
B.4.5.4 Artikel des Tages (N) 130
B.4.5.5 In den Nachrichten (O) 131
B.4.6 Workspaces (O) 131
B.4.6.1 Private Arbeitsbereiche (O) 131
B.4.6.2 Gemeinsame Arbeitsbereiche (O) 131
B.4.6.3 Visualisierung und Spatial Hypertext (O) 131
B.4.7 Benutzerzertifizierung und -verwaltung (M/N) 131
B.4.7.1 Benutzer befördern oder degradieren (M) 131
B.4.7.2 Vorschlagen neuer zertifizierter Experten (M) 132
B.4.7.3 Mit vordefinierten Benutzergruppen arbeiten (N) 132
B.4.7.4 Eigene Gruppen anlegen und verwalten (N) 132
B.4.8 Qualitäts-Feedback und Review Mechanismus (N) 132
B.5 Systemfunktionen 133
B.5.1 Suche und Clustering (N) 133
B.5.2 Blacklisting (M/N) 133
B.5.3 Whitelisting (M/N) 134
B.5.4 Transclusions (N/O) 134
B.5.5 Statistische Auswertungen (N) 134
B.5.6 Personalisierung (O) 135
B.6 Potentielle Design Entscheidungen 135
B.6.1 Link Datenbank (O) 135
B.6.2 Persistenz und Konsistenz von Links (M/N) 135
B.6.3 Versionskontrolle (N) 135
B.6.4 Identifizierung von Objekten (N) 136
B.6.5 Metadaten zu Objekten (N) 136

Appendix C: Data from the Analysis of Anonymous Authorship in Wikipedia 137

Appendix D: Data from the Analysis of HTML-Encoded Dublin Core Metadata 155

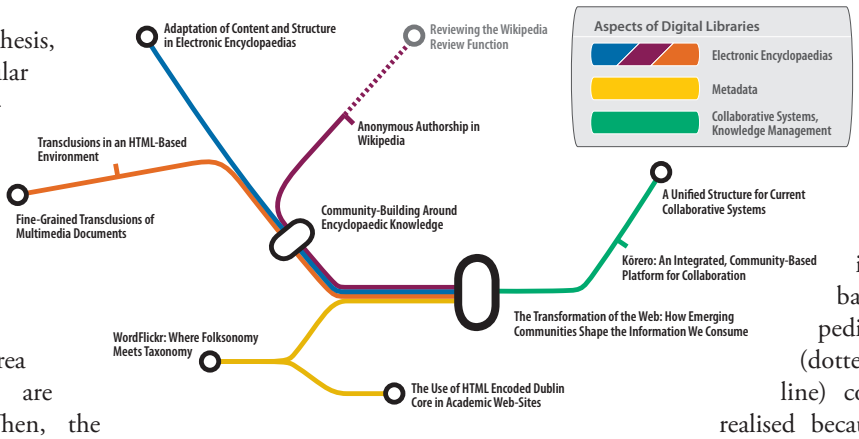
Appendix E: Data from Experiments with WordFlickr and FlickrClustr 159

References 161

Introduction

Today, digital libraries comprise a wide range of systems and services from classic libraries and electronic encyclopaedias to computer-supported collaborative work environments, image databases, and documentation systems. Recently, novel forms of digital libraries were developed and introduced to a wide audience on the World Wide Web: wikis, weblogs, social networks, file sharing services, etc. These types of libraries have unconventional and sometimes chaotic approaches to authoring and organising content. Although they manage to attract millions of users, many of their characteristics are not well understood, and some of their implications might not even have been discovered yet.

In this thesis, three particular facets of digital libraries are investigated. First, electronic encyclopaedias as a popular application area of libraries are addressed. Then, the use of metadata is discussed and finally, collaborative features in libraries and functionality known from knowledge management systems are highlighted.



Electronic Encyclopaedias

While traditional encyclopaedias are authored by experts, articles in Wikipedia, a free, collaborative encyclopaedia, may be produced and modified by anybody on the Internet. Hence with Wikipedia, a new and unusual concept is introduced to encyclopaedias, which brings a renaissance to research in this area. However, most research projects focus on the unconventional approach taken by Wikipedia. Other aspects of electronic encyclopaedias such as the lack of functionality for supporting the users' needs are not considered. Therefore in this thesis, several topics that are frequently overlooked are dealt with. An emphasis is put on the quality and structure in encyclopaedias, and a number of demands for upcoming advanced electronic encyclopaedias are defined. Technical designs and implementations for the key requirements are detailed in this thesis: transclusions for advanced authoring facilities (orange lines in the figure on this page), adaptation of content and structure (blue line), and an analysis of the role of anonymous authors in Wikipedia (purple line). Unfortunately, another investigation based on Wikipedia's content (dotted purple line) could not be realised because the corresponding functionality—the article validation feature—was not implemented in Wikipedia.

Metadata

Similar to electronic encyclopaedias, the use of metadata has recently undergone changes. Well-established standards auch as the Dublin Core metadata initiative face the "competition" of tagging systems with their community-authored, loose, unstructured metadata. On the one hand, the actual use of established technologies including Dublin Core is unclear.

On the other hand, it is a known fact that tagging systems have a number of shortcomings. Therefore, the usage of Dublin Core metadata in academic and educational settings, on of the most popular application areas of this standard, is analysed in this thesis (yellow line in the figure on the previous page). Moreover, a brief overview of tagging systems is given, and a solution to a frequent problem in these environments is offered.

Collaboration and Knowledge Management

The recent library-like applications including wikis, weblogs, and social networks are usually individual systems. Neither are they integrated into a single system nor are synergetic effects that might emerge utilised. Therefore in this thesis, an integrated community-based platform for collaboration is designed—Kōrero.

The underlying structure of Kōrero is a social network storing information on all users in the system and the relations among the users. Based on this infrastructure, technologies for content generation such as wikis, blogs, and similar techniques can be implemented. The concept of the system enables the discovery of tacit knowledge such as implicit links between chunks of information or implicit relationships among users.

The way Kōrero is designed, an underlying data model that unifies the structures of existing services is required. Therefore a unified infrastructure for current collaborative systems is detailed.

Organisation of this Thesis

This thesis, *Aspects of Digital Libraries,* is subdivided into three parts and an introductory chapter. The first chapter provides the foundations for the subsequent parts of the thesis and gives a detailed overview of recent developments on the Internet and the transformations the Web has undergone.

Most chapters of this thesis were published individually. For every publication, the title of the journal or conference, information on the submission date and on co-authors are supplied.

Part 1

Part 1 of the thesis focuses on electronic encyclopaedias. In chapter 2, state-of-the-art electronic encyclopaedias are described, and seven demands for upcoming advanced encyclopaedias are formulated. In addition to this, the importance of allowing users of the encyclopaedia to work actively with the content and to participate in the content creation process is underlined. The design of a system implementing community-building around encyclopaedic information is proposed.

One of the demands for enhanced electronic encyclopaedias is the use of adaptation mechanisms. Chapter 5 details dynamic adaptation of content and structure are introduced.

Another requirement for electronic encyclopaedias stated in chapter 2 are transclusions for creating new content. Chapters 4 and 5 present concepts and implementations of both textual and multimedia transclusions in HTML-based environments. In these implementations, transclusions can be made on the finest granularity possible—for text on the level of single characters, for images on pixel-basis, etc.

The concept of community-building in electronic encyclopaedias is expert-based and does not necessarily allow anonymous authors to modify or create content. However, the role and importance of anonymous authors in collaborative environments are not known. Therefore an analysis of anonymous authorship in Wikipedia, one of the largest community-based environments focusing on content production, is presented in chapter 6.

Part 2

Part 2 of the thesis is concerned with metadata in digital libraries. As mentioned above, one of the leading metadata standards is Dublin Core. In chapter 7, the results of an investigation of Dublin Core metadata in academic and educational settings are presented. By contrast, popular tagging systems with their radically different approach to the creation of loose and unstructured metadata involve a number of unresolved issues. An answer for one of these deficiencies is detailed in section 8.

Part 3

The third part of the thesis aims at integrating existing collaborative concepts into a single, unified infrastructure. Section 9 introduces Kōrero and the components required for an integrated, collaborative platform. The underlying data models of technologies such as wikis and weblogs are analysed in section 10, and a unified structure that is supportive of Kōrero is presented.

Appendixes

The appendixes at the end of this thesis include an additional chapter on the importance of online communities on the Web and an outlook on their potential future development. Moreover, a basic functional specification for the implementation of the community-based electronic encyclopaedia in chapter 2 is provided (project “Alexander”). In addition to this,

the detailed results of the investigations in chapters 6 (anonymous authorship in Wikipedia), 7 (use of Dublin Core metadata), and 8 (issues in tagging systems) are listed in the appendixes.

1

The Transformation of the Web

How Emerging Communities Shape the Information We Consume

A version of this chapter was published in the Journal of Universal Computer Science, Volume 12, Issue 2, pp. 187-213.

A version of this chapter was also published on www.masternewmedia.org, a publishing platform for professionals in new media.

Publication submitted in January 2006, co-authored by Hermann Maurer.

To date, one of the main aims of the World Wide Web has been to provide users with information. In addition to private homepages, large professional information providers, including news services, companies, and other organisations have set up web-sites. With the development and advance of recent technologies such as wikis, blogs, podcasting and file sharing this model is challenged and community-driven services are gaining influence rapidly. These new paradigms obliterate the clear distinction between information providers and consumers. The lines between producers and consumers are blurred even more by services such as Wikipedia, where every reader can become an author, instantly.

This chapter presents an overview of a broad selection of current technologies and services: blogs, wikis including Wikipedia and Wikinews, social networks such as Friendster and Orkut as well as related social services like del.icio.us, file sharing tools such as Flickr, and podcasting. These services enable user participation on the Web and manage to recruit a large number of users as authors of new content. It is argued that the transformations the Web is subject to are not driven by new technologies but by a fundamental mind shift that encourages individuals to take part in developing new structures and content. The evolving services and technologies encourage ordinary users to make their knowledge explicit and help a collective intelligence to develop.

1.1 Introduction

The World Wide Web has grown into a truly worldwide computer-based media network. Previously, most information was, by and large, offered by professional information providers such as companies advertising their products and services, organisations or news services. In addition, private users on the Web had the option of establishing personal homepages as well. However, technological obstacles—complicated tools, lack of infrastructure and technical background knowledge—prevented many users from producing web-pages (e.g., [Lindahl and Blount 2003]).

The infrastructure of successful corporate web-sites often relies on content management systems. These environments take their organisational structures from traditional media such as newspapers or television channels, where authors, editors, an editor-in-chief, etc. are in charge. Hence it can be argued that, to date, the Web has mapped structures from the physical world to the hypertext domain. It has not been able to deliver other qualities than traditional media; the World Wide Web has been far from being interactive, and users rarely had a chance to *participate.*

Recently, however, the Web has undergone changes. Although it has gone partly unnoticed, these transformations are profound as they give ordinary users ability to get more involved in the content creation process. As a result, community-driven initiatives such as blogs, wikis and podcasts have emerged.

This chapter gives an overview of a number of popular community-based technologies and services. The following sub-sections detail the nature of self-organising systems and briefly outline current developments on the Internet. Section 1.2 introduces blogs, a form of web-based journals. The subsequent sections address wikis, types of collaborative content development systems, and two particular instances of wikis: Wikipedia and Wikinews. While blogs and wikis allow communities to generate mainly textual content, podcasting addressed in section 1.6 introduces a way for users to distribute audio content. Section 1.7 discusses file sharing tools including Flickr used by communities for sharing documents such as photos. In section 1.8, social networks as communication and meeting platforms for communities and friends are presented. The last section of this chapter discusses the

impact of the services introduced and gives an outlook on further developments.

1.1.1 Self-Organising Structures

Oftentimes it seems to be necessary to introduce hierarchies in order to make large amounts of data and complex structures manageable and comprehensible. Hence most information systems, as computer-based environments or systems in the engineering sciences in general, have hierarchical structures. Examples are filesystems, web-sites, newsgroups, and programming.

Daily life, however, has numerous examples of systems that do not have a clearly defined hierarchy and follow the rules of organised complexity that may yield emerging structures (e.g., [Johnson 2001]). Examples include the formation of neighbourhoods and cities, the growth of plants, and the natural balance of ecosystems.

Probably the most prominent examples of self-organising systems, however, are ant colonies (see [Gordon 2000; Gordon 1999]). Although every colony has its queen, the term queen is misleading because she is not a leader of the colony. The ant queen lays the eggs, but she does not communicate any particular orders to the workers. They communicate with each other using a vocabulary of about ten to twenty pheromones. In addition to this, they are believed to have a small set of "built-in" rules they follow in order to be able to complete tasks including building a nest, protecting it, and foraging.

Every ant starts the day with a particular task such as collecting food. When many foragers return to the nest with food, ants performing other tasks will also start foraging. ("Today is a good day for collecting food.") When, on the other hand, many foragers return without food, the ants in the nest will continue doing their current work, and foragers will, for instance, go back into the nest and remain idle. ("Today is not a good day for finding food, try again tomorrow.")

Although these simple decisions foragers make might not be ideal, and individual ants make wrong decisions, the large number of ants in colonies assures that decisions are ultimately correct. This can be explained as a variation of the principle of evolution that holds true for a sufficiently large amount of time and a critical mass of individuals (e.g., [Rechenberg 1973]).

Another remarkable aspect of ant colonies is that they change their behaviour over the course of generations. Foragers of a three or four year old colony, for example, are likely to fight over food with ants from neighbouring colonies. In contrast to this, ants of older colonies will steer clear of foreign foragers on their encounters. On subsequent days, foragers will avoid the corresponding region altogether and attempt to find food in different areas.

This change in behaviour is particularly astonishing since, apart from the queen, ants do not live for more than a year. The ant queen, however, is unable to pass any knowledge on to the workers. Although the behaviour of single ants seems to remain unchanged, the entire colony becomes more mature towards the end of its lifetime.

1.1.2 Recent Developments on the Internet

Technological advancements in recent years have yielded systems with entirely new qualities. Although similar applications have existed in the past, the new developments have amalgamated distinct features of their "precursors" and enable entirely novel applications.

The common characteristic, which all these systems share, is that the approach is "bottom-up" rather than "top-down". This means that in these environments content and structure are not determined by professional, corporate information providers. Both content and structure are defined by the individuals of the community. This facilitates self-organisation in these systems and makes the emergence of advanced structures possible. The result is a system where the knowledge of the community is "larger" than the sum of knowledge and experience of all individuals.

For this approach to work, a critical mass (of users in the community) is required (e.g., [Andrus 2005] and [Rechenberg 1973]). This requirement can be seen as an analogy to ant colonies. Individual ants may be unable to build a nest and defend it, may be incapable of providing all fellow-ants with food, etc. With a large number of ants, however, they become more than their sum—they form a community, an ant colony (cf., [Surowiecki 2005]).

1.2 Blogs

Weblogs, often simply called blogs, are web pages that contain newsgroup-like articles in a chronological order with the newest article listed first. Postings to blogs are frequent, typically once a day. They are usually produced by one author or by a small group of authors and are open to the public for reading. Both in style and content, blogs resemble a cross between diaries, newsgroups, newspaper editorials, and hotlists where owners write down information important to them on a regular basis (see [Blood 2002]). Blogs are, however "owned" and maintained by a single person or group of users. They are not open to the public for authoring, and there is no well-defined publishing process as in newspapers (e.g., [Herring et al. 2004]).

Blog entries frequently cite a current event such as a law recently passed, a news story, or the release of a new product. Individuals write comments and their opinion on the event in their blog. Hence, blogs are usually opinionated and reflect the author's views on certain topics.

1.2.1 Types of Blogs and Applications

Basically, two major sorts of weblogs can be distinguished: diaries or personal journals and filters. Journals amount to approximately seventy percent of all blogs, and filters to about ten to fifteen percent ([Herring et al. 2004]).

In the first class of blogs, authors keep readers informed on their work, their social lives, they post holiday photos, etc. The first diary-style weblog believed to have been published was started by Justin Hall, a college student, in January 1994 (e.g., [Pollock 2001]). He employed it to keep people informed about his daily life. Nowadays, for many users, weblogs are a replacement for homepages because they can be used in similar ways but are easier to maintain.

Filters are collections of links to external web-sites that are supplemented with abstracts or brief comments on the contents of the corresponding page. They are usually dedicated to particular topics that can be as diverse as computer hardware, politics, or sports. One of the best-known filter style blogs is Slashdot, a web-site focusing on technology ([Slashdot 2006]).

Slashdot has a large number of authors, and hundreds of new articles are added every day. This was a potential problem for readers because it is hard to find out which articles are interesting, and almost impossible to read all new articles. Therefore Slashdot introduced a rating system: every entry in the blog is rated by readers of the blog. At the same time, readers can choose to have only articles with a certain average rating displayed. Thus, the community of readers determines which articles are significant and hence is capable of establishing a sort of quality control.

Blogging technology is employed in both professional and personal areas. Companies, for instance, make use of weblogs in order to keep employees informed about new products and strategies or on the progress of projects. Furthermore, they are a means to foster cooperation between various departments (see [Treese 2004]). Such blogs are usually only available within the network of the company and not publicly accessible. Authors are frequently project leaders and heads of departments.

1.2.2 The Blogging Community

With several free tools and services available on the Internet, basically anyone can set up their own blogs relatively easily (e.g., [Blogger 2005]). Hence, readers can also comment on other authors' blog articles in their own blogs. This network of more or less loosely interconnected weblogs is called *blogosphere.*

Connections among various blogs are a type of community-building that becomes possible through a set of technologies including permalinks, trackback and RSS feeds (see [Efimova and de Moor 2004]). A *permalink* is a persistent URL to a single posting in a weblog. When author *A* refers in her article a_A to an article a_B by author *B,* a permalink to a_B can be used. If the blogs are *trackback*-enabled, a link from a_B to a_A is appended to a_B. Thus, a_A and a_B are linked bidirectionally, and authors of cited articles are informed about their content being used (cf., [Maurer and Tochtermann 2002] and [Bharat et al. 2001]).

RSS (RDF Site Syndication or Really Simple Syndication, [RDF 2004; RSS 1999; RSS 2001]) is a relatively uncomplicated way for users to find out about the most recent changes on a blog, or a web-site in general, in a given period of time. The *RSS feed* for a site presents a list of changes and additions that typically contains the title of an article, a short summary and the URL to the full entry.

1.2.3 Advantages and Drawbacks of Blogs

Weblogs are an easy way for users to express themselves on the Web and are a valuable tool for companies and organisations to communicate information to employees. Critics, however, claim that they are essentially nothing new (e.g., [Herring et al. 2004]). Hotlists, discussion forums, and "what's new" pages have existed before; however, their usage was more complicated than writing an entry for a blog.

Blogs are sometimes perceived as authoritative works—which they are not. Their contents may be flawed due to a bias. Depending on the purpose of the blog, this can be an advantage or a shortcoming. In systems where a blog is utilised in order to give users the opportunity to comment on articles on a web-site, for instance, opinionated entries can be of value to other readers. Filter-style blogs, on the other hand, offer links to external information complemented with comments. In this case, biased comments are undesirable.

When analysing the blogosphere technological drawbacks of HTTP and HTML become obvious. Permalinks become necessary because it is not possible to identify and locate information at the required level of detail. Since the Web merely implements unidirectional links a technology like "trackback" has to be introduced (cf., [Kappe 1995]). The Web is a passive media that provides content on request; it cannot inform users whenever an existing document is altered

or a new page is added. Therefore RSS is employed in order to notify users of new or modified content.

Blogs are usually not used by themselves but in conjunction with several other technologies ([Efimova and de Moor 2004]). Most frequently, they are combined with e-mail and instant messaging for "out-of-band" communication or wikis (see below). From this perspective, weblogs are a new sort of media that is complemented with various other technologies.

1.3 Wikis

The term *wiki wiki* is Hawaiian for "quick" and reflects Ward Cunningham's intent to create a concept that makes rapid development and organisation of web-pages possible (see [Leuf and Cunningham 2001]). The first wiki was started in 1995 as a collaborative authoring environment (see [WikiWikiWeb 2006]). Wikis in general are self-organising web-sites, where anyone on the Internet can edit existing pages and add new documents any time they wish. This means that every reader can instantly become an author.

This characteristic is interesting because initial authors of articles allow other users to edit "their" content. The fundamental idea behind wikis is that a vast number of users read and edit the content, and therefore errors will be found and corrected. Although modifications to the original article can introduce errors, the principle of evolution determines that in the course of time, after a number of changes, the document will become complete (cf., section 1.1.1).

The aim of wikis is to reach an agreement among the authors. Through the iterations an article undergoes, and the numerous editors, the content is generally agreed upon. For the same reason, wikis tend to be unbiased, which differentiates them from blogs.

1.3.1 Technical Aspects of Wikis

From a technical perspective, a wiki is a web-based content management system (CMS) for generating web-pages that can contain text, images, sound and similar media objects as well as hyperlinks to internal and external resources. Unlike a regular CMS, wikis usually do not contain sophisticated rights management. Thus apart from a few users with administrative privileges, every user in the system has the same permissions.

When the content of a page is modified a wiki-specific source code has to be employed. The complexity of this markup language determines, for example, if the wiki can be used to display tables, mathematical formulas or different fonts. The visual design of the wiki articles is determined by templates that define the placement of the content on the page, the font to be used, etc. When a wiki page is requested by a user the content entered using the wiki-specific markup is translated into HTML code and is inserted into the template. Thus, a conventional HTML page with a pre-defined design is sent to the client.

The articles of a wiki are stored in a database. However, not only the most current versions of articles but their entire history is retained. Therefore wikis inherently provide version tracking, and users can have access to a list of recent changes of a given page. Moreover, the differences to a previous version of an article can be pointed out.

1.3.2 Application Areas

The concept of wikis is applied in numerous fields, from learning environments to documentation systems. Many companies including computer businesses and car manufacturers offer online documentation and help for business customers and consumers. Traditionally, these support databases contain information provided by engineers and customer support. Other valuable information such experiences with the actual product gathered by users are handled in discussion forums.

With wikis, engineers could provide the first version of a product description, and users could modify the initial content and append complementary information when needed. This approach can make documentation systems much more effective since the essential description of a product, for instance, is available in a single document. The information can be kept up-to-date by the user community, and new developments such as the influence of a new operating system on an existing software product can be dealt with potentially more effectively. Discussions about topics such as the usability of the product can take place in forums or blogs attached to articles in the wiki.

Another area in which wikis are successfully applied are "knowledge bases" used in companies and organisations for internal communication and documentation. In such repositories information needed for doing everyday business but also information on competitive products, on the use of technologies, etc. can be retained. With wikis not only a small group of editors but everyone can contribute even small portions of information to articles in an uncomplicated way. When a programmer, for instance, finds a more effective solution to a problem, she can add it to the corresponding article in the wiki, and the result is available to other programmers immediately.

A research project conducted by the CIA suggests a similar concept (see [Andrus 2005; Tomlin 2005]). Intelligence officers collecting data could insert their information into a wiki and thus make it available to the entire organisation. Since an editor does not have to approve the content, the information can be offered

faster, and actions such as re-structuring of articles can be performed more easily.

Many state-of-the-art learner-support systems make extensive use of digital libraries. In most cases, the information in these libraries is authored by teachers, lecturers, and professional information providers. More recent projects, however, rely heavily on the students who generate content as part of their homework or lecture. Although the first version of the digital library will most likely be rather imperfect, subsequent versions—after a few iterations, after a few semesters—will become more and more complete. An example is a wiki that is used by university students of structural engineering to create an online library of lecture material for reinforced concrete construction science ([Ebner and Zechner 2005]).

Further examples for wikis are Wikipedia and Wikinews presented in sections 1.4 and 1.5. Wikipedia, by far the largest wiki in use today, is of special interest because phenomena of large-scale communities can be observed (source: [Wikimedia 2006a]).

1.3.3 Benefits and Shortcomings of Wikis

Similar to ant colonies, wikis are self-organising systems with a large number of individuals at work (cf. section 1.1.1). As ant colonies manage to succeed in tasks such as foraging and building nests, wiki communities can successfully author content and create organisational structures.

The fact that basically everyone on the Internet can contribute to wikis in an uncomplicated way makes them more flexible than static editor-based web-sites. Content can be created and published by users easily and, unlike regular web-sites, without profound technical background. In addition, features such as easy and open access as well as version control make them particularly well suited for collaborative work.

Users being both readers and authors at the same time is one of the strengths of wikis but also one of their major drawbacks. Although the wiki concept makes the development of content highly flexible and a system versatile, it makes maintaining high quality standards for entire wikis almost impossible. Since basically everyone with access to a wiki can modify its content, the credibility of the information provided can be questioned. Users might inadvertently add incorrect information to a page in a wiki, and readers might mistake the content provided for reliable. Vandalism is also a problem experienced in wikis: wrong data, advertisements, defamatory content are inserted deliberately, existing content is deleted, etc. Although acts of vandalism are usually found and corrected relatively fast they are pestering communities and can impede the authors' motivation to contribute to the system.

After a period of time and several evolutionary cycles, single articles in wikis usually become authoritative, and their level of accuracy and completeness is high. This does, however, not mean that the wiki as a whole becomes authoritative, which might be confusing especially for users not fully aware of the wiki concept. Further issues related to the wiki concept are discussed in section 1.4.2.

1.3.4 Counter Measures

The open access to the content in wikis is one of their disadvantages, particularly when the information presented is critical. Therefore a distinct form of wikis with various levels of permissions is proposed, while the core features of the concept remain unaltered (cf., chapter 2).

For users on the highest level, the "hierarchical wiki" can be used in exactly the same way as traditional wikis, i.e., these users can edit any section in any article. Users on lower levels, however, may only modify parts authored by users on the same level or on lower levels but not portions of text, for instance, written by users on the highest level. Users on the lowest level can only edit information initially authored by users of this level.

The key for making such a system successful is to restrict the majority of all users to the lowest level. While a few users in higher level groups supply essential information, start new articles, and work actively on maintaining quality standards, the large mass of users work only on the actual development of content.

This approach makes it possible to have the most important information provided by high level users, whereas supplementary information is authored by writers on lower levels. An example is a product-related wiki, where engineers as high-level users publish a user's manual, and consumers add their experiences or information on the use of the product. Although consumers can add new information and can alter pieces of content authored by other low-level users, they are not capable of modifying the content provided by the engineers.

Potential further application areas for hierarchical wikis include the communities around intelligence agencies (see [Andrus 2005]). One of the main processes in these environments is gathering pieces of information. As soon as new information becomes available it can, for instance, be inserted into an existing article in a wiki by the officer of the intelligence community. While the majority of the article can be edited and viewed by everyone of the community, certain parts might be confidential and only be viewed by members of higher levels. Yet other sections such as the "core" of the article might only be edited by users on the highest level.

1.4 Wikipedia

The tradition of trading knowledge in the form of professionally authored encyclopaedias goes back to the 17th and 18th centuries. This is in stark contrast to Wikipedia, where articles are neither written by acknowledged experts nor are they reviewed by editors. Wikipedia could be coined "the people's encyclopaedia". It is a free, wiki-based encyclopaedia that anyone can edit. Every user is reader, author, and editor at the same time ([Wikipedia 2006a]).

The success of Wikipedia builds on the tight involvement of the users, the sense of the community, and a dedication to developing a knowledge repository of unprecedented breadth and depth. The project is growing rapidly: from its founding in 2001 until October 2006, Wikipedia has been established in more than 250 languages with more than five million articles. The largest editions are the English one with about 1.5 million articles and the German one with almost 500,000 articles (source: [Wikipedia 2006b]).

1.4.1 Advantages of Wikipedia

The concept and architecture of Wikipedia make it much more flexible than a print version or an edited online version of an encyclopaedia. When events like the Olympic games, for example, take place the most current results are published just minutes after they become available. These articles are sometimes even complemented with tables, photos and links to external resources.

The main argument against the Wikipedia project is that with an open editing process the correctness of the information provided cannot be guaranteed. However, a recent study conducted by Nature shows that, in terms of accuracy, science articles in Wikipedia are fully comparable with corresponding information in the Encyclopædia Britannica (see [Giles 2005; Britannica 2005]). Both encyclopaedias contain a number of misinterpretations of concepts, omissions, misleading statements, and factual errors.

Wikipedia is probably doing comparatively well because it endorses guidelines to ensure that articles are written in an objective and unbiased way. One of the main policies for writers is the "neutral point of view". It urges authors to write content from a neutral perspective so that "ideas and facts [are presented] in such a fashion that both supporters and opponents can agree" (from [Wikipedia 2006c]).

1.4.2 Drawbacks of the Approach

Despite all guidelines, the concept of Wikipedia is prone to a number of complications. This section presents selected problems that can be observed frequently (see also [Denning et al. 2005]). They are most likely not only to be encountered in Wikipedia but in any wiki with many articles and users.

1.4.2.1 Quality and Authority

Encyclopaedias and dictionaries are typically reference works. They are used by researchers, librarians, students, journalists, etc. in order to obtain precise definitions and explanations. Since articles in Wikipedia are written by a large number of users, and currently mechanisms to approve the expertise of authors or to verify the reliability of content do not exist, the quality in Wikipedia is not equal for all articles. Therefore it can be precarious to use Wikipedia as a sole source of reference. An editor of the New York Times has even warned the journalists of the newspaper to use Wikipedia with caution (see [Ingrassia 2005]).

The Wikipedia initiative is aware of the problem of a lack of quality, but instead of having articles approved by experts, a peer review and rating mechanism is favoured (e.g., [Giles 2005]). The "article validation feature", announced for January 2006 but not yet implemented, would allow users of the encyclopaedia to assess the quality of articles (see [Wales 2005; Wikipedia 2006d; Wikimedia 2006b]). Since the mean value of all ratings for a given document version indicates its quality, it will be easier for readers to judge whether to trust the information provided by Wikipedia or not.

1.4.2.2 Background and Balance

Wikipedia has several policies in place to ensure, for instance, that articles are unbiased (see section 1.4.1). However, even if an article is written in compliance with the "neutral point of view" the varying cultural, social, national and lingual backgrounds can have an enormous influence. Hence, content in Wikipedia can only be as professional and balanced as its authors and their demography are.

On February 5th, 2005, the English article on the American chess player Paul Morphy, for example, had 5,466 words, contained a photo, citations and references to external resources. In contrast to this, the German version consisted of only 290 words and did not offer any additional information. This example shows, on the one hand, that Paul Morphy is an important person for Americans. On the other hand, it distorts reality and creates an imbalance in that it emphasises "local heroes".

Edited encyclopaedias meet imbalance and bias by introducing guidelines for the creation of content. The length of articles or the number of references to external sources, for instance, might be limited. These measures make it possible to create articles in a given

"class" with the same structure and similar volume (e.g., [Korica et al. 2006]).

In Wikipedia, disproportionate weight is given to topics such as controversial scientific matters, disputed hypotheses, science fiction, and conspiracy theories.

1.4.2.3 Sensitive Information and Privacy

Incorrect information in Wikipedia articles is particularly problematic when sensitive information is covered. A recent, startling example is the case of John Seigenthaler. An anonymous user published a biography for Seigenthaler on Wikipedia. It related him to the assassinations of John F. Kennedy and his brother and accused Seigenthaler of collaboration with the Soviet Union in the 1970s (see [Seigenthaler 2005]). An intense discussion followed and as a consequence of the Seigenthaler case, the creation of new articles is restricted to registered users only, i.e., anonymous users are not able to start new articles. Becoming a registered user is, however, not complicated, being only a matter of several minutes.

Although defamatory content can cause much harm, sometimes incomplete articles and articles with deliberate omissions are just as bad. If an article states that an author has written books *A* and *B* but does not mention that the same author has written another five books it leaves the impression that only two books were written. In some cases, however, correct but controversial information is published, and the person concerned wishes to have it corrected or removed. Daniel Brandt, privacy activist and critical of Wikipedia, wanted to have the Wikipedia article on him removed (see [Brandt 2005]). After a lengthy, sometimes provocative discussion, Daniel Brandt was blocked from using Wikipedia, and his article was not removed.

1.4.2.4 Wiki-Related Phenomena

Malicious modifications of articles, including the deletion of information, appending incorrect or inappropriate content, insertion of vulgarities and the insertion of advertisements, happens occasionally in Wikipedia. Research shows that these acts of vandalism are often repaired within only a few minutes after they occur (see [Viégas et al. 2004]). Spiteful deletions, in particular, are reverted very fast.

Another example of something that is common to wikis in general are "edit wars": a number of paragraphs of articles are repeatedly inserted and deleted or modified and reverted by two users or groups of users. Most likely this is a social problem, where two parties are unable to reach a consensus over a piece of content. Usually such a dispute is ended by a democratic vote that is attached to the article.

1.4.2.5 Awareness of the Concept

Although in theory everybody can edit articles, only a small percentage of users actually do—even though they probably know that the content is incorrect or incomplete. Some users might not even be aware that Wikipedia is not an edited work and that basically every reader can edit the content provided. This is true especially for users that are relatively new to the Internet and are directed to Wikipedia by search engines such as Google.

If users are not acquainted with the underlying concepts of Wikipedia they do not know that the content may not be authoritative. However, even if users do not rely solely on the information provided by Wikipedia and do consult other resources, the content provided might be identical because several services including Answers.com retrieve information from Wikipedia ([Answers 2006]). Thus a situation could occur where misinformation originating from Wikipedia is used as a basis for a new piece of work, utilised by Wikipedia authors to argue for the incorrect information in the encyclopaedia.

1.4.2.6 Volume of Wikipedia

The number of articles is not necessarily a yardstick for the completeness of the encyclopaedia. The English language edition of Wikipedia with approximately 2.5 million articles has far more articles than the Encyclopædia Britannica with about 120,000 articles ([Britannica 2006]), but it also contains many articles about movies, rock groups, etc. These kinds of articles are not usually part of a general encyclopaedia but of more specific works. Although it can be seen as an advantage that detailed articles on a very wide range of topics are present in a single encyclopaedia, it is sometimes cumbersome (cf., section 1.4.2.2). In a general encyclopaedia an article on The Beatles, for instance, is not expected to exceed more than a few paragraphs in length. In Wikipedia, however, the corresponding article is several printed pages long and includes a complete discography, a history and a set of photos of the band, etc. Thus, the Wikipedia article might be better suited for a specialised encyclopaedia on pop music.

These differences make it hard to compare Wikipedia to traditional encyclopaedias. On the one hand, the topics covered by Wikipedia vary greatly, and it might have to be compared to a set of specialised encyclopaedias. On the other hand, articles in Wikipedia are sometimes much longer and more detailed than corresponding information in a conventional encyclopaedia or dictionary. It can be believed that Wikipedia is likely to become a new type of encyclopaedia incommensurable to existing ones.

1.5 Wikinews

Most community-based news services on the Web reverse the order of the traditional publishing process. In conventional publishing, a board of editors select a set of stories from the vast amount of information produced every day. The number of stories is usually determined by the volume of the newspaper, by the time available for a TV or radio programme, etc. By contrast, community-based news providers make every piece of news accessible, and filtering techniques such as filter-style blogs are employed to present only relevant articles to the consumers.

1.5.1 Goals of Wikinews

In November 2004, Wikinews, a community-based, participatory news project linked to Wikipedia, was started ([Wikinews 2006a]). Wikinews is not only a news provider but a journalistic service. The aim is to publish complete news stories and to act as a counterpart to commercial news agencies such as Reuters and United Press or as competitor to other independent new services like Korean community-based OhmyNews (see [Wikinews 2006b] and [OhmyNews 2006]). Neither does Wikinews offer only news headlines with short abstracts like Slashdot does, nor is it restricted to a specific topic or does it present an opinion in its articles as services such as Indymedia do (see [Slashdot 2006] and [Indymedia 2006]). Instead, Wikinews articles are written conforming to Wikipedia's "neutral point of view" guideline.

As a news service where everyone can contribute information, it has the potential to have an impact on the information made available to consumers. Content that might not be relevant enough to be presented by large news providers or information deliberately suppressed by mainstream media can still be made available on Wikinews. Especially in countries where freedom of speech and freedom of press are restricted Wikinews can become an important service.

1.5.2 The Relationship between Wikinews and Wikipedia

Both Wikinews and Wikipedia build on the same concept and infrastructure, and both share the same benefits and disadvantages. Wikinews however takes a different approach to the publication of information. While Wikipedia articles are usually open for editing any time, stories in Wikinews are set "read-only" after editing has been completed and their content has been approved by the community. After editing has finished, a permanent and stable version of the articles is archived in the system.

This means that the convergence criteria applied in Wikipedia is not valid for Wikinews. Wikipedia articles are typically long-lived, therefore the probability to achieve completeness and accuracy is higher because the more time available, the more readers will access an article, and the more likely it is that errors will be corrected. Such an approach cannot be taken for news sites since news need to be published quickly—otherwise they will be obsolete. Therefore information edited on Wikinews has to reach maturity rapidly, which is not always successful and sometimes results in rather short articles.

1.5.3 Success of Wikinews

Up to the October 28th, 2006, 7,000 articles were published on the English edition of Wikinews, i.e., on average about nine or ten articles were produced per day (source: [Wikinews 2006c]). This makes Wikinews not nearly as successful as Wikipedia. One of the reasons might be the directive to write articles from a neutral perspective, without bias and opinion. This makes Wikinews rather monotonous to read. Also articles are often collections of news and different views on a given topic gathered from various mainstream news providers (see [Bruns 2005a]).

One reason for the limited success of Wikinews might be the fact that there is no way of commenting on news articles within the Wikinews system. Although it is possible to attach discussions to Wikinews articles ("talk pages") these postings are strictly confined to discussing details on authoring. Once a Wikinews article is completed, however, users cannot debate its content or add complementary information. In contrast to this, the popularity of blogs and Slashdot-like news services is *based on* the comments added by readers. In these systems, views on a news article shared by readers are sometimes more enlightening and more important than the actual news item because they can offer a different perspective on the story, details on the topic, related information, etc. This type of commentary is not permitted in Wikinews. As one of the administrators of the system explains, "It's deliberate – opinion or commentary is banned. There are enough blogs already." (from: [Bruns 2005b]).

1.6 Podcasts

The expression *podcasting* is a combination of two terms: iPod, a popular MP3 player, and broadcasting. The word appropriately describes the nature of podcasts. On the one hand, they offer audio content that can be listened to on demand—like music on an MP3 player. On the other hand, it is a system that provides content resembling radio programmes.

Podcasting basically means blogging audio content, where the content producers post audio content regularly on a server in the MP3 audio format (just as us-

ers post short articles on blogs). In a fashion similar to readers using RSS feeds to stay informed on the most recent articles on a blog, podcasting allows users to subscribe to various audio content producers. Each podcast offers a list of audio clips that are available for download complemented with metadata such as a brief description of the actual content. By subscribing to several podcasts, users are able to accumulate material from numerous sources. The content, however, is only retrieved on the users' request, hence podcasting can be seen as an "audio on demand" service (e.g., [Biever 2005]).

Topics covered by podcasts range from music and cultural programmes, mainstream entertainment, business, politics, science, technology, and travel to religious programmes. Podcasts are typically either person-centred or dedicated to specific topics. "Personal" podcasts are usually produced and published by a single person and offer the person's views on various subjects, present the person's favourite music, etc. Podcasts geared to particular topics are often created by a small group of users and contain a selection of separate "stories". Examples are news programmes, regular discussions on political topics, or science-related shows such as the Nature podcast (see [Nature 2006]).

1.6.1 Use of Podcasts and Application Areas

In September 2004, the concept of podcasting started to take off. The initial idea was to offer anyone on the Internet a platform for publishing their own radio show. Soon a large number of amateur shows emerged, one of the favourite shows being *"Daily Source Code"* by Adam Curry, one of the earliest adopters of podcasting (e.g., [Pod411 2006; PodStats 2006]). Although the target group were amateur users, and the largest proportion of podcasted content is still produced by amateurs, the technology was soon also employed by professional content providers.

Nowadays podcasting is, for instance, applied in education in order to enable distance learning or simply to provide the possibility of listening to a lecture again (e.g., [Downes 2005]). Even organisations such as Duke University or the Washington College of Law fully endorse podcasting technology. Both universities make a range of content, including lectures and discussions, freely available in podcast format (see [DukeCast 2006; WCL 2006]). Moreover, podcasts can also be offered as supplementary material to the proceedings at conferences. Two workshops of the IEEE Symposium on High Performance Interconnects, for example, can be downloaded freely as podcasts (e.g., [HotI 2006; IEEE 2006]).

1.6.2 Similar Technologies

The idea of publishing audio content using blogging technology can also be applied to other types of media such as photos or video content. With *photocasting,* for example, users can share and distribute their photos using an RSS feed. This enables uses such as photos diaries or sharing entire photo albums with friends on the Internet. With the required functionality being included in popular applications such as Apple's iPhoto, photocasting can be expected to become a fashionable technology among users (see [iPhoto 2006]).

Videocasting, sometimes called *vodcasting,* applies the blogging concept to video content. Vodcasting is, in fact, an acronym, where "vod" stands for "video on demand". With vodcasting, content producers can create video clips and inform users about new episodes using RSS feeds. Consumers subscribing to a vodcast have access to a list of video clips that can be played at the users' request. Vodcasting can include both downloadable video files and content streamed from a streaming video server.

The technology receives attention from various business areas. Recently, German car manufacturer BMW, for example, made a videocast available for presenting new products and for disseminating interviews (see [BMW 2006]).

1.7 File Sharing Tools

For most readers, maybe, file sharing has a negative connotation. It is often synonymous with downloading music and movies illegally from the Internet and with distributing pirated digital content. Napster and Kazaa are popular examples for tools that let users share files (mostly illegally) over the Internet. However, lately also legal peer-to-peer file sharing networks have evolved (see [Rodriguez et al. 2005]). The BBC, for instance, has started a service, based on file sharing technology, for the legal dissemination of TV shows (see [BBC 2006a]).

Recently, a new type of file sharing has emerged. This class of systems are web-based, offer users a private space to store their documents and a public space for sharing files with other users, as well as helping them to organise their information. The prime example of such an application is Flickr, a portal for managing and sharing photos (see [Flickr 2006]).

Flickr lets users store, organise, and share photos. Users can upload their photos to a server, add comments and leave notes inside pictures. The key element in the system, however, are arbitrary tags attached to photos (e.g., [Weiss 2005], see also chapter 8). These tags represent loose metadata and are utilised to describe the content of the photo. A photo depicting a tree can, for instance, have the tags "tree", "my holidays in Iceland", and "winter". When users search

Flickr for "winter", the photo of the tree is part of the results. Users can also browse the photos in Flickr. For every photo displayed the tags defined by the author are shown. Instantly, users can have all images in the same "category" presented (i.e., pictures with the same tag).

Flickr is a self-organising community where the system does not tell users how to tag their photos or impose any structures on the organisation of content. This approach is in contrast to the conventional way of generating metadata. In traditional "editor-based" systems, professionals assign metadata based on a well-defined taxonomy and a set of guidelines (e.g., [Mathes 2004]). In Flickr, however, the choice of tags is entirely up to the user. Although this concept lowers the barriers to entry and is a major incentive for people to store their content and metadata in the system, it raises the problem of ambiguity (see chapter 8). Since there are no formal taxonomies users can use ambiguous terms and synonyms when tagging photos. The tag "apple", for example, can stand for the fruit or the computer manufacturer. On the other hand, there are several synonyms for Apple computers including macintosh and mac. As these examples illustrate, the free-form taxonomy can sometimes make it difficult to find the desired content.

The content in Flickr is largely published under a Creative Commons license (see [CC 2006]). With this type of license content is freely available while protecting the owners' copyrights. Therefore Flickr is an increasingly attractive resource for web designers, publishers, etc.

1.8 Social Networks

In 1967, American psychologist Stanley Milgram conducted the "small world experiment", in which he sent letters to sixty volunteers in Kansas and asked them to forward the envelopes to a specific person in Massachusetts—by hand and through friends or friends of friends. The letters that reached the addressee were, on average, relayed by five to seven people. This is seen as an empirical proof that arbitrary people in our society are related to each other through friends and friends of friends (see [Milgram 1967]).

The small world hypothesis based on Milgram's findings states that the number of personal acquaintances needed to connect two random persons on the planet is small. The hypothesis led to the expression "the six degrees of separation", meaning that any two random persons are associated with each other by a chain of about six individuals. The "six degrees of separation" is one of the underlying concepts of social networks on the Internet.

Social networking services offer friends a space where they can maintain their relationships, chat with each other and share information. Moreover, they offer the opportunity to build new relationships through existing friends. On the first use of the system, users are required to submit a profile containing personal information such as their name, date of birth, and a photo. The personal information is made available to other users of the system, and is used to identify friends on the network and to add them to a list of contacts. In most systems, users can not only view their friends but also second degree friends (friends of their friends). Some networks follow an "invitation only" approach. Hence, every person in the system is automatically connected to at least one other person.

Examples for common social networks are Friendster with about 21 million users (November 2005, [Friendster 2006]), MySpace with about 105 million users (August 2006, [MySpace 2006]), and Google's Orkut with about 21 million users (January 2006, see [Orkut 2006]). In addition to these general-purpose networks, specialised services have evolved in order to establish a community of like-minded individuals. OpenBC, for instance, is a professional networking service that attempts to create a web of trusted experts and business partners (see [OpenBC 2006]).

1.8.1 Large Communities

When communities grow larger, self-organisation tendencies emanate, and frequently sub-communities covering more specific topics or smaller groups of friends are established. Several services including Orkut facilitate creating new sub-groups as a core functionality of the system. In these smaller communities users chat, have lively discussions in dedicated forums, and exchange pictures and other documents (e.g., [O'Murchu et al. 2004]).

The formation of smaller groups within a large collective can probably be described with the rule of 150. This axiom refers to the social channel capacity, the ability of the human brain to relate factual, emotional, and social details to people. A series of social studies show that the average person can remember these features for approximately 150 individuals (see [Dunbar 1993]). Psychologists explain this characteristic by using the evolution of human societies: early settlements did not comprise more than 100-150 people, and therefore the brain developed only to the point where it was able to store the information on all people in this social network. Thus, a "genuine" social network is limited to about 150 people.

1.8.2 Technical Aspects

Most popular social networks in use ask users explicitly for personal information (e.g., [Adamic et al. 2003]). Hence users fill out profiles and provide personal data

as well details on their likes and dislikes. As mentioned above, users add their friends manually to the list of contacts. So the social network is generated manually, which usually results in a high accuracy of the connections made.

A system that forms a large social network without the users' explicitly knowing it, although users provide the required information voluntarily, is Skype. Skype is a provider of free internet telephony (see [Skype 2006]). Every user in the system has a user profile that can contain the name, address, phone number, e-mail address, a photo, etc. When person A wants to call person B, usually the profile of person B is added to the contacts of person A. Calling a person is a strong indication of a personal or professional relationship. Thus, the information stored in Skype represents a large, manually generated social network.

An alternative approach to manual generation relies on fully automatic creation of networks. E-mails of a group of users, their postings in newsgroups and blogs, links on their homepages and similar resources are analysed. An e-mail from user A to user B, for instance, indicates a connection between users A and B. In the same way, a follow-up by user B to a newsgroup posting by user A can be interpreted as a (weak) relation between the two users (e.g., [Kautz et al. 1997]). All connections detected by the generative algorithm are accumulated and utilised to form a graph of weighted edges between "user nodes". Edges whose weight is over a given threshold correspond to the connections in the network to be generated. The advantage of this method is that it does not require user interaction. Moreover, it is capable of unveiling connections that might otherwise have remained implicit or hidden. The drawback is, however, that automatic generation of the network cannot be as precise as manually adding contacts. Furthermore, a fully automated process is usually not able to collect the personal data provided by users.

Automatic generation of a type of "social networks" is also possible for services such as eBay or Amazon. In eBay, for instance, information is retained on who bought from whom, which buyer rated which seller, etc. This information can be used to generate a network of weighted connections, where the weight depends on positive, neutral or negative ratings between buyers and sellers. In Amazon, users' buying a book, writing a review or giving a recommendation for a book imply that they have an interest in the author or the topic. Although this data does not form a traditional social network, it can be interpreted as a social structure in the broader sense. On the one hand, clusters of users with similar interests are formed, and clusters and users are connected with each other; a friendly contact and direct communication among users is, however, not possible.

1.8.3 Use of Social Networks

The obvious aim of social networks is to give users a way to stay in touch with friends, colleagues, and acquaintances. Services such as OpenBC also let users browse through their contacts and second degree contacts (contacts of contacts). Additionally, users in OpenBC can search for people with certain skills or other attributes. When an appropriate person is found, the chain of contacts to this person is displayed. Thus, users can, for example, ask their friends and friends' friends on the person's qualification.

Potentially, one of the biggest application areas of social networks might be personalised searching on the World Wide Web (e.g., [Freyne and Smyth 2004]). Whereas today's search engines provide largely anonymous or uncredited information, future versions might highlight or recommend web pages created by recognised or familiar individuals. The integration of search engines and social networks could also enable queries such as *"Has any of my acquaintances been on holidays in New Zealand?"* or *"Recent articles on hypertext authored by people associated with Ted Nelson"*.

It should be noted that real concerns regarding the privacy of members of social networks exist. Information on consumers that privacy activists have been trying to protect from companies are nowadays provided willingly by inexperienced users. The detailed personal information stored in user profiles, for instance, could be utilised to disseminate personalised fraudulent advertisements, automatically sign users up to services matching their profiles or even sell the personal data to third parties. Moreover, the service providers have the ability to monitor and store the information communicated among users and make use of ideas expressed and data transferred during users' discussions (e.g., [Orlowski 2004]).

1.8.4 Other Community-Based Networks

Although not directly associated with social networks, this section introduces three community-based networking services: del.icio.us, Furl, and Eurekster. del.icio.us is a social bookmarking and classification service that enables collecting and sharing favourite web pages (see [del.icio.us 2006]). Users can add bookmarks of web pages to del.icio.us, attach tags or keywords and choose if they are to be publicly available or private. The keywords assigned by users are used for non-hierarchical categorisation of the bookmarks. Hence, clusters of bookmarks for various topics are created in the system. When users access a bookmark, they can also look at the public bookmarks of users that have the same web page in their portfolio. Moreover, users can search for "similar" pages—bookmarks that share certain keywords or are in the same bookmark cluster (e.g., [Millen et al. 2005]).

Furl, a similar service, takes the concept a step further and stores bookmarked articles in an internal database (see [Furl 2006]). Thus, users can create their own "Personal Web" that only contains the pages they store. As in del.icio.us, pages can be private or publicly accessible. In addition to keywords users can also assign topics, give ratings and attach comments to pages. Furl also creates an index of all documents stored in its database and offers full-text search functionality.

The social bookmarking concept is not only used by private users but also in companies, by researchers, and in computer-supported learning environments (cf., [Ham 2005]). Some researchers, for example, organise their references to resources on the Web in social bookmarking services and make it also available to other researchers. This can significantly reduce the time required for finding scientific publications and similar literature. Social bookmarking can also enhance learning environments (e.g., [Vuorikari 2005]). A lecturer, for instance, makes a list of links to resources available. By browsing the associated metadata students can, for example, find similar pages on the Internet and, therefore, get a more comprehensive view of the topic.

Both Wikipedia and services like del.icio.us are employed by some users as an alternative to conventional search engines such as Google. Wikipedia is a good starting point for many topics, since it can give an overview of a topic and frequently offers manually selected links to more detailed resources. Similarly, a query in del.icio.us yields a number of web pages that have been selected by users as one of their favourite pages on the Web. Although Google's search and ranking algorithms are very sophisticated and mostly offer relevant results first, they are currently unable to offer documents that were evaluated and chosen by individuals.

Eurekster is a collaborative search engine whose concept is a blend of social networking and social bookmarking (see [Eurekster 2006]). People sign up to the system and form communities of users with similar interests. When a user searches the Web, information on the query and the documents actually chosen from the result are stored in the system. These data are used in order to introduce a prioritisation of topics within the community and perform a ranking of relevant articles within a topic. Thus, the system eventually "knows" which topics and web pages are relevant for a community. A user part of a community of archaeologists searching for adobe, for instance, might be confronted with results including historic sites in Peru. By contrast, in a technology centred community documents on the computer software company Adobe might be the result of the query (e.g., [Freyne and Smyth 2004]).

1.9 Summary and Outlook

During the last few years, new forms of content generation and organisation on the World Wide Web have emerged. Services such as blogs, wikis and podcasting give users the opportunity to become authors and to express themselves. For the first time, even users lacking the knowledge of the underlying technologies can participate in contributing content to the Web. In a way, these new services have finally brought a form of democracy to the Internet, and the traditional distinction between content producers and consumer is blurred (e.g., [Miller 2005]).

With these new technologies, the flow of content is no longer strictly "top-down", from classic producers to readers, but an increasing number of users become writers and contribute new content (e.g., [Lenhart et al. 2004]). Thus, a new "bottom-up" movement can be observed—consumers start producing information that is distributed among other users until it is picked up by mainstream media. The aspect that makes such an approach work is the critical mass of users that allows self-organisation to take place (e.g., [Johnson 2001]). This resembles ant colonies (see section 1.1.1) when they are, for example, building nests: while single ants can only contribute small pieces, the collective establishes an extremely complex and efficient structure (see [Gordon 2000]). In analogy, new services support individuals in making their knowledge explicit and help collective intelligence unfold.

The novel forms of content development have sparked a "revolution" across all types of media. Classic web-sites are complemented with, or even replaced by, wikis; services such as Wikinews and blogs offer an alternative to conventional news providers and commentary; traditional knowledge repositories are challenged by Wikipedia; and radio broadcasting is supplemented by podcasting. The next logical step seems to be "video-blogging" as an approach to the creation and distribution of television shows.

The community-based types of media introduced in this chapter allow for alternative perspectives and views that are not suitable for traditional media. Furthermore, topics that are inappropriate for the mass of users served by broadcast media, as well as news that are possibly not relevant enough for the majority of consumers can be addressed by the new services.

1.9.1 Major Transformations

The new services that have recently emerged have indeed spawned a series of transformations on the Web. The transformation, however, is not only based on *technological* changes, but more importantly on a fundamental mind shift. The aspects that Web communities (such as blogs, wikis, file sharing and social networks, bookmarking services and podcasting groups) have in

common are user participation and openness. Basically every user on the Internet can start new blogs, can readily produce podcasts, and can edit the content in wikis. So, the attitude of users has changed insofar as they now enthusiastically make the information they produce available to the public (see [O'Reilly 2005]). In addition to this, even companies make their content repositories publicly accessible and enable new and sometimes unconventional uses of existing data: the content made available by the BBC, for instance, is used in a dictionary of English phrases (e.g., [BBC 2006b]). To put it a different way, the services recently developed on the Web are based on "an attitude not a technology" (from [Davis 2005]).

Although it is not within the scope of this chapter, it should be mentioned that the example of the BBC making available content which is reused for implementing new applications is a movement that is becoming increasingly popular. Such hybrid applications, or "mash-ups", make use of one or more content providers and employ their data for creating new applications and services. By combining the best of various existing services, these mash-ups are sometimes far more useful than the original applications they "inherit" from.

Apart from making the Web more democratic and enabling user participation, the community-based services have opened up entirely new opportunities. Wikis, for example, have the potential to alter the way collaboration among users and groups happens. It is no longer necessary to send text documents as e-mail attachments or to employ an expensive groupware solution in order to enable collaborative work on a common body of content. In similar ways, blogs make it possible for users to utilise the Web to express their views—without having to purchase web authoring software or to get acquainted with hypertext technologies.

1.9.2 Opportunities and Future Trends

Wikis including Wikipedia, blogs, podcasting, file sharing and similar techniques can react faster to recent events and new developments than conventional infrastructures. When an event happens, it can be published instantly on a blog or in a wiki. In contrast, a traditional news service article has to undergo fact-checking and an editorial process prior to publication. Articles in Wikipedia are often updated only minutes after new information becomes available. For instance, only shortly after the spacecraft for the Nasa mission to Pluto was launched photos and other details were included in Wikipedia articles. In contrast to this, a classic encyclopaedia requires an editorial cycle (usually at least a year) in order to incorporate such information.

With the tools and services at hand, users become more independent from classic information providers. Therefore in the future, probably a smaller percentage of information will be authored by professional editors, and distributed by (media) companies. Moreover, new structures might become mainstream: wikis, blogs and podcasts are the environments that produce content. When the authors' names are known they can be looked up in social networks in an attempt to verify their expertise. Finally, social bookmarking services and filter-style blogs are utilised as aggregators and filters in order to offer a balanced selection of reliable information. Hence, individuals as well as large organisations have the potential to establish a network of trust, where information can be accountable to users.

Moreover, systems which rely on a large user community can facilitate the "accidental" encountering of new information. Although environments such as learner support systems or digital libraries explicitly include functionality that enables accidental information encountering (e.g., [Marshall and Bly 2004]), community-based systems provide this feature intrinsically. Examples are Flickr and del.icio.us (see above). Since users in these large communities have varying opinions and interests, they are likely to access diverse resources on the Internet. The information they gather from these contrasting sources can easily be made available within their communities.

1.9.3 Challenges and Concerns

The changes the Web is undergoing raise a number of concerns. Most can be clearly observed in very large community-based environments such as Wikipedia. One of the most problematic issues is the lack of accuracy and, connected with this, the lack of accountability. Several evolutionary cycles are required to make information accurate and complete, especially in wikis. In addition, both blogs and wikis do not have the means to indicate the completeness and correctness of articles, which makes it difficult for users to judge the content provided.

Moreover, in most community-based systems it is not a requirement to provide a real name when authoring content. Authors can usually hide behind self-assigned synonyms, or only their IP addresses are shown (as for anonymous authors in Wikipedia, see also chapter 6). Therefore it becomes almost impossible for average readers to find out who the content authors are, and even simple enquiries such as asking for the source of a quotation might be impossible.

Despite the advantages that new technologies have, readers have to learn how to deal with the new media. Users have to get used to the fact that not everything published on the World Wide Web is true and that it is necessary to find at least another, independent source

that corroborates the initial document. Visitors have to realise that the same process is even more relevant where content is authored by numerous, potentially anonymous users.

1.9.4 Technological Aspects

As indicated above, the technologies introduced in combination with community-based services make it clear that the design of the Web does not allow for these types of interaction per se. There are no bidirectional hyperlinks, therefore a technology like trackback has to be used. The URI and URL scheme and the composition of documents on the Web do not permit to identify and locate an exact portion of content. Hence permalinks have to be employed. The implementation of the Web does not consider a notification mechanism for updated or new documents, which makes a method like RSS necessary. Version management is not part of the Web, and so services such as wikis have to implement version tracking systems, which results in incompatible implementations. Furthermore, content is regularly duplicated in order to be able to quote portions of the original document. By duplicating instead of virtually including content from the original resource, both the context and the reference to the source document are lost.

Although the new services seem to require new technologies, they underline the shortcomings of the Web. In the 1960s, Ted Nelson presented the concept of a hypertext system that supported multidirectional links, identification and location of content on the level of single characters, notification techniques, and the virtual inclusion of remote documents (see also chapters 3 and 4). The environment allowed for collaborative authoring, various levels of access to documents, and had versioning functionality built-in (see [Nelson 1981]). Since then systems offering similar functionality as Xanadu have been implemented (e.g., [Maurer 1996; Andrews et al. 1995]). The technologies were, however, not included into the infrastructure of the Web.

1.10 Conclusion

"From chaos comes order" is an expression accredited to chaos theory. It can also be applied to the services introduced in this chapter. Although the concept of wikis, for instance, might seem utterly chaotic, Wikipedia is the principal example that such an anarchic system can yield structure and to a certain extent even high quality content. The new and successful Web services range from free encyclopaedias to free and independent news services, amateur radio shows, free and legal photo sharing tools and social networks. Since the attitude of professionals and non-professionals has changed in that they are willing to make their content available, still more collaborative services can be expected in the near future.

Despite the popularity of recent, a number of issues that might have a negative influence on the quality of content remain unsolved. An example is Wikipedia which is troubled by incidents such as the Seigenthaler case. Therefore an alternative approach to collaboration in encyclopaedic environments is presented in the subsequent chapter.

Part 1 Electronic Encyclopaedias

Concepts and Organisation
Technical Aspects
Encyclopaedic Applications

2 Community-Building Around Encyclopaedic Knowledge

A version of this chapter was published in the Journal of Communication and Information Technology, Volume 14, Issue 3, pp. 175-190.

Publication submitted in February 2005, co-authored by Hermann Maurer.

This chapter gives a brief overview of current technologies in systems handling encyclopaedic knowledge. Since most of the electronic encyclopaedias currently available are rather static and inflexible, in the first part of this chapter an attempt is made to utilise enhanced functionality that enables more effective and collaborative work. Users have the ability, for instance, to add annotations to every kind of object, can have private and shared workspaces, and are provided with navigational aids.

The second part of this chapter details community building around encyclopaedic knowledge with the aim to involve "plain" users and experts in environments with largely editorial content. The foundations for building a user community are specified along with significant facets such as retaining the high quality of content, rating mechanisms and social aspects. A system that implements large portions of the community-related concepts in a heterogeneous environment of several largely independent data sources is proposed.

2.1 Introduction

Numerous electronic encyclopaedic systems are available, however most are rather conservative, inflexible, and static. Their focus is on technical aspects such as data storage, retrieval, and links to the Internet.

The following sub-section gives a brief overview of the capabilities of traditional electronic encyclopaedias. Sections 2.1.2 and 2.1.3 introduce two popular state-of-the-art encyclopaedias, and section 2.1.3 addresses a substantially different approach to encyclopaedic knowledge. Section 2.2, the first half of this chapter, presents seven requirements for advanced electronic encyclopaedias. The second half, section 2.3, deals with the integration of collaborative functionality in electronic encyclopaedias. Potential applications areas are outlined in section 2.4.

2.1.1 Traditional Electronic Encyclopaedias

Electronic Encyclopaedias were introduced in the 1980s (e.g., [Glushko et al. 1988]). Early versions were published on CD-ROMs in order to host the large amount of data. In the 1990s the Internet was discovered as a way to offer access to electronic encyclopaedias. The approach of most electronic encyclopaedias available on the market is rather conservative in that they focus on simplicity of information retrieval, embedding into applications such as word processors, links to search engines on the Internet, and the inclusion of multimedia content (images, videos, 3D animations).

Hence, typical electronic encyclopaedias are digital collections of material that used to be available in a series of books. Currently, the benefits of a digital representation are mainly ease of search and retrieval, that electronic encyclopaedias are relatively easily kept up-to-date, that they are available anytime anywhere, and that they can offer media such as animations or video clips. However, many things people usually do in printed dictionaries and encyclopaedias cannot be done in their digital counterparts. Users cannot make annotations, do not have the possibility to highlight parts of the content, cannot make clippings, etc.

More importantly, digital encyclopaedias can open up entirely new perspectives and modes of application. The power of distributed environments, for instance, and the potentiality for users to collaborate are still mostly ignored.

2.1.2 Brockhaus Digital Encyclopaedia

The Brockhaus Digital Encyclopaedia is undoubtedly one of the most advanced electronic encyclopaedias

currently available (see [Brockhaus 2006]). Apart from the functionality of conventional electronic encyclopaedias, the Brockhaus encyclopaedia includes several innovative features such as the "knowledge network" (see [Maurer and Tochtermann 2002; Mülner 2001]). The software finds similarities and connections and graphically depicts them in an intuitive way. The encyclopaedia also offers a large number of links to external articles, to a large image database, to specialised resources on the Internet, and to popular general-purpose search engines. Related to the Brockhaus Digital Encyclopaedia is Xipolis, a large collection of several dictionaries, encyclopaedias, and similar resources (see [Xipolis 2005]).

2.1.3 Wikipedia—A Radically Different Concept

In the modern age, the tradition to distribute and trade knowledge in the form of professionally authored encyclopaedias goes back to the 17th and 18th centuries. The probably best known example for one of the first modern encyclopaedias is "Encyclopédie ou Dictionnaire raisonné des sciences, des arts et des métiers" by Jean Baptiste le Rond d'Alembert and Denis Diderot (authored 1751-1772).

This is in stark contrast to Wikipedia. As thoroughly discussed in section 1.4, Wikipedia is the prime example for a free, Web-based, community-driven encyclopaedia. The principle of Wikipedia stems from the open source community: work in a community is done to meet one's own needs, for the good of the public, and for free (see also [CC 2006; GFDL 2005]).

The Wikipedia project that has a potentially deep "philosophical" and historical impact. In Wikipedia, not only the content is free, but the entire editorial process is open to all people on the Internet. Hence, articles are neither written exclusively by experts nor are they reviewed by editors, and there is no fact-checking. Reliable mechanisms to approve the expertise of authors or to verify the authenticity of content do not exist.

Although the underlying concept of Wikipedia is revolutionary and Wikipedia's aim to develop an encyclopaedia of the entire human knowledge is unprecendent, it is still a rather conventional encyclopaedia. Despite its collaborative power, Wikipedia "only" makes use of the community to produce new content, and therefore merely mimics existing encyclopaedias. This is, in fact, confirmed by member of the community who compare Wikipedia to tranditional encyclopaedias such as the Encyclopædia Britannica.

2.1.4 Google Answers

A radically different approach to encyclopaedic knowledge is taken in knowledge brokering environments such as Google Answers ([Google 2005]). With this service, users choose a problem domain, ask a specific question, and determine how much they are willing to spend on an answer. Domain experts reading the question can post appropriate answers along with references and other resources. They are remunerated for every satisfactory answer. Also registered users have the possibility to submit answers; they will not be paid, though. Questions and their corresponding answers are archived in the system, and other users can freely browse the repository.

The question *"Why is the sky blue?"*, for instance, is answered by a registered user. Later, a domain expert, a physicist, appends a detailed explanation of the phenomena leading to the blue colour of the sky. The question and both answers are stored in the "geophysics" category of the system and are made available to all readers.

Although a knowledge brokering system largely deals with encyclopaedic knowledge there are fundamental differences to conventional encyclopaedias. The information is not structured, does not necessarily have to be persistent, and its authenticity cannot be guaranteed. However, both experts and registered users attempt to maintain a high quality and professional style. Thus, knowledge brokering systems like Google Answers are largely community-driven, and quality as well as content depend on experts and users participating in the project. (It should be noted that these characteristics strongy resemble Wikipedia.)

2.2 Functionality Beyond Traditional Electronic Encyclopaedias

The approaches to electronic encyclopaedias described above are too static, offer too little flexibility, do not support users in their work, and do not open new prospects as they could. In order to counter these deficiencies, seven distinct demands that are essential in an elaborate electronic encyclopaedia system are pointed out in the following sub-sections.

It is quite surprising that despite the fact that these technologies have been employed in hypertext and digital libraries for years, they have not yet been introduced in the field of electronic encyclopaedias. Together with the collaborative functionality described in section 2.3, truly new applications may be realised.

2.2.1 Annotations

People working with physical encyclopaedias, and books in general, are used to highlighting sections of text with markers, writing down comments with pencils or attaching post-it notes (see [Marshall 1997]). Knowledge management systems, modern digital li-

braries, and enhanced journals such as the Journal of Universal Computer Science offer support for annotations (see [JUCS 2005] and [Krottmaier 2003]).

A modern electronic encyclopaedia should allow users to "annotate everything"—the textual content of articles, images, video clips, sound files, other users' contributions (see section 2.3 below), even the links to external references, etc. Users should also have the possibility to highlight certain portions of text or mark sections of images, videos or sound files. Moreover, users should be able to attach (potentially varicoloured) labels such as "important" or "for project A" to all kinds of objects.

In order to support collaboration among users annotations can be on different levels of access: private annotations are only available to the one user, group annotations can be seen within a specified group of users, and public annotation are visible for everyone. Private and group annotations may simply be created by users themselves, whereas public annotations need some kind of regulations by administrators of the system in order to prevent confusion (see [Maurer 2001; Maurer et al. 2006]).

2.2.2 Active Multimedia Documents

The idea of active documents is that users can ask arbitrary questions to documents, and answers are provided immediately and seemingly by the document itself (see [Heinrich and Maurer 2000]). The implementation of this concept includes an "online" and an "offline" component. If a semantically equivalent question can be found in the system its answer is presented to the user. Thus, the answer is provided online.

If an appropriate, existing question cannot be retrieved the user's request is forwarded to a human expert, and the user gets an apologetic message that an answer will be provided as soon as possible. In this case, the answer is provided offline by an expert. In the course of time, typically after some 500 to 1,000 users per document, answers are available for the most significant and most frequently asked questions. Therefore human experts are no longer required, and answers can be provided by the online component.

In an environment that handles encyclopaedic knowledge, it should be possible to ask questions to every piece of content, i.e., most objects in the system are active documents. Answers are provided either by an editor or by a user of the system (see section 2.3 below).

A large number of questions to an editorial article causes a notification to be sent to the respective editor, and the article has to be reviewed. If too many questions to an article written by a user from the community occur the article is (at least temporarily) disabled and not visible to other users.

2.2.3 Links to External Resources

An electronic encyclopaedia should be massively hyperlinked—both internally and to external resources. External resources do not only comprise general-purpose search engines such as Google and specialised databases such as the Internet Movie Database, or a dedicated encyclopaedia on chemistry, for instance, but also links to other kinds of (relatively) persistent information such as newspapers, magazines and journals. With these resources, an encyclopaedia can offer features such as "in the news". Moreover editorial articles from the encyclopaedia can be supplemented with high-quality, up-to-date information from the news. Content on the Ukraine, for example, could be complemented with a report from the BBC on the recent election in the country.

The inclusion of external resources, however, poses interesting problems such as the consistency of links and quality control. Link consistency is relevant internally as well as externally (cf., [Kappe 1995]). A hyperlink to the current programme of the Globe Theatre in an article on London has to be removed, for instance, when it becomes unavailable. On the other hand, it has to be ensured that bookmarks users make to articles in the system can be accessed even if the article has been modified. Persistent URLs might offer a partial solution to this problem (e.g., [PURL 2005]).

When external resources are offered quality becomes an issue as well. Hence, techniques such as blacklisting and whitelisting might have to be employed (see section 2.3.3 below).

2.2.4 Private and Shared Workspaces

Electronic encyclopaedias should allow for the users' demand to retain clippings of articles, pieces of content and other information (e.g., [Marshall and Bly 2004]). Therefore private and shared workspaces are proposed for electronic encyclopaedias by which a central archive for storing material is provided. In order to make use of the visual memory and pattern recognition skills of humans a two-dimensional workspace similar to a spatial hypertext seems favourable (see [Buchanan et al. 2004])

Personal workspaces may contain textual content as well as pictures, video clips, animations, etc. Users can either copy and paste the material they discover or, alternatively, they can extract parts thereof by way of transclusions (see [Nelson 1981; Nelson 1996], [Krottmaier and Maurer 2001], and chapters 4 and 5). In the latter approach, the information is not duplicated but the original data is "cited", i.e., every time a transcluded piece of information is accessed it is retrieved from the corresponding original document. Moreover, users can make annotations to objects in their personal workspaces, can save hyperlinks

to external resources, can store search queries, can put external documents into their workspace, can start to assemble new information and perform similar tasks. The personal workspace also provides an overview of articles recently read and a history of most recent comments and discussions.

Shared workspaces are essential for collaboration in user groups (e.g., [Rubart et. al 2001]). They provide the basic functionality of personal workspaces but have to take access control and communication facilities into consideration. To the benefit of ease of use, only two access modes are provided: private (to the author) and public (to the group).

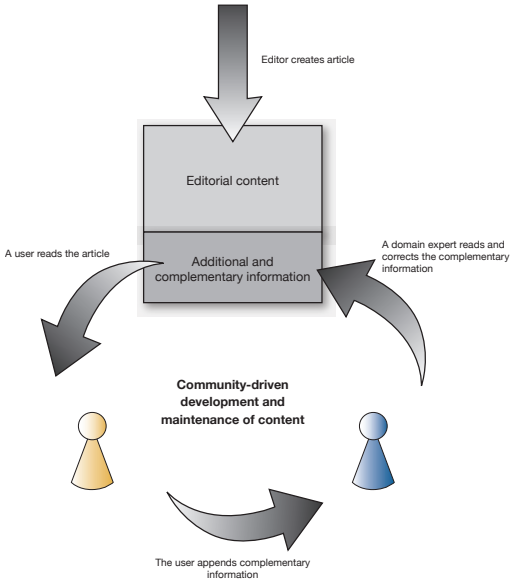


Figure 2.1: Community-driven development and maintenance of content. Editors create an article and users can add complementary content, references, etc.

Functions for supporting communication among group members in a shared workspace are simplistic. User can leave a notes for individual users or for an entire group. When a user is actively working in the workspace and receives a message it is displayed instantly. Otherwise it is presented as soon as the user returns to the workspace (see also section 2.3.6 below).

2.2.5 User Profiles and Adaptation

Electronic encyclopaedias should adapt to their users in order to provide better search results in shorter time and to facilitate the discovery of knowledge that might otherwise not have been found. Adaptation to the user allows for differing levels of experience and knowledge as well as distinct interests and varying aims. Students doing research for their homework, for instance, need to be confronted with different tools and need to get different search results than a teacher looking up a particular detail of the colonisation of New Zealand. Therefore adaptation has to take place in various aspects of an encyclopaedic application: from the user interface to the presentation and level of detail of search results (content-level adaptation) and the material supplied (link-level adaptation; see chapter 3).

User profiles are utilised to create an adaptive environment. Several differing strategies are suggested for generating a profile. These approaches include the submission of general profiles and dynamic profiling (see chapter 3). A general user profile, for instance, can be filled out by the user on the first use of the encyclopaedia. It contains general questions about the user's experience and interests in certain domains such as geography, history, literature, science, etc. The type of user interface—ranging from a simplistic, Google-like search window to an expert mode—can be selected as well (e.g., [Brusilovsky 1996; Brusilovsky 2000]).

An advanced version could include dynamic profiling and adaptation, which basically means that the user's actions in the environment, the kind of content displayed, the areas from which articles are retrieved, etc. are tracked. Based on these data, the system can adjust the user interface and display more appropriate content. The system might also put up personalised messages such as *"Did you know that ..."* or *"You might also be interested in ..."* where applicable. The encyclopaedia's featured article of the day can be adapted or recent news items might be pointed out to the user.

2.2.6 Trails

Electronic encyclopaedias should make use of an advanced tracking and navigation technique. The suggested mechanism in this chapter is the use of trails (cf., [Bush 1945; DeRoure et al. 2001; Furuta et al. 1997]). Trails are graphs that usually depict the most recent nodes (documents) a user accessed in a hypertext system. These previously viewed articles are complemented with the most relevant (most popular) articles from the perspective of the current article (see figure 3.2). The user can navigate to these potentially interesting pages by simply clicking on the trail.

Trails require a tracking and navigation component (see chapter 3). The tracking component analyses the users' navigational "behaviour" in the encyclopaedia. The navigation component utilises this data to generate new trails and presents them to the users.

2.2.7 Quality Feedback and Review Mechanism

Electronic encyclopaedias should retain detailed statistics on the users' queries and the articles retrieved in order to enhance the performance of the system. The findings derived from the system logs can be employed to find out which articles are currently very popular

and, consequently, might have to be revised or extended. On the other hand, the statistics might reveal that certain topics are requested rarely and, therefore, do not need to be updated anymore.

In addition to this, the statistical analysis can point out missing articles and instantly notify an editor. Alternatively, the system can offer a function for users to request a new article explicitly. This relatively simple technique can improve the flexibility and topicality of edited encyclopaedias (cf. section 2.1.3 above).

2.3 Community Building in Electronic Encyclopaedias

As pointed out above, traditional applications handling encyclopaedic knowledge are rarely based on user communities. A few exceptions, though, manage to be quite successful—Wikipedia and knowledge brokering systems such as Google Answers. However, these system have drawbacks such as the lack of authenticity of the content provided and the unknown identity of the authors.

Therefore a novel concept is introduced that allows communities to be built around editorial encyclopaedic knowledge. The following sub-sections address the purpose and notion of the concept, detail foundations of a user community as well as social issues and concerns about the quality of content. Furthermore discussion and communication facilities along with techniques for presenting the content are briefly outlined.

2.3.1 Purpose and Notion of a User Community

In an encyclopaedic environment, users should not only have the capability to read articles and comment on them, but they should be able to contribute actively to the development of the "knowledge base", to keep it up-to-date, and to make the environment more flexible. Thus, the community built around an electronic encyclopaedia is an opportunity for users to share knowledge and experiences in various ways.

Users may not only read but also append information to existing pages or write entirely new articles, can include or reference external resources, and have discussions. A community enables collaboration, so users can actually work *together* on a given project.

2.3.2 Foundations of a User Community

As the focus of the user community is to work on the knowledge base, an emphasis has to be put on the quality of the articles. In order to avoid problems known from Wikipedia—users cannot be sure whether articles they read are correct—at least the editorial content has to be authentic.

Hence, editorial content is provided by the board of editors and cannot be modified by the user community. When users from the community read an article and would like to supply additional content such as text, photos or sound they may do so. Subsequent users can provide yet more content or change the information that was provided by other users. This "cycle" of community-driven content development is depicted in figure 2.1.

The user community is based on three levels of identified (non-anonymous) users: certified experts, domain experts and plain users (see figure 2.2). Certified experts are typically experts in their domain (e.g., university professors or researchers) that are certified either by an editor or by a certain number of other certified experts. Potentially new experts can submit an application or can be recommended by other users. Usually their application is discussed by a group of experts in the same domain, and they have to pass an "entrance exam", in which their expertise is to be confirmed.

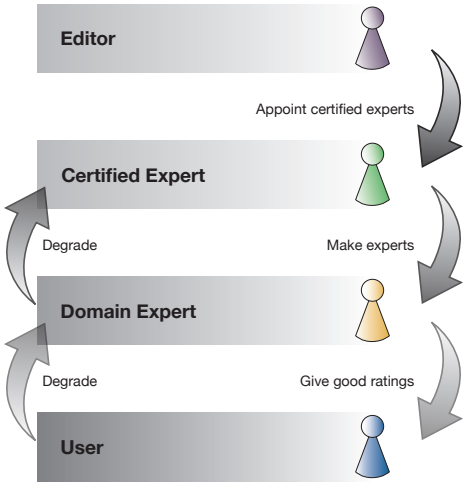


Figure 2.2: Different levels of users and basic mechanisms to grant and retract user statuses. Certified experts can, e.g., make users experts, and by means of poor ratings users can degrade experts.

Certified experts have the ability to make a plain user a domain expert when the user, for example, posts a number of high-quality contributions. Plain users can also become domain experts when they have a very large number of positive ratings (see also section 2.3.3). On the other hand, certified experts or domain experts unable to meet the required standards will have their certification withdrawn.

In an attempt to highlight the community's professionalism, domain experts and especially certified experts are encouraged to provide a publicly available web-page with some personal information and documentation of their expertise (e.g., a list of publications, professional experience).

The combination of a certification mechanism and providing personal and professional information is capable of ensuring high quality standards and can keep vandalism as observed in Wikipedia low (see [Viégas et al. 2004; Wikipedia 2006b]) while it still remains easy to submit information to the system. Another measure to foster development of content and deter malice are the quality control functions introduced in section 2.3.3.

2.3.3 Quality Control

The quality control mechanism is based on blacklisting, certification by experts, and a rating system similar to those known from discussion forums or online auctioning systems. Blacklisting means that the system retains a list of certain words or phrases, for example vulgar expressions, that must not be contained within articles published by users of the community (see [Krottmaier 2003; Weiss 2003], see also section 3.3.3). If an article includes a blacklisted expression, it cannot be submitted to the system. Blacklisting is applied to text-based content, annotations, comments, hyperlinks, in discussion forums, etc.

For various other kinds of information, such as links to external resources, whitelisting is utilised (see [Lennon and Maurer 2003]). This means that only certain, well-chosen external resources can be accessed, and all other sources are rejected (see section 3.3.4).

The second approach to quality control includes the intervention of experts. Every article has to fall into a particular category. Whenever a contribution is published by a plain user or by a domain expert, it has to be approved either by a certified expert or by an editor working in the corresponding category.

Although this technique is certainly rather expensive, it fuels the dialogue between experts and authors. An expert could, for example, suggest to the user that a certain section of the article be changed so that it can be appended to the content. Thus, consensus needs to be reached prior to the publication of a new article, which is, in a way, similar to the review process in scientific publishing.

Another measure for quality control is an advanced, credit-based rating system. Credits can be positive, negative or neutral and can be given for any kind of information including text, images, annotations and hyperlinks from both editors and users of the community. Credits from different users have different importance. A positive credit from a certified expert might be worth a value of 1.2, whereas the credit of a domain expert has a value of 1.1, and the credit of a plain user is 1.0.

The overall ratings for each user are calculated in an accumulative manner. Five positive credits by certified experts and three positive credits by domain experts, for example, result in an total rating of 9.3 (5 * 1.2 + 3 * 1.1). If users receive a very large number of positive credits, the value of their own credits are increased. The credit of a plain user A, for instance, is increased to 1.05 if A's total rating exceeds a certain threshold value. Every credit given by A is worth a value of 1.05 from that point onwards. In the case where A's total rating deteriorates, the value of A's credits decreases as well.

Although this approach is somewhat complex, it represents a "fairer" rating system that takes the users' commitment into consideration. Moreover, it encourages users to contribute valuable content because it potentially makes them more "powerful".

Ratings have an influence on both the users' articles and their user level. If plain users, for instance, receive a great number of positive ratings they might be promoted to domain experts. If domain experts, on the other hand, receive a lot of negative credits, they might be downgraded to plain user level. Ratings can also trigger certain functions on content (see also section 2.3.5):

- if a user-authored article receives a large number of positive credits, the encyclopaedia automatically suggests to an editor to make the article part of the editorial content;
- if the number of negative credits for a user-authored article exceeds a certain threshold it will be automatically removed from the system;
- if certified experts or domain experts have a certain number of articles removed due to negative ratings their certification might be withdrawn;
- if editorial content receives too many negative credits the board of editors is notified and the article has to be reviewed.

Together with statistical analysis functionality introduced in section 2.2.7, the concept can be refined. Articles that contain a lot of data provided by the community are automatically reported to the board of editors. When the article is reviewed by editors they can accept it. This means that the information provided by a user is automatically added to the editorial content and credited to the corresponding user. The user is not only associated with the section in the article but also gets an extra positive rating or a particular reward.

2.3.4 Active Knowledge Brokering

Active knowledge brokering is an essential component in a community environment. It lets users choose a particular domain and ask arbitrary questions within this field of knowledge. Answers to the users' questions are provided either immediately by the system or by experts and editors within a certain period of time.

The concept combines multiple technologies such as natural language queries similar to [AskJeeves 2005] or [GuruNet 2005], syntactic, semantic or heuristic analysis of questions, clustering, active document filters, and conventional knowledge brokering as detailed in section 2.1.5. An overview of the paradigm is given in figure 2.3.

Users start by selecting a certain area of interest such as "astronomy" and formulate an arbitrary question in natural language, for example, in plain English. The question triggers a database query in the internal knowledge repositories. The results from the various sources are combined, clustered and displayed to the user as "articles from the archives".

However, the user's question also triggers an active documents and answer brokering component. First, an analysis of the question is performed (see [Heinrich et al. 2001]). If semantically equivalent questions that have already been answered by human experts are available the user is asked *"Did you mean …?"* If this

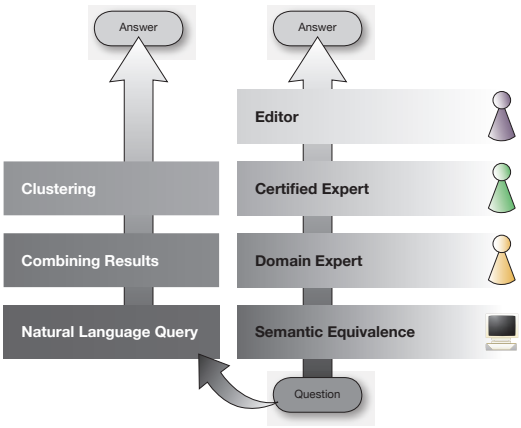


Figure 2.3: Answer brokering mechanism. When a question is asked the system searches for semantically equivalent previous queries. If none can be found, the question is passed on to domain experts, certified experts and, finally, to editors.

active document filter fails, i.e., none of the suggestions is satisfactory or none are obtainable, the question is passed on to human experts. In case an appropriate answer cannot be provided by domain experts within a certain time, the question is passed on to a certified expert. If an answer still cannot be provided, it is forwarded to an editor.

At any point, questions can be rejected as inappropriate. Moreover, experts and editors have the capability of combining semantically equivalent questions that are not recognised by the system. So if a question is similar to a prior one, it is specified as "equivalent" to an existing question and linked to the corresponding answer. Thus, the semantic filter is able to learn.

From a technical perspective, active knowledge brokering is a generalisation of the active documents approach. With active documents, question-answer pairs are always attached to particular articles or other pieces of information. Active knowledge brokering, on the other hand, is question-centered: questions are largely autonomous and are usually not attached to articles. An article or a piece thereof, however, can be part of an answer. Apart from that, an answer to a question can, by means of transclusions, lead to a new article (see also chapter 3).

2.3.5 Social Aspects

Research has shown that at least two types of users exist in environments where digital content is shared and disseminated: "givers" and "receivers" (see [Marshall and Bly 2004]). This implies that some users are willing to update incorrect articles, answer other users' questions, or even collaborate on certain issues. By contrast, other users will rather receive information than make an effort to reply to questions, etc.

In order to animate *all* users to contribute to the community—and in addition to the rating system—these socially disparate characters should be visualised in the environment. A straightforward approach is the use of distinct icons for different types of users and characters. A teacher, for instance, can be depicted as a man or woman in front of a blackboard. A teacher writing many articles has a pile of notes in the icon, and a professor answering a lot of other users' questions has a warm smile, for instance.

2.3.6 Mixed and Split Content

An article on the history of film, for example, has a large number of annotations and links to external references attached. In order to display such an extensive page in a comprehensible yet intuitive way, two modes for displaying articles are proposed—split and mixed content.

In split content mode, editorial content is strictly separated from user authored content. As depicted in figure 2.4 (c), the editorial section of an article is always displayed on top. Complementary information added by users is displayed below in an individual area of the window; it can be revealed and hidden by clicking on the triangle symbol. This example also shows two additional sections for discussions and external resources; both are "collapsed". Instead of collapsing areas in the window, icons can be used to indicate additional content such as annotations or hyperlinks. By clicking on an icon the corresponding annotation or hyperlink is revealed.

The main benefit of the split content approach is that it is apparent for users which content was published by editors (and therefore is authentic) and which content was added by other users of the system. Moreover

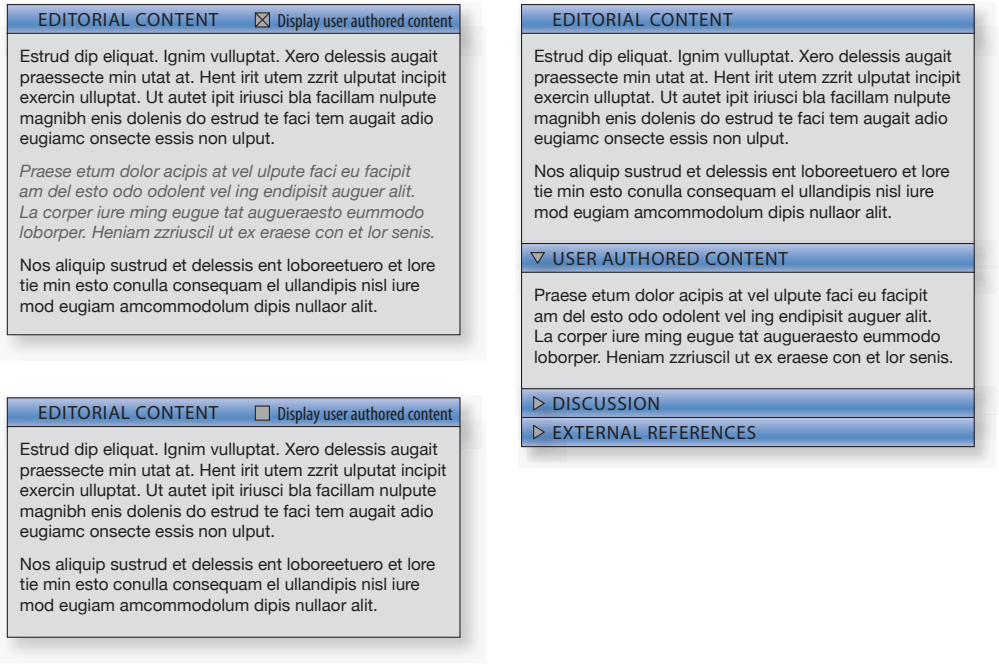


Figure 2.4: Mixed and split content modes. (a) Left, top: mixed mode with user authored content. (b) Left, bottom: mixed mode without user authored content. (c) Right: split mode with user authored content displayed and discussions and external references hidden.

this functionality can be used relatively easily and is straightforward to implement.

A more innovative technique for displaying articles is the mixed content mode. When an expert user, for instance, wants to add a paragraph to an editorial article the entire article is presented in a text editor. The expert can add content including text, images and video clips at any position in the existing article. The editorial content, however, is protected and cannot be changed. Users viewing the article can choose whether the user authored content is displayed (see the checkbox in the title bar of figures 2.4 (a) and (b)). In figure 2.4 (a), the piece of text provided by the community is displayed using a different font.

This approach ensures that the quality of editorial content cannot not be tampered with, while users of the community can easily add material at the most appropriate position in the article. The context is preserved, which can result in a seamless integration of editorial and community-provided content. Readers, on the other hand, can choose to view only the editorially approved information (figure 2.4 (a)) or can also have the data supplemented by the community displayed (figure 2.4 (b)).

2.3.7 Notification and Communication among Users

Notification mechanisms and communication among users are two significant aspects in the community. Notification mechanisms are necessary to let users know that one of the articles they authored was certified or that information has been added to an area of interest. In both cases, "pull" as well as "push" technologies can be employed. Push technology implies active dissemination where the encyclopaedia sends a message, e.g., an e-mail, to the users informing them of news. With pull techniques, the user has to retrieve the corresponding information explicitly from the encyclopaedia. A technology that is very well suited for this kind of information retrieval is RSS (see section 1.2.2 and [RSS 1999; RSS 2001]).

It is important that direct user-user and user-group communication can take place within the system and does not require external technologies such as e-mail. Therefore a simple system of „leaving notes" for users is suggested as the primary way of personal communication. When user A sends a message to a user or user group B, and B is online, the message is presented immediately. Otherwise the note is retained in the system, and as soon as B logs on to the system all messages stored are displayed. A more open way to communicate are discussion forums briefly outlined in the next sub-section.

2.3.8 Discussion Forums

Discussion forums based on the paradigm of conventional newsgroups should be available in any situation and at any point in the environment. Users have the possibility to discuss matters related to content in forums and discussions can, in fact, be attached to any kind of information in the system—to articles, to annotations, hyperlinks, images, etc. Although discus-

sions are usually related to a particular piece of information users are also able to have "meta-discussions" in which general strategies such as the reorganisation of a encyclopaedic category or the need for new experts can be argued.

Discussions take place in the system but are out-of-band rather than directly in the article (see also figure 2.4 (c) where discussions are presented in a separate area). The advantage of this approach is that articles strictly contain content and do not include discussions about content. Thus readers can mainly concentrate on the actual content and still have the option to join a discussion, if they are interested.

2.4 System Proposal

A web-based system that builds a community around an encyclopaedic body of knowledge consisting of at least one electronic encyclopaedia, several archives of articles from magazines and newspapers, and potentially other sources of persistent information of high quality is proposed in this section. The target group is a fairly closed user group of people wishing to have simple access to in-depth, high quality information on a wide range of topics.

The following sub-sections provide a brief overview of the architecture and the main concepts of an enhanced, community-based encyclopaedic environment. Several aspects have been implemented as project "Alexander" and are tested with a closed user group of more than 700 individuals (see Appendix B and [Alexander 2006]).

A pilot test of the Alexander's core functionality started in September 2006 and is scheduled for at least three months. The test phase is designed to collect data on the acceptance of Alexander's key features and to gain experience in encyclopaedias with both collaborative features. New features are gradually added to Alexander.

2.4.1 Overview

The proposed system is based on several sources of data from potentially heterogeneous architectures. The knowledge repositories contain entries from encyclopaedias, newspaper articles and similar data. Moreover, the contributions of the community—annotations, complementary content, discussions, etc.—are retained in the databases of the system.

Building on these repositories, the integral components of the system are implemented. User management, different levels of users, the rating system and adaptation are essential to the environment (see sections 2.4.2 and 2.4.5). Communication facilities and answer brokering (see section 2.4.6) are further components that are deeply rooted in the system core.

Basic functionality such as discussions, annotations and links to external resources is built on the foundations provided by the system core. More complex concepts such as transclusions or workspaces require the basic functionality.

The features and functionality offered by the system are presented to editors, experts as well as users in a graphically appealing and easy-to-use, intuitive web-based user interface. An overview of the generic system architecture is given in figure 2.5.

2.4.2 User Community

The system is based on a user community as introduced in sections 2.3.2 and 2.3.3. The community can contribute to the knowledge base, develop content, and work together with the information provided by the system. The system comprises five levels of users:

- administrators that have access to all functions and data in the system, mainly for performing maintenance and organisational tasks;
- editors that may create, modify and delete user accounts as well as all kinds of content in the system, may answer questions, and have full access to discussions; they have also access to whitelists (see below);
- certified experts that can create, modify, remove and certify user authored content, have discussions, answer questions (see below), promote plain users, and certify new experts (together with other certified experts); they have access to whitelists;
- domain experts that can create new content and modify or remove their own articles, have discussions, and answer questions;
- plain users that can create new content and modify or remove their own articles, and have discussions; and
- all user can rate other users and their contributions.

During a prototype phase, when only a small number of representative users work with the system, the experts in the community are simulated with a set of well-educated people in various areas. Each of these experts is responsible for several domains and has premium encyclopaedias, specialised databases, and similar archives of high quality at hand.

This approach is used to ensure the vividness of the community from the very beginning of the project and to "jump-start" the acquisition of experts and new users.

2.4.3 Basic Functionality

The basic functionality of the proposed system resembles the one of traditional encyclopaedias and digital

libraries. It provides full-text search over all information stored in the system with fault tolerant input. In addition to searching, articles can be accessed through browsing.

The system permits users to attach annotations and labels such as "important" or "for my project" to any object in the system. As discussed in section 2.2.1, for annotations and labels different access levels are taken into consideration. They can be private, accessible within a certain group of users, or public.

Furthermore, users are capable of bookmarking articles and any other information items in the system. Thus, they can make direct references to content in the system.

2.4.4 Authoring New Content, Linking, and Transclusions

Users of the community, especially certified experts and domain experts, have the ability to generate new content. This is mainly done by authoring new articles, writing annotations, providing links to external resources, and having discussions.

New content, however, can also be authored by means of transclusions (see chapters 4 and 5). An expert writing an article on the history of tea might, for instance, want to include a paragraph on the Boston tea party. A brief summary that is well-suited can be found in an article about settlements on the North American west coast of the 18th century. Instead of copying and pasting the paragraph from this article, the user simply transcludes it. Thus, by transcluding pieces of information from various articles completely "new" content can be produced.

Although transclusions generally pose a number of difficulties in web-based environments, their implementation can be made easier because it will only be possible to make transclusions from the internal knowledge repositories of the system. Hence, consistency of transcluded information and potential copyright-related issues can be controlled.

As described in section 2.3.3, in order to retain a high standard whitelisting is employed. The system stores a list of external references (servers) that may be used, and all other resources are automatically blocked.

2.4.5 Adaptation

The system can adapt to its users. The adaptation is based both on a general profile that can be filled out when a user registers with the system and on ad hoc profiles. Dynamic profiling as described in section 2.2.5 will not be implemented in the prototype of the system but rather in a forthcoming version (see also Appendix B).

The profiles provided by users are employed for adaptation on both content and link levels. Hence, search results, the links provided, the source of information, the types of user-authored articles displayed, etc. vary based on the users' preferences. Users have the option to disable the adaptation in order to get a "default view" that would have been used without any adaptation. This ensures that users still have access to *all* information stored in the system and that no content is filtered by the adaptation component inadvertently.

2.4.6 Clustering and Knowledge Maps

When users search for terms in the encyclopaedia, clustering techniques are employed to combine and display the results in a way similar to [Autonomy 2005] and [Clusty 2005]. A query for the term "blue sky" might return a number of results from the encyclopaedia as well as articles from newspapers and other repositories. These results are displayed to the user in clusters such as "atmosphere", "electromagnetic field", "pollution" and "space".

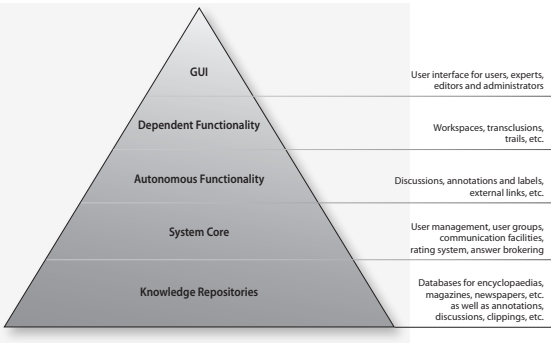


Figure 2.5: Generic system architecture.

However, clustering is not only utilised in search queries but also when users view articles. Every article displayed is complemented with document clusters that offer a brief overview of pages available on a topic (cf., [A9 2005]). This approach facilitates the discovery of information and makes it easier to get an overview of a certain area of interest.

Knowledge maps as mentioned in section 2.1.2 are basically a graphic depiction of information clusters. Every article in the system is supplemented with a knowledge map on the user's demand.

2.4.7 Active Knowledge Brokering

An active knowledge brokering module as described in section 2.3.4 will be included as a fundamental part of the system. The prototype version will require users to select both a category and a sub-category for their queries, for example, "astronomy" and "solar system".

Subsequently the user can ask a question in natural language.

The system will search the internal repositories including the encyclopaedia and the newspaper archive for matching articles. The results are combined, organised in clusters and presented to the user. Additionally, the system will perform a syntactic and heuristic analysis on the question in order to determine if similar questions have been asked before. In case the system is unable to detect appropriate questions-answer pairs from previous queries, it will forward the question to human domain experts of the corresponding problem domain. If the question is not answered within one day by a domain expert it is passed on to a certain group of certified experts. If, after two more days, still no answer is provided the user's query is forwarded to editors in the respective area. (See also figure 2.3.)

2.4.8 Discussion Forums and Communication

Similar to the "annotate everything" paradigm (see section 2.2.1), every piece of information in the system can be discussed. As explained in section 2.3.7, discussion forums are typically attached to objects in the system, whereas meta-discussions do not have to be related to any specific kind of object.

The system includes communication facilities as outlined in section 2.3.6. Users have the ability to write messages to other users in the system. They may include content from articles, make references to objects such as images or annotations, quote contributions to a discussion, etc. The messages are displayed immediately if the recipient is currently online (similar to instant messaging) or presented as soon as the user logs on to the system (similar to traditional e-mail).

2.5 Application Areas

The concept of the proposed system can prove to be valuable in numerous fields ranging from e-learning systems to enhanced electronic encyclopaedias. The following sub-sections briefly outline a few potential application areas.

2.5.1 Enhanced Encyclopaedic Environments

The proposed system clearly targets heterogeneous encyclopaedias consisting of several data sources such as encyclopaedias and archives of newspapers or magazines. It is obvious that especially communities that have the opportunity to develop and maintain content can be competitive in comparison to Wikipedia while ensuring a high standard.

However, not only online encyclopaedias but also DVD-based encyclopaedias can make use of parts of the proposed functionality. Annotations, private workspaces, trails as well as user profiles and adaptation can be fully implemented without the need for a connection to the Internet.

2.5.2 E-Learning

Many state-of-the-art e-learning environments make extensive use of encyclopaedic knowledge, and students utilise electronic encyclopaedias when they are doing their homework or learning for exams. Therefore e-learning environments are particularly well-suited for the functionality introduced above. Students can use private and shared workspaces to work together on projects (see section 2.2.4), they can make information they encounter available to other students, and are able to use annotations to share their insights. When articles from the encyclopaedia are active documents students can ask questions to documents and get instant answers.

Discussion forums can be employed to exchange views on articles and the information retrieved from the system, and the notification mechanisms described in section 2.3.5 keep students informed when new articles or comments are added, or a discussion forum is updated. Lecturers and teachers can, of course, use the same infrastructure as well. They can publish new documents, make annotations to articles, initiate discussions, attach additional content. In this case, teachers (cf. editors in an encyclopaedia) and students (domain experts) form a community.

2.6 Summary and Outlook

This chapter introduced several demands that future systems dealing with encyclopaedias knowledge should meet. Annotations should be omnipresent, and documents should be really active documents. Furthermore, users should have the opportunity to store the information encountered and their comments in a private workspace—and they have to be able to share it with other users in the system. User profiles with the aim of adaptating to individual users, statistical analysis, and navigation and tracking technologies based on trails will provide more accurate and more detailed search results.

The following chapters provide in-depth research on several topics covered in this chapter. The adaptation of structure and content in electronic encyclopaedias is discussed in chapter 3. Chapters 4 and 5 focus on transclusions and offer approaches to implementing both textual and multimedia transclusions in HTML-based environments such as the elecotronic community presented.

Although the concept for community-building around encyclopaedic knowledge is based on itendified users it is unclear which role anonymity play in

collaborate systems. Therefore anonymous authorship in Wikipedia, the largest collaborative electronic encyclopaedia currently available, is analysed in chapter 6.

Note

Only a few weeky before the final publication of this thesis Larry Sanger, one of the founders of Wikipedia, announced a new project called "Citizendium" (see [Citizendium 2006]). Citizedium is a collaborative encyclopaedia quite similar to the community described in this chapter.

Citizendium is based on registered users and experts (editors). Initially, Citizendium will make Wikipedia's articles available. Eventuelly the content will be modified by the Citizendium's experts, which might lead to significantly different articles. The aim is to increase the standard of the information provided and to be capable of offering "expert approved" content.

3 Adaptation of Content and Structure in Electronic Encyclopaedias

A version of this chapter was submitted to the Journal of Digital Information.

Publication submitted in October 2005, co-authored by Christian Safran and Hermann Maurer.

Adaptive functionality has been applied successfully in many areas ranging from user interfaces to hypermedia systems. Digital libraries and electronic encyclopaedias, however, have rarely made use of the power of adaptation.

In this chapter, an approach to include adaptation in encyclopaedic environments is presented. The proposal covers a set of adaptation techniques. They enable the system to explain technical terms and replace domain specific expressions with "plain" words automatically. Moreover, specific terms can be linked to further articles automatedly. Blacklisting, whitelisting and general link alteration are employed in order to assure quality standards and to provide users with more appropriate hyperlinks. With navigation support based on the automatic insertion of trails and suggestions of potentially interesting articles, the users' navigation in encyclopaedias can be facilitated.

A flexible framework for the implementation of this adaptive functionality in a web-based environment is proposed. The system is based on a traditional client-server architecture, where the server-side components perform the actual adaptation. The concept is modular and easily extensible.

3.1 Introduction

Electronic encyclopaedias were introduced in the 1980s, and numerous products with varying coverage and focuses exist today. Most electronic encyclopaedias, however, are rather conventional, inflexible, and static with an emphasis on technical aspects such as data storage, information retrieval, and multimedia content (see chapter 2). Moreover, many systems are monolithic in that they offer one body of knowledge that has to be appropriate for *all* readers.

Users, however, are often interested in different aspects of the same article, or they approach the same piece of information from different perspectives (cf., [Vester 1978]). An article on the ginkgo tree, for instance, might be read by a passionate gardener and by a person interested in alternative medicine. The gardener might expect references to articles on how to grow the tree adequately, whereas the person interested in medicine might want to know more about its curativeness.

As this example shows, users have varying goals, diverse backgrounds and different degrees of knowledge. With today's largely inadaptable encyclopaedias readers are frequently not satisfied because too little, too much, or inappropriate information is provided. One possible solution is to own several encyclopaedias in order to have the same knowledge at different levels of detail at hand.

In order to counter these deficiencies, adaptive functionality and personalisation for electronic encyclopaedias is proposed. Section 2 of this chapter gives a brief overview of adaptive systems, describes what kind of information can be adapted and how the adaptation can be implemented. The key part of this chapter is section 3, where various adaptation techniques for the use in electronic encyclopaedias are presented. Technical aspects are addressed in section 4, and section 5 details a number of potential further application areas.

3.2 User Adaptive Systems

Traditionally, information providers used to follow a "one size fits all" approach in order to minimise the effort of creation and maintenance, to decrease time-to-market and reduce complexity and expenses. Information, in this context, comprises actual content, organisational structures, user interfaces, network connections, etc. In such static systems, users of an

online news service, for instance, are provided with the same news stories—independent of their interests and background. In web-shops, all potential buyers are presented the same featured article of the day. Visitors of online museums are offered the same explanations and the same guided tour. Users of software products have to utilise the same user interface, no matter what their experience is (e.g., [Brusilovsky and Maybury 2002]).

Adaptive environments break up these constraints of static systems and attempt to offer information in ways better suitable for users (see [Benyon and Murray 1993; Holland 1962]). They are capable of providing adapted user interfaces, adapted link structures in hypermedia systems, content adapted to the users' knowledge, their aims, or the capabilities of the devices they use, etc.

Adaptation technologies and adaptation mechanisms can be distinguished. Technologies of adaptation determine the types of information that can be adapted, whereas adaptation mechanisms describe approaches used to implement them. The following sub-sections provide an overview of the technologies and mechanisms relevant to incorporate adaptive functionality into electronic encyclopaedias.

3.2.1 Adaptation Technologies

Many adaptation technologies have been proposed and implemented. Some of the more prominent examples are the adaptation of user interfaces, adaptation of content (content itself is modified), selection of content (which content is presented), the ranking of search results (which content is more relevant), presentation, structural adaptation (alteration of links in hypermedia structures), and navigational support (insertion of useful links).

In order to give a brief introduction to adaptation technologies, the following sub-sections focus on three particular techniques: adaptation of user interfaces, content adaptation and structural adaptation.

3.2.1.1 Adaptation of User Interfaces

User interfaces, especially graphical ones, dominate the users' work with computers. Most user interfaces allow users to make settings in order to adjust certain aspects of the interface, provide default answers to frequent queries, or simply make the work with a software application more comfortable. A refined approach is to allow users to modify menus, short-cuts, and other user interface elements.

However, user interfaces can be adapted in more sophisticated ways. Microsoft, for instance, adapts user interfaces dynamically in several products (e.g., [Microsoft 2005a]). Menus are re-ordered so that frequently used menu items are among the first, and less frequently used ones among the latter items in a menu. Although the intention is to provide easy and fast access to frequently used functions, this technique is inconsistent with the concept of human visual memory. Users can usually remember the position of menu items, icons and other symbols in a two-dimensional space ([Marshall and Shipman 1995; Findlater and McGrenere 2004]). When the position of menu items is altered constantly even experienced users can feel lost in the user interface. This example shows that not every type of adaptation is sensible in certain application areas.

Moreover, some products make use of "agents" that attempt to identify the users' aims and actively assist them in completing their tasks. Additionally, agents can inform users of simpler ways to accomplish a task (e.g., [Xiao et al. 2004; Schlimmer and Hermens 1993]).

A similar approach is taken in recommendation and decision support systems, where the system makes suggestions to users. An example is the highly complex control software of power plants that highlights the required tools and recommended responses in case of a critical situation (cf., [Langley 1997]). User interfaces for physically impaired people are another example. The adaptive software attempts to predict what the user is about to do and adapts the user interface to proactively present the tools required for the task.

3.2.1.2 Content Adaptation

Various aims and requirements can lead to systems that implement content adaptation. E-learning systems are one of the first areas in which this technology was employed. It was recognised in the early stages of computer-supported learning systems that skills and knowledge of the learners vary greatly. Hence, environments that are able to adjust lessons and other learning material to the users' needs were developed (e.g., [Mödritscher et al. 2004; Pivec and Baumann 2004]).

Mobile applications and ubiquitous computing are a field of growing importance for the use of content adaptation. Common demands for mobile devices are small size and light weight, which results in limited capabilities such as small screen sizes, relatively slow network connections, and restricted input methods. Therefore content is adapted to the capabilities of client devices prior to transmission. This means that video clips, for example, are transcoded to lower resolutions and reduced frame rates, and images are transmitted with less colours (e.g., [Fu et al. 2001; Lum and Lau 2002]).

3.2.1.3 Structural Adaptation

Structural adaptation has its roots in hypertext. One of the fundamental elements of hypertext systems are links between objects, where the structure of a hypertext collection frequently represents a graph with edges (links) and nodes (objects). Structural adaptation corresponds to the alteration of edges in the graph (e.g., [Stotts and Furuta 1991]).

Three basic operations can be distinguished: modification, removal, and insertion of links. When users in a web-based environment select a link that has been altered they are directed to a different resource. A possible application area are e-learning systems, where students may only access the next lesson if they have successfully completed an exam. If they fail the exam the link is adjusted and points to the beginning of the previous lesson.

From a user's perspective, the removal of links from a hypertext structure can be seen as blacklisting (see below). In corporate computer networks, for instance, this technology can be employed in order to prevent users from accessing web-sites of competitors or resources that are not related to their work.

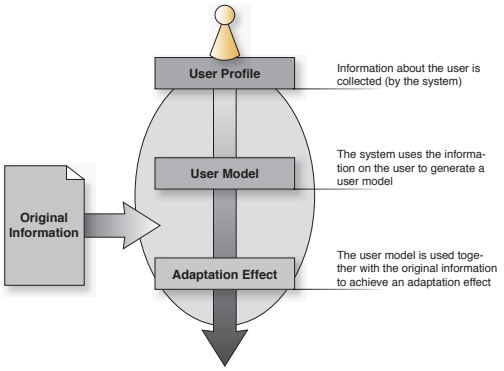


Figure 3.1: The common process employed in most adaptive systems to achieve an adaptation effect. The user data stored in the user profile are used to generate a user model that is applied to the original information in order to accomplish the adaptation.

The insertion of links is often used in navigation support systems. Links to potentially significant information that might otherwise have gone unnoticed or links to resources that have been recommended by other users of the system are presented to the reader (e.g., [Weber and Specht 1997]).

3.2.2 Adaptation Mechanisms

Adaptation mechanisms determine how the actual adaptation is performed by the system. Most adaptive environments accomplish an adaptation in three steps (see figure 3.1). First, an attempt is made to gather information about the user; the data is stored in a *user profile.* The data that comprises the user profile can either be entered explicitly by users or can be collected gradually by the system itself. The adaptive system makes use of the data retained in the user profile in order to create a *user model.* By applying the user model to the original piece of information—a user interface, textual content, a hypertext structure, etc.—an *adaptation effect* is accomplished.

In this thesis, three generalised kinds of adaptation mechanisms are distinguished: static, dynamic, and flexible dynamic adaptation. The differences between these three mechanisms are the generation of user profiles and user models, and the format in which original information is provided. A comprehensive overview of adaptation mechanisms can be found in [Brusilovsky 1996].

3.2.2.1 Static Adaptation

Systems that make use of static adaptation are sometimes denominated adaptable systems. The characteristic of static adaptation is that all possible adaptations are defined explicitly by authors and, in contrast to adaptive systems, retained statically in the system. Moreover, decisions on the use of adaptation are often not made automatically by the system but by the user. Thus, the user has to specify explicitly what is to be adapted and how the adaptation is carried out.

An example are web-sites available in multiple languages. The author prepares a page in various languages, and therefore multiple instances of one information object exist and are stored separately in the system. When a page is to be displayed, readers may choose from the set of languages. Alternatively, the web browser's language settings can be used to make an automated decision.

Similar to hierarchical file systems, the information objects in a static adaptation environment can be depicted graphically as a tree. The tree has several main branches—in this example, one for each language. By making a language preference, the user enters one particular branch of the tree, and the subsequent navigation takes place within this specific sub-tree.

Another example for static adaptation are adaptable user interfaces. Apple's DVD authoring suite "DVD Studio Pro", for instance, offers three pre-defined user interfaces: beginner, intermediate and advanced (see [DSP 2005]). The user interface for beginners includes a basic set of tools that let even unskilled users easily author DVDs. A very high level of abstraction is employed so that users do not have to be familiar with technical details. With the interface for intermediate users, authors have additional tools and options at hand. They can define a number of advanced parameters but have to have a basic knowledge of the underlying technologies. The interface for advanced users makes the full range of tools available. Options

and parameters can be modified on their lowest levels, and every aspect of a project can be optimised. Users, however, have to have an in-depth knowledge of the technologies employed.

In this example, the software designer prepares three different, static user interface variants. The user can choose between these pre-defined options. However, once one of the three user interfaces is chosen the system does not further adapted to the user.

3.2.2.2 Dynamic Adaptation

A more advanced approach to adjusting information to the users' needs is dynamic adaptation. With this mechanism, authors do not define statically which information is to be altered in a given situation, but the *software* makes an assessment which portion of information is to be adapted. This decision is derived from a number of parameters. In a learning environment, for instance, a user's learning objective together with the information on the categorisation of a lesson and the user's latest test results can be used to determine that an explanation in the lesson content is inappropriate and needs to be adjusted.

In most cases, dynamic adaptation requires a higher level of structure in the information to be specifically customised. Moreover, the information usually has to be constructed or prepared by authors. Depending on the actual implementation, authors might have to give hints on which information can be adapted and which pieces of information are suitable replacements. In other approaches, particular cases and conditions, in which a certain adaptation is performed, are defined. Thus, authors need both experience and special skills in producing information, which makes development and maintenance more demanding.

Another requirement for dynamic adaptation is a user profile for every user of the system. This information is typically provided by the users on the first use of the system in the form of answers to questions about their experience, background, favourites, and aims. This data is retained in the system and utilised for producing a user model that can be applied in the adaptation process. Most systems allow users to modify the settings in their profile in order to reflect changes in their aims, etc.

An alternative to persistent user profiles is the use of ad-hoc profiles. Every time the user wants to employ the system a question such as *"What would you like to do today?"* is asked. Depending on the user's answer aspects such as the user interface, the results of a database query, or the suggestions of a recommendation system are adapted.

Dynamic adaptation is a popular mechanism that is employed in areas ranging from learner support systems to online help systems, and general hypertext systems (e.g., [Pivec and Baumann 2003; Moore et al. 2001]). Another example for the use of dynamic adaptation are adaptive user interfaces (see section 3.2.1.1).

3.2.2.3 Flexible Dynamic Adaptation

Flexible dynamic adaptation can be seen as an enhanced variant of dynamic adaptation. With this approach, the adaptation mechanism itself, the selections of content to be adapted, and the user profiles can be adapted. Hence, an author, for instance, gives hints on which portions of information can be used in an adaptation. The system, however, has the ability to find data more suitable. Users, on the other hand, fill out user profiles, but the system can adjust profiles in order to express the users' preferences in a better way.

Thus, user profiles are not static but dynamic. As with dynamic adaptation, users may be asked to provide basic information when they use a service for the first time, and the system attempts to adjust the profile. Even more advanced systems generate a user profile eventually making explicit contribution of information by the user unnecessary.

Techniques from the field of artificial intelligence are employed in order to implement the adaptive functionality of such systems. Neural networks and machine learning, for instance, are technical foundations of many implementations (see [Annunziato et al. 2002; Narendra and Parthasarathy 1989]).

An example for adaptive adaptation can be found in recommendation systems. When the user submits a query, various selected recommendations are presented by the system. If the users follow any of the recommended answers, it can be assumed that the recommendation was appropriate. Thus, both the correctness of the adaptation algorithm can be confirmed under the given conditions and the user's profile can be adjusted.

A major problem of the flexible dynamic adaptation approach is that it might be difficult to obtain the data necessary to adapt the adaptation process and user profiles. In traditional hypertext environments, for instance, only little information can be gathered automatically—the last document requested by the reader, the time spent viewing the document, etc. However, this information might not be accurate, simply not sufficient to generate a user profile, or expensive to collect or compute (e.g., [Langley and Fehling 1998]).

Therefore flexible dynamic adaptation is used frequently in conjunction with features of dynamic adaptation, i.e., the system attempts to adjust the adaptation mechanism with the feedback manually provided by users. In a database system, for example, the system might ask the user after each query, if the results were appropriate.

3.3 Adaptation in Electronic Encyclopaedias

As mentioned above, most electronic encyclopaedias such as Wikipedia, the Encyclopædia Britannica, and the Brockhaus Digital Encyclopaedia offer little or no ways at all do adapt to their users (see [Wikipedia 2006a; Britannica 2005; Brockhaus 2006]). Therefore in this chapter, a number of adaptation techniques are employed, some of which are successfully incorporated and tested in various application areas such as e-learning environments or recommendation systems.

Both adaptation of content and structure of articles in the encyclopaedia is utilised, whereas adaptation of the user interface is not considered, yet. Mainly dynamic adaptation mechanisms are used. The adaptation is performed at run-time, automatically, and based on user profiles. In this thesis, the focus is on the types of adaptation that can be performed and their use rather than on actual adaptation mechanisms or user modelling.

The proposed functionality is to be part of the larger web-based encyclopaedic environment introduced in chapter 2. For this reason articles in the encyclopaedia are HTML documents that can:

- contain hyperlinks to articles in the internal repositories;
- include multimedia data such as images and video clips;
- include references to pages and other content from external resources; and
- be organised in a hierarchy of categories and subcategories.

Not only pages from internal repositories but basically every page available on the World Wide Web and documents from other external resources can be used as the source of the adaptation procedure.

3.3.1 Replacement, Explanation, and Linking

One of the most significant types of adaptation in this concept is the fully automated explanation and replacement of terms combined with linking to appropriate resources. Both explanation and replacement can prove to be useful for expressions such as technical terms and domain-specific words. The article on the human heart in an encyclopaedia, for instance, might contain the expression "angina pectoris". Although this term is familiar for most adults, various groups of readers including school children might not understand it.

Therefore three approaches to the explanation and replacement of terms that suit the varying needs of most users are proposed. Explanations are provided as appositions or in brackets and are available on two levels of detail. Users with a certain experience in the domain are offered a more elaborate description of the term (an *abstract),* whereas for users with no appropriate background including school children only a *short description* or a synonym is provided. Moreover, hyperlinks to appropriate resources are inserted automatically. When the same article is viewed by an expert user it does not include explanations but only links to specialised knowledge.

In order to illustrate the approach, the phrase *"leading to angina pectoris …"* is adapted. The following examples show the results of the adaptation for a school child, an intermediate user, and a domain expert:

- *"leading to a heart disease …"* where "heart disease" is a hyperlink to an article on diseases of the heart in a children's encyclopaedia;
- *"leading to angina pectoris, an ischemic disease of the heart causing pain in the chest due to a lack of oxygen supply, …"* where "angina pectoris" is a hyperlink to an article on angina pectoris in a general encyclopaedia;
- *"leading to angina pectoris …"* where "angina pectoris" is a hyperlink to a list of selected publications and recommended reading including current research results regarding angina pectoris.

In the examples above, two content adaptations are performed. For the novice user, the more general term "heart disease" is inserted as a replacement. In the second example, "an isechemic disease of the heart …" is used as an explanation. All examples make structural adaptations by including various hyperlinks.

Two questions are crucial for the feasibility of an implementation and need to be answered. Which terms are to be replaced, explained and linked? Where do explanations, replacements, and links stem from?

The solution for the first problem is based on two observations: (1) terms that occur relatively seldom in all articles of the encyclopaedia are usually technical terms, and (2) in many cases technical terms specific to a particular category are not very well known in other domains.

The first finding means, for example, that it does not make sense to explain the relatively frequent word "heart" in an article of a general encyclopaedia, while it is reasonable to describe the term "ischemic". The second approach is more complex, as the following example demonstrates. An article on Puhutukawa trees, for instance, can be found in the category Science and Nature – Biology – Botany – Botany of Australasia. Moiti Island is the home of a variety of Pohutukawa trees and is contained in the article on Pohutukawa trees. The term "Moiti Island", however, is classified in the category History and Geography – Australasia – New Zealand and is not part of the category Botany of Australasia. Thus, this term is part of a distinctly different category and therefore a potential candidate for an explanation.

The solution to the second problem, the source of the information required for an adaptation, attempts to involve the user community (see chapter 2). It is conceivable that the data structure of articles in environments such as Wikipedia, where content is solely developed by the community, is altered. Two additional fields for a short description and an abstract could be appended to every article and could be filled out gradually by the users of the encyclopaedia.

Alternatively, both the short description and the abstract can be generated in a semi-automatic process. In numerous encyclopaedias including Wikipedia, the first sentence of an article often resembles an abstract. Although this information can be extracted automatically, the data has to be proof-read and confirmed by members of the community in order to ensure the accuracy of the information.

3.3.2 Translation

Translation of expressions in foreign languages is an important aspect that is often overlooked. Many authors, for example, do not want to use translations because foreign-language terms are sometimes more precise or simply established in a domain. Therefore terms such as English words in computer-related articles or French expressions used in articles on cuisine might not be fully comprehended by readers.

Hence technologies for the automatic translation of terms from foreign languages are utilised in this approach to adaptation in encyclopaedias. In order to enable translations words from foreign languages have to be detected. In a first step, terms that occur infrequently in the categories or in the entire encyclopaedia are determined. If they cannot be found in a dictionary of the current language they are potential candidates for a translation.

For the subsequent actual translation, a simplified approach is taken; more advanced implementations can be found, for instance, in [Hutchins 2001] and in [Ide and Véronis 1998]. Terms are reduced to their principal form and looked up in a number of foreign-language dictionaries. The results of these queries are either provided as inline text (in brackets, for example) or in separate windows that can be accessed through hyperlinks.

Although several approaches have been developed automatic translations are usually not accurate (e.g., [Yarkowsky 1992]). Therefore users should be informed that the translations provided are automatic and might not be precise. In environments where users are actively involved, community feedback can be used to improve the performance of the approach. Similar to rating systems, users can decide whether translations are correct. After a certain number of such ratings the system can dispose of inappropriate translations.

3.3.3 Blacklisting

Blacklisting is a technique that prevents the access to, or the use of, services, resources, content or information defined in an exclusion list. It is usually employed as means of censorship, parental control, or in order to filter unwanted and unsolicited information (e.g., [Balkin et al. 1999]). Especially in environments such as the Wikipedia, where content is largely developed by the community, blacklisting can become necessary to maintain the quality.

Implementations exist on various levels including:

- access to networks: physical or logical access to certain networks is not possible;
- access to services: only services such as HTTP can be used in many Internet cafés, for instance, whereas file sharing is blocked;
- access to resources: URLs to certain service providers on the WWW, for example, are filtered in corporate networks; and
- access to information: e.g., certain words and phrases are removed from textual content.

In this proposal, blacklisting is employed on both the resource and information levels. This means that URLs and phrases can be blacklisted.

The approach to implementing resource filtering is straightforward. The system retains a list of URLs that are not permitted in the system. Entries in the list can consist of exact addresses or URLs with wildcards. If an article in the encyclopaedia contains a blacklisted URL the link is removed from the article. This strategy makes it possible to prevent users from accessing links to resources that were provided by the community but do not conform to required standards.

For blacklisting on the information level, an advanced, context-sensitive filtering mechanism is proposed. Conventional blacklisting removes the word "sex", for instance, from an article, if it is a member of the blacklist. With context-sensitive blacklisting, it is possible to define if a word is mandatorily or conditionally blacklisted. While mandatory blacklisting leads to the same results as conventional blacklisting, the decision if a conditionally blacklisted word is removed from an article depends on the context. The article on "sex", for example, might be part of two categories: biology and psychology. If the word "sex" is used in another article in one of these categories, it is not filtered because it can be assumed that, within the context of the category, the word is necessary, appropriate, and not offensive. Whenever the word is used in articles that are not in these categories it is filtered.

How "broad" the context for conditional blacklisting is depends on the actual implementation. It can be

rather narrow by limiting the context to the exact subcategories of articles, or rather broad by choosing top-level categories. Although it is possible to let authors assign the context for words on the blacklist *manually,* the cost for doing so is most likely too high.

3.3.4 Whitelisting

In consideration of the sheer number of documents and resources available on the World Wide Web and the amount of information produced every day, it is almost impossible to maintain lists of all unwanted resources (e.g., [Zakon 2005]). Hence, blacklisting is not always an ideal approach and whitelisting can be favourable.

With whitelisting, a list of all resources that may be accessed is retained. This technique can be used to enforce rather restrictive control of external resources in order to be capable of maintaining quality standards. However, it also enables parental control and can be particularly valuable when encyclopaedias are used in learning environments (see [Lennon and Maurer 2003]). In this case, linking to external material can become problematic, and whitelisting can be employed efficiently by allowing access only to a small number of accredited external content providers.

Although from an ethical perspective this notion is worrisome, it should be mentioned that both blacklisting and whitelisting can be used to prevent users from accessing content and services provided by critics and competitors.

3.3.5 Link Alteration

The automatic alteration of links during run-time can be used to provide better explanations and offer more appropriate further articles. In this context, links do not only comprise hyperlinks in text documents but also links to inline images, sound files and similar media documents. As such, link alteration can basically be seen as generalisation of blacklisting and whitelisting. While these two techniques are used to disable links that meet certain criteria, general link alteration is employed to modify links based on a given set of rules.

An article on heart diseases, for instance, contains a hyperlink to the article on echocardiography. Depending on the user's skills that are retained in the user profile, the link can point to the corresponding article in various sources such as a general encyclopaedia or a children's dictionary. Based on the user's preferences, hyperlinks can also point to different areas of interest. For an electrical engineer, for example, the term echocardiography could be presented as a hyperlink pointing to a detailed description of the technical design of the apparatuses available, whereas for a medical doctor the same link could point to an article describing cases in which echocardiography is an appropriate diagnostic measure. In these examples primarily hyperlinks to other articles are modified.

For an implementation, the system has to determine both the experience of users in various categories (e.g., beginner in history) and the level of complexity of resources in the system. Information on the users' skills might be part of the users' profiles. The content source (e.g., children's encyclopaedia), on the other hand, can be an indication for the complexity of resources.

When links point to resources whose levels are more advanced the system attempts to find appropriate resources matching the users' skills. If such replacements can be found links are modified, otherwise they remain unaltered. Alternatively, links can be removed (cf., blacklisting) in order to avoid confronting users with unsuitable information.

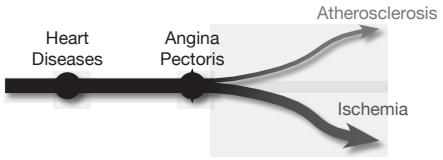


Figure 3.2: A trail, where the user has already viewed the articles on heart diseases and angina pectoris. The suggestions for further articles are "atheroschlerosis" and ischemia, where the connection from angina pectoris to ischemia is more popular.

This strategy is, of course, not only applicable to hyperlinks that point to further articles or external web pages. It can also be employed for links to other resources such as inline images, multimedia animations or video clips. However, this approach has to be evaluated extensively in order to avoid unwanted results.

3.3.6 Navigation Support

Adaptive systems can enable the implementation of enhanced navigational aids. With trails, an established method for presenting popular or pre-defined paths through hyperstructures is made use of in this chapter (see [Bush 1945; DeRoure et al. 2001]).

This mechanism usually consists of a tracking module and a navigation component. The tracking module collects information on the users' navigational behaviour in a hypertext system. It retains weights for every actual connection C_{AB} between any two nodes A and B, and whenever the user navigates from A to B the weight of C_{AB} is incremented. As an alternative to the dynamic generation of trails, they can be defined manually by authors as ordered series of edges in a directed graph.

The maximum number of connection weights to be stored is n * (n – 1), with n being the total number

of articles in the encyclopaedia. In order to improve the appropriateness of suggested articles in the trail the length of the tracking paths can be extended from two to three (e.g., C_{ABC} instead of C_{AB}). However, longer paths consume considerably more space in teh database.

When a user views an article the navigation component offers a trail containing (or starting at) the current node. The trail consists of the articles most recently requested by the user and a number of suggested further articles (see figure 3.2). The selection of articles suggested in a trail is based on the weight of connections starting at the current node. However, more advanced selection mechanisms might also take user profiles into consideration and produce trails that contain further articles that were requested by users with similar experience and preferences.

3.3.7 Information Encountering

While trails are a graphic depiction of recommended articles and can serve as direct navigational aids, the use of text-based suggestions is a method that can facilitate the accidental encountering of information (e.g., [Elderez 1997]). With this approach, potentially interesting articles are presented to the user in personalised messages such as *"Did you know that …"* or *"You might also be interested in …"*. A user interested in electrical engineering reading an article on heart diseases, for example, might be confronted with the suggestion *"You might also be interested in how a cardiac pacemaker works."* The links to these articles are generated automatically by the system on the basis of the user's aims and statistical information on the navigational behaviour of all users (cf., section 3.3.6). Thus, readers are actively supported in accidentally encountering information that might not have been found otherwise.

In addition to this, articles are complemented with a "recently in the news" section. When the user reads an article on, for instance, biochemistry a news service such as Google News is queried in order to find out if biochemistry was in the news lately. If so, references to the most relevant news stories from various news services are appended to the article in the encyclopaedia on-the-fly. This approach makes it possible to provide users with a range of external resources and up-to-date information (see also chapter 2).

3.4 Technical Aspects and Proposed Implementation

An implementation that provides a flexible framework for the application-level adaptation of articles in encyclopaedic environments is proposed. Before technical aspects are discussed, several requirements for articles and user accounts stated implicitly in section 3.3 have to be addressed.

3.4.1 Structure and Categorisation of Articles

Articles in the encyclopaedia are stored as HTML pages or in a different, potentially proprietary format that can easily be converted to HTML documents. A particular structure is not required, simple formattings such as "heading" or "paragraph" suffice. Links to other articles or to external resources as well as references to images and similar media objects are included as anchors or can be stored in a link database.

Every article is part of at least one category. Articles can, however, be part of multiple categories and subcategories. Articles can originate from various sources including general and specialised encyclopaedias, children's dictionaries, or scientific journals. This means that an article on the human heart may exist in several instances, each from a different source.

In order to be able to perform context and user sensitive link alteration (see section 3.3.5) a level of specialisation has to be assigned to articles. Specialisation is specified in three nuances reflecting the user's skills—beginner, intermediate, and expert. Thus, an article on the human heart could, for example, be categorised as "medicine, intermediate".

Additionally, articles have to have a small set of descriptive metadata including a short description and an abstract attached. These fields are required to enable the replacement and explanation of technical terms (see section 3.3.1).

The set of data including categorisation, level of specialisation, origin of the content and descriptions is stored in a feature vector. It may be extended with additional metadata when necessary. Authors could, for example, define manually which adaptations are to be applied or that certain articles are not to be adapted at all.

3.4.2 Profile Matrices for Users

As in many other adaptive systems, the user's knowledge is represented in an overlay model (e.g., [Brusilovsky 1996]). The overlay model uses the same structure as the subject domain, i.e., the user model is based on the same categories and features as the articles retained in the system.

In the proposed system, users can define both their skills and experience, and their aims and interests. These data are retained in two separate user profiles.

A user profile is a structured set of data stored in a matrix. The first dimension of the matrix consists of categories and sub-categories. The second dimension represents the three levels of specialisation (beginner, intermediate, and expert). With this approach, us-

ers can define simple attributes such as "beginner in medicine" (profile: skills, category: Medicine, level of specialisation: beginner) or "interest in geography of Australia" (profile: aims, category: History and Geography — Australasia — Australia, level of specialisation: expert).

3.4.3 System Architecture

The proposed system architecture follows a classic client-server approach where the computational logic lies within the server. The server-side system core consists of the adaptation engine and a number of adaptation modules. The adaptation engine acts as a request handler and is mainly responsible for data input and output as well as the invocation of the adaptation modules. Adaptation modules are small, special-purpose programs that perform a simple task such as blacklisting or the inclusion of trails into an article.

A typical adaptation process in the system starts with the client's request that contains the address of the desired resource as well as information uniquely identifying the user (see figure 3.3). These data are passed on to the adaptation engine that resides on the server. The engine analyses the request and retrieves the original resource together with its feature vector from the system's repositories. The information identifying the user is used to load the related user profile. This set of data comprising the original resource, its feature vector and the user profile is transferred to the various adaptation modules of the system. The modules perform the actual adaptation and return the results to the adaptation engine that might forward the results together with the original feature vector and the user profile to another adaptation module. After every registered module was called on the content, the final results of the adaptation are returned to the user as response to the initial query.

This approach resembles [Phan et al. 2002], an extensible, pipeline-based adaptive system used at UCLA Medical Systems. The advantages of such an architecture are flexibility and modularity. Modules could, for example, be added or removed during runtime without the need to modify the overall system. A drawback, however, is the increased complexity and potentially slower adaptation, which might result in deteriorating response times (see also section 3.4.6).

Independent from the architecture, adaptation can be computationally expensive. In order to increase performance caching is utilised, i.e., every adapted article is stored in an internal cache. When users with similar profiles request the same article within a given period of time, the article retained in the cache is returned to the user. Further measures for improving performance are described in section 3.4.6.

3.4.4 System Components

The system can be divided into client-side and server-side components. The client system is a traditional Web browser displaying the HTML-based articles of the encyclopaedia. Modifications or extensions of the client are not required.

The server-side components include a Web server, the adaptation engine and modules, content repositories and metadata archives. The Web server is a conventional, yet extensible HTTP server. The adaptation engine can be implemented as an extension module to a traditional HTTP server, or as an external program such as a CGI script that is exclusively used by the HTTP server (e.g., [Thau 2003]). In the latter case, the CGI program is similar to a non-transparent proxy application (see section 5).

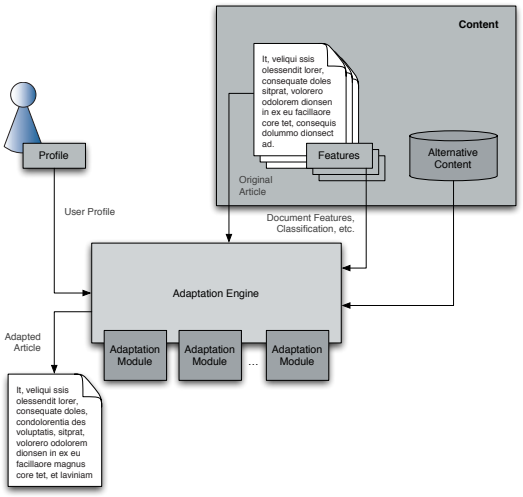


Figure 3.3: The adaptation engine and its modules utilise the feature vector of the original article and the user's profile in order to determine which parts of the content have to be adapted. Alternative content can be retrieved from a content database and can be used to assemble an adapted article.

The adaptation engine has various purposes. It handles client-side requests that are forwarded by the HTTP server and decides whether content is to be adapted. Moreover, it calls adaptation modules and determines the order in which they are invoked (see below). Modules have access to data objects provided by the adaptation engine including the user profile and the feature vector of the article currently being adapted.

The content repositories store the original, unadapted articles of the encyclopaedia. Metadata archives, on the other hand, contain not only user profiles and article feature vectors but also alternative content required for mechahnisms such as replacement and explanation.

It should be mentioned that the order in which the individual adaptations are applied is relevant. Blacklisting before automatic insertion of links, for exam-

ple, might lead to different results than the opposite order. Therefore a specific sequence for the application of adaptation modules is suggested: translation, replacement, explanation and linking, generalised link adaptation, blacklisting and whitelisting, and methods for information encountering.

3.4.5 Source of Required Data

In many innovative systems, especially those requiring metadata, the source of the required additional information remains unclear. This section highlights this aspect for the most significant types of data in the system.

The metadata for user profiles is provided by the users on the first use of the system. Users give weighted answers to a set of questions, which enables the system to automatically generate profiles. Eventually the system can update the profile by collecting data on the users' behaviour in the system.

In most encyclopaedias, a categorisation of articles is available (e.g., [Wikipedia 2006a; Brockhaus 2006]). In addition to this, large encyclopaedic environments such as Xipolis contain articles not only from a single source but from a number encyclopaedias and domain specific dictionaries (see [Xipolis 2005]). In this case, an article on one topic exists in several instances, and therefore information on both the category and the level of specialisation of articles are known.

As detailed in section 3.3.1, generation of abstracts and short descriptions can be based on an automatic extraction from the articles in the encyclopaedia. Human intervention will, however, be necessary in order to ensure the accuracy and quality of the information.

3.4.6 Performance Issues

Especially for online encyclopaedias, performance becomes an issue. Adaptation must not consume too much time because otherwise the responsiveness deteriorates. Therefore two approaches to increasing the performance are suggested: offline adaptation and delayed adaptation.

With offline adaptation, the data that is required for performing the actual adaptation is extracted and collected when the system load is low (e.g., during the night). The data is stored in an internal cache in a format immediately suitable for the adaptation process. When a document is requested, data is fetched from the cache (rather than from the original documents) and employed for the adaptation.

Recent technologies such as AJAX make delayed adaptation possible (e.g., [van Veen 2006]). With this approach, initially a largely unadapted document is sent to the user. However, at the same time a separate process for the generation of the information required for adapting and complementing this document is forked. As soon as the information becomes available it is "sent" to the user and inserted in the document on-the-fly. The results generated in this process are stored in the internal cache described above.

3.5 Application Areas

Although electronic encyclopaedias are the prime application area for the proposed functionality, there are several other fields that can make use of this concept. In fact, most systems dealing with encyclopaedic knowledge, and digital libraries in general, can benefit from the adaptive features presented.

A database storing user's manuals of technical devices, for example, usually offers the same information to a general audience, no matter what the readers' backgrounds and aims are. Senior citizens, however, might have to be addressed in a different way than electrical engineers. In such a scenario adaptive features including explanation and linking might lead to a better comprehension of the content.

Domain specific databases that focus on a general audience might be confronted with similar problems. In medical databases, for instance, many different types of users want to look up diseases—not only physicians. Since articles in these archives are often intended for medical doctors they contain numerous technical terms, Latin words, and references to rather specific further articles. In this case, explanation, translation, the automatic insertion of links and redirection of existing links to "plain" resources can facilitate understanding the complex material.

In learning environments such as web-based training or learner support systems, adaptation can be employed to provide more detailed descriptions where needed. This is mainly achieved through explanation and linking. Moreover, with blacklisting and whitelisting it becomes possible to have unsuitable content and links to unaccredited resources removed from learning material (see [Lennon and Maurer 2003]).

3.6 Summary

In this chapter, an approach to adaptation in electronic encyclopaedias was presented. This makes the implementation of techniques such as automatic explanation of terms, link alteration, and navigational support possible. With this concept, both the content and the structure of articles stored in an encyclopaedic "knowledge base" can be accommodated to the aptitudes of the individual users. Moreover, it is an attempt to provide advanced functionality in electronic encyclopaedias.

Various techniques that make the information provided more suitable for readers are employed. Technical and domain-specific terms, for instance, are explained or replaced with expressions that can be understood more easily. This concept may be combined with the automatic insertion of hyperlinks that offer further information on a topic. Translation of terms in foreign languages such as Latin, for example, can be applied in a similar fashion.

Blacklisting and whitelisting, successfully employed in learner support systems, are used as means to maintain the quality of the content included into the encyclopaedia. Especially hyperlinks pointing from articles in the encyclopaedia to external resources can be filtered accordingly. In addition to this, generalised link alteration attempts to modify hyperlinks in order that they point to more appropriate resources.

Furthermore trails and text-based suggestions are made use of so that users can encounter potentially interesting information. This mechanism also enables users to find articles frequently visited or recommended by other users of the system.

An architecture outlining the implementation of the proposed functionality is presented. The design relies on a classic client-server paradigm in a web-based environment. The system core with the adaptation engine residing on the server can be seen as request handler and decision engine. The actual adaptation is carried out by a number of adaptation modules, where each adaptive function is realised as a separate module that can be "plugged into" the adaptation engine. This modular design renders the architecture highly flexible and extensible.

4 Transclusions in an HTML-Based Environment

A version of this chapter was published in the Journal of Communication and Information Technology, Volume 14, Issue 2, pp. 161-174.

Publication submitted in May 2005, co-authored by Hermann Maurer.

Transclusions are an advanced technique for the inclusion of existing content into new documents without the need to duplicate it. Although originally described in the early 1960s, transclusions have still not been made available to users and authors on the World Wide Web.

This chapter describes the prototype implementation of a system that allows users to write articles that may contain transclusions. The system offers a simple web-based interface where users can compose new articles. With a simple button the user has the ability to insert a transclusion from any HTML page available on the Web.

While other approaches introduce new markups for the HTML specification, make use of technologies such as XML and XLink or employ authoring systems that internally support transclusions and can generate web pages as output, this implementation solely relies on the techniques provided by an HTML-based environment. Therefore HTML, Javascript, the Document Object Model, CGI scripts, and HTTP are the core technologies utilised in the prototype.

4.1 Introduction

In 1965, Ted Nelson presented *"a file structure for the complex, the changing and the indeterminate",* in which he introduced the term hypertext (see [Nelson 1965]). One of the fundamental concepts in Nelson's notion of hypertext is a technique called *transclusions.* Transclusions allow authors to include portions of existing documents into their own articles without duplicating them. Basically, a transclusion in document A is a reference to a portion of the content of a potentially remote document B that is virtually included into document A (see figure 4.1).

The following sub-sections discuss the background of transclusions and their use in HTML. Section 4.2 addresses attempts to implement transclusion. The design and implementation of this approach to transclusions in HTML-based environments are detailed in section 4.3. Issues that were encountered during the implementation are described in section 4.4. Finally, section 4.5 discusses several aspects of this implementation.

4.1.1 Background of Transclusions

Transclusions are designed as complete replacement for all *cut-and-paste* mechanisms in use. Nelson argues that cut-and-paste is not what people actually want to do but that it is a restriction imposed upon authors by the nature of paper. Writers actually do not want to make a copy of an existing document, cut out the piece they want to reuse and paste it in their document. They want to include the original content and let readers know what the source and the context of the quote is (e.g., [Nelson 1981]).

Reference lists at the end of a scientific publication, for instance, are usually not what is intended by writers and desired by readers. They are rather a pragmatic solution to the problem that both the source and the context of the quotation are lost by copying-and-pasting a portion of content printed on paper.

What used to be physical restrictions of paper was embraced by most computing systems in an attempt to resemble the work environments and common processes in offices (cf., [Yocom 2004]). Therefore most current graphical operating systems make use of metaphors such as a desktop, folders and documents; a document has to be put in exactly one folder; there is a clipboard, and content from a different document is included using copy-and-paste mechanisms (see [Nelson 1996]).

4.1.2 Implications of Transclusions

Transclusions are, however, not only a mere replacement for copy-and-paste. They assure that the original context of a quotation is preserved and can provide a visible link to the source of the transclusion. Ted Nelson's approach to realising this functionality is based on *transpointing windows* (e.g., [Nelson 1995]).

Moreover, authors of documents can be notified when their articles are transcluded. Thus they can, for instance, find out about other researchers in the same area. Authors using transclusions, on the other hand, can be informed automatically about modifications in source documents (see [Krottmaier 2002]).

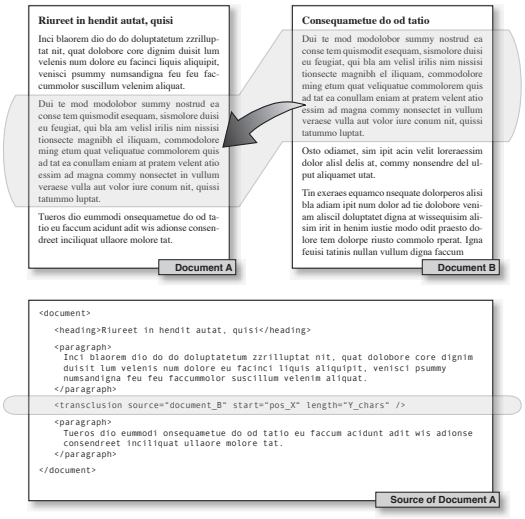


Figure 4.1: Exemplary transclusion. Part of document B (top right) is transcluded into document A (top left). Bottom: the source code of document A does not contain the actual text of document 2 but only the data required to retrieve it from the original document.

Apart from obvious improvements in authoring and publishing systems, transclusions can also offer a solution to some copyright issues experienced today on the World Wide Web—authors include content into their documents by means of transclusions. Whenever a reader views a transclusion a note about the rights associated with the transcluded content is added, and a micropayment is made to the corresponding owner ([Nelson 1999]). Nelson names this model *transcopyright* (see [Nelson 1998]).

4.1.3 Transclusions and HTML

The Hypertext Markup Language (HTML, [HTML 1999]) is a relatively simple language for describing platform independent hypertext pages. In the early stages of its development, the focus of HTML was on style and graphical presentation rather than on functionality and underlying paradigms. Therefore many innovative ideas such as bidirectional hyperlinks and issues already known at that time including broken links were not considered in the implementation (e.g., [Pam 1995; Maurer and Lennon 1996]).

In principle, transclusions are used in HTML. Designated markups including `<img>`, `<object>` and `<embed>` incorporate content such as images, Java applets and animations into HTML pages by means of linking. Thus, these elements basically make use of the concept of transclusions.

Transclusions in HTML are very limited, though. Only certain media such as images can be virtually included, whereas textual content, in general, cannot be transcluded. Moreover, the transclusion mechanisms available in HTML can only be applied to entire documents. Fine-grained transclusions such as a small spatial selection of an image are not implemented.

4.2 Attempts to Implement Transclusions

Although the idea of transclusions was proposed some forty years ago, only a few attempts to implement this advanced technique have been made. The following sections give an overview of several notable approaches to the realisation of transclusions.

4.2.1 Xanadu

Transclusions are an integral part of Xanadu, Ted Nelson's original hypertext system (e.g., [Nelson 1981]). Their implementation relies on a document model, though, that is radically different from what is widely used today. In Xanadu, documents *(versions)* do not contain content but references to the actual content. Content is both stored and referenced with the highest granularity possible—on the level of single characters. All content is retained in (potentially remote) content repositories.

Any document is made up of a list of references to content stored in the system, e.g., a document consists of "characters 124 to 729 and 1276 to 1301 from the repository". When content from document B is transcluded into document A, the corresponding references to the actual content in the repositories are added to the reference list of document A.

Thus, the creation and retrieval of transclusions in Xanadu are trivial list operations. Ted Nelson also details a number of functions related to transclusions and the handling of situations in which documents are modified or large portions of documents are deleted (e.g., [Nelson 1999]). Basically, these functions can be seen as more complex list operations.

4.2.2 Proposal to Amend the HTML Specification

Since the idea of transclusions is already present in HTML for media such as images and multimedia animations, a sensible approach to text-based transclusions is to introduce a new tag that allows users to transclude text. Therefore [Pam 1997] suggests an amendment to the HTML specification. A new markup, `<text>`, is proposed with the intention to offer an element of the same significance as `<img>` or `<embed>`.

The main attributes of the markup are the URI of the source document and the start position and length of the text to be transcluded. The web browser analyses the tag, loads the source document of the transclusion, extracts the portion of text given by the attributes of the `<text>` tag, and inserts it into the document. Thus, a transclusion is handled in a similar way as an inline image.

Although the proposal seems rational it has not been accepted, and no web browser to date has the feature implemented.

4.2.3 Transclusions with IFrames and Embedded Objects

The recommendation for HTML 4 includes markups for inline frames and embedded objects (e.g., [HTML 1999]). Both inline frames and embedded objects define areas within a given HTML document that can be used to display potentially remote resources. Inline frames can merely contain HTML pages and images, whereas objects may contain resources of arbitrary type.

```
<my:transclusion
   xmlns:my="http://www.kolbitsch.org/"
   xmlns:xlink="http://www.w3.org/1999/xlink"
   xlink:type="simple"
   xlink:href="source.xml"
   xlink:actuate="onLoad"
   xlink:show="embed" />
```

Listing 4.1: Fragmentary transclusion with XLink.

Transcluding document A into document B can, for instance, be achieved by inserting an `<object>` tag with a reference to document A into document B ([Krottmaier and Maurer 2001]). The capabilities of this technique are rather limited, though. Only entire documents can be referenced. Moreover, the context is lost because a link from the document containing the embedded object to the source of the transclusion is not provided by these markups. Therefore this approach is not well suited for realising transclusions.

4.2.4 XML-Based Transclusions

The Extensible Markup Language, XML, is a flexible language for describing documents that contain structured information (see [XML 2003]). In contrast to other markup languages such as HTML, where both syntax and semantics are determined, neither a set of tags nor the semantics are defined in XML. Therefore, XML per se does not contain a distinct markup for links; a separate linking language is used instead.

XLink, the XML Linking Language (e.g., [XLink 2001]), provides a framework for describing the syntax and semantics of even complex linking structures between resources. An XLink link typically contains a number of attributes that describe, for instance, what resource is to be loaded and when it is to be displayed.

Three attributes are essential for the implementation of transclusions using XLink:

- `href`: the document to be loaded. Set to the source document of the transclusion;
- `actuate`: when the resource is to be loaded. When set to `onLoad`, the resource is loaded when the document containing the XLink link is loaded;
- `show`: in which manner the resource is to be displayed. When set to `embed`, the resource is displayed practically instead of the XLink tag.

The skeleton of the XLink link shown in listing 4.1 transcludes the entire document `source.xml` into the document containing the link at the position of the link. A similar approach to transclusions in XML is described by [Wilde and Lowe 2002].

With the XML Pointer Language (XPointer, [XPointer 2002]) fragments of XML documents can be identified and addressed as well. Thus, a combination of XML, XLink, and XPointer can be employed to make the use of fine-grained transclusions in XML-based environments possible (see [Krottmaier and Maurer 2001; Krottmaier and Helic 2002]).

4.2.5 Recent Projects Involving Transclusions

Currently, several mostly academic projects that experiment with transclusions exist. The following paragraphs introduce three selected systems.

4.2.5.1 University of Nottingham

The University of Nottingham, UK, has proposed a technology-based learning environment that adapts to its users (see [Moore et al. 2001]). The information retained in the system is organised in small "chunks" that are stored as XML files.

Since the system is adaptive, lessons are not static but assembled dynamically on the basis of a lesson plan. When a user requests a particular lesson, appropriate chunks of information are retrieved and included into

a virtual document by means of transclusion. Thus, the system facilitates the reuse of small pieces of information for a number of lessons or for students with various differing standards of knowledge.

4.2.5.2 University of Bologna

At the University of Bologna, Italy, researchers attempt to combine existing software products such as the Internet Explorer and Microsoft Word to offer a collaborative editing environment for the World Wide Web (see [Di Iorio and Vitali 2003]). The implemented tool, XanaWord, allows users to edit any web page they view in their web browser—even if they do not have write permissions for the resource.

Any page displayed in Internet Explorer can be opened with a word processor such as Microsoft Word, where the user can make arbitrary changes to the document. When the user saves the page, only the changes to the original document are stored in the XanaWord repository. Whenever a document is retrieved from the repository, the modifications made by the user and the content from the original resource are included in a dynamically generated document by means of transclusion. Finally, the dynamic document is sent to the user's web browser.

4.2.5.3 IICM, Graz University of Technology

The Institute for Information Systems and Computer Media in Graz, Austria, proposed an environment capable of handling transclusions in various output document formats (see [Krottmaier 2002]). The system includes three components:

- the Latex typesetting system that allows users to create documents and save them in a number of document formats including Postscript, PDF and HTML;
- an extension to Latex that allows users to create transclusions; and
- a Hyperwave Information Server (see [Hyperwave 2005]) that handles issues such as linking and versioning.

In the proposed environment, users can insert a special markup that designates a transclusion in Latex documents. Then, the user has to upload the file to a Hyperwave Information Server that extracts links and saves them in a link database, etc. When the document is requested by a user, the transclusions and links are inserted into the file saved on the server. The resulting intermediate file is processed by Latex in order to generate the requested document format. Ultimately, the document containing the transclusion is sent to the client.

4.3 Implementation

In contrast to several approaches to transclusions illustrated above, this project does not present a proposal but an actual implementation of a system that lets users take advantage of transclusions. It is designed as part of a larger system that offers communities instruments to work actively with content from digital libraries and electronic encyclopaedias (see chapter 2).

A prototype is implemented offering a tool for authoring new articles that can contain transclusions. It is available online at [Kolbitsch 2005a].

4.3.1 Design Goals and Requirements

The environment for creating and retrieving transclusions aims at facilitating the reuse of information readily available on the Web—even by novice users. Therefore a number of design goals have to be taken into consideration:

- ease of use: the tool for making transclusions must be as easy to use as traditional copy-and-paste mechanisms;
- use of any document on the Web: not only documents from a closed repository but basically any web page may be the source of a transclusion;
- level of granularity: any portion of text may be transcluded from a document—from a single character to the entire content of a page.

Browser plug-ins or special software tools should not be required. Therefore, this implementation of transclusions solely relies on technologies available and widely utilised on the World Wide Web:

- HTML: transclusions can be made from any HTML formatted document available on the web. Moreover, documents containing transclusions are presented to the reader as traditional HTML documents (see [HTML 1999]);
- Javascript, DOM: internally, most current web browsers represent HTML pages as trees of objects. The underlying technology is the Document Object Model (DOM, [DOM 1998; DOM 2005]). Javascript is used to access individual objects in the DOM tree of the HTML page to be transcluded and enables fine-grained transclusions (see [Netscape 2004; ECMA 1999]);
- HTTP: documents containing transclusions are transmitted to the readers using the Hypertext Transport Protocol (see [HTTP 1999]).

4.3.2 System Overview

Since the system for creating and retrieving transclusions consists of a number of components, a brief overview is given. The following description is made in the order of actions taken by a user in authoring and reading a document including transclusions.

Two fundamental actions can be distinguished in the system: the creation of a transclusion when the article is authored, and its evaluation when the page containing the transclusion is to be displayed. When the user wants to create an article with a transclusion the web browser presents a frameset with two frames. One frame contains a conventional text area for authoring HTML content and an additional button for adding a transclusion. When the user presses this button, the URL of a page can be entered, and the corresponding page is loaded through an HTTP proxy application into the second frame. The user can either transclude content from this page or can use the second frame to browse to a different page—again through the HTTP proxy application. The document to be transcluded is complemented with a button that inserts the transclusion into the text area of the first frame, when pressed. An illustration of the two frames is given in figure 4.5.

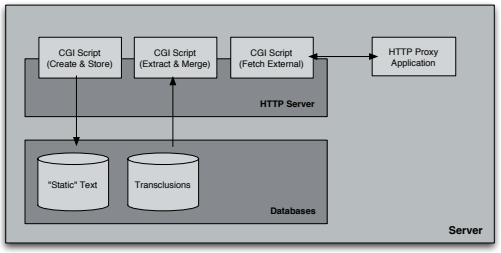


Figure 4.2: Overview of the server-side components.

The user selects the portion of text to be transcluded in the second frame and presses the button to have the transclusion actually inserted into the article. The button calls a Javascript function that determines the start and end positions of the selection made. Together with the URL of the page in the second frame these values are used to generate an intermediate markup that is inserted into the article (see section 4.3.4).

Once a user has finished authoring an article and chooses to save the new document, the contents of the text area including the intermediate transclusion markup are sent to a CGI script on the server. The server stores the "static" text and the values provided through the transclusion tag to a database. In addition to this, metadata on the source of the transclusion is collected and stored in the database.

Whenever the article containing the transclusion is requested, a second CGI script is invoked. The script retrieves the contents of the article and the parameters of the transclusion tag from the database. The parameters defining the transclusion are used to load the original page from its original location. If it is unchanged, the transcluded portion of text is extracted from the original page, combined with the static text and sent to the client's web browser of the client (see figure 4.5).

The next section gives a brief introduction to the architecture of the implementation and its components.

4.3.3 System Architecture

This implementation of transclusions follows a classic client-server paradigm. A conventional HTTP server, a relational database, several server-side CGI programs, a non-transparent HTTP proxy application and client-side Javascript code are the main components of the system.

The CGI script "Create and Store" in figure 4.2 receives the data submitted by users. It analyses the content of the article, extracts transclusions and stores both content and transclusions in the internal database of the system (see section 4.3.4).

The "Extract and Merge" script, on the other hand, reads the content of an article together with the information on the transclusion from the database, fetches the source document of the transclusion, assembles the complete article and sends it to the user (see section 4.3.5).

The third CGI script, "Fetch External", is utilised during the authoring process for loading the page to be transcluded. This programme is basically necessary to insert a button and a small portion of Javascript code into the corresponding page (see section 4.4.1). It relies on a specialised, non-transparent proxy application developed for this project.

In the current prototype implementation the relational database consists of only two tables. While one table contains the static content of the article, the other one stores detailed information on the transclusion as well as a rich set of metadata and a fingerprint of the source document.

```
<transclusion src="{url}"
   atag="{tag}" aindex="{int}" aoffset="{int}"
   ftag="{tag}" findex="{int}" foffset="{int}" />


<transclusion src="http://www.kolbitsch.org/about/"
   atag="H1" aindex="1" aoffset="0"
   ftag="P" findex="4" foffset="29" />
```

Listing 4.2: Syntax of a intermediate transclusion tag (top) and an example (bottom).

4.3.4 Creating a Transclusion

As explained above, the interface for authoring new articles consists of a frameset with a frame for writing an article in an HTML form and a separate frame for displaying the content to be transcluded (figure 4.5). When users wish to insert a transclusion they select

the portion of text with the pointer device and click on the button provided in the window.

The button calls a Javascript function which is essentially the only operation carried out on the client computer. It accesses the document object model to determine the exact start and end positions of the selection made by the user and generates an intermediate tag that is inserted into the article. The Javascript interface to selections provided by most browsers is somewhat peculiar in that it determines these start and end positions in the way the user actually marked the text. I.e., of the second H1 heading and ends at the 30th character of the fifth paragraph in the given document.

When the article containing the transclusion is saved by the user the data is sent to the server, and the "Create and Store" CGI program is invoked (see figure 4.3). It extracts the transclusion from the article, determines the attributes of the transclusion and writes the information to the database. The `<transclusion>` markup in the original article is replaced with a transclusion object that refers to the transclusion stored in the database.

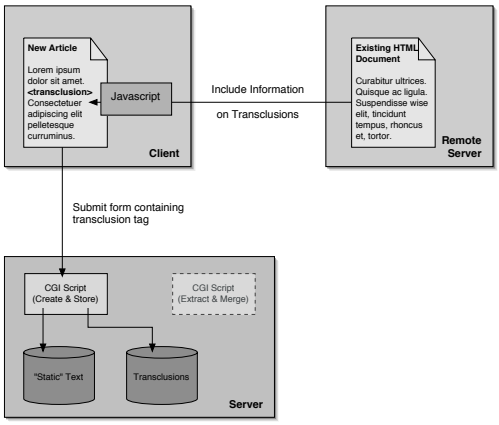


Figure 4.3: Simplified schematic illustration of the process of creating a new article containing a transclusion.

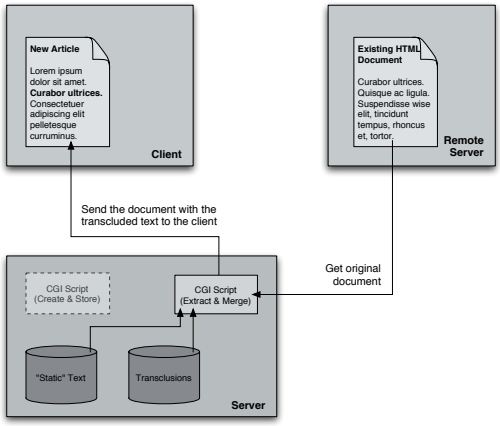


Figure 4.4: Handling a request for a page containing a transclusion (simplification).

the *anchor* of the selection is the position where the user clicked to indicate the beginning of the selection. Then, the user drags the mouse, for instance, to the end of the selection and releases the mouse button to denote the end of the selection. The end position is the *focus.* In the following paragraphs, anchor and focus are denoted by prefixes "a" and "f".

The syntax of the intermediate `<transclusion>` markup with its seven parameters is rather complex (see listing 4.2). This level of detail is required to be able to determine the exact start and end positions of transclusions. Values in curly braces describe the type of attribute values:

- `src`: the URL of the document to be transcluded;
- `atag`, `ftag`: the names of the tags in which the transclusion starts and ends, e.g., "P" for a paragraph;
- `aindex`, `findex`: the index of the tags in which the transclusion starts and ends, e.g., the seventh paragraph in the document;
- `aoffset`, `foffset`: the offset within the start and end tags, e.g., the transclusion starts at the second character of the seventh paragraph in the document.

The exemplary tag shown in listing 4.2, for instance, describes a transclusion that starts at the first character

The source document of the transclusion is not stored in the internal repository. However, its URL, the creation and modification dates as well as an MD5 hash value of the entire page content are retained as fingerprint. These values are necessary to determine if the source document has changed when the transclusion is retrieved.

It should be noted that, in contrast to [Pam 1997], where an amendment to the HTML specification is suggested, the `<transclusion>` markup in this implementation is only used during the authoring process. It is inserted when the user makes a transclusion, is evaluated by the system and replaced by a transclusion object. When the page containing the transclusion is to be displayed, the transclusion object is replaced with the corresponding content from the original page (see below). Hence, the `<transclusion>` tag is only visible within the system but not externally to the user.

4.3.5 Retrieving a Transclusion

The "computational logic" of this implementation mainly lies within the component that retrieves transclusions. Whenever an article is requested, its body is analysed for the presence of transclusion objects. For each transclusion object the following steps have to be carried out (see also figure 4.4):

- resolve the object and retrieve the information on both the transclusion and on the source document from the database;
- check if the given URL of the source document can be loaded;
- if it can be retrieved check if the metadata, i.e., the creation and modification dates as well as the MD5 hash values, have changed;
- if the fingerprint of the source document is valid retrieve the resource and extract the portion of text determined by the start and end positions of the transclusion;
- replace the transclusion object in the article with the transcluded content;
- if any of the operations above fails insert an apologetic error message.

Every transclusion is formatted in a way that readers can distinguish between authentic and transcluded content. In figure 4.5, transcluded text is highlighted using a light-gray background. Transclusions are complemented with a hyperlink to the original source of the content.

4.4 Issues Encountered

During the implementation and evaluation of the prototype a number of difficulties were experienced. A few substantial issues are addressed in the following sections.

4.4.1 Javascript Restrictions

As described in section 4.3.4, this implementation relies on Javascript code that detects which portions of a document are selected by the user; when the user presses a button, the start and end positions of the selection are determined.

Restrictions imposed by the security mechanisms of most modern web browsers (e.g., [Microsoft 2005b]) prevent Javascript functions from accessing selections in "foreign" frames and documents. This means that the button that reads the user's selection has to be present in the same frame as the selection.

Since a premise was that transclusions can be made from any HTML document on the Web, it has to be ensured that the Javascript code required is inserted into any page the user wants to transclude. The approach in the current implementation is to use of a non-transparent proxy application. So when users enter the URL of the page they wish to transclude, the page is not loaded directly by the web browser but by a CGI script on the server that acts an HTTP proxy. The CGI script appends the demanded Javascript code and sends the document to the client.

The proxy application could be omitted if transclusions were only made in documents from an internal repository such as an online journal or a content management system. The system generating the documents could automatically insert the essential Javascript code when the resource is requested, for example, with a particular parameter.

4.4.2 Browser Specific Implementation

The function for accessing the user's selection poses yet another problem. Different implementations of the corresponding function exist in the various web browsers available today. In the Mozilla family, the selection can be accessed through the `document.getSelection()` method, whereas in Internet Explorer, a dedicated `document.selection` object has to be used (e.g., [Koch 2004]).

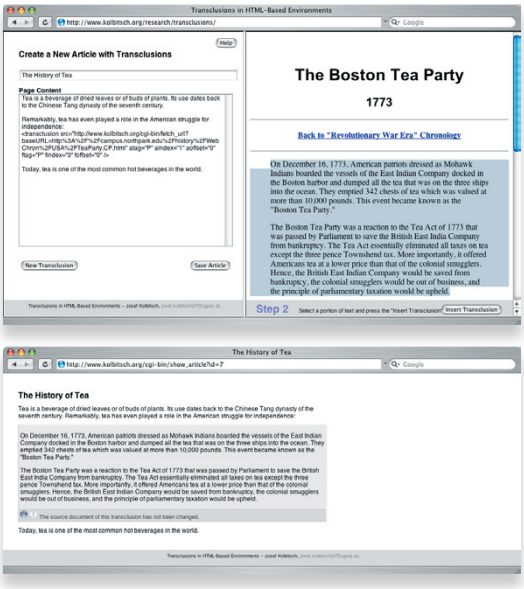


Figure 4.5: Screenshots from a prototype of transclusions in an HTML-based authoring environment.

Due to the use of the `document.getSelection()` method in this implementation the prototype is only compatible with Mozilla-based browsers. With minor modifications in the client-side Javascript code, however, the prototype should work with a wide range of web browsers including Internet Explorer.

4.4.3 Modified Documents and Unavailable Resources

Similar to broken links in web pages, documents that are modified and resources that become unavailable can pose a problem for transclusions. One reason for this deficiency is the use of Uniform Resource Locators on the World Wide Web (URLs, [Berners-Lee 1994]).

URLs identify an object *and* describe its physical location. Defining the physical location of a document

determines that only one instance of the document may exist at a time. Different resource identification and allocation mechanisms allow for multiple locations of the same document, i.e., several instances of the same document may exist in different physical locations. When a resource with a certain object identifier is requested, it is retrieved from one of the locations that retain a copy (e.g., [Pam 1995]). This can, for instance, be the location with the fastest network connection, the one with the lowest load, or the one with the shortest distance.

The design of Xanadu takes a similar approach, in which a resource may exist in several locations (e.g., [Nelson 1999]). Thus, when a transclusion is requested and one instance of the source data becomes unavailable it is retrieved from another repository containing the same information.

For this implementation, an analogue mechanism is proposed that makes use of the Wayback Machine (e.g., [Koman 2002]), a very large archive of currently about forty billion web pages, and local caching. When a transclusion is requested whose source document has undergone major changes or has become unavailable, the Wayback machine is queried for the resource. The query includes the URL and the creation date of the transclusion as access date.

Alternatively, a local cache or Google Cache can be employed. Local caching means that a copy of a resource has to be made when it is transcluded; the local copy is retained in an internal repository of the system. In case of local caches, however, legal issues may arise. [Bahr 2002], for instance, discusses whether services such as Google Cache are in conflict with German copyright laws.

Figure 4.6 illustrates the suggested retrieval strategy for transclusions. When the original source of a transclusion is available and has not been changed, it is retrieved from the original location. Otherwise an attempt is made to load the page from the Wayback Machine or from a similar cache. If this attempt fails as well, the user is notified that the transclusion cannot be made at this time.

4.5 Discussion

The implementation of transclusions in a purely HTML-based environment has shown interesting perspectives, and various aspects need to be investigated in detail. A few selected topics are pointed out in the following sections.

4.5.1 Robustness

The prototype presented in this chapter offers ease of use and relative overall stability. The robustness, however, can still be improved. Under certain conditions, for example, transclusions can be imprecise. Content transcluded from a document by the "Extract and Merge" component can be slightly different from what a user originally selected—a few characters too many or too little are extracted.

An issue that generally affects the robustness of the implementation and demands in-depth analysis is modified content. The shortcoming partly arises from an optimization that improves the system performance. When modifications in the source document

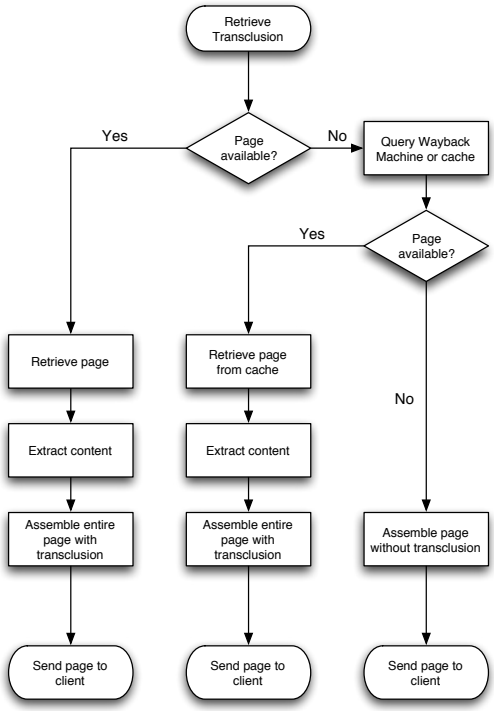


Figure 4.6: Flow chart of a document retrieval strategy where the source document of a transclusion can become unavailable or can be modified.

of a transclusion are to be detected, only the creation and modification dates as well as the content length in the HTTP header of the resource are scanned. Some servers do not return these values at all, though, and a small percentage of hosts return invalid date values. So if creation and modification dates or the content length are not available, the entire resource is retrieved and an MD5 hash value is generated. When the content to be retrieved is very large, the system load is high or the network connection is slow, it might take too long to calculate the hash value. In this case, the process might terminate with a time-out signal, and the transclusion cannot be made.

As pointed out above, modifications in transclusion sources are a general problem. Especially dynamic content such as pages from content management systems and from digital libraries can be critical. In many cases, these documents contain advertisements or other

frequently changing information such as references to the most recent articles. Although the actual content of the document is not altered, the system component that analyses the state of transclusion sources would detect a modification.

It is desirable to have modifications in documents and their importance—was only an advertisement changed or has the meaning of the article changed?—detected automatically. However, this functionality is presently computationally not feasible. [Krottmaier 2002] suggests leaving the decision to the user: despite the modifications the content from the altered source document is transcluded, and the user has to determine if the transclusion and the context are still appropriate.

In any case, authors of modified transclusion sources should be notified that the content they virtually included into their articles might not be suitable anymore, and that it has to be reviewed.

4.5.2 Aspects of the Design

The design of the implementation, the use of a transclusion object in particular, open up exciting opportunities. Since the transclusion object is associated directly with the source document of a transclusion, it is possible to determine which other articles in the system include content from the same source. This information indicates that the corresponding articles might deal with a similar topic and that they could be of interest for both authors and readers. More importantly, this information denotes that the authors of these articles might work in a similar area. In a scientific setting, for instance, these authors can be researchers working on similar projects. Thus, information exchange can be enabled. From a more general perspective, collaboration can be fostered and organisational knowledge management can be facilitated (e.g., [Maurer and Tochtermann 2002]).

Therefore a simple function like *"Which other articles transclude the same document?"* or *"Who else uses the same document?"* is proposed that can help readers and writers discover new information.

This principle can be applied in the "opposite direction" as well. Authors can easily find out which other articles in the system transclude the articles they produced. This information can basically be used for the same purposes as pointed out above. Hence, another function that complements every article in the system is proposed: *"Which articles in the system transclude this article"* or simply put, *"Who transcludes 'us?"*

In a more sophisticated approach, the system could pro-actively point out resources and authors that are related to the article being displayed.

4.5.3 Aspects of the Proxy Application

This implementation relies on a non-transparent HTTP proxy application that makes it possible to insert a small portion of Javascript code into every page the user wishes to transclude. Although the application was initially intended for a very specific purpose, its design is so flexible that it a whole range of other, largely unrelated applications become feasible.

Blacklisting of words and hyperlinks, highlighting of text and dynamic insertion of annotations are just a few simple examples. Advanced techniques may include dynamic adaptation of content and on-the-fly insertion of complementing information (see chapter 3).

4.6 Conclusion

This paper briefly outlined Ted Nelson's notion of hypertext and one of its prime concepts—transclusions. Although HTML has been influenced by the notion of transclusions for the inclusion of external objects such as images, they have not been implemented consistently. Therefore a number of proposals have been made on how to implement transclusions with the technologies available today. A few of the most important approaches have been discussed.

Based merely on the technologies provided by a web-based environment, in this chapter a system was designed that offers users to author articles that may contain transclusions. A first prototype utilises plain HTML, Javascript and server-side components including CGI scripts and a specialised HTTP proxy application.

Although a number of issues were encountered during the implementation phase, valuable results were collected that promise that robustness and stability of this implementation can be imcreased.

The innovative design of the transclusion structures as well as the architecture of system components open up new perspectives and can lead to more advanced functionality. Facilitating information discovery, proactive dissemination of related content and the stimulation of community-building are only a few possibilities among others.

After this successful implementation of text-based transclusions the next chapter deals with an implementation of transclusions of multimedia documents on the Web.

5 Fine-Grained Transclusions of Multimedia Documents in HTML

A version of this chapter was published in the Journal of Universal Computer Science, Volume 11, Issue 6, pp. 926-943.

Publication submitted in June 2005.

In principle, transclusions are used in HTML for the inclusion of entire text documents, images, movies and similar media. However, the HTML specification only takes transclusions of entire documents into account. Hence it is not possible, for instance, to include a part of an existing image into an HTML document.

In this chapter, fine-grained transclusion of multimedia documents on the Web are proposed, which presents a logical realisation of the concept of transclusions in HTML. The proposal makes it possible, for instance, to include sections of existing images or small portions of entire movies into HTML documents.

Two different approaches to implementing the functionality presented are detailed. The first architecture is based on a transparent extension module to conventional HTTP servers, whereas the alternative design makes use of a CGI program. Both approaches are fully self-contained, reside on an HTTP server and do not require browser plugins or any other special software components to be installed on client computers. An amendment to the HTTP specification is not required either. A prototype implementation demonstrates the proposal for a number of document types.

5.1 Introduction

As mentioned in section 4.1.3, HTML makes use of transclusions in various aspects. Markups such as `<iframe>`, `<img>`, `<object>`, or `<embed>` virtually include content like entire HTML files, images, video clips and animations into HTML pages by means of linking. However, transclusive functionality in HTML is very limited. The transclusion mechanisms available can only be applied to entire documents, and fine-grained transclusions such as a small spatial selection of an image are not implemented. This is most likely due to concessions made in favour of ease of implementation, in order to reduce the computing power required, and to increase performance on the Web.

Therefore fine-grained transclusions of multimedia documents as a logic extension of to the infrastructure used in HTML-based environments are proposed. This makes it possible, for example, to include only a part of an existing sound file into a web page. The proposed functionality is an addition to the work presented in the previous chapter. By combining these two projects, it becomes possible to transclude text documents and a wide range of multimedia documents in conventional HTML pages at the finest level of granularity and without the need to additional install software on client computers.

Section 5.2 describes for a number of media types how fine-grained transclusions can be implemented. Two approaches to implementing this functionality are introduced in section 5.3 along with a working prototype. Section 5.4 gives an overview of several application areas.

5.2 Fine-Grained Transclusions of Multimedia Documents

The concept of fine-grained transclusions can be applied to numerous content types. Therefore, this section briefly explains a variety of multimedia document types, and their use in connection with transclusions is addressed.

The syntax used in all examples resembles the syntax for conventional HTTP GET requests and the common HTML syntax, for instance, of `<area>` markups.

5.2.1 Drawings, Vector Graphics

Drawings and vector graphics are used very often in the technical domain and when abstract concepts are graphically depicted. Examples are CAD drawings such as plans of buildings, cars and machinery in general; both two- and three-dimensional models of objects such as molecules in chemistry; flowchart diagrams, and organisational charts. A common property of most vector graphic formats including SVG (scalable vector graphics, [SVG 2003]) is that all vectors and all other objects are stored in way that they can be addressed independently. This means that even after saving a document, a line can still be selected as a line, and its length or position can be modified.

The same is also true of some formats that are employed to describe three-dimensional models or "virtual reality" scenes. The recent X3D format, for instance, stores every object separately in the file, and it is possible to modify every object individually (see [X3D 2005]).

These characteristics can be made use of when creating transclusions. Fine-grained transclusions of vector graphics can be based on spatial selections or on object selections. Thus when a transclusion is created, the user can select either certain objects or a region of the drawing. The component generating the actual transclusion (see below) has to select the given objects from a drawing and subsequently interpret, i.e., render, the resulting data. In case of a spatial selection, the vector graphic has to be interpreted first (window-viewport transformation), and only then a selection can be made.

Object-based selection is most likely only reasonable when a small number of objects have to be dealt with. Real-world models often contain millions of polygons, which makes selecting groups of objects impracticable. In such a scenario, spatial region selction seems to be more appropriate. The first example in table 5.1 describes the transclusion of a rectangular shape 170 points wide and 20 points high, starting at 10 points from the left top corner.

5.2.2 Photos and Images

Photos and rasterised images are particularly wide spread on the World Wide Web. They are usually produced by conventional photo cameras, specialized cameras such as infrared cameras, and other imaging devices such as ultrasound and x-ray detectors or radar units. Diverse areas such as medical imaging, satellite imaging, microscopy, (print) publishing make use of rasterised images. Some technologies for the description of three-dimensional models and virtual reality scenes such as Apple's Quicktime VR utilise rasterised information as well (e.g., [QTVR 2005]). A series of images are stored and displayed in a particular way so that users have the impression that they are viewing a three-dimensional object. (The "object" is, however, merely a recombination of two-dimensional images from different perspectives to a new two-dimensional image.)

For rasterised images it is usually not possible to select separate objects, only pixels or regions of pixels can be addressed. A more general technique is the usage of normalised device co-ordinates instead of pixels (NDC; e.g., [Foley et al. 1997]). With NDC, the left-bottom corner of the image is described with the co-ordinate pair (0,0) and the top-right corner corresponds to (1,1). This method makes references independent, for example, from actual devices and implementations.

A straightforward approach for creating fine-grained transclusions of images and photos are selections based on regions. When users want to make a fine-grained transclusion of an image they can mark a certain (spatial) region of the image, where regions can be rectangles, other geometric objects and arbitrary shapes defined with polygons or Bézier curves.

The examples in table 5.1 list two types of spatial selections. The fist one is a simple rectangular selection 200 pixels wide and 200 pixels high, starting at coordinates (10, 20). The second example is an arbitrary polygon consisting of six co-ordinate pairs.

5.2.3 Video Content

Most digital video content is represented as a series of frames, where each frame is a rasterised image. Therefore basically all methods that can be utilised with images and photos are also applicable to video content including spatial selections using areas of pixels or normalised device coordinates. In addition to this, two further region selection mechanisms can be identified:

- temporal selection and
- spatio-temporal selection.

When using a temporal selection, the user specifies a certain period of time on the time line of a video clip in order to denote that a temporal region of a movie is to be transcluded. The combination of a spatial and a temporal selection leads to a spatio-temporal selection. This means that a fixed area in one frame of a video clip is selected also in a number of consecutive frames. This method is useful, for instance, when only a certain area and a certain section filmed by a surveillance camera are relevant.

Table 5.1 provides two examples for temporal and spatio-temporal selections. The first video file is a temporal selection with a length of 4 minutes and 59 seconds, starting at 3 minutes and 42 seconds. In the second example, an area 320 pixels wide and 200 pixel high is cropped, and a temporal selection with a

length of 47 minutes and 12 seconds, starting at the beginning of the video file, is made.

5.2.4 Sound and Music

A number of different approaches for describing sound and musical data exist. Most technologies available to date such as AIFF (see [AIFF 1989]) are frame-based. All current audio formats include some sort of timecode, and some of them support the use of index points.

Therefore temporal region selection is the apparent method for providing fine-grained transclusions of sound files. Users can define a start position and either the length of the content to be transcluded or the end position. As illustrated in table 5.1, positions and lengths can be given as absolute time codes (e.g., `1m13s`), as index points (e.g., `index1`), or as number of frames where applicable (e.g., `1825f`).

5.2.5 Compound Multimedia Documents

Multimedia animations are much more complex than the media types introduced above. Although they can be seen as compound documents they can usually not be treated as composites of basic media types. One reason is that the most common documents formats such as Macromedia Flash are more or less proprietary (e.g., [Macromedia 2005]). This means that even if a video clip within a multimedia animation is represented as an individual object in the document format, it might not be possible to access it separately or to extract it. Thus, even relatively simple operations such as cropping the area to be displayed might be very hard to implement.

Despite these difficulties and in order to underline the fundamental concept, it should be mentioned that, at least in theory, it is possible to make fine-grained transclusions of compound multimedia documents in HTML. Examples are:

- spatial, temporal, or spatio-temporal selections of entire animations;
- selections of single objects or a group of objects within complex animations; or
- spatial, temporal, or spatio-temporal selections of individual objects or a group of objects of complex animations.

An actual implementation, however, would most probably require active support from the companies maintaining the respective document format.

5.3 Approaches to an Implementation

Fine-grained transclusions are not only an abstract concept in hypermedia theory but there are several potential application areas that are particularly well suited for the proposed functionality (see section 5.4). Therefore two approaches to an implementation are introduced, the design goals are detailed, and a prototype implementation that can be accessed over the Internet is presented.

5.3.1 Design Goals and Requirements

The idea of fine-grained transclusions attempts to foster the reuse of information that is already available on the World Wide Web. A wide range of differing media formats should be supported, and their characteristic features should be made use of. Therefore a number of design goals have to be considered:

- use of any document on the Web: it should be possible to make fine-grained transclusions of any media document accessible on the Web—not only of files from closed repositories;
- extensibility: a system for creating and retrieving fine-grained transclusions should be extensible, i.e., adding support for new document types should be taken into consideration;
- flexibility: the system should make use of the features offered by the various media types, e.g., the types of selections (spatial region, object selection, etc.) should be adapted to the document and encoding types;

Doc. Type	Selection	URL
Drawing	rectangular shape	`image.svg?shape=rect&coords=10pt,10pt,180pt,30pt`
Image	rectangular shape	`image.jpeg?shape=rect&coords=10,20,210,220`
	arbitrary shape	`image.jpeg?shape=poly&coords=10,10,30,30,50,10,50,90,30,50,10,90`
Video	temporal region	`movie.mpeg?start=3m42s&length=4m59s`
	spatio-temporal region	`movie.mpeg?shape=rect&coords=0,20,320,220&start=0m0s&end=47m12s`
Sound	start/end positions	`song.aiff?start=0m0s&end=1m13s`
	relative length, time	`song.aiff?start=0m0s&length=1m13s`
	relative lenght, frames	`song.aiff?start=0m0s&length=1825f`
	index points	`song.aiff?start=index1&end=index3`

Table 5.1: Examples of selections in different multimedia document types.

- use of existing standards: the definition of fine-grained transclusions should not require the introduction of new HTML markups;
- reuse of existing infrastructure: it should be possible to reuse existing infrastructure including browser plug-ins. Moreover, a seamless integration into existing frameworks must be provided, i.e., the use of fine-grained transclusions of multimedia documents must not break existing plug-ins, etc.;
- no additional software: users should not have to install additional software on their client computers to be able to make use of the proposed functionality;
- transparency: an implementation should be transparent to both web browsers and users;
- ease of use: for authors, it should be easy to create fine-grained transclusions of media documents.

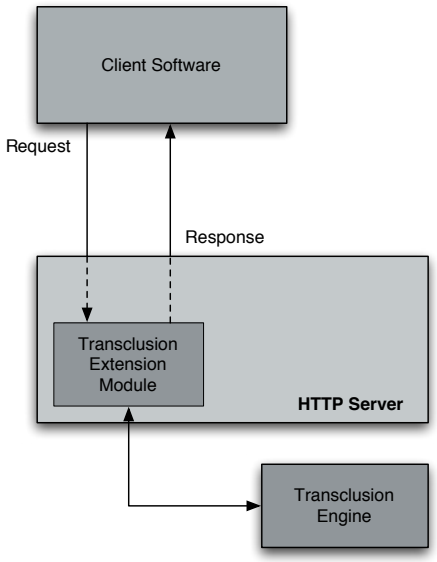


Figure 5.1: Architecture of an approach, where the functionality of the HTTP server is extended with an external module.

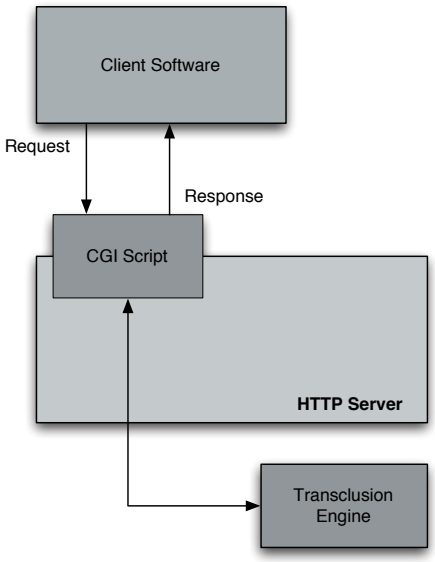


Figure 5.2: System architecture of a CGI-based approach. The client sends a request to a CGI script that, in turn, forwards the request to the transclusion engine.

An implementation can be based on plain HTML, without any additional markups, and plain HTTP, without the need to alter the communication protocol used between client and server. The basic requirement is a conventional HTTP server that can be extended with either external extension modules or CGI programs. Most current web servers such as the popular Apache HTTP server incorporate both capabilities (e.g., [Apache 2005]).

5.3.2 System Architecture

Two implementations are proposed. An implementation that involves an extension module for the HTTP server, and a CGI-based approach (see sections 5.3.3 and 5.3.4). Independent of the actual approach used, the system follows a particular architecture that is described in the following paragraphs.

A fine-grained transclusion is defined by a number of additional parameters that describe which portion of the content is to be extracted. They are appended, for instance, to the `src` attribute of the `<img>` tag in an HTML document (see table 5.1 for a number of examples). The parameters defining the fine-grained transclusion can be provided "manually" by the author, can be generated using dedicated authoring tools, or a simple web interface.

When a web page containing fine-grained transclusions is requested, a component on the server (extension module or CGI program) analyses the definition of the fine-grained transclusion and retrieves the original document. Subsequently, the specified portion of content is extracted from the original document using an appropriate plug-in module for the corepsonding document type and sent to the client (see figues 5.1 and 5.2 and section 5.3.5).

Thus, the implementation follows a client-server approach, where the complexity and logic lie within the server-side component.

The major benefit of this approach is that it is completely transparent to HTTP clients. Furthermore the current infrastructure for playing back sound and videos already present in most web browsers can be reused. No additional software components need to be installed on client computers. A video clip, for instance, can still be included into a web page by means of the `<object>` markup. Only the URL given in the corresponding tag has to be altered—it has to define the piece of media to be transcluded.

Another advantage is that the content does not have to be stored in any particular database or special repository. Content can be stored as files in directories of conventional filesystems, but also in remote filesystems, in databases, etc. The architecture presented also avoids unnecessary network usage by only transferring the data actually requested from the server

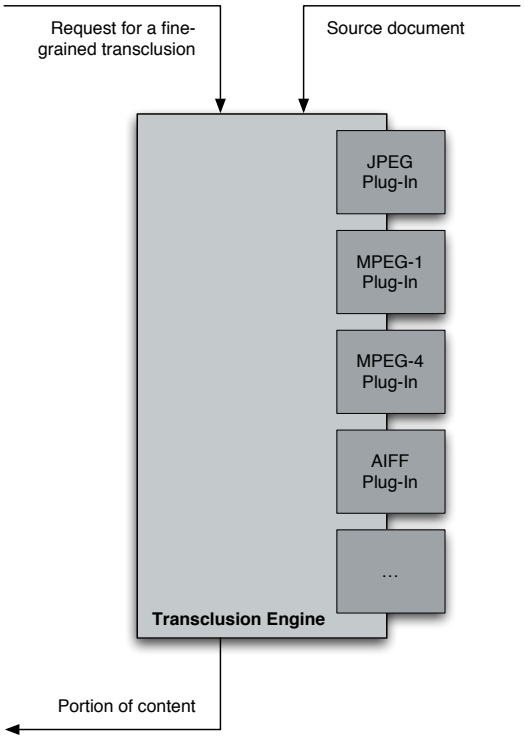


Figure 5.3: Basic structure of the plug-in architecture. The "transclusion engine" receives a request, fetches the given source document and processes it with an appropriate plug-in module. The extracted content is returned.

to the client.

This generic structure of the system is the foundation of the implementations of the server-side components detailed in the following two sections.

5.3.3 Extension Module for HTTP Servers

The functionality of many HTTP servers including the widely-used Apache web server can be extended by means of modules (see [Apache 2005] and [Thau 2003]). Modules can perform various tasks from authentication to enabling access to databases and handling various error conditions. Frequently employed modules include `mod_perl`, a module that implements an interpreter for the Perl programming language, and `mod_php`, an extension for using the PHP scripting language.

This proposed variant for implementing fine-grained transclusions relies on an extension module that is invoked whenever data is requested from the server (see figure 5.1). It analyses the request, and if multimedia content is requested, the module checks whether the entire file is requested or only a part thereof, i.e., if a fine-grained transclusion is to be made. If so, the extension module checks if a plug-in for the corresponding document type exists (see section 5.3.5) and generates the requested portion of content. Finally, the content selected from the entire multimedia document is returned to the HTTP client.

If the HTTP request does not refer to a multimedia document, the request is passed on to the HTTP server. If a plug-in for a given media type is not present or the transclusion cannot be made, an error message is returned to the client.

This approach is completely transparent to both HTTP clients and users. Only the parameters defining the portion of content to be transcluded can be "seen" by users; otherwise a fine-grained transclusion can, for instance, not be distinguished from a conventional inline image. The disadvantage of this approach is that an extension module has to be installed on the HTTP server. In environments where the configuration of the HTTP server cannot or must not be changed, this implementation is not favourable.

5.3.4 CGI Program

Similar results can be achieved by using a CGI program. When the CGI-based approach is utilised, every fine-grained transclusion has to be loaded through a particular CGI program (see figure 5.2). The CGI program carries out the same operations as the HTTP server extension module: it checks if the request is valid, finds an appropriate plug-in, extracts the requested portion of content, and sends it to the client.

The difference becomes obvious when looking at the structure of an HTML tag referring to a fine-grained transclusion. In listing 5.1 an example is illustrated. The first markup is a traditional `<img>` tag. The second markup defines a fine-grained transclusion that is loaded through an HTTP server extension module (cf., table 5.1). It is transparent to users. The third markup is the same fine-grained transclusion loaded through a CGI program. Although it is transparent to web browsers—they do not "recognise" that they do not load the original image—it is not transparent to users because they can see that an intermediate component is used when the image is retrieved.

Since CGI programs can almost always be added to web servers, even if extension modules cannot be installed, this approach can be advantageous when the setup of servers cannot be modified. Moreover the CGI program can be implemented in a way that not only local documents can be transcluded but also files from remote URLs (see section 5.3.6). So even users

that do not have the CGI program installed on their servers could make use of fine-grained transclusions.

5.3.5 Plug-In Architecture

Both the extension module to HTTP servers and the CGI programs are mere frameworks that offer the functionality needed to handle the requests of clients. The requested portion of content is generated by plug-in modules.

When a client retrieves a fine-grained transclusion of a certain document the HTTP server extension module or the CGI program analyse the request. If the requested document is a multimedia document and the request is valid, an appropriate plug-in that can handle the given document type (e.g., a GIF image) is searched for. If a plug-in can be found the original document is retrieved and passed on to the corresponding plug-in along with the parameters describing the content to be selected. The plug-in extracts the demanded portion of content and returns it to the HTTP server extension module or the CGI program. The basic structure of the plug-in architecture is depicted in figure 5.3; the retrieval strategy with plug-ins is shown in figure 5.4.

Although this approach may seem overly complex, it has various advantages. Every file type can be realised as separate plug-in. Therefore it is fairly easy to offer support fort new document types or encoding standards. The approach is flexible in that the plug-ins can take the particularities of certain media types into account and deal with their peculiarities.

The implementation of a plug-in is relatively uncomplicated for simple, open document formats such as JPEG/JFIF and PNG images or MP3 sound files. The prototype illustrated in section 5.3.6, for instance, implements a plug-in module for JPEG, PNG and GIF images. The implementation is somewhat more complex, though, when media types such as MPEG-1 are to be handled.

The encoding is frame-based, where frames only contain the segments different from a particular base-frame. So when a temporal selection of an MPEG-1 video file starts at an arbitrary frame, it might be necessary to reconstruct the entire frame first. An implementation becomes even more difficult when proprietary document formats such as Microsoft's WMV-encoded video files or complex media files such as animations based on Macromedia Flash are to be handled.

5.3.6 Prototype Implementation

The prototype implemented follows the CGI-based approached introduced above. It consists of two components: a script for generating a URL that can be used to include the fine-grained transclusion into an HTML document, and a second CGI program that actually produces the requested portion of content.

Currently, users can make fine-grained transclusions of three frequently employed document types—JPEG, GIF and PNG images. It can be assumed that a small number of content providers will create fine-grained transclusions, and a large number of users will retrieve them. Therefore only a very simplistic interface for that task is provided for authors. When users want to create a fine-grained transclusion they are asked to supply the URL of the original file as well as the coordinates that describe the region to be extracted (see figure 5.5). Only rectangular selections can be made; more complex geometric or even arbitrary shapes are not supported at the moment. In the course of time, tools that simplify the generation of fine-grained transclusions and offer enhanced functionality should be made available.

Upon submitting this data set to the server, the first

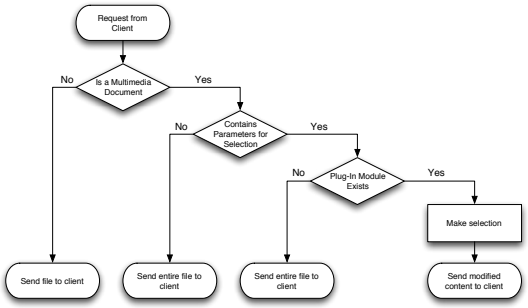


Figure 5.4: Retrieval strategy for multimedia documents with the ability to make fine-grained transclusions.

CGI program creates a URL and displays an example of how to use the URL to integrate the fine-grained transclusion into HTML documents. Whenever a fine-grained transclusion is retrieved, i.e., an image whose source URL contains certain parameters is requested, the second CGI script is invoked. It loads the original file from its location and extracts the requested spatial region of the image according to the given parameters (see also listing 5.1). Finally, a file with the selected region is sent to the client.

The prototype implemented impressively demonstrates the power and ease-of-use of the proposed technology. Along with several examples, it is available online at [Kolbitsch 2005b].

A drawback of the current implementation is that it is realised as CGI application. Hence, it is not transparent to users. This means users can recognise that images are loaded through a CGI program. An advantage of the approach is, though, that even users that do not have the two CGI programs installed on their HTTP servers can make use of the functional-

ity because the software allows the use of images with remote URLs.

A further shortcoming of the implementation is that original resources have to be loaded onto the server that hosts the CGI scripts because image files need to be present in order to be able to crop the desired region. This makes retrieval of a fine-grained transclusion of an image rather inefficient and slow. Moreover it causes increased network traffic.

Therefore, ideally the CGI program for retrieving the content resides on the same server as the content. In this case, downloading the original file can be reduced to a simple "file open" function call.

An alternative approach to increasing performance is the use of cachine in a way similar to the method detailed in sections 4.4.3 and 4.5.1 (see also section 5.4.4). When a transclusion of a document is made for the first time, the server caches the resulting transclusion as well as a fingerprint of the original document. On subsequent requests, the cached results are sent to the client if the fingerprint of the original document is unchanged. Whenever the fingerprint of the original resource changes, a new transclusion of the document is made.

5.4 Application Areas

Fine-grained transclusions of multimedia documents in web-based environments can have numerous applications. The prototype implementation, for instance, is designed as part of the encyclopaedic environment described in chapter 2. In this system, users have the ability to make text-based transclusions in order to quote, refer to, and re-assemble existing articles. With the implementation of this proposal they will also be able to make transclusions of multimedia content at a high granularity.

The following sections give a brief overview of further, potential applications in differing areas from museums to news and learning environments.

5.4.1 News Providers

News channels usually retain a large number of audio and video interviews. In many cases, they are not made available on the web-site of the news providers because files are too large, because the content is simply too long, or for economic reasons.

With fine-grained transclusions news providers could store one large file with the original interview in their internal database. On their web-site they insert a fine-grained transclusion of the original content. The transclusion is, for instance, a temporal selection of an audio file that contains the most significant key issues of the interview. The entire interview (in full length) can, for example, be made available only to subscribers of the news service.

Thus news providers have hardly any extra effort in creating the short version of the interview, more or less no additional disk space is required, and providers are able to offer a complimentary service, while having the capability to retain full control of the content.

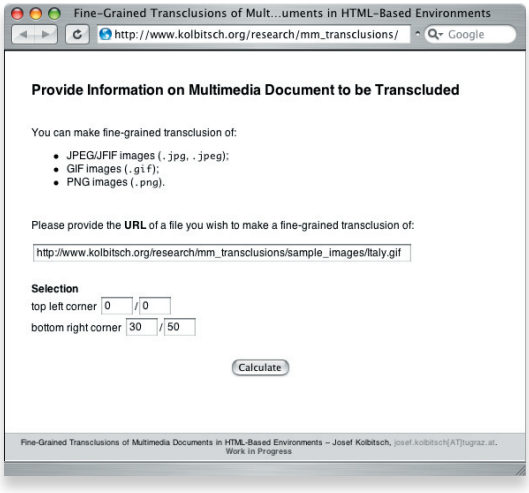


Figure 5.5: Screenshot of the prototype implementation. The user has to provide the URL of an image and the coordinates necessary to select a spatial region.

5.4.2 Galleries and Museums

Visitors in museums and galleries frequently encounter situations in which they require professional help:

- they look at paintings and do not know what to "see";
- they look at paintings and do not know how to interpret them;
- they are not familiar with the artist's life and social surroundings;
- they do not have adequate historical background;
- etc.

Therefore audio guides, and more recently PDA-based multimedia guides, are readily available in museums in order to assist visitors in their striving for comprehension.

Basically the same is true of people looking at paintings in online galleries or online museums (e.g., [Getty 2004]). With fine-grained transclusions it becomes possible to store the original painting *once*—as high quality image. When certain parts of the painting are explained, a transclusion of a spatial region is made, and the selection is described in detail. Thus, it is sufficient to store an image once, and when only a small section is referred to, a transclusion can be made.

5.4.3 Learning Environments

Similar to museums and galleries, web-based learning environments can utilise fine-grained transclusions

to explain details of drawings and other media documents. In medical education, details of X-rays can be explained thoroughly, in biology in-depth descriptions of parts of photos resulting from electron microscopy can be given, etc.

In case of a sound file, only one minute of an entire opera can be transcluded into an e-learning lesson. The selection contains the key scene of the work and is supplemented with the historical background, information on the composer, and a reference to the score of the opera.

In addition to this, lesson authors could virtually include a part of an external multimedia resource into a local lesson. A lesson for highschool students that explains a chemical experiment, for instance, can transclude a particular temporal selection of a related video available on the server of a university.

5.4.4 Movie Archives and Music Stores

In large archives of movies such as the Prelinger Archive (see [Prelinger 2005]) fine-grained transclusions can be employed for producing trailers and previews of the content on-the-fly. An entire movie, for instance, is stored in the system. When a user requests a description of the movie, a temporal or spatio-temporal selection of the movie is made and transcluded into an HTML page that gives a summary of the movie, lists a number of comments and provides a rating.

The same technology can also be employed in online music stores such as the iTunes Music Store (e.g., [iTunes 2005]). Previews of songs are created from the original sound file when users request them. Although the short preview clips would usually not be stored they can be retained within the system in order to increase performance. So when a preview is requested for the first time, the fine-grained transclusion is generated and cached in the system. On subsequent requests, the transclusion is not newly generated but the preview is retrieved from the server cache instead (see also section 5.3.6).

Note that the use of cached files does not limit the generality of the concept of fine-grained transclusions. The relationship between the transclusion and its cached version is identical; caches are merely utilised in order to reduce loading time and improve overall system performance (see also [Nelson 1996; Nelson 1999]).

5.5 Conclusion

In this chapter, an approach to introducing fine-grained transclusions of multimedia documents in HTML was presented, which constitutes a consistent extension to the concept of transclusions in HTML introduced in chapter 4. With this technology, users have the ability to virtually include only small regions of images, only short clips of an entire movies, etc. into their web pages.

Two approaches to an implementation were discussed—a CGI-based implementation and an extension module for HTTP servers. Both methods offer the same functionality in terms of fine-grained transclusions while the levels of transparency vary, and their integration into existing systems (HTTP servers) is different. The suggested plug-in architecture offers an easy way to make an implementation open to third-party developers and facilitate the support of new document types.

A prototype implementation proves that the concept is reasonable and that the proposed technology can make reusing multimedia documents on the Web more efficient. Various application areas from news to museums and learning environments can have immediate benefits from the techniques presented in this chapter.

6 Anonymous Authorship in Wikipedia

A version of this chapter was submitted to the Journal of Computer-Mediated Communication.

Publication submitted in March 2006.

In Wikipedia, the free online encyclopaedia, every reader can become an author and edit content—even anonymously. In this chapter, the role of anonymous authors in Wikipedia is investigated. A set of data is collected for a pseudo-random sample of articles. The data includes the number of edits and the number of unique registered and anonymous users per article. First, this information is utilised to make general statements on the authorship in Wikipedia. In a second step, articles are arranged in classes in order to determine if articles covering topics perceived as "delicate", awkward, socially controversial, or socially taboo show significantly different characteristics in authorship than "regular" articles. Finally, articles are arranged in categories such as science and politics. The categories are analysed to find out if anonymous users work more actively in particular topics.

6.1 Introduction

The vast amount of information available in Wikipedia, the collaborative online encyclopaedia described in section 1.4, is developed by its vigorous community whose aim is to create a knowledge repository of unprecedented depth and breadth. Although Wikipedia's approach may seem chaotic and incapable of yielding structure and quality content, many people including students, researchers and journalists use it as a regular source of reference.

A major part of Wikipedia's popularity is due to the fact that users have an opportunity to contribute their knowledge and participate in the development of the encyclopaedia. Everyone who detects an error or has information enhancing an article can easily modify existing content. Wikipedia is open to *all* users and they do not even have to register—users can edit articles anonymously. Many articles are, in fact, authored by a majority of anonymous users.

As the case of John Seigenthaler demonstrates anoymous, though, users are also pestering the community (see section 1.4.2.3). Thus it could be argued that collaborative systems should only let identified users edit content, and that anonymous users be restricted to viewing content. However, the importance of anonymous authors within the collaborative environment is unclear and has not yet been investigated.

Information on anonymous users is an essential aspect in understanding how collaborative environments work. Therefore in this chapter, anonymous authorship in Wikipedia is analysed and an attempt is made to determine if anonymous users are an important part of the community. Section 6.2 gives an overview of previous research that has been done on Wikipedia, and section 6.3 gives and introduction to anonymity in information systems. The methodology and aim of this analysis are described in section 6.4. Section 6.5 presents results and provides an interpretation.

6.2 Previous Work

With its unconventional approach, its ambitious aims, the large user base, and the effects that can occur, Wikipedia offers a basis and an incentive for extensive research. Wikipedia openly supports researchers by making data required for analyses available to the public. The content of all articles including their version histories can be downloaded freely from the Internet.

Although the data sources for analyses are readily accessible, comparatively little research has been done on Wikipedia. The following sub-sections highlight several studies that address Wikipedia.

6.2.1 Wikipedia Statistics

An overview of the Wikipedia project, its history, "size" and structure is given in [Voß 2005]. The interesting aspect of this investigation, however, is an analysis of data from the official Wikipedia statistics (e.g., [Zachte 2006]). Voß, tightly involved in the German Wikipedia project, points out primarily positive facets of Wikipedia such as its volume, its flexibility and topicality. He shows, for example, that Wikipedia's growth in terms of the number of articles is initially linear but becomes exponential after a short of time. In addition, the sizes of articles, the number of edits per article, and the number of registered users in the system are examined.

6.2.2 Wikipedia and the Media

Wikipedia articles are regularly cited in the media. In order to analyse the impact that media coverage has on the development of Wikipedia articles, [Lih 2004] establishes a possibly debatable set of metrics for automatically determining the quality of Wikipedia articles and applies them to entries referred to in the media. The metrics for describing the quality of an article are defined using its "rigour" (total number of edits per article) and its "diversity" (total number of unique users per article). The basic assumption in this study is that high values for rigour and diversity correspond to high quality of the content of articles. This means that more edits lead to a more detailed and more refined article. A larger number of unique users, on the other hand, usually suggests that more users can agree with the content of an article, and its accuracy is therefore higher.

Lih's analysis finds that Wikipedia articles have, on average, received increased attention after citations in the media. This results in a significant rise in quality according to his metrics.

6.2.3 Social Phenomena in Wikipedia

The fact that basically everyone on the Internet can edit content in Wikipedia, makes it vulnerable to malicious edits, commonly referred to as vandalism (see also section 1.4.2.4). *History flow* is a technique developed by [Viégas et al. 2004] for visualising the history of modifications of Wikipedia articles and for making patterns in the authors' editing behaviour explicit.

History flow visualisations for a set of articles illustrate interesting effects in Wikipedia. Acts of vandalism including mass deletions and insertions of vulgarities, for example, occur frequently, but they are usually corrected within minutes by reverting articles to their previous version. The visualisations can also impressively illustrate social phenomena such as "edit wars", where a number of paragraphs of content are repeatedly modified and reverted, or inserted and deleted, by two users or groups of users.

The history flow technique can be employed to depict the growth and stability of Wikipedia entries. This method also demonstrates that articles continue to undergo changes so that content usually does not stabilise over time. Furthermore, articles tend to grow and frequently do not converge in size.

6.2.4 Qualitative Comparison

In December 2005, *Nature* published a study comparing the quality of entries in Wikipedia and Encyclopædia Britannica (see [Giles 2005]). The content of 42 science articles present in both encyclopaedias was manually analysed and compared by experts.

The findings of the survey showed that Encyclopædia Britannica contained, on average, about three inaccuracies per article, whereas the corresponding articles in Wikipedia contained about four. In both encyclopaedias, four misinterpretations of important concepts were discovered. While in Encyclopædia Britannica 123 factual errors, misleading statements, and omissions were found, Wikipedia contained 162.

These results indicate that the content of Wikipedia's science articles meets similar standards in terms of completeness and accuracy as entries in Encyclopædia Britannica. Nevertheless, many of the experts reviewing Wikipedia articles noted that their structure was poor and sometimes confusing.

6.2.5 Formality of Content

A recent investigation was concerned with the style and volume of articles in online encyclopaedias including Wikipedia, community-based Everything2, and editor-based Columbia Encyclopedia (see [Everything2 2006; Columbia 2006]). The aim of the project was to find differences in the content of encyclopaedias with different approaches to the development of content (see [Emigh and Herring 2005]).

By comparing a selection of articles from the three encyclopaedias, the formality of content was analysed. For every article, the number of words per article and the number of characters per word were calculated automatically. Moreover, the number of terms and phrases indicating a scientific style (concise, formal) as well as the number of expressions denoting a rather prosaic style (lenient, colloquial) were determined. In addition to this, a manual qualitative analysis of the style and professionalism of articles was performed.

[Emigh and Herring 2005] conclude that the automatically generated data such as the word length suggests that Wikipedia articles are, in their style, close to professional, editor-based encyclopaedias and "better" than other free online encyclopaedias. In terms of sty-

listic features and formality, Wikipedia could not be distinguished from the editor-based Columbia Encyclopedia. In contrast to other statements, these findings imply that the collaboratively authored content in Wikipedia is written professionally and conforming to high stylistic standards (cf., [McHenry 2004]).

6.3 Anonymity in Information Systems

Many state-of-the art information systems allow users to access and modify information anonymously. This section gives an overview of the various levels of anonymity currently in use, outlines the impact of anonymity, and describes the approach to anonymity in Wikipedia.

6.3.1 Varying Levels of Anonymity

In most modern information systems, and computer systems in general, varying levels of anonymity are implemented. They range from a level of identity where users are identified physically to true anonymity, where users cannot be identified or traced at all (cf., [Flinn and Maurer 1995]). Which approach is actually employed in a system usually depends on the level of security required, the sensitivity of information involved, and the "openness" of the system.

6.3.1.1 Physical Identification

Users wishing to employ a system or a service with physical identification mechanisms have to give proof of their physical identity. Hence, in such systems users are not anonymous at all. The identity is established on the basis of physical features of users such as fingerprints, irises, or similar biometric characteristics (e.g., [Liu and Silverman 2001; Ma et al. 2003]).

One copy of the users' biometric information is retained in the system as reference. Whenever a user wishes to make use of the service, the same biometric feature has to submitted to the system (e.g., using a fingerprint scanner). Only if the two sets of biometric data match access to the service is granted.

It is commonly believed that physical identification makes "identity fraud" impossible and prevents unauthorised access to systems and services (e.g., [Liu and Silverman 2001]). Such an approach is usually utilised in highly sensitive systems such as power plants, banks, or security agencies, where anonymous access is unwanted.

6.3.1.2 Strict Authentication

Another concept that attempts to prevent anonymous use of systems involves strict authentication techniques. Only after a user's identity has been verified by a certified authority a set of credentials is issued. These credentials can include username-password pairs, key cards with digital signatures, and one-time-password generators.

Whenever users want to employ the system, they have to provide the credentials assigned to them. Since the system can associate the credentials with a real person every action in the system can be related to an individual. However, this approach does not prevent users from passing their credentials on to other users.

An example is online banking, where a person has to be registered with a bank in order to get the credentials for accessing the online service—a one-time-password generator, for example. If the user's one-time-password generator is stolen, the thief can access the online banking service. (In this case, the system would still identify the thief as the owner of the account.)

6.3.1.3 Lax Authentication

Lax authentication is a method commonly used on the Internet. Users can freely choose a (unique) username and a password. The system stores this tuple, and whenever users want to employ the given service they have to submit their username and password.

This approach permits partial anonymity because neither the physical identity of users is verified nor is an attempt made to prevent passing on the credentials to someone else. A person with the username "ted.nelson", for example, is not necessarily Ted Nelson, the hypertext pioneer. Even if "ted.nelson" was Ted Nelson he would be able to give his password to one of his assistants, and consequently the assistant could author a wiki article, for instance, under Nelson's name.

6.3.1.4 Pseudo Anonymity

Pseudo anonymity describes the situation where users have the impression they work anonymously but the system is still able to track them or (at least vaguely) identify them. Users work pseudo-anonymously when they read web-sites such as CNN.com. Although users do not have to provide a username or similar information for reading articles on the web-site, CNN.com's web server could trace and identify readers using their IP addresses.

For every piece information transmitted over the Internet, both the sender's and the recipient's IP addresses have to be attached to the actual content in order to enable routing. Since IP addresses usually have an owner (an internet access provider, for instance) they can, in many cases but with considerable effort, be traced back to an individual.

In addition to IP addresses, data such as HTTP cookies or session keys stored in hidden form fields

of web pages can be employed to identify users on the Web.

6.3.1.5 True Anonymity

Users working truly anonymously on the Internet cannot be traced or identified at all. This implies that the "obstacle" of being identified on the basis of IP addresses (see section 6.3.1.4) has to be overcome. Several research projects on anonymising users' actions on the Internet have been conducted, and a number of software packages for anonymising Internet connections are available (see [Reiter and Rubin 1997; Berthold et al. 2000] and [Anonymizer 2006; JAP 2006]).

6.3.2 Impact of Anonymity

Currently, most systems on the Internet are based on lax authentication or pseudo anonymity, some services make use of strict authentication, and hardly any systems utilise physical identification or offer true anonymity (cf., [Flinn and Maurer 1995]). This means that most services on the Internet allow for at least partial anonymity.

As previous research shows, anonymity is an important aspect in many Internet-based services (e.g., [Joinson 2001]). In many cases, users posting information on the Internet anonymously or under a pseudonym disclose more facts about themselves than in settings where their identity is known. Students using pseudonyms in discussion forums, for example, are generally less inhibited than in classroom discussions. They overcome their shyness and are more confident, more extrovert, and talkative (see [Chester and Gwynne 1998]). Aspects that might be obstacles in face-to-face communication including language, looks, and social differences or racial prejudice are not relevant in computer-mediated communication.

Medical information systems are another example where anonymity is popular. Patients might feel uncomfortable disclosing symptoms and other personal information when they discuss a particular disease online. Therefore discussion forums on diseases, for instance, tend to attract more users when questions and articles can be posed anonymously (cf., [Maurer 1988]). Research even indicates that patients give more accurate descriptions of their medical conditions and that they reveal more personal details when they are interviewed by a computer instead of a face-to-face interview.

However, anonymity comes at a price. It does not only enable information exchange between people who might otherwise choose not to communicate with other people. Anonymity and pseudonyms are also employed by vandals, trolls (people posting rude or offensive information), and spammers (users posting advertisments and similar unwanted information). In many discussion forums, flaming is a common practice (cf., [Burnett and Buerkle 2004]). In Wikipedia, on the other hand, articles with defamatory content are often authored by anonymous users, and also other malicious edits are performed anonymously.

6.3.3 Anonymity in Wikipedia

As indicated above (see section 6.1), the Wikipedia user model includes lax authentication and pseudo anonymity. On the one hand, authors in Wikipedia can register, where registration requires users to choose a username and a password. Whenever users wish to work as registered authors, they have to log on to Wikipedia using their usernames and passwords. Any subsequent modifications to content in the encyclopaedia are associated with the users' names.

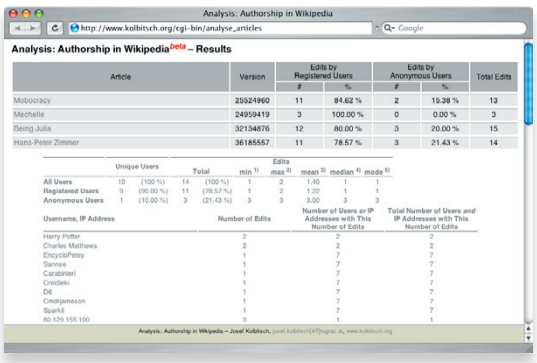


Figure 6.1: Screenshot of the web-based tool used for collecting data from Wikipedia history pages.

Users not wishing to register with Wikipedia can author content pseudo-anonymously. Instead of associating usernames with modifications, the authors' IP addresses are retained. The only restriction imposed upon anonymous users is that they cannot create new articles. This measure was introduced as a consequence of the Seigenthaler case in order to avoid the creation of malicious and defamatory content (see section 1.4.2.3).

In this investiagation, "unique users" are considered. A unique user is either a registered user that can be identified by a unique username or an anonymous user represented by a unique IP address (see sections 6.4 and 6.5).

6.4 Aim and Methodology

As mentioned above, Wikipedia allows modification of articles by both registered and anonymous users, and it is unclear what role anonymous users play in authoring articles or if they are an important part of the Wikipedia community (see section 6.1). In an at-

tempt to foster the understanding of user communities, this study addresses authorship in Wikipedia by comparing editing behaviour of anonymous and registered users for a selection of Wikipedia articles.

The basic hypothesis is that articles covering "delicate", awkward, socially controversial, and socially taboo topics are more likely to be authored by anonymous users. Therefore two sets of "regular" and "delicate" articles were selected from the English language edition of Wikipedia and their authorship was analysed. A total number *n=346* articles (0.035 % of all articles) were sampled. They were assigned categories (topics) and classes (see below). For each of the sampled articles a range of features including the number of edits and the number of unique users per article were calculated automatically.

#	Category	#	Category
1	All Articles	8	Miscellaneous
2	Computer and Technology	9	People and Society
3	Drugs	10	Politics
4	Economy and Law	11	Religion
5	Entertainment and Sports	12	Sexuality
6	Humanities and Arts	13	Science
7	Medicine and Diseases	14	Terrorism

Table 6.1: List of categories used in this study.

The aim of this paper is to find patterns in the collected data, particularly deviations between registered and anonymous users. The results were used for determining if anonymous users are predominantly editing certain categories or classes of articles. The following sub-sections explain how articles for the analysis were selected, how the data was collected, and which data analyses were conducted.

6.4.1 Selection of Articles

The 346 Wikipedia articles utilised for this analysis were selected using five mechanisms. Articles were chosen in the following order (articles chosen in one "class" could not be chosen in any subsequent class):

- 47 articles chosen by the author: the author of this thesis selected 47 articles randomly;
- 76 "delicate" articles: 76 articles that might be perceived as "delicate" or socially controversial (see above) selected by both the author and five volunteers;
- 52 articles chosen by volunteers: the volunteers were asked to collect a set of Wikipedia articles covering arbitrary topics. The volunteers were told that they did not have to read the actual articles and that neither the volume of articles nor their correctness or completeness mattered;
- 76 "special" articles: a number of articles were selected from Wikipedia's "Special" pages (see [WPspecial 2006]). They include a selection of the longest, oldest and shortest articles, several featured articles, a number of the articles with most revisions, a range of most referenced and most visited articles, as well as articles lacking sources;
- 95 random articles: 95 articles were chosen using Wikipedia's „Random article" function (see [WPrandom 2006]).

Articles are chosen by the author and by volunteers in order to obtain a selection of articles that are of interest for actual users. The combination of volunteers and the author is used as a minimal measure to reduce bias in the manual selection of articles. Articles retrieved using Wikipedia's random article function represent truly arbitrary articles that might not be of actual use to readers (e.g., article stubs, pages clarifying ambiguous terms, etc.). Special Wikipedia pages are included in the study in order to find differences between ordinary and uncommon, atypical articles—very long or very old entries, for instance.

6.4.2 Classification and Categorisation of Articles

For these analyses, articles are arranged into classes and categories. Article *classes* correspond to mechanisms that were used during the selection of articles. They include "chosen by author", "chosen by volunteers", "Wikipedia random" articles, "Wikipedia special" articles, and "delicate" articles.

In addition to this, the same set of articles were assigned one of the following *categories:* computer and technology, drugs, economy and law, entertainment and sports, humanities and arts, medicine, miscellaneous, people and society, politics, religion, sexuality, science, and terrorism (see table 6.1). These categories reflect "delicate" topics such as drugs, sexuality, and terrorism as well as generic categories such as people and society or science.

This approach makes it possible to have two different views of the same set of data. On the one hand, the set of delicate articles can easily be compared to random articles and articles selected by the author and by volunteers. Alternatively, a range of topics can be analysed.

6.4.3 Tools

Since Wikipedia employs version tracking, not only the most current versions of articles but their entire history is retained. This means that every article in Wikipedia has a history page that lists all modifications that were made to the page content. History

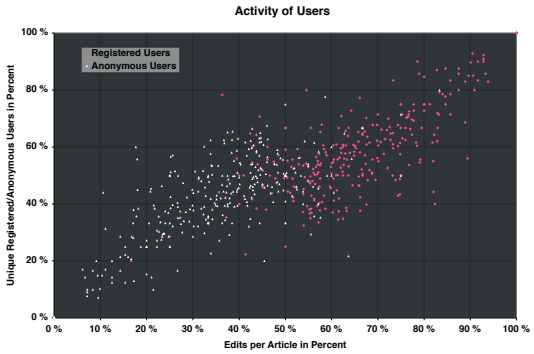


Figure 6.2: This scatter plot illustrating the distribution of edits per article in relation to the percentage of unique users shows that the percentage of edits is approxiamtely proportional to the percenatage of users. Registered users tend to perform more edits on articles (top right quadrant) than anonymous users.

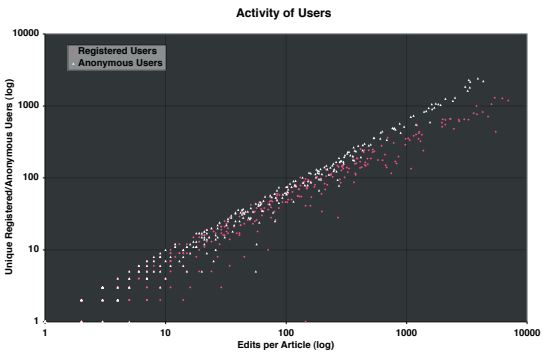


Figure 6.3: Number of edits per article vs. number of unique users following a power law on a log-log scale.

pages include the usernames of registered users or the IP addresses of anonymous users for every edit.

For this analysis, a web-based tool for collecting the required data was implemented. In this tool, the user provides a list of URLs of history pages of Wikipedia articles as input. Subsequently, the software retrieves the entire history of every article in the list, determines the names of registered users and IP addresses of anonymous users, and finds out how many edits were done by each user. The output data of the application contains the number of unique users (in absolute numbers and percent), the number of edits (in absolute numbers and percent) as well as the minimum, maximum, mean, median, and mode values of edits per unique user for each page. The data is calculated for registered users, anonymous users, and all users (registered and anonymous users combined; see figure 6.1).

The current version of the tool is available online at [Kolbitsch 2006a]. It can be used as a tool for collecting data for upcoming research projects on Wikipedia.

The information generated by the web-based application is imported into Microsoft Excel. Excel is utilised for the analysis of the data and for producing the graphs.

6.4.4 Data Analysis

The key values utilised in this analysis are the number of edits per article and the number of unique users per article. The analysis is limited to simple methods such as the comparison of mean or median values and frequency distributions. These techniques are sufficient for proving or rejecting the hypothesis detailed above.

Since a first evaluation shows that the collected data is afflicted with positive skewness, median values of the characteristic features are preferred over mean values. In this case, median values are a more appropriate measure to describe a "typical" article in a class or category. However, the qualitative results are the same for mean and median values. Hence, only the quantitative information is different, while the general trends of mean and median values are similar.

In order to analyse the class of delicate articles, its features (such as edits per article, edits per user, and percentage of anonymous users of delicate articles) are compared to the four other classes. In addition to comparing articles classes, the same analyses are performed on article categories.

6.4.5 Repeated Analysis

Wikipedia is a highly flexible and volatile system. Thousands of existing articles are edited every day, and therefore even within a short period of time a shift in authorship trends may occur.

In order to determine the impact of short-term fluctuations in authorship, the data for this analysis was collected twice. In a second run, sixteen days after the initial data was obtained, the same set of data for the same list of articles was retrieved. The results of the comparison of the two data sets are outlined in section 6.5.2.1 (see also table 6.2).

Data Set	Value	Unique Users (reg'd)	Edits (reg'd)
Feb 9th	Mean	245.42 (64.09 %)	615.68 (69.75 %)
	Median	38.50 (59.84 %)	75.00 (66.67 %)
Feb 25th	Mean	260.60 (64.04 %)	656.89 (69.87 %)
	Median	42.00 (60.00 %)	81.00 (67.27 %)

Table 6.2: Comparing unique users and number of edits per article from the data initially collected and a second run 16 days later. Values in brackets show the percentage of registered users.

Although five year old Wikipedia is a relatively mature system, it is a known fact that communities sometimes undergo long-term changes; they show a different "behaviour" when they get older (cf., chapter 1). Therefore the results of this analysis can only describe the current state of Wikipedia articles. More light will be shed on the long-term development of authorship in Wikipedia by repeating this study in the future.

By "going back" in time and examining old article histories, a long-term analysis could be simulated. However, this poses the problem that some of the articles in this study might not have been created at the point of analysis in the past.

6.4.6 Methodological Concerns

The methodology for this analysis has some shortcomings that should to be pointed out. They include the actual approach to the study, the selection of articles, and the way anonymous users are identified.

6.4.6.1 Approach to the Analysis

The approach to this study is suboptimal because only a subset of articles is covered. The best method for conducting this analysis would be to generate characteristic values for *all* articles in Wikipedia and find commonalities in the values collected. These patterns could then be used to describe authorship trends in Wikipedia. However, this approach is not practicable because it would require the assignment of appropriate categories to every article. Moreover, the analysis could be performed fully automatically, which poses a problem when dealing with such large numbers of articles.

Instead, for this analysis a pseudo-random sample was selected and used to prove or refute the initial hypothesis—that "delicate" articles are more likely to be written by anonymous users. Selecting a subset of articles, however, can introduce bias and lead to a distortion of both the quantitative and the qualitative outcome of the study. Although articles were not only chosen by the author but also by volunteers and Wikipedia's random article function, some bias may remain.

6.4.6.2 Anonymous Users

Unlike usernames, IP addresses as they are employed in Wikipedia are not an adequate means for identifying anonymous users. When using a proxy server or when network address translation (NAT) technology is utilised, for example, one IP address can be shared by several users. Thus one IP address corresponds to more than one unique user. With dial-up connections, DHCP services and similar technologies, on the other hand, users may be assigned different IP addresses every time they log onto the network. Therefore one unique user corresponds to more than one IP address.

Reliable data on how many actual users correspond to one IP address are not readily available. Therefore it is assumed in this study that, on a large scale, the two effects—dial-up connections and NAT—tend to cancel each other out and that one IP address represents one user. If this assumption proves not to be valid, the analyses involving the number of unique users (e.g., edits per unique anonymous user) may be incorrect. However, analyses based on the number of edits by anonymous users are unaffected.

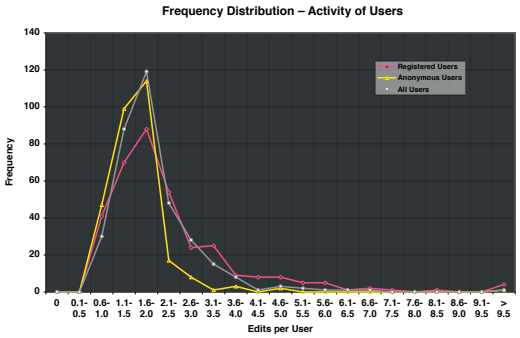


Figure 6.4: Frequency distribution of edits per user for all articles. Most articles are, on average, edited about 1.5-2.0 times by a unique user.

6.5 Results and Interpretation

This section provides qualitative and quantitative results and their interpretation for the collected data and highlights the key findings in a number of charts. The two sets of data together with about fifty additional charts that could not be presented within the scope of this paper are available online at [Kolbitsch 2006a].

6.5.1 Anticipated Results

It is anticipated that articles chosen by the author and by volunteers show similar characteristics because they were both selected manually based on the same process. Significantly different characteristic values for these classes would indicate bias in the selection of articles or methodological shortcomings.

Wikipedia random and Wikipedia special articles might yield different results because random articles can contain anything—from an article "stub" to old articles with hundreds of edits. Wikipedia special articles, on the other hand, should have the most "extreme" features because they include the longest, oldest and shortest articles.

6.5.2 General Findings

Before the results in article classes and categories are discussed, several general findings are addressed. They include the deviations between the two data sets that were collected and the activity of users in the encyclopaedia.

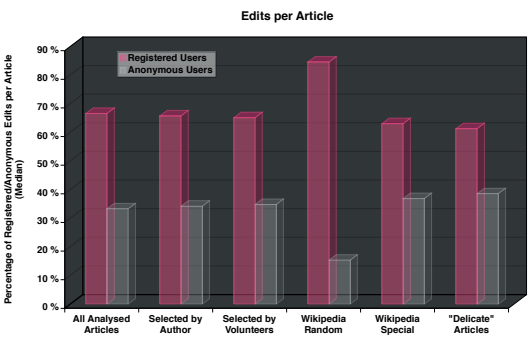


Figure 6.5: Proportion of edits per article for article classes, registered users vs. anonymous users.

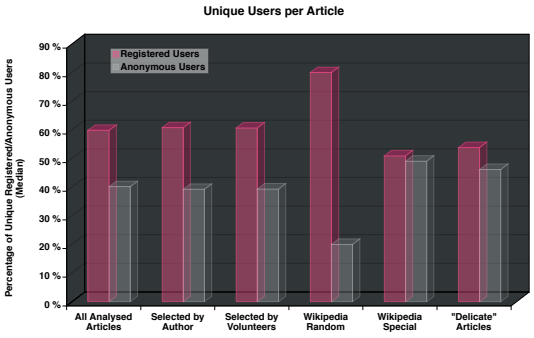


Figure 6.6: Proportion of unique users per article for article classes, registered users vs. anonymous users.

6.5.2.1 Comparison of Data Sets

As mentioned in section 6.4.5, two sets of data for the same list of articles were collected in order to observe short-term changes in article authorship. The main differences (the number of unique users and the number of edits per article for all articles in this study) are shown in table 6.2.

Both mean and median values of the number of unique users and edits per article increased. This finding is trivial because it is known that Wikipedia articles do not stabilise and continue to grow (see section 6.2.3). However, the percentages of registered and anonymous users for both edits and unique users per article remain roughly the same. This implies that the modifications Wikipedia articles undergo in a short term do not have a significant influence on this study. Therefore only one data set, the older one, is used in all investigations.

As an aside, it is interesting to note that 66.18 % of all articles were modified at least once within the sixteen days from the first to the second data collection process. While only 34.74 % of all articles in the Wikipedia random class were edited, 88.16 % of articles in the "delicate" class were modified.

6.5.2.2 Skewness of Data

For all article classes, except for Wikipedia random articles, the mean values of edits and unique users per article were significantly higher than the corresponding median values (see section 6.4.4). This means that in all classes (except for Wikipedia random) a number of entries with considerably more edits and unique users than average exists. A manual analysis of the number of edits per article confirms this finding.

6.5.2.3 Edits per Article

According to Wikipedia statistics, the mean value for edits per article is 24.9 (see [WPstats 2006]). In this analysis, both mean and median values for edits per article are much higher. The mean number of edits for all articles is 615.68, and the median value is 75.00. The mean number of edits per article for entries selected by volunteers, for example, is 294.29. The only class of articles that confirms the official mean value is the class of articles selected by the Wikipedia random function; their mean number of edits per article is 23.65.

It is unclear why such a large discrepancy occurs (factor ten and more). It might be due to the fact that Wikipedia contains a potentially large number of article "stubs" that have not received much attention by authors so far, and therefore greatly reduce the mean value of edits per article. Another possible explanation is the automated mass insertion of articles in Wikipedia. In October 2002, for example, about 36,000 Wikipedia entries covering towns and cities in the U.S.A. were generated automatically by a "bot"—a computer program appearing as registered user in Wikipedia. Since their addition to Wikipedia, these entries may not have been edited often.

6.5.2.4 Activity of Users

For this analysis, the user activity is defined as the quotient of edits per article and unique users per article. It describes how many modifications, on average, are made to an article by every unique user.

The scatter plot in figure 6.2 depicts the percentage of edits per article on the axis of abscissae and the percentage of unique registered and anonymous users per article on the axis of ordinates for all articles. An element in the chart can, for example, define that for an article 60 % of all edits were done by the 40 % registered users that modified the article.

Figure 6.2 also shows that the percentage of edits per article is approximately proportional to the percentage of unique users per article. The highest density of data elements is in the centre of the plot. A large number of registered users can also be found in the top-right quadrant, whereas anonymous users have a concentration in the bottom-left quadrant. This means that reg-

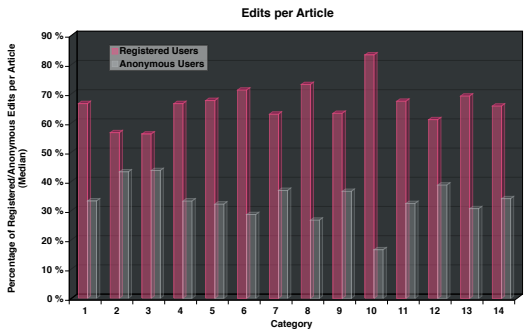


Figure 6.7: Proportion of edits per article for article categories, registered users vs. anonymous users

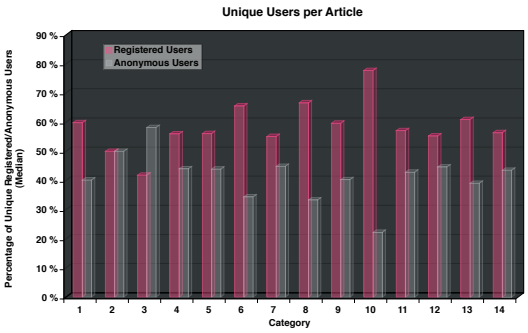


Figure 6.8: Proportion of unique users per article for article categories, registered users vs. anonymous users

istered users are, on average, more active and perform more edits per user than anonymous users.

In figure 6.3, the absolute numbers of edits per article and unique users are plotted on a log-log scale. The distribution of data shows a power law relationship between edits and unique users per article. Data elements corresponding to registered users occur frequently "south" of elements of anonymous users which, again, indicates a higher activity of registered users. These findings are confirmed by [Voß 2005] and [Lih 2004].

The chart in figure 6.4 depicts the frequency distribution of user activity in steps of 0.5 edits per article. The graph is right-skewed (leptokurtic). This shows that most articles in Wikipedia are, on average, edited about 1.5-2.0 times by every unique user. Another small peak can be seen at around 2.5-5.5 edits per user. Such articles have a large number of edits (between a few hundred and several thousand) with a significantly lower number of users. The outlier indicating more than 9.5 edits per registered user is due to an article that was edited 145 times by a single user.

6.5.3 Analysis of Classes

In the analysis of article classes, the median values of edits per article and unique users per article for the five classes are compared. Instead of collating the absolute numbers of edits and unique users, the *percentages* of edits by registered and anonymous users as well as percentages of unique registered and unique anonymous users are employed.

Figures 6.5 and 6.6 show edits and unique users per article for the five classes. Both charts include an average value of all articles as a reference. A first evaluation of article classes confirms the general tendency of registered users being more active than anonymous users, i.e., registered users perform more edits.

Articles chosen by the author and by volunteers exhibit similar values for edits per article and unique users per article (see figures 6.5 and 6.6). Articles selected using the Wikipedia random article function are, indeed, quite different from manually selected articles. Articles in this class have a much larger percentage of registered users and significantly fewer edits by anonymous users than any other class. This characteristic might be partially due to automatic bulk addition of articles (see above). The class of Wikipedia special articles has about ten percent more unique anonymous users and about three percent more edits by anonymous users than average articles in this study. These figures are notably higher than the average values for articles in this analysis, which leads to the conclusion that Wikipedia special entries are of particular interest for anonymous editors. The reason for this deviation is unclear.

Both in absolute numbers (not shown) and in percentages, delicate articles attract more users than average articles in this study. Whereas "average" articles are edited 65.00 times by 38.50 unique users, delicate articles are modified 350.00 times by 182.50 users. Not only the overall number of users is higher, but there is also an increased *percentage* of anonymous users: compared to average articles, delicate articles have about five percent more anonymous edits and authorship includes about six percent more anonymous users.

Both five percent more anonymous edits and six percent more unique anonymous users constitute a highly significant deviation from average articles *(Z_{edits}=3.42, Z_{users}=4.07,* for both *$p<0.001$).* This proves the initial hypothesis that delicate articles are edited by an increased number of anonymous users.

An apparent resemblance of characteristic values for delicate articles and Wikipedia special articles can be observed. This peculiarity cannot be explained at this stage and may require further research.

6.5.4 Analysis of Categories

Basically, the same methods used for analysing article classes are applied to article categories, i.e., median values of edits and unique users per article are investigated. Figures 6.7 and 6.8 illustrate the percentages of edits and unique registered and anonymous users

per article for the categories. The list of categories is provided in table 6.1.

A first glance at the two charts seems to prove the previous findings that articles are, in general, edited by a majority of registered users. However, figure 6.8 shows an interesting deviation: the category on drugs has 58.17 % unique anonymous users. Obviously, users resort to anonymous authorship when it comes to editing articles covering this sensitive topic.

In article categories, the investigation of the proportion of edits performed by anonymous users and the percentage of unique anonymous users do not show a clear tendency and do not yield significant deviations. This leads to the conclusion that anonymous users work zealously on delicate articles but not necessarily more devotedly on the entire category a delicate article belongs to. To put it a different way, an entire category a delicate article belongs to is obviously not of significant interest to anonymous users. The only exception is the category on drugs, as described above.

6.5.5 Anecdotal Note

The tool for collecting the data for this analysis was used for retrieving and extracting information from the *entire* history of Wikipedia articles (see above). When an attempt was made to analyse the history of the article on George W. Bush, Wikipedia reported an "out of memory" error while rendering the HTML output for the entire history. I.e., the list of modifications for this article is so long (more than 80 MB, this is where the server presents the error message) that the server-side Wikipedia software runs out of memory.

6.6 Conclusion

With thousands of registered and anonymous authors, Wikipedia is a highly popular, collaborative environment. Since little research on Wikipedia has been done in the past, hardly any information is available on the editing "behaviour" of its users. The analyses conducted in this paper shed light on this topic and make it possible to make important new statements about authorship in Wikipedia.

In several aspects, deviations between registered and anonymous users can be ascertained. Generally, registered users seem to be more active than anonymous users, i.e., they perform more edits per article. Most articles being edited about 1.5-2.0 times by every single user indicates that users tend to "come back" and do more work on articles.

Every class of articles (except for the Wikipedia random class) has a few articles with a much higher number of edits and users. In terms of statistics, this characteristic becomes evident in the deviations of mean and median values. Moreover, considerably more edits per article were found than the official Wikipedia statistics state. According to the statistics, an article is edited 24.9 times on average, whereas articles in this investigation are changed more than 600 times (mean value).

Even more importantly, statements about the role of anonymous authors in Wikipedia can be made. On average, articles have about forty percent anonymous users, and more than 33 % of all edits are made anonymously. These numbers per se describe the importance of anonymous users. Some delicate articles, though, have even significantly more anonymous users and anonymous edits. Thus, anonymity is important for users when dealing with delicate articles. They can resort to anonymous authorship in order to contribute knowledge and experience to socially taboo or awkward topics without being identified. Examples include articles on sexuality, drugs, terrorism, and abuse.

The same is, however, not true of entire "delicate categories". Although particular articles on sexuality receive more attention from anonymous authors, the entire category of sexuality does not exhibit an increased number of anonymous contributions.

Anonymous users are a very important component of the Wikipedia community, a factor in the popularity of Wikipedia, and therefore it might be recommended to make anonymity part of future collaborative environments. These findings should be taken into consideration when designing collaborative content development environments (see chapter 9; cf., chapter 2).

Part 2 Digital Libraries and Collaborative Systems

Metadata and Taxonomies

7 The Use of HTML Encoded Dublin Core in Academic Web-Sites

A version of this chapter was submitted to the ACM/IEEE Joint Conference on Digital Libraries (JCDL'07).

Publication submitted in January 2007, co-authored by Harald Krottmaier.

This chapter presents the results of a survey on the usage of HTML-encoded Dublin Core metadata in web pages from 48 academic and educational institutions. A total of 118,900 documents are analysed. Of these 118,900 pages 12,881 documents (10.83 %) contain HTML-encoded Dublin Core elements. The majority of these 12,881 pages originate from only four web-sites. Three of these four organisations are from Australia, which is most likely due to a national initiative for promoting Dublin Core metadata at Australian universities. 36 of 48 web-sites contain less than three percent, and 30 of these 48 sites less than one percent HTML documents with Dublin Core elements.

Pages with Dublin Core metadata contain, on average, 6.62 metadata elements. The most frequently used elements in all pages analysed were title, creator, description, language, and publisher.

The use of HTML-encoded Dublin Core does not significantly increase the size of the headers of HTML documents. This means that the overhead produced by Dublin Core in HTML pages can be neglected. Web-sites with higher proportions of Dublin Core elements frequently rely on content or document management systems. However, there is no strong evidence that pages with Dublin Core metadata make use of such systems more often than documents without Dublin Core. Moreover, all pages were checked with the W3C validator software for the correctness of HTML syntax. The findings indicate that documents with Dublin Core elements are valid almost twice as often as pages without Dublin Core metadata.

7.1 Introduction

Dublin Core (DC) is a widely used metadata standard for describing resources on the Internet. It aims at providing a simple, easy-to-use, and easily maintainable format for enabling resource discovery on the Web. The Dublin Core Metadata Element Set is a small set of descriptive elements for characterising the content of document-like objects. DC elements are coarse-grained and not domain specific. This means that authors do not have to use words from a specific dictionary or a particular taxonomy but may describe the content they produce in their own words. DC elements include, for instance, the title of a document, its creator and language (see section 7.1.1).

DC is widely used for describing the content of web pages, the permissions associated with them, and details of their actual instantiations. Probably the easiest way for adding DC elements to web pages is including the metadata directly into the HTML code (see section 7.1.2).

7.1.1 Core Elements

Dublin Core is available in simple (unqualified) and qualified variants. The unqualified version allows for `<attribute, value>` pairs. The qualified version, on the other hand, makes use of additional qualifiers to further refine the description of a resource. In this scheme, elements might be more sophisticated than simple `<attribute, value>` tuples.

The fifteen elements of the unqualified DC cover the content, information related to intellectual property and digital rights, as well as the distribution process. These elements are (see [DC 2004; DC 2005]):

- title: the title of the document-like object;
- subject: the topic of the object's content;
- description: a short description or abstract of the content;
- type: the *"nature or genre"* of the document-like object (e.g., image or physical object);
- source: a resource on which the current document-like object is based;
- relation: a reference to a related object;

- coverage: the time span, physical area, jurisdiction, etc. the object covers;
- creator: the creator of the object;
- publisher: the person or organisation responsible for the publication of the content;
- contributor: people or organisations contributing to the creation of the content (e.g., co-authors);
- rights: the digital rights associated with the object or its content;
- date: a date *"in the life cycle of the resource"* such as publication of the object;
- format: resource or document format (e.g., MIME types for computer media);
- identifier: a unique identifier for the object (e.g., ISBN or URI); and
- language: the language of the object content.

All elements are optional and repeatable. A document, for example, does not necessarily have to have an author (the "creator" element might be missing). If, on the other hand, a document has two authors two "creator" elements are listed in the document's metadata record.

7.1.2 HTML-Encoded Dublin Core

The DC guidelines do not impose a certain format or syntax on authors. This is why there are several ways to associate a DC metadata record with, or embed it into, a digital resource. However, several formats for storing DC metadata are suggested. One recommendation is Dublin Core metadata encoded in RDF/XML or in plain XML (see [DC 2006a]). An alternative way of encoding DC metadata is the use of `<meta>` and `<link>` tags in the headers of HTML documents (see [Powell 2003; Kunze 1999]).

HTML-encoded DC makes use of the `<link>` markup for elements that make references to other resources such as "relation" or "rights". The attribute `rel` defines the name of the DC element and the attribute `href` provides its value. An example is `<link rel="DC.rights" href="http://www.kolbitsch.org/rights.xml" />`.

All other DC elements are encoded using the `<meta>` tag, where the attribute `name` describes the name of the DC element, and the attribute `content` holds the actual metadata: `<meta name="DC.creator" content="Josef Kolbitsch" />`.

Although encoding DC in the headers of HTML documents can be done relatively easily, this approach has several drawbacks. Compared to RDF/XML-encoded metadata, for instance, additional effort for extracting the metadata from documents is required. Another shortcoming is that HTML-encoded DC is not as easily machine readable as metadata in RDF/XML or plain XML.

7.2 Related Research

To the author's knowledge, only two surveys on the actual use of DC have been conducted to date. [Guinchard 2002] presents the results of an investigation on the usage of DC in libraries. A questionnaire containing questions on usage scenarios, DC implementations, and tools employed was disseminated among librarians. 29 organisations, mainly university libraries, responded and provided an insight in the use and usefulness of DC metadata.

Results immediately relevant to the study in this chapter include the techniques used for encoding DC elements and the most frequently used elements (see below): 61 percent of all organisations encode DC in HTML, 48 percent in XML, and 30 percent in RDF. The most frequently used DC elements in the 29 libraries are creator, title, rights, and publisher.

In 2002, an empirical study on the usage of DC elements by 82 data providers listed with the Open Archives Initiative was conducted (see [Ward 2003]). The Open Archives Initiative attempts to facilitate the dissemination of content and the interoperability among content providers (see [OAI 2006]). DC is one of the technologies that help achieve this aim.

In Ward's investigation, the frequencies of the fifteen unqualified DC elements were analysed. The examination yields that the most frequently used DC elements are creator, identifier, title, and date. The least frequently used elements are relation, coverage, and source. As the results detail, about one half of the 82 data providers predominantly make use of only two DC elements—creator and identifier. Furthermore in Ward's study, more than seventy percent of the entire DC metadata volume were due to the five most frequently used elements.

7.3 Aim and Hypotheses

As mentioned above, DC can be utilised for describing any document-like object. Moreover, DC elements can be included directly into HTML files. It is unclear, however, if this technology is widely used and which consequences the use of HTML encoded DC has on HTML documents and their structure.

The following sub-sections describe the motivation and aims for the study detailed in this chapter as well as the research hypotheses. Section 7.4 addresses the methodology and tools used.

7.3.1 Motivation and Aims

The DC projects overview page lists 75 entries of which 30 projects (40 %) refer to academic or educational institutions ([DC 2006b]). Thus, DC is frequently used in academic and educational settings.

Initially, the aim was to do an analysis on the usage of DC in web-sites of academic institutions in Austria. Soon it became clear, though, that it would be necessary to compare the results from Austrian web-sites to similar international sites. Therefore a decision was made to examine the usage of DC in academic and educational web-sites around the globe, but with an emphasis on Austrian and European sites.

The objective of this investigation is to find out how popular Dublin Core actually is in web-sites of universities and similar organisations. Moreover, the implications of the use of DC and an evaluation of which DC elements are particularly popular are of interest. The emphasis of this study is not on complex statistical analyses but on qualitatively answering one question: How popular is Dublin Core in academic settings?

7.3.2 Hypotheses

With the data collected in this study, it might be possible to find evidence regarding the use of Dublin Core in web pages. Hence, the following hypotheses are put forward and and attempt to find support for them is made in section 7.5.

7.3.2.1 Number and Proportion of Pages with DC Elements

As mentioned in section 7.3.1, a large portion of services and sites making use of Dublin Core are academic or educational, indicating that DC might be widely used in these settings. In order to find proof for this statement, the first hypothesis is: *Most of the tested domains will include pages with DC elements.* This means that hardly any domains will consist only of pages without DC elements, i.e., most domains will have at least one page containing DC elements.

7.3.2.2 Overhead Produced through DC

The use of Dublin Core, and metadata in general, consumes space. In the case of HTML-encoded DC, the metadata is stored in the header of the document, which makes it easy to analyse the overhead produced. Hence, the second hypothesis is: *The use of DC will increase the absolute and relative sizes of the headers of HTML documents, where DC elements are stored.*

7.3.2.3 Creation of Pages with DC Elements

By and large, there are two types of software utilised by "average users" for developing web pages: WYSIWYG HTML editors like Macromedia Dreamweaver or Microsoft FrontPage and content management systems like Typo3 or Apache Lenya (see [Dreamweaver 2006; FrontPage 2006] and [Typo3 2006; Lenya 2006]).

Many users employing Dreamweaver and similar applications do not know the intricate details of Web technologies and are most likely not aware of standards like Dublin Core. Thus, if such HTML editors do not insert DC metadata automatically (e.g., by asking for the name of the author or for a description) pages are most likely created without DC metadata. Even if extension modules are available for making the HTML editor DC aware users will most likely not install them because they are not expert users.

For an increasing number of content management systems DC extension modules are available (e.g., [Dunning 2000]). Adding DC functionality to a content management system means that content editors can utilise DC metadata in all documents generated by the system.

Since most HTML editors like Dreamweaver do not provide functionality for easily adding DC metadata to HTML documents, and since the use of content management systems is probably a more convenient approach for this task (cf., [Maurer 1998]), the third hypothesis is formulated: *In this investigation, the majority of pages containing DC elements will be generated by content management systems.*

7.3.2.4 Syntactic Correctness of Pages with DC Elements

All web pages in this study are checked with both the W3C and the WDG HTML markup validation software (see [W3C 2006; WDG 2006]). Since authoring DC metadata usually requires specialised tools or skills (see above), which might increase the "quality" and syntactic correctness of the pages created, the fourth hypothesis is put forward: *Pages containing DC elements will be more often valid HTML than pages without DC elements.*

7.3.2.5 Most Frequently Used DC Elements

As reported in section 7.2, both [Guinchard 2002] and [Ward 2003] provide statistics on the usage of DC elements. The intersection of the most frequently used DC elements from these two statistics is produced and the fifth hypothesis is stated: *The DC elements most frequently used in all pages of this analysis will be creator, identifier, title, and type. In addition to this, metadata that can be determined automatically like "format" will be present frequently.*

7.4 Methodology and Tools of this Study

The following sections detail how Internet domains and web pages were selected. Moreover, the custom-made tools for downloading and analysing web pages for this study are briefly described.

7.4.1 Selection of Sites

Based on Google searches, Internet domains from international academic and educational institutions were selected pseudo-randomly. The list of domains includes 48 organisations. About one half of all web-sites (25 out of 48) are European sites, about one quarter (10 out of 48) are American sites, and about one quarter of all sites (13 out of 48) are from the Asia-Pacific region. This selection of sites from a wide variety of geographic regions make it possible to give evidence on the adoption of DC in different regions.

Between March 19th and May 29th, 2006, a total of 118,900 web pages were collected (see section 7.4.2.1 for details). Pages are grouped by their domain names (e.g., stanford.edu or cam.ac.uk). A full listing of all sites and the number of web pages per site are provided in table 7.1 and Appendix D.

7.4.2 Tools

A set of tools were implemented for retrieving web pages from the World Wide Web and for analysing their content and metadata. This section gives an overview of the tools used in this project and their functionality.

7.4.2.1 Page Crawler

A specialised program for crawling web sites was implemented in Perl. Starting at a given URL, the program retrieves a certain number of documents with the content type "text/html" from the Web and stores them in a relational database. The maximum number of pages to be loaded and the URL range are determined by the user. Calling the page crawler, for instance, with the parameters `http://www.tugraz.at/index.html tugraz.at 5000` makes the software load a maximum of 5,000 pages from the Internet domain tugraz.at starting at http://www.tugraz.at/index.html.

For each page, the URL, the HTTP response header, the entire page content, and the date of retrieval are stored in the database. Pages are downloaded recursively using a breadth first algorithm. This means that a page is retrieved, all links within this page are extracted and also retrieved, etc. The breadth first approach makes it possible to crawl a wider range of sub-domains within a given web-site. Hence, the "diversity" of pages can be increased.

Although the initial aim was ambitious—to download at least 5,000 pages per web site—the number had to be reduced to approximately 2,500 pages. The reason is that the page crawler was programmed to stop loading pages when a given timeout (e.g., 12 hours) was reached for the entire operation. Since the connection to some web-sites was too slow, it was not possible to download the requested 5,000 content within the set time span.

Internet Domain	Pages		Internet Domain	Pages	
	Total	With DC Elem.		Total	With DC Elem.
rmit.edu.au	2505	95,77%	fh-burgenland.at	2534	0,24%
curtin.edu.au	2723	75,65%	ethz.ch	3487	0,20%
uibk.ac.at	3674	73,60%	uiuc.edu	2551	0,20%
usyd.edu.au	2528	72,67%	berkeley.edu	2518	0,16%
hu-berlin.de	1407	36,39%	u-tokyo.ac.jp	2556	0,16%
ox.ac.uk	2519	35,09%	uni-klu.ac.at	1896	0,11%
imperial.ac.uk	2727	32,23%	fh-hagenberg.at	2513	0,04%
princeton.edu	2557	25,73%	hku.hk	2662	0,04%
massey.ac.nz	2535	8,40%	aaue.dk	2769	0,04%
auckland.ac.nz	2685	7,15%	boku.ac.at	5003	0,02%
yale.edu	421	4,75%	uni-salzburg.at	5038	0,02%
otago.ac.nz	2535	3,12%	fh-joanneum.at	2508	0,00%
univie.ac.at	5035	2,84%	fh-kaernten.at	1709	0,00%
uni-muenchen.de	975	2,36%	tudelft.nl	2515	0,00%
tum.de	2515	2,23%	columbia.edu	362	0,00%
tugraz.at	2873	1,57%	stanford.edu	2574	0,00%
cam.ac.uk	1927	1,45%	harvard.edu	609	0,00%
unige.ch	2598	1,35%	upenn.edu	550	0,00%
unibas.ch	2530	0,99%	brown.edu	2545	0,00%
uni-graz.at	5423	0,48%	keio.ac.jp	2532	0,00%
nottingham.ac.uk	2510	0,40%	kyoto-u.ac.jp	2551	0,00%
tuwien.ac.at	1079	0,37%	pku.edu.cn	1354	0,00%
uni-heidelberg.de	2528	0,36%	shu.edu.cn	1582	0,00%
dartmouth.edu	2537	0,28%	ntu.edu.sg	2636	0,00%

Table 7.1: Web-sites analysed in this investigation, number of pages per site, and proportion of pages with DC elements.

Several sites like yale.edu or columbia.edu seem to have detected the crawler, have classified it as a potentially malicious piece of software, and have blocked it from retrieving further pages. Thus, from some web-sites only a very limited number of pages (significantly below the aspired target of 2,500 pages) could be retrieved.

7.4.2.2 Tools for Data Extraction

The purpose of the page crawler is to download web pages from the Internet and to store them in a relational database. As "pre-processors" for the data retained in the database, a set of tools for the extraction of significant information for the analyses were written. These tools gather characteristic values from all web pages and store them in a separate table in the database. The pre-processors include:

- page head extraction: extract the headers of all HTML pages, parse them, and write the entire

page header, the absolute and relative lengths of the header, the unique names of all tags found in the header and their frequencies, the names and values of all tags in the header, and the names and frequencies of all DC elements to the database;
- page validation/W3C: validate all pages in the database using a locally installed version of the W3C HTML markup validation service 0.7.2;
- page validation/WDG: validate all pages using a locally installed version of the WDG HTML markup validator 1.6.2.

The entire data set (both web pages and characteristic values) is about 3.0 GB in size. It consists of approximately 1.2 million database records.

7.4.2.3 Tools for Data Analysis

In addition to the pre-processors, several post-processors were implemented generating numbers that can actually be analysed and interpreted. These tools are utilised for:

- analysis of page validation results: analyse the results of the page validation processes;
- DC element analysis: analyse the number of DC elements;
- simple DC element analysis: analyse the frequency of each of the fifteen simple (unqualified) DC elements;
- head length analysis: analyse the absolute and relative sizes of the headers of HTML pages;
- page head analysis: analyse the frequency of each tag in the page headers. Unfortunately, the usefullness of results from this analysis is limited because of noise. Too many meta elements that are simply "invented" by web designers are present in the page headers (see table 7.2).

Where applicable, analyses are carried out for both all pages of a web-site and all pages of the entire investigation. This makes it possible to determine if individual sites follow general trends.

7.4.3 Methodological Concerns

Although this study was carefully designed, the methodology has some shortcomings. Most importantly, the selection of web-sites is suboptimal. On the one hand, the selection is not truly random because web-sites were chosen using the Google search engine. On the other hand, the selection is imbalanced since the number of domains in each geographic region was determined by the author of this thesis.

On average, approximately 2,500 pages were collected per web-site from 48 Internet domains. However, there are potentially hundreds of millions of web pages in thousands of academic and educational Internet domains. Hence, the sample is too small to make a solid statement about the use of DC in academic settings.

Moreover, the number of pages per web-site is not the same for all domains. This might cause results to be biased.

7.5 Results and Interpretation

This section provides both quantitative and qualitative results and their interpretation. See Appendix D for further selected results.

7.5.1 Use of DC Elements

Out of the 118,900 collected pages a total of 12,881 pages (10.83 %) include DC elements. However, only eight web-sites contain more than ten percent pages with DC elements and only four web-sites account for more than two thirds (69.87 %) of web pages with DC elements. In fact, more than one quarter of all sites (13 of 48) do not contain DC elements at all. 30 web-sites out of 48 contain less than one percent pages with DC elements, and 36 sites less than three percent.

DC Element	Elements	Sites	Metadata Elements	Elements
dc.title	8,827	27	content-type	84,877
dc.creator	8,807	24	keywords	41,877
dc.description	7,237	25	description	35,906
dc.language	7,219	28	author	21,815
dc.publisher	7,027	28	robots	21,640
dc.subject	6,793	24	generator	19,368
dc.format	6,529	27	revisit-after	10,074
dc.identifier	5,515	22	copyright	9,975
dc.date	3,126	24	dc.title	8,827
dc.contributor	3,077	13	dc.creator	8,807
dc.source	2,771	8	imagetoolbar	8,017
dc.rights	2,750	21	pragma	7,781
dc.type	1,757	23	publisher	7,245
dc.coverage	200	9	dc.description	7,237
dc.relation	58	6	dc.language	7,219

Table 7.2: Number of the fifteen simple DC elements in all pages and the number of sites that employ these elements as well as the fifteen most frequently used elements in page headers of all pages.

Although an overall usage of DC of almost eleven percent is quite notable this is mainly due to a few pages with very high proportions of pages with DC elements. While the majority of sites has an almost negligible proportion of pages with DC elements, a small number of sites make extensive use of metadata (see figure 7.1).

Another way to look at the distribution of DC elements across web-pages is to examine the number of DC elements per page. A total of 85,245 DC elements (both qualified and unqualified) were detected. When considering all pages collected, an average of 0.72 DC elements per page can be derived. This number rises to

6.62 if only pages containing Dublin Core elements are analysed. Since these two values are skewed it can be stated that only a small number of pages contain the majority of DC elements, whereas a majority of documents contain very little DC metadata (or none at all).

Based on these findings the conclusion is reached that a many of the Internet domains analysed do not contain a significant amount of HTML encoded DC metadata. *Hence, the results are not consistent with hypothesis 1.*

Generator	All Pages	Generator	DC Pages
typo3 3.7 cms	4,048	teratext database system	2,376
teratext database system	2,376	plone	573
microsoft frontpage 5.0	1,186	oxford university central administration	551
microsoft frontpage 4.0	1,160	tei xslt stylesheets	226
typo3 3.8 cms	589	text encoding initiative consortium xslt stylesheets	45
mshtml 6.00.2900.2668	578	siteswift-cms	44
plone	573	movable type 2.661	37
oxford university central administration	551	cpsskins 2.3	21
mshtml 6.00.2800.1170	402	armadillo 2.0	20
dotnetnuke 3.2.0	360	typo3 3.6 cms	19

Table 7.3: The ten most frequently used page generators (where provided) for all pages in the analysis and only for pages containing DC elements.

Although relatively few web-sites account for the majority of pages with DC elements *all* of the Australian web-sites analysed are among these sites (see table 7.1). The rationale for this geographic imbalance is that Australia founded a national initiative for the adoption of DC metadata at the country's universities (e.g., [Campbell 1999]). According to [Ivanova 2004], two of the three Australian universities examined (rmit.edu.au and curtin.edu.au) are in the stage of advanced implementation of DC, and at the third one (usyd.edu.au), implementation is underway.

7.5.2 Frequency of Simple DC Elements

The frequencies of the fifteen simple DC elements is shown in table 7.2. For a full listing of DC element frequencies see Appendix D.

The most frequently used metadata elements are title, creator, description, language, and publisher with more than 7,000 occurrences each. Only two of these elements (title and creator) are among the most frequently used elements in both [Guinchard 2002] and [Ward 2003]. Furthermore, elements that can be generated largely automatically including "date" and "format" were not among the most frequently used elements. *Therefore, support for hypothesis 5 cannot be found.*

The least frequently used elements are relation, coverage, type, rights, and source with less than 3,000 elements in all pages each. These results are largely consistent with [Ward 2003].

Of the 85,245 DC elements found in all pages, 71,693 elements (84.10 %) were part of the unqualified DC. The five most frequently used elements listed above account for 54.56 percent of all simple DC elements. The two least frequently used elements (relation and coverage) amount to only 0.36 percent, while the five least used DC elements add up to 10.51 percent. These findings, by and large, confirm the results detailed in [Ward 2003] although Ward's numbers are more "extreme".

Table 7.2 also provides a list of the fifteen most frequently used general metadata elements in the headers of the 118,900 HTML documents collected. Among these fifteen elements are four DC elements. Interestingly, unstandardised metadata elements that are either inserted by authors or automatically by HTML editing software are very popular. Elements like "mssmarttagspreventparsing" or "pageservedby" can be found several thousand times (not shown in table 7.2).

7.5.3 Page Head Size

For all pages, the relative and absolute sizes of the headers are analysed. The results are shown in Appendix D.

In general, the relative size of the headers of HTML pages is significantly larger for pages including DC elements than for pages without DC elements *($Z=35.56$, $p<0.001$).* However, Internet domains with particularly many pages containing DC elements do not expose extraordinarily large headers in absolute size. Websites like rmit.edu.au, usyd.edu.au, and uibk.ac.at, for instance, actually have smaller headers than the mean value for all pages in this analysis (888.40, 1,111.21, and 1,057.55 bytes vs. 2,399 bytes, see Appendix D).

Thus, a definite statement on hypothesis 2 cannot be made. Although the headers of HTML pages with DC elements seem to have larger relative sizes, their absolute sizes do not differ from pages without DC elements.

7.5.4 Page Generation

Works as early as [Maurer 1998] describe that document and content management systems can facilitate, and partly automate, the generation of metadata. In order to look into the topicality of this statement the widely-used, but not DC-based "generator" metadata element (`<meta name="generator" content="…">`) is analysed both for pages containing DC elements and for all web pages in this analysis. This metadata field describes the software that was employed for producing an HTML document and is inserted automatically by numerous software products. The list of the ten

most frequently encountered document generators is depicted in table 7.3.

19,720 of all 118,900 pages contain the "generator" metadata element and yield 612 unique names of products. Among the most commonly used document generators are content management systems like Typo3, TeraText DBS, and Plone as well as the WYSIWYG HTML editor Microsoft FrontPage.

Of the 12,881 pages with DC elements 3,971 documents contain information about their generators. The unique names of 34 products were encountered and denote the use of TeraText DBS (rmit.edu.au), Plone (mainly hu-berlin.de), Movable Type (univie.ac.at), and the Text Encoding Initiative guidelines (ox.ac.uk), which implies the presence of some sort of document management system.

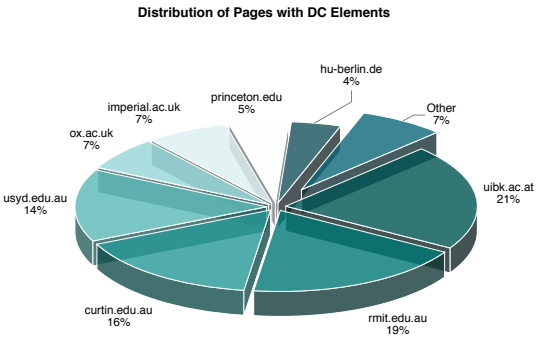


Figure 7.1: Pie chart showing the distribution of all DC elements. The Internet domain rmit.edu.au, for instance, contains 19 percent of all pages with DC elements in this analysis.

Since information on the document generator is missing for a very high proportion of pages, only a vague statement can be made on the use of content management systems and the relation between content management systems and DC. When considering all document collected, at least 8,500 of 19,720 pages that contain data on the document generator (43.10 %) are produced by some sort of content management system. This is in contrast to documents with DC elements, where at least 3,800 of 3,971 pages with document generator information (95.69 %).

Although it is not possible to draw a valid conclusion based on these findings they seem to imply that a relatively high proportion of pages with DC elements are generated by content or document management systems. *Hence although a definitive statement cannot be made, there seems to be weak support for hypothesis 3.*

7.5.5 HTML Syntax Validity

The two tools utilised for validating the syntax of the web pages in this investigation produced quantitatively diverging results. Qualitatively, however, the figures were similar. For the W3C validator, 7.12 percent of all pages, 12.38 percent of pages with DC elements, and 5.78 percent of pages without DC elements show valid HTML syntax. In the WDG validation, 9.22 percent of all pages, 13.66 percent of pages with DC elements, and 7.74 percent of pages without DC elements pass the HTML validation without errors or warnings.

In both cases, pages with DC elements are almost twice as often valid as pages without DC elements. *These findings are supportive of hypothesis 4.*

7.6 Discussion

This survey provides various insights related to the use of DC. First and foremost, DC as a means of describing resources on academic and educational web-sites is not "evenly distributed". While some organisations make extensive use of this metadata standard, most sites only have very few documents that contain DC metadata or none at all. Probably content authors and service providers perceive adding metadata to documents as additional effort and not as a benefit or aid.

As the results underline, campaigns like the national Australian initiative for promoting the DC metadata standard seem to be capable of convincing organisations of the positive effects metadata can have—from enhancements in resource discovery and interoperability to the semantic Web. Hence, it might be necessary to clarify the usefulness of DC, and metadata in general.

Some sites might not use HTML-encoded DC but metadata stored in external files. Since these techniques are not studied in this chapter the actual use of DC metadata might be higher. Further investigation will be required to shed light on this topic.

However, there might be further reasons for not using DC metadata. Probably the use of structured metadata is "outdated" and more unconventional approaches to resource discovery, especially on the Web, have to be considered. Tagging systems such as CiteULike or del.icio.us, for instance, allow readers (not only authors) to attach loose metadata to documents on the Internet and make them available to other users (see [CiteULike 2006; del.icio.us 2006]). Instead of asking authors to find appropriate keywords and other metadata elements, readers can attach the information that is most suitable for them. This concept is rapidly gaining popularity (see the next chapter and sections 1.7 and 1.8).

7.6.1 Additional Findings

In addition to the findings directly concerned with the use of DC, several other statements can be made based on the data collected. The syntactic validity of web pages, for instance, is generally low. Merely 7.12

% of all web pages pass the W3C syntax validation check without errors or warnings (see section 7.4.5). Thus, the vast majority of HTML documents are created with flawed or poorly configured tools, or proprietary markups are still employed. The on average higher validity rates of pages with DC elements might indicate either higher expertise of authors, betters tools, or well-configured software. However, the syntactic correctness of pages with DC elements is also rather rare.

The analysis of the "generator" metadata field in the headers of HTML documents shows that content and document management systems are employed to an increasing extent (see section 7.5.4). Such systems make it easier to insert metadata into a high proportion of pages of a web-site. Elements such as author, date, format, and title can even be inserted fully automatedly. Moreover, content management systems are capable of ensuring the syntactic validity of virtually all pages of a web-site.

Hardly any overhead is produced by the use of DC metadata (see section 7.5.3). For some web-sites with high percentages of pages with DC elements, the size of headers is even smaller than for pages without DC elements. A potential explanation is that a consciously chosen set of well-defined, standardised elements allows authors to define metadata concisely and at a required level of granularity. Proprietary, non-standardised metadata sets, on the other hand, may be ambiguous, non-interoperable, and might require multiple similar definitions in order to achieve the same expressiveness as standardised metadata. These multiple definitions can consume more space.

7.7 Conclusion

The analyses conducted in this chapter elucidate the role of DC metadata in academic and educational web-sites. The key finding from this investigation is that almost eleven percent of all web pages contain DC elements. However, this relatively high proportion is due to the very high usage of DC metadata in a few web-sites, whereas most sites contain isolated pages with DC elements or none at all. Hence, DC metadata does not seem to be widely used in academic and educational web-sites.

While reasons for this relatively low usage of DC metadata are unclear, this study suggests that DC might not have been promoted properly or that DC is probably not the optimal way to associate metadata with web pages. Possibly the use of recent, popular technologies like tagging systems is an alternative. Therefore, the next chapter focuses on aspects of tagging systems and on one of the currently most popular services available in particular—Flickr.

8

WordFlickr

Where Folksonomy Meets Taxonomy

A version of this chapter was submitted to I-media '07.

Publication submitted in January 2007.

Allowing users to publish and share photos on the Internet makes Flickr one of the most popular tagging services currently available. The organisation of images in Flickr is based on Folksonomies, where users attach loose metadata—instead of well-defined terms from a taxonomy—to their images. Although this lowers the barrier to participation it has a number of negative effects and can make searching, for instance, more difficult.

This chapter offers a solution to a particular issue that can be encountered in Flickr—the Vocabulary Problem. The suggested approach is based on the use of a semantic lexical database for Flickr queries. WordFlickr, a prototype implementation of this concept, is presented together with FlickrClustr, a related tool for clustering Flickr search results. The results of first informal tests with these two tools are provided, and characteristics on tag usage can be derived.

8.1 Introduction

Social bookmarking, file sharing, and similar services are becoming increasingly popular. One of the underlying concepts of these services is tagging, which refers to the users' attaching loose metadata to the information items they store (see also sections 1.7 and 1.8.4 as well as chapter 9).

This chapter gives a brief overview of the nature of tagging systems and their shortcomings. In addition to this, a solution to a frequently encountered issue is offered. The following sub-sections describe tagging systems and two popular examples: del.icio.us and Flickr. Section 8.2 summarises previous research on tagging systems. With WordFlickr, a response to the Vocabulary problem in tagging systems, and Flickr in particular, is presented in section 8.3. Section 8.4 discusses WordFlickr's concept and the results of an informal comparison between Flickr and WordFlickr.

8.1.1 Tagging Systems

In conventional libraries all items including books and magazines are stored in a catalogue. For every resource, the catalogue usually retains information on the author, a primary category, optional secondary categories, and a list of keywords, and other metadata. The categorisation in such catalogues is frequently based on strict, hierarchical classification systems, and the keywords often originate from controlled vocabularies. In many of the taxonomies currently in use every category has exactly one parent, i.e., it can be sub-concept of exactly one super-concept.

Examples for well-established taxonomies are the Universal Decimal Classification, the Dewey Decimal Classification, and the ACM Computing Classification Systems ([UDCC 2006; DDC 2006; ACM 2006]). In the latter system, for instance, works in the field of digital libraries are classified as H.3.7.

Traditional libraries employ experts for generating accurate keywords and for classifying resources. A slightly different approach is to let content authors assign keywords and categories to the information they produce, and have experts approve the authors' input.

A recent trend in metadata generation and information classification is the participation of the community. Users may attach loose, unstructured metadata to resources. These metadata fields, named *tags,* serve both as keywords and as "classes". When a user attaches the tags "Sherlock Holmes" and "fiction" to a

resource other users can find the resource, for instance, by searching for "Sherlock Holmes" or by browsing the all resources tagged "fiction". If a tag is assigned to a resource several times by different users its weight increases, and it might be more significant than other tags also assigned to the same information item.

In systems based on tagging, any string can be used as a tag. Hence, metadata creation does not rely on controlled vocabularies and does usually not yield classic, hierarchical taxonomies.

8.1.1.1 Structure of Tagging Systems

A tagging system has at least three types of entities: information items, tags attached to these items, and users creating information items and tags. Every information item has (at least) one author and has usually at least one tag assigned. Depending on the service, items can be web pages, links to resources on the Web, images, video clips, etc. Although tags may, in theory, be arbitrary strings many tagging system restrict the users' freedom and impose rules on the use of tags. In some systems, every tag may only consist of a single word, while other systems allow multiple terms per tag, in some systems tags are case sensitive, while in others they are not, etc.

Relation	Description
synonym	different words with the same meaning
homonym	one word with different meanings (e.g., spring)
antonym	A is antonym of B is A is not B (A=hot, B=cold)
hypernym	A is hypernym of B if B is a *sort of* A (A=fruit, B=apple)
hyponym	A is hyponym of B if A is a *sort of* B (A=apple, B=fruit)
holonym	A is holonym of B if A is *part of* B (A=bud, B=flower)
meronym	A is meronym of B if B is *part of* A (A=flower, B=bud)

Table 8.1: Selected semantic relations in WordNet.

The organisational structures that emerge from the user-authored, free-form tags are social classification systems that resemble taxonomies. However, they are not conventional, professionally created, hierarchical taxonomies with terms from controlled vocabularies. They are flat structures consisting solely of terms chosen by users, where terms do not even have to be words listed in dictionaries.

Although the term is not accurate, many authors refer to the structures in tagging systems as folk taxonomies or *folksonomies* (e.g., [Mathes 2004]). Since this term is disputed, it will not used be throughout this chapter.

8.1.1.2 Benefits of Tagging Systems

Tagging systems share some features of the recently developed technologies detailed in chapter 1. They allow users to participate in the content production process, let *them* assign metadata and classify content. Emerging terms from the youth culture, for instance, can be used in tagging systems with no additional effort. In closed vocabularies, on the other hand, an editorial process is required for adding new terms. This makes tagging systems more flexible than traditional classification systems.

Moreover, tags are a relatively simple way for managing information. They might not only contain traditional classifications but also reminders such as "to read" or personal information such as "wishlist". Therefore the application areas of tagging systems are broader than those of classic taxonomies and metadata systems.

In terms of user base, taxonomies usually do not scale well. The reason is that all users have to be aware of the vocabulary, its usage, and similar guidelines. Hence, special training might be required which is not feasible for a use base of hundreds of thousands of people. One of the strengths of tagging systems is that such a special kind of training is basically not required because there are no guidelines for tag usage.

A limitation of taxonomies is that every item has to be assigned an unabiguous primary category. Additionally, it can be assigned one or more secondary categories. [Golder and Huberman 2005] give the example of an article on cats in Africa. When relying on hierarchical systems, the article is either assigned the sub-category "cats" in the "Africa" category or vice versa; it cannot be assigned both categories at the same time. In tagging systems, however, this is possible because the tag space is flat and all tags are equally important.

The implementations of most tagging systems support serendipity. When an information item is viewed, the name of the author and all tags attached to the item are displayed. By clicking on the user's name or on an tag, potentially interesting and related items are presented (see sections 8.1.2 and 8.1.3).

8.1.1.3 Restrictions and Shortcomings

Tagging systems, however, also have a number of shortcomings. Although tagging lowers the barrier to entry the quality of the classification can be low. Since there are no controlled vocabularies and since users do not even have to agree upon a set of terms, they can choose tags of their liking. These metadata fields might suit their authors' needs but might be illogical and worthless for other users. Such tags include, for instance, the personal tags mentioned above. "To read" might be valuable information for the person who attached tag but might be irrelevant for other people.

Due to the lack of well-organised vocabularies polysemic and synonymous terms might occur, and

vernavular vocabularies might be employed. An example for polysemy is the word "spring" (fountain vs. season vs. suspension, see also table 8.1). The word "apple", used as an example in chapter 1, is both polysemic and synonymous. On the one hand, it can refer to the fruit or to the computer manufacturer. On the other hand, apple, mac, and macintosh can be used as synonyms for computers produced by Apple.

Moreover, the structures yielded by tagging systems are flat (see section 8.1.1.1). Therefore it is not possible to browse information items in tagging systems like the data in directories such as Yahoo's directory ([Yahoo 2006]).

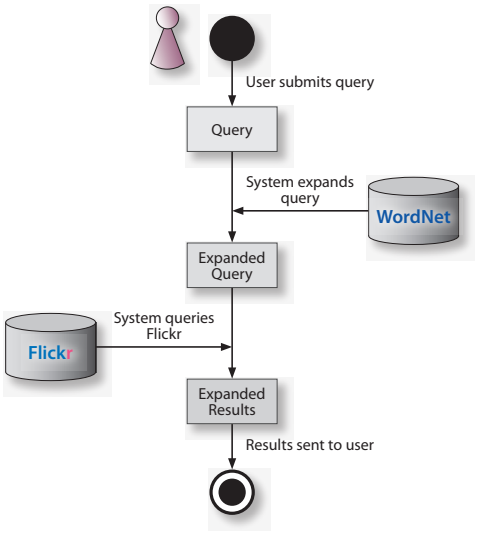


Figure 8.1: Structure of WordFlickr. The user's query is expanded using the WordNet database and is sent to Flickr.

In some services, design decisions and the implementation cause limitations. In del.icio.us, for example, a tag can only consist of a single word (see section 8.1.2). This means that "to do", for instance, is not a valid tag in del.icio.us. Users would have to employ alternative spellings such as "todo", "to_do", or "ToDo" to be able to use this term as a tag. This, however, makes it difficult for other users to find such items.

8.1.2 del.icio.us

Numerous tagging systems with varying functionality exist today. Popular examples are del.icio.us, a social bookmark manager, CiteULike, an online service for managing academic literature, and Flickr, a photo sharing service (see [del.icio.us 2006; CiteULike 2006] and section 8.1.3).

In del.icio.us, users can store bookmarks and assign tags to every bookmark (see [del.icio.us 2006]). Users can make their bookmarks private so that only they can access them. Alternatively, bookmarks can be made public so that all people on the Internet can view them.

People exploring a bookmark in del.icio.us are presented a list of all users that tagged this reference as well as all tags assigned by all users. When selecting a user's name people can browse all public bookmarks of this user. On choosing a tag, all bookmarks stored with the same tag in del.icio.us are displayed.

8.1.3 Flickr

As described in section 1.7, Flickr is a community-based web-site for organising and sharing photos (see [Flickr 2006]). Users can upload images to Flickr's server, attach tags, and insert notes into pictures.

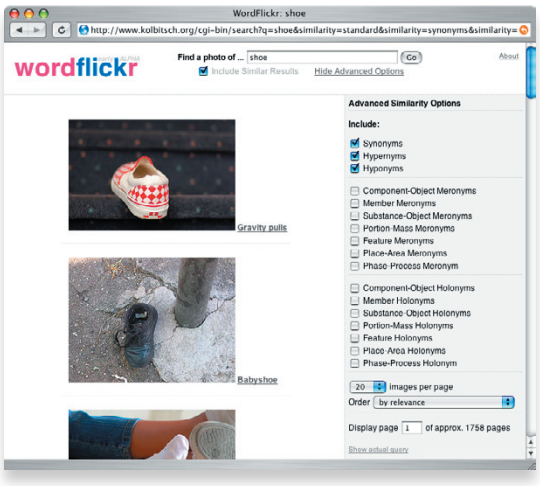


Figure 8.2: Screenshot of WordFlickr. An exemplary query for "shoe" is expanded using synonyms, hypernyms, and hyponyms.

Whenever an image is displayed all associated tags and the photographer's name are provided. Similar to del.icio.us, selecting a user's name shows all public photos of this user. Selecting a tag results in a list of all images that have the same tag attached.

In contrast to other information items, almost all photos have distinct properties that relate to the physical world. Photos are, for instance, taken on a given date and at a physical location. Hence, photos can be timestamped and assigned geographic coordinates. Although these characteristics are obvious, Flickr only recently implemented a feature called *geo tagging*.

With geo tagging, every image can be assigned two special tags whose values are the geographic coordinates of the scene depicted in the photo. These tags are `geo:lat=<coords>` for the latitude and `geo:lon=<coords>` for the longitude with `<coords>` being the actual geographic coordinates.

Although not directly related to the aim of this chapter, the following example illustrates the usefulness of geo tagging. The availability of geographic data can enable innovative applications such as a dynamically and

fully automatically generated travel guidebook. Users simply enter the name of a city or a country they want to go to. The guidebook software retrieves information from Wikipedia or from Wikitravel ([Wikipedia 2006a; Wikitravel 2006]). The Wikipedia article, for instance, contains information on the history, politics, and culture of the city as well as the geographic coordinates. Alternatively, coordinates can be retrieved from the Getty Thesaurus of Geographic Names

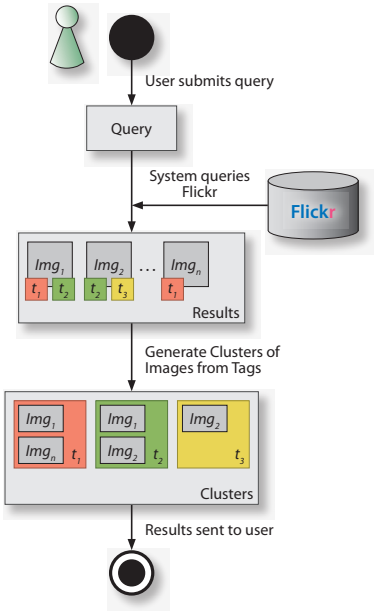


Figure 8.3: Structure of FlickrClustr. The user's query is sent directly to Flickr. The tags of the search results are used to generate clusters of images.

([Getty 2006a]). These coordinates are used for querying Flickr. Hence, a set of photos from the city can be provided. In addition to this, satellite images and maps are retrieved from Google Earth ([Google 2006]). By including a service such as Eventful, the automatic travel guidebook can even offer an "upcoming events" section ([Eventful 2006]).

8.1.4 Motivation

On close inspection, it can be seen that the concept of del.icio.us and its consequences are rather different from Flickr's approach. Although both services are tagging services, del.icio.us allows multiple users to store the same item in the system and lets users attach potentially different tags. Hence, eventually numerous tags with varying frequencies might be assigned to the same item. While a small number of tags might be dominant, a wide range of disparate terms can be available (see [Golder and Huberman 2005]). This means that users would be capable of finding a given item using many different queries. This is a type of "divergence criterion".

In Flickr, however, the same item (image) is usually only stored once, and the tags initially assigned by the photographer are rarely modified or expanded (e.g., [Marlow et al. 2006]). Therefore the divergence criterion explained above cannot be satisfied.

This leads to a variation of the *Vocabulary Problem.* [Furnas et al. 1987] describes that users interacting with computing systems frequently employ very different terms for naming the same objects, tasks, commands, actions, etc. When asked, for example, to assign a term to an object in an information system user A might use an expression not obvious to user B and vice versa. This is the Vocabulary Problem.

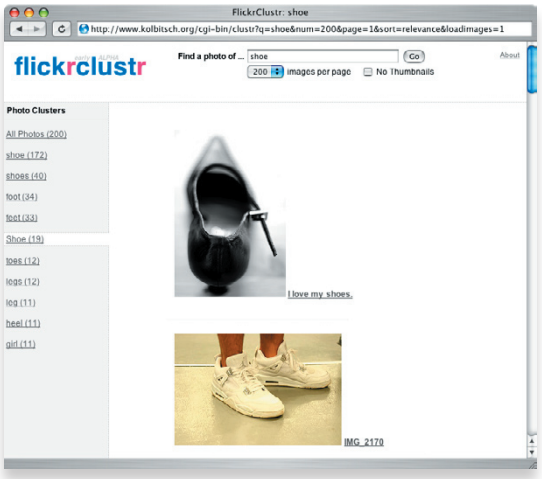


Figure 8.4: Screenshot of FlickrClustr. The results of a query for "shoe" with a maximum of 200 results are displayed.

In Flickr, a variant of this problem can be encountered. When users upload photos to Flickr they perceive the tags they assign to their images as apparent. Other users, however, might not think of these tags when searching for photos and might therefore not be able to find them.

This Vocabulary Problem is widely considered a problem in tagging system (e.g., [Golder and Huberman 2005; Furnas et al. 2006; Mathes 2004]). In this chapter, an attempt is made to counter this deficiency. Section 8.3 presents a potential solution for Flickr and similar systems. Moreover, a prototype implementation of the concept is provided.

8.2 Previous Research

So far, most research in tagging systems focused on the nature and impact of community-based classifiations and the use of loose metadata (e.g., [Tonkin 2006; Hammond et al. 2005]). One of the first comprehensive overview of tagging is provided in [Mathes 2004]. The following sub-sections present two notable research projects related to del.icio.us.

8.2.1 Structural Analysis of Tagging Systems

At the Hewlett-Packard Information Dynamics Lab, the structure of collaborative tagging systems such as del.icio.us was analysed (see [Golder and Huberman 2005]). Several shortcomings including polysemy, synonymy, and the "basic level variation" are described. The latter means that users with varying backgrounds, knowledge, and experience consider terms on different levels of detail as "basic" terms. A dog lover, for instance, might tag a photo of a dog "beagle", whereas someone else might simply use the tag "dog".

The study also investigates the usage of tags in del.icio.us. The results show that there are several types of tags describing, for instance, what the bookmarked item is, what a bookmarked items is about, who owns it, information for organising tasks, or self-referential information (see [Golder and Huberman 2005]). Some of these tags are useful for the general public, while other tags are only useful for their author.

Further findings of the investigation are that the majority of newly added bookmarks in del.icio.us receive most attention within the first ten days. Moreover, tags usage is fairly "stable", i.e., after about one hundred users tagged the same bookmark, hardly any new tags are added by subsequent users.

Another result from the analyses is that although a relation between the number of tags employed by a user and the number of bookmarks stored by the same user cannot be found, the user's tag list is getting longer in the course of time. This might be due to the fact that the user's understanding of a topic eventually gets better, which can result in more refined tags. Probably another cause is a change in interests.

8.2.2 Hierarchies in Tagging Systems

Since tagging systems are flat structures users cannot browse their content like in categorised directories such as Yahoo directory (see section 8.1.1.2). Taking this deficiency seriously, researchers at Stanford University designed an application that automatically generates hierarchies from the tags stored in del.icio.us and CiteULike (see [Heymann and Garcia-Molina 2006]).

The software creates tag vectors that store how often each tag t_u is attached to an information item i_v. Cosine similarity measures are employed for determining the closeness of two vectors. The subsequent applicationg of a threshold filter to the resulting vectors yields a graph of "similar" or "related" tags. Using an algorithm based on graph centrality, the similarity graph is transformed to a hierarchical structure. This hierarchy of tags can be used like taxonomies or like hierarchical directories.

8.3 Enhancing Flickr Queries

8.3.1 WordFlickr

WordFlickr is a dictionary-based approach to solving the Vocabulary Problem in Flickr (see section 8.1.1). Since this approach requires some type of dictionary and the majority of tags in tagging systems are in English (see [Guy and Tonkin 2006]), the concept and implementation focus on the English language.

8.3.1.1 Concept

Users querying Flickr for the term "shoe", for instance, are usually presented with photos of all kinds of shoes. However, if the photographers publishing photos use tags such as "slippers" or "boot" (instead of "shoe") users would not be able to find these photos with their query for "shoe".

WordFlickr offers a solution by allowing users to expand their queries with semantically related terms. In WordFlickr, the users' queries are analysed and expanded using a database called WordNet. WordNet is a sort of a lexical ontology developed at Princeton University. It contains words of the English language together with semantic relations and is both a dictionary and a thesaurus (see [WordNet 2006]). However it does not only contain relations such as synonyms and antonyms but also more complex relations including holonyms, meronyms, hypernyms, and hyponyms (see table 8.1).

When users submit queries to WordFlickr they can choose which types of relations are used for expanding their initial queries. The default setting is to include synonyms, hypernyms, and hyponyms. Users can change these parameters any time and modify additional settings such as the sort order (see figure 8.2).

8.3.1.2 Implementation

The realisation of the concept forms a hybrid web application and can be seen as a "mashup". As depicted in figure 8.1, the user's query is analysed and looked up in WordNet. After word stemming is applied, terms such as synonyms, hypernyms, and hyponyms are retrieved from the WordNet database according to the user's settings. The modified, expanded query is submitted to Flickr, and Flickr's response is sent to the user.

The implementation of the prototype relies on a CGI program implemented in Perl. A WordNet database resides on the same server as the CGI program and is converted from its proprietary format to the Berkeley DB format, which results in significantly faster access. The database is accessed through the

Lingua::Wordnet module, a Perl API to WordNet. The Flickr service is contacted using the Perl Flickr:: API module.

The stable prototype implementation of WordFlickr was tested with Mozilla Firefox 2.0 and Safari 2.0. It is almost as fast and as responsive as the original Flickr service. The WordFlickr prototype is available online at [Kolbitsch 2006b]; its interface is depicted in figure 8.2.

8.3.1.3 Discussion of the Implementation

WordFlickr's solution to the Vocabulary Problem depends on the content stored in tags. Therefore a dictionary such as WordNet required, which has a negative effect on the scalability of the service. On the one hand, WordNet offers a lexical database with powerful semantic relations that can be used well for finding semantically equivalent and similarly related terms. On the other hand, WordNet restricts WordFlickr to the English language. Words from other languages as well as technical terms and personal tags cannot be expanded.

Separate ontologies would be required for supporting further languages. Additional resources to be incorporated include lexical databases such as EuroWordNet available, for instance, in Dutch, Spanish, and Italian (see [Vossen 1999]). Moreover, specialised thesauri like the Getty Thesaurus of Geographic Names and the Art & Architecture Thesaurus can be integrated into an enhanced version of WordFlickr (see [Getty 2006a,b]).

8.3.2 FlickrClustr

First informal experiments with WordFlickr showed that it was hard to "analyse" the results of the queries (see also section 8.4). It was not possible, for instance, to tell if images found through WordFlickr queries were more closely related to the terms searched for than in traditional Flickr searches. Therefore another tool was implemented that can make it easier to analyse search results from both Flickr and WordFlickr—FlickrClustr.

8.3.2.1 Concept

FlickrClustr is an approach to make use of simple clustering in Flickr. FlickrClustr analyses the search results of Flickr queries and the tags assigned to the images in the results. The most frequently used tags are the basis for forming image clusters.

Although Flickr introduced a type of clustering functionality in August 2005, their approach differs considerably from FlickrClustr (cf., [Butterfield 2005]). On the user's request, Flickr's clustering service offers clusters of similar tags for a given tag. These "recommended" tags are most likely generated using the frequencies of tags that are assigned to images in combination with the given tag. Hence, clusters in Flickr serve as a kind of recommendation system for related tags.

The aim of FlickrClustr, on the other hand, is to produce clusters from any search results in Flickr. This enables users to find the most frequently used tags within the results, to group large numbers of search results into a small set of clusters, and also gives them easy access to potentially related tags (like Flickr's clustering).

8.3.2.2 Implementation

The implementation of FlickrClustr is similar to WordFlickr. It is implemented as a CGI program in Perl and makes use of the Flickr::API module for querying Flickr.

The user's query is submitted directly to Flickr as illustrated in figure 8.3. For every image in Flickr's response, all tags attached to the image are retrieved. Subsequently, clusters for the ten most frequently used tags in all images are generated.

Despite the high stability and robustness of FlickrClustr the performance is, unfortunately, rather poor. This is due to limitations of the Flickr API. Tags can only be retrieved for individual images, i.e., when the results for a query contain fifty images, fifty individual queries have to be posted to Flickr in order to collect all tags. Hence, queries with a large number of results are rather inefficient and slow in FlickrClustr.

The prototype implementation of FlickrClustr was tested in Internet Explorer 6, Mozilla Firefox 2.0, and Safari 2.0. It is available online at [Kolbitsch 2006c]. A screenshot of clustered search results for the term "shoe" is provided in figure 8.4.

8.3.2.3 Remark on the Implementation

At the moment only identical tags are clustered, so even tags spelt uppercase and lowercase are regarded as separate tags. Clusters could be made "more tolerant" by transforming all terms lowercase and by employing a word stemming mechanism. Even the latter can be implemented fairly easily by using software such as the Perl module Lingua::Stem.

However, this functionality is deliberately not incorporated in the current research prototype in order to be able to collect information on the usage of tags—upper- vs. lowercase, singular vs. plural forms, etc. Details are provided below.

8.4 Discussion

8.4.1 Concept and Implementation

With WordFlickr, the impact of the Vocabulary Problem can be reduced. Hence, a solution to a single, yet common issue in tagging systems can be offered. Moreover, the basic level variation described in [Golder and Huberman 2005] can be dealt with by including hypernyms and hyponyms in the expansion of the users' queries.

However, WordFlickr cannot consider further shortcomings such as the use misspelt tags, personal tags such as "my dog", and individually encoded tags like "VacationInHongKong" (see [Guy and Tonkin 2006]). Moreover, the concept of WordFlickr is not capable of providing a solution to tag ambiguity, polysemy, and synonym control (see [Mathes 2004]).

8.4.2 Informal Comparison

An informal comparison was performed between Flickr and WordFlickr. The sample consists of a pseudo-random selection of 21 tags (see Appendix E). These 21 terms were selected by the author of this thesis, which can be seen as a methodological concern. However, the intention is not to conduct a formal experiment and to present quantiative results. The aim is to find out in which way results from WordFlickr differ from those of traditional Flickr queries.

For each of the 21 words both Flickr and WordFlickr were queried. The queries were submitted on November 11th and 13th, 2006. "Query by tag" was chosen as search option, the sort order was set to "by relevance", and the number of results was limited to a maximum of 200 images. Synonyms, hypernyms, and hyponyms were included in WordFlickr queries. For *all* queries, clusters of the ten most frequently used tags in the search results were generated with FlickrClustr.

The results for four representative tags are listed in table 8.2. The full list of results is provided in Appendix E.

8.4.2.1 General Findings

As the clustered tags show (see table 8.2), the use WordFlickr has an impact on the diversity of tags. In some cases, WordFlickr's overall range of tags is narrower while the diversity in the requested "category" is wider. Flickr's results for "wine", for instance, include the tags "glass" and "party". Hence, the overall range of terms is rather wide including even word that are not directly related to the original query. WordFlickr's results, on the other hand, include tags such as "chardonnay" and "rose". These terms refine the term "wine", and increase the (vertical) variety of terms directly related to the user's query.

This aspect also has an effect on serendipity. While users can still explore the tags in WordFlickr's search results and find potentially interesting and new resources, their scope might be limited to a narrower context than in regular Flickr queries.

Without providing evidence, [Shirky 2005] claims that making use of thesauri in tagging systems increases "noise". However, this assertion seems to be without substance. As the first informal experiments show, WordFlickr's results are, at worst, about as good as Flickr's (see table 8.2 and Appendix E). In many cases, the tags in the results of WordFlickr actually seem to be closer to the users' initial queries than in Flickr queries.

8.4.2.2 Tag Inconsistencies

As described elsewhere, tag usage in systems like Flickr and del.icio.us is inconsistent (see [Mathes 2004; Guy and Tonkin 2006]). The results from the comparison between Flickr and WordFlickr support these find-

Query	wine		red		rock		shoe	
System	Flickr	WordFlickr	Flickr	WordFlickr	Flickr	WordFlickr	Flickr	WordFlickr
Ten Most Frequent Tags	wine (161)	wine (174)	red (159)	red (183)	rock (125)	rock (186)	shoe (168)	shoe (161)
	Wine (39)	chardonnay (80)	Red (29)	cherry (116)	music (42)	limestone (172)	shoes (50)	pump (52)
	party (20)	rose (49)	green (24)	ruby (110)	Rock (39)	ocean (28)	foot (33)	shoes (38)
	vino (16)	vin (37)	blue (20)	burgundy (85)	ROCK (36)	geology (28)	feet (31)	flipflop (31)
	glass (12)	pinot (28)	2006 (17)	rouge (68)	live (29)	water (27)	Shoe (22)	foot (21)
	2004 (11)	france (27)	macro (15)	scarlet (62)	metal (26)	cave (25)	legs (11)	feet (20)
	2005 (10)	white (27)	orange (13)	maroon (61)	show (23)	coast (23)	leg (10)	heel (17)
	yen (9)	bottle (22)	yellow (13)	cerise (58)	punk (22)	climb (22)	toes (10)	flip-flop (14)
	brokenwood (9)	cabernet (22)	black (13)	color (57)	musica (22)	landscape (22)	heel (9)	red (14)
	roadtrip (9)	Wine (22)	white (13)	pink (54)	york (20)	sea (22)	black (9)	sneakers (13)

Table 8.2: Selected results from a comparison between Flickr and WordFlickr queries. For every term 200 results "ordered by relevance" were retrieved. For the ten most frequently used tags, the numbers in brackets describe how often each tags occurred in the results. All queries submitted on November 11th, 2006.

ings. The tags in the query results for "shoe" in table 8.2 illustrate several discrepancies that may occur. Both singular and plural forms of nouns are present: shoe and shoes, foot and feet, leg and legs. Moreover, both upper and lower case are used: shoe and Shoe.

As mentioned above, FlickrClustr could be easily modified so that inconsistent terms are merged into single clusters. Although this would "correct" the clusters generated by the system, the actual problem would remain unsolved because a query for "shoe", for instance, might still not include photos tagged "shoes".

8.4.2.3 Tag Usage and Interpretation

This sub-section qualitatively analyses the tags present in the results of the Flickr and WordFlickr queries listed in table 8.2 and offers an interpretation. The first finding is that, in some cases, WordFlickr yields less tags that are only relevant to owners of photos than Flickr. When Flickr is queried for "wine", for instance, personal tags such as "2005", "party", and "roadtrip" are frequently used among the resulting images. These tags are often not relevant for people searching images but only for the owners of photos.

The WordFlickr query for "wine", on the other hand, is able to eliminate such tags. WordFlickr lists tags such as "cardonnay", "rose", and "pinot" which are vine varieties (i.e., hyponyms). Hence, WordFlickr is capable of offering tags in the results that are semantically closer to the user's query. (However, it cannot be stated if the images provided by WordFlickr are more appropriate for the user.)

Another example where WordFlickr can provide results that are semantically more closely related to the user's query is the term "red". Although Flickr's search results frequently contain other colours such as "green" and "blue" WordFlickr's results more often include shades of red such as "cherry", "burgundy", "rouge", and "scarlet" (synonyms and hyponyms). Moreover, WordFlickr filters out the personal tag "2006". Therefore WordFlickr's queries are, again, probably more appropriate results for the user's query.

The query for "rock" shows an interesting effect. While Flickr includes tags such as "music", "live", "metal", and "show", WordFlickr provides tags like "limestone", "geology", and "climb". This means that Flickr's query relates the term to rock music, whereas WordFlickr interprets the term "rock" as stone. This is particularly astonishing because a manual WordNet query shows that the lexical database retains both meanings of the word "rock".

However, in this example WordFlickr also includes tags such as "ocean", "water", and "sea". Since they are not directly connected with the user's query some of WordFlickr's most frequently used tags are semantically even more distant that Flickr's tags.

The last of the four terms analysed is the example used throughout this chapter—"shoe". In this case, the results of both Flickr and WordFlickr resemble each other. The influence of the WordNet database can be seen in the tags produced by WordFlickr: they include tags such as "pump", "flip-flop", and "sneakers" (i.e., hyponyms). However, both queries contain significant numbers of tags that are not directly associated with the user's query: "foot", "leg", and "heel". Hence, in this situation WordFlickr cannot provide more appropriate results than Flickr.

8.5 Conclusion

Tagging systems are becoming an increasingly popular way of organising content and generating metadata. In a sense, tagging services are a competition for conventional classification systems and metadata initiatives. However, services based on collaborative tagging have a number of disadvantages including the Vocabulary Problem.

In this chapter, WordFlickr was introduced as a solution to the Vocabulary Problem. The concept of WordFlickr is based on WordNet, a lexical database with semantic relations between words from the English language.

A first informal experiment compares search results from the prototype implementation of WordFlickr with results from Flickr. It is hardly possible to quantise the differences between Flickr and WordFlickr. The qualitative results, however, imply that WordFlickr excels Flickr in some cases, where the most frequent tags in WordFlickr's search results are semantically closer to the users' initial query than in Flickr's results. On other occassions, though, the results of Flickr and WordFlickr are alike. Hence, it can be concluded that WordFlickr yields results that are, at worst, as good as Flickr's search results. Therefore the use of WordFlickr's concept can be a valuable addition to tagging systems such as Flickr.

In addition to this, the experiment conducted supports other authors' findings that tags are often inconsistent and contain both in singular and plural forms, uppercase and lowercase, etc. Moreover, there is no substantial evidence for the assertion made in [Shirky 2005] that the use of a thesaurus increases noise in tagging systems.

Part 3 Collaborative Systems and Knowledge Management

Integration of Collaborative Concepts

9 Kōrero: An Integrated Community-Based Platform for Collaboration

A version of this chapter was submitted to the International Conference on Knowledge Management (ICKM 2007).

Publication submitted in January 2007.

Various types of community-based systems including blogs, wikis, file sharing tools, social bookmarking tools and social networks have emerged recently. These services are usually only loosely connected and rarely make use of synergetic effects. In this chapter, a concept for a platform that adds value by *integrating* the benefits of existing community-based systems is proposed.

The proposed concept is user-centred and puts an emphasis on communication and collaborative content development. The approach presented in this chapter facilitates the discovery of implicit knowledge by making proactive, personalised suggestions about content and the availability of experts. Application areas include learner-support systems, corporate and organisational environments, and communities sharing common interests.

9.1 Introduction

Recent developments on the World Wide Web have yielded systems and services with unprecedented qualities. Nowadays, large numbers of people use blogs, wikis, and social networks every day (cf., chapter 1). These environments foster collaborative work, and allow users to express themselves, participate in the development of content, and enable new forms of communication among individuals.

The following sub-section explains the methods available to date for combining these individual systems. In section 9.2, an alternative approach is presented that integrates the separate technologies and makes use of synergetic effects. The architecture of such a platform is described in section 9.3, and section 9.4 addresses organisational aspects. Several application areas for the proposed concept are presented in section 9.5.

9.1.1 Blended Systems

With millions of users participating and millions of information objects produced, the technologies introduced above are quite popular and successful. As mentioned in chapter 1, these developments—blogs, wikis, social networks—are independent and autonomous systems.

Several approaches to combining the various community-based technologies in a single system exist. Examples are environments employing both wikis for the actual development of content and blogs for discussion and commentary (e.g., [Su 2005]). However, this results in a system where blogs merely complement wikis.

A system implementing a similar approach is Elgg, a feature-rich learning environment (see [Elgg 2006; Anderson 2005]). It includes a wiki, blogs and podcasting. Moreover, it makes use of external systems for sharing images, for social bookmarking, directories and news, for example. Elgg, however, does not integrate these "modules" in a unified system but is a framework that allows administrators to combine individual systems under a single user interface.

9.2 Concept

When the technologies described in section 9.1 are not only combined but *integrated* in a consistent environment, synergetic effects can be obtained. These include the guided discovery of knowledge, system gen-

erated knowledge, forming relationships with friends, and experts who would otherwise be overlooked.

The information stored in a wiki, for example, consists basically of the actual content and data on its authors. This elementary set of information on its own can be of relatively limited value. In conjunction with data from other content generating components and a social network that contains relations among users the significance of this information can be increased. In this case, techniques from knowledge management can be applied to generate *new* information (e.g., [Maurer and Tochtermann 2002]).

Methods employed in well-accepted knowledge management systems make use of explicit relations (such as the names of blog authors and the names of their friends) to find implicit linkages among information entities and make them accessible (cf., [Novak and Wurst 2004]). Hence, tacit knowledge that might otherwise be lost is communicated. This approach can, of course, not only involve textual content but also other media such as images and more abstract information objects such as users in a social network.

9.2.1 System Generated Information

When connections among information entities are generated, the users' relationships in the social network, the data they produce and the information they consume are analysed. Moreover existing links between pieces of content are examined: the links from blog entries to wiki articles, the links between blogs, etc. By creating a graph containing all explicit connections, implicit connections can be found. The information discovered in this process is retained and can be used for further inferences.

Figure 9.1 illustrates this approach and highlights a number of features of a system implementing the proposal. It contains four users named *A* through *D,* two blogs b_m and b_n, and two wiki articles w_x and w_y. In the social network, direct connections between users *A* and *B, B* and *D* as well as *C* and *D* exist (black lines in figure 9.1). Since *D* is directly connected to both users *B* and *C,* there is an indirect relation between *B* and *C.* This means that *B* and *C* are second degree friends in the network (wavy line in figure 9.1).

Users *C* and *D* maintain blogs b_m and b_n. Although the two users are explicitly connected through the social network, there is no link between the two blogs. The system can, however, deduce a connection between the two blogs, based on the relation between their owners. Consequently, links between b_m and b_n can be inserted automatically into the blogs, and their owners can be notified of the implicit connection between the blogs.

Between users *A* and *C,* on the other hand, there is no explicit connection, although both users are editing the same wiki article. Particularly, when several such congruences occur the system can identify these implicit connections and point them out. These relations might indicate that both individuals share a common interest or work in the same area. Hence, the users' contact list (part of the social network) might, for instance, include a dynamically generated category of "professional colleagues".

In addition to this, the system can find somewhat weaker links among entities. User *C,* for instance, works on both blog b_m and wiki article w_y. Hence, there might be a possible connection between b_m and w_y such as a similar topic. (Note that the connection between b_m and w_y would be a strong link if one of the blog entries of b_m cited w_y.)

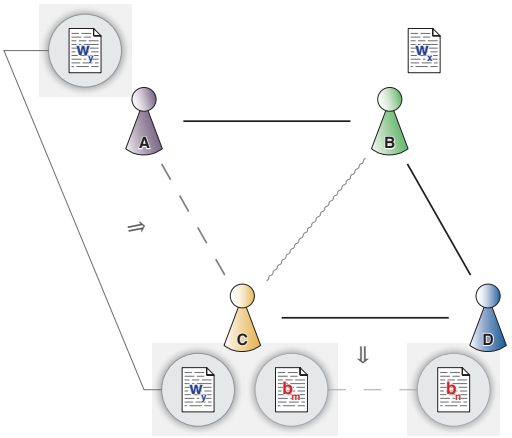


Figure 9.1: Social network with four users *A* through *D,* two blogs b_m and b_n, and two wiki articles w_x and w_y.

Further weak links could be established between b_m and users *A* and *B.* The link between b_m and *A* is based on the implicit connection between *A* and *C* due to w_y. The relation between b_m and *B* can be explained by the fact that *B* is a second degree friend of *C,* the owner of blog b_m.

For the example outlined in figure 9.1 only the most important connections that can be discovered are presented. Additional linkages, and thus knowledge, can be found even in such a simple setting as the one above. It should be mentioned, though, that linkages based on implicit and indirect connections may be insignificant. Therefore in these cases, links should only be created if several "indicators" for a possible connection are present.

9.2.2 Guided Information Discovery

In addition to the automatic generation of connections between information entities, users can be notified directly of relevant information. These notifications should be proactive, yet unobtrusive. A user reading several wiki articles on a particular topic, for instance,

might be informed (verbally, visually or iconically): *"You probably haven't met user C yet. C has contributed to several articles in this field and might be a domain expert. You might also be interested in C's blog."*

With guided information discovery, the system operates the same way as described above, i.e., the system makes, for example, potential links between wiki articles, users and their blogs explicit (see section 9.2.1). The newly generated information, however, is not merely stored in the underlying structure or inserted into documents, but proactively disseminated to users. The system points out that there may be a connection between entities in the environment and lets users observe connections and decide whether the links are relevant. The users' feedback—following a link or not—can be utilised for improving the quality of suggestions made by the system.

By including information from the social network "personalised" data can be provided, i.e., information which users can relate to is given priority. Information not directly relevant to users, on the other hand, is de-prioritised.

9.2.3 Visualisation

Both explicit and implicit knowledge can be visualised in order to facilitate knowledge retrieval and understanding the complexity of information in a collaborative environment (e.g., [Burkhard 2004; Viégas and Donath 2004]). In this proposal, visualisation methods can be used to show both the most relevant connections between content and users and their significance. Extensive research has been done in this area, and therefore only two possible visualisation techniques are outlined—knowledge nets and document maps.

Knowledge nets are a technique for visualising relations among pieces of content. One node is displayed in the centre of the graph, and its connections to associated information objects are depicted in a "star topology" (e.g., [Maurer and Tochtermann 2002]). The further a node is away from the centre of the graph the less it is related to the central node (see figure 9.2).

This visualisation technology is, for instance, successfully applied in advanced electronic encyclopaedias such as the Brockhaus Digital Encyclopaedia (see section 2.1.2). In this proposal, knowledge nets can be utilised for visualising connections from the information item currently displayed by a user to related objects and their "distance". Users would, for example, easily recognise which blog entries or users are associated to the wiki article they are currently reading, and which discussions are attached to it.

In contrast to this, document maps do not focus on a single node but on groups of information objects, typically documents. Such maps are a way to point out clusters of objects that share similar characteristics (e.g., [Novak and Wurst 2004]). Features used for similarity recognition can, for instance, include information on the categorisation of objects, topics of articles, authorship, and link structures. Document maps usually contain a number of clusters that are depicted in a two-dimensional space. The distance between clusters indicates how "different" clusters are.

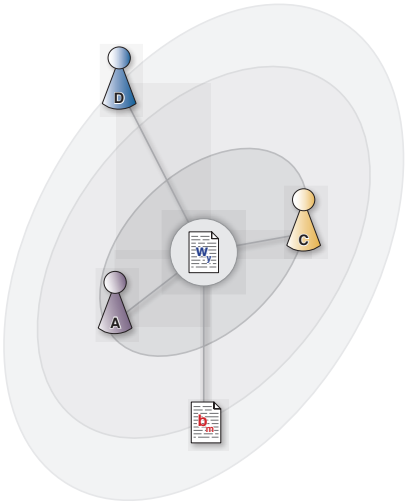


Figure 9.2: Examplary knowledge net showing the information objects connected to wiki article w_y: users *A, B* and *C,* and blog b_m.

When creating an integrated collaborative platform, document maps can be used to present clusters of related information items. This makes it relatively easy for readers to find out which other users participate in a given topic, which wiki articles are available for the topic, whether related postings in blogs or discussion forums have occurred, etc.

9.3 The Proposed Platform

This is the proposal of Kōrero—a web-based platform integrating all the most important current community-based and collaborative technologies. *Kōrero* is Māori for "talk, chat, communicate". The term adequatly describes the nature of the platform: users should be able to employ a single, uncomplicated user interface to meet friends and experts in the system, to communicate with them, to make personal and professional commentary, and to author and edit existing content. These diverse technologies are integrated so seamlessly that users do not "see" or "feel" that they are making use of different functional parts of a system.

The proposed platform includes several functional components and supporting technical foundations. Unlike blended systems, where a framework is provided for enabling the parallel use of existing products, in this apporach a system that incorporates a set of

functions in a unifying environment is designed (cf., section 9.1.1).

The platform has a layered structure that is outlined in the following sub-sections (see figure 9.3). The organisation of information and users in the system is discussed in section 9.4. A detailed account of the underlying technical infrastructure is presented in the next chapter.

9.3.1 Core Technologies

The technology layer, the lowest layer in figure 9.1, provides the underpinnings and essential infrastructure for the platform. All other components of the platform share the functionality provided and build on these foundations.

This layer contains repositories for the data stored in the system including content and information on users. It contains the basic functions for user management (e.g., authentication) and features such as full-text and metadata searching. Also part of this layer are the inference machine and visualisation techniques which make the discovery of new knowledge possible (sections 9.2.1 and 9.2.2).

The fundamental component of the system is the all-embracing social network. Since it is an essential part of the infrastructure and also generates new information at the same time, the social network is included in both the technology and the content layers.

9.3.2 Content Generating Components

The content layer is the second layer in figure 9.2. It builds on the core technologies and contains several functional components that can be employed by users for generating new content.

These components include wikis, podcasting and vodcasting, and potentially other technologies for generating actual content. Blogging and forum components are used for commentary and discussion, and for generating supplementary content (see also section 9.4.1). In addition to this, the social bookmarking service and the social network are modules that produce information in the system.

9.3.3 Communication Facilities

The communication layer is the third layer in figure 9.3. It focuses on the direct contact and information exchange between users in the system and allows users to discuss content.

Both synchronous and asynchronous communication are supported. When multiple users are logged onto the system instant messaging, chat, and technologies similar to the ones described in section 2.3.7. The system can also encourage users to communicate with each other by proactively suggesting: *"One of the authors of this article is currently online. Would you like to talk to her?"* In an advanced variant of this approach, the system can suggest an "expert discussion" on a particular topic involving, for example, the five authors with most contributions to a certain topic. (Only relatively "long-lived" contributions are considered in order to exclude users involved in flame wars.) The discussion is associated with the corresponding information objects in the system and made available to all users.

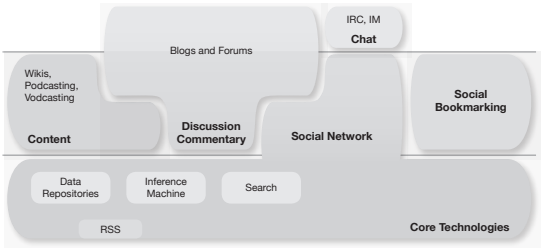


Figure 9.3: Overall structure of the proposed concept. It includes wikis as the main component for generating content, blogs and forums for discussion, a social network including communication facilities as its core component, and social bookmarking.

Besides direct user-to-user communication, the concept also incorporates system-to-user information delivery. This includes mainly syndication technologies such as RSS (cf., sections 1.2.2 and 2.3.7) but also simple methods like e-mail notifications. The system can inform users of recent changes to content they have contributed to and notify them of new content. The system can send brief summaries to users in order to keep them updated on content they are interested in, for example, when new articles on a favourite blog are posted or a new podcast episode is published in the system.

9.4 Aspects of Organisation

A platform encompassing various community-based technologies should take the complexities of content and user management into account. Aspects of these two fields are described in the following sub-sections.

9.4.1 Organisation of Content

As mentioned above (section 9.3.2), content in this proposal involves data as diverse as wiki articles, podcast and vodcast episodes as actual content, blogs and forums for commentary and discussion, or links in the social bookmarking service. The approach chosen for resolving this complexity is to abstract the different kinds of content in the system to a degree that they can be treated as one generalised type of content (see next chapter). These generalised content objects consist of a title, a "body" (the actual content), a version,

at least one author, and potentially further metadata such as keywords. The body can contain text-based as well as binary data, making the approach open enough to support types of content other than the ones described above.

Two aspects of content storage are particularly noteworthy: versioning and link structures. All content in the system is subject to version tracking. Although this is an inherent feature of wikis, other technologies such as blogs, podcasting, links in a social bookmarking service usually do not make use of versioning. However, in this approach for every information item several versions can be retained. When a user decides to update a podcasting episode, for example, both the old and the new versions are available (cf., [Nelson 1981; Haake and Hicks 1996]).

Link structures in the proposal include both hyperlinks explicitly created by authors and connections discovered by the system. Links are not necessarily visible for users. While hyperlinks inserted by authors are always shown, connections encountered by the system are only displayed when they are relevant for users. (It should be noted that the relevance of linkages found by the system may change over time for one and the same user.)

Every information item in the system can be the target of a link—not only actual content but also users, for instance. Moreover, every object in the system can be the source of link. For those kinds of information that usually cannot support hyperlinks, the system *attaches* links to the content. A podcast episode, for instance, can have a hyperlink attached that points to a wiki article offering complimentary information.

All links in the system are bidirectional. This permits users to determine which information objects in the system link to the document they are currently viewing.

9.4.2 Organisation of Users

Previous research shows that anonymous authorship is a very important aspect in successful large-scale collaborative systems (see chapter 6). Social networks, however, require registered users and are intrinsically incompatible with unregistered users.

Therefore a "dual" approach is employed. Every user wishing to participate as an author has to register and log in to the system (cf., [Flinn and Maurer 1995]). Once logged on, the user can edit existing information or produce new content. However, for *every* action in the system users can decide whether they want to perform it anonymously. I.e., a user can choose to post a blog entry, for instance, under their name while editing a wiki article two minutes later anonymously.

For such "quasi-anonymously" authored information, neither a username nor an IP address (like in Wikipedia) are displayed to other users. Moreover, the system does not consider actions of anonymous authors when attempting to discover new information. So when user *C,* for example, edits wiki article w_y anonymously, the system will not point out connections between w_y and *C* or w_y and *C*'s blog.

With this approach, users can work anonymously while the system still "knows" the authors of information items (their usernames and potentially also their IP addresses). This makes the system accountable and can help prevent its misuse.

Users not wishing to contribute content to the system, i.e., mere readers, do not have to register or log in. In this case, however, only a limited set of connections might be pointed out because of the missing "background" information on the user. Moreover, the quality of connections discovered by the system may be suboptimal because it is difficult to generate personalised information for an unknown user.

9.5 Application Areas

The concept proposed in this paper can prove to be a valuable tool in diverse fields. The following subsections illustrate a set of potential application areas: learner-support systems, corporate and organisational environments, and special interest groups.

9.5.1 Learner-Support Systems

With functionality similar to blended systems (section 9.1.1), this proposal also supports similar application areas. Examples are learning environments, where the presented concept provides a unified platform for lecturers and students to author content collaboratively, a space for commentary and discussion, and a place for communication as well as social and professional contacts (cf., [Elgg 2006; Su 2005]).

Also in contrast to existing systems, this concept actively encourages users to discover new information and makes users aware of connections between pieces of content that might otherwise have gone unnoticed. Relatively simple visualisation techniques can help learners to identify relations between different parts of the content and enable them to find references to associated information in the system. Furthermore, meeting new (relevant) people is actively supported through suggestions made by the system (see section 9.2.2).

9.5.2 Corporate and Organisational Environments

As mentioned above, companies make use of wikis for documentation and employ blogs, for example, for communicating news, progress on projects and similar

information to employees. Rarely, however, these technologies are combined into a single application. One of the few attempts in this field is a research project carried out by the CIA. It describes a concept that utilises wikis for creating a general body of knowledge by members of the organisation, and offers blogging functionality for comments, potentially opinionated views and complementary information (see [Andrus 2005] and section 1.3.4). Though combined, the system does not make use of synergetic effects and cannot add value to its information in the individual components.

With the illustrated approach, users have similar capabilities for creating content and for expressing personal views. However, this system can depict explicit relations between information items and reveal implicit linkages between the objects stored in the system. Moreover, experts whose expertise in certain topics might previously have been unknown can be found. System-to-user communication keeps users informed, for example, when new users contribute to a topic or when wiki articles are updated.

9.5.3 Special Interest Groups

The proposed platform can be a useful tool for building communities of people with shared interests. Imagine a group of snorkellers in a particular geographic region. The members of the group can use the wiki component of the system to work collaboratively on an encyclopaedia of the region's underwater world. Eventually the encyclopaedia will become increasingly complete and ultimately represent a comprehensive work encompassing the area's flora and fauna.

Moreover, users can maintain photoblogs or vodcasts to present their most spectacular sightings. In podcasts, "senior snorkellers" can give advice to beginners or can point out the highlights of the region to visitors. Users can, for instance, make use of the built-in chat functionality of discussion forums to communicate with each other and discuss the occurrence of an uncommon variety of fish. In addition to this, the system can let new users know who the experts in the various topics are and can recommend photoblogs or postings in discussion forums that are associated with wiki articles for beginners.

9.6 Conclusion

Recent community-based technologies such as wikis, blogs or social networks attract millions of users that contribute their knowledge and experience to producing and maintaining content. The spectrum of these developments has evolved into a patchwork of autonomous, unrelated systems. Although a small number of approaches to combining these technologies into a framework exist, these systems have failed to make use of the synergetic effects.

In this paper, a system that bridges this gap was proposed. The proposal tightly integrates the currently most popular community-based technologies into a single, unified, and easy-to-use platform—Kōrero. The fundamental part of the Kōrero is a social network that connects functional components such as wikis, blogs or a social bookmarking service. With the use of features from knowledge management in this integrated environment, discovery of implicit linkages and tacit knowledge can be attained. Moreover, the system can make personalised recommendations of newly discovered information to users.

The proposed platform is based on a unifying data structure that can be used to store content produced by individual technologies such as wikis, blogs, and social networks. Explorations in designing such a structure are presented in the next chapter.

10

A Unified Structure for Current Collaborative Systems

This chapter introduces an infrastructure that unifies the data model of existing collaborative environments such as blogs, wikis, file sharing services, and social networks. The proposed model consists of a small number of generalised entities that basically represent content, users, links between pieces of content, and links between users of the system. By assigning appropriate permissions to content, linking pieces of information, and associating users with content, most of today's collaborative systems can be implemented.

The relative simplicity of the model makes it flexible and extensible. Therefore, future developments in collaborative systems are likely to be realisable with the proposed model. Moreover, the unified structure presented in this chapter is the basis for Kōrero, the integrated, community-based platform for collaboration.

10.1 Introduction

Recent developments in collaborative systems including weblogs and wikis are based on largely autonomous infrastructures. The platform introduced in the previous chapter attempts to combine and integrate these services on the application level, i.e., users employ a single interface for accessing the various functionalities. In this chapter, the underlying core data model for such an integrated platform is presented. It aims at unifying the data models of various collaborative services available to date.

Section 10.2 gives a brief overview of the characteristics of weblogs, wikis, podcasts, social bookmarking and file sharing services, and social networks. The proposed unified structure for recent collaborative systems is described in section 10.3, and a selection of exemplary applications of the infrastructure is provided in section 10.4.

10.2 Characteristics of Collaborative Systems

The following sub-sections detail the most important aspects of current collaborative systems that are relevant for their underlying data structures. Although the systems discussed offer further features that might be relevant for users, only the core functionality is considered in this chapter.

10.2.1 Blogs

Blogs consist of reverse chronologically ordered lists of entries, where every entry is basically an HTML page that can contain text, unidirectional links to images and similar media, and conventional hyperlinks (see section 1.2). A blog entry has an owner, a timestamp, and a unique identifier that can serve as permalink. Moreover, comments may be attached to blog entries. Comments are also listed in reverse chronological order and can usually be authored by anyone on the Internet.

In general, blog entries are written once and remain unedited thereafter. Entires are publicly readable, while only blog owners may create them. However, articles on blogs may have permissions attached so that only a certain user or group of users have access or that not only the owners of blogs may edit their articles.

Users can subscribe to blogs using RSS feeds. Feeds contain a unique identifier (URL) of the complete blog entry together with information on the author, a title, the publication date, and additional metadata.

10.2.2 Wikis

Wikis are collections of independent pages that are connected by hyperlinks. Hence, they form a directed graph (see section 1.3). Every page can be seen as HTML document and may consist of text, unidirectional links to external resources and to other pages in the wiki.

Wiki articles may be edited by several users. All versions of all articles are retained, where every version of an article has exactly one author. (A subsequent edit results in a new version.) Every wiki article has a unique identifier, a version, a timestamp, and an author. When an article is requested by a user the most recent version is displayed by default.

10.2.3 Podcasts

As noted earlier, podcasts are essentially blogs with a special kind of content (see section 1.6). Therefore podcasts share some of the properties of blogs: they are listed in reverse chronological order, every episode has a title, an author, a timestamp, a unique identifier (usually a URL), optionally additional metadata, and may have comments attached.

Podcast episodes, however, do not contain textual content but MP3 audio data, i.e., binary data. They can usually not contain other kinds of media or links to external resources. Moreover, episodes are normally not edited after publication.

Podcasts are published using RSS feeds. For every episode, feeds contain the identifier (URL) of the binary audio content, the title of the episode, the timestamp, and optional metadata such as a short description of the content.

10.2.4 Social Bookmarking and File Sharing

Both social bookmarking and file sharing tools such as Flickr or YouTube are tagging systems (see section 1.8 and chapter 8). As detailed in chapter 8, a piece of content in tagging systems is stored together with its title, a set of tags, the name of the person adding the content to the system, a timestamp, optional metadata, and permissions for accessing this information.

A piece of content can be the URL of a bookmark (del.icio.us), an actual web-site (Furl), a photo (Flickr), a video clip (YouTube), etc. Tags are textual metadata attached to the content. The type of optional metadata depends on the various implementations of tagging systems. An example is Flickr where users may leave notes in pictures.

10.2.5 Social Networks

Social networks such as MySpace, Friendster, Orkut, and OpenBC basically contain users, user profiles, and potentially weighted links between users. For every user, the user's name, an e-mail address, and a photo are retained. Moreover, a list of hobbies, likes and dislikes, etc. is stored in the user profile.

Users maintain a list of their friends and have access to friends of their friends. In some social networking services, these connections to friends can be weighted, i.e., users are capable of defining how closely related they are to someone else.

Users in social networks can send messages to their contacts and to other users in the network. Moreover, several services provide their users with guestbooks and facilities for authoring blogs.

10.3 Proposed Structure

Although the technologies outlined above have disparate concepts and serve distinct purposes, their underyling structures are, in many cases, rather similar. Both wikis and blogs, for instance, deal with articles. While wikis allow for versioning and multiple authors, blogs usually have only a single author and different permissions for editing content (cf., [Tonkin 2005]).

The concept of the community-based platform for collaboration introduced in the previous chapter relies on a unified structure for all types of content in the system. Hence, a novel data model for a relational database system is suggested.

The proposed model includes generalised entities that can be used for the services described in section 10.2. Entities basically include content, persons, and links. Content is of a given content type, has permissions associated, and can have additional metadata attached. Persons have profiles and may belong to groups. Links can be established between pieces of content, between persons, and also exist between persons and content.

The following sub-sections describe the entities and relations in the system in greater detail. Figure 10.1 illustrates a basic entity relationship model for the core entities of the unified structure.

10.3.1 Entities

10.3.1.1 Content

`Content` entities can store all kinds of content including blog entries, wiki articles, podcast episodes, photos in file sharing services, etc. The `content` entity type contains the following attributes:

- unique identifier: a unique identifier for every element of the data set. It is used in permalinks of blogs, for pointing to single podcast episodes in RSS feeds, etc.;
- title: the title of the piece of content. This can be the title of the blog entry, the name of a wiki

article, the title of a podcast episode, the name of an image, etc.;
- content: the actual content. For blog entries and wiki articles, this is the text of the article (including markups for images, etc.). For podcasts and images, it is the binary data, etc.;
- version: the version of a particular instance of the content. In wikis, for instance, this is the version of an article. Moreover, the version attribute enables versioning in blogs, podcasts, and other applications. If versioning is undesired, the version attribute can be left undefined or can be set to a default value;
- timestamp: the timestamp when this instance was created;
- enabled: whether this instance of content is enabled (accessible). The notion is to keep all versions of all content. Instead of deleting obsolete or unwanted instances, the status of a piece of content is set to "disabled".

10.3.1.2 Content Type

The `content type` entity is used for differentiating between the various types of content in the system. Such an entity becomes necessary because all content is stored in the same type of entity—regardless of its purpose on the application level. This entity type has only two attributes:
- type identifier: a unique identifier for the content type;
- type name: the name of this content type, e.g., "blog entry", "wiki article", "podcast episode".

In addition to identifying the type of content, it can also be used for defining the types of link sources and destinations, metadata, and permissions. (This might require the introduction of further relations.) Thus, the `type name` attribute might, for instance, contain values such as "image tag", or "external link".

10.3.1.3 Metadata

The `metadata` entity type is used for attaching additional descriptive data to content. Although the primary aim is to attach metadata to content, it can also be attached to other entities such as links by using appropriate relations. The attributes of the `metadata` entity type are:
- metadata id: a unique identifier for the metadata element;
- metadata type: foreign key relation to the unique identifier of `content type`. This attribute might, for instance, reference the content type "image tag";
- metadata value: the actual metadata value, e.g., a tag attached to an image in a file sharing service.

10.3.1.4 Person and Group

The `person` and `group` entity types represent users of the systems and their memberships in user groups. A person is a registered user in the system and may, according to the permissions granted, access and edit existing content, author new content, etc. The person entity type includes:
- user id: a unique number identifying a person;
- user name: a person's unique short name;
- first name: a person's firstname;
- last name: a person's lastname;
- photo: foreign key relation to a unique identifier of the `content` entity type. A photo of a user stored as content in the system;
- e-mail: the e-mail address of the person;
- etc.

A group is a set of one or more persons. It might be used for granting rights to an entire group (instead of individuals), for formings special interest groups in the social network, etc. The `group` entity type consists of the following attributes:
- group id: the unique identifier of a group;
- group name: the name of a group.

10.3.1.5 Profile

The user's profile commonly present in social networking services and other community-based environments is realised with the `profile` entity type. Among other attributes, it includes:
- profile id: a unique value identifying the profile;
- likes: textual representations of a person's favourites;
- dislikes: a person's distastes;
- hobbies: a person's hobbies;
- etc.

10.3.1.6 Permission

The `permission` entity type describes a set of privileges for accessing and modifying content. It does, however, not define which content, for instance, the permissions apply for or which person or group these permissions are granted. This entity type contains the following attributes:
- permission id: unique identifier of the permission;
- permission name: a name for this permission, e.g., "may edit blog article";
- rights granted: the actual privileges granted constitute a set of zero or more values of "read", "edit", "create", and "delete".

10.3.2 Relations

The following sub-sections only describe the most significant relations of the entity relationship model

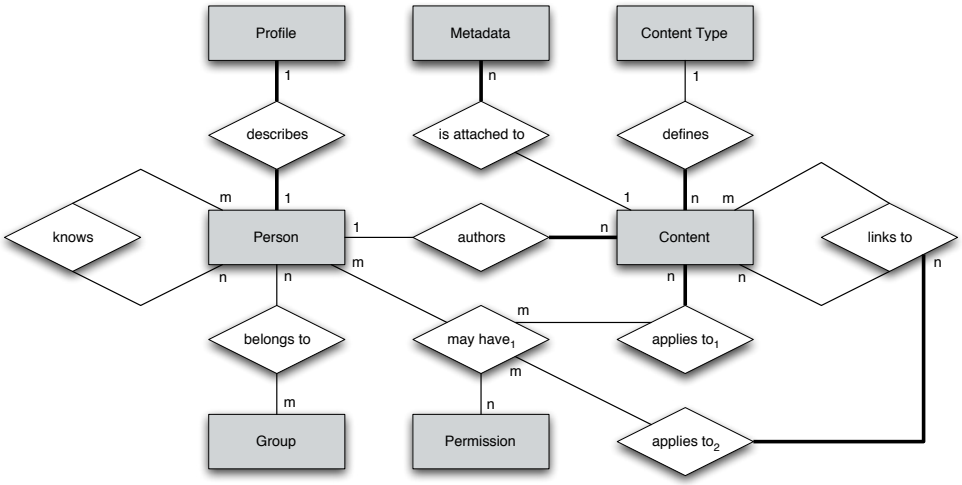


Figure 10.1: The entity relationship model for the most significant entities in the system. For ease of readability, attributes are not shown. Bold lines indicate total relationships. "applies to" relations associated with the "group" entity are not shown.

in figure 10.1. Relations such as `profile describes person` are considered self-explanatory. (This relation could, in fact, be replaced by an attribute in the `person` entity.)

10.3.2.1 Authorship

The authorship of articles is determined in the `authors` relation that connects an author to a version of a piece of content. This binary relation consists of the following attributes:

- authorship id: unique identifier of the relation;
- author: a reference to the unique identifier of a person;
- content: the unique identifier of a piece of content.

The participation of `content` in the `authors` relation is total (bold line in figure 10.1). This means that every instance of the `content` entity has to be related to an instance of the `person` entity. In other words, every version of a piece of content needs to have an author. The participation of `person` in the `authors` relation, on the other hand, is partial, i.e., not every person needs to author content.

10.3.2.2 Content Link

Content links defined in the `links to` relation are used for connecting two pieces of content with a directed link. Links can be deep links referring to content on its most granular level. Moreover, links can be implemented to be bidirectional on the application level. The `links to` relation contains the following attributes:

- link id: a unique identifier of the link;
- link title: a title of the link;
- link source: foreign key relation to the identifier of a piece of content that is used as the source of the anchor. In classic HTML, this is the document containing an anchor;
- link source location: the optional posistion a link originates from. This attribute enables deep link such as "character 123" in a text document, "x=123px, y=234px" in an image, or "1m23s" in a sound file;
- link source length: the optional "length" of a link. this might be "10 characters" for text-based links, "width=10px, height=20px" for an image, or "12s" for a sound file;
- link source type: a foreign key relation to the identifier of the `content type` entity. This attribute is used for describing the type of the link source. A link can, for instance, be embedded into the document when transmitted to the user (like in HTML documents), the link can be an "external link" provided as supplementary information to a sound file that can normally not contain links, etc.;
- link destination: foreign key relation to the identifier of a piece of content that the link points to;
- link destination location: the optional position the link points to;
- link destination length: the optional length of the link destination;
- link destination type: the type of the link destination (cf., link source type);
- author: foreign key relation to the identifier of a person defining the author of the link;
- version: the version of this link (cf., version attribute in the `content` entity type);
- timestamp: the creation date and time of the link.

Both participations of `content` entities in this relation are partial.

10.3.2.3 Person Link

Links between two persons are defined in the knows relation. Such links are mainly a necessity in social networking services, where the connections between users are stored. Person links are directed and can, on the application level, be implemented as bidirectional links. The knows relation includes:

- link id: the unique identifier of a link;
- link name: the name of a link;
- link source: foreign key relation to the unique identifier of a person. This attribute describes the person the link originates from;
- link destination: the person the link points to (cf., link source);
- link weight: the weight of a connection between two persons as an integer value. A value of 0 might, for instance, denote a neutral relationship, whereas +2 could be a friend and -3 might be a not very well liked person;
- link type: foreign key relation to the identifier of an instance of the content type entity. The type determines, for instance, if the relationship between two persons is professional or personal. Hence, several types of social networking services such as OpenBC (a professional network) and Friendster (a personal network) can be implemented with the same infrastructure and using the same data set.

Similar to content links, the participation of person entities as link sources or link destinations in the knows relation is partial.

10.3.2.4 Metadata

The is attached to relation used for assigning metadata to content is a binary relation consisting of:

- metalink id: unique identifier of the relation;
- metadata: the unique identifier of a metadata element;
- content: the unique identifier of a piece of content;
- author: the unique identifier of the person associating the mtadata element with the content.

Notice that the participation of metadata in is attached to is total, while the participation of content is partial.

10.3.2.5 Permission

Several relations are associated with the permission entity. These relations describe which persons and groups hold certain permissions. Furthermore, they determine which pieces of content and which links these rights are applied to.

The may have 1 relation is used for granting permissions to persons. This relation contains the following attributes:

- person permission id: unique identifier of the relation;
- person: the unique identifier of a person;
- permission: the unique identifier of a permission.

In analogy to the may have 1 relation, the may have 2 relation is utilised for granting permissions to groups (for the benefit of readability not shown in figure 10.1). It contains:

- group permission id: unique identifier of the relation;
- group: unique identifier of a group;
- permission: unique identifier of a permission.

The participations of all entities in these two relations are partial, i.e., users and groups may exist without having permissions, and there may be permissions without being granted to users or groups.

The applies to 1 relation assigns the rights granted in may have 1 relations to pieces of content. Every piece of content can have multiple rights assigned.The relation includes:

- content permission id: the unique identifier of the relation;
- person permission: the unique identifier of a may have 1 relation;
- content: the unique identifier of a piece of content.

Similary, the applies to 2 relation associates content links with permissions granted to users. This relation contains:

- link permission id: the unique identifier of the relation;
- person permission: the unique identifier of a may have 1 relation;
- content link: the unique identifier of a links to relation.

The analogous relations do not only exist for users but also for groups. They connect permissions granted to groups (in the may have 2 relation) to pieces of content and content links.

The participation of may have in the applies to relations is partial, while both the content entity and the links to relation are total participations.

10.4 Exemplary Applications

The following sub-sections give a few examples that illustrate how the proposed data model can be utilised for implementing existing community-based technologies. These examples emphasise some key aspects of content generation and information retrieval and do not include functionality required in real-world systems such as user authentication and aspects of the user interface.

10.4.1 Blogs

With the proposed data model, blog entries can basically be realised by introducing appropriate permissions and a new content type "blog entry". One set of permissions called "blog owner" contains the rights read, create, and edit. Another set named "blog reader" contains only read permissions.

The blog owner's person entity is related to the "blog owner" permission by means of the `may have 1` relation. A group containing all users of the system (e.g., "world") is linked to the "blog reader" permissions. Every blog entry is associated with the relation granting "blog reader" permissions to the "world" group, and every blog entry created by the blog owner is connected with the relation that grants "blog owner" permissions to the author's person entity. Moreover, the version of blog entries is set to "1", and the timestamp is set to the creation date of the content.

When A's blog is accessed by users, for instance, and the most recent entries are to be displayed a query similar to the following pseudo code has to be processed: "select all records from content where (content type = 'blog entry') and (author = 'A') and (author is associated with content) and (group permissions = 'blog reader') and (group permissions are associated with content) and (group permissions are associated with group 'world') order by timestamp descending limit 20".

An RSS feed, for example, can be generated by retrieving all blog entries authored by a person from the database. Subsequently the unique content identifier, the content title, the timestamp, and further metadata from the query results are inserted into a feed and sent to the user.

When users want to write comments to a blog entry, a new piece of content is created with content type "blog comment". The comment is connected with the original blog entry using the `links to` relation, where the link source is the the blog entry and the destination is the comment. Locations do not have to be provided, unless deep links are desired.

10.4.2 Wikis

For the use of wikis, a new content type "wiki article" is introduced. The set of permissions for a "wiki author" includes read, create, and edit rights. These permissions are granted to the group "world", i.e., everyone may edit existing articles. For every modification to a piece of content, a new version is stored with a potentially new owner. Thus, after a number of edits, a piece of content might exist in several versions with various authors.

Links from one wiki article to another article are stored in `links to` relations. The link source is the wiki article containing the link, the source location is the position, where the link is to be displayed in the document, the length is the number of characters that are to be displayed as link, and the link type is set to "inline link". The link destination, for instance, might be set to a blog entry.

When wiki articles are displayed, only the most recent version of the corresponding piece of content is retrieved from the database. Links and similar information can be inserted on-the-fly.

10.4.3 Podcasts

Since podcasts resemble blogs, the explanation of the basic functionality in section 10.4.1 applies. However, an aspect of the proposed data model enables features not available in conventional podcasts.

The content link model allows for links to and from media types that do not support links per se. Suppose the destination of a link is a podcast episode, and the link destination location is, for instance, set to "1m23s". In this case, the audio player can jump directly to the specified position in the audio content when the link is selected (given the audio player supports such a function).

However, even a podcast episode can be made the source of a link. To achieve this, the source link is set to the podcast episode, and the link type is set to "external link". The link can be provided as additional metadata to the podcast episode and an audio player can display it.

10.4.4 Tagging Systems

As described in section 10.2.4, tagging systems contain content that is complemented with loose metadata. With the proposed data model, the `content` entity can be used for storing the actual content—bookmarks, images, video clips, etc. The author of the content is the person adding the content to the system. The version of the content is set to "1", and usually remains unaltered.

Tags are added by users in the system by creating instances of the `metadata` entity and by associating these metadata instances with pieces of content through the `is attached to` relation. This makes it possible to retrieve all tags attached to a piece of content and all pieces of content with a given tag. However, it also enables queries for all tags a user has used so far. Additionally, all users using a common tag can be listed.

Notes in photos, available in Flickr, for example, can also be implemented. A note is stored as content with a special content type such as "photo note". A content link with the note as link source and the image as link destination is created. The link destination location and destination length define the spatial area the note refers to.

10.4.5 Social Network

With the `person` entity and the `knows` relation, the most significant requirements for a social network are built into the proposed data structure. With relatively simple queries all friends of a user or all people who have a person on their list of contacts can be retrieved. Exemplary pseudo code for these queries is: "select link destination from person links where (link source = 'A') and (link weight > 0) and (link type = 'personal')" or "select link source from person links where link destination = 'A'".

Further queries can, for instance, produce a list of friends of friends: "select distinct link destination from person links where (link type = 'personal') and (link weight > 0) and link source in (select link destination from person links where (link type = 'personal') and (link weight > 0) and link source = 'A')".

10.5 Conclusion

This chapter introduced a data model that can be used as the underlying infrastructure for several current, collaborative, community-based environments such as weblogs, wikis, file sharing services, and social networks. The model consists of a small number of generalised entities and a set of relations among these entities. This makes it possible to store various types of content, potentially even of systems and services developed in the future. By choosing appropriate permissions for reading, creating, and editing content a range of concepts can be implemented on top of the data model. Moreover, the use of links can result in structures such wikis, connections between blogs and wikis, or social networks of users of the system.

The infrastructure presented in this chapter serves as the underlying model for Kōrero, the collaborative platform discussed in the previous chapter. The functionality of Kōrero—making implicit knowledge explicit, revealing hidden connections, finding people working in related areas, etc.—can be implemented more easily with this unified model than with the inconsistent infrastructures of the systems and services available to date. With the proposed data model, tasks such as finding links between blog authors and wiki editors can be reduced to relatively simple queries.

Conclusion

Summary

In this work, various aspects of new and unconventional digital libraries were discussed. The following sections summarise the designed concepts and the key findings of the research conducted.

Electronic Encyclopaedias

The first part of this thesis addressed electronic encyclopaedias and collaboration in such systems. As previous research shows, Wikipedia, currently the most prominent example of collaborative encyclopaedias, has a number of weaknesses. Since anyone on the Internet may edit content in Wikipedia the information provided, for instance, can be inaccurate and the quality of articles might be affected.

In a move to counter these and other deficiencies, the concept for a web-based collaborative encyclopaedic environment with quality control was introduced in chapter 2 (see also Appendix B). The proposed system, which is in part being implemented in project "Alexander", includes several levels of expert users and a rating mechanism. While the quality of content can be ensured through experts, the rating mechanism can be seen as an incentive for users to contribute to the community and to the encyclopaedic "knowledge base".

The proposed architecture does not allow for the participation of anonymous authors. Since it had previously been unclear whether anonymous users are an essential part in collaborative environments, authorship in Wikipedia was examined in this thesis. The findings of this investigation imply that anonymity is particularly appreciated by users when it comes to delicate topics. Hence it was suggested that all authors in the system have to authenticate (i.e., they have to log on to the system), while they may decide for each article they modify or create if it is "signed" with their name or not. This is a type of pseudo-anonymity.

Most traditional electronic encyclopaedias mimic classic print-published encyclopaedias and do not make use of the full power of distributed, networked environments. Despite its innovative approach to collaboratively authoring content and other novelties, these shortcomings of current encyclopaedias are not considered in Wikipedia either. Therefore a number of technologies that can enhance conventional encyclopaedias were detailed in this thesis. The idea of transclusions as an advanced technique for including existing content into new documents, for instance, was applied to the domain of the World Wide Web. The technology was implemented for both text-based and multimedia documents. Furthermore, an approach to realising adaptation on the structural and content levels in encyclopaedias was presented.

Combining these "components" results in a community-based adaptive system that allows its users to reuse existing information and author new content, while high quality standards for content can be assured. Hence flaws of Wikipedia as well as conventional electronic encyclopaedias could be positively dealt with.

Metadata

The second part of this thesis focused on an important aspect in digital libraries and current trends in this area—metadata. Both conventional approaches to metadata creation such as Dublin Core and the relatively uncommon tagging concept were analysed. Although Dublin Core is a well-established standard for describing resources (on the Web), little was known about the acceptance of this technology. Therefore an investigation on the use of HTML-encoded Dublin Core metadata was carried in academic and educational settings, one of the key application areas. The results show that, although a few institutions make extensive use of this kind of structured metadata, the

overall usage is relatively low. Potential explanations for these results are the lack of suitable tools, the additional effort required, and insufficient awareness of the benefits of structured metadata.

Tagging systems, on the other hand, are very successful tools giving users the opportunity to attach loose, unstructured metadata to all kinds of resources. However, as several authors pointed out before, the simple nature of tagging systems is prone to a number of deficiencies, one of which is the Vocabulary Problem. In this thesis, a solution to this problem on the basis of a semantic lexical database was suggested. This concept was implemented for the Flickr photo sharing service. First experiments with the implementation showed that the query results of the research prototype ("WordFlickr") are at worst as good as conventional Flickr queries. Hence, it might be sensible to include WordFlickr's functionality into next generation tagging systems.

Collaborative Systems

The last part of this thesis introduced a platform that integrates existing collaborative concepts—Kōrero. In the design of Kōrero, technologies such weblogs, wikis, podcasting, social bookmarking, file sharing services, and social networking are combined, and the synergetic effects that might emerge can be utilised. This means that implicit links between content authors, for instance, can be made explicit, links between related pieces of content can be pointed out, etc.

In the proposed design, Kōrero's components for authoring content and for communication are built on top of an all-embracing social network that contains all users of the system. The architecture of Kōrero is based on a data model that unifies the models of existing technologies. The advantage of such a structure lies, for example, in the reduced complexity for finding implicit links. An appropriate data model capable of providing compatibility for both existing systems and Kōrero complemented this work.

Conclusion and Further Research

The scientific contributions in this thesis helped shed light on previously not very well known aspects of library applications (anonymous authorship, metadata). Also, implementations of long-standing proposals (transclusions) and solutions to recently discovered issues (Vocabulary Problem) could be provided, which may lead to improved future services. Moreover, the concept of an enhanced, community-based electronic encyclopaedia with support for quality assurance mechanisms was introduced.

The latter is particularly noteworthy because one of the founders of Wikipedia recently announced a new, international, collaborative encyclopaedia with a concept similar to the one described in this thesis. This can be seen as both a confirmation and an endorsement of the results of this work.

Meanwhile a system with similar, yet more ambitious, aims is developed at the Institute for Information Systems and Computer Media at Graz University of Technology, Austria. Although "Alexander" is based on this thesis, further research will be required for making it a successful, innovative platform with a thriving community. Especially further aspects of the user community and of collaborative functionality will have to be investigated.

It is almost certain, for instance, that a community-based systems with editors (or experts) can only work if the overwhelming majority of users are "plain users" (i.e., the critical mass of participants develops the content). However, there are no reliable data for this statement yet. Moreover, it is unclear if a small number of editors could actually handle the large amount of information constantly produced by users of the community. Therefore a system like Alexander or Citizendium needs to be fully implemented and tested with a large user base in order that the efficiency of experts and a rating system can be assessed.

Additionally, previously unknown and unexpected phenomena will almost inevitably emanate when a system like Alexander is used in a large-scale environment. Therefore further research will become necessary for analysing the effects of the concepts detailed in this thesis.

Not only the proposed encyclopaedic environment can serve as the basis for further research. Although tagging systems, for example, are still relatively "young" they are a booming area, in which unprecedented effects of global communities can be observed. Also, investigations can be carried out on the structures that emerge from tagging systems and representations of these structures. Moreover, instead of solving the problems that are present in tagging systems an attempt could be made to bridge the seemingly insurmountable gap between taxonomies with their controlled vocabularies and tagging systems with the users' arbitrary metadata terms.

As a closing statement, it can be said the uncommon library applications that are currently available will most probably have a major impact on future developments in the field of digital libraries. However, it will remain to be seen if systems like Wikipedia will be capable of overcoming quality issues in the long run or if classic digital libraries will remain the standard in quality publishing.

Appendixes

A The Growing Importance of E-Communities on the Web

A version of this appendix was submitted to the Journal of Universal Science and Technology of Learning.

Publication submitted in May 2006, co-authored by Hermann Maurer.

Until recently, one of the main aims of the World Wide Web has been to offer users a wide range of information. This information was authored, by and large, by professional information providers. Recent advancements on the Internet, however, have changed this paradigm and the clear distinction between information producers and consumers is becoming blurred. New technologies such as weblogs, wikis, file sharing services, podcasting and social networks allow users to become an active part on the Web and let them participate in developing content.

In this chapter, an overview of several successful community-based concepts and services is given. The fact that many of these concepts have existed before is pointed out together with true novelties of their current implementations. Moreover, a critical view of recent communities, their importance and impact is presented. Especially the potential loss of individuality and the movement towards an "integrated society" with a common shared memory is discussed. A look is taken at the future development of e-Communities in the light of ubiquitous access to information with technologies such as "always-on" wearable cameras and E-Ink.

A.1 Introduction

During the last decade, the World Wide Web has evolved into a truly worldwide computer network. Traditionally, most information on the Web was published by professional information providers such as news services, companies advertising their products and offering support, or research institutions. Moreover, personal homepages could be established by users.

Although millions of individuals make use of the Web every day, in the past only a small percentage was capable of actually authoring content and participating on the Web. Primarily technological obstacles including the lack of technical background and complicated tools prevented users from producing web-pages and from participating in other services on the Web (e.g., [Lindahl and Blount 2003]). The only successful exceptions to this rule are discussion forums, communities for diseases and disabilities, e-learning systems, and dating services.

Recently, however, concepts and services that let users become a part of the content creation and distribution process have been introduced on the Web. "Novel" systems including blogs, wikis, file sharing services, and social networks have started a movement towards more user participation, and users are not only information consumers but are capable of authoring content, modifying existing content, and sharing it with other users on the Internet.

This chapter gives an overview of a number of popular, community-based services on the Internet, discusses their impact on our lives, and gives an outlook on future applications and their significance. Section A.2 introduces blogs, wikis with Wikipedia being the most prominent example, file sharing services including podcasting, and social networks. The novelty and effects of these technologies are discussed in sections A.3 and A.4. Section A.5 makes an attempt to forecast the role of these concepts in conjunction with forthcoming technical developments.

A.2 Community-Based Services on the Web

In recent years, a host of new, mainly community-based concepts and services was introduced on the Internet. By some, these new technologies were coined the "Web 2.0", emphasising both the evolutionary

Figure A.1: Blogger, one of the free and easy-to-use services for creating blogs. (All screenshots taken on May 6th and 7th, 2006, from URLs shown in the address bars of the respective images.)

Figure A.2: Slashdot—probably the best known and most popular filter-style blog currently available.

process the Web is undergoing and the innovation of the novel products (see [O'Reilly 2006] and chapter 1). The following sub-sections describe a selection of relevant community-based services on the Web.

A.2.1 Blogs

Blogs, short for weblogs, are a form of web pages that contain articles similar to newsgroup postings in a reverse chronological order. Blog entries are usually produced by a single author or by a small group of authors and cannot be edited by the public. Postings on blogs are regular, typically once a day (see [Blood 2002]). Their content is similar to a combination of diaries, editorials in magazines, "hotlists", and the "breaking news" section on news channels. Contributions frequently refer to a current event such as a news story, a discussion in parliament, the release of a new record, etc.

In May 2006, Technorati, a service tracking blogs and the links between blogs, indexes almost 40 million blogs (source: [Technorati 2006]). About eleven percent of American Internet users have read blogs, and two percent have actually maintained blogs in 2004 (see [Lenhart et al. 2004]).

A.2.1.1 Blog Styles

Currently, two notable types of blogs are available: diaries or personal journals (accounting for about seventy percent of all blogs) and filters (about ten to fifteen percent, see [Herring et al. 2004]). In diaries and personal journals, authors make details of their personal lives and their views on various topics public. The first diary-style blog believed to have been published was started in January 1994 by Justin Hall, then a college student (e.g., [Pollock 2001]). Nowadays, personal journals are particularly well-liked among young people who want to tell friends (and absolute strangers) about their experiences. An example for a young woman's blog discussing taboo and provocative topics is "Miss Izzy", a rather popular web-site in Singapore (see figure A.3; [MissIzzy 2006]).

Filter-style blogs aggregate links to noteworthy resources on the Internet. Links are usually complemented with short summaries of the respective resources' content and comments added by the author. The scope of such blogs is often limited to a particular topic such as globalisation, music, or computers and technology. One of the best known filter-style blogs is Slashdot, a web-site with a very large user base focussing on technology (see figure A.2; [Slashdot 2006]).

Due to their nature, blogs are intrinsically opinionated. They allow users to express themselves and present their views to a broad audience. However, weblogs are not only employed in personal environments but also in organisations and enterprises. They are utilised for keeping employees informed on the status of projects, of new policies or similar news (e.g., [Treese 2004]). Moreover, they can be used to encourage the communication and co-operation between various departments in large organisations.

A.2.1.2 Technical Aspects

A major part of the success of blogs is their ease-of-use—even for novices. Using blogs is about as difficult as writing e-mails and organising them in mailboxes. Moreover, free services like Blogger make it possible to start a new weblog within a few minutes (see figure A.1; [Blogger 2005]). Hence, they are often used as a replacement for traditional homepages.

Another aspect that makes blogs a popular means for communication is a set of technologies that greatly enhance community-building among users: permalinks, trackback, and RSS (see [Efimova and de Moor 2004]). *Permalinks* are persistent URLs to single postings on a blog. When an author refers to another blogger's article, the permalink to this entry can be used. If the two blogs are *trackback*-enabled a link from the newer blog entry to the existing one is established automatically. Thus with trackback, blog entries can be linked

Figure A.3: Missizzy with its sometimes controversial topics is one of the most popular diary-style blogs in Singapore.

Figure A.4: No other wiki has more pages or more authors than Wikipedia.

practically bidirectionally, and blog authors are notified about being cited by someone else (cf., [Maurer and Tochtermann 2002]).

RSS ("RDF Site Syndication" or "Really Simple Syndication") is a technology that lets users retrieve a list of changes made to a blog, or a web-site in general. Users subscribing to an *RSS feed* are provided with the titles of new articles, short summaries, and the URLs to the full blog entries. When users read an interesting article on a blog they have subscribed to (using RSS) they can write a blog entry in their own blog and refer to the original posting using a permalink. The author of the initial posting is informed trough trackback.

This combination of technologies builds a network of more or less loosely connected blogs—the *blogosphere.* RSS feeds in particular foster community-building among bloggers. This relatively simple mechanism helps users stay up-to-date on blogs and people they are interested in, and transforms occasional visitors into frequent readers.

A.2.2 Wikis

The term "wiki" is derived from *wiki wiki,* which is Hawaiian for "quick". This word is an appropriate description for Ward Cunningham's notion of a concept for the rapid development and organisation of web pages (see [Leuf and Cunningham 2001]). Wikis are collaborative, web-based authoring environments, where anyone on the Internet can edit existing content and add new pages any time they wish. In other words, every reader can instantly become a writer.

This concept is in stark contrast to authoring systems previously widely available on the Web. Content management systems, the de-facto standard for large web-sites, for instance, make use of hierarchical rights management, and a publishing process similar to the one employed in newspaper publishing. Such a system usually incorporates administrators, editors, authors, and mere readers. A wiki, on the other hand, does not distinguish between readers, authors and editors; they have the same capabilities in the system.

This aspect is of particular interest because initial authors of articles allow other users to modify "their" content. Although this approach may seem utterly chaotic, there are several very large wiki sites offering quality content provided by the community (see below). One characteristic that makes wikis work is the aim to reach an agreement among all authors of an article. Hence, the content of single wiki articles is usually agreed upon, unbiased, and neutral.

A.2.2.1 Advantages and Drawbacks of Wikis

Their "open" nature makes wikis more flexible than conventional, editor-based web-sites. When new information becomes available it can be added to the wiki immediately, without an editor's approval. Similarly, when an error is found by a reader it can be corrected by the reader, without the need to contact the site's administrator or the author of the document. Moreover, wiki documents can be written using a relatively uncomplicated syntax, and features such as version control make wikis well-suited for collaborative environments.

At the same time, the openness of wikis poses a number of problems. Since quality control through editors is not in place errors might be inserted accidentally, or even deliberately. Readers, on the other hand, might mistake the information provided on a wiki site for reliable. Another problem is vandalism, where incorrect information, defamatory content, and advertisements are inserted, existing content is deleted or overwritten, etc. In many cases, however, such acts of vandalism are repaired within minutes by reverting a page to its previous version (cf., [Viégas et al. 2004]).

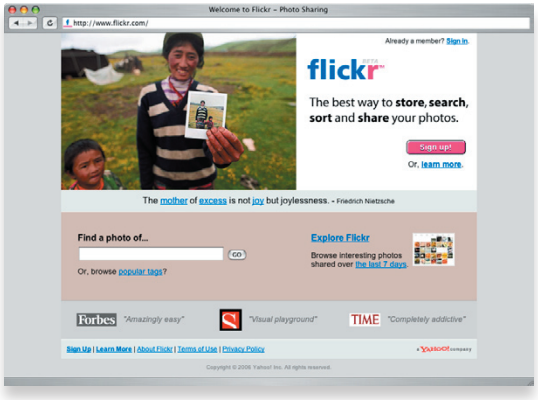


Figure A.5: Flickr—a popular photo sharing service that makes use of tagging.

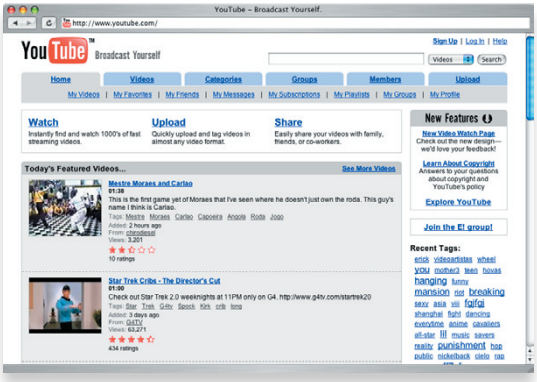


Figure A.6: YouTube is a service where users of the community can post the video clips they produced.

A.2.2.2 Wikipedia

The largest wiki to date is Wikipedia, a free online encyclopaedia available in more than 200 languages (see figure A.4; [Wikipedia 2006a]). Since it is based on the wiki concept, every reader is also an author and can add or modify information instantly—even anonymously. This is one of the reasons for the project's rapid and steady growth: from Wikipedia's founding in 2001, more than 1.1 million articles have been written in the English edition, and about 400.000 pages in the German edition. Wikipedia offers more material than many other, commercial encyclopaedias, and can deliver more supplementary content such as hyperlinks to information on the Web than many other works of reference. This makes people trust in Wikipedia more than in other resources.

However, as in any other wiki, any information provided might be erroneous because quality assurance mechanisms are not available. Research shows, though, that Wikipedia articles contain, on average, only about 25 percent more errors than renowned, for-profit encyclopaedias (see [Giles 2005]).

One of the most striking examples of incorrect data in Wikipedia is the case of journalist John Seigenthaler. A false biography published on Wikipedia associated him with the assassination of John F. Kennedy and alleged that he collaborated with the Soviet Union in the 1970s (see [Seigenthaler 2005]). As a consequence, the rights of anonymous authors have been restricted. In addition to this, a peer review mechanism for articles in Wikipedia has been discussed. (Although this feature was due for January 2006 (see [Wales 2005]), it has not been realised yet.)

In some cases, incomplete content can be just as bad as wrong information. An article that lists a politician's successes while deliberately omitting the promises that were not implemented is obviously not balanced and leaves a wrong impression. Similarly, due to systematic bias it is difficult to provide unified views in Wikipedia. Although one of aims of the encyclopaedia is to be unbiased, social and cultural differences as well as different national and lingual backgrounds might have an influence on the content. On May 14th, 2006, the English article on the Hungarian scientist John von Neumann, for instance, was about five times as long as the corresponding article in the German edition and included detailed accounts of his research and a comprehensive list of external references. Thus, even if both articles are written in an unbiased and objective way, an imbalance due to the background of the authors and the target group can be observed. In professional, editor-based encyclopaedias this kind of systematic bias is countered by authoring guidelines that set standards for the length of articles, etc.

A.2.3 File Sharing Services

File sharing is probably best known in conjunction with applications that allow users to share any kind of files over the Internet such as Napster or Kazaa. These services are primarily used for downloading music and movies illegally. Recently, however, also several legal file sharing services have been introduced. These systems are usually web-based, provide users with a private space for storing documents, help users organise content, and let them make documents publicly available.

A popular file sharing tool is Flickr, a service for sharing and organising photos (see figure A.5; [Flickr 2006]). With Flickr, users upload their photos to a server, can add comments and leave notes inside images. Additionally, users can attach tags to every photo uploaded. Tags resemble keywords that loosely describe the content of the corresponding image. A photo of a family can, for instance, be tagged "wedding", "May 2006", and "Vienna". Consequently, a query for "Vienna" would also list this photo as a result. Besides searching, users can also browse the vast archive using tags. Every photo shown in Flickr is supplemented with the tags assigned by the author. By selecting a tag, all images with the same tag are displayed.

Figure A.7: CurrentTV does not only offer home-made videos on their web-site but also broadcasts a selection on a traditional TV channel.

Figure A.8: The podcast directory integrated in Apple's iTunes Music player.

A service extending this concept to the domain of motion pictures is YouTube (see figure A.6; [YouTube 2006]). With YouTube, users can share video clips they produced and employ a tagging mechanism similar to the one in Flickr. CurrentTV takes video sharing a step further. Not only can viewers produce their own video clips and publish them on CurrentTV's web-site, but a selection of video clips is also broadcast on conventional television channels (see figure A.7; [YouTube 2006]). Thus, viewers may even become TV producers.

A.2.3.1 Podcasting

A slightly different approach to sharing content on the Internet is podcasting. Basically, podcasting means blogging audio content. Content producers regularly upload audio files in MP3 audio format to a server and insert references to these new files into an RSS feed. Listeners subscribing to a podcast (actually to the RSS feed) have access to the full list of audio files made available by the producer and are notified about newly published content. On the users' request, audio files are downloaded. Therefore, podcasting can be seen as a type of "audio on demand" service (see [Biever 2005]).

A directory integrated in Apple's iTunes music player software catalogues thousands of podcasts (see figure A.8). Podcasts are available on a wide range of topics ranging from self-made music to amateur talk and radio shows, from religious programmes and masses to professionally produced shows such as the Nature Podcast (see [Nature 2006]). Lately, podcasting has also been identified as a technology for enhancing existing e-Learning applications and distance learning initiatives. Lectures and discussions are recorded and provided free-of-charge as podcasts (e.g., [DukeCast 2006]). In a similar fashion, conference presentations are disseminated as podcasts on the Internet (e.g., [JISC-CETIS 2005]).

A.2.4 Social Networks

Social networks are structures that describe the social relations among individuals. Every node in the network is a person, and edges between nodes are the connections among individuals, where the weight of edges can be used to denote the degree of "amity". In recent developments, the concept of social networks, previously mainly used for describing existing social structures, was successfully applied to the online world.

On the Web, social networks are chiefly utilised for maintaining relations with friends and acquaintances and for making new friends. Such services offer basic functionality for chatting with members of the network, for sharing information, etc. Users joining a social network have to fill out a profile containing information such as the person's name, date of birth, and a photo. These data are made available to members of the network in order that they can find their friends. Moreover, most social networks do not only let users view their friends but also their friends' friends (second degree friends). This feature clearly facilitates creating new connections in the network.

Well-known examples for general-purpose social networks are MySpace, the sixth most popular website worldwide (source: [Alexa 2006]) with more than 78 million registered users, and Friendster with about 27 million users (see figures A.9 and A.10; [MySpace 2006; Friendster 2006]). In addition to this, specialised services for people with similar interests have been established. OpenBC, for instance, is a social network of professionals with the aim of creating a web of trusted business partners and experts (see figure A.11; [OpenBC 2006]).

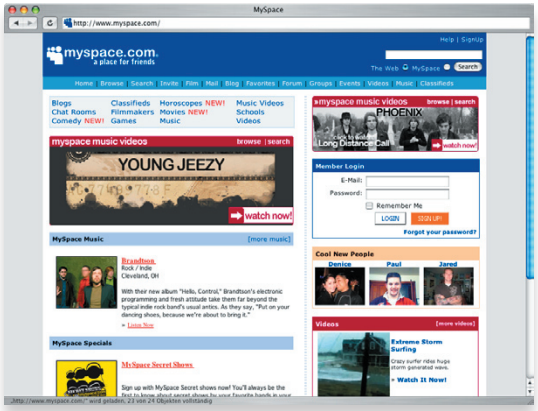


Figure A.9: MySpace, a social network with more than 78 mn registered users—the sixth most popular web-site worldwide.

Figure A.10: With about 25 mn members, Friendster is one of the most successful social networks.

A.2.5 Other Community-Based Applications

A new class of applications combines features of social networks and file sharing systems. del.icio.us, for example, is a social bookmarking service (see figure A.12; [del.icio.us 2006]). Users can retain bookmarks of favourite pages in the del.icio.us database instead of storing them on their local computers. Like in Flickr, users can attach tags to bookmarks and use these tags for finding similar bookmarks in the system. Since users can see who else bookmarked the same web page, it is possible to find people with similar interests. Hence, del.icio.us is not only a platform for sharing information but also includes mechanisms from social networks.

Furl is a service similar to del.icio.us (see figure A.13; [Furl 2006]). In Furl, not only bookmarks but the actual resources from the Internet are stored in an internal database. This means that users can create a space only containing the web pages they want to store—their own "Private Web".

Further services driven entirely by the community are, for instance, Eventful and OhmyNews. Eventful is a web-site listing events for almost any region in the world (see figure A.14; [Eventful 2006]). The events offered by this service together with a short description, the exact location and additional information are submitted by members of the community. The second application, Ohmy News, is a blog-like news service in which articles are authored by "citizen reporters"—amateur journalists from the global community ([OhmyNews 2006]). Ohmy News is often faster than traditional news providers and can offer in-depth information written by locals and first-hand witnesses.

A.3 What Is Really New? What Is Different?

On close inspection, it can be seen that the "novel" concepts and technologies introduced above are essentially nothing new (cf., [O'Reilly 2005]). Similar services have been in use earlier, for example, in hypermedia systems such as Xanadu, Microcosm or Hyperwave ([Nelson 1981; Fountain et al. 1990; Maurer 1996]).

However, there are certain aspects that distinguish these new applications from previous implementations. The novelty is not what these services do but how they achieve it. Moreover, all of the concepts and services introduced in this chapter have one aspect in common: they get better the more people use them (see [O'Reilly 2005]). The more people get involved in environments such as Wikipedia or Ohmy News, the more respectable the results get.

A.3.1 Blogs

Newsgroups, letters to the editor, editorials, and "what's new" pages as parts of larger systems have existed before weblogs were conceived. Blogs, however, let users only write short articles and comments, while they are not offering functionality beyond these simple operations. Blogging software is usually a lightweight application that is not overloaded with functions users rarely make use of. Moreover, most blogging tools are free, easy to use, and hardly any special skills or technical background knowledge are required.

Most importantly, weblogs give users on the Web a chance to participate. Editorials, for example, are "passive" for most users. There is a small group of authors and a large mass of readers. Readers, however, do usually not have a way to comment on editorials on the same level; they could write a letter to the editor, but this letter might be shortened or not published at all. With blogs, however, every reader can start a new blog and discuss, or comment on, someone else's article. Additionally, a small set of technologies including trackback and RSS helps forming a blogging community.

Figure A.11: OpenBC is a social network of professionals.

A.3.2 Wikis

Although wikis put forward a new concept that was previously unknown on the Web, the basic idea is far from original. In one of the early designs of the World Wide Web, Tim Berners-Lee describes the system as the "read/write Web", where users are not only able to read documents but can also author documents (cf., [BBC 2005; Gillmor 2004]). Even earlier, Ted Nelson's concept for Xanadu, the genuine hypertext system, involved versioning and allowed any user of the system to produce new content and share it with other users. Software such as Hyperwave implemented these features.

Further, Wikipedia is not the first attempt to establish a free encyclopaedia on the Internet. This notion dates back to October 1993 when the Interpedia Project was proposed on the Usenet (e.g., [Foust 1994]). While letting users participate in developing the content of the encyclopaedia, Interpedia offered an approach to quality assurance by providing seals of approval (SOAP). With this mechanism, various independent organisations could rate articles in the encyclopaedia and confirm the accuracy and quality of content. Displaying an article would also present the various seals of approval granted by organisations, making it easier for users to trust the information provided by the community.

A.3.3 File Sharing

Although the file sharing concept on a large scale is relatively new, the basic technologies for enabling file sharing have existed since the early days of the Internet. Anyone can set up an FTP or HTTP server on their computers, for example, and offer any kind of content to other users on the Internet. Software such as Napster employs proprietary protocols for the same purpose and add indexing and search functionality to the service. This makes the application purpose-built and easier to use. The same is true of podcasts: even years ago it was possible to make audio files publicly available on a web server. However, only with technologies such as RSS this became attractive and relatively consumer-friendly.

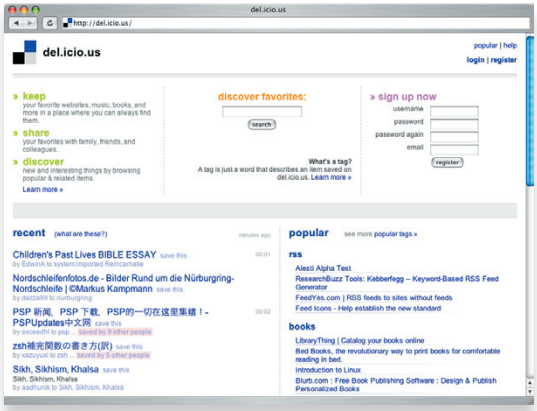


Figure A.12: del.icio.us, one of the pioneers in social bookmarking services.

An innovative feature in recent web-based file sharing services is tagging. This functionality distinguishes Flickr from other approaches to organising large amounts of data (cf., [Mathes 2004]). In previous environments, strict taxonomies were employed, which usually limited the use of such systems to (domain) experts. Although annotations could have been employed for organising content in the same way as tags are used nowadays, a classification of data based on such loose metadata did not seem reasonable.

A.3.4 Social Networks

Social networks in the physical world have existed for a long time. Clubs, associations of people with shared interests, workgroups, and similar societies were successful even before the Internet was developed. In the 1980s, characteristics of social networks were introduced in computer-mediated systems. Among other functionality, these "computer supported cooperative work" environments allow for collaboratively authoring content, sharing and organising information, maintaining relations among members of the systems, and direct communication (e.g., [Schmidt and Bannon 1992]).

Today's social networks include significantly less functionality. Most systems currently available only focus on their main purpose—communication—and do not incorporate diverse functionality such as collaborative authoring. This lowers the barrier to entry and makes social networks easier to handle, even for novice users.

A.4 Impact of Recent Community-Based Developments

The driving factor behind the transformations the Web is undergoing is probably not a set of new technologies but a fundamental mind shift in users and organi-

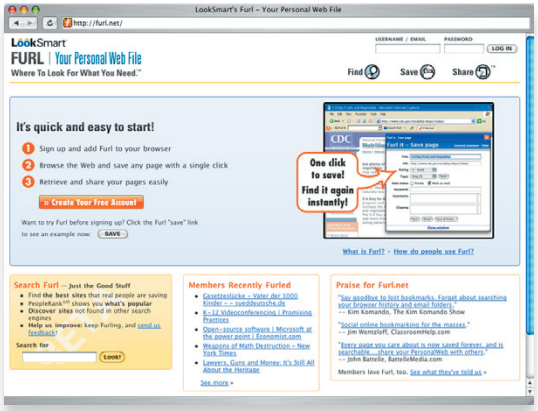


Figure A.13: Furl is a social bookmarking service that lets users not only store references to pages but the actual resources.

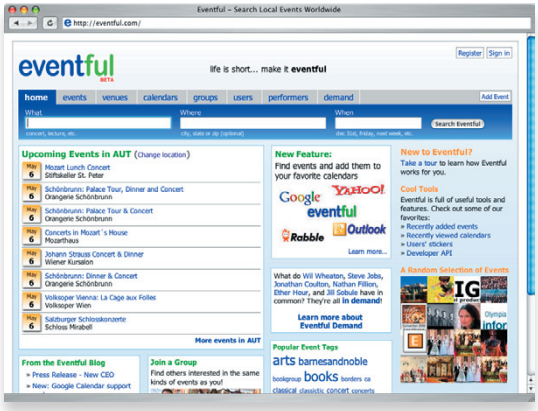


Figure A.14: A worldwide event directory, where users submit information on upcoming events.

sations alike. Users wish to participate instead of using the Web only passively as readers. Furthermore, there is a willingness to share content, and even companies grant access to their content databases. One of the best examples is Google Earth, where satellite images and geographic information are made publicly available ([Google 2006]).

For private users, this new tendency "materialises" in a combination of blogs, file sharing services and social networks. Regular postings on weblogs provide a continuous stream of thoughts, experiences and emotions, while services such as Flickr or YouTube deliver complementing photos and videos. A social network offers the infrastructure for maintaining the ties with friends and acquaintances (cf., chapter 9).

An immediate effect of these developments is that the world is getting yet "smaller". Teenagers in Europe, for example, can have friends in New Zealand and Singapore and can be a part of their lives as if they were next-door neighbours. Although, from a technical perspective, this was possible years ago it has become reality only recently with the services detailed above (and other technologies such as free Internet telephony). Especially the blend of social networks and blogging is intriguing: in social networks, it is possible to find new friends or rediscover "old" friends one has not been in touch with for a long time. Friends' blogs, on the other hand, are the means to stay informed on their daily activities—from minor events such as buying a new CD or doing a mountain bike tour to a three-month trip to South America.

From a less enthusiastic perspective, it can be argued that our society is heading in a direction where individual experiences become increasingly rare. Imagine holidays in Papua New Guinea. On the Internet, there are travel-related blogs written by people with first-hand experience, there is a wide range of photos and videos from the country, etc. Thus, even before actually going on holidays to a country we have not been before, we will have a very good impression of what to expect, what we will see and what it will be like.

A.5 Future Advancements

Looking back at the development of communication among human beings, one can see that written language started about 6,000 years ago. About 600 years ago, the invention of book printing increased the importance of written language, and some 200 years ago the introduction of compulsory education further spread reading and writing. About 150 years ago, photography and telephony were invented, and some 100 years ago moving pictures were introduced. About 80 years ago, radio broadcasting was started. 70 years ago, anyone interested and able to afford it, could buy a camera; forty years ago, the same was true of video cameras. Also about forty years ago, television was widely accepted. About twenty-five years ago, the Walkman was introduced. Twenty years ago, amateur video cameras became affordable and widely available. Fifteen years ago, computer networks were implemented in numerous organisations. Ten years ago, the Web (and the Internet in general) took off. About five to ten years ago, concepts such as wikis, weblogs, and social networks were introduced.

Considering that reading and writing became available to the public only about 200 years ago, that widespread technologies such as television have been accepted for only forty years or so, and that blogging came into being only ten years ago, it can be assumed that recent technologies permitting user participation will have a deep impact on almost anyone on the Internet in only a few years.

It can be believed that, in the future, most information will be accessed through a networked computer. Moreover, most people on the Web will be members of at least one social network or specialised community. Thus, it can be assumed that the technologies employed today and their usage are just the beginning of

a movement towards an "integrated society". Already today users reveal more personal and intimate details on the Internet (even to absolute strangers) than most people would have expected a few years ago. [MissIzzy 2006], mentioned above, is a good example (see section A.2.1.1). On the one hand, the young female author posts very personal experiences and intimacies on her weblog. On the other hand, readers can relate to the experiences and feelings, sometimes even to a degree where they have the perception that *they* had the experiences, although they only read about it.

To an increasing degree, individuals merge their own experiences with the experiences of other users on the Internet. This is an important side-effect of modern communication systems. Consequently, in the future people might find it hard to have any truly individual experiences.

Figure A.15: Demonstration of e-Ink technology. e-Ink paper is thin, flexible, and content can be changed as if it were a regular display (source: [eInk 2005]).

A.5.1 Upcoming Technologies

A key to the further success of the recent community-based services, and the formation of an "integrated society", might be ubiquity. In order to be able to have a "live experience" of someone else's life, this other person has to provide a stream of data (text, images, video, sound, etc.) from virtually anywhere. At the same time, viewers need to have ubiquitous access to this information.

Both aspects will be possible in the future, as current technologies readily demonstrate. Hewlett-Packard, for instance, conducts research on an "always-on" wearable camera that captures what the user sees (see [HP 2004]). The data is recorded continuously and is to be stored in data centres. With such a camera, individuals could provide a continuous stream of their life, from their perspective, on the Internet.

E-Ink paper is another technology that has the potential to revolutionise the way we have access to information. E-Ink paper is a material that requires electricity to "load" an image into a matrix-based display but does not consume any energy while displaying the image. The content of a page can be changed as if it were a regular computer display. However, electronic paper is thin, flexible, can display both monochrome and colour images, and is low in power consumption (see figure A.15; [eInk 2006]).

With high resolution wearable cameras, foldable displays, and fast, wireless network connections, users can publish information anytime from anyplace and can have ubiquitous access to information on the Internet. Imagine, for example, your partner sitting at home, watching your first presentation at a conference—from your perspective. Or imagine being on holidays in Papua New Guinea. With the upcoming technologies, your friends can "tune in" to your holidays anytime they want to (when you make the video stream publicly available). After a few weeks or months, absolute strangers can watch part of the video stream on your video blog. For these users, your video blog can be a replacement for a conventional travel guidebook. Or, even more radically, your video stream can be a complete substitute for the actual trip!

A.6 Conclusion

In recent years, the Web has grown into a network for community-based systems and global collaboration. Concepts for creating and managing information such as weblogs and wikis, file sharing services and social networks attract millions of users on the Web. With these novel developments, users are both willing to participate and willing to share content, experiences, thoughts and emotions. Although these advances are able to offer users unprecedented opportunities and are generally viewed positively, we should be well aware of the risks that may be involved.

The new forms of communication are part of a transformation that goes almost unnoticed. In fact, mankind is about to develop an integrated society with a "shared memory" stored on the Internet. As a consequence, a reduction of unique individual experiences seems unavoidable. What is more, our society relies increasingly on network-based services, even to a degree where we become dependent on the Internet. This can be particularly problematic in such sensitive areas as interpersonal communication. Hence, it will be exciting, on the one hand, to see future developments in electronic communities on the Web. On the other hand, it remains to be seen which effects the growing influence of information technology will have on our society and everyday lives.

B

Alexander: A Basic Functional Specification

Alexander is a Web-based, collaborative environment that combines encyclopaedic knowledge and current information based on the ideas and concepts presented in this thesis. This text is the basic functional specification for project "Alexander" (in German). In many aspects, it resembles chapter 2 of this thesis—"Community-Building around Encyclopaedic Knowledge".

A first prototype of Alexander was made available to a closed user group of about 700 users in September 2006.

B.0 Konventionen

Ziel dieses Dokumentes ist es nicht nur, das Produkt, das aus dem Projekt Alexander hervorgehen soll, in seiner Funktionalität grob zu beschreiben, sondern auch, das Projekt möglichst anschaulich nach außen hin zu präsentieren. Zu diesem Zweck wird eine informelle Darstellung gewählt, mittels derer die Funktionen und Module des Systemes anhand verschiedener Benutzerszenarien geschildert werden. Diese benutzerorientierte, funktionale Grobspezifikation soll auch als Grundlage für die Erarbeitung einer technischen Spezifikation der einzelnen Module des zu entwickelnden Produktes dienen.

Die Umsetzung des Projektes ist in drei Stufen geplant. Bis Dezember 2005 soll ein lauffähiger Prototyp erstellt werden, der die wesentlichen Funktionen des Systems enthält und von einer geringen Anzahl von Benutzern verwendet werden kann. In einer zweiten Stufe wird ein hochskalierbares, kommerzielles System entwickelt, das die Funktionalität des Prototypes anbietet, aber von einer großen Anzahl von Benutzern angewendet werden kann. In einer dritten Phase wird das kommerzielle System um zusätzliche Funktionen erweitert, die weder für den Prototyp noch für die erste öffentliche Version essentiell sind.

Um dieser Realisierungsstrategie Rechnung zu tragen, werden die in den nachfolgenden Abschnitten beschriebenen Funktionen bezüglich ihrer Umsetzung nach drei Kategorien priorisiert:

- M – „mandatory": die Funktion kommt sowohl im Prototyp als auch in der kommerziellen Version vor;
- N – „nice to have": es wird erwartet, dass die Funktion in der kommerziellen Version umgesetzt wird, im Prototyp hat sie aber eine niedrigere Priorität;
- O – „optional": diese Funktion hat weder im Prototyp noch in der kommerziellen Version hohe Priorität, die Umsetzung kann in Phase 3 erfolgen.

B.1 Einleitung und Ziele

Alexander ist ein Projekt, aus dem ein System hervorgehen soll, das verschiedene Datenquellen wie der *Brockhaus Enzyklopädie Digital* (Enzyklopädie) und *Die Presse* (Tageszeitung) zusammenführt und eine Community um die kombinierten Ressourcen aufbaut. Eines der obersten Ziele ist es, den Benutzern des Systemes zu ermöglichen, ihr eigenes Wissen in vielfältiger Weise in die Datenbasis einzubringen und dabei gleichzeitig die Qualität der gesammelten Informationen auf möglichst hohem Niveau halten zu können.

Beschrieben wird ein Web-basiertes System, das zurückgreift auf verschiedene Datenbestände wie sämtliche Artikel und Zusatzinformationen einschließlich Bildern, Tondokumenten und Videoclips

der *Brockhaus Enzyklopädie Digital* und das Archiv von *Die Presse.* Über herkömmliche Web-Browser ist es Benutzern möglich, datenbestandsübergreifend nach Informationen zu suchen, diese anzeigen zu lassen und mit Unterstützung innovativer Technologien weiterführende Informationen zu finden. Benutzer haben die Möglichkeit, selbst Daten hinzuzufügen, Diskussionen zu führen oder in Kontakt mit anderen Benutzern zu treten.

Eine weitere Kerntechnologie ist das Active Knowledge Brokering, das es Benutzern erlaubt, innerhalb einer Kategorie beliebige Fragen an das System zu stellen (siehe Abschnitt „Active Knowledge Brokering" unten). Fragen werden mittels Active Documents Funktionen entweder selbständig vom System beantwortet oder werden von Experten oder Redakteuren beantwortet.

Abschnitt 2 gibt einige allgemeine Definitionen über häufig verwendete Begriffe in diesem Dokument, und der dritte Teil beschäftigt sich mit allgemeinen organisatorischen Aspekten des Systemes. Teil 4 beschreibt Funktionen, die Benutzern direkt zugänglich sind, während der fünfte Abschnitt die Funktionalität, die vom System intern durchgeführt wird, beleuchtet. Der sechste und letzte Abschnitt weist auf einige potentielle Entscheidungen bezüglich des technischen Designs des Systemes hin.

B.2 Allgemeine Definitionen

B.2.1 Benutzer

Ein Benutzer ist eine reale Person, die mit dem geplanten System arbeitet. Ein Benutzer hat einen Benutzernamen und ein Kennwort. Um das System verwenden zu können, muss sich jeder Benutzer mit diesen Accountdaten anmelden. Innerhalb des Systemes ist der Benutzer durch seinen Benutzernamen oder seinen realen Namen identifizierbar.

Es werden zumindest fünf Arten von Benutzern unterschieden: Administratoren (AD), Redakteure (ED), zertifizierte Experten (CE), Domänenexperten (DE) und einfach Benutzer (PU; siehe unten).

B.2.2 Artikel

Artikel werden als HTML-Dokumente an den Client gesendet und können Bilder, Tondateien, Videoclips und Animationen enthalten. Jeder Artikel gehört zumindest einer Kategorie an, kann aber auch mehreren Kategorien zugeordnet werden.

Der Grundinhalt von Artikeln wird von ED verfasst. CE, ED und PU haben die Möglichkeit, selbst Artikel zu erweitern, Annotationen oder Links zu externen Ressourcen hinzuzufügen, etc. Artikel, die von PU oder DE geschrieben werden, müssen von CE oder ED bestätigt werden. Erst durch die Bestätigung werden Artikel publiziert und nur dadurch werden sie anderen Benutzern zugänglich gemacht.

Funktionen, die im Folgenden für Artikel beschrieben werden, sind in der Regel sinngemäß auch auf andere Arten von Inhalten wie etwa Bilder, Tondaten, Videoclips, Animationen und ähnliche Medien anwendbar.

B.2.3 Kategorien

Das System verwendet zur Organisation von Inhalten eine Hierarchie von Kategorien und Unterkategorien. Jede Kategorie kann eine nicht-begrenzte Anzahl von Unterkategorien enthalten. Jeder Artikel wird entweder einer Kategorie oder einer Unterkategorie zugeordnet (vgl. derzeitige Umsetzung in der *Brockhaus Enzyklopädie Digital).*

An Kategorien können – wie an Artikel – Annotationen, Links zu externen Ressourcen und Diskussionen angefügt werden.

B.2.4 Annotation

Eine Annotation kann ein Kommentar eines Benutzers sein, ein Label wie „wichtig" oder „für Projekt A" oder auch ein Hyperlink zu einer externen Ressource. Jegliche Art von Annotation ist an bestehende Artikel oder Kategorien angehängt; eine Annotation kann nicht alleine für sich stehen. Annotationen gibt es auf verschiedenen Zugriffsebenen:

- private Annotationen, auf die nur der Autor Zugriff hat;
- Annotationen, auf die eine gewisse Gruppe von Benutzern Zugriff hat; und
- öffentliche Annotationen, auf die jeder Benutzer des Systems zugreifen kann.

Beim Anlegen von Annotationen wird eine Variante des Blacklistings verwendet.

B.2.5 Hyperlink

Ein Hyperlink (oder einfach „Link") ist eine Referenz auf Content im internen System oder auf eine externe Ressource. Diese Links sind nicht notwendigerweise unidirektional.

Das System unterstützt auch „tiefe Links" (O). Das heißt, der Benutzer hat die Möglichkeit, einen beliebigen Teil eines bestehenden Artikels oder eines anderen inhaltsbezogenen Objektes (jede Art von Content) als Ziel eines Hyperlink zu bestimmen. Damit kann der Benutzer beispielsweise einen Link auf einen Absatz in einem längeren Artikel erstellen – und nicht nur auf den Artikel selbst.

Beim Anlegen von Hyperlinks wird eine Whitelisting Strategie verfolgt.

B.3 Organisation

B.3.1 Benutzerrollen

Das System basiert auf zumindest fünf verschiedenen, hierarchisch angelegten Arten von Benutzern:

- Administratoren (administrator, AD),
- Redakteuren (editor, ED),
- zertifizierten Experten (certified expert, CE),
- Domänenexperten (domain expert, DE) und
- normalen Benutzern (plain user, PU).

Wenn nicht anders erwähnt, können sämtliche Funktionen von Benutzern aller Rollen verwendet werden. Administratoren haben *immer* Zugriff auf *jede* vom System angebotene Funktion und auf alle Inhalte im System (also auch auf Annotationen anderer Benutzer, etc.).

B.3.2 Kategorien

Das System verwendet eine Hierarchie von Kategorien und Unterkategorien, um Inhalte bestimmten Themenbereichen zuzuordnen. Jeder Artikel wird zumindest einer Kategorie zugeordnet; jedes andere inhaltsbezogene Objekt (Annotationen, etc.) kann einem Artikel oder wiederum einer Kategorie zugeordnet werden.

Um die hierarchische Organisation von Artikeln in der *Brockhaus Enzyklopädie Digital* nachzubilden, sind derzeit nicht mehr als zwei Ebenen von Kategorien vorgesehen; es ist aber möglich, die Granularität bei Bedarf zu erhöhen.

B.4 Benutzerfunktionen

B.4.1 Artikel und Kategorien

Artikel sind das zentrale inhaltsbezogene, und Kategorien das wesentliche strukturierende Element im System. Die folgenden Abschnitte beschreiben die wesentlichen Operationen, die Benutzern im Zusammenhang mit Artikeln und Kategorien vom System angeboten werden.

B.4.1.1 Artikel suchen (M)

Actor: Alle Benutzer.

Interaction: Suche nach Informationen (insbesondere Artikeln) im System.

Goal: Auffinden der gewünschten Information.

Scenario: Ein Benutzer gibt in einer Suchmaske einen Begriff ein. Nach Durchführung einer Suche in allen verfügbaren internen Datenbeständen werden die Ergebnisse in einer übersichtlichen, intuitiven und optisch ansprechenden Weise dargestellt.

Benutzer haben die Möglichkeit, das gesamte System nach Informationen zu durchsuchen. Dem Benutzer stehen dabei zwei Eingabemasken zur Verfügung: eine bewusst schlicht gehaltene Suchmaske bestehend aus einem einzigen Texteingabefeld (vgl. Google) und eine Eingabemaske mit verschiedenen Feldern (beispielsweise „Kategorie“, „Jahr“, etc.) , über die detaillierte Suchabfragen durchgeführt werden können (M).

Es handelt sich um eine fehlertolerante Volltextsuche, die Technologien wie Fuzzy Search, Sounds-Like Queries, Thesauri und Funktionen wie *„Meinen Sie vielleicht ...?“* einsetzt (O). Neben den eigentlichen Inhalten werden auch Metadaten wie Jahr, Autor oder Schlüsselwörter (etwa bei Bildern aus Zeitungsarchiven) durchsucht (O).

Ergebnisse können dabei aus den Datenbeständen der Enzyklopädie und des Zeitungsarchivs stammen, aus Annotationen, Diskussionsbeiträgen, etc. Diese werden geclustered angezeigt (O), wobei zwei Varianten möglich sind:

- Clustering rein nach Inhalten, das heißt, es werden Cluster verschiedener Themengebiete angelegt, und die Ergebnisse diesen Clustern zugeordnet (siehe beispielsweise [Autonomy 2005; Clusty 2005]); oder
- Clustering nach Typen von Informationen bzw. deren Quellen, das heißt, es werden beispielsweise alle Bilder oder alle Dokumente aus dem Zeitungsarchiv zu einem Cluster zusammengefasst (vgl. [A9 2005]).

Cluster können visualisiert und wie in [Brockhaus 2006] in Form von Wissensnetzen dargestellt werden (N).

B.4.1.2 Durch Artikel und Kategorien browsen (M)

Actor: Alle Benutzer.

Interaction: Öffnen von Kategorien, Verwenden von Links, etc.

Goal: Auffinden der gewünschten Information.

Scenarios: (1) Ein Benutzer verwendet eine Zeitleiste, um durch einen Überblick über das 20. Jahrhundert zu bekommen und stößt auf diese Weise auf einen Artikel, den er anzeigen lässt. (2) Ein Benutzer verschafft sich einen Überblick über die verschiedenen Themengebiete, indem er die Hauptkategorien der Artikel anzeigen lässt. Durch Auswahl einer Unterkategorie gelangt der Benutzer zu einer Auswahl von Artikeln, aus der er einen bestimmten anzeigen lässt. (3) Ein Benutzer lässt sich einen Artikel anzeigen. Durch Verwendung des eingeblendeten Trails gelangt der Benutzer zu einem weiterführenden Artikel. (4) Ein Benutzer lässt sich einen Artikel anzeigen. Durch anklicken eines Hyperlinks gelangt er zu einem weiterführenden Artikel.

Ähnlich wie in der Brockhaus Ezyklopädie Digital ist es möglich, durch Artikel zu browsen. Benutzer können dazu sämtliche Hauptkategorien (Themengebiete) anzeigen lassen. Durch Auswählen einer Hauptkategorie werden alle enthaltenen Unterkategorien dargestellt.

Durch Auswählen einer Unterkategorie werden die darin enthaltenen Artikel angezeigt. Auf diese Weise kann der Benutzer Artikel anzeigen lassen.

Benutzern steht auch eine Zeitleiste zur Verfügung, in der Themengebiete und Artikel in chronologischer Reihenfolge aufgelistet werden (N). Über die Zeitleiste kann ein direkter Zugriff auf Artikel erfolgen.

Weiters kann ein Benutzer den Links in Artikeln, den vorgeschlagenen Artikeln in eingeblendeten dynamischen Trails *(„Benutzer, die diesen Artikel gelesen haben, haben sich auch für … interessiert“,* N) und statischen Trails („geführte Touren“, O) oder den Begriffen im Wissensnetz folgen und damit zu anderen Artikeln oder Themengebieten gelangen.

B.4.1.3 Artikel anzeigen (M)

Actor: Alle Benutzer.

Interaction: Auswählen von Artikeln.

Goal: Darstellen der gewünschten Information.

Scenario: Ein Benutzer führt eine Suche durch und erhält eine Anzahl von Ergebnisse. Der Benutzer wählt ein Suchergebnis aus, und es wird angezeigt. Der Benutzer gibt an, dass nur Informationen von Redakteuren und zertifizierten Experten angezeigt werden; andere Beiträge werden ausgeblendet. Zusätzlich lässt der Benutzer zu jedem Artikel das entsprechende Wissensnetz anzeigen.

Nach einer Suche oder durch Browsen werden gewünschte Artikel angezeigt. Dargestellt wird der Artikel selbst und zusätzlich (textuell) Cluster (O), die inhaltlich ähnliche Artikel in den internen Datenbeständen anbieten. Der Benutzer kann zu jedem Artikel das entsprechende Wissensnetz anzeigen lassen (N).

Der Benutzer kann angeben, welche Teile des Artikel angezeigt werden sollen: beispielsweise nur redaktionelle Inhalte oder auch Inhalte, die von CE, DE oder PU verfasst worden sind. Weiters kann ausgewählt werden, ob Annotationen, Links zu externen Referenzen oder Diskussionsbeiträge angezeigt werden.

B.4.1.4 Artikel, Annotationen, etc. verfassen (M/N)

Actor: Alle Benutzer.

Interaction: Erstellen eines neuen Beitrages.

Goal: Neue Inhalte im System verfügbar machen.

Scenarios: (1) Ein Benutzer sucht einen Artikel und kann ihn im System nicht finden. Daraufhin schreibt er selbst einen neuen Artikel zu diesem Thema und ordnet ihn einer Kategorie zu. (2) Ein Benutzer schreibt eine themenübergreifende Zusammenfassung. Anstatt Inhalte zu kopieren, verwendet er Transclusions, um Teile von bereits existierenden Artikeln zu übernehmen.

Benutzer haben die Möglichkeit, neue Inhalte in das System einzubringen. Benutzer können selbst auf sehr einfache Weise – wie etwa in einem herkömmlichen Textverarbeitungsprogramm (N) – neue Artikel verfassen und Hyperlinks, Bilder und andere Medien einfügen. Der Benutzer kann auch mittels Transclusions neue Artikel erstellen (O), das heißt, es werden im neuen Artikel Daten aus bestehenden Artikeln „zitiert“.

Jeder neue Artikel muss zumindest *einer* möglichst genauen Kategorie zugeordnet werden. Diese Zuordnung zu einer Kategorie kann später geändert werden.

Wenn ein Artikel von einem ED oder CE verfasst wird, wird er automatisch publiziert. Das heißt, der Artikel ist im System sichtbar und verfügbar, sobald er abgespeichert wird. Wird der Artikel von einem DE oder PU erstellt, dann wird er an die CE in dieser Domäne weitergeleitet. Der Artikel wird auf Korrektheit und Einhaltung der Qualitätsstandards und anderer Richtlinien wie etwa Blacklisting geprüft und gegebenenfalls freigegeben (publiziert).

ED und CE haben zusätzlich die Möglichkeit, Artikel zu zertifizieren, um deren Authentizität zu unterstreichen (siehe unten).

B.4.1.5 Artikel ändern (M)

Actor: Alle Benutzer.

Interaction: Ändern eines bestehenden Artikels.

Goal: Erweitern, korrigieren oder verbessern von bestehenden Inhalten.

Scenario: Ein von einem Benutzer verfasster Artikel enthält inkorrekte Daten. Der Benutzer wird von einem zertifizierten Experten darauf aufmerksam gemacht und ändert daraufhin den Artikel ab.

Artikel können nur vom jeweiligen Autor selbst geändert werden. ED und CE können unpassende Artikel von DE oder PU beispielsweise nicht selbst ändern. Sie können den Autor im Rahmen einer Diskussion dazu bewegen, inkorrekte oder unangebrachte Passagen zu überarbeiten. Sobald ein Konsens erreicht wird, kann der geänderte Artikel publiziert und gegebenenfalls auch zertifiziert werden (siehe unten). Wird kein Konsens erreicht, wird der Artikel von ED und CE nicht publiziert oder dessen Publikation rückgängig gemacht.

Unter der Voraussetzung, dass ein Versionskontrollsystem verwendet wird (siehe unten), kann ein eindeutiger Artikel nicht geändert werden (N). Soll ein Artikel geändert werden, so wird eine neue, geänderte Version des Artikels angelegt, und der ursprüngliche Artikel bleibt unverändert im System erhalten. Diese Vorgehensweise ist beispielsweise notwendig, um Funktionen wie Transclusions aber auch die Konsistenz von Bookmarks umzusetzen.

B.4.1.6 Artikel entfernen (M)

Actor: Alle Benutzer.

Interaction: Publikation eines bestehenden Artikels widerrufen.

Goal: Entfernen eines Artikels aus dem System.

Scenario: Ein Artikel enthält unpassende Informationen und soll deshalb aus dem System entfernt werden. Der Autor, CE und ED können die Publikation des Artikel widerrufen.

Es ist nicht möglich, Artikel aus dem System zu entfernen (siehe Abschnitt „Versionskontrolle“ weiter unten). Es kann lediglich die Publikation eines Artikels vom Autor, von ED oder CE zurückgezogen werden, wodurch er weder über die Suchfunktion noch über Browsen oder das Folgen von Links abgerufen werden kann.

B.4.1.7 Artikel zertifizieren (N)

Actor: Redakteure und zertifizierte Experten.

Interaction: Einen bestehenden Artikel zertifizieren.

Goal: Die Authentizität eines verfassten Artikels bestätigen und damit die inhaltliche Verantwortung übernehmen.

Scenario: Ein Redakteur liest einen Artikel, der von einem Benutzer auf einem sehr hohen Niveau verfasst worden ist. Um die Korrektheit der Information zu bestätigen, die Qualität des Artikels zu unterstreichen und den Artikel aufzuwerten, kann er den Artikel zertifizieren.

ED und CE haben die Möglichkeit, von anderen Benutzern erstellte Artikel zu zertifizieren. Damit wird die Authentizität des Artikels bestätigt, und der ED bzw. CE übernimmt die inhaltliche Verantwortung für einen Artikel. Artikel können mehrfach zertifiziert werden, was die Wahrscheinlichkeit für dessen Korrektheit erhöht.

Eine (mehrfache) Zertifizierung eines Artikels hat auch eine Auswirkung auf dessen Reihung innerhalb von Suchergebnissen (Ranking; N).

B.4.1.8 Artikel bewerten (M)

Actor: Alle Benutzer.

Interaction: Vergeben einer Bewertung für einen Artikel oder eine andere Information im System.

Goal: Die Qualität von Beiträgen bewerten und dadurch indirekt den jeweiligen Benutzer bewerten.

Scenario: Ein Benutzer liest einen von einem anderen Benutzer verfassten Diskussionsbeitrag. Da der Beitrag von schlechter Qualität ist, vergibt der Benutzer für diesen Beitrag eine negative Bewertung. Diese Bewertung hat einen Einfluss auf die Bewertung des Benutzers.

Jeder Benutzer im System kann eine Bewertung für Beiträge anderer Benutzer abgeben. Bewertungen können für jegliche Art von Inhalten abgegeben werden: für Artikel, Annotationen, Links, Diskussionsbeiträge, etc. Ähnlich [eBay 2005] fügt der Benutzer zusätzlich zur Bewertung (positiv, negativ oder neutral) einen kurzen Kommentar mit einer Begründung für die jeweilige Bewertung an.

Bewertungen für eine bestimmte Version eines Artikels können nicht geändert werden.

Wird ein Artikel, der beispielsweise durch inkorrekte Angaben eine schlechte Bewertung erhalten hat, geändert, so wird folgendermaßen vorgegangen:

- der ursprüngliche Artikel wird archiviert;
- die Bewertung des ursprünglichen Artikels wird archiviert, bleibt aber mit dem ursprünglichen Artikel assoziiert;
- die aktualisierte Version des Artikels wird publiziert und hat noch keine Bewertungen;
- sämtliche Benutzer, die eine Bewertung zum ursprünglichen Artikel abgegeben haben, werden informiert, dass eine neuere Version des Artikels verfügbar ist, und werden aufgefordert, eine neue Bewertung abzugeben.

Die Bewertungen der Artikel werden ihren Autoren angerechnet. Eine große Anzahl positiver Bewertungen von Artikeln ergibt eine positive Bewertung für den Autor. Die Gesamtbewertung des Benutzers ist ähnlich [eBay 2005] für alle anderen Benutzer im System sichtbar.

Die Bewertungen von Artikeln haben auch eine Auswirkung auf deren Reihung in Suchergebnissen (Ranking), etc.

Da eine genauere Beschreibung des Bewertungssystem für dieses Dokument zu umfangreich wäre, wird an dieser Stelle nur auf Kapitel 2 dieser Arbeit verwiesen.

B.4.1.9 Artikel bookmarken (M)

Actor: Alle Benutzer.

Interaction: Setzen eines Lesezeichens für eine im System gefundene Information.

Goal: Im System gefundene Information durch setzen eines Lesezeichens wiederauffindbar machen.

Scenarios: Der Benutzer liest einen Artikel, der auch zu einem späteren Zeitpunkt von Interesse sein könnte. Der Benutzer setzt deshalb ein Lesezeichen auf den Artikel. (1) Das Lesezeichen wird im System in einem privaten Bereich des Benutzers gespeichert und damit kann jederzeit wieder abgerufen werden. (2) Das Lesezeichen wird vom Web-Browser abgespeichert und kann damit jederzeit wieder abgerufen werden.

Benutzer können für jeden Artikel im System und für jedes andere inhaltsbezogene Objekt wie beispielsweise eine Annotation oder einen Beitrag in einem Diskussionsforum Bookmarks anlegen. Bookmarks können sowohl systemintern – etwa in einem privaten Arbeitsbereich – oder systemextern im Web-Browser gespeichert werden.

Vom Benutzer angelegte Lesezeichen sind persistent.

B.4.1.10 Fragen an Artikel stellen (Active Documents) (M)

Actor: Alle Benutzer.

Interaction: Eine Frage an einen Artikel stellen.

Goal: Der Benutzer stellt eine Frage an einen Artikel, die von der Active Documents Komponente beantwortet wird.

Scenarios: Ein Benutzer stellt eine beliebige Frage an einen Artikel oder an ein anderes inhaltsbezogenes Objekt im System. (1) Wenn bereits eine äquivalente Frage gestellt wurde und dazu eine Antwort verfügbar ist, wird die Antwort angezeigt. (2) Wenn keine äquivalente Frage verfügbar ist, wird sie an einen Experten weitergeleitet, der innerhalb einer gewissen Zeit die Frage beantwortet.

Benutzer können beliebige Fragen an Dokumente stellen. Diese Fragen werden von einer Active Documents Komponente bearbeitet. Zunächst wird versucht, eine äquivalente Frage im System zu finden. Die Äquivalenz kann mittels syntaktischer Analyse (N), Heuristik (M) oder semantischer Analyse (O) festgestellt werden. Ist eine äquivalente, bereits beantwortete Frage im System bereits gespeichert, dann wird die entsprechende Frage angezeigt. Andernfalls wird die Frage an einen Experten (zuerst DE, dann CE, später ED) zur Bearbeitung weitergeleitet (siehe Abschnitt "Antwort durch einen Experten" unten). (Für Details siehe auch [Heinrich and Maurer 2000].)

Fragen, die an einen Artikel gestellt worden sind, werden innerhalb dieses Artikels als Symbol dargestellt. Dadurch ist es für andere Benutzer ersichtlich, dass bereits Fragen an den Artikel gestellt wurden. Wird eine große Anzahl von Fragen an einen Artikel gestellt, so muss der Artikel möglicherweise überarbeitet werden, oder Benutzer haben großes Interesse an diesem Themenbereich (siehe Abschnitt „Qualitäts-Feedback und Review Mechanismus" weiter unten).

B.4.1.11 Artikel Kategorien zuordnen (M)

Actor: Redakteure und zertifizierte Benutzer.
Interaction: Einen Artikel einer oder mehreren Kategorien zuordnen.
Goal: Die thematische Zuordnung eines Artikels treffen oder ändern.
Scenarios: (1) Ein neuer Artikel wird verfasst und vom Autor wird eine Kategorie dafür vorgeschlagen. Ein zertifizierter Experte überprüft den neuen Artikel und weist ihn der vorgeschlagenen Kategorie zu. (2) Ein bestehender Artikel ist einer Kategorie zugewiesen. Nachdem einen neue Kategorie eingeführt wird, die ebenso passend für den Artikel wäre, wird dieser von einem zertifizierten Experten einer zweiten Kategorie zugewiesen.

ED und CE können Artikel zu neuen Kategorien zuordnen sowie bestehende Zuordnungen ändern oder aufheben. DE und PU haben lediglich die Möglichkeit, Zuordnungen vorzuschlagen.

Eine Änderung der Zuordnung eines Artikels zu Kategorien hat keinen Einfluss auf Verweise auf diesen Artikel, auf angehängte Annotation, etc. oder auf Bookmarks.

B.4.1.12 Kategorien anlegen (M)

Actor: Redakteure.
Interaction: Neue Kategorie anlegen.
Goal: Hinzufügen einer Kategorie oder Unterkategorie.
Scenario: Ein Redakteur legt eine neue Kategorie an, um die Erschließung eines neuen Themengebietes zu ermöglichen.

Neue Kategorien und Unterkategorien können von ED angelegt werden. Andere Benutzer des Systems – einschließlich CE – können neue Kategorien und Unterkategorien nur vorschlagen, aber nicht selbst anlegen.

B.4.1.13 Kategorien umbenennen (M)

Actor: Redakteure.
Interaction: Umbenennen einer Kategorie.
Goal: Die Bezeichnung einer Kategorie ändern.
Scenario: Die Bezeichnung einer Kategorie ist nicht mehr passend. Der Redakteur ändert die Bezeichnung.

Die Bezeichnungen von Kategorien können von ED beliebig geändert werden. Von allen anderen Benutzern im System kann das Umbenennen einer Kategorie nur vorgeschlagen werden.

Das Umbenennen einer Kategorie hat keine Auswirkung auf in der Kategorie enthaltene Unterkategorien und Artikel. Auch Annotationen, Diskussionen, etc., die an die Kategorie angehängt sind, sind von der Änderung der Bezeichnung nicht betroffen.

B.4.1.14 Kategorien zusammenlegen (M)

Actor: Redakteure.
Interaction: Auswählen von zwei Kategorien oder Unterkategorien und Zusammenlegen derselben.
Goal: Zwei bestehende Kategorien oder Unterkategorien zu einer einzigen zusammenfügen.
Scenario: Zwei Unterkategorien ähneln sich sehr stark und haben nur eine kleine Anzahl von Artikeln. Ein Redakteur legt die beiden Unterkategorien zusammen, wodurch eine größere Unterkategorie mit einer größeren Zahl von Artikeln entsteht.

ED haben die Möglichkeit, jeweils zwei Kategorien zusammenzulegen. Dabei kann eine Kategorie als „dominant" angesehen werden: sie bleibt unverändert erhalten, während sämtliche Unterkategorien und Artikel aus der anderen Kategorie der dominanten zugeordnet werden. Um die Konsistenz von Bookmarks und anderen Verweisen nicht zu gefährden, wird die nun leere Kategorie nicht gelöscht sondern vom System als symbolischer Link zur dominanten Kategorie behandelt.

B.4.1.15 Kategorien aufteilen (N)

Actor: Redakteure.
Interaction: Artikel innerhalb einer Kategorie auswählen und die Titel der neuen Kategorien angeben.
Goal: Aufteilen einer bestehenden Kategorie in zwei neue.
Scenario: Eine Unterkategorie enthält sehr viele Artikel. Ein Redakteur wählt eine Anzahl von Artikeln aus, gibt die neuen Namen für die beiden Kategorien an. Alle ausgewählten Artkel werden in die eine Kategorie verschoben, alle anderen Artikel in die andere, neue Kategorie.

Enthält eine Kategorie zuviele Artikel, kann es notwendig werden, diese aufzuteilen. ED haben die Möglichkeit, eine bestehende Kategorie in zwei neue Kategorien aufzuteilen und die ursprünglichen Artikel diesen beiden Kategorien zuzuordnen.

Sämtliche Verweise auf die ursprüngliche Kategorie werden auf die neu angelegten Kategorien umgeleitet (O). Das Aufteilen von Kategorien hat keine Auswirkungen auf Bookmarks auf darin enthaltene Artikel oder Unterkategorien.

B.4.1.16 Kategorien löschen (M)

Actor: Redakteure.
Interaction: Löschen einer Kategorie oder Unterkategorie.
Goal: Entfernen einer Kategorie oder Unterkategorie.
Scenario: Eine Unterkategorie enthält sehr wenige oder keine Artikel. Ein Redakteur verschiebt die etwaig verbleibenden Artikel in eine andere, passende Unterkategorie oder die übergeordnete Kategorie und entfernt die leere Unterkategorie.

Wird eine Kategorie entfernt, werden zuerst die darin enthaltenen Unterkategorien und Artikel sowie die angehängten Annotationen, etc. in eine andere, passende Kategorie oder Unterkategorie verschoben bzw. angehängt. Ist die Kategorie leer, dann kann die Kategorie entfernt werden.

Sämtliche Verweise auf die entfernte Kategorie werden auf eine entsprechende andere Kategorie „umgeleitet".

B.4.1.17 Kategorien verschieben (M)

Actor: Redakteure.
Interaction: Verschieben einer Kategorie oder Unterkategorie.
Goal: Reorganisation der Themenhierarchie.
Scenario: Die Kategorie „Nanotechnologie" wird als Unterkategorie des Bereiches „Physik" geführt. Da diese Unterkategorie sehr rasch wächst, bereits sehr viele Artikel enthält und eine Verfeinerung der Unterkategorie wünscheswert wäre, wird sie um eine Ebene „nach oben" verschoben und damit zu einer Hauptkategorie gemacht.

ED haben die Möglichkeit, Kategorien zu verschieben. Es können beispielsweise Unterkategorien zu (Haupt-)Kategorien gemacht werden, Unterkategorien in andere (Haupt-)Kategorien verschoben werden, oder (Haupt-)Kategorien zu Unterkategorien gemacht werden. Es muss darauf geachtet werden, dass die vom System vorgegebene maximale Verschachtelungstiefe von Kategorien eingehalten wird.

B.4.2 Diskussionen (M)

Das System stellt dem Benutzer Diskussionsforen nach dem Prinzip von Newsgruppen bzw. Web-Diskussionsforen zur Verfügung. Ein Diskussionforum ist dabei eindeutig einer Informationseinheit bzw. einem Objekt im System zugeordnet. So kann zum Beispiel ein Diskussionsforum einem Artikel oder auch einer Kategorie zugeordnet sein.

B.4.2.1 Beiträge lesen (M)

Actor: Alle Benutzer.
Interaction: Auswählen einer Diskussion.
Goal: Lesen der Diskussionsbeiträge zur ausgewählten Diskussion.
Scenario: Ein Benutzer hat einen Artikel gelesen und möchte nun die zu diesem Artikel bereits verfassten Diskussionsbeiträge einsehen.

Das einem Objekt zugeordnete Diskussionsforum wird durch ein entsprechendes Symbol visualisiert und ist dadurch anwählbar. Durch die Auswahl wird eine konfigurierbare Standardsicht auf die Beiträge dargestellt. Eine Sicht stellt die Beiträge sequentiell in chronologischer Reihenfolge dar, wobei der volle Inhalt aller Beiträge angezeigt wird. Eine alternative Visualisierung stellt einerseits alle Beiträge in einer Baumstruktur und andererseits einen aktuell angewählten Beitrag in voller Detailstufe dar.

B.4.2.2 Beiträge verfassenen (M)

Actor: Alle Benutzer.
Interaction: Schreiben eines Beitrages.
Goal: Der Benutzer fügt zu einer Diskussion einen Beitrag hinzu.
Scenarios: (1) Ein Benutzer liest einen Beitrag und möchte seine Meinung dazu kundtun. Er eröffnet eine neue Diskussion, indem er einen ersten Diskussionsbeitrag zu diesem Artikel schreibt. (2) Ein Benutzer liest einen Diskussionsbeitrag zu einem Artikel. Er möchte sich mit dem Autor in Verbindung setzen und schreibt eine Antwort auf den Diskussionsbeitrag.

In Diskussionsforen wird eine Funktion „Beitrag verfassen" angeboten. Bei Auswahl dieser Funktion wird ein Dialog für die Eingabe eines neuen Beitrages zur Verfügung gestellt – ähnlich wie zum Schreiben eines neuen Artikels.

B.4.2.3 Beiträge bewerten (M)

Actor: Alle Benutzer.
Interaction: Bewerten eines Beitrages.
Goal: Ein Benutzer bewertet einen Beitrag eines anderen Benutzers im Diskussionsforum.
Scenario: Der Benutzer liest einen sehr konstruktiven Beitrag in einem Diskussionsforum und gibt dazu eine Bewertung ab.

Zu jedem Beitrag in einem Diskussionsforum wird die Funktion „Bewertung abgeben" angeboten. Auf diese Weise kann der Benutzer den Diskussionsbeitrag eines anderen Benutzers bewerten.

B.4.2.4 Beiträge löschen (M)

Actor: Alle Benutzer.
Interaction: Zurückziehen der Publikation eines Beitrages aus einem Diskussionsforum.
Goal: Einen Beitrag aus einem Diskussionsforum entfernen.
Scenarios: (1) Ein Benutzer möchte einen von ihm verfassten Beitrag löschen. (2) Ein zertifizierter Experte möchte einen unpassenden Diskussionsbeitrag eines PU löschen.

Diskussionsbeiträge können nicht aus dem System entfernt werden. Ähnlich wie für Artikel wird für einen zu entfernenden Beitrag lediglich die Publikation zurückgezogen (siehe auch weiter oben).

Jeder Benutzer hat die Möglichkeit, die von ihm verfassten Diskussionsbeiträge zu „entfernen", wenn noch keine Antworten darauf geschrieben wurden. ED und CE haben auch die Möglichkeit, Beiträge von DE oder PU zu entfernen, wenn diese unpassend sind.

B.4.3 Active Knowledge Brokering (M)

In den folgenden Abschnitten werden die einzelnen Teile des Active Knowledge Brokering Systems dargestellt. Obwohl die einzelnen Punkte in diesem Zusammenhang stets in Kombination verwendet werden und nur gemeinsam ein sinnvolles System ergeben, werden sie zur Verdeutlichung separat behandelt.

Ein Benutzer wählt eine Kategorie (und eine Unterkategorie) oder ein Dokument aus und stellt eine Frage in natürlicher Sprache. Daraufhin werden die internen Datenbestände nach passenden Artikeln durchsucht. Zusätzlich wird eine Active Documents Komponente verwendet, um nach ähnlichen, bereits beantworteten Fragen zu suchen. Verfügbare Fragen werden dem Benutzer präsentiert. Sind keine Antworten vorhanden oder passend, dann wird die Frage an einen Experten weitergeleitet. (Siehe auch Kapitel 2.)

B.4.3.1 Fragen stellen (M)

Actor: Alle Benutzer.
Interaction: Frage an das System stellen.
Goal: Der Benutzer formuliert eine Frage.
Scenario: Ein Benutzer wählt eine Kategorie und eine Unterkategorie aus. In natürlicher Sprache formuliert der Benutzer eine Frage und schickt Sie ab.

Ein Benutzer hat die Möglichkeit, beliebige Fragen in natürlicher Sprache an das System zu stellen. Dazu muss der Benutzer zuerst eine genaue Kategorie und Unterkategorie auswählen. In einer bewusst schlicht gehaltenen Eingabemaske (vgl. Google) kann der Benutzer eine Frage eingeben.

B.4.3.2 Vorgeschlagener Content aus dem System (M)

Actor: System.
Interaction: Frage in natürlicher Sprache an die Suchkomponente übergeben.
Goal: Ergebnisse aus den internen Datenbeständen werden geliefert.
Scenario: Ein Benutzer hat in natürlicher Sprache eine Frage formuliert. Mittels der Frage wird eine Abfrage in den internen Datenbeständen durchgeführt. Sämtliche Ergebnisse werden dem Benutzer in Form von Clustern dargestellt.

Die vom Benutzer gestellte Frage löst eine Abfrage in den internen Datenbeständen aus (vgl. [AskJeeves 2005]). Artikel und andere Beiträge aus den verschiedenen Datenquellen werden kombiniert und geclustered (N). Die Cluster werden dem Benutzer als „Artikel aus den Archiven" präsentiert.

B.4.3.3 Bereits vorhandene Antworten (Active Documents) (M)

Actor: System.
Interaction: Anfrage an die Active Documents Komponente des Systems übergeben.
Goal: Abfragen, ob bereits eine ähnliche Frage gestellt wurde und gegebenenfalls die Antwort anzeigen.
Scenario: Ein Benutzer hat in natürlicher Sprache eine Frage formuliert. Die Active Documents Komponente überprüft, ob bereits ähnliche Fragen gestellt wurden. Es werden alle ähnlichen Fragen mit den entsprechenden Antworten angezeigt.

Zusätzlich zur Abfrage der internen Datenbestände wird eine Active Documents Anfrage an das System gestellt. Das heißt, es wird mittels syntaktischer Analyse (N), Heuristik (M) oder semantischer Analyse (O) versucht, eine ähnliche Frage innerhalb der angegebenen Kategorie oder Unterkategorie zu finden.

Sämtliche passenden Fragen werden dem Benutzer zusammen mit den entsprechenden Antworten vom System mit der Frage *„Meinten Sie vielleicht ...?"* angeboten.

B.4.3.4 Weiterleiten der Frage an Experten (M)

Actor: System.
Interaction: Frage eines Benutzers an Experten weiterleiten.
Goal: Die vom Benutzer gestellte Frage wird an Experten zur Beantwortung weitergeleitet.
Scenario: Ein Benutzer hat eine Frage in natürlicher Sprache formuliert, und die Active Documents Komponente hat keine passenden Antworten finden können. Die ursprüngliche Frage wird an einen Experten weitergeleitet.

Wenn keine der Frage-Antwort-Paare, die dem Benutzer von der Active Documents Komponente angeboten werden, passend sind, wird die Frage an einen menschlichen Experten im System weitergeleitet.

Eine Frage wird zuerst an alle DE der entsprechenden Unterkategorie weitergeleitet. Wird innerhalb einer bestimmten Zeit (beispielsweise innerhalb eines Tag-

es) keine Antwort geliefert, wird die Frage automatisch an die Gruppe der CE und schließlich an ED der entsprechenden Kategorie oder Unterkategorie verschickt.

B.4.3.5 Antwort durch einen Experten (M)

Actor: Domänen-Experten, zertifizierte Experten und Redakteure.

Interaction: Beantworten von von Benutzern gestellten Fragen.

Goal: Beantworten von Fragen, die von Benutzern gestellt worden sind.

Scenario: Ein Benutzer hat eine Frage in natürlicher Sprache gestellt, und die Active Documents Komponente hat keine passenden Antworten finden können. Die ursprüngliche Frage wird an einen Experten weitergeleitet, der eine Antwort zur Verfügung stellt.

Wird die Frage eines Benutzers an Experten weitergeleitet, so werden diese von der neu gestellten Frage verständigt. Sobald ein Experte um die Beantwortung einer Frage in Angriff nimmt, wird die Frage für alle anderen Experten gesperrt.

Ist die Beantwortung der Frage abgeschlossen, wird

- die Frage gemeinsam mit der entsprechenden Antwort im Archiv abgelegt;
- dem Benutzer die Antwort auf die Frage zugestellt;
- die Aufforderung zur Beantwortung der Frage zurückgezogen.

Andere Experten haben jedoch die Möglichkeit, die erste Antwort auf die Frage zu ergänzen.

Eine Antwort hat keine bestimmten Formvorschriften. Das Erstellen einer Antwort erfolgt analog zum Verfassen eines neuen Artikels (siehe oben).

B.4.4 Kommunikation (M)

Das System bietet verschiedenen Möglichkeiten, um mit anderen Benutzern des Systems zu kommunizieren. Dadurch ist gesichert, dass diese nicht auf externe Applikationen angewiesen sind.

Diese Kommunikationsfunktionen sind nicht an Artikel geknüpft sondern unabhängig davon verwendbar.

B.4.4.1 Synchrone Kommunikation mit anderen Benutzern (M)

Actor: Alle Benutzer.

Interaction: An einer synchronen Kommunikation (Chat) teilnehmen.

Goal: Benutzer kommunizieren synchron miteinander innerhalb des Systems.

Scenario: Ein Benutzer chattet mit einem anderen Benutzer.

Das System bietet dem Benutzer die Möglichkeit, mit einem anderen Benutzer synchron zu kommunizieren, wie es beispielsweise aus Internet Chats bekannt ist.

B.4.4.2 Asynchrone Kommunikation mit anderen Benutzern (M)

Actor: Alle Benutzer.

Interaction: Nachricht versenden.

Goal: Einem Benutzer des Systems eine Nachricht zukommen lassen.

Scenario: Ein Benutzer lässt einem anderen Benutzer eine Nachricht innerhalb des Systems zukommen.

Ein Benutzer kann einem anderen Benutzer eine Nachricht schicken. Diese Funktionalität ist vergleichbar mit dem Verschicken von e-Mails. Jeder Benutzer verfügt über ein Postfach („persönlicher Nachrichtenordner"), in dem die an ihn adressierten Nachrichten abgelegt werden.

B.4.4.3 Externe Kommunikation (N)

Actor: Alle Benutzer.

Interaction: Versenden von Inhalten.

Goal: Informationen an eine Person außerhalb des Systems versenden.

Scenarios: (1) Ein Benutzer liest einen interessanten Artikel (oder beliebige andere „versendbare" Information) und verschickt diese an eine Person außerhalb des Systems. (2) Ein Benutzer möchte Nachrichten an Personen außerhalb des Systems versenden.

Das System bietet eine Funktion zum verschicken von e-Mails an. Auf diese Weise können Benutzer des Systems beispielsweise auch Artikel an Personen außerhalb des Systems verschicken oder beliebige e-Mails verfassen, ohne eine andere externe Applikation verwenden zu müssen. (Anmerkung: Urheberrechtliche Aspekte müssen in diesem Zusammenhang geklärt werden.)

B.4.4.4 Personalisierter Benachrichtigungsdienst (N)

Actor: Alle Benutzer.

Interaction: Benachrichtigungen abonnieren.

Goal: Abonnierte Benachrichtigungen in gewählter Form Benutzern zur Verfügung stellen.

Scenarios: (1) Der Benutzer bekommt Benachrichtigungen über Annotationen und Diskussionen zu von ihm verfassten Artikeln per e-Mail zugesandt (Push). (2) Der Benutzer bekommt Benachrichtigungen über Annotationen und Diskussionen zu von ihm verfassten Artikeln in den persönlicher Nachrichtenordner im System zugestellt (Push). (3) Der Benutzer fragt Informationen über neue Annotationen und Diskussionen zu von ihm verfassten Artikeln mittels Syndication-Diensten wie RSS ab (Pull).

Zur Illustration der Idee werden zunächst einige Beispiele für abonnierbare Benachrichtigungen gegeben:

- Es wurde ein neuer Artikel zu einer entsprechend markierten Kategorie hinzugefügt.
- Artikel zu Themen aus den Nachrichten werden aktualisiert.

- Ein vom Benutzer erstellter Artikel, Beitrag im Forum, etc. wird bewertet.
- Ein vom Benutzer erstellter Artikel wird kommentiert.
- Eine Antwort auf einen vom Benutzer erstellten Diskussionsbeitrag wird veröffentlicht.

Diese Änderungen und neuen Inhalte werden automatisiert gesammelt und dem Benutzer in der gewählten Form zur Verfügung gestellt. Wird beispielsweise eine Antwort auf einen Diskussionsbeitrag geschrieben, so wird der Autor des ursprünglichen Beitrages davon verständigt.

Es werden hierbei sowohl „Pull"- als auch „Push"-Technologien verwendet. Informationen werden also entweder automatisch per e-Mail verschickt oder in den persönlichen Nachrichtenordner des Benutzers innerhalb des Benutzers abgelegt. Alternativ kann der Benutzer auch selbst Informationen explizit aus dem System abrufen. Dazu werden Syndication-Technolgien wie RSS verwendet (siehe [RSS 1999; RSS 2001] und [RDF 2004]).

B.4.5 Personalisierung (N/O)

Das System ist adaptiv seinen Benutzern gegenüber. Das heißt, das System passt sich dem Verhalten bzw. den Wünschen seiner Benutzer an. Erreicht wird dies durch verschiedene Arten von Profilen:

- globale Profile werden vom Benutzer bei der ersten Verwendung des System ausgefüllt (N);
- ad hoc Profile werden vom Benutzer etwa vor dem Ausführen einer Suchabfrage ausgefüllt (N);
- dynamische Profile werden vom System selbständig aus dem Verhalten der Benutzer erstellt (O).

Die folgenden Unterpunkte beschreiben eine Auswahl von Funktionen, die durch Personalisierung ermöglicht werden. Eine Reihe weiterer Funktionalitäten ist denkbar.

B.4.5.1 Profile ausfüllen (N)

Actor: Alle Benutzer.
Interaction: Der Benutzer liefert ein Profil über seine Person, Interessen und Kenntnisse ab.
Goal: Sammeln von Benutzerdaten, um eine Adaptierung und Personalisierung der angebotenen Information zu erreichen.
Scenarios: (1) Der Benutzer verwendet zum ersten Mal das System und wird aufgefordert, ein allgemeines Profil über seine Interessen und Kenntnisse auszufüllen. (2) Der Benutzer will eine Suchabfrage im System durchführen. Er füllt ein ad hoc Profil aus und erhält damit personalisierte Suchergebnisse.

Benutzer füllen ein Profil aus, um dem System Informationen über Interessen und Kenntnisse zu liefern, die für Adaptierungen verwendet werden können. Das System implementiert in einer ersten Version lediglich globale Profile und ad hoc Profile (siehe oben).

B.4.5.2 Suchergebnisse (N)

Actor: System.
Interaction: Anpassen von Suchergebnissen.
Goal: Anpassen der Suchergebnisse an die Bedürfnisse und Wünsche des Benutzers.
Scenario: Ein Benutzer führt eine Suchabfrage durch. Die Ergebnisse werden entsprechend des ausgefüllten Profiles adaptiert.

Das System verwendet die Informationen, die der Benutzer in Form eines Profiles angegeben hat, um die Auswahl, das Ranking und die Präsentation der Suchergebnisse anzupassen. Ein besonderes Interesse in einer bestimmten Domäne hat beispielsweise eine höhere Reihung von Suchergebnissen aus dieser Kategorie zur Folge.

Der Benutzer kann jederzeit die Adaptierung aufheben, um damit eine „Standard-Anzeige" der Suchergebnisse zu erreichen.

B.4.5.3 Anzeige von Artikeln (O)

Actor: System.
Interaction: Anpassen der Darstellung von Artikeln.
Goal: Anpassen der Darstellung von Artikeln an die Bedürfnisse und Wünsche des Benutzers.
Scenario: Ein Benutzer lässt einen Artikel aus dem System anzeigen. Obwohl der Artikel Annotationen von ED und PU enthält, werden diese nicht angezeigt, weil dies nicht dem Profil des Benutzers entspricht (er möchte nur Annotationen von ED und CE sehen).

Der Benutzer kann auswählen, auf welche Weise Artikel im System dargestellt werden, und welche Informationen angezeigt werden. Ein Benutzer kann beispielsweise nur von Beiträge von ED und CE anzeigen lassen, während Inhalte von DE und PU ausgeblendet werden. Es können beispielsweise auch lediglich Artikel angezeigt werden; Diskussionsbeiträge und Annotationen werden ausgeblendet.

Der Benutzer kann jederzeit die Adaptierung aufheben, um damit eine „Standard-Anzeige" der Suchergebnisse zu erreichen.

B.4.5.4 Artikel des Tages (N)

Actor: System.
Interaction: Finden und Anzeigen eines passenden Artikels.
Goal: Entsprechend den Interessen des Benutzers wird täglich ein Artikel gefunden, der dem Benutzer als Artikel des Tages präsentiert wird.
Scenario: Der Benutzer wählt sich in das System ein. Auf der Startseite wird entsprechend seines Profils ein Artikel des Tages präsentiert.

Das System verwendet das Profil des Benutzers, um in den internen Datenbeständen aus einer Kategorie, für die sich der Benutzer interessiert, einen Artikel abzurufen. Dieser täglich andere Artikel wird beispielsweise auf der Startseite des Benutzers angezeigt.

B.4.5.5 In den Nachrichten (O)

Actor: System.

Interaction: Finden eines interessanten Artikels in den Nachrichten.

Goal: Entsprechend den Interessen des Benutzers werden laufend neue Artikel in den Nachrichten gefunden und dem Benutzer präsentiert.

Scenario: Der Benutzer wählt sich in das System ein. Auf der Startseite wird entsprechend seines Profils ein Artikel aus den aktuellen Nachrichten präsentiert.

Ähnlich wie bei der Funktion „Artikel des Tages" sucht das System basierend auf den Interessen des Benutzers potentiell interessante Artikel in den Nachrichten. Diese Artikel werden beispielsweise auf der Startseite des Benutzers angezeigt und mit passenden Informationen aus den Datenbeständen des Systems ergänzt.

B.4.6 Workspaces (O)

Workspaces sind persönliche Arbeitsbereiche für Benutzer und Gruppen von Benutzern. Sie dienen als persönlicher Ablagebereich innerhalb des Systems bzw. als Möglichkeit, Informationen innerhalb einer Gruppe zu sammeln und zu verwalten.

In diesen Bereichen können beispielsweise Bookmarks angelegt und von dort aus aufgerufen werden, eigene Dokumente geschrieben, Ausschnitte aus bestehenden Artikeln abgelegt (als Kopie) werden, Transclusions von bestehenden Artikeln angelegt werden, etc.

B.4.6.1 Private Arbeitsbereiche (O)

Actor: Alle Benutzer.

Interaction: Verwenden des privaten Arbeitsbereiches.

Goal: Ablegen von Informationen und Produzieren neuer Information in persönlichen Arbeitsbereichen.

Scenarios: (1) Ein Benutzer findet einen Artikel. Er kopiert einen interessanten Teil des Artikels und legt ihn in seinem persönlichen Arbeitsbereich ab. (2) Ein Benutzer legt Bookmarks auf interessante Artikel in seinem persönlichen Arbeitsbereich ab. (3) Ein Benutzer sammelt in den internen Datenbeständen Informationen zu einem Thema. Diese legt er im persönlichen Arbeitsbereich ab und beginnt dort auch, einen neuen Artikel aus den gesammelten Informationen zu verfassen.

Ein privater Arbeitsbereich wird – wie oben erwähnt – verwendet, um Informationen in einem persönlichen Bereich abzulegen. Der Benutzer hat alleine Zugriff auf diesen Bereich und kann sämtliche darin enthaltenen Daten selbst verwalten.

Die Darstellung des Arbeitsbereiches und die genaue Funktionalität werden an dieser Stelle nicht näher beschrieben und sind im Rahmen einer technischen Spezifikation des Sytemes zu diskutieren.

B.4.6.2 Gemeinsame Arbeitsbereiche (O)

Actor: Alle Benutzer, Benutzergruppen.

Interaction: Verwenden eines Gruppenarbeitsbereiches.

Goal: Ablegen von Informationen und Produzieren neuer Information in Gruppen-Arbeitsbereichen.

Scenario: Benutzer arbeiten an einem gemeinsamen Projekt. Sie legen gesammelte Informationen im gemeinsamen Arbeitsbereich ab und erstellen in diesem Arbeitsbereich gemeinsam neue Dokumente und Artikel.

Zusätzlich zu den Funktionen und Möglichkeiten privater Arbeitsbereiche wird bei gemeinsamen Arbeitsbereichen eine Zugriffskontrolle eingeführt. So können Benutzer beispielsweise auch innerhalb eines gemeinsamen Arbeitsbereiches private Artikel ablegen, die von anderen Gruppenmitgliedern nicht gelesen oder geändert werden können.

B.4.6.3 Visualisierung und Spatial Hypertext (O)

Actor: System.

Interaction: Aufbereiten der Informationen in privaten und gemeinsamen Arbeitsbereichen.

Goal: Graphische Darstellung von Informationen in einem zweidimensionalen Bereich.

Scenario: Die von Benutzern gesammelten Informationen aus Workspaces werden in einem zweidimensionalen Bereich graphisch dargestellt. Der Benutzer kann die positionierten Objekte im zweidimensionalen Bereich beliebig verschieben und nach seinen Wünschen anordnen.

Das System verwendet Technologien aus dem Feld des Spatial Hypertexts, um Informationen aus Workspaces graphisch darzustellen. Benutzer können dadurch Elemente wie Bookmarks, selbst geschriebene Artikel oder Ausschnitte aus anderen Artikeln beliebig im Arbeitsbereich positionieren und somit beispielsweise sinngemäß verwandte Artikel auch optisch näher ablegen (siehe auch [Buchanan et al. 2004]).

B.4.7 Benutzerzertifizierung und -verwaltung (M/N)

An dieser Stelle werden nur für das System speziell entwickelte Techniken der Benutzerverwaltung dargestellt. Grundfunktionen, die aus anderen Systemen bekannt sind – etwa Login mit Benutzername und Kennwort, Logout oder das Ändern des eigenen Kennwortes – sollen in diesem Dokument nicht näher behandelt werden.

B.4.7.1 Benutzer befördern oder degradieren (M)

Actor: Redakteure und zertifizierte Experten.

Interaction: Befördern oder Degradieren eines Benutzers.

Goal: Benutzer abhängig von ihrer Qualifikation befördern oder degradieren.

Scenarios: (1) Ein Benutzer schreibt mehrere qualitativ hochwertige Artikel und bekommt dafür positive Bewertungen. Das System erkennt dies und verständigt einen zertifizierten Experten, dass dieser Benutzer möglicherweise eine Beförderung verdient. Nach einer Überprüfung durch einen zertifizierten Experten wird der Benutzer vom Status PU zum Status DE befördert. (2) Der Benutzer schreibt als DE einige Beiträge, die negative Bewertungen bekommen. Daraufhin informiert das System einen zertifizierten Experten, der den Sachver-

halt überprüft und den Status des Benutzers von DE auf PU zurücksetzt.

Zertifizierte Experten und Redakteure können den Benutzerstatus anderer Benutzer im System verändern. Benutzer, die qualitativ hochwertige Beiträge verfassen, können auf diese Weise zu DE oder CE befördert werden, während Benutzer, die minderwertige Beiträge schreiben, zu DE oder PU degradiert werden können.

Das System schlägt CE und ED automatisch Benutzer vor, die befördert oder degradiert werden sollten. Vorschläge des Systems sind abhängig von den Bewertungen der Artikel, die Benutzer geschrieben haben. Überwiegt beispielsweise die Anzahl von positiven Bewertungen für Artikel eines Autors mit großer Mehrheit, dann wird der Benutzer vom System für eine Beförderung vorgeschlagen.

B.4.7.2 Vorschlagen neuer zertifizierter Experten (M)

Actor: Alle Benutzer.

Interaction: Benutzer schlagen andere Benutzer vor, die zu zertifizierten Experten gemacht werden sollen.

Goal: Vorschlagen neuer zertifzierter Experten.

Scenarios: (1) Ein Domänenexperte hat sich innerhalb Gruppe von Benutzern verdient gemacht und wird deshalb von der Gruppe als zertifizierter Experte vorgeschlagen. (2) Ein Benutzer glaubt von sich selbst, als zertifizierter Experte geeignet zu sein. Er „bewirbt" sich als zertifizierter Experte, das heißt, er schlägt sich selbst vor.

Benutzer des Systemes haben die Möglichkeit, andere Benutzer oder sich selbst als zertifizierte Experten für eine bestimmte Domäne vorzuschlagen. Ein Vorschlag wird an alle CE und ED einer Domäne zur Diskussion geschickt. CE und ED entscheiden über den Vorschlag und veranlassen gegebenenfalls eine Aufnahmeprüfung für den Bewerber. Letztendlich entscheidet die Gruppe über die Zertifizierung des Bewerbers.

B.4.7.3 Mit vordefinierten Benutzergruppen arbeiten (N)

Actor: Alle Benutzer.

Interaction: Arbeiten mit vordefinierten Benutzergruppen.

Goal: Verwenden von und Beitreten zu vordefinierten Benutzergruppen.

Scenarios: (1) Ein Benutzer sendet eine Nachricht an die vordefinierte Gruppe von zertifizierten Experten einer bestimmten Domäne. (2) Ein Benutzer tritt einer bestimmten vordefinierten Gruppe zu einem Thema bei. Dadurch bekommt er Zugriff auf Informationen, die nur innerhalb der Gruppe „sichtbar" sind: Annotationen, spezielle Links, etc.

Das System kennt gewisse vordefinierte Benutzergruppen. Dies ist beispielsweise die Gruppe aller Editoren oder die Gruppe aller zertifizierten Experten aus einem Gebiet. Zudem werden themenbezogene Gruppen wie etwa eine Gruppe „Geographie Südamerikas" oder „Astronomie" angeboten.

Benutzer Benutzer können beispielsweise Nachrichten an vordefinierte Benutzergruppen senden. Außerdem kann jeder Benutzer gewissen themenbezogenen Gruppen beitreten um so Zugriff auf gruppenspezifische Annotationen zu haben, an Gruppen-Diskussionen teilzunehmen und gruppenspezifische Mailings zu erhalten.

B.4.7.4 Eigene Gruppen anlegen und verwalten (N)

Actor: Redakteure und zertifizierte Experten, alle Benutzer.

Interaction: Anlegen von neuen und Verwalten von existierenden Benutzergruppen.

Goal: Management von frei definierbaren Benutzergruppen.

Scenarios: (1) Mehrere verschiedene Benutzer wollen sich zu einer Benutzergruppe zusammenschließen. Ein Vorschlag wird an einen zertifizierten Experten geschickt, der die Gruppe anlegt. (2) Ein Benutzer stößt auf eine Gruppe, die für ihn von Interesse ist. Er tritt der Gruppe bei.

Im System können eigene Benutzergruppen frei angelegt werden, um beispielsweise an einem Projekt zusammenzuarbeiten. Diese Gruppen können beliebig um Mitglieder erweitert werden. Mitglieder können aus einer Gruppe nicht entfernt werden, sie können lediglich „ruhend" geschalten werden. Bestehende Gruppen können auch nicht gelöscht werden.

Es wird beispielsweise eine neue Gruppe gegründet, die einen CE aus dem Gebiet, einige DE und mehrere PU enthält. Nun können etwa Annotationen angelegt werden, die nur innerhalb dieser Gruppe sichtbar sind, Nachrichten an alle Gruppenmitglieder verschickt werden oder gar ein eigener Arbeitsbereich für die Gruppe verwendet werden (siehe Abschnitt „Workspaces").

Neue Gruppen können nur von ED und CE angelegt werden; DE und PU können sie lediglich vorschlagen.

B.4.8 Qualitäts-Feedback und Review Mechanismus (N)

Actor: Redakteur, zertifizierter Experte, System.

Interaction: Auswertung der Zugriffsstatistik und der Verwendung von gewissen Ressourcen.

Goal: Erhöhen der Qualität, Aktualität und Flexibilität der Informationen in den Datenbeständen.

Scenarios: (1) Ein Artikel wird sehr oft abgerufen; es besteht möglicherweise die Notwendigkeit, den Artikel zu aktualisieren oder zu erweitern. ED und CE werden von einem derartigen Artikel verständigt und können ihn bei Bedarf erweitern oder abändern. (2) Ein Artikel hat sehr viele Annotationen; es besteht möglicherweise die Notwendigkeit, den Artikel zu verbessern. ED und CE werden von einem derartigen Artikel verständigt und können auf etwaige Missstände reagieren. (3) Ein Artikel enthält einen Link zu einer externen Quelle, die nicht mehr verfügbar ist. Das System erkennt selbständig einen derartigen Link, deaktiviert ihn und verständigt einen CE oder ED.

ED und CE werden über Artikel, die aus ihrer Domäne stammen und sehr oft abgefragt oder annotiert werden, verständigt. ED und CE haben auf diese Weise die Möglichkeit, auf die Wünsche der Benutzer zu reagieren. Zudem wird es möglich, Artikel, die unverständlich oder unvollständig sind und dadurch mit einer großen Anzahl von Annotationen versehen sind, einem Review zu unterziehen.

Weiters werden ED und CE über Artikel verständigt, die von Benutzern gesucht werden, aber im System noch nicht vorhanden sind. Damit wird eine Flexibilisierung der Datenbasis erreicht.

Der Qualitäts-Feedback Mechanismus wird auch verwendet, um ED und CE über nicht mehr verfügbare Ressourcen zu informieren. Wird beispielsweise auf einen externen Link zugegriffen, der nicht mehr verfügbar ist, wird der Link automatisch deaktiviert, und der entsprechende Autor bzw. ED und CE aus der Domäne werden verständigt. (Siehe auch Abschnitt „Persistenz und Konsistenz von Links" weiter unten.)

Siehe auch Abschnitt „statistische Auswertung" unten.

B.5 Systemfunktionen

Als Systemfunktionen werden all jene Funktionen bezeichnet, die vom System verwendet werden, dem Benutzer aber nicht direkt zugänglich sind, oder nur Teil einer bestimmten Benutzerfunktion sind.

B.5.1 Suche und Clustering (N)

Actor: System.

Interaction: Das System führt eine Suche in den internen Datenbeständen durch und bereitet die Ergebnisse auf.

Goal: Verbesserung der Suchergebnisse und deren Präsentation.

Scenario: Eine Suchanfrage wird an das System gestellt. Es wird eine Suche in allen internen Datenbeständen durchgeführt, die ergebnisse werden kombiniert, geclustered und dem Benutzer graphisch präsentiert.

Das System bietet vielfältige Möglichkeiten zur Verbesserung von Suchergebnissen. Technologien wie Fuzzy Search, Sounds-Like Queries, Thesauri und Funktionen wie *„Meinen Sie vielleicht ...?"* werden verwendet, um die Eingabe der Benutzer besser auswerten und für Suchabfragen besser einsetzen zu können (O). Suche ist nicht nur in textuellen Inhalten möglich sondern kann auch in Bilddaten oder Filmen durchgeführt werden (O). Benutzer können dabei direkt auf die jeweilige „Region" des Inhaltes zugreifen, das heißt, der Benutzer erhält eine Szene eines Filmen als Suchergebnis oder nur einen Ausschnitt eines Bildes. Um das Formulieren von nicht-textuellen Suchabfragen zu vereinfachen, werden Techniken wie Query-by-Example und andere Ähnlichkeitssuchen verwendet (O).

Suchergebnisse aus den verschiedenen Datenquellen wie der *Brockhaus Enzyklopädie Digital* oder dem Archiv von *Die Presse* werden dynamisch in Cluster eingeteilt. Hierbei sind Cluster auf inhaltlicher Ebene (siehe [Autonomy 2005; Clusty 2005]) oder auf Basis von Dokumenttyp oder Dokumentursprung (etwa [A9 2005]) möglich (N). Inhaltliche Cluster können dabei mittels Heuristik (N) und anderen Ähnlichkeitsanalysen (O) erzeugt werden.

Neben noch näher zu untersuchenden graphischen Darstellungen von Suchergebnissen und Clustern ist vor allem die Visualisierung in Form von Wissensnetzen (vgl. [Brockhaus 2006]) eine Variante zur intuitiven Präsentation von Informationen (N).

B.5.2 Blacklisting (M/N)

Actor: System.

Interaction: Das System prüft Eingaben des Benutzers auf nicht zugelassene Ausdrücke und verständigt gegebenenfalls einen Redakteur oder einen zertifizierten Experten.

Goal: Unterdrücken von von Benutzern erstellten Beiträgen mit unerwünschten Ausdrücken.

Scenario: Ein Benutzer erstellt einen Artikel mit einem unerwünschten Ausdruck. Der Artikel kann deshalb nicht sofort publiziert werden. Zudem wird ein zertifizierter Experte oder ein Redakteur vom Vorfall verständigt. CE und ED können den Artikel publizieren oder den Autor zu einer Änderung auffordern.

Blacklisting bedeutet, dass sämtliche Wörter oder Ressourcen bis auf jene, die in einer bestimmten Liste (black list) vermerkt sind, verwendet werden können (M). Dieses Paradigma kann verfeinert werden, indem der Kontext einbezogen wird (N):

- Wenn ein ursprünglicher Artikel ein Wort, das auf der Blacklist angeführt ist, enthält, und ein Benutzer etwa eine Annotation oder Erweiterung zu diesem Artikel schreibt, dann ist das Wort auch im vom Benutzer verfassten Inhalt erlaubt;
- In bestimmten Kategorien sind bestimmte Wörter, die auf der Blacklist verzeichnet sind, dennoch erlaubt.

Wenn ein Benutzer beispielsweise einen Kommentar zu einem Artikel über Freud schreibt, dann ist es zulässig, das Wort „Sex" zu verwenden, weil es bereits im ursprünglichen Artikel vorkommt. Außerdem kann der Benutzer beispielsweise auch in jedem Artikel in den Kategorien „der Mensch" und „Zoologie" das Wort „Sex" verwenden.

Von PU und DE verfasste Artikel und andere Beiträge werden Blacklisting unterworfen. Welche Methoden hierbei genau verwendet oder entwickelt werden, wird in einem separaten Paper beschrieben (siehe die erwähnten Verfeinerungen oben). Liefert der Blacklisting-Filter bei einem Artikel ein positives Ergebnis,

dann muss der jeweilige Artikel einem Review unterzogen werden.

Er wird an einen CE weitergeleitet. Der CE entscheidet, ob der Artikel in dieser Form publiziert werden kann, abgelehnt werden muss, oder mit kleineren Änderungen publiziert werden kann. Im letzten Fall nimmt der CE Kontakt mit dem Autor des Artikels auf und versucht, einen Konsens zu finden.

Blacklisting wird nicht nur auf Artikel angewendet sondern auch auf Annotationen, Diskussionsbeiträge, Fragen. Eine Ausnahme bildet die persönliche Kommunikation zwischen zwei Benutzern bzw. einem Benutzer und einer Benutzergruppe.

B.5.3 Whitelisting (M/N)

Actor: System.
Interaction: Das System prüft beim Erstellen von Links, ob diese auf einer Whitelist verzeichnet sind und schreitet gegebenenfalls ein.
Goal: Verhindern von Links zu externen Ressourcen niedriger oder unbekannter Qualität.
Scenario: Ein Benutzer versucht, einen Link zu einer externen Ressource anzulegen. Dieser Vorgang kann nur dann erfolgreich abgeschlossen werden, wenn der Link (bzw. ein Teil davon) auf einer Whitelist verzeichnet ist.

Whitelisting bedeutet, dass keine Wörter oder Ressourcen bis auf jene, die in einer bestimmten Liste (white list) vermerkt sind, verwendet werden können (siehe [Lennon and Maurer 2003]).

Beim Hinzufügen von Links zu externen Ressourcen wird Whitelisting verwendet. Das heisst, es können nur Links zu bestimmten, in einer Liste vermerkten Servern hinzugefügt werden. Links zu anderen Servern sind nicht möglich (M).

Wenn der Benutzer einen externen Link anklickt, wird er innerhalb des Systemes verarbeitet (N). Das heißt:

- die entsprechende Ressource wird vom System geladen;
- Links innerhalb dieser Seite, die nicht auf der Whitelist stehen, werden entfernt;
- Links innerhalb dieser Seite, die auf der Whitelist stehen, werden vom System derart angepasst, dass auch sie vom System geladen werden;
- und die Seite wird innerhalb des Systems angezeigt.

Die Whitelist wird von ED und CE gewartet. DE und PU können lediglich Vorschläge an CE (und ED; nach einem Timeout automatisch weitergeleitet) für hinzuzufügende Ressourcen machen.

B.5.4 Transclusions (N/O)

Actor: Alle Benutzer.
Interaction: Der Benutzer erstellt einen neuen Artikel oder ein anderes inhaltsbezogenes Objekt mit Hilfe von Transclusions.
Goal: Verhindern von unnötigem Kopieren von Inhalten.
Scenario: Ein Benutzer schreibt einen Beitrag in einem Diskussionsforum innerhalb des Systems. Um einen Aspekt besser argumentieren zu können, transcluded der er einen Teil eines Artikels in sein Posting anstatt den entsprechenden Text mittels copy-and-paste einzufügen.

Benutzern werden Transclusions als Werkzeug zur Verfügung gestellt, um die nicht notwendige Duplizierung von Inhalten zu vermeiden und den ursprünglichen Kontext zu erhalten (siehe [Nelson 1981; Nelson 1996]). An jeder Stelle, wo neuer Inhalt generiert wird, also beispielsweise beim Schreiben eines neuen Artikels oder beim Verfassen eines Diskussionsbeitrages, sind Transclusions möglich.

Da ein Versionskontrollsystem verwendet wird (siehe unten), ist es nicht möglich, einen Artikel aus dem System zu entfernen. Damit ist die wichtigste Vorbedingung für die Realisierung von Transclusions erfüllt. Das bedeutet aber auch, dass ein inhaltlich inkorrekter Artikel nicht gelöscht werden kann. Wohl aber kann dessen Publikation rückgängig gemacht werden. Obwohl der Zugriff auf diesen Artikel über die Suchfunktion oder über Browsen in Kategorien nicht mehr möglich ist, und auch keine neuen Transclusions mehr angelegt werden können, ist der Zugriff durch bestehende Transclusions dennoch möglich.

Damit bleibt die Implementierung von Transclusions konsistent (Artikel können nicht gelöscht werden), und auch der Kontext bleibt erhalten: Obwohl ein zurückgezogener Artikel möglicherweise falsche Informationen enthält, wurde er mit genau diesen Inhalten für eine Transclusion verwendet. Genau diese – möglicherweise inkorrekten – Informationen sind für bereits angelegte Transclusions noch immer verfügbar, in diesem Kontext vielleicht auch immer noch korrekt und decken sich mit der Intention des Autors.

B.5.5 Statistische Auswertungen (N)

Actor: System.
Interaction: Auswerten von statistischen Daten.
Goal: Auffinden von Schwachstellen, fehlenden Artikeln, etc. im System.
Scenarios: (1) Bestimmte Artikel werden sehr oft abgerufen. Da diese Artikel von großem Interesse sind, ist es möglicherweise notwendig, sie zu aktualisieren oder zu erweitern. (2) Eine Suche nach einem bestimmten Begriff verläuft erfolglos, weil der entsprechende Artikel nicht vorhanden ist. Das System erkennt dies und meldet es an einen Redakteur.

Das System führt Statistiken über Zugriffe auf bestimmte Seiten, die Häufigkeit und Art von Suchabfragen und ähnliche Aktionen. Eine automatisierte, statistische Auswertung dieser Daten gemeinsam mit einem Benachrichtigungsdienst, der ED und CE beispielsweise über zu aktualisierende Artikel informiert, erhöhen die Flexibilität des Systems.

Es werden beispielsweise Zeitstempel verwendet, um sowohl die letzte Änderung als auch den letzten

Zugriff auf einen Artikel zu vermerken (N). Wird ein Artikel lange nicht aktualisiert, wird der Autor verständigt, um die Gültigkeit der Information zu überprüfen.

Siehe auch Abschnitt „Qualitäts-Feedback und Review Mechanismus“ weiter oben.

B.5.6 Personalisierung (O)

Actor: System.
Interaction: Anpassen der Benutzeroberfläche, von Suchergebnissen, etc. basierend auf den Einstellungen und Profilen des Benutzers.
Goal: Anpassen an Wünsche und Arbeitsweise des Benutzers.
Scenario: Das System passt das User Interface an die Erfahrung des Benutzers mit dem System an.

Das System ist auf den wesentlichen Ebenen der Benutzerinteraktion adaptierbar. So werden beispielsweise Suchalgorithmen und die Darstellung der Ergebnisse dem Profil eines Benutzers entsprechend angepasst. Auch die Anzeige von Artikeln ist personalisierbar: Es werden etwa alle Artikel zusammen mit den dazugeförenden Diskussionsforen angezeigt oder nur redaktionelle Einträge mit Links zu externen Ressourcen; es werden Artikel von ED und CE angezeigt und mit Clustern ähnlicher Dokumente ergänzt oder es werden lediglich die Artikel mit einer Menge von Metadaten angezeigt.

Zusätzlich kann auch die Benutzerschnittstelle adaptiert werden. Verwendet ein Benutzer zum ersten Mal das System, dann wird beispielsweise ein vereinfachtes Interface angezeigt, in dem komplexere Funktionen nicht verfügbar sind. Nach mehrmaliger Verwendung werden schrittweise komplexere Funktionen eingeblendet. Der Benutzer hat jederzeit die Möglichkeit, diese Personalisierung aufzuheben und den vollen Funktionsumfang anwenden zu können.

Siehe auch Abschnitt „Personalisierung“ weiter oben.

B.6 Potentielle Design Entscheidungen

Einige Design-Entscheidungen haben potentiell weitreichende Konsequenzen bezüglich des implementierbaren Funktionsumfanges und der Funktionalität sowie der Komplexität und Flexibilität eines Systemes. Dennoch müssen einige Entscheidungen über Architektur und Design des Systems getroffen werden, um die Implementierung von Konzepten wie Transclusions zu ermöglichen und beispielsweise die Konsistenz und Persistenz von Links und Bookmarks zu erleichtern.

Die folgenden Abschnitte stellen eine Auswahl von Aspekten dar, die Auswirkungen auf die Umsetzbarkeit und die konkrete Implementierung einiger oben genannter Funktionen haben und im Rahmen einer technischen Spezifikation des Systems genauer diskutiert werden sollten.

B.6.1 Link Datenbank (O)

Im klassischen HTML-Paradigma sind nur unidirektionale Hyperlinks möglich, und es kann passieren, dass Links nicht mehr verfügbar sind. Diese Einschränkungen sind nicht wünschenswert und können etwa durch Verwendung einer Link Datenbank behoben werden (vgl. [Kappe 1995]).

Mit Hilfe einer Link-Datenbank können beispielsweise bidirektionale oder potentiell auch m-zu-n Links realisiert werden. Zudem ist es möglich, die Konsistenz von Links zu erhalten und deren Persistenz zu kontrollieren.

Bei Links ins WWW ist es notwendig, deren Konsistenz und Persistenz zu erhalten (siehe unten). Zusätzlich werden Qualitätssicherungs-Strategien wie Blacklisting und Whitelisting angewendet (siehe oben).

B.6.2 Persistenz und Konsistenz von Links (M/N)

Innerhalb des Systems sind sämtliche Links persistent. Das heißt, dass Links auf Kategorien, Artikel oder beispielsweise auch Annotationen oder Diskussionsbeiträge immer verfügbar sind und auch im Laufe der Zeit nicht „vergehen“. Zudem sind sämtliche Links konsistent. Das heißt, durch Änderung oder Verschieben eines Artikels in eine Kategorie kann es beispielsweise nicht geschehen, dass ein Link zu einer internen Ressource nicht mehr verfügbar ist.

Konsistenz und Verfügbarkeit von Links gilt aber auch für externe Ressourcen. Bevor ein Artikel angezeigt wird, müssen demnach sämtliche Links zu externen Datenquellen auf ihre Verfügbarkeit überprüft werden (N). Es kann auch notwendig sein, zu überprüfen, ob sich der Inhalt einer externen Ressource geändert hat (O).

B.6.3 Versionskontrolle (N)

Ein Versionskontrollsystem organisiert die verschiedenen Versionen von inhaltsbezogenen Objekten im System. Versionskontrolle ist notwendig, um die „Geschichte“ eines Artikels und dessen Änderungen nachvollziehen zu können. Weiters wird sie von Funktionen wie Transclusions und Konsistente Bookmarks benötigt.

Die Verwendung eines Versionskontrollsystem hat beispielsweise Auswirkungen auf das Erstellen und Ändern von Artikeln. Ein Artikel etwa kann nicht geändert werden. Stattdessen wird eine neue, geänderte Version des Artikels angelegt, während der ursprüngliche Artikel unverändert bleibt.

Auch bei von Benutzern angelegten Bookmarks muss ein Versionskontrollsystem berücksichtigt werden. Ein Bookmark eines Benutzers bezieht sich auf eine konkrete Version eines Artikels. Wenn das Bookmark aufgerufen wird und in der Zwischenzeit eine neuere Version angelegt wurde, wird zwar der ursprüngliche, nun ältere Artikel angezeigt, der Benutzer muss aber informiert werden, dass in der Zwischenzeit auch eine neuere Version verfügbar ist. Alternativ kann beim Anlegen des Bookmarks dem Benutzer angeboten werden, bei dessen Aufruf die jeweils aktuellste Version des Artikels anzeigen zu lassen.

Aufgrund der geringen Informationen über die vorhandenen Datenquellen ist es noch unklar, wie ein Versionskontrollsystem in die bestehenden Architekturen eingebaut werden kann. Es muss beispielsweise festgelegt werden, wo eine neue, geänderte Version eines Artikels abgespeichert wird.

Ein Beispiel: Ein Artikel aus dem Archiv von *Die Presse* wird von einem CE erweitert um aktuelle Informationen und einen Link zu einer externen Ressource. Wo wird diese neue Version des Artikels abgespeichert? Am sinnvollsten scheint derzeit die Speicherung in einem internen Archiv, weil damit der Zusammenhang zu den neuen Informationen bzw. der Kontext der Änderung selbst eher gegeben ist.

B.6.4 Identifizierung von Objekten (N)

Das System verwendet eine Methode zur eindeutigen Identifikation von sowohl internen als auch externen Ressourcen. Vorgeschlagen wird ein System ähnlich DOI (Digital Object Identified, [DOI 2003]) oder cIDf (Content ID Forum, [cIDf 2003]). Jedes Object kann durch ein 4-Tupel identifiziert werden:

- Datenquelle (etwa eine interne Datenbank);
- Objekt ID innerhalb der originalen Datenquelle (beispielsweise die ID innerhalb der Datenbank);
- Versionsnummer des Objektes;
- Qualität des Objektes (etwa eine niedrige Auflösung bei Bildern oder hohe Bitrate bei Ton).

Durch die Verwendung eines derartigen 4-Tupels können mehrere Aspekte erreicht werden: Es können beispielsweise die ursprünglichen Objekt-IDs weiterverwendet werden, weil durch die zusätzliche Angabe der Datenquelle als Prefix eine Art „Namespace" eingeführt wird. Versionsnummern erlauben eine systemweite Versionskontrolle (siehe oben) und ermöglichen damit Funktionen wie Transclusions.

Die Angabe der Qualität eines Objektes schafft neue Anwendungsgebiete. Beispielsweise können auf Computern mit permanenten Internetverbindungen Bilder, Ton und Videos in einer hohen Qualität dargestellt werden. Zu Computern mit Modemverbindungen hingegen werden dieselben Dokumente in niedrigeren Qualitätsstufen übertragen. Auf mobile Geräte wie PDAs oder Handies können generell andere Objekte übertragen werden: etwa ein Ausschnitt eines Bildes anstelle des gesamten Materials. Auf ein Handy wird nicht ein kompletter Artikel übertragen sondern nur eine (automatisch) generierte Zusammenfassung.

Zudem wird es durch die Angabe von verschiedenen Qualitätsstufen auch möglich, Material über dieses System kommerziell zu vertreiben. Ein Benutzer sieht sich beispielsweise eine Landkarte aus dem Lexikon auf seinem Computer an. Das gleiche Objekt ist mit einer weit höheren Auflösung und versehen mit DRM-Einschränkungen (etwa „nur zwei Ausdrucke möglich") im System verfügbar und kann vom Benutzer käuflich erworben werden.

Das System zur Objektidentifizierung greift auf einen Resolver-Dienst ähnlich dem von DOI zurück. Sowohl innerhalb des Systems als auch nach außen hin werden ausschließlich die gewählten Objektidentifizierungen (4-Tupel) verwendet. Wird ein Objekt angefordert, dann löst der Resolver-Dienst die Objekt ID auf und gibt an, auf welche Weise von woher die Ressource bezogen werden kann (vgl. das Internet Domain Name System). Der Resolver-Dienst speichert demnach drei Werte ab:

- die verwendete Objekt ID;
- den Ort, von wo die eigentliche Ressource bezogen werden kann (beispielsweise ein URL); und
- die Methode, wie die Ressource zu beziehen ist (beispielsweise über das Protokoll HTTPs mit gegebenem Benutzernamen und Kennwort).

Damit kann die Persistenz von Links innerhalb des Systemes gewährleistet werden, auch wenn sich Links von externen Ressourcen ändern.

B.6.5 Metadaten zu Objekten (N)

Es wird versucht, zu sämtlichen im System gespeicherten, inhaltsbezogenen Objekten Metadaten zu sammeln. Dies gestaltet sich relativ einfach für Inhalte, die im System produziert werden, weil zumindest der Autor, das Datum und eine Kategorie verfügbar sind. Dies kann aber auch relativ unkompliziert für externe Daten sein: Bilder aus Zeitungsarchiven enthalten in der Regel Metadaten wie den Photographen, das Datum und den Ort, an dem das Photo aufgenommen wurde, sowie Schlagwörter. Zudem können auch Referenzen auf Artikel, in denen das Photo vorkommt, als Metadaten zum Bild gespeichert werden.

Metadaten werden vom System in der Suche verwendet, aber auch für die Erstellung von Clustern und Knowledge Maps genutzt.

Es wird versucht, Metadaten zumindest im Ausmaß der Dublin Core Metadata Initiative zu sammeln ([DC 2004; DCMI 2006]). Eine umfangreichere Beschreibung etwa mittels MPEG-7 Descriptors ist optional ([MPEG-7 2002]).

C Data from the Analysis of Anonymous Authorship in Wikipedia

The following tables contain the results from the analysis of anonymous authorship in Wikipedia detailed in chapter 6. The figures listed in this appendix were collected on February 9th, 2006, and are the basis for all charts in chapter 6.

The first table lists the Wikipedia articles that were analysed and their versions as well as the classes and categories they were assigned in the investigation. The second table presents the characteristic values—number of unique users per article, number of edits per article, and the minimum, maximum, mean, median, and mode values for edits per user per article—for registered and anonymous users combined ("all users"). Tables three and four contain these characteristic values for registered users and anonymous users, respectively.

Articles, Classes, and Categories

#	Class	Sub-Class	Category	Article	Article URL	Article Version
1			Physics	Boltzmann Constant	http://en.wikipedia.org/wiki/Boltzmann_constant	37273204
2			Physics	Force	http://en.wikipedia.org/wiki/Force	38360622
3			Mathematics	Wave Equation	http://en.wikipedia.org/wiki/Wave_equation	36341017
4			Mathematics	Hilbert's Problems	http://en.wikipedia.org/wiki/Hilbert%27s_problems	35970666
5			Biology	Brown Rat	http://en.wikipedia.org/wiki/Brown_Rat	38375851
6			Biology	Cotyledon	http://en.wikipedia.org/wiki/Cotyledon	37805723
7			Geology	Cretaceous	http://en.wikipedia.org/wiki/Cretaceous	38683956
8			Geography	Beijing	http://en.wikipedia.org/wiki/Beijing	38844666
9			Chemistry	Nicolaou Taxol Total Synthesis	http://en.wikipedia.org/wiki/Nicolaou_Taxol_total_synthesis	38135999
10			Chemistry	Haloalkane	http://en.wikipedia.org/wiki/Haloalkane	38017967
11			Computer Science	Mach Kernel	http://en.wikipedia.org/wiki/Mach_kernel	37363731
12			Computer Science	Turing Test	http://en.wikipedia.org/wiki/Turing_Test	15926722
13			History	Historical Capitals of China	http://en.wikipedia.org/wiki/Historical_capitals_of_China	38032312
14			History	Battle of Caporetto	http://en.wikipedia.org/wiki/Battle_of_Caporetto	36893840
15			Chemistry	Geometric Isomerism	http://en.wikipedia.org/wiki/Geometric_isomerism	28568074
16			Chemistry	Baking Powder	http://en.wikipedia.org/wiki/Baking_powder	38638258
17			Chemistry	Casein	http://en.wikipedia.org/wiki/Casein	37736588
18			Mathematics	Mode (Statistics)	http://en.wikipedia.org/wiki/Mode_%28statistics%29	37625865
19			Medicine	Anemia	http://en.wikipedia.org/wiki/Anemia	38857667
20			Drugs, Medicine, Chemistry	Nadolol	http://en.wikipedia.org/wiki/Nadolol	30876959
21			Medicine	Anencephaly	http://en.wikipedia.org/wiki/Anencephaly	37205042
22			Society, Art	Tattoo	http://en.wikipedia.org/wiki/Tattoo	38836429
23			Society, Illegal Societies	Triad	http://en.wikipedia.org/wiki/Triad	37853283
24			Law	Intellectual Property	http://en.wikipedia.org/wiki/Intellectual_property	38841023
25			Politics	Protectionism	http://en.wikipedia.org/wiki/Protectionism	38214150
26			Economics	Classic Economics	http://en.wikipedia.org/wiki/Classical_economics	32516940
27			Philosophy, Economics	Ceteris Paribus	http://en.wikipedia.org/wiki/Ceteris_paribus	36384037
28			Philosophy, Economics	Parsimony	http://en.wikipedia.org/wiki/Parsimony	33399331
29			History	Gorgidas	http://en.wikipedia.org/wiki/Gorgidas	38390457
30			Philosophy	Meta-Ethics	http://en.wikipedia.org/wiki/Meta-ethics	36343695
31			Philosophy	Moral Relativism	http://en.wikipedia.org/wiki/Moral_relativism	37082310
32			Philosophy	Transcendental Idealism	http://en.wikipedia.org/wiki/Transcendental_idealism	38298718
33			Society	Debunker	http://en.wikipedia.org/wiki/Debunker	38543378
34			Psychology	Hypnagogic Hallucination	http://en.wikipedia.org/wiki/Hypnagogic_hallucination	21455394
35			Society	Ig Nobel Prize	http://en.wikipedia.org/wiki/Ig_Nobel_Prize	38619265
36			Miscellaneous	Five-Second Rule	http://en.wikipedia.org/wiki/Five-second_rule	38825167
37			Miscellaneous	Loukoumia	http://en.wikipedia.org/wiki/Loukoumia	18523754
38			People, Culture	Dr. Seuss	http://en.wikipedia.org/wiki/Dr._Seuss	38727466
39			Geography	La Jolla	http://en.wikipedia.org/wiki/La_Jolla	28655483
40			Chemistry	Radiocarbon Dating	http://en.wikipedia.org/wiki/Radiocarbon_dating	38767688
41			History, War	Decimation	http://en.wikipedia.org/wiki/Decimation	30679352
42			Psychology	Procrastination	http://en.wikipedia.org/wiki/Procrastination	38115133
43			Economics	NASDAQ	http://en.wikipedia.org/wiki/NASDAQ	38803752
44			Miscellaneous	TLA	http://en.wikipedia.org/wiki/TLA	38424603
45			Biology, Botany	Bristlecone Pine	http://en.wikipedia.org/wiki/Bristlecone_pine	38067867
46			Religion	Kaballah	http://en.wikipedia.org/wiki/Kabbalah	38766225
47			Geography	Girona	http://en.wikipedia.org/wiki/Girona	38248667
48			Politics	Mobocracy	http://en.wikipedia.org/wiki/Mobocracy	25524960
49			Miscellaneous	Mechelle	http://en.wikipedia.org/wiki/Mechelle	24959419
50			Entertainment	Being Julia	http://en.wikipedia.org/wiki/Being_Julia	32134876
51			People, Culture, Arts	Hans-Peter Zimmer	http://en.wikipedia.org/wiki/Hans-Peter_Zimmer	36185557
52			Entertainment	Pinochle	http://en.wikipedia.org/wiki/Pinochle	38009589
53			Chemistry	Carbonyldiimidazole	http://en.wikipedia.org/wiki/Carbonyldiimidazole	35355940
54			Geography	Tjurabalan	http://en.wikipedia.org/wiki/Tjurabalan	34108220
55			Entertainment	End of Time (Chrono Trigger)	http://en.wikipedia.org/wiki/End_of_Time_%28Chrono_Trigger%29	31115398
56			Architecture	Baker Memorial Library	http://en.wikipedia.org/wiki/Baker_Memorial_Library	33350095
57			History, War	French Battleship Redoutable (1876)	http://en.wikipedia.org/wiki/French_battleship_Redoutable_%281876%29	32902962
58			Chemistry	Gel Electrophoresis	http://en.wikipedia.org/wiki/Gel_electrophoresis	37845418
59			People, Culture, Arts	Albrecht Altdorfer	http://en.wikipedia.org/wiki/Albrecht_Altdorfer	37416649
60			Chemistry	Rare Earth Elements	http://en.wikipedia.org/wiki/Rare_earth_elements	16212908
61			Geography	Dolomites	http://en.wikipedia.org/wiki/Dolomites	37996357
62			Technology, Mech. Engineering	Glow Plug	http://en.wikipedia.org/wiki/Glow_plug	34815681
63			Geography	Tahiti	http://en.wikipedia.org/wiki/Tahiti	36482714
64			Biology, Botany	Caper	http://en.wikipedia.org/wiki/Caper	37417739
65			Religion, Islam	Sharia	http://en.wikipedia.org/wiki/Sharia	36716012
66			Politics	Anarchism	http://en.wikipedia.org/wiki/Anarchism	38399178
67			People, Science	Alexander von Humboldt	http://en.wikipedia.org/wiki/Alexander_von_Humboldt	38642551
68			Biology, Chemistry	Photorespiration	http://en.wikipedia.org/wiki/Photorespiration	34216375
69			Music	Richard Clayderman	http://en.wikipedia.org/wiki/Richard_Clayderman	38433513
70			Geography	Rovaniemi	http://en.wikipedia.org/wiki/Rovaniemi	38361937
71			Politics	Oligarchy	http://en.wikipedia.org/wiki/Oligarchy	38790907
72			Culture, Sports	Shotokan	http://en.wikipedia.org/wiki/Shotokan	38793448
73			Biology, Zoology	Stromateidae	http://en.wikipedia.org/wiki/Stromateidae	37698871
74			Geography	Easter Islands	http://en.wikipedia.org/wiki/Easter_island	20540651
75			Philiophy	Nihilism	http://en.wikipedia.org/wiki/Nihilism	38780732
76			Geography	Sinai Peninsula	http://en.wikipedia.org/wiki/Sinai_Peninsula	38133732
77			People, Politics	Vaclav Havel	http://en.wikipedia.org/wiki/V%C3%A1clav_Havel	38208911
78			Religion, Society	Sect	http://en.wikipedia.org/wiki/Sect	37684399
79			Religion, Mythology	Holy Grail	http://en.wikipedia.org/wiki/Holy_Grail	38733545
80			Society, History	Flagellation	http://en.wikipedia.org/wiki/Flagellation	37340140
81			Miscellaneous	Ambigram	http://en.wikipedia.org/wiki/Ambigram	38211409
82			Philosophy	Fatalism	http://en.wikipedia.org/wiki/Fatalism	36471303
83			Geography	Melaka	http://en.wikipedia.org/wiki/Melaka	38748558
84			Science	Nanotechnology	http://en.wikipedia.org/wiki/Nanotechnology	38877751
85			Miscellaneous	Pai Mu Tan Tea	http://en.wikipedia.org/wiki/Pai_Mu_Tan_tea	23347149
86			Literature	Manufacturing Consent	http://en.wikipedia.org/wiki/Manufacturing_Consent	37101609
87			Miscellaneous	Special K (Disambiguation)	http://en.wikipedia.org/wiki/Special_K_%28disambiguation%29	26447858
88			Computer Science	USB Mass Storage Device Class	http://en.wikipedia.org/wiki/USB_mass_storage_device_class	36986591
89			Science	Roman Numerals	http://en.wikipedia.org/wiki/Roman_numerals	38396369
90			Geography	Damascus	http://en.wikipedia.org/wiki/Damascus	38730007
91			History, War	Use of Poison Gas in World War I	http://en.wikipedia.org/wiki/Use_of_poison_gas_in_World_War_I	32081772
92			Entertainment	Peanuts	http://en.wikipedia.org/wiki/Peanuts	38670130
93			People, Literature	Jack London	http://en.wikipedia.org/wiki/Jack_London	38825359
94			Science	Communication	http://en.wikipedia.org/wiki/Communication	38046831
95			Medicine	Nervous Tissue	http://en.wikipedia.org/wiki/Nervous_tissue	16142015
96			Literature	The Hobbit	http://en.wikipedia.org/wiki/The_Hobbit	38826711
97			History, War	Opium War	http://en.wikipedia.org/wiki/Opium_War	15958687
98			Medicine	Parkinson's Disease	http://en.wikipedia.org/wiki/Parkinson%27s_disease	38735301

Manually Selected by Authors

Manually Selected by Volunteers

#	Class	Sub-Class	Category	Article	Article URL	Article Version
99			Music	Slayer	http://en.wikipedia.org/wiki/Slayer	38879076
100			Drugs	Methamphetamine	http://en.wikipedia.org/wiki/Methamphetamine	38778939
101			Drugs, Society	Recreational Drug Use	http://en.wikipedia.org/wiki/Recreational_drug_use	38057817
102			Society, Sexuality	Statutory Rape	http://en.wikipedia.org/wiki/Statutory_rape	38698832
103			Society, Sexuality	Academic Seduction	http://en.wikipedia.org/wiki/Academic_seduction	27971511
104			Society, Sexuality	Sexual Harassment	http://en.wikipedia.org/wiki/Sexual_harassment	38714040
105			Society, Politics	Human Experimentation	http://en.wikipedia.org/wiki/Human_experimentation	38677550
106			War, Politics	Abu Ghraib Torture and Prisoner Abuse	http://en.wikipedia.org/wiki/Abu_Ghraib_torture_and_prisoner_abuse	38719567
107			Society, Sociology	Bully	http://en.wikipedia.org/wiki/Bully	38766627
108			Society, Sociology	Spousal Abuse	http://en.wikipedia.org/wiki/Spousal_abuse	35241134
109			Society, Sexuality	Chickenhawk (sexuality)	http://en.wikipedia.org/wiki/Chickenhawk_%28sexuality%29	38723981
110			Society, Sociology	Child Pornography	http://en.wikipedia.org/wiki/Child_pornography	38816524
111			Sexuality	Shotacon	http://en.wikipedia.org/wiki/Shotacon	38394784
112			Drugs	LSD	http://en.wikipedia.org/wiki/LSD	38872325
113			Drugs	Psychedelic Mushroom	http://en.wikipedia.org/wiki/Psychedelic_mushroom	38714785
114			Medicine, Science	Anti-Psychiatry	http://en.wikipedia.org/wiki/Anti-psychiatry	38625914
115			Sexuality	Necrophilia	http://en.wikipedia.org/wiki/Necrophilia	38856704
116			Sexuality	Petticoat Punishment	http://en.wikipedia.org/wiki/Petticoat_Punishment	38056231
117			Sexuality	Anal Masturbation	http://en.wikipedia.org/wiki/Anal_masturbation	38198115
118			Sexuality	Autoerotic Asphyxiation	http://en.wikipedia.org/wiki/Autoerotic_asphyxiation	37752048
119			Sexuality	Fifi (Masturbation Aid)	http://en.wikipedia.org/wiki/Fifi_%28masturbation_aid%29	33282952
120			Pornography	Jeff Stryker	http://en.wikipedia.org/wiki/Jeff_Stryker	38117120
121			Religion	Ku Klux Klan	http://en.wikipedia.org/wiki/Ku_Klux_Klan	38870644
122			Society, Politics	Gay Rights	http://en.wikipedia.org/wiki/Gay_rights	38748456
123			Religion, Society	Militant Islam	http://en.wikipedia.org/wiki/Militant_Islam	35815548
124			Religion	Scientology	http://en.wikipedia.org/wiki/Scientology	38876995
125			Religion, Espionage	Operation Snow White	http://en.wikipedia.org/wiki/Operation_Snow_White	37261477
126			Terrorism	Ahmed Yassin	http://en.wikipedia.org/wiki/Ahmed_Yassin	38113877
127			Politics, History, Terrorism	Nakba	http://en.wikipedia.org/wiki/Nakba	38206861
128			History, Science Fiction	Roswell Incident	http://en.wikipedia.org/wiki/Roswell_incident	21921781
129			Religion, Politics, History	Jyllands-Posten Muhammad Cartoons Controversy	http://en.wikipedia.org/wiki/Jyllands-Posten_Muhammad_cartoons_controvers	38876376
130			Religion	Satanism	http://en.wikipedia.org/wiki/Satanism	38839411
131			Sexuality	Zoophilia	http://en.wikipedia.org/wiki/Zoophilia	38208106
132			Sexuality	Prince Albert Piercing	http://en.wikipedia.org/wiki/Prince_Albert_piercing	38368233
133			Religion	Sun Myung Moon	http://en.wikipedia.org/wiki/Sun_Myung_Moon	38835843
134			Religion, Politics, History	New Anti-Semitism	http://en.wikipedia.org/wiki/New_anti-Semitism	38703805
135	"Delicate" Artikel		Intelligence Organisation	Project ARTICHOKE	http://en.wikipedia.org/wiki/Project_ARTICHOKE	37927325
136			Intelligence Organisation	Black Site	http://en.wikipedia.org/wiki/Black_site	38387548
137			Terrorism	Guantanamo Bay	http://en.wikipedia.org/wiki/Guantanamo_Bay	38874766
138			Politics, History, War	2003 War on Iraq	http://en.wikipedia.org/wiki/2003_war_on_Iraq	24244545
139			Society, Sociology	Euthanasia	http://en.wikipedia.org/wiki/Euthanasia	38828950
140			Society, Psychology	Suicide	http://en.wikipedia.org/wiki/Suicide	38845887
141			Sociology	Ted Bundy	http://en.wikipedia.org/wiki/Ted_Bundy	38846842
142			Religion	Heaven's Gate Cult	http://en.wikipedia.org/wiki/Heaven%27s_Gate_%28cult%29	38769165
143			Politics, Terrorism	Hezbollah	http://en.wikipedia.org/wiki/Hezbollah	38830221
144			Politics, Terrorism	Al-Aqsa Intifada	http://en.wikipedia.org/wiki/Al-Aqsa_Intifada	38855951
145			Intelligence Organisation	N44982	http://en.wikipedia.org/wiki/N44982	37356372
146			Politics, Terrorism	War on Terror	http://en.wikipedia.org/wiki/War_on_Terror	20213625
147			Society	Jafa	http://en.wikipedia.org/wiki/Jafa	37818305
148			Terrorism	al-Qaeda	http://en.wikipedia.org/wiki/Al-Qaeda	38862411
149			Society, Religion, Politics	Abortion	http://en.wikipedia.org/wiki/Abortion	38806159
150			Society, Politics	Racism	http://en.wikipedia.org/wiki/Racism	38766395
151			Politics, History	American Nazi Party	http://en.wikipedia.org/wiki/American_Nazi_Party	37857558
152			Religion	Judaism	http://en.wikipedia.org/wiki/Judaism	38605886
153			Occultism	Necromancy	http://en.wikipedia.org/wiki/Necromancy	38049338
154			Biochemistry	Genetically Modified Food	http://en.wikipedia.org/wiki/Genetically_modified_food	38813725
155			Society	Domestic Violence	http://en.wikipedia.org/wiki/Domestic_violence	38809236
156			History	Holocaust Denial	http://en.wikipedia.org/wiki/Holocaust_denial	38869319
157			Sexuality, Medicine	Prostate Milking	http://en.wikipedia.org/wiki/Prostate_milking	37162347
158			Occultism	Hermetic Order of the Golden Dawn	http://en.wikipedia.org/wiki/Hermetic_Order_of_the_Golden_Dawn	38877261
159			Science	Nuclear Power	http://en.wikipedia.org/wiki/Nuclear_power	38768437
160			Public Protest	Gay Nigger Association of America	http://en.wikipedia.org/wiki/Gay_Nigger_Association_of_America	38700820
161			Medicine	Gonorrhea	http://en.wikipedia.org/wiki/Gonorrhea	15956283
162			Medicine	Scabies	http://en.wikipedia.org/wiki/Scabies	37229330
163			Biology, Politics	Whaling	http://en.wikipedia.org/wiki/Whaling	38711832
164			Society, Sociology	Child Sexual Abuse	http://en.wikipedia.org/wiki/Child_sexual_abuse	38835597
165			Medicine	Syphilis	http://en.wikipedia.org/wiki/Syphilis	38722440
166			History, War	The Holocaust	http://en.wikipedia.org/wiki/The_Holocaust	38862374
167			Conspiracy Theory	Apollo Moon Landing is a Hoax	http://en.wikipedia.org/wiki/Apollo_moon_landing_hoax_accusations	38822817
168			History	Appeasement	http://en.wikipedia.org/wiki/Appeasement	38520190
169			Religion, History	Jesus	http://en.wikipedia.org/wiki/Jesus	38876713
170			History	Adolf Hitler	http://en.wikipedia.org/wiki/Adolf_Hitler	38749532
171			Geography	United States	http://en.wikipedia.org/wiki/United_States	38868438
172			Geography	Iraq	http://en.wikipedia.org/wiki/Iraq	38833491
173			Geography	Israel	http://en.wikipedia.org/wiki/Israel	38872729
174			Religion, History	Land of Israel	http://en.wikipedia.org/wiki/Land_of_Israel	37223158
175			Philosophy	Eugenics	http://en.wikipedia.org/wiki/Eugenics	38645935
176			History, Sports	France at the 1896 Summer Olympics	http://en.wikipedia.org/wiki/France_at_the_1896_Summer_Olympics	34027968
177			Geography, Politics	Prefectures of the Central African Republic	http://en.wikipedia.org/wiki/Prefectures_of_the_Central_African_Republic	36022243
178			Miscellaneous	NatureServe	http://en.wikipedia.org/wiki/NatureServe	27386620
179			Science, Astronomy	La Posta Astro-Geophysical Observatory	http://en.wikipedia.org/wiki/La_Posta_Astro-Geophysical_Observatory	31006975
180			Economics, Chemistry	Diosynth	http://en.wikipedia.org/wiki/Diosynth	27115196
181			Architecture, Entertainment	Royal Theatre (Victoria)	http://en.wikipedia.org/wiki/Royal_Theatre_(Victoria)	29485491
182			Geography	Bitterne Railway Station	http://en.wikipedia.org/wiki/Bitterne_railway_station	32639226
183			History, Politics	Albert L'Ouvrier	http://en.wikipedia.org/wiki/Albert_L'Ouvrier	35283176
184			Earth Sciences, Geology	Vernal Pool	http://en.wikipedia.org/wiki/Vernal_pool	30784586
185			History, War	Navy Commendation Star	http://en.wikipedia.org/wiki/Navy_Commendation_Star	36775678
186			People, Sports	Enefiok Udo-Obong	http://en.wikipedia.org/wiki/Enefiok_Udo-Obong	37754382
187			Geography	Long Lake	http://en.wikipedia.org/wiki/Long_Lake	36966895
188			Literature	Bernardo de la Paz	http://en.wikipedia.org/wiki/Bernardo_de_la_Paz	37165415
189			History, War	Harlem Hellfighters	http://en.wikipedia.org/wiki/Harlem_Hellfighters	38757486
190			Entertainment	Ivan Vasilievich: Back to the Future	http://en.wikipedia.org/wiki/Ivan_Vasilievich:_Back_to_the_Future	32608173
191			Religion	Mount Angel Abbey	http://en.wikipedia.org/wiki/Mount_Angel_Abbey	37316069
192			Geography	Geography of the Philippines	http://en.wikipedia.org/wiki/Geography_of_the_Philippines	36820786
193			Entertainment	Joey Fatone	http://en.wikipedia.org/wiki/Joey_Fatone	38451830
194			History	Cephas Washburn	http://en.wikipedia.org/wiki/Cephas_Washburn	32577018
195			Entertainment	Group Dance	http://en.wikipedia.org/wiki/Group_dance	36409956
196			Geography	Two-Mile Borris	http://en.wikipedia.org/wiki/Two-Mile_Borris	37167970
197			History	Afghanistan Timeline July 2001	http://en.wikipedia.org/wiki/Afghanistan_timeline_July_2001	36876876

#	Class	Sub-Class	Category	Article	Article URL	Article Version
198	Randomly Selected Articles (http://en.wikipedia.org/wiki/Special:Randompage)		Geography	Viscount Chandos	http://en.wikipedia.org/wiki/Viscount_Chandos	16277634
199			History	Charles X of France	http://en.wikipedia.org/wiki/Charles_X_of_France	35607312
200			Mathematics	Square Pyramidal Number	http://en.wikipedia.org/wiki/Square_pyramidal_number	27573986
201			Geography	Ludgate Hill	http://en.wikipedia.org/wiki/Ludgate_Hill	22665154
202			Miscellaneous	Edmonton Journal	http://en.wikipedia.org/wiki/Edmonton_Journal	38034442
203			Society, Miscellaneous	Rotary International	http://en.wikipedia.org/wiki/Rotary_International	38631482
204			Music	Burningn'n Tree	http://en.wikipedia.org/wiki/Burningn'n_Tree	36704980
205			People, Sports	Lev Yashin	http://en.wikipedia.org/wiki/Lev_Yashin	37737410
206			Science, Philosophy	Philosophical Institute	http://en.wikipedia.org/wiki/Philosophical_Institute	34009119
207			Entertainment	List of Protagonists in Xenosaga	http://en.wikipedia.org/wiki/List_of_protagonists_in_Xenosaga	38117658
208			History, War	Military Unit Mottos: United States	http://en.wikipedia.org/wiki/Military_Unit_Mottos:_United_States	35630475
209			Entertainment	Sarafina	http://en.wikipedia.org/wiki/Sarafina	37292319
210			Sports	Aerobatics	http://en.wikipedia.org/wiki/Aerobatics	37803650
211			Miscellaneous	ITV Evening News	http://en.wikipedia.org/wiki/ITV_Evening_News	35718828
212			Biology	Blaarkop	http://en.wikipedia.org/wiki/Blaarkop	33008328
213			Science	Residency (Medicine)	http://en.wikipedia.org/wiki/Residency_(medicine)	37210564
214			Society	Mechitza	http://en.wikipedia.org/wiki/Mechitza	36713816
215			Miscellaneous	Ramage	http://en.wikipedia.org/wiki/Ramage	28989805
216			People, Medicine	Tom Pashby	http://en.wikipedia.org/wiki/Tom_Pashby	32544502
217			Miscellaneous	Mo	http://en.wikipedia.org/wiki/Mo	37240709
218			Economics	100 Best Global Brands	http://en.wikipedia.org/wiki/100_Best_Global_Brands	36940977
219			Miscellaneous	The Media Institute of Southern Africa	http://en.wikipedia.org/wiki/The_Media_Institute_of_Southern_Africa	27149358
220			History	434	http://en.wikipedia.org/wiki/434	33582014
221			Music	Principles of Lust	http://en.wikipedia.org/wiki/Principles_of_Lust	35232883
222			Biology, Botany	Meineckia	http://en.wikipedia.org/wiki/Meineckia	37646858
223			Geography	St Brendan's Island	http://en.wikipedia.org/wiki/St_Brendan's_Island	36187706
224			Entertainment	Hungry Hungry Hippos	http://en.wikipedia.org/wiki/Hungry_Hungry_Hippos	37278425
225			People	Lee Radziwill	http://en.wikipedia.org/wiki/Lee_Radziwill	38218751
226			People, Entertainment	Gary Perkins	http://en.wikipedia.org/wiki/Gary_Perkins	33969893
227			Geography	Longmont, Colorado	http://en.wikipedia.org/wiki/Longmont,_Colorado	36610620
228			Biology, History	Aurochs	http://en.wikipedia.org/wiki/Aurochs	38845373
229			People, Literature	Assoucy	http://en.wikipedia.org/wiki/Assoucy	37574287
230			Computer Science	ProjectForum	http://en.wikipedia.org/wiki/ProjectForum	29934036
231			Sports	Danish Longball	http://en.wikipedia.org/wiki/Danish_longball	27059972
232			People	Godfrey the Bearded	http://en.wikipedia.org/wiki/Godfrey_the_Bearded	19829119
233			History	Portuguese Colonization of the Americas	http://en.wikipedia.org/wiki/Portuguese_colonization_of_the_Americas	36578547
234			History	Australian Referendum, 1910 (Surplus Revenue)	http://en.wikipedia.org/wiki/Australian_referendum,_1910_(Surplus_Revenue	26850010
235			Technology	Drilling Machine	http://en.wikipedia.org/wiki/Drilling_machine	35140416
236			Entertainment	Here Comes the Grump	http://en.wikipedia.org/wiki/Here_Comes_The_Grump	32395313
237			Mythology	Thiru Nadana Ula	http://en.wikipedia.org/wiki/Thiru_Nadana_Ula	31738726
238			Politics	Foreign Relations of Saint Lucia	http://en.wikipedia.org/wiki/Foreign_relations_of_Saint_Lucia	34913688
239			Geography	San Quintin, Abra	http://en.wikipedia.org/wiki/San_Quintin,_Abra	28043254
240			Entertainment	The Aenar (Enterprise Episode)	http://en.wikipedia.org/wiki/The_Aenar_(Enterprise_episode)	32179849
241			Computer Science	Zapf Dingbats	http://en.wikipedia.org/wiki/Zapf_dingbats	31103834
242			Entertainment	WFTT-TV	http://en.wikipedia.org/wiki/WFTT-TV	36582093
243			Entertainment	Procrastination (SongeBob SquarePants Episode)	http://en.wikipedia.org/wiki/Procrastination_(SpongeBob_SquarePants_episod	38569862
244			Architecture, History	Watchtower (Fortification)	http://en.wikipedia.org/wiki/Watchtower_(fortification)	28269972
245			Sports	International Karate Association	http://en.wikipedia.org/wiki/International_Karate_Association	36330620
246			Science	Friendly Artificial Intelligence	http://en.wikipedia.org/wiki/Friendly_artificial_intelligence	33172210
247			Entertainment	Tammy Layne Winslow	http://en.wikipedia.org/wiki/Tammy_Layne_Winslow	36548817
248			People, Politics	Ulrich Graf von Brockdorff-Rantzau	http://en.wikipedia.org/wiki/Ulrich_Graf_von_Brockdorff-Rantzau	37799116
249			Music	The Mean Reds	http://en.wikipedia.org/wiki/The_Mean_Reds	36986416
250			Physics	Shallow Donor	http://en.wikipedia.org/wiki/Shallow_donor	26402254
251			Entertainment	List of Happy Days Episodes	http://en.wikipedia.org/wiki/List_of_Happy_Days_episodes	30047093
252			Politics	Chief Minister of Gujarat	http://en.wikipedia.org/wiki/Chief_Minister_of_Gujarat	35933096
253			Arts	Mercer Union	http://en.wikipedia.org/wiki/Mercer_Union	37527146
254			People, Medicine	Alfred Downing Fripp (Surgeon)	http://en.wikipedia.org/wiki/Alfred_Downing_Fripp_(surgeon)	34660645
255			Miscellaneous	Catupiry	http://en.wikipedia.org/wiki/Catupiry	36020664
256			Society	Gehenna in Popular Culture	http://en.wikipedia.org/wiki/Gehenna_in_popular_culture	37046217
257			Computer Science	UTF-1	http://en.wikipedia.org/wiki/UTF-1	24411621
258			Arts	Contemporary Baroque Art	http://en.wikipedia.org/wiki/Contemporary_Baroque_Art	32846606
259			Geography	Big Belt Mountains	http://en.wikipedia.org/wiki/Big_Belt_Mountains	32518509
260			Miscellaneous	U of A	http://en.wikipedia.org/wiki/U_of_A	32455903
261			People, Entertainment	Suzanne Rogers	http://en.wikipedia.org/wiki/Suzanne_Rogers	38330065
262			Miscellaneous	Board Foot	http://en.wikipedia.org/wiki/Board_foot	38503164
263			Miscellaneous	Whizzinator	http://en.wikipedia.org/wiki/Whizzinator	30394029
264			Music	San Francisco Tape Music Center	http://en.wikipedia.org/wiki/San_Francisco_Tape_Music_Center	36208985
265			People, Medicine	Barney Clark	http://en.wikipedia.org/wiki/Barney_Clark	35620429
266			Religion	Church of the Gesu	http://en.wikipedia.org/wiki/Church_of_the_Gesu	38620524
267			Geography, History	Hisingen Eastern Hundred	http://en.wikipedia.org/wiki/Hisingen_Eastern_Hundred	37942089
268			Miscellaneous	Scotch College	http://en.wikipedia.org/wiki/Scotch_College	38276583
269			Society, Psychology	Child Discipline	http://en.wikipedia.org/wiki/Child_discipline	34710054
270			History, Politics	John Maclean MA	http://en.wikipedia.org/wiki/John_Maclean_MA	37011049
271		Long Articles	Religion	Prophets of Islam	http://en.wikipedia.org/wiki/Prophets_of_Islam	38346418
272			Politics	State Leaders by Year	http://en.wikipedia.org/wiki/State_leaders_by_year	33946271
273			Politics	Guatemala Constitution	http://en.wikipedia.org/wiki/Guatemala_Constitution	34726472
274			Biology, Philosophy	Race	http://en.wikipedia.org/wiki/Race	38881833
275			History, War	First English Civil War	http://en.wikipedia.org/wiki/First_English_Civil_War	37994356
276			Literature	Italian Literature	http://en.wikipedia.org/wiki/Italian_literature	38834427
277			Literature	The Idler (1758-1760)	http://en.wikipedia.org/wiki/The_Idler_%281758-1760%29	38007160
278			History, War	Continuation War	http://en.wikipedia.org/wiki/Continuation_War	38730042
279			Entertainment	Short-Lived Recurring Characters on Saturday Night Live	http://en.wikipedia.org/wiki/Short-lived_recurring_characters_on_Saturday_N	37920995
280			Philosophy	Friedrich Nietzsche	http://en.wikipedia.org/wiki/Friedrich_Nietzsche	38881785
281		Old Articles	Law	Software Law	http://en.wikipedia.org/wiki/Software_law	16368299
282			History	Czech Lands 880s-1198	http://en.wikipedia.org/wiki/Czech_lands:_880s-1198	16276141
283			Technology	Effective Height	http://en.wikipedia.org/wiki/Effective_height	15936587
284			Terrorism, History	Terrorist Attacks Against Israelis in 2000	http://en.wikipedia.org/wiki/Terrorist_attacks_against_Israelis_in_2000	16303508
285			Technology	Path Quality Analysis	http://en.wikipedia.org/wiki/Path_quality_analysis	15936984
286			Literature	Some Tame Gazelle	http://en.wikipedia.org/wiki/Some_Tame_Gazelle	15994079
287			Literature	The Clandestine Marriage	http://en.wikipedia.org/wiki/The_Clandestine_Marriage	16140455
288			Sports	Opponents Batting Average	http://en.wikipedia.org/wiki/Opponents_batting_average	16503128
289			Sports	Hong Kong Open (Snooker)	http://en.wikipedia.org/wiki/Hong_Kong_Open_%28snooker%29	16489325
290			History, Medicine	History of the French School of Anatomy	http://en.wikipedia.org/wiki/History_of_the_French_school_of_anatomy	38836575
291		Short Articles	Miscellaneous	Neema	http://en.wikipedia.org/wiki/Neema	23530164
292			Medicine	Sipple Syndrome	http://en.wikipedia.org/wiki/Sipple_syndrome	32188907
293			Miscellaneous	Suggestio Falsi	http://en.wikipedia.org/wiki/Suggestio_falsi	36687034
294			Music	Cymbaline	http://en.wikipedia.org/wiki/Cymbaline	32669617
295			Miscellaneous	Webrovka	http://en.wikipedia.org/wiki/Webrovka	31918095
296			Miscellaneous	Suppressio Veri	http://en.wikipedia.org/wiki/Suppressio_veri	36687241
297		Featured Articles	People, Art	Henry Moore	http://en.wikipedia.org/wiki/Henry_Moore	37336243
298			Medicine	Multiple Sclerosis	http://en.wikipedia.org/wiki/Multiple_sclerosis	38880952

#	Class	Sub-Class	Category	Article	Article URL	Article Version
299	Wikipedia "Special" Articles		Chemistry	Ammolite	http://en.wikipedia.org/wiki/Ammolite	37448848
300			Computer Science	Commodore 64	http://en.wikipedia.org/wiki/Commodore_64	38242041
301			Economics	Gold Standard	http://en.wikipedia.org/wiki/Gold_standard	38478230
302			Miscellaneous	Cheese	http://en.wikipedia.org/wiki/Cheese	38817612
303			History, People	Ferdinand Magellan	http://en.wikipedia.org/wiki/Ferdinand_Magellan	38875495
304			Law	Equal Protection Clause	http://en.wikipedia.org/wiki/Equal_Protection_Clause	38702165
305			Mathematics	Eigenvalue, Eigenvector and Eigenspace	http://en.wikipedia.org/wiki/Eigenvalue%2C_eigenvector_and_eigenspace	38533087
306			Music	Sex Pistols	http://en.wikipedia.org/wiki/Sex_Pistols	38779219
307		Articles with Most Revisions	People, Music	Michael Jackson	http://en.wikipedia.org/wiki/Michael_Jackson	38883303
308			Earth Sciences	Hurricane Katrina	http://en.wikipedia.org/wiki/Hurricane_Katrina	38883339
309			People, Politics	John Kerry	http://en.wikipedia.org/wiki/John_Kerry	38871311
310			History, War	World War II	http://en.wikipedia.org/wiki/World_War_II	38882408
311			People, Religion	Pope Benedict XVI	http://en.wikipedia.org/wiki/Pope_Benedict_XVI	38839269
312			Computer Science	Wiki	http://en.wikipedia.org/wiki/Wiki	38857205
313			Geography	Canada	http://en.wikipedia.org/wiki/Canada	38879764
314			Geography	India	http://en.wikipedia.org/wiki/India	38882043
315			People, History	Joseph Stalin	http://en.wikipedia.org/wiki/Joseph_Stalin	38872097
316			Entertainment	Star Wars	http://en.wikipedia.org/wiki/Star_Wars	38878187
317		Most Referenced Articles	Geography	Population Density	http://en.wikipedia.org/wiki/Population_density	38448322
318			Society	Marriage	http://en.wikipedia.org/wiki/Marriage	38863425
319			Geography	Japan	http://en.wikipedia.org/wiki/Japan	38841002
320			Technology	Television	http://en.wikipedia.org/wiki/Television	38676359
321			Biology	Animal	http://en.wikipedia.org/wiki/Animal	38722239
322			Entertainment	Actor	http://en.wikipedia.org/wiki/Actor	38791354
323			Miscellaneous	Latin	http://en.wikipedia.org/wiki/Latin	38748360
324			Science	Mathematics	http://en.wikipedia.org/wiki/Mathematics	38751995
325			Biology	Chordate	http://en.wikipedia.org/wiki/Chordate	37926199
326			History, War	American Civil War	http://en.wikipedia.org/wiki/American_Civil_War	38880177
327		Most Visited Articles	Terrorism, History	World Trade Center	http://en.wikipedia.org/wiki/World_Trade_Center	38868415
328			People, Pornography	Aria Giovanni	http://en.wikipedia.org/wiki/Aria_Giovanni	38567158
329			Science	Western Philosophy	http://en.wikipedia.org/wiki/Western_philosophy	35416997
330			Computer Science	BASIC Programming Language	http://en.wikipedia.org/wiki/BASIC_programming_language	38432624
331			Science	Chemistry	http://en.wikipedia.org/wiki/Chemistry	38848167
332			Entertainment	Nude Celebrities on the Internet	http://en.wikipedia.org/wiki/Nude_celebrities_on_the_Internet	38434727
333			People, History	Thomas Jefferson	http://en.wikipedia.org/wiki/Thomas_Jefferson	38869811
334			Geography	Sweden	http://en.wikipedia.org/wiki/Sweden	38853661
335			Computer Science	NP-Complete	http://en.wikipedia.org/wiki/NP-complete	38583812
336			History, War	World War I	http://en.wikipedia.org/wiki/World_War_I	38861312
337		Articles Lacking Sources	People, Entertainment	Adam Sandler	http://en.wikipedia.org/wiki/Adam_Sandler	38628112
338			Geography	Bethlehem	http://en.wikipedia.org/wiki/Bethlehem	38620086
339			Religion	Iconostasis	http://en.wikipedia.org/wiki/Iconostasis	33287214
340			History, Miscellaneous	La Marseillaise	http://en.wikipedia.org/wiki/La_Marseillaise	38492917
341			Science, Art	Nanoart	http://en.wikipedia.org/wiki/Nanoart	36900971
342			Music	Oasis (Band)	http://en.wikipedia.org/wiki/Oasis_%28band%29	38846457
343			Technology	Radio Frequency	http://en.wikipedia.org/wiki/Radio_frequency	37830416
344			Religion, Mythology	Samara	http://en.wikipedia.org/wiki/Sarama	36075192
345			Economics, Technology	VeriPic	http://en.wikipedia.org/wiki/Veripic	34225025
346			People, History, War	Yukio Seki	http://en.wikipedia.org/wiki/Yukio_Seki	37060688

All Users

#	Article	Article Version	All Users (Registered and Anonymous Combined) # Total Unique (int)	# Total Unique (%)	# Edits total (int)	# Edits total (%)	# Edits per user MIN	# Edits per user MAX	# Edits per user mean	# Edits per User median	# Edits per User mode
1	Boltzmann Constant	37273204	64	100,00%	123	100,00%	1	15	1,92	1	1
2	Force	38360622	183	100,00%	336	100,00%	1	18	1,84	1	1
3	Wave Equation	36341017	56	100,00%	97	100,00%	1	9	1,73	1	1
4	Hilbert's Problems	35970666	77	100,00%	161	100,00%	1	34	2,09	1	1
5	Brown Rat	38375851	72	100,00%	111	100,00%	1	4	1,54	1	1
6	Cotyledon	37805723	24	100,00%	39	100,00%	1	12	1,62	1	1
7	Cretaceous	38683956	97	100,00%	157	100,00%	1	10	1,62	1	1
8	Beijing	38844666	348	100,00%	921	100,00%	1	122	2,65	1	1
9	Nicolaou Taxol Total Synthesis	38135999	9	100,00%	24	100,00%	1	13	2,67	1	1
10	Haloalkane	38017967	82	100,00%	142	100,00%	1	27	1,73	1	1
11	Mach Kernel	37363731	94	100,00%	164	100,00%	1	19	1,74	1	1
12	Turing Test	15926722	7	100,00%	8	100,00%	1	1	1,14	1	1
13	Historical Capitals of China	38032312	64	100,00%	115	100,00%	1	13	1,80	2	1
14	Battle of Caporetto	36893840	34	100,00%	44	100,00%	1	4	1,29	1	1
15	Geometric Isomerism	28568074	28	100,00%	38	100,00%	1	3	1,36	1	1
16	Baking Powder	38638258	43	100,00%	52	100,00%	1	2	1,21	1	1
17	Casein	37736588	51	100,00%	72	100,00%	1	5	1,41	1	1
18	Mode (Statistics)	37625865	16	100,00%	17	100,00%	1	2	1,06	1	1
19	Anemia	38857667	115	100,00%	211	100,00%	1	17	1,83	1	1
20	Nadolol	30876959	9	100,00%	13	100,00%	1	3	1,44	1	1
21	Anencephaly	37205042	14	100,00%	19	100,00%	1	3	1,36	1	1
22	Tattoo	38836429	354	100,00%	807	100,00%	1	124	2,28	1	1
23	Triad	37853283	251	100,00%	586	100,00%	1	29	2,33	1	1
24	Intellectual Property	38841023	292	100,00%	633	100,00%	1	57	2,17	1	1
25	Protectionism	38214150	89	100,00%	133	100,00%	1	4	1,49	1	1
26	Classic Economics	32516940	31	100,00%	40	100,00%	1	2	1,29	1	1
27	Ceteris Paribus	36384037	32	100,00%	46	100,00%	1	5	1,44	1	1
28	Parsimony	33399331	28	100,00%	40	100,00%	1	5	1,43	1	1
29	Gorgidas	38390457	23	100,00%	29	100,00%	1	2	1,26	1	1
30	Meta-Ethics	36343695	66	100,00%	122	100,00%	1	18	1,85	1	1
31	Moral Relativism	37082310	123	100,00%	209	100,00%	1	30	1,70	1	1
32	Transcendental Idealism	38298718	36	100,00%	57	100,00%	1	5	1,58	1	1
33	Debunker	38543378	22	100,00%	59	100,00%	1	10	2,68	1	1
34	Hypnagogic Hallucination	21455394	4	100,00%	4	100,00%	1	1	1,00	1	1
35	Ig Nobel Prize	38619265	84	100,00%	113	100,00%	1	13	1,35	1	1
36	Five-Second Rule	38825167	94	100,00%	142	100,00%	1	15	1,51	1	1
37	Loukoumia	18523754	3	100,00%	3	100,00%	1	1	1,00	1	1
38	Dr. Seuss	38727466	268	100,00%	456	100,00%	1	29	1,70	1	1
39	La Jolla	28655483	4	100,00%	4	100,00%	1	1	1,00	1	1
40	Radiocarbon Dating	38767688	133	100,00%	243	100,00%	1	38	1,83	1	1
41	Decimation	30679352	1	100,00%	2	100,00%	2	2	2,00	2	2
42	Procrastination	38115133	174	100,00%	488	100,00%	1	53	2,80	1	1
43	NASDAQ	38803752	138	100,00%	233	100,00%	1	17	1,69	1	1
44	TLA	38424603	187	100,00%	286	100,00%	1	10	1,53	1	1
45	Bristlecone Pine	38067867	29	100,00%	70	100,00%	1	24	2,41	1	1
46	Kaballah	38766225	359	100,00%	747	100,00%	1	63	2,08	1	1
47	Girona	38248667	71	100,00%	122	100,00%	1	9	1,72	1	1
	Median Values for this Class of Articles		64	100,00%	113	100,00%	1	10	1,69	1	1
	Mean Values for this Class of Articles		93,26	100,00%	181,66	100,00%	1,02	19,21	1,71	1,04	1,02
48	Mobocracy	25524960	9	100,00%	13	100,00%	1	3	1,44	1	1
49	Mechelle	24959419	2	100,00%	3	100,00%	1	2	1,50	1	2
50	Being Julia	32134876	11	100,00%	15	100,00%	1	3	1,36	1	1
51	Hans-Peter Zimmer	36185557	10	100,00%	14	100,00%	1	2	1,40	1	1
52	Pinochle	38009589	56	100,00%	83	100,00%	1	3	1,48	1	1
53	Carbonyldiimidazole	35355940	5	100,00%	6	100,00%	1	2	1,20	1	1
54	Tjurabalan	34108220	6	100,00%	9	100,00%	1	2	1,50	1	2
55	End of Time (Chrono Trigger)	31115398	8	100,00%	13	100,00%	1	5	1,62	1	1
56	Baker Memorial Library	33350095	13	100,00%	21	100,00%	1	3	1,62	1	1
57	French Battleship Redoutable (1876)	32902962	16	100,00%	25	100,00%	1	7	1,56	1	1
58	Gel Electrophoresis	37845418	75	100,00%	110	100,00%	1	8	1,47	1	1
59	Albrecht Altdorfer	37416649	22	100,00%	27	100,00%	1	4	1,23	1	1
60	Rare Earth Elements	16212908	2	100,00%	2	100,00%	1	1	1,00	1	1
61	Dolomites	37996357	42	100,00%	50	100,00%	1	2	1,19	1	1
62	Glow Plug	34815681	19	100,00%	29	100,00%	1	5	1,53	1	1
63	Tahiti	36482714	87	100,00%	136	100,00%	1	6	1,56	1	1
64	Caper	37417739	52	100,00%	64	100,00%	1	3	1,23	1	1
65	Sharia	36716012	293	100,00%	580	100,00%	1	42	1,98	1	1
66	Anarchism	38399178	990	100,00%	6693	100,00%	1	1228	6,76	1	1
67	Alexander von Humboldt	38642551	120	100,00%	181	100,00%	1	7	1,51	1	1
68	Photorespiration	34216375	17	100,00%	27	100,00%	1	1	1,59	1	1
69	Richard Clayderman	38433513	38	100,00%	84	100,00%	1	13	2,21	1	1
70	Rovaniemi	38361937	59	100,00%	96	100,00%	1	7	1,63	1	1
71	Oligarchy	38790907	155	100,00%	235	100,00%	1	11	1,52	1	1
72	Shotokan	38793448	108	100,00%	347	100,00%	1	51	3,21	1	1
73	Stromateidae	37698871	2	100,00%	4	100,00%	1	3	2,00	1	3
74	Easter Islands	20540651	2	100,00%	2	100,00%	1	1	1,00	1	1
75	Nihilism	38780732	361	100,00%	818	100,00%	1	263	2,27	1	1
76	Sinai Peninsula	38133732	118	100,00%	169	100,00%	1	5	1,43	1	1
77	Vaclav Havel	38208911	119	100,00%	162	100,00%	1	7	1,36	1	1
78	Sect	37684399	48	100,00%	63	100,00%	1	8	1,31	1	1
79	Holy Grail	38733545	290	100,00%	509	100,00%	1	59	1,76	1	1
80	Flagellation	37340140	62	100,00%	95	100,00%	1	10	1,53	1	1
81	Ambigram	38211409	96	100,00%	161	100,00%	1	13	1,68	1	1
82	Fatalism	36471303	23	100,00%	33	100,00%	1	4	1,43	1	1
83	Melaka	38748558	34	100,00%	63	100,00%	1	11	1,85	1	1
84	Nanotechnology	38877751	544	100,00%	964	100,00%	1	15	1,77	1	1
85	Pai Mu Tan Tea	23347149	1	100,00%	3	100,00%	3	3	3,00	3	3
86	Manufacturing Consent	37101609	31	100,00%	45	100,00%	1	3	1,45	1	1
87	Special K (Disambiguation)	26447858	6	100,00%	11	100,00%	2	3	1,83	2	2
88	USB Mass Storage Device Class	36986591	33	100,00%	59	100,00%	1	16	1,79	1	1
89	Roman Numerals	38396369	238	100,00%	382	100,00%	1	14	1,61	1	1
90	Damascus	38730007	129	100,00%	270	100,00%	1	41	2,09	1	1
91	Use of Poison Gas in World War I	32081772	1	100,00%	2	100,00%	2	2	2,00	2	2
92	Peanuts	38670130	273	100,00%	545	100,00%	1	11	2,00	1	1
93	Jack London	38825359	216	100,00%	513	100,00%	1	140	2,38	1	1
94	Communication	38046831	204	100,00%	318	100,00%	1	27	1,56	1	1
95	Nervous Tissue	16142015	3	100,00%	3	100,00%	1	1	1,00	1	1
96	The Hobbit	38826711	210	100,00%	361	100,00%	1	20	1,72	1	1
97	Opium War	15958687	5	100,00%	9	100,00%	1	3	1,80	1	1
98	Parkinson's Disease	38735301	237	100,00%	419	100,00%	1	68	1,77	1	1

#	Article	Article Version	All Users (Registered and Anonymous Combined) # Total Unique (int)	# Total Unique (%)	# Edits total (int)	# Edits total (%)	# Edits per user MIN	# Edits per user MAX	# Edits per user mean	# Edits per User median	# Edits per User mode
99	Slayer	38879076	245	100,00%	457	100,00%	1	22	1,87	1	1
	Median Values for this Class of Articles		50	100,00%	74	100,00%	1	7	1,60	1	1
	Mean Values for this Class of Articles		110,50	100,00%	294,29	100,00%	1,08	42,19	1,76	1,08	1,15
100	Methamphetamine	38778939	451,00	100,00%	754,00	100,00%	1,00	29,00	1,67	1,00	1,00
101	Recreational Drug Use	38057817	153,00	100,00%	260,00	100,00%	1,00	17,00	1,70	1,00	1,00
102	Statutory Rape	38698832	69,00	100,00%	129,00	100,00%	1,00	10,00	1,87	1,00	1,00
103	Academic Seduction	27971511	10,00	100,00%	13,00	100,00%	1,00	1,00	1,30	1,00	1,00
104	Sexual Harassment	38714040	102,00	100,00%	244,00	100,00%	1,00	33,00	2,39	1,00	1,00
105	Human Experimentation	38677550	46,00	100,00%	111,00	100,00%	1,00	19,00	2,41	1,00	1,00
106	Abu Ghraib Torture and Prisoner Abuse	38719567	280,00	100,00%	668,00	100,00%	1,00	29,00	2,39	1,00	1,00
107	Bully	38766627	220,00	100,00%	407,00	100,00%	1,00	6,00	1,85	1,00	1,00
108	Spousal Abuse	35241134	38,00	100,00%	53,00	100,00%	1,00	4,00	1,39	1,00	1,00
109	Chickenhawk (sexuality)	38723981	23,00	100,00%	88,00	100,00%	1,00	8,00	3,83	1,00	1,00
110	Child Pornography	38816524	257,00	100,00%	475,00	100,00%	1,00	25,00	1,85	1,00	1,00
111	Shotacon	38394784	114,00	100,00%	154,00	100,00%	1,00	3,00	1,35	1,00	1,00
112	LSD	38872325	734,00	100,00%	1394,00	100,00%	1,00	60,00	1,90	1,00	1,00
113	Psychedelic Mushroom	38714785	202,00	100,00%	320,00	100,00%	1,00	17,00	1,58	1,00	1,00
114	Anti-Psychiatry	38625914	165,00	100,00%	427,00	100,00%	1,00	58,00	2,59	1,00	1,00
115	Necrophilia	38856704	180,00	100,00%	262,00	100,00%	1,00	11,00	1,46	1,00	1,00
116	Petticoat Punishment	38056231	9,00	100,00%	13,00	100,00%	1,00	2,00	1,44	1,00	1,00
117	Anal Masturbation	38198115	59,00	100,00%	78,00	100,00%	1,00	10,00	1,32	1,00	1,00
118	Autoerotic Asphyxiation	37752048	69,00	100,00%	100,00	100,00%	1,00	10,00	1,45	1,00	1,00
119	Fifi (Masturbation Aid)	33282952	2,00	100,00%	3,00	100,00%	1,00	2,00	1,50	1,00	2,00
120	Jeff Stryker	38117120	28,00	100,00%	51,00	100,00%	1,00	9,00	1,82	1,00	1,00
121	Ku Klux Klan	38870644	1362,00	100,00%	3430,00	100,00%	1,00	524,00	2,52	1,00	1,00
122	Gay Rights	38748456	335,00	100,00%	672,00	100,00%	1,00	84,00	2,01	1,00	1,00
123	Militant Islam	35815548	47,00	100,00%	80,00	100,00%	1,00	10,00	1,70	1,00	1,00
124	Scientology	38876995	1497,00	100,00%	3793,00	100,00%	1,00	133,00	2,53	1,00	1,00
125	Operation Snow White	37261477	27,00	100,00%	38,00	100,00%	1,00	6,00	1,41	1,00	1,00
126	Ahmed Yassin	38113877	195,00	100,00%	373,00	100,00%	1,00	28,00	1,91	1,00	1,00
127	Nakba	38206861	29,00	100,00%	54,00	100,00%	1,00	5,00	1,86	1,00	1,00
128	Roswell Incident	21921781	1,00	100,00%	2,00	100,00%	2,00	2,00	2,00	2,00	2,00
129	Jyllands-Posten Muhammad Cartoons Controversy	38876376	1069,00	100,00%	4157,00	100,00%	1,00	165,00	3,89	2,00	1,00
130	Satanism	38839411	428,00	100,00%	758,00	100,00%	1,00	42,00	1,77	1,00	1,00
131	Zoophilia	38208106	326,00	100,00%	1106,00	100,00%	1,00	374,00	3,39	1,00	1,00
132	Prince Albert Piercing	38368233	86,00	100,00%	127,00	100,00%	1,00	14,00	1,48	1,00	1,00
133	Sun Myung Moon	38835843	212,00	100,00%	556,00	100,00%	1,00	153,00	2,62	1,00	1,00
134	New Anti-Semitism	38703805	169,00	100,00%	839,00	100,00%	1,00	87,00	4,96	1,00	1,00
135	Project ARTICHOKE	37927325	13,00	100,00%	33,00	100,00%	1,00	16,00	2,54	1,00	1,00
136	Black Site	38387548	84,00	100,00%	257,00	100,00%	1,00	37,00	3,06	1,00	1,00
137	Guantanamo Bay	38874766	400,00	100,00%	805,00	100,00%	1,00	44,00	2,01	1,00	1,00
138	2003 War on Iraq	24244545	7,00	100,00%	7,00	100,00%	1,00	1,00	1,00	1,00	1,00
139	Euthanasia	38828950	514,00	100,00%	875,00	100,00%	1,00	68,00	1,70	1,00	1,00
140	Suicide	38845887	559,00	100,00%	1018,00	100,00%	1,00	30,00	1,82	1,00	1,00
141	Ted Bundy	38846842	266,00	100,00%	520,00	100,00%	1,00	41,00	1,95	1,00	1,00
142	Heaven's Gate Cult	38769165	105,00	100,00%	169,00	100,00%	1,00	5,00	1,61	1,00	1,00
143	Hezbollah	38830221	363,00	100,00%	1224,00	100,00%	1,00	139,00	3,37	1,00	1,00
144	Al-Aqsa Intifada	38855951	261,00	100,00%	898,00	100,00%	1,00	93,00	3,44	1,00	1,00
145	N44982	37356372	34,00	100,00%	99,00	100,00%	1,00	10,00	2,91	1,00	1,00
146	War on Terror	20213625	3,00	100,00%	3,00	100,00%	1,00	1,00	1,00	1,00	1,00
147	Jafa	37818305	35,00	100,00%	56,00	100,00%	1,00	6,00	1,60	1,00	1,00
148	al-Qaeda	38862411	715,00	100,00%	1567,00	100,00%	1,00	100,00	2,19	1,00	1,00
149	Abortion	38806159	1079,00	100,00%	4034,00	100,00%	1,00	362,00	3,74	1,00	1,00
150	Racism	38766395	952,00	100,00%	2232,00	100,00%	1,00	58,00	2,34	1,00	1,00
151	American Nazi Party	37857558	98,00	100,00%	143,00	100,00%	1,00	9,00	1,46	1,00	1,00
152	Judaism	38605886	830,00	100,00%	2037,00	100,00%	1,00	120,00	2,45	1,00	1,00
153	Necromancy	38049338	132,00	100,00%	226,00	100,00%	1,00	28,00	1,71	1,00	1,00
154	Genetically Modified Food	38813725	219,00	100,00%	432,00	100,00%	1,00	31,00	1,97	1,00	1,00
155	Domestic Violence	38809236	185,00	100,00%	355,00	100,00%	1,00	46,00	1,92	1,00	1,00
156	Holocaust Denial	38869319	572,00	100,00%	1460,00	100,00%	1,00	72,00	2,55	1,00	1,00
157	Prostate Milking	37162347	14,00	100,00%	25,00	100,00%	1,00	9,00	1,79	1,00	1,00
158	Hermetic Order of the Golden Dawn	38877261	109,00	100,00%	212,00	100,00%	1,00	18,00	1,94	1,00	1,00
159	Nuclear Power	38768437	309,00	100,00%	1456,00	100,00%	1,00	242,00	4,71	1,00	1,00
160	Gay Nigger Association of America	38700820	616,00	100,00%	1807,00	100,00%	1,00	130,00	2,93	1,00	1,00
161	Gonorrhea	15956283	5,00	100,00%	7,00	100,00%	1,00	1,00	1,40	1,00	1,00
162	Scabies	37229330	75,00	100,00%	129,00	100,00%	1,00	3,00	1,72	1,00	1,00
163	Whaling	38711832	162,00	100,00%	477,00	100,00%	1,00	99,00	2,94	1,00	1,00
164	Child Sexual Abuse	38835597	104,00	100,00%	274,00	100,00%	1,00	36,00	2,63	1,00	1,00
165	Syphilis	38722440	247,00	100,00%	496,00	100,00%	1,00	36,00	2,01	1,00	1,00
166	The Holocaust	38862374	1578,00	100,00%	3630,00	100,00%	1,00	134,00	2,30	1,00	1,00
167	Apollo Moon Landing is a Hoax	38822817	378,00	100,00%	960,00	100,00%	1,00	38,00	2,54	1,00	1,00
168	Appeasement	38520190	168,00	100,00%	310,00	100,00%	1,00	8,00	1,85	1,00	1,00
169	Jesus	38876713	3002,00	100,00%	10326,00	100,00%	1,00	638,00	3,44	1,00	1,00
170	Adolf Hitler	38749532	3688,00	100,00%	10132,00	100,00%	1,00	479,00	2,75	1,00	1,00
171	United States	38868438	3470,00	100,00%	8785,00	100,00%	1,00	201,00	2,53	1,00	1,00
172	Iraq	38833491	1269,00	100,00%	2368,00	100,00%	1,00	37,00	1,87	1,00	1,00
173	Israel	38872729	1676,00	100,00%	4636,00	100,00%	1,00	211,00	2,77	1,00	1,00
174	Land of Israel	37223158	68,00	100,00%	195,00	100,00%	1,00	48,00	2,87	1,00	1,00
175	Eugenics	38645935	289,00	100,00%	926,00	100,00%	1,00	167,00	3,20	1,00	1,00
	Median Values for this Class of Articles		183	100,00%	390	100,00%	1	31	1,96	1	1
	Mean Values for this Class of Articles		442,72	100,00%	1152,50	100,00%	1,01	76,87	2,23	1,03	1,03
176	France at the 1896 Summer Olympics	34027968	3	100,00%	16	100,00%	1	14	5,33	1	1
177	Prefectures of the Central African Republic	36022243	8	100,00%	13	100,00%	1	3	1,62	1	1
178	NatureServe	27386620	3	100,00%	3	100,00%	1	1	1,00	1	1
179	La Posta Astro-Geophysical Observatory	31006975	3	100,00%	4	100,00%	1	2	1,33	1	1
180	Diosynth	27115196	12	100,00%	15	100,00%	1	2	1,25	1	1
181	Royal Theatre (Victoria)	29485491	2	100,00%	2	100,00%	1	1	1,00	1	1
182	Bitterne Railway Station	32639226	3	100,00%	3	100,00%	1	1	1,00	1	1
183	Albert L'Ouvrier	35283176	8	100,00%	8	100,00%	1	1	1,00	1	1
184	Vernal Pool	30784586	6	100,00%	12	100,00%	1	7	2,00	1	1
185	Navy Commendation Star	36775678	3	100,00%	5	100,00%	1	3	1,67	1	1
186	Enefiok Udo-Obong	37754382	2	100,00%	4	100,00%	1	3	2,00	1	1
187	Long Lake	36966895	13	100,00%	24	100,00%	1	7	1,85	1	1
188	Bernardo de la Paz	37165415	4	100,00%	7	100,00%	1	3	1,75	2	1
189	Harlem Hellfighters	38757486	24	100,00%	66	100,00%	1	11	2,75	1	1
190	Ivan Vasilievich: Back to the Future	32608173	3	100,00%	11	100,00%	1	6	3,67	4	4
191	Mount Angel Abbey	37316069	13	100,00%	25	100,00%	1	4	1,92	1	1
192	Geography of the Philippines	36820786	39	100,00%	75	100,00%	1	15	1,92	1	1
193	Joey Fatone	38451830	27	100,00%	33	100,00%	1	2	1,22	1	1
194	Cephas Washburn	32577018	14	100,00%	21	100,00%	1	6	1,50	1	1
195	Group Dance	36409956	10	100,00%	14	100,00%	1	3	1,40	1	1
196	Two-Mile Borris	37167970	4	100,00%	8	100,00%	1	4	2,00	1	4
197	Afghanistan Timeline July 2001	36876876	7	100,00%	8	100,00%	1	2	1,14	1	1

#	Article	Article Version	All Users (Registered and Anonymous Combined) # Total Unique (int)	# Total Unique (%)	# Edits total (int)	# Edits total (%)	# Edits per user MIN	# Edits per user MAX	# Edits per user mean	# Edits per User median	# Edits per User mode
198	Viscount Chandos	16277634	4	100,00%	6	100,00%	1	2	1,50	2	1
199	Charles X of France	35607312	111	100,00%	170	100,00%	1	5	1,53	1	1
200	Square Pyramidal Number	27573986	13	100,00%	18	100,00%	1	4	1,38	1	1
201	Ludgate Hill	22665154	10	100,00%	12	100,00%	1	2	1,20	1	1
202	Edmonton Journal	38034442	9	100,00%	15	100,00%	1	6	1,67	1	1
203	Rotary International	38631482	80	100,00%	120	100,00%	1	13	1,50	1	1
204	Burningn'n Tree	36704980	11	100,00%	43	100,00%	1	19	3,91	1	1
205	Lev Yashin	37737410	70	100,00%	144	100,00%	1	5	2,06	1	1
206	Philosophical Institute	34009119	4	100,00%	9	100,00%	2	3	2,25	2	2
207	List of Protagonists in Xenosaga	38117658	20	100,00%	97	100,00%	1	50	4,85	1	1
208	Military Unit Mottos: United States	35630475	2	100,00%	7	100,00%	3	4	3,50	3	4
209	Sarafina	37292319	13	100,00%	16	100,00%	1	2	1,23	1	1
210	Aerobatics	37803650	42	100,00%	76	100,00%	1	18	1,81	1	1
211	ITV Evening News	35718828	11	100,00%	14	100,00%	1	1	1,27	1	1
212	Blaarkop	33008328	4	100,00%	6	100,00%	1	2	1,50	2	1
213	Residency (Medicine)	37210564	20	100,00%	29	100,00%	1	4	1,45	1	1
214	Mechitza	36713816	12	100,00%	16	100,00%	1	3	1,33	1	1
215	Ramage	28989805	2	100,00%	2	100,00%	1	1	1,00	1	1
216	Tom Pashby	32544502	6	100,00%	11	100,00%	1	3	1,83	2	1
217	Mo	37240709	26	100,00%	44	100,00%	1	4	1,69	1	1
218	100 Best Global Brands	36940977	14	100,00%	21	100,00%	1	3	1,50	1	1
219	The Media Institute of Southern Africa	27149358	2	100,00%	2	100,00%	1	1	1,00	1	1
220	434	33582014	20	100,00%	32	100,00%	1	4	1,60	1	1
221	Principles of Lust	35232883	6	100,00%	12	100,00%	1	7	2,00	1	1
222	Meineckia	37646858	2	100,00%	3	100,00%	1	2	1,50	1	2
223	St Brendan's Island	36187706	14	100,00%	15	100,00%	1	1	1,07	1	1
224	Hungry Hungry Hippos	37278425	27	100,00%	36	100,00%	1	3	1,33	1	1
225	Lee Radziwill	38218751	22	100,00%	44	100,00%	1	4	2,00	1	1
226	Gary Perkins	33969893	7	100,00%	28	100,00%	1	17	4,00	2	1
227	Longmont, Colorado	36610620	27	100,00%	38	100,00%	1	3	1,41	1	1
228	Aurochs	38845373	103	100,00%	175	100,00%	1	14	1,70	1	1
229	Assoucy	37574287	6	100,00%	7	100,00%	1	2	1,17	1	1
230	ProjectForum	29934036	5	100,00%	5	100,00%	1	1	1,00	1	1
231	Danish Longball	27059972	5	100,00%	11	100,00%	1	3	2,20	1	1
232	Godfrey the Bearded	19829119	2	100,00%	2	100,00%	1	1	1,00	1	1
233	Portuguese Colonization of the Americas	36578547	31	100,00%	39	100,00%	1	6	1,26	1	1
234	Australian Referendum, 1910 (Surplus Revenue)	26850010	5	100,00%	5	100,00%	1	1	1,00	1	1
235	Drilling Machine	35140416	5	100,00%	8	100,00%	1	3	1,60	2	1
236	Here Comes the Grump	32395313	4	100,00%	7	100,00%	1	4	1,75	1	1
237	Thiru Nadana Ula	31738726	2	100,00%	4	100,00%	1	3	2,00	1	1
238	Foreign Relations of Saint Lucia	34913688	13	100,00%	14	100,00%	1	2	1,08	1	1
239	San Quintin, Abra	28043254	2	100,00%	5	100,00%	1	4	2,50	1	1
240	The Aenar (Enterprise Episode)	32179849	3	100,00%	4	100,00%	1	2	1,33	1	1
241	Zapf Dingbats	31103834	2	100,00%	4	100,00%	1	1	2,00	1	1
242	WFTT-TV	36582093	5	100,00%	15	100,00%	1	6	3,00	2	1
243	Procrastination (SongeBob SquarePants Episode)	38569862	11	100,00%	14	100,00%	1	2	1,27	1	1
244	Watchtower (Fortification)	28269972	5	100,00%	8	100,00%	1	4	1,60	1	1
245	International Karate Association	36330620	2	100,00%	9	100,00%	3	6	4,50	3	3
246	Friendly Artificial Intelligence	33172210	32	100,00%	47	100,00%	1	5	1,47	1	1
247	Tammy Layne Winslow	36548817	5	100,00%	17	100,00%	1	13	3,40	1	13
248	Ulrich Graf von Brockdorff-Rantzau	37799116	13	100,00%	15	100,00%	1	2	1,15	1	1
249	The Mean Reds	36986416	5	100,00%	6	100,00%	1	1	1,20	1	1
250	Shallow Donor	26402254	6	100,00%	7	100,00%	1	2	1,17	1	1
251	List of Happy Days Episodes	30047093	7	100,00%	39	100,00%	1	27	5,57	1	1
252	Chief Minister of Gujarat	35933096	5	100,00%	11	100,00%	1	7	2,20	1	1
253	Mercer Union	37527146	3	100,00%	4	100,00%	1	1	1,33	1	1
254	Alfred Downing Fripp (Surgeon)	34660645	2	100,00%	5	100,00%	1	4	2,50	1	1
255	Catupiry	36020664	6	100,00%	9	100,00%	1	2	1,50	1	1
256	Gehenna in Popular Culture	37046217	8	100,00%	11	100,00%	1	4	1,38	1	1
257	UTF-1	24411621	3	100,00%	5	100,00%	1	3	1,67	1	1
258	Contemporary Baroque Art	32846606	8	100,00%	14	100,00%	3	4	1,75	3	3
259	Big Belt Mountains	32518509	4	100,00%	4	100,00%	1	1	1,00	1	1
260	U of A	32455903	4	100,00%	4	100,00%	1	1	1,00	1	1
261	Suzanne Rogers	38330065	13	100,00%	23	100,00%	1	10	1,77	1	1
262	Board Foot	38503164	14	100,00%	21	100,00%	1	3	1,50	1	1
263	Whizzinator	30394029	6	100,00%	8	100,00%	1	1	1,33	1	1
264	San Francisco Tape Music Center	36208985	2	100,00%	7	100,00%	1	6	3,50	1	1
265	Barney Clark	35620429	7	100,00%	10	100,00%	1	3	1,43	1	1
266	Church of the Gesu	38620524	21	100,00%	41	100,00%	1	13	1,95	1	1
267	Hisingen Eastern Hundred	37942089	2	100,00%	2	100,00%	1	1	1,00	1	1
268	Scotch College	38276583	16	100,00%	35	100,00%	1	6	2,19	2	1
269	Child Discipline	34710054	11	100,00%	28	100,00%	1	15	2,55	1	1
270	John Maclean MA	37011049	42	100,00%	75	100,00%	1	17	1,79	1	1
	Median Values for this Class of Articles		7	100,00%	12	100,00%	1	3	1,53	1	1
	Mean Values for this Class of Articles		13,45	100,00%	23,65	100,00%	1,07	5,41	1,85	1,19	1,28
271	Prophets of Islam	38346418	84,00	100,00%	197,00	100,00%	1,00	20,00	2,35	1,00	1,00
272	State Leaders by Year	33946271	29,00	100,00%	91,00	100,00%	1,00	11,00	3,14	2,00	1,00
273	Guatemala Constitution	34726472	3,00	100,00%	3,00	100,00%	1,00	1,00	1,00	1,00	1,00
274	Race	38881833	436,00	100,00%	1684,00	100,00%	1,00	202,00	3,86	1,00	1,00
275	First English Civil War	37994356	41,00	100,00%	211,00	100,00%	1,00	84,00	5,15	1,00	1,00
276	Italian Literature	38834427	41,00	100,00%	90,00	100,00%	1,00	12,00	2,20	1,00	1,00
277	The Idler (1758-1760)	38007160	1,00	100,00%	145,00	100,00%	145,00	145,00	145,00	145,00	145,00
278	Continuation War	38730042	166,00	100,00%	409,00	100,00%	1,00	50,00	2,46	1,00	1,00
279	Short-Lived Recurring Characters on Saturday Night Live	37920995	23,00	100,00%	72,00	100,00%	1,00	19,00	3,13	3,00	1,00
280	Friedrich Nietzsche	38881785	831,00	100,00%	1854,00	100,00%	1,00	88,00	2,23	1,00	1,00
281	Software Law	16368299	2,00	100,00%	3,00	100,00%	2,00	2,00	1,50	2,00	2,00
282	Czech Lands 880s-1198	16276141	4,00	100,00%	5,00	100,00%	1,00	1,00	1,25	1,00	1,00
283	Effective Height	15936587	4,00	100,00%	4,00	100,00%	1,00	1,00	1,00	1,00	1,00
284	Terrorist Attacks Against Israelis in 2000	16303508	3,00	100,00%	4,00	100,00%	1,00	2,00	1,33	1,00	2,00
285	Path Quality Analysis	15936984	4,00	100,00%	4,00	100,00%	1,00	1,00	1,00	1,00	1,00
286	Some Tame Gazelle	15994079	3,00	100,00%	3,00	100,00%	1,00	1,00	1,00	1,00	1,00
287	The Clandestine Marriage	16140455	3,00	100,00%	3,00	100,00%	1,00	1,00	1,00	1,00	1,00
288	Opponents Batting Average	16503128	3,00	100,00%	4,00	100,00%	1,00	2,00	1,33	1,00	1,00
289	Hong Kong Open (Snooker)	16489325	2,00	100,00%	4,00	100,00%	3,00	3,00	2,00	3,00	3,00
290	History of the French School of Anatomy	38836575	7,00	100,00%	15,00	100,00%	1,00	8,00	2,14	1,00	1,00
291	Neema	23530164	2,00	100,00%	2,00	100,00%	1,00	1,00	1,00	1,00	1,00
292	Sipple Syndrome	32188907	3,00	100,00%	3,00	100,00%	1,00	1,00	1,00	1,00	1,00
293	Suggestio Falsi	36687034	2,00	100,00%	3,00	100,00%	1,00	2,00	1,50	1,00	2,00
294	Cymbaline	32669617	2,00	100,00%	2,00	100,00%	1,00	1,00	1,00	1,00	1,00
295	Webrovka	31918095	3,00	100,00%	3,00	100,00%	1,00	1,00	1,00	1,00	1,00
296	Suppressio Veri	36687241	2,00	100,00%	2,00	100,00%	1,00	1,00	1,00	1,00	1,00
297	Henry Moore	37336243	149,00	100,00%	278,00	100,00%	1,00	55,00	1,87	1,00	1,00
298	Multiple Sclerosis	38880952	248,00	100,00%	568,00	100,00%	1,00	62,00	2,29	1,00	1,00

#	Article	Article Version	All Users (Registered and Anonymous Combined) # Total Unique (int)	# Total Unique (%)	# Edits total (int)	# Edits total (%)	# Edits per user MIN	# Edits per user MAX	# Edits per user mean	# Edits per User median	# Edits per User mode
299	Ammolite	37448848	52,00	100,00%	100,00	100,00%	1,00	10,00	1,92	1,00	1,00
300	Commodore 64	38242041	343,00	100,00%	813,00	100,00%	1,00	116,00	2,37	1,00	1,00
301	Gold Standard	38478230	288,00	100,00%	666,00	100,00%	1,00	100,00	2,31	1,00	1,00
302	Cheese	38817612	1126,00	100,00%	2351,00	100,00%	1,00	235,00	2,09	1,00	1,00
303	Ferdinand Magellan	38875495	428,00	100,00%	937,00	100,00%	1,00	95,00	2,19	1,00	1,00
304	Equal Protection Clause	38702165	50,00	100,00%	302,00	100,00%	1,00	232,00	6,04	1,00	1,00
305	Eigenvalue, Eigenvector and Eigenspace	38533087	174,00	100,00%	555,00	100,00%	1,00	32,00	3,19	1,00	1,00
306	Sex Pistols	38779219	316,00	100,00%	673,00	100,00%	1,00	123,00	2,13	1,00	1,00
307	Michael Jackson	38883303	2452,00	100,00%	7507,00	100,00%	1,00	731,00	3,06	1,00	1,00
308	Hurricane Katrina	38883339	3274,00	100,00%	9393,00	100,00%	1,00	233,00	2,87	1,00	1,00
309	John Kerry	38871311	1889,00	100,00%	7174,00	100,00%	1,00	559,00	3,80	1,00	1,00
310	World War II	38882408	2849,00	100,00%	6874,00	100,00%	1,00	250,00	2,41	1,00	1,00
311	Pope Benedict XVI	38839269	2073,00	100,00%	6358,00	100,00%	1,00	192,00	3,07	1,00	1,00
312	Wiki	38857205	2964,00	100,00%	5600,00	100,00%	1,00	135,00	1,89	1,00	1,00
313	Canada	38879764	2052,00	100,00%	5706,00	100,00%	1,00	415,00	2,78	1,00	1,00
314	India	38882043	1733,00	100,00%	4764,00	100,00%	1,00	703,00	2,75	1,00	1,00
315	Joseph Stalin	38872097	1446,00	100,00%	4133,00	100,00%	1,00	235,00	2,86	1,00	1,00
316	Star Wars	38878187	1722,00	100,00%	4131,00	100,00%	1,00	424,00	2,40	1,00	1,00
317	Population Density	38448322	169,00	100,00%	253,00	100,00%	1,00	12,00	1,50	1,00	1,00
318	Marriage	38863425	546,00	100,00%	973,00	100,00%	1,00	43,00	1,78	1,00	1,00
319	Japan	38841002	1487,00	100,00%	3599,00	100,00%	1,00	185,00	2,42	1,00	1,00
320	Television	38676359	722,00	100,00%	1235,00	100,00%	1,00	69,00	1,71	1,00	1,00
321	Animal	38722239	440,00	100,00%	848,00	100,00%	1,00	38,00	1,93	1,00	1,00
322	Actor	38791354	206,00	100,00%	316,00	100,00%	1,00	14,00	1,53	1,00	1,00
323	Latin	38748360	482,00	100,00%	739,00	100,00%	1,00	12,00	1,53	1,00	1,00
324	Mathematics	38751995	828,00	100,00%	1774,00	100,00%	1,00	69,00	2,14	1,00	1,00
325	Chordate	37926199	113,00	100,00%	166,00	100,00%	1,00	9,00	1,47	1,00	1,00
326	American Civil War	38880177	966,00	100,00%	2218,00	100,00%	1,00	111,00	2,30	1,00	1,00
327	World Trade Center	38868415	473,00	100,00%	777,00	100,00%	1,00	27,00	1,64	1,00	1,00
328	Aria Giovanni	38567158	176,00	100,00%	314,00	100,00%	1,00	13,00	1,78	1,00	1,00
329	Western Philosophy	35416997	277,00	100,00%	468,00	100,00%	1,00	15,00	1,69	1,00	1,00
330	BASIC Programming Language	38432624	366,00	100,00%	854,00	100,00%	1,00	126,00	2,33	1,00	1,00
331	Chemistry	38848167	478,00	100,00%	866,00	100,00%	1,00	31,00	1,81	1,00	1,00
332	Nude Celebrities on the Internet	38434727	82,00	100,00%	133,00	100,00%	1,00	8,00	1,62	1,00	1,00
333	Thomas Jefferson	38869811	873,00	100,00%	2008,00	100,00%	1,00	65,00	2,30	1,00	1,00
334	Sweden	38853661	844,00	100,00%	1744,00	100,00%	1,00	112,00	2,07	1,00	1,00
335	NP-Complete	38583812	98,00	100,00%	186,00	100,00%	1,00	22,00	1,90	1,00	1,00
336	World War I	38861312	1709,00	100,00%	3615,00	100,00%	1,00	138,00	2,12	1,00	1,00
337	Adam Sandler	38628112	217,00	100,00%	291,00	100,00%	1,00	9,00	1,34	1,00	1,00
338	Bethlehem	38620086	119,00	100,00%	227,00	100,00%	1,00	16,00	1,91	1,00	1,00
339	Iconostasis	33287214	37,00	100,00%	63,00	100,00%	1,00	8,00	1,70	1,00	1,00
340	La Marseillaise	38492917	152,00	100,00%	231,00	100,00%	1,00	19,00	1,52	1,00	1,00
341	Nanoart	36900971	4,00	100,00%	10,00	100,00%	1,00	4,00	2,50	4,00	1,00
342	Oasis (Band)	38846457	536,00	100,00%	1544,00	100,00%	1,00	308,00	2,88	1,00	1,00
343	Radio Frequency	37830416	117,00	100,00%	177,00	100,00%	1,00	8,00	1,51	1,00	1,00
344	Samara	36075192	4,00	100,00%	5,00	100,00%	1,00	2,00	1,25	1,00	1,00
345	VeriPic	34225025	4,00	100,00%	5,00	100,00%	1,00	2,00	1,25	1,00	1,00
346	Yukio Seki	37060688	3,00	100,00%	4,00	100,00%	1,00	2,00	1,33	1,00	1,00
	Median Values for this Class of Articles		168	100,00%	297	100,00%	1	21	1,93	1	1
	Mean Values for this Class of Articles		524,47	100,00%	1307,18	100,00%	2,93	93,36	3,94	3,01	2,96
	Mean Values for all Classes of Articles		245,42	100,00%	615,68	100,00%	1,46	47,83	2,36	1,52	1,54
	Minimum Value for all Classes of Articles		1	100,00%	2	100,00%	1	1	1,00	1	1
	Maximum Value for all Classes of Articles		3688	100,00%	10326	100,00%	145	1228	145,00	145	145
	Median Value for all Classes of Articles		39	100,00%	75	100,00%	1	9	1,75	1	1

Registered Users

#	Article	Article Version	Registered Users # Total Unique (int)	# Total Unique (%)	# Edits total (int)	# Edits total (%)	# Edits per user MIN	# Edits per user MAX	# Edits per user mean	# Edits per User median	# Edits per User mode
1	Boltzmann Constant	37273204	46	71,88 %	102	82,93 %	1	15	2,22	1	1
2	Force	38360622	106	57,92 %	215	63,99 %	1	18	2,03	1	1
3	Wave Equation	36341017	39	69,64 %	75	77,32 %	1	9	1,92	1	1
4	Hilbert's Problems	35970666	55	71,43 %	134	83,23 %	1	34	2,44	1	1
5	Brown Rat	38375851	47	65,28 %	73	65,77 %	1	4	1,55	1	1
6	Cotyledon	37805723	16	66,67 %	31	79,49 %	1	12	1,94	1	1
7	Cretaceous	38683956	64	65,98 %	113	71,97 %	1	10	1,77	1	1
8	Beijing	38844666	176	50,57 %	633	68,73 %	1	122	3,60	1	1
9	Nicolaou Taxol Total Synthesis	38135999	5	55,56 %	19	79,17 %	1	13	3,80	1	1
10	Haloalkane	38017967	50	60,98 %	104	73,24 %	1	27	2,08	1	1
11	Mach Kernel	37363731	46	48,94 %	99	60,37 %	1	19	2,15	1	1
12	Turing Test	15926722	2	28,57 %	2	25,00 %	1	1	1,00	1	1
13	Historical Capitals of China	38032312	36	56,25 %	71	61,74 %	1	13	1,97	2	1
14	Battle of Caporetto	36893840	22	64,71 %	32	72,73 %	1	4	1,45	1	1
15	Geometric Isomerism	28568074	19	67,86 %	22	57,89 %	1	3	1,16	1	1
16	Baking Powder	38638258	23	53,49 %	28	53,85 %	1	2	1,22	1	1
17	Casein	37736588	31	60,78 %	45	62,50 %	1	5	1,45	1	1
18	Mode (Statistics)	37625865	7	43,75 %	8	47,06 %	1	2	1,14	1	1
19	Anemia	38857667	71	61,74 %	149	70,62 %	1	17	2,10	1	1
20	Nadolol	30876959	6	66,67 %	8	61,54 %	1	3	1,33	1	1
21	Anencephaly	37205042	12	85,71 %	15	78,95 %	1	3	1,25	1	1
22	Tattoo	38836429	152	42,94 %	449	55,64 %	1	124	2,95	1	1
23	Triad	37853283	138	54,98 %	411	70,14 %	1	29	2,98	1	1
24	Intellectual Property	38841023	160	54,79 %	382	60,35 %	1	57	2,39	1	1
25	Protectionism	38214150	46	51,69 %	60	45,11 %	1	4	1,30	1	1
26	Classic Economics	32516940	17	54,84 %	19	47,50 %	1	2	1,12	1	1
27	Ceteris Paribus	36384037	21	65,62 %	29	63,04 %	1	5	1,38	1	1
28	Parsimony	33399331	15	53,57 %	23	57,50 %	1	5	1,53	1	1
29	Gorgidas	38390457	20	86,96 %	24	82,76 %	1	2	1,20	1	1
30	Meta-Ethics	36343695	38	57,58 %	87	71,31 %	1	18	2,29	1	1
31	Moral Relativism	37082310	56	45,53 %	125	59,81 %	1	30	2,23	1	1
32	Transcendental Idealism	38298718	21	58,33 %	37	64,91 %	1	5	1,76	1	1
33	Debunker	38543378	17	77,27 %	39	66,10 %	1	10	2,29	1	1
34	Hypnagogic Hallucination	21455394	3	75,00 %	3	75,00 %	1	1	1,00	1	1
35	Ig Nobel Prize	38619265	60	71,43 %	85	75,22 %	1	13	1,42	1	1
36	Five-Second Rule	38825167	61	64,89 %	104	73,24 %	1	15	1,70	1	1
37	Loukoumia	18523754	3	100,00 %	3	100,00 %	1	1	1,00	1	1
38	Dr. Seuss	38727466	142	52,99 %	273	59,87 %	1	29	1,92	1	1
39	La Jolla	28655483	4	100,00 %	4	100,00 %	1	1	1,00	1	1
40	Radiocarbon Dating	38767688	74	55,64 %	149	61,32 %	1	38	2,01	1	1
41	Decimation	30679352	1	100,00 %	2	100,00 %	2	2	2,00	2	2
42	Procrastination	38115133	80	45,98 %	220	45,08 %	1	53	2,75	1	1
43	NASDAQ	38803752	78	56,52 %	132	56,65 %	1	17	1,69	1	1
44	TLA	38424603	132	70,59 %	219	76,57 %	1	10	1,66	1	1
45	Bristlecone Pine	38067867	18	62,07 %	58	82,86 %	1	24	3,22	1	1
46	Kaballah	38766225	141	39,28 %	425	56,89 %	1	63	3,01	1	1
47	Girona	38248667	43	60,56 %	71	58,20 %	1	9	1,65	1	1
	Median Values for this Class of Articles		39	60,78 %	71	65,77 %	1	10	1,77	1	1
	Mean Values for this Class of Articles		51,49	62,41 %	115,13	67,30 %	1,02	19,21	1,89	1,04	1,02
48	Mobocracy	25524960	7	77,78 %	11	84,62 %	1	3	1,57	1	1
49	Mechelle	24959419	2	100,00 %	3	100,00 %	1	2	1,50	1	2
50	Being Julia	32134876	8	72,73 %	12	80,00 %	1	3	1,50	1	1
51	Hans-Peter Zimmer	36185557	9	90,00 %	11	78,57 %	1	2	1,22	1	1
52	Pinochle	38009589	28	50,00 %	34	40,96 %	1	3	1,21	1	1
53	Carbonyldiimidazole	35355940	5	100,00 %	6	100,00 %	1	2	1,20	1	1
54	Tjurabalan	34108220	6	100,00 %	9	100,00 %	1	2	1,50	1	2
55	End of Time (Chrono Trigger)	31115398	6	75,00 %	11	84,62 %	1	5	1,83	1	1
56	Baker Memorial Library	33350095	7	53,85 %	10	47,62 %	1	3	1,43	1	1
57	French Battleship Redoutable (1876)	32902962	10	62,50 %	19	76,00 %	1	7	1,90	1	1
58	Gel Electrophoresis	37845418	38	50,67 %	66	60,00 %	1	8	1,74	1	1
59	Albrecht Altdorfer	37416649	16	72,73 %	21	77,78 %	1	4	1,31	1	1
60	Rare Earth Elements	16212908	2	100,00 %	2	100,00 %	1	1	1,00	1	1
61	Dolomites	37996357	25	59,52 %	31	62,00 %	1	2	1,24	1	1
62	Glow Plug	34815681	11	57,89 %	18	62,07 %	1	5	1,64	1	1
63	Tahiti	36482714	53	60,92 %	83	61,03 %	1	6	1,57	1	1
64	Caper	37417739	34	65,38 %	43	67,19 %	1	3	1,26	1	1
65	Sharia	36716012	135	46,08 %	333	57,41 %	1	42	2,47	1	1
66	Anarchism	38399178	438	44,24 %	5493	82,07 %	1	1228	12,54	1	1
67	Alexander von Humboldt	38642551	87	72,50 %	139	76,80 %	1	7	1,60	1	1
68	Photorespiration	34216375	12	70,59 %	12	44,44 %	1	1	1,00	1	1
69	Richard Clayderman	38433513	23	60,53 %	53	63,10 %	1	13	2,30	1	1
70	Rovaniemi	38361937	34	57,63 %	55	57,29 %	1	7	1,62	1	1
71	Oligarchy	38790907	76	49,03 %	117	49,79 %	1	11	1,54	1	1
72	Shotokan	38793448	41	37,96 %	159	45,82 %	1	51	3,88	1	1
73	Stromateidae	37698871	2	100,00 %	4	100,00 %	1	3	2,00	1	3
74	Easter Islands	20540651	1	50,00 %	1	50,00 %	1	1	1,00	1	1
75	Nihilism	38780732	155	42,94 %	477	58,31 %	1	263	3,08	1	1
76	Sinai Peninsula	38133732	80	67,80 %	118	69,82 %	1	5	1,48	1	1
77	Vaclav Havel	38208911	76	63,87 %	103	63,58 %	1	7	1,36	1	1
78	Sect	37684399	29	60,42 %	41	65,08 %	1	8	1,41	1	1
79	Holy Grail	38733545	142	48,97 %	293	57,56 %	1	59	2,06	1	1
80	Flagellation	37340140	37	59,68 %	62	65,26 %	1	10	1,68	1	1
81	Ambigram	38211409	47	48,96 %	77	47,83 %	1	13	1,64	1	1
82	Fatalism	36471303	15	65,22 %	19	57,58 %	1	4	1,27	1	1
83	Melaka	38748558	22	64,71 %	42	66,67 %	1	11	1,91	1	1
84	Nanotechnology	38877751	192	35,29 %	359	37,24 %	1	15	1,87	1	1
85	Pai Mu Tan Tea	23347149	1	100,00 %	3	100,00 %	3	3	3,00	3	3
86	Manufacturing Consent	37101609	21	67,74 %	31	68,89 %	1	3	1,48	1	1
87	Special K (Disambiguation)	26447858	3	50,00 %	7	63,64 %	2	3	2,33	2	2
88	USB Mass Storage Device Class	36986591	16	48,48 %	41	69,49 %	1	16	2,56	1	1
89	Roman Numerals	38396369	148	62,18 %	267	69,90 %	1	14	1,80	1	1
90	Damascus	38730007	83	64,34 %	209	77,41 %	1	41	2,52	1	1
91	Use of Poison Gas in World War I	32081772	1	100,00 %	2	100,00 %	2	2	2,00	2	2
92	Peanuts	38670130	148	54,21 %	297	54,50 %	1	11	2,01	1	1
93	Jack London	38825359	120	55,56 %	325	63,35 %	1	140	2,71	1	1
94	Communication	38046831	103	50,49 %	181	56,92 %	1	27	1,76	1	1
95	Nervous Tissue	16142015	1	33,33 %	1	33,33 %	1	1	1,00	1	1
96	The Hobbit	38826711	127	60,48 %	253	70,08 %	1	20	1,99	1	1
97	Opium War	15958687	5	100,00 %	9	100,00 %	1	3	1,80	1	1
98	Parkinson's Disease	38735301	112	47,26 %	232	55,37 %	1	68	2,07	1	1

#	Article	Article Version	Registered Users								
			# Total Unique (int)	# Total Unique (%)	# Edits total (int)	# Edits total (%)	# Edits per user MIN	# Edits per user MAX	# Edits per user mean	# Edits per User median	# Edits per User mode
99	Slayer	38879076	113	46,12 %	237	51,86 %	1	22	2,10	1	1
	Median Values for this Class of Articles		29	60,66 %	43	65,17 %	1	7	1,71	1	1
	Mean Values for this Class of Articles		56,02	64,15 %	200,81	67,94 %	1,08	42,19	1,99	1,08	1,15
100	Methamphetamine	38778939	185,00	41,02 %	390,00	51,72 %	1,00	29,00	2,11	1,00	1,00
101	Recreational Drug Use	38057817	96,00	62,75 %	186,00	71,54 %	1,00	17,00	1,94	1,00	1,00
102	Statutory Rape	38698832	34,00	49,28 %	65,00	50,39 %	1,00	10,00	1,91	1,00	1,00
103	Academic Seduction	27971511	6,00	60,00 %	6,00	46,15 %	1,00	1,00	1,00	1,00	1,00
104	Sexual Harassment	38714040	55,00	53,92 %	129,00	52,87 %	1,00	33,00	2,35	1,00	1,00
105	Human Experimentation	38677550	28,00	60,87 %	75,00	67,57 %	1,00	19,00	2,68	1,00	1,00
106	Abu Ghraib Torture and Prisoner Abuse	38719567	185,00	66,07 %	473,00	70,81 %	1,00	29,00	2,56	1,00	1,00
107	Bully	38766627	97,00	44,09 %	145,00	35,63 %	1,00	6,00	1,49	1,00	1,00
108	Spousal Abuse	35241134	23,00	60,53 %	32,00	60,38 %	1,00	4,00	1,39	1,00	1,00
109	Chickenhawk (sexuality)	38723981	18,00	78,26 %	32,00	36,36 %	1,00	8,00	1,78	1,00	1,00
110	Child Pornography	38816524	139,00	54,09 %	320,00	67,37 %	1,00	25,00	2,30	1,00	1,00
111	Shotacon	38394784	54,00	47,37 %	74,00	48,05 %	1,00	3,00	1,37	1,00	1,00
112	LSD	38872325	307,00	41,83 %	784,00	56,24 %	1,00	60,00	2,55	1,00	1,00
113	Psychedelic Mushroom	38714785	84,00	41,58 %	137,00	42,81 %	1,00	17,00	1,63	1,00	1,00
114	Anti-Psychiatry	38625914	86,00	52,12 %	284,00	66,51 %	1,00	58,00	3,30	1,00	1,00
115	Necrophilia	38856704	89,00	49,44 %	142,00	54,20 %	1,00	11,00	1,60	1,00	1,00
116	Petticoat Punishment	38056231	5,00	55,56 %	6,00	46,15 %	1,00	2,00	1,20	1,00	1,00
117	Anal Masturbation	38198115	32,00	54,24 %	45,00	57,69 %	1,00	10,00	1,41	1,00	1,00
118	Autoerotic Asphyxiation	37752048	38,00	55,07 %	57,00	57,00 %	1,00	10,00	1,50	1,00	1,00
119	Fifi (Masturbation Aid)	33282952	2,00	100,00 %	3,00	100,00 %	1,00	2,00	1,50	1,00	2,00
120	Jeff Stryker	38117120	17,00	60,71 %	33,00	64,71 %	1,00	9,00	1,94	1,00	1,00
121	Ku Klux Klan	38870644	514,00	37,74 %	1972,00	57,49 %	1,00	524,00	3,84	1,00	1,00
122	Gay Rights	38748456	168,00	50,15 %	412,00	61,31 %	1,00	84,00	2,45	1,00	1,00
123	Militant Islam	35815548	33,00	70,21 %	54,00	67,50 %	1,00	10,00	1,64	1,00	1,00
124	Scientology	38876995	549,00	36,67 %	2097,00	55,29 %	1,00	133,00	3,82	1,00	1,00
125	Operation Snow White	37261477	19,00	70,37 %	29,00	76,32 %	1,00	6,00	1,53	1,00	1,00
126	Ahmed Yassin	38113877	109,00	55,90 %	234,00	62,73 %	1,00	28,00	2,15	1,00	1,00
127	Nakba	38206861	23,00	79,31 %	45,00	83,33 %	1,00	5,00	1,96	1,00	1,00
128	Roswell Incident	21921781	1,00	100,00 %	2,00	100,00 %	2,00	2,00	2,00	2,00	2,00
129	Jyllands-Posten Muhammad Cartoons Controversy	38876376	646,00	60,43 %	3263,00	78,49 %	1,00	165,00	5,05	2,00	1,00
130	Satanism	38839411	188,00	43,93 %	387,00	51,06 %	1,00	42,00	2,06	1,00	1,00
131	Zoophilia	38208106	159,00	48,77 %	827,00	74,77 %	1,00	374,00	5,20	1,00	1,00
132	Prince Albert Piercing	38368233	50,00	58,14 %	84,00	66,14 %	1,00	14,00	1,68	1,00	1,00
133	Sun Myung Moon	38835843	110,00	51,89 %	379,00	68,17 %	1,00	153,00	3,45	1,00	1,00
134	New Anti-Semitism	38703805	116,00	68,64 %	746,00	88,92 %	1,00	87,00	6,43	1,00	1,00
135	Project ARTICHOKE	37927325	8,00	61,54 %	27,00	81,82 %	1,00	16,00	3,38	1,00	1,00
136	Black Site	38387548	60,00	71,43 %	197,00	76,65 %	1,00	37,00	3,28	1,00	1,00
137	Guantanamo Bay	38874766	228,00	57,00 %	522,00	64,84 %	1,00	44,00	2,29	1,00	1,00
138	2003 War on Iraq	24244545	7,00	100,00 %	7,00	100,00 %	1,00	1,00	1,00	1,00	1,00
139	Euthanasia	38828950	228,00	44,36 %	442,00	50,51 %	1,00	68,00	1,94	1,00	1,00
140	Suicide	38845887	280,00	50,09 %	621,00	61,00 %	1,00	30,00	2,22	1,00	1,00
141	Ted Bundy	38846842	127,00	47,74 %	287,00	55,19 %	1,00	41,00	2,26	1,00	1,00
142	Heaven's Gate Cult	38769165	60,00	57,14 %	82,00	48,52 %	1,00	5,00	1,37	1,00	1,00
143	Hezbollah	38830221	169,00	46,56 %	806,00	65,85 %	1,00	139,00	4,77	1,00	1,00
144	Al-Aqsa Intifada	38855951	150,00	57,47 %	672,00	74,83 %	1,00	93,00	4,48	1,00	1,00
145	N44982	37356372	22,00	64,71 %	42,00	42,42 %	1,00	10,00	1,91	1,00	1,00
146	War on Terror	20213625	3,00	100,00 %	3,00	100,00 %	1,00	1,00	1,00	1,00	1,00
147	Jafa	37818305	20,00	57,14 %	33,00	58,93 %	1,00	6,00	1,65	1,00	1,00
148	al-Qaeda	38862411	355,00	49,65 %	1032,00	65,86 %	1,00	100,00	2,91	1,00	1,00
149	Abortion	38806159	483,00	44,76 %	2465,00	61,11 %	1,00	362,00	5,10	1,00	1,00
150	Racism	38766395	437,00	45,90 %	1276,00	57,17 %	1,00	58,00	2,92	1,00	1,00
151	American Nazi Party	37857558	57,00	58,16 %	84,00	58,74 %	1,00	9,00	1,47	1,00	1,00
152	Judaism	38605886	348,00	41,93 %	1255,00	61,61 %	1,00	120,00	3,61	1,00	1,00
153	Necromancy	38049338	70,00	53,03 %	128,00	56,64 %	1,00	28,00	1,83	1,00	1,00
154	Genetically Modified Food	38813725	103,00	47,03 %	244,00	56,48 %	1,00	31,00	2,37	1,00	1,00
155	Domestic Violence	38809236	98,00	52,97 %	195,00	54,93 %	1,00	46,00	1,99	1,00	1,00
156	Holocaust Denial	38869319	289,00	50,52 %	974,00	66,71 %	1,00	72,00	3,37	1,00	1,00
157	Prostate Milking	37162347	11,00	78,57 %	21,00	84,00 %	1,00	9,00	1,91	1,00	1,00
158	Hermetic Order of the Golden Dawn	38877261	50,00	45,87 %	111,00	52,36 %	1,00	18,00	2,22	1,00	1,00
159	Nuclear Power	38768437	134,00	43,37 %	1089,00	74,79 %	1,00	242,00	8,13	1,00	1,00
160	Gay Nigger Association of America	38700820	324,00	52,60 %	1303,00	72,11 %	1,00	130,00	4,02	1,00	1,00
161	Gonorrhea	15956283	3,00	60,00 %	3,00	42,86 %	1,00	1,00	1,00	1,00	1,00
162	Scabies	37229330	42,00	56,00 %	51,00	39,53 %	1,00	3,00	1,21	1,00	1,00
163	Whaling	38711832	87,00	53,70 %	366,00	76,73 %	1,00	99,00	4,21	1,00	1,00
164	Child Sexual Abuse	38835597	69,00	66,35 %	211,00	77,01 %	1,00	36,00	3,06	1,00	1,00
165	Syphilis	38722440	134,00	54,25 %	298,00	60,08 %	1,00	36,00	2,22	1,00	1,00
166	The Holocaust	38862374	642,00	40,68 %	2103,00	57,93 %	1,00	134,00	3,28	1,00	1,00
167	Apollo Moon Landing is a Hoax	38822817	210,00	55,56 %	661,00	68,85 %	1,00	38,00	3,15	1,00	1,00
168	Appeasement	38520190	89,00	52,98 %	159,00	51,29 %	1,00	8,00	1,79	1,00	1,00
169	Jesus	38876713	1195,00	39,81 %	6977,00	67,57 %	1,00	638,00	5,84	1,00	1,00
170	Adolf Hitler	38749532	1274,00	34,54 %	6223,00	61,42 %	1,00	479,00	4,88	1,00	1,00
171	United States	38868438	1298,00	37,41 %	5406,00	61,54 %	1,00	201,00	4,16	1,00	1,00
172	Iraq	38833491	536,00	42,24 %	1224,00	51,69 %	1,00	37,00	2,28	1,00	1,00
173	Israel	38872729	642,00	38,31 %	2719,00	58,65 %	1,00	211,00	4,24	1,00	1,00
174	Land of Israel	37223158	41,00	60,29 %	145,00	74,36 %	1,00	48,00	3,54	1,00	1,00
175	Eugenics	38645935	140,00	48,44 %	684,00	73,87 %	1,00	167,00	4,89	1,00	1,00
	Median Values for this Class of Articles		97	53,81 %	223	61,37 %	1	31	2,24	1	1
	Mean Values for this Class of Articles		194,58	55,80 %	724,96	63,45 %	1,01	76,87	2,71	1,03	1,03
176	France at the 1896 Summer Olympics	34027968	3	100,00 %	16	100,00 %	1	14	5,33	1	1
177	Prefectures of the Central African Republic	36022243	7	87,50 %	11	84,62 %	1	3	1,57	1	1
178	NatureServe	27386620	3	100,00 %	3	100,00 %	1	1	1,00	1	1
179	La Posta Astro-Geophysical Observatory	31006975	3	100,00 %	4	100,00 %	1	2	1,33	1	1
180	Diosynth	27115196	10	83,33 %	11	73,33 %	1	2	1,10	1	1
181	Royal Theatre (Victoria)	29485491	2	100,00 %	2	100,00 %	1	1	1,00	1	1
182	Bitterne Railway Station	32639226	3	100,00 %	3	100,00 %	1	1	1,00	1	1
183	Albert L'Ouvrier	35283176	7	87,50 %	7	87,50 %	1	1	1,00	1	1
184	Vernal Pool	30784586	3	50,00 %	9	75,00 %	1	7	3,00	1	1
185	Navy Commendation Star	36775678	3	100,00 %	5	100,00 %	1	3	1,67	1	1
186	Enefiok Udo-Obong	37754382	2	100,00 %	4	100,00 %	1	3	2,00	1	1
187	Long Lake	36966895	13	100,00 %	24	100,00 %	1	7	1,85	1	1
188	Bernardo de la Paz	37165415	4	100,00 %	7	100,00 %	1	3	1,75	2	1
189	Harlem Hellfighters	38757486	17	70,83 %	41	62,12 %	1	11	2,41	1	1
190	Ivan Vasilievich: Back to the Future	32608173	3	100,00 %	11	100,00 %	1	6	3,67	4	4
191	Mount Angel Abbey	37316069	7	53,85 %	11	44,00 %	1	4	1,57	1	1
192	Geography of the Philippines	36820786	26	66,67 %	57	76,00 %	1	15	2,19	1	1
193	Joey Fatone	38451830	15	55,56 %	17	51,52 %	1	2	1,13	1	1
194	Cephas Washburn	32577018	13	92,86 %	19	90,48 %	1	6	1,46	1	1
195	Group Dance	36409956	9	90,00 %	13	92,86 %	1	3	1,44	1	1
196	Two-Mile Borris	37167970	2	50,00 %	5	62,50 %	1	4	2,50	1	4
197	Afghanistan Timeline July 2001	36876876	6	85,71 %	7	87,50 %	1	2	1,17	1	1

#	Article	Article Version	Registered Users # Total Unique (int)	# Total Unique (%)	# Edits total (int)	# Edits total (%)	# Edits per user MIN	# Edits per user MAX	# Edits per user mean	# Edits per User median	# Edits per User mode
198	Viscount Chandos	16277634	4	100,00 %	6	100,00 %	1	2	1,50	2	1
199	Charles X of France	35607312	58	52,25 %	88	51,76 %	1	5	1,52	1	1
200	Square Pyramidal Number	27573986	9	69,23 %	14	77,78 %	1	4	1,56	1	1
201	Ludgate Hill	22665154	9	90,00 %	11	91,67 %	1	2	1,22	1	1
202	Edmonton Journal	38034442	9	100,00 %	15	100,00 %	1	6	1,67	1	1
203	Rotary International	38631482	48	60,00 %	75	62,50 %	1	13	1,56	1	1
204	Burningn'n Tree	36704980	6	54,55 %	34	79,07 %	1	19	5,67	1	1
205	Lev Yashin	37737410	45	64,29 %	63	43,75 %	1	5	1,40	1	1
206	Philosophical Institute	34009119	3	75,00 %	7	77,78 %	2	3	2,33	2	2
207	List of Protagonists in Xenosaga	38117658	10	50,00 %	71	73,20 %	1	50	7,10	1	1
208	Military Unit Mottos: United States	35630475	2	100,00 %	7	100,00 %	3	4	3,50	3	4
209	Sarafina	37292319	5	38,46 %	7	43,75 %	1	2	1,40	1	1
210	Aerobatics	37803650	26	61,90 %	50	65,79 %	1	18	1,92	1	1
211	ITV Evening News	35718828	6	54,55 %	6	42,86 %	1	1	1,00	1	1
212	Blaarkop	33008328	4	100,00 %	6	100,00 %	1	2	1,50	2	1
213	Residency (Medicine)	37210564	17	85,00 %	26	89,66 %	1	4	1,53	1	1
214	Mechitza	36713816	9	75,00 %	13	81,25 %	1	3	1,44	1	1
215	Ramage	28989805	2	100,00 %	2	100,00 %	1	1	1,00	1	1
216	Tom Pashby	32544502	4	66,67 %	7	63,64 %	1	3	1,75	2	1
217	Mo	37240709	16	61,54 %	30	68,18 %	1	4	1,88	1	1
218	100 Best Global Brands	36940977	9	64,29 %	14	66,67 %	1	3	1,56	1	1
219	The Media Institute of Southern Africa	27149358	2	100,00 %	2	100,00 %	1	1	1,00	1	1
220	434	33582014	17	85,00 %	29	90,62 %	1	4	1,71	1	1
221	Principles of Lust	35232883	5	83,33 %	11	91,67 %	1	7	2,20	1	1
222	Meineckia	37646858	2	100,00 %	3	100,00 %	1	2	1,50	1	2
223	St Brendan's Island	36187706	7	50,00 %	7	46,67 %	1	1	1,00	1	1
224	Hungry Hungry Hippos	37278425	16	59,26 %	21	58,33 %	1	3	1,31	1	1
225	Lee Radziwill	38218751	13	59,09 %	22	50,00 %	1	4	1,69	1	1
226	Gary Perkins	33969893	5	71,43 %	24	85,71 %	1	17	4,80	2	1
227	Longmont, Colorado	36610620	16	59,26 %	20	52,63 %	1	3	1,25	1	1
228	Aurochs	38845373	69	66,99 %	127	72,57 %	1	14	1,84	1	1
229	Assoucy	37574287	6	100,00 %	7	100,00 %	1	2	1,17	1	1
230	ProjectForum	29934036	2	40,00 %	2	40,00 %	1	1	1,00	1	1
231	Danish Longball	27059972	4	80,00 %	6	54,55 %	1	3	1,50	1	1
232	Godfrey the Bearded	19829119	2	100,00 %	2	100,00 %	1	1	1,00	1	1
233	Portuguese Colonization of the Americas	36578547	21	67,74 %	29	74,36 %	1	6	1,38	1	1
234	Australian Referendum, 1910 (Surplus Revenue)	26850010	5	100,00 %	5	100,00 %	1	1	1,00	1	1
235	Drilling Machine	35140416	4	80,00 %	7	87,50 %	1	3	1,75	2	1
236	Here Comes the Grump	32395313	4	100,00 %	7	100,00 %	1	4	1,75	1	1
237	Thiru Nadana Ula	31738726	2	100,00 %	4	100,00 %	1	3	2,00	1	1
238	Foreign Relations of Saint Lucia	34913688	12	92,31 %	13	92,86 %	1	2	1,08	1	1
239	San Quintin, Abra	28043254	2	100,00 %	5	100,00 %	1	4	2,50	1	1
240	The Aenar (Enterprise Episode)	32179849	3	100,00 %	4	100,00 %	1	2	1,33	1	1
241	Zapf Dingbats	31103834	1	50,00 %	1	25,00 %	1	1	1,00	1	1
242	WFTT-TV	36582093	5	100,00 %	15	100,00 %	1	6	3,00	2	1
243	Procrastination (SongeBob SquarePants Episode)	38569862	8	72,73 %	9	64,29 %	1	2	1,12	1	1
244	Watchtower (Fortification)	28269972	5	100,00 %	8	100,00 %	1	4	1,60	1	1
245	International Karate Association	36330620	2	100,00 %	9	100,00 %	3	6	4,50	3	3
246	Friendly Artificial Intelligence	33172210	18	56,25 %	25	53,19 %	1	5	1,39	1	1
247	Tammy Layne Winslow	36548817	2	40,00 %	14	82,35 %	1	13	7,00	1	13
248	Ulrich Graf von Brockdorff-Rantzau	37799116	11	84,62 %	12	80,00 %	1	2	1,09	1	1
249	The Mean Reds	36986416	1	20,00 %	1	16,67 %	1	1	1,00	1	1
250	Shallow Donor	26402254	5	83,33 %	6	85,71 %	1	2	1,20	1	1
251	List of Happy Days Episodes	30047093	3	42,86 %	29	74,36 %	1	27	9,67	1	1
252	Chief Minister of Gujarat	35933096	4	80,00 %	10	90,91 %	1	7	2,50	1	1
253	Mercer Union	37527146	2	66,67 %	2	50,00 %	1	1	1,00	1	1
254	Alfred Downing Fripp (Surgeon)	34660645	2	100,00 %	5	100,00 %	1	4	2,50	1	1
255	Catupiry	36020664	4	66,67 %	5	55,56 %	1	2	1,25	1	1
256	Gehenna in Popular Culture	37046217	5	62,50 %	8	72,73 %	1	4	1,60	1	1
257	UTF-1	24411621	3	100,00 %	5	100,00 %	1	3	1,67	1	1
258	Contemporary Baroque Art	32846606	2	25,00 %	7	50,00 %	3	4	3,50	3	3
259	Big Belt Mountains	32518509	4	100,00 %	4	100,00 %	1	1	1,00	1	1
260	U of A	32455903	3	75,00 %	3	75,00 %	1	1	1,00	1	1
261	Suzanne Rogers	38330065	8	61,54 %	18	78,26 %	1	10	2,25	1	1
262	Board Foot	38503164	11	78,57 %	18	85,71 %	1	3	1,64	1	1
263	Whizzinator	30394029	4	66,67 %	4	50,00 %	1	1	1,00	1	1
264	San Francisco Tape Music Center	36208985	2	100,00 %	7	100,00 %	1	6	3,50	1	1
265	Barney Clark	35620429	5	71,43 %	7	70,00 %	1	3	1,40	1	1
266	Church of the Gesu	38620524	14	66,67 %	34	82,93 %	1	13	2,43	1	1
267	Hisingen Eastern Hundred	37942089	2	100,00 %	2	100,00 %	1	1	1,00	1	1
268	Scotch College	38276583	12	75,00 %	28	80,00 %	1	6	2,33	2	1
269	Child Discipline	34710054	10	90,91 %	26	92,86 %	1	15	2,60	1	1
270	John Maclean MA	37011049	25	59,52 %	55	73,33 %	1	17	2,20	1	1
	Median Values for this Class of Articles		5	80,00 %	9	84,62 %	1	3	1,56	1	1
	Mean Values for this Class of Articles		9,09	77,76 %	16,56	79,51 %	1,07	5,41	2,00	1,19	1,28
271	Prophets of Islam	38346418	54,00	64,29 %	162,00	82,23 %	1,00	20,00	3,00	1,00	1,00
272	State Leaders by Year	33946271	26,00	89,66 %	82,00	90,11 %	1,00	11,00	3,15	2,00	1,00
273	Guatemala Constitution	34726472	3,00	100,00 %	3,00	100,00 %	1,00	1,00	1,00	1,00	1,00
274	Race	38881833	240,00	55,05 %	1380,00	81,95 %	1,00	202,00	5,75	1,00	1,00
275	First English Civil War	37994356	34,00	82,93 %	198,00	93,84 %	1,00	84,00	5,82	1,00	1,00
276	Italian Literature	38834427	34,00	82,93 %	80,00	88,89 %	1,00	12,00	2,35	1,00	1,00
277	The Idler (1758-1760)	38007160	1,00	100,00 %	145,00	100,00 %	145,00	145,00	145,00	145,00	145,00
278	Continuation War	38730042	102,00	61,45 %	306,00	74,82 %	1,00	50,00	3,00	1,00	1,00
279	Short-Lived Recurring Characters on Saturday Night Live	37920995	8,00	34,78 %	45,00	62,50 %	1,00	19,00	5,62	3,00	1,00
280	Friedrich Nietzsche	38881785	335,00	40,31 %	1062,00	57,28 %	1,00	88,00	3,17	1,00	1,00
281	Software Law	16368299	1,00	50,00 %	2,00	66,67 %	2,00	2,00	2,00	2,00	2,00
282	Czech Lands 880s-1198	16276141	2,00	50,00 %	2,00	40,00 %	1,00	1,00	1,00	1,00	1,00
283	Effective Height	15936587	2,00	50,00 %	2,00	50,00 %	1,00	1,00	1,00	1,00	1,00
284	Terrorist Attacks Against Israelis in 2000	16303508	2,00	66,67 %	3,00	75,00 %	1,00	2,00	1,50	1,00	2,00
285	Path Quality Analysis	15936984	2,00	50,00 %	2,00	50,00 %	1,00	1,00	1,00	1,00	1,00
286	Some Tame Gazelle	15994079	2,00	66,67 %	2,00	66,67 %	1,00	1,00	1,00	1,00	1,00
287	The Clandestine Marriage	16140455	2,00	66,67 %	2,00	66,67 %	1,00	1,00	1,00	1,00	1,00
288	Opponents Batting Average	16503128	3,00	100,00 %	4,00	100,00 %	1,00	2,00	1,33	1,00	1,00
289	Hong Kong Open (Snooker)	16489325	1,00	50,00 %	3,00	75,00 %	3,00	3,00	3,00	3,00	3,00
290	History of the French School of Anatomy	38836575	6,00	85,71 %	14,00	93,33 %	1,00	8,00	2,33	1,00	1,00
291	Neema	23530164	1,00	50,00 %	1,00	50,00 %	1,00	1,00	1,00	1,00	1,00
292	Sipple Syndrome	32188907	2,00	66,67 %	2,00	66,67 %	1,00	1,00	1,00	1,00	1,00
293	Suggestio Falsi	36687034	2,00	100,00 %	3,00	100,00 %	1,00	2,00	1,50	1,00	2,00
294	Cymbaline	32669617	1,00	50,00 %	1,00	50,00 %	1,00	1,00	1,00	1,00	1,00
295	Webrovka	31918095	1,00	33,33 %	1,00	33,33 %	1,00	1,00	1,00	1,00	1,00
296	Suppressio Veri	36687241	2,00	100,00 %	2,00	100,00 %	1,00	1,00	1,00	1,00	1,00
297	Henry Moore	37336243	80,00	53,69 %	174,00	62,59 %	1,00	55,00	2,17	1,00	1,00
298	Multiple Sclerosis	38880952	115,00	46,37 %	375,00	66,02 %	1,00	62,00	3,26	1,00	1,00

#	Article	Article Version	Registered Users								
			# Total Unique (int)	# Total Unique (%)	# Edits total (int)	# Edits total (%)	# Edits per user MIN	# Edits per user MAX	# Edits per user mean	# Edits per User median	# Edits per User mode
299	Ammolite	37448848	30,00	57,69 %	63,00	63,00 %	1,00	10,00	2,10	1,00	1,00
300	Commodore 64	38242041	165,00	48,10 %	454,00	55,84 %	1,00	116,00	2,75	1,00	1,00
301	Gold Standard	38478230	147,00	51,04 %	458,00	68,77 %	1,00	100,00	3,12	1,00	1,00
302	Cheese	38817612	419,00	37,21 %	1265,00	53,81 %	1,00	235,00	3,02	1,00	1,00
303	Ferdinand Magellan	38875495	186,00	43,46 %	505,00	53,90 %	1,00	95,00	2,72	1,00	1,00
304	Equal Protection Clause	38702165	28,00	56,00 %	270,00	89,40 %	1,00	232,00	9,64	1,00	1,00
305	Eigenvalue, Eigenvector and Eigenspace	38533087	96,00	55,17 %	280,00	50,45 %	1,00	32,00	2,92	1,00	1,00
306	Sex Pistols	38779219	157,00	49,68 %	373,00	55,42 %	1,00	123,00	2,38	1,00	1,00
307	Michael Jackson	38883303	822,00	33,52 %	4290,00	57,15 %	1,00	731,00	5,22	1,00	1,00
308	Hurricane Katrina	38883339	1062,00	32,44 %	5068,00	53,96 %	1,00	233,00	4,77	1,00	1,00
309	John Kerry	38871311	710,00	37,59 %	4747,00	66,17 %	1,00	559,00	6,69	1,00	1,00
310	World War II	38882408	998,00	35,03 %	3810,00	55,43 %	1,00	250,00	3,82	1,00	1,00
311	Pope Benedict XVI	38839269	767,00	37,00 %	3830,00	60,24 %	1,00	192,00	4,99	1,00	1,00
312	Wiki	38857205	662,00	22,33 %	2316,00	41,36 %	1,00	135,00	3,50	1,00	1,00
313	Canada	38879764	793,00	38,65 %	3607,00	63,21 %	1,00	415,00	4,55	1,00	1,00
314	India	38882043	651,00	37,56 %	3015,00	63,29 %	1,00	703,00	4,63	1,00	1,00
315	Joseph Stalin	38872097	612,00	42,32 %	2735,00	66,17 %	1,00	235,00	4,47	1,00	1,00
316	Star Wars	38878187	660,00	38,33 %	2297,00	55,60 %	1,00	424,00	3,48	1,00	1,00
317	Population Density	38448322	94,00	55,62 %	153,00	60,47 %	1,00	12,00	1,63	1,00	1,00
318	Marriage	38863425	247,00	45,24 %	525,00	53,96 %	1,00	43,00	2,13	1,00	1,00
319	Japan	38841002	584,00	39,27 %	1969,00	54,71 %	1,00	185,00	3,37	1,00	1,00
320	Television	38676359	362,00	50,14 %	696,00	56,36 %	1,00	69,00	1,92	1,00	1,00
321	Animal	38722239	234,00	53,18 %	540,00	63,68 %	1,00	38,00	2,31	1,00	1,00
322	Actor	38791354	131,00	63,59 %	214,00	67,72 %	1,00	14,00	1,63	1,00	1,00
323	Latin	38748360	252,00	52,28 %	414,00	56,02 %	1,00	12,00	1,64	1,00	1,00
324	Mathematics	38751995	385,00	46,50 %	1145,00	64,54 %	1,00	69,00	2,97	1,00	1,00
325	Chordate	37926199	79,00	69,91 %	120,00	72,29 %	1,00	9,00	1,52	1,00	1,00
326	American Civil War	38880177	398,00	41,20 %	1342,00	60,50 %	1,00	111,00	3,37	1,00	1,00
327	World Trade Center	38868415	242,00	51,16 %	444,00	57,14 %	1,00	27,00	1,83	1,00	1,00
328	Aria Giovanni	38567158	86,00	48,86 %	162,00	51,59 %	1,00	13,00	1,88	1,00	1,00
329	Western Philosophy	35416997	146,00	52,71 %	290,00	61,97 %	1,00	15,00	1,99	1,00	1,00
330	BASIC Programming Language	38432624	177,00	48,36 %	579,00	67,80 %	1,00	126,00	3,27	1,00	1,00
331	Chemistry	38848167	243,00	50,84 %	486,00	56,12 %	1,00	31,00	2,00	1,00	1,00
332	Nude Celebrities on the Internet	38434727	46,00	56,10 %	93,00	69,92 %	1,00	8,00	2,02	1,00	1,00
333	Thomas Jefferson	38869811	402,00	46,05 %	1198,00	59,66 %	1,00	65,00	2,98	1,00	1,00
334	Sweden	38853661	349,00	41,35 %	981,00	56,25 %	1,00	112,00	2,81	1,00	1,00
335	NP-Complete	38583812	56,00	57,14 %	125,00	67,20 %	1,00	22,00	2,23	1,00	1,00
336	World War I	38861312	651,00	38,09 %	1996,00	55,21 %	1,00	138,00	3,07	1,00	1,00
337	Adam Sandler	38628112	101,00	46,54 %	131,00	45,02 %	1,00	9,00	1,30	1,00	1,00
338	Bethlehem	38620086	75,00	63,03 %	171,00	75,33 %	1,00	16,00	2,28	1,00	1,00
339	Iconostasis	33287214	26,00	70,27 %	48,00	76,19 %	1,00	8,00	1,85	1,00	1,00
340	La Marseillaise	38492917	78,00	51,32 %	144,00	62,34 %	1,00	19,00	1,85	1,00	1,00
341	Nanoart	36900971	4,00	100,00 %	10,00	100,00 %	1,00	4,00	2,50	4,00	1,00
342	Oasis (Band)	38846457	198,00	36,94 %	859,00	55,63 %	1,00	308,00	4,34	1,00	1,00
343	Radio Frequency	37830416	63,00	53,85 %	101,00	57,06 %	1,00	8,00	1,60	1,00	1,00
344	Samara	36075192	3,00	75,00 %	4,00	80,00 %	1,00	2,00	1,33	1,00	1,00
345	VeriPic	34225025	3,00	75,00 %	4,00	80,00 %	1,00	2,00	1,33	1,00	1,00
346	Yukio Seki	37060688	3,00	100,00 %	4,00	100,00 %	1,00	2,00	1,33	1,00	1,00
	Median Values for this Class of Articles		90	50,94 %	186	63,11 %	1	21	2,32	1	1
	Mean Values for this Class of Articles		207,20	56,31 %	768,28	66,60 %	2,93	93,36	4,55	3,01	2,96
	Mean Values for all Classes of Articles		106,16	64,09 %	378,36	69,75 %	1,46	47,83	2,70	1,52	1,54
	Minimum Value for all Classes of Articles		1	20,00 %	1	16,67 %	1	1	1,00	1	1
	Maximum Value for all Classes of Articles		1298	100,00 %	6977	100,00 %	145	1228	145,00	145	145
	Median Value for all Classes of Articles		25	59,84 %	45	66,67 %	1	9	1,89	1	1

Anonymous Users

#	Article	Article Version	Anonymous Users # Total Unique (int)	# Total Unique (%)	# Edits total (int)	# Edits total (%)	# Edits per user MIN	# Edits per user MAX	# Edits per user mean	# Edits per User median	# Edits per User mode
1	Boltzmann Constant	37273204	18	28,12 %	21	17,07 %	1	2	1,17	1	1
2	Force	38360622	77	42,08 %	121	36,01 %	1	9	1,57	1	1
3	Wave Equation	36341017	17	30,36 %	22	22,68 %	1	3	1,29	1	1
4	Hilbert's Problems	35970666	22	28,57 %	27	16,77 %	1	3	1,23	1	1
5	Brown Rat	38375851	25	34,72 %	38	34,23 %	1	5	1,52	1	1
6	Cotyledon	37805723	8	33,33 %	8	20,51 %	1	1	1,00	1	1
7	Cretaceous	38683956	33	34,02 %	44	28,03 %	1	4	1,33	1	1
8	Beijing	38844666	172	49,43 %	288	31,27 %	1	10	1,67	1	1
9	Nicolaou Taxol Total Synthesis	38135999	4	44,44 %	5	20,83 %	1	2	1,25	1	1
10	Haloalkane	38017967	32	39,02 %	38	26,76 %	1	2	1,19	1	1
11	Mach Kernel	37363731	48	51,06 %	65	39,63 %	1	6	1,35	1	1
12	Turing Test	15926722	5	71,43 %	6	75,00 %	1	2	1,20	1	1
13	Historical Capitals of China	38032312	28	43,75 %	44	38,26 %	1	4	1,57	1	1
14	Battle of Caporetto	36893840	12	35,29 %	12	27,27 %	1	1	1,00	1	1
15	Geometric Isomerism	28568074	9	32,14 %	16	42,11 %	1	6	1,78	1	1
16	Baking Powder	38638258	20	46,51 %	24	46,15 %	1	2	1,20	1	1
17	Casein	37736588	20	39,22 %	27	37,50 %	1	3	1,35	1	1
18	Mode (Statistics)	37625865	9	56,25 %	9	52,94 %	1	1	1,00	1	1
19	Anemia	38857667	44	38,26 %	62	29,38 %	1	5	1,41	1	1
20	Nadolol	30876959	3	33,33 %	5	38,46 %	1	2	1,67	2	2
21	Anencephaly	37205042	2	14,29 %	4	21,05 %	1	3	2,00	1	1
22	Tattoo	38836429	202	57,06 %	358	44,36 %	1	16	1,77	1	1
23	Triad	37853283	113	45,02 %	175	29,86 %	1	7	1,55	1	1
24	Intellectual Property	38841023	132	45,21 %	251	39,65 %	1	22	1,90	1	1
25	Protectionism	38214150	43	48,31 %	73	54,89 %	1	5	1,70	1	1
26	Classic Economics	32516940	14	45,16 %	21	52,50 %	1	3	1,50	1	1
27	Ceteris Paribus	36384037	11	34,38 %	17	36,96 %	1	3	1,55	1	1
28	Parsimony	33399331	13	46,43 %	17	42,50 %	1	2	1,31	1	1
29	Gorgidas	38390457	3	13,04 %	5	17,24 %	1	2	1,67	2	2
30	Meta-Ethics	36343695	28	42,42 %	35	28,69 %	1	4	1,25	1	1
31	Moral Relativism	37082310	67	54,47 %	84	40,19 %	1	4	1,25	1	1
32	Transcendental Idealism	38298718	15	41,67 %	20	35,09 %	1	3	1,33	1	1
33	Debunker	38543378	5	22,73 %	20	33,90 %	1	15	4,00	1	1
34	Hypnagogic Hallucination	21455394	1	25,00 %	1	25,00 %	1	1	1,00	1	1
35	Ig Nobel Prize	38619265	24	28,57 %	28	24,78 %	1	4	1,17	1	1
36	Five-Second Rule	38825167	33	35,11 %	38	26,76 %	1	3	1,15	1	1
37	Loukoumia	18523754	0	0,00 %	0	0,00 %	n/a	n/a	n/a	n/a	n/a
38	Dr. Seuss	38727466	126	47,01 %	183	40,13 %	1	8	1,45	1	1
39	La Jolla	28655483	0	0,00 %	0	0,00 %	n/a	n/a	n/a	n/a	n/a
40	Radiocarbon Dating	38767688	59	44,36 %	94	38,68 %	1	6	1,59	1	1
41	Decimation	30679352	0	0,00 %	0	0,00 %	n/a	n/a	n/a	n/a	n/a
42	Procrastination	38115133	94	54,02 %	268	54,92 %	1	134	2,85	1	1
43	NASDAQ	38803752	60	43,48 %	101	43,35 %	1	13	1,68	1	1
44	TLA	38424603	55	29,41 %	67	23,43 %	1	4	1,22	1	1
45	Bristlecone Pine	38067867	11	37,93 %	12	17,14 %	1	2	1,09	1	1
46	Kaballah	38766225	218	60,72 %	322	43,11 %	1	8	1,48	1	1
47	Girona	38248667	28	39,44 %	51	41,80 %	1	5	1,82	1	1
	Median Values for this Class of Articles		22	39,22 %	27	34,23 %	1	4	1,38	1	1
	Mean Values for this Class of Articles		41,77	37,59 %	66,53	32,70 %	1,00	7,95	1,50	1,05	1,05
48	Mobocracy	25524960	2	22,22 %	2	15,38 %	1	1	1,00	1	1
49	Mechelle	24959419	0	0,00 %	0	0,00 %	n/a	n/a	n/a	n/a	n/a
50	Being Julia	32134876	3	27,27 %	3	20,00 %	1	1	1,00	1	1
51	Hans-Peter Zimmer	36185557	1	10,00 %	3	21,43 %	3	3	3,00	3	3
52	Pinochle	38009589	28	50,00 %	49	59,04 %	1	7	1,75	1	1
53	Carbonyldiimidazole	35355940	0	0,00 %	0	0,00 %	n/a	n/a	n/a	n/a	n/a
54	Tjurabalan	34108220	0	0,00 %	0	0,00 %	n/a	n/a	n/a	n/a	n/a
55	End of Time (Chrono Trigger)	31115398	2	25,00 %	2	15,38 %	1	1	1,00	1	1
56	Baker Memorial Library	33350095	6	46,15 %	11	52,38 %	1	6	1,83	1	1
57	French Battleship Redoutable (1876)	32902962	6	37,50 %	6	24,00 %	1	1	1,00	1	1
58	Gel Electrophoresis	37845418	37	49,33 %	44	40,00 %	1	3	1,19	1	1
59	Albrecht Altdorfer	37416649	6	27,27 %	6	22,22 %	1	1	1,00	1	1
60	Rare Earth Elements	16212908	0	0,00 %	0	0,00 %	n/a	n/a	n/a	n/a	n/a
61	Dolomites	37996357	17	40,48 %	19	38,00 %	1	2	1,12	1	1
62	Glow Plug	34815681	8	42,11 %	11	37,93 %	1	3	1,38	1	1
63	Tahiti	36482714	34	39,08 %	53	38,97 %	1	8	1,56	1	1
64	Caper	37417739	18	34,62 %	21	32,81 %	1	3	1,17	1	1
65	Sharia	36716012	158	53,92 %	247	42,59 %	1	11	1,56	1	1
66	Anarchism	38399178	552	55,76 %	1200	17,93 %	1	49	2,17	1	1
67	Alexander von Humboldt	38642551	33	27,50 %	42	23,20 %	1	3	1,27	1	1
68	Photorespiration	34216375	5	29,41 %	15	55,56 %	1	9	3,00	2	2
69	Richard Clayderman	38433513	15	39,47 %	31	36,90 %	1	10	2,07	1	1
70	Rovaniemi	38361937	25	42,37 %	41	42,71 %	1	5	1,64	1	1
71	Oligarchy	38790907	79	50,97 %	118	50,21 %	1	7	1,49	1	1
72	Shotokan	38793448	67	62,04 %	188	54,18 %	1	27	2,81	1	1
73	Stromateidae	37698871	0	0,00 %	0	0,00 %	n/a	n/a	n/a	n/a	n/a
74	Easter Islands	20540651	1	50,00 %	1	50,00 %	1	1	1,00	1	1
75	Nihilism	38780732	206	57,06 %	341	41,69 %	1	45	1,66	1	1
76	Sinai Peninsula	38133732	38	32,20 %	51	30,18 %	1	6	1,34	1	1
77	Vaclav Havel	38208911	43	36,13 %	59	36,42 %	1	8	1,37	1	1
78	Sect	37684399	19	39,58 %	22	34,92 %	1	2	1,16	1	1
79	Holy Grail	38733545	148	51,03 %	216	42,44 %	1	7	1,46	1	1
80	Flagellation	37340140	25	40,32 %	33	34,74 %	1	4	1,32	1	1
81	Ambigram	38211409	49	51,04 %	84	52,17 %	1	9	1,71	1	1
82	Fatalism	36471303	8	34,78 %	14	42,42 %	1	7	1,75	1	1
83	Melaka	38748558	12	35,29 %	21	33,33 %	1	5	1,75	1	1
84	Nanotechnology	38877751	352	64,71 %	605	62,76 %	1	17	1,72	1	1
85	Pai Mu Tan Tea	23347149	0	0,00 %	0	0,00 %	n/a	n/a	n/a	n/a	n/a
86	Manufacturing Consent	37101609	10	32,26 %	14	31,11 %	1	3	1,40	1	1
87	Special K (Disambiguation)	26447858	3	50,00 %	4	36,36 %	1	2	1,33	1	1
88	USB Mass Storage Device Class	36986591	17	51,52 %	18	30,51 %	1	2	1,06	1	1
89	Roman Numerals	38396369	90	37,82 %	115	30,10 %	1	4	1,28	1	1
90	Damascus	38730007	46	35,66 %	61	22,59 %	1	8	1,33	1	1
91	Use of Poison Gas in World War I	32081772	0	0,00 %	0	0,00 %	n/a	n/a	n/a	n/a	n/a
92	Peanuts	38670130	125	45,79 %	248	45,50 %	1	15	1,98	1	1
93	Jack London	38825359	96	44,44 %	188	36,65 %	1	27	1,96	1	1
94	Communication	38046831	101	49,51 %	137	43,08 %	1	7	1,36	1	1
95	Nervous Tissue	16142015	2	66,67 %	2	66,67 %	1	1	1,00	1	1
96	The Hobbit	38826711	83	39,52 %	108	29,92 %	1	6	1,30	1	1
97	Opium War	15958687	0	0,00 %	0	0,00 %	n/a	n/a	n/a	n/a	n/a
98	Parkinson's Disease	38735301	125	52,74 %	187	44,63 %	1	13	1,50	1	1

#	Article	Article Version	Anonymous Users # Total Unique (int)	# Total Unique (%)	# Edits total (int)	# Edits total (%)	# Edits per user MIN	# Edits per user MAX	# Edits per user mean	# Edits per User median	# Edits per User mode
99	Slayer	38879076	132	53,88 %	220	48,14 %	1	21	1,67	1	1
	Median Values for this Class of Articles		19	39,35 %	25	34,83 %	1	6	1,39	1	1
	Mean Values for this Class of Articles		54,48	35,85 %	93,48	32,06 %	1,05	8,66	1,53	1,07	1,07
100	Methamphetamine	38778939	266,00	58,98 %	364,00	48,28 %	1,00	10,00	1,37	1,00	1,00
101	Recreational Drug Use	38057817	57,00	37,25 %	74,00	28,46 %	1,00	6,00	1,30	1,00	1,00
102	Statutory Rape	38698832	35,00	50,72 %	64,00	49,61 %	1,00	12,00	1,83	1,00	1,00
103	Academic Seduction	27971511	4,00	40,00 %	7,00	53,85 %	1,00	3,00	1,75	2,00	1,00
104	Sexual Harassment	38714040	47,00	46,08 %	115,00	47,13 %	1,00	16,00	2,45	1,00	1,00
105	Human Experimentation	38677550	18,00	39,13 %	36,00	32,43 %	1,00	7,00	2,00	1,00	1,00
106	Abu Ghraib Torture and Prisoner Abuse	38719567	95,00	33,93 %	195,00	29,19 %	1,00	42,00	2,05	1,00	1,00
107	Bully	38766627	123,00	55,91 %	262,00	64,37 %	1,00	23,00	2,13	1,00	1,00
108	Spousal Abuse	35241134	15,00	39,47 %	21,00	39,62 %	1,00	3,00	1,40	1,00	1,00
109	Chickenhawk (sexuality)	38723981	5,00	21,74 %	56,00	63,64 %	1,00	47,00	11,20	1,00	1,00
110	Child Pornography	38816524	118,00	45,91 %	155,00	32,63 %	1,00	5,00	1,31	1,00	1,00
111	Shotacon	38394784	60,00	52,63 %	80,00	51,95 %	1,00	8,00	1,33	1,00	1,00
112	LSD	38872325	427,00	58,17 %	610,00	43,76 %	1,00	11,00	1,43	1,00	1,00
113	Psychedelic Mushroom	38714785	118,00	58,42 %	183,00	57,19 %	1,00	13,00	1,55	1,00	1,00
114	Anti-Psychiatry	38625914	79,00	47,88 %	143,00	33,49 %	1,00	16,00	1,81	1,00	1,00
115	Necrophilia	38856704	91,00	50,56 %	120,00	45,80 %	1,00	6,00	1,32	1,00	1,00
116	Petticoat Punishment	38056231	4,00	44,44 %	7,00	53,85 %	1,00	3,00	1,75	2,00	1,00
117	Anal Masturbation	38198115	27,00	45,76 %	33,00	42,31 %	1,00	3,00	1,22	1,00	1,00
118	Autoerotic Asphyxiation	37752048	31,00	44,93 %	43,00	43,00 %	1,00	4,00	1,39	1,00	1,00
119	Fifi (Masturbation Aid)	33282952	0,00	0,00 %	0,00	0,00 %	n/a	n/a	n/a	n/a	n/a
120	Jeff Stryker	38117120	11,00	39,29 %	18,00	35,29 %	1,00	6,00	1,64	1,00	1,00
121	Ku Klux Klan	38870644	848,00	62,26 %	1458,00	42,51 %	1,00	37,00	1,72	1,00	1,00
122	Gay Rights	38748456	167,00	49,85 %	260,00	38,69 %	1,00	25,00	1,56	1,00	1,00
123	Militant Islam	35815548	14,00	29,79 %	26,00	32,50 %	1,00	5,00	1,86	1,00	1,00
124	Scientology	38876995	948,00	63,33 %	1696,00	44,71 %	1,00	36,00	1,79	1,00	1,00
125	Operation Snow White	37261477	8,00	29,63 %	9,00	23,68 %	1,00	2,00	1,12	1,00	1,00
126	Ahmed Yassin	38113877	86,00	44,10 %	139,00	37,27 %	1,00	7,00	1,62	1,00	1,00
127	Nakba	38206861	6,00	20,69 %	9,00	16,67 %	1,00	4,00	1,50	1,00	1,00
128	Roswell Incident	21921781	0,00	0,00 %	0,00	0,00 %	n/a	n/a	n/a	n/a	n/a
129	Jyllands-Posten Muhammad Cartoons Controversy	38876376	423,00	39,57 %	894,00	21,51 %	1,00	17,00	2,11	1,00	1,00
130	Satanism	38839411	240,00	56,07 %	371,00	48,94 %	1,00	9,00	1,55	1,00	1,00
131	Zoophilia	38208106	167,00	51,23 %	279,00	25,23 %	1,00	17,00	1,67	1,00	1,00
132	Prince Albert Piercing	38368233	36,00	41,86 %	43,00	33,86 %	1,00	3,00	1,19	1,00	1,00
133	Sun Myung Moon	38835843	102,00	48,11 %	177,00	31,83 %	1,00	12,00	1,74	1,00	1,00
134	New Anti-Semitism	38703805	53,00	31,36 %	93,00	11,08 %	1,00	9,00	1,75	1,00	1,00
135	Project ARTICHOKE	37927325	5,00	38,46 %	6,00	18,18 %	1,00	2,00	1,20	1,00	1,00
136	Black Site	38387548	24,00	28,57 %	60,00	23,35 %	1,00	15,00	2,50	1,00	1,00
137	Guantanamo Bay	38874766	172,00	43,00 %	283,00	35,16 %	1,00	20,00	1,65	1,00	1,00
138	2003 War on Iraq	24244545	0,00	0,00 %	0,00	0,00 %	n/a	n/a	n/a	n/a	n/a
139	Euthanasia	38828950	286,00	55,64 %	433,00	49,49 %	1,00	19,00	1,51	1,00	1,00
140	Suicide	38845887	279,00	49,91 %	397,00	39,00 %	1,00	11,00	1,42	1,00	1,00
141	Ted Bundy	38846842	139,00	52,26 %	233,00	44,81 %	1,00	19,00	1,68	1,00	1,00
142	Heaven's Gate Cult	38769165	45,00	42,86 %	87,00	51,48 %	1,00	10,00	1,93	1,00	1,00
143	Hezbollah	38830221	194,00	53,44 %	418,00	34,15 %	1,00	16,00	2,15	1,00	1,00
144	Al-Aqsa Intifada	38855951	111,00	42,53 %	226,00	25,17 %	1,00	14,00	2,04	1,00	1,00
145	N44982	37356372	12,00	35,29 %	57,00	57,58 %	1,00	16,00	4,75	2,00	1,00
146	War on Terror	20213625	0,00	0,00 %	0,00	0,00 %	n/a	n/a	n/a	n/a	n/a
147	Jafa	37818305	15,00	42,86 %	23,00	41,07 %	1,00	3,00	1,53	1,00	1,00
148	al-Qaeda	38862411	360,00	50,35 %	535,00	34,14 %	1,00	10,00	1,49	1,00	1,00
149	Abortion	38806159	596,00	55,24 %	1569,00	38,89 %	1,00	438,00	2,63	1,00	1,00
150	Racism	38766395	515,00	54,10 %	956,00	42,83 %	1,00	26,00	1,86	1,00	1,00
151	American Nazi Party	37857558	41,00	41,84 %	59,00	41,26 %	1,00	4,00	1,44	1,00	1,00
152	Judaism	38605886	482,00	58,07 %	782,00	38,39 %	1,00	24,00	1,62	1,00	1,00
153	Necromancy	38049338	62,00	46,97 %	98,00	43,36 %	1,00	7,00	1,58	1,00	1,00
154	Genetically Modified Food	38813725	116,00	52,97 %	188,00	43,52 %	1,00	10,00	1,62	1,00	1,00
155	Domestic Violence	38809236	87,00	47,03 %	160,00	45,07 %	1,00	16,00	1,84	1,00	1,00
156	Holocaust Denial	38869319	283,00	49,48 %	486,00	33,29 %	1,00	18,00	1,72	1,00	1,00
157	Prostate Milking	37162347	3,00	21,43 %	4,00	16,00 %	1,00	2,00	1,33	1,00	1,00
158	Hermetic Order of the Golden Dawn	38877261	59,00	54,13 %	101,00	47,64 %	1,00	20,00	1,71	1,00	1,00
159	Nuclear Power	38768437	175,00	56,63 %	367,00	25,21 %	1,00	63,00	2,10	1,00	1,00
160	Gay Nigger Association of America	38700820	292,00	47,40 %	504,00	27,89 %	1,00	35,00	1,73	1,00	1,00
161	Gonorrhea	15956283	2,00	40,00 %	4,00	57,14 %	2,00	2,00	2,00	2,00	2,00
162	Scabies	37229330	33,00	44,00 %	78,00	60,47 %	1,00	26,00	2,36	1,00	1,00
163	Whaling	38711832	75,00	46,30 %	111,00	23,27 %	1,00	7,00	1,48	1,00	1,00
164	Child Sexual Abuse	38835597	35,00	33,65 %	63,00	22,99 %	1,00	12,00	1,80	1,00	1,00
165	Syphilis	38722440	113,00	45,75 %	198,00	39,92 %	1,00	33,00	1,75	1,00	1,00
166	The Holocaust	38862374	936,00	59,32 %	1527,00	42,07 %	1,00	17,00	1,63	1,00	1,00
167	Apollo Moon Landing is a Hoax	38822817	168,00	44,44 %	299,00	31,15 %	1,00	36,00	1,78	1,00	1,00
168	Appeasement	38520190	79,00	47,02 %	151,00	48,71 %	1,00	15,00	1,91	1,00	1,00
169	Jesus	38876713	1807,00	60,19 %	3349,00	32,43 %	1,00	72,00	1,85	1,00	1,00
170	Adolf Hitler	38749532	2414,00	65,46 %	3909,00	38,58 %	1,00	26,00	1,62	1,00	1,00
171	United States	38868438	2172,00	62,59 %	3379,00	38,46 %	1,00	14,00	1,56	1,00	1,00
172	Iraq	38833491	733,00	57,76 %	1144,00	48,31 %	1,00	13,00	1,56	1,00	1,00
173	Israel	38872729	1034,00	61,69 %	1917,00	41,35 %	1,00	74,00	1,85	1,00	1,00
174	Land of Israel	37223158	27,00	39,71 %	50,00	25,64 %	1,00	6,00	1,85	1,00	1,00
175	Eugenics	38645935	149,00	51,56 %	242,00	26,13 %	1,00	8,00	1,62	1,00	1,00
	Median Values for this Class of Articles		87	46,19 %	147	38,64 %	1	13	1,70	1	1
	Mean Values for this Class of Articles		248,14	44,20 %	427,54	36,55 %	1,01	22,31	1,88	1,06	1,01
176	France at the 1896 Summer Olympics	34027968	0	0,00 %	0	0,00 %	n/a	n/a	n/a	n/a	n/a
177	Prefectures of the Central African Republic	36022243	1	12,50 %	2	15,38 %	2	2	2,00	2	2
178	NatureServe	27386620	0	0,00 %	0	0,00 %	n/a	n/a	n/a	n/a	n/a
179	La Posta Astro-Geophysical Observatory	31006975	0	0,00 %	0	0,00 %	n/a	n/a	n/a	n/a	n/a
180	Diosynth	27115196	2	16,67 %	4	26,67 %	1	3	2,00	1	1
181	Royal Theatre (Victoria)	29485491	0	0,00 %	0	0,00 %	n/a	n/a	n/a	n/a	n/a
182	Bitterne Railway Station	32639226	0	0,00 %	0	0,00 %	n/a	n/a	n/a	n/a	n/a
183	Albert L'Ouvrier	35283176	1	12,50 %	1	12,50 %	1	1	1,00	1	1
184	Vernal Pool	30784586	3	50,00 %	3	25,00 %	1	1	1,00	1	1
185	Navy Commendation Star	36775678	0	0,00 %	0	0,00 %	n/a	n/a	n/a	n/a	n/a
186	Enefiok Udo-Obong	37754382	0	0,00 %	0	0,00 %	n/a	n/a	n/a	n/a	n/a
187	Long Lake	36966895	0	0,00 %	0	0,00 %	n/a	n/a	n/a	n/a	n/a
188	Bernardo de la Paz	37165415	0	0,00 %	0	0,00 %	n/a	n/a	n/a	n/a	n/a
189	Harlem Hellfighters	38757486	7	29,17 %	25	37,88 %	1	16	3,57	1	1
190	Ivan Vasilievich: Back to the Future	32608173	0	0,00 %	0	0,00 %	n/a	n/a	n/a	n/a	n/a
191	Mount Angel Abbey	37316069	6	46,15 %	14	56,00 %	1	4	2,33	2	4
192	Geography of the Philippines	36820786	13	33,33 %	18	24,00 %	1	3	1,38	1	1
193	Joey Fatone	38451830	12	44,44 %	16	48,48 %	1	3	1,33	1	1
194	Cephas Washburn	32577018	1	7,14 %	2	9,52 %	2	2	2,00	2	2
195	Group Dance	36409956	1	10,00 %	1	7,14 %	1	1	1,00	1	1
196	Two-Mile Borris	37167970	2	50,00 %	3	37,50 %	1	2	1,50	1	1
197	Afghanistan Timeline July 2001	36876876	1	14,29 %	1	12,50 %	1	1	1,00	1	1

#	Article	Article Version	Anonymous Users # Total Unique (int)	# Total Unique (%)	# Edits total (int)	# Edits total (%)	# Edits per user MIN	# Edits per user MAX	# Edits per user mean	# Edits per User median	# Edits per User mode
198	Viscount Chandos	16277634	0	0,00 %	0	0,00 %	n/a	n/a	n/a	n/a	n/a
199	Charles X of France	35607312	53	47,75 %	82	48,24 %	1	7	1,55	1	1
200	Square Pyramidal Number	27573986	4	30,77 %	4	22,22 %	1	1	1,00	1	1
201	Ludgate Hill	22665154	1	10,00 %	1	8,33 %	1	1	1,00	1	1
202	Edmonton Journal	38034442	0	0,00 %	0	0,00 %	n/a	n/a	n/a	n/a	n/a
203	Rotary International	38631482	32	40,00 %	45	37,50 %	1	4	1,41	1	1
204	Burningn'n Tree	36704980	5	45,45 %	9	20,93 %	1	3	1,80	1	1
205	Lev Yashin	37737410	25	35,71 %	81	56,25 %	1	22	3,24	1	1
206	Philosophical Institute	34009119	1	25,00 %	2	22,22 %	2	2	2,00	2	2
207	List of Protagonists in Xenosaga	38117658	10	50,00 %	26	26,80 %	1	8	2,60	1	1
208	Military Unit Mottos: United States	35630475	0	0,00 %	0	0,00 %	n/a	n/a	n/a	n/a	n/a
209	Sarafina	37292319	8	61,54 %	9	56,25 %	1	2	1,12	1	1
210	Aerobatics	37803650	16	38,10 %	26	34,21 %	1	8	1,62	1	1
211	ITV Evening News	35718828	5	45,45 %	8	57,14 %	1	4	1,60	1	1
212	Blaarkop	33008328	0	0,00 %	0	0,00 %	n/a	n/a	n/a	n/a	n/a
213	Residency (Medicine)	37210564	3	15,00 %	3	10,34 %	1	1	1,00	1	1
214	Mechitza	36713816	3	25,00 %	3	18,75 %	1	1	1,00	1	1
215	Ramage	28989805	0	0,00 %	0	0,00 %	n/a	n/a	n/a	n/a	n/a
216	Tom Pashby	32544502	2	33,33 %	4	36,36 %	2	2	2,00	2	2
217	Mo	37240709	10	38,46 %	14	31,82 %	1	3	1,40	1	1
218	100 Best Global Brands	36940977	5	35,71 %	7	33,33 %	1	2	1,40	1	1
219	The Media Institute of Southern Africa	27149358	0	0,00 %	0	0,00 %	n/a	n/a	n/a	n/a	n/a
220	434	33582014	3	15,00 %	3	9,38 %	1	1	1,00	1	1
221	Principles of Lust	35232883	1	16,67 %	1	8,33 %	1	1	1,00	1	1
222	Meineckia	37646858	0	0,00 %	0	0,00 %	n/a	n/a	n/a	n/a	n/a
223	St Brendan's Island	36187706	7	50,00 %	8	53,33 %	1	2	1,14	1	1
224	Hungry Hungry Hippos	37278425	11	40,74 %	15	41,67 %	1	3	1,36	1	1
225	Lee Radziwill	38218751	9	40,91 %	22	50,00 %	1	8	2,44	1	1
226	Gary Perkins	33969893	2	28,57 %	4	14,29 %	1	3	2,00	1	1
227	Longmont, Colorado	36610620	11	40,74 %	18	47,37 %	1	4	1,64	1	1
228	Aurochs	38845373	34	33,01 %	48	27,43 %	1	7	1,41	1	1
229	Assoucy	37574287	0	0,00 %	0	0,00 %	n/a	n/a	n/a	n/a	n/a
230	ProjectForum	29934036	3	60,00 %	3	60,00 %	1	1	1,00	1	1
231	Danish Longball	27059972	1	20,00 %	5	45,45 %	5	5	5,00	5	5
232	Godfrey the Bearded	19829119	0	0,00 %	0	0,00 %	n/a	n/a	n/a	n/a	n/a
233	Portuguese Colonization of the Americas	36578547	10	32,26 %	10	25,64 %	1	1	1,00	1	1
234	Australian Referendum, 1910 (Surplus Revenue)	26850010	0	0,00 %	0	0,00 %	n/a	n/a	n/a	n/a	n/a
235	Drilling Machine	35140416	1	20,00 %	1	12,50 %	1	1	1,00	1	1
236	Here Comes the Grump	32395313	0	0,00 %	0	0,00 %	n/a	n/a	n/a	n/a	n/a
237	Thiru Nadana Ula	31738726	0	0,00 %	0	0,00 %	n/a	n/a	n/a	n/a	n/a
238	Foreign Relations of Saint Lucia	34913688	1	7,69 %	1	7,14 %	1	1	1,00	1	1
239	San Quintin, Abra	28043254	0	0,00 %	0	0,00 %	n/a	n/a	n/a	n/a	n/a
240	The Aenar (Enterprise Episode)	32179849	0	0,00 %	0	0,00 %	n/a	n/a	n/a	n/a	n/a
241	Zapf Dingbats	31103834	1	50,00 %	3	75,00 %	3	3	3,00	3	3
242	WFTT-TV	36582093	0	0,00 %	0	0,00 %	n/a	n/a	n/a	n/a	n/a
243	Procrastination (SongeBob SquarePants Episode)	38569862	3	27,27 %	5	35,71 %	1	2	1,67	2	2
244	Watchtower (Fortification)	28269972	0	0,00 %	0	0,00 %	n/a	n/a	n/a	n/a	n/a
245	International Karate Association	36330620	0	0,00 %	0	0,00 %	n/a	n/a	n/a	n/a	n/a
246	Friendly Artificial Intelligence	33172210	14	43,75 %	22	46,81 %	1	7	1,57	1	1
247	Tammy Layne Winslow	36548817	3	60,00 %	3	17,65 %	1	1	1,00	1	1
248	Ulrich Graf von Brockdorff-Rantzau	37799116	2	15,38 %	3	20,00 %	1	2	1,50	1	2
249	The Mean Reds	36986416	4	80,00 %	5	83,33 %	1	2	1,25	1	1
250	Shallow Donor	26402254	1	16,67 %	1	14,29 %	1	1	1,00	1	1
251	List of Happy Days Episodes	30047093	4	57,14 %	10	25,64 %	1	4	2,50	3	4
252	Chief Minister of Gujarat	35933096	1	20,00 %	1	9,09 %	1	1	1,00	1	1
253	Mercer Union	37527146	1	33,33 %	2	50,00 %	2	2	2,00	2	2
254	Alfred Downing Fripp (Surgeon)	34660645	0	0,00 %	0	0,00 %	n/a	n/a	n/a	n/a	n/a
255	Catupiry	36020664	2	33,33 %	4	44,44 %	2	2	2,00	2	2
256	Gehenna in Popular Culture	37046217	3	37,50 %	3	27,27 %	1	1	1,00	1	1
257	UTF-1	24411621	0	0,00 %	0	0,00 %	n/a	n/a	n/a	n/a	n/a
258	Contemporary Baroque Art	32846606	6	75,00 %	7	50,00 %	1	2	1,17	1	1
259	Big Belt Mountains	32518509	0	0,00 %	0	0,00 %	n/a	n/a	n/a	n/a	n/a
260	U of A	32455903	1	25,00 %	1	25,00 %	1	1	1,00	1	1
261	Suzanne Rogers	38330065	5	38,46 %	5	21,74 %	1	1	1,00	1	1
262	Board Foot	38503164	3	21,43 %	3	14,29 %	1	1	1,00	1	1
263	Whizzinator	30394029	2	33,33 %	4	50,00 %	2	2	2,00	2	2
264	San Francisco Tape Music Center	36208985	0	0,00 %	0	0,00 %	n/a	n/a	n/a	n/a	n/a
265	Barney Clark	35620429	2	28,57 %	3	30,00 %	1	2	1,50	1	1
266	Church of the Gesu	38620524	7	33,33 %	7	17,07 %	1	1	1,00	1	1
267	Hisingen Eastern Hundred	37942089	0	0,00 %	0	0,00 %	n/a	n/a	n/a	n/a	n/a
268	Scotch College	38276583	4	25,00 %	7	20,00 %	1	3	1,75	2	1
269	Child Discipline	34710054	1	9,09 %	2	7,14 %	2	2	2,00	2	2
270	John Maclean MA	37011049	17	40,48 %	20	26,67 %	1	2	1,18	1	1
	Median Values for this Class of Articles		1	20,00 %	3	15,38 %	1	2	1,40	1	1
	Mean Values for this Class of Articles		4,36	22,24 %	7,09	20,49 %	1,22	3,06	1,59	1,30	1,35
271	Prophets of Islam	38346418	30,00	35,71 %	35,00	17,77 %	1,00	3,00	1,17	1,00	1,00
272	State Leaders by Year	33946271	3,00	10,34 %	9,00	9,89 %	1,00	4,00	3,00	4,00	4,00
273	Guatemala Constitution	34726472	0,00	0,00 %	0,00	0,00 %	n/a	n/a	n/a	n/a	n/a
274	Race	38881833	196,00	44,95 %	304,00	18,05 %	1,00	14,00	1,55	1,00	1,00
275	First English Civil War	37994356	7,00	17,07 %	13,00	6,16 %	1,00	4,00	1,86	2,00	2,00
276	Italian Literature	38834427	7,00	17,07 %	10,00	11,11 %	1,00	3,00	1,43	1,00	1,00
277	The Idler (1758-1760)	38007160	0,00	0,00 %	0,00	0,00 %	n/a	n/a	n/a	n/a	n/a
278	Continuation War	38730042	64,00	38,55 %	103,00	25,18 %	1,00	9,00	1,61	1,00	1,00
279	Short-Lived Recurring Characters on Saturday Night Live	37920995	15,00	65,22 %	27,00	37,50 %	1,00	5,00	1,80	1,00	1,00
280	Friedrich Nietzsche	38881785	496,00	59,69 %	792,00	42,72 %	1,00	17,00	1,60	1,00	1,00
281	Software Law	16368299	1,00	50,00 %	1,00	33,33 %	1,00	1,00	1,00	1,00	1,00
282	Czech Lands 880s-1198	16276141	2,00	50,00 %	3,00	60,00 %	1,00	2,00	1,50	1,00	1,00
283	Effective Height	15936587	2,00	50,00 %	2,00	50,00 %	1,00	1,00	1,00	1,00	1,00
284	Terrorist Attacks Against Israelis in 2000	16303508	1,00	33,33 %	1,00	25,00 %	1,00	1,00	1,00	1,00	1,00
285	Path Quality Analysis	15936984	2,00	50,00 %	2,00	50,00 %	1,00	1,00	1,00	1,00	1,00
286	Some Tame Gazelle	15994079	1,00	33,33 %	1,00	33,33 %	1,00	1,00	1,00	1,00	1,00
287	The Clandestine Marriage	16140455	1,00	33,33 %	1,00	33,33 %	1,00	1,00	1,00	1,00	1,00
288	Opponents Batting Average	16503128	0,00	0,00 %	0,00	0,00 %	n/a	n/a	n/a	n/a	n/a
289	Hong Kong Open (Snooker)	16489325	1,00	50,00 %	1,00	25,00 %	1,00	1,00	1,00	1,00	1,00
290	History of the French School of Anatomy	38836575	1,00	14,29 %	1,00	6,67 %	1,00	1,00	1,00	1,00	1,00
291	Neema	23530164	1,00	50,00 %	1,00	50,00 %	1,00	1,00	1,00	1,00	1,00
292	Sipple Syndrome	32188907	1,00	33,33 %	1,00	33,33 %	1,00	1,00	1,00	1,00	1,00
293	Suggestio Falsi	36687034	0,00	0,00 %	0,00	0,00 %	n/a	n/a	n/a	n/a	n/a
294	Cymbaline	32669617	1,00	50,00 %	1,00	50,00 %	1,00	1,00	1,00	1,00	1,00
295	Webrovka	31918095	2,00	66,67 %	2,00	66,67 %	1,00	1,00	1,00	1,00	1,00
296	Suppressio Veri	36687241	0,00	0,00 %	0,00	0,00 %	n/a	n/a	n/a	n/a	n/a
297	Henry Moore	37336243	69,00	46,31 %	104,00	37,41 %	1,00	7,00	1,51	1,00	1,00
298	Multiple Sclerosis	38880952	133,00	53,63 %	193,00	33,98 %	1,00	6,00	1,45	1,00	1,00

#	Article	Article Version	Anonymous Users								
			# Total Unique (int)	# Total Unique (%)	# Edits total (int)	# Edits total (%)	# Edits per user MIN	# Edits per user MAX	# Edits per user mean	# Edits per User median	# Edits per User mode
299	Ammolite	37448848	22,00	42,31 %	37,00	37,00 %	1,00	10,00	1,68	1,00	1,00
300	Commodore 64	38242041	178,00	51,90 %	359,00	44,16 %	1,00	76,00	2,02	1,00	1,00
301	Gold Standard	38478230	141,00	48,96 %	208,00	31,23 %	1,00	8,00	1,48	1,00	1,00
302	Cheese	38817612	707,00	62,79 %	1086,00	46,19 %	1,00	8,00	1,54	1,00	1,00
303	Ferdinand Magellan	38875495	242,00	56,54 %	432,00	46,10 %	1,00	22,00	1,79	1,00	1,00
304	Equal Protection Clause	38702165	22,00	44,00 %	32,00	10,60 %	1,00	7,00	1,45	1,00	1,00
305	Eigenvalue, Eigenvector and Eigenspace	38533087	78,00	44,83 %	275,00	49,55 %	1,00	149,00	3,53	1,00	1,00
306	Sex Pistols	38779219	159,00	50,32 %	300,00	44,58 %	1,00	49,00	1,89	1,00	1,00
307	Michael Jackson	38883303	1630,00	66,48 %	3217,00	42,85 %	1,00	143,00	1,97	1,00	1,00
308	Hurricane Katrina	38883339	2212,00	67,56 %	4325,00	46,04 %	1,00	101,00	1,96	1,00	1,00
309	John Kerry	38871311	1179,00	62,41 %	2427,00	33,83 %	1,00	159,00	2,06	1,00	1,00
310	World War II	38882408	1851,00	64,97 %	3064,00	44,57 %	1,00	23,00	1,66	1,00	1,00
311	Pope Benedict XVI	38839269	1306,00	63,00 %	2528,00	39,76 %	1,00	31,00	1,94	1,00	1,00
312	Wiki	38857205	2302,00	77,67 %	3284,00	58,64 %	1,00	10,00	1,43	1,00	1,00
313	Canada	38879764	1259,00	61,35 %	2099,00	36,79 %	1,00	62,00	1,67	1,00	1,00
314	India	38882043	1082,00	62,44 %	1749,00	36,71 %	1,00	56,00	1,62	1,00	1,00
315	Joseph Stalin	38872097	834,00	57,68 %	1398,00	33,83 %	1,00	28,00	1,68	1,00	1,00
316	Star Wars	38878187	1062,00	61,67 %	1834,00	44,40 %	1,00	46,00	1,73	1,00	1,00
317	Population Density	38448322	75,00	44,38 %	100,00	39,53 %	1,00	4,00	1,33	1,00	1,00
318	Marriage	38863425	299,00	54,76 %	448,00	46,04 %	1,00	8,00	1,50	1,00	1,00
319	Japan	38841002	903,00	60,73 %	1630,00	45,29 %	1,00	42,00	1,81	1,00	1,00
320	Television	38676359	360,00	49,86 %	539,00	43,64 %	1,00	10,00	1,50	1,00	1,00
321	Animal	38722239	206,00	46,82 %	308,00	36,32 %	1,00	11,00	1,50	1,00	1,00
322	Actor	38791354	75,00	36,41 %	102,00	32,28 %	1,00	8,00	1,36	1,00	1,00
323	Latin	38748360	230,00	47,72 %	325,00	43,98 %	1,00	11,00	1,41	1,00	1,00
324	Mathematics	38751995	443,00	53,50 %	629,00	35,46 %	1,00	15,00	1,42	1,00	1,00
325	Chordate	37926199	34,00	30,09 %	46,00	27,71 %	1,00	4,00	1,35	1,00	1,00
326	American Civil War	38880177	568,00	58,80 %	876,00	39,50 %	1,00	20,00	1,54	1,00	1,00
327	World Trade Center	38868415	231,00	48,84 %	333,00	42,86 %	1,00	17,00	1,44	1,00	1,00
328	Aria Giovanni	38567158	90,00	51,14 %	152,00	48,41 %	1,00	24,00	1,69	1,00	1,00
329	Western Philosophy	35416997	131,00	47,29 %	178,00	38,03 %	1,00	4,00	1,36	1,00	1,00
330	BASIC Programming Language	38432624	189,00	51,64 %	275,00	32,20 %	1,00	10,00	1,46	1,00	1,00
331	Chemistry	38848167	235,00	49,16 %	380,00	43,88 %	1,00	10,00	1,62	1,00	1,00
332	Nude Celebrities on the Internet	38434727	36,00	43,90 %	40,00	30,08 %	1,00	2,00	1,11	1,00	1,00
333	Thomas Jefferson	38869811	471,00	53,95 %	810,00	40,34 %	1,00	24,00	1,72	1,00	1,00
334	Sweden	38853661	495,00	58,65 %	763,00	43,75 %	1,00	26,00	1,54	1,00	1,00
335	NP-Complete	38583812	42,00	42,86 %	61,00	32,80 %	1,00	4,00	1,45	1,00	1,00
336	World War I	38861312	1058,00	61,91 %	1619,00	44,79 %	1,00	14,00	1,53	1,00	1,00
337	Adam Sandler	38628112	116,00	53,46 %	160,00	54,98 %	1,00	6,00	1,38	1,00	1,00
338	Bethlehem	38620086	44,00	36,97 %	56,00	24,67 %	1,00	5,00	1,27	1,00	1,00
339	Iconostasis	33287214	11,00	29,73 %	15,00	23,81 %	1,00	3,00	1,36	1,00	1,00
340	La Marseillaise	38492917	74,00	48,68 %	87,00	37,66 %	1,00	4,00	1,18	1,00	1,00
341	Nanoart	36900971	0,00	0,00 %	0,00	0,00 %	n/a	n/a	n/a	n/a	n/a
342	Oasis (Band)	38846457	338,00	63,06 %	685,00	44,37 %	1,00	55,00	2,03	1,00	1,00
343	Radio Frequency	37830416	54,00	46,15 %	76,00	42,94 %	1,00	4,00	1,41	1,00	1,00
344	Samara	36075192	1,00	25,00 %	1,00	20,00 %	1,00	1,00	1,00	1,00	1,00
345	VeriPic	34225025	1,00	25,00 %	1,00	20,00 %	1,00	1,00	1,00	1,00	1,00
346	Yukio Seki	37060688	0,00	0,00 %	0,00	0,00 %	n/a	n/a	n/a	n/a	n/a
	Median Values for this Class of Articles		75	49,06 %	103	36,90 %	1	8	1,48	1	1
	Mean Values for this Class of Articles		317,28	43,69 %	538,91	33,40 %	1,00	20,59	1,51	1,06	1,06
	Mean Values for all Classes of Articles		139,25	35,91 %	237,32	30,25 %	1,06	13,53	1,62	1,11	1,11
	Minimum Value for all Classes of Articles		0	0,00 %	0	0,00 %	1	1	1,00	1	1
	Maximum Value for all Classes of Articles		2414	80,00 %	4325	83,33 %	5	438	11,20	5	5
	Median Value for all Classes of Articles		15	40,16 %	22	33,33 %	1	5	1,51	1	1

D Data from the Analysis of HTML-Encoded Dublin Core Metadata

The following two tables present the key results from the analysis of Dublin Core Metadata in academic web-sites detailed in chapter 7. The first table contains the means values from the data collected for the investigation: number of pages with Dublin Core (DC) elements, number of DC elements per page, length of the heads of HTML pages, the proportion of the headers of the HTML files to the entire file size, and the results from the W3C and WDG HTML syntax validation. In addition to the mean values, the minimum, maximum, median, and mode values were calculated. However, these values were not used in the chapter 7.

The second table lists the frequencies of the fifteen simple (unqualified) Dublin Core elements for the individual web-sites analysed. The total number of occurences for each DC element are also shown in table 7.2.

Key Results from the Investigation

#	Domain	Total Number of Pages	Pages with DC Elements		Total Number of DC Elements	Number of DC Elements per Page		Head Length		Head-to-Page Ratio		Passed W3C HTML Validation without Errors or Warnings		Passed WDG HTML Validation without Errors or Warnings	
			Number of Pages	Proportion of All Pages		All Pages	Pages with DC Elements Only	All Pages	Pages with DC Elements Only	All Pages	Pages with DC Elements Only	All Pages	Pages with DC Elements Only	All Pages	Pages with DC Elements Only
1	tugraz.at	2873	45	1,57%	171	0,06	3,80	1357,32	2033,38	17,15%	11,27%	6,72%	0,00%	9,85%	24,44%
2	uni-graz.at	5423	26	0,48%	326	0,06	12,54	596,19	2156,31	6,07%	29,90%	9,55%	0,00%	10,09%	0,00%
3	univie.ac.at	5035	143	2,84%	1392	0,28	9,73	1924,29	5704,86	16,61%	28,40%	19,90%	0,00%	42,09%	26,57%
4	tuwien.ac.at	1079	4	0,37%	29	0,03	7,25	548,09	2279,00	7,96%	39,38%	3,71%	0,00%	7,69%	0,00%
5	uibk.ac.at	3674	2704	73,60%	26694	7,27	9,87	1057,55	1115,75	13,17%	12,20%	3,95%	3,51%	6,53%	6,66%
6	boku.ac.at	5003	1	0,02%	5	0,00	5,00	3612,07	364,00	20,51%	35,10%	0,02%	0,00%	0,24%	0,00%
7	uni-klu.ac.at	1896	2	0,11%	12	0,01	6,00	1430,95	1349,00	19,73%	13,74%	0,74%	100,00%	1,00%	100,00%
8	uni-salzburg.at	5038	1	0,02%	5	0,00	5,00	6615,23	3791,00	24,00%	10,89%	9,13%	0,00%	10,24%	0,00%
9	fh-joanneum.at	2508	0	0,00%	0	0,00		1808,62		8,94%		0,08%		0,08%	
10	fh-hagenberg.at	2513	1	0,04%	3	0,00	3,00	639,90	559,00	7,35%	7,89%	0,20%	0,00%	0,24%	0,00%
11	fh-kaernten.at	1709	0	0,00%	0	0,00		6626,34		13,98%		0,06%		0,06%	
12	fh-burgenland.at	2534	6	0,24%	96	0,02	16,00	5942,56	2212,33	33,88%	39,90%	0,67%	0,00%	1,58%	0,00%
13	ethz.ch	3487	7	0,20%	16	0,00	2,29	2219,04	3912,57	17,66%	22,06%	2,41%	0,00%	3,10%	0,00%
14	unibas.ch	2530	25	0,99%	297	0,12	11,88	2866,73	1962,00	11,72%	21,62%	0,71%	0,00%	0,95%	0,00%
15	unige.ch	2598	35	1,35%	183	0,07	5,23	841,27	3613,63	20,68%	25,18%	1,12%	28,57%	5,20%	28,57%
16	uni-heidelberg.de	2528	9	0,36%	36	0,01	4,00	520,80	1266,78	6,84%	15,00%	2,02%	0,00%	14,79%	11,11%
17	uni-muenchen.de	975	23	2,36%	118	0,12	5,13	1460,24	1702,70	13,18%	11,79%	2,36%	0,00%	5,03%	0,00%
18	tum.de	2515	56	2,23%	320	0,13	5,71	2411,52	5995,57	10,14%	24,82%	0,16%	0,00%	0,44%	5,36%
19	hu-berlin.de	1407	512	36,39%	3133	2,23	6,12	2046,76	4357,95	19,25%	22,49%	30,14%	74,41%	30,77%	74,61%
20	tudelft.nl	2515	0	0,00%	0	0,00		922,90		5,74%		9,15%		9,15%	
21	imperial.ac.uk	2727	879	32,23%	7911	2,90	9,00	14847,89	34025,65	39,71%	76,00%	0,29%	0,00%	0,33%	0,00%
22	nottingham.ac.uk	2510	10	0,40%	80	0,03	8,00	1257,38	1097,90	11,07%	28,18%	3,59%	10,00%	3,67%	10,00%
23	aaue.dk	2769	1	0,04%	1	0,00	1,00	684,30	1177,00	12,79%	7,77%	12,17%	0,00%	19,10%	0,00%
24	ox.ac.uk	2519	884	35,09%	3485	1,38	3,94	1230,82	1715,56	15,37%	16,99%	41,05%	61,76%	42,16%	62,78%
25	cam.ac.uk	1927	28	1,45%	212	0,11	7,57	587,87	2127,57	8,09%	25,07%	54,13%	82,14%	55,79%	82,14%
26	columbia.edu	362	0	0,00%	0	0,00		2031,96		15,11%		4,14%		4,14%	
27	stanford.edu	2574	0	0,00%	0	0,00		849,67		9,95%		2,91%		2,99%	
28	berkeley.edu	2518	4	0,16%	15	0,01	3,75	1150,91	1097,25	7,86%	14,41%	1,87%	50,00%	1,91%	50,00%
29	yale.edu	421	20	4,75%	75	0,18	3,75	2744,33	473,70	20,07%	2,55%	0,71%	0,00%	0,71%	0,00%
30	harvard.edu	609	0	0,00%	0	0,00		1661,17		13,71%		8,37%		8,70%	
31	princeton.edu	2557	658	25,73%	4617	1,81	7,02	1038,39	1297,26	12,13%	11,46%	13,34%	38,30%	14,08%	40,27%
32	dartmouth.edu	2537	7	0,28%	37	0,01	5,29	1232,58	2837,43	11,23%	13,18%	14,43%	42,86%	15,49%	42,86%
33	upenn.edu	550	0	0,00%	0	0,00		942,04		6,87%		0,73%		0,73%	
34	brown.edu	2545	0	0,00%	0	0,00		1038,42		14,00%		3,26%		3,54%	
35	uiuc.edu	2551	5	0,20%	68	0,03	13,60	2139,46	4173,80	11,61%	27,42%	4,55%	0,00%	4,70%	0,00%
36	keio.ac.jp	2532	0	0,00%	0	0,00		879,69		13,40%		3,91%		5,06%	
37	u-tokyo.ac.jp	2556	4	0,16%	23	0,01	5,75	1059,45	1379,25	10,80%	26,27%	4,77%	0,00%	8,96%	0,00%
38	kyoto-u.ac.jp	2551	0	0,00%	0	0,00		774,57		10,21%		1,92%		2,94%	
39	hku.hk	2662	1	0,04%	1	0,00	1,00	1220,92	661,00	14,54%	41,25%	1,05%	0,00%	1,24%	0,00%
40	pku.edu.cn	1354	0	0,00%	0	0,00		1060,92		12,17%		0,00%		0,37%	
41	shu.edu.cn	1582	0	0,00%	0	0,00		618,70		11,33%		0,06%		1,07%	
42	ntu.edu.sg	2636	0	0,00%	0	0,00		1225,77		10,63%		0,46%		0,46%	
43	rmit.edu.au	2505	2399	95,77%	4699	1,88	1,96	888,40	888,50	5,47%	5,18%	0,68%	0,00%	0,68%	0,00%
44	usyd.edu.au	2528	1837	72,67%	12741	5,04	6,94	1111,21	1153,48	8,69%	8,24%	1,23%	0,44%	1,27%	0,49%
45	curtin.edu.au	2723	2060	75,65%	15939	5,85	7,74	2311,51	2881,51	23,15%	18,88%	12,63%	12,33%	12,71%	12,43%
46	auckland.ac.nz	2685	192	7,15%	396	0,15	2,06	1882,72	1455,83	12,31%	9,67%	2,35%	0,00%	2,38%	0,00%
47	otago.ac.nz	2535	79	3,12%	625	0,25	7,91	898,73	2519,15	8,82%	20,77%	32,35%	22,78%	33,02%	22,78%
48	massey.ac.nz	2535	213	8,40%	1484	0,59	6,97	12050,75	1270,89	39,67%	8,18%	0,16%	0,00%	0,16%	0,00%
	Mean Value	2477	268	10,13%	1776	0,64	4,62	2184,77	3045,96	14,28%	20,94%	6,87%	15,06%	8,49%	17,17%
	All Domains Together	118900	12881	10,83%	85245	0,72	6,62	2399,14	3894,68	15,05%	16,80%	7,12%	12,38%	9,22%	13,66%

Frequency of Unqualified Dublin Core Elements

#	Domain	Frequencies of Simple DC Elements															Total
		dc.contributor	dc.coverage	dc.creator	dc.date	dc.description	dc.format	dc.identifier	dc.language	dc.publisher	dc.relation	dc.rights	dc.source	dc.subject	dc.title	dc.type	
1	tugraz.at	0	0	17	27	17	19	2	28	1	0	0	0	1	2	19	133
2	uni-graz.at	25	24	25	25	25	25	25	25	25	0	25	0	25	25	25	324
3	univie.ac.at	112	73	115	36	136	39	65	142	112	29	139	73	139	110	39	1359
4	tuwien.ac.at	2	1	2	1	1	2	1	2	2	1	4	1	2	2	2	26
5	uibk.ac.at	2632	0	2666	31	2665	2665	10	2688	2692	0	2	2669	0	2682	3	21405
6	boku.ac.at	0	0	1	1	0	1	0	1	0	0	0	0	0	1	0	5
7	uni-klu.ac.at	0	0	0	2	0	0	2	2	2	0	2	0	0	0	0	10
8	uni-salzburg.at	0	0	1	0	1	0	0	1	0	0	1	1	0	0	0	5
9	fh-joanneum.at	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	fh-hagenberg.at	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0	3
11	fh-kaernten.at	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	fh-burgenland.at	0	0	0	6	0	6	6	6	6	0	0	0	6	6	6	48
13	ethz.ch	1	0	0	0	6	0	0	1	1	0	0	0	6	1	0	16
14	unibas.ch	0	25	25	1	25	24	24	25	24	24	25	0	25	25	25	297
15	unige.ch	0	0	24	0	0	33	4	2	5	0	2	0	8	6	33	117
16	uni-heidelberg.de	0	0	7	0	2	1	1	1	1	0	7	1	8	1	1	31
17	uni-muenchen.de	0	0	23	0	3	4	0	19	0	0	0	0	0	0	23	72
18	tum.de	0	0	33	1	2	46	21	22	17	0	17	0	9	21	33	222
19	hu-berlin.de	1	3	505	7	13	503	5	498	10	1	3	6	41	12	505	2113
20	tudelft.nl	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	imperial.ac.uk	0	0	879	879	879	879	879	879	879	0	0	0	879	879	0	7911
22	nottingham.ac.uk	0	0	0	10	10	10	10	0	10	0	0	0	9	10	10	79
23	aaue.dk	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
24	ox.ac.uk	0	34	281	36	46	98	40	272	39	0	38	0	598	865	319	2666
25	cam.ac.uk	0	0	0	23	7	7	28	21	21	0	0	0	28	28	7	170
26	columbia.edu	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
27	stanford.edu	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
28	berkeley.edu	3	0	0	0	2	3	2	0	1	0	0	0	1	2	0	14
29	yale.edu	0	0	1	1	2	2	2	1	20	0	20	0	20	2	2	73
30	harvard.edu	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
31	princeton.edu	0	0	0	0	0	142	657	657	656	0	657	0	0	657	657	4083
32	dartmouth.edu	0	0	6	0	6	2	0	0	4	0	0	0	4	5	2	29
33	upenn.edu	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
34	brown.edu	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
35	uiuc.edu	2	2	3	3	3	3	5	5	3	2	2	2	3	5	5	48
36	keio.ac.jp	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
37	u-tokyo.ac.jp	0	0	3	3	1	1	0	1	1	0	1	0	1	4	1	17
38	kyoto-u.ac.jp	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
39	hku.hk	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
40	pku.edu.cn	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
41	shu.edu.cn	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
42	ntu.edu.sg	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
43	rmit.edu.au	1	1	22	22	22	22	2132	23	23	0	1	0	2398	23	1	4691
44	usyd.edu.au	3	0	1832	140	1382	25	16	25	1817	1	138	0	809	1461	24	7673
45	curtin.edu.au	221	37	2045	1654	1697	1563	1578	1601	367	0	1662	18	1694	1700	13	15850
46	auckland.ac.nz	0	0	0	2	0	192	0	192	0	0	2	0	0	0	2	390
47	otago.ac.nz	73	0	77	2	77	0	0	78	74	0	2	0	78	79	0	540
48	massey.ac.nz	0	0	213	212	207	212	0	0	213	0	0	0	1	213	0	1271
	Mean Value	64	4	183	65	151	136	115	150	146	1	57	58	142	184	37	1494
	All Domains Together	3077	200	8807	3126	7237	6529	5515	7219	7027	58	2750	2771	6793	8827	1757	71693
	Used in Domains	13	9	24	24	25	27	22	28	28	6	21	8	24	27	23	309

E Data from Experiments with WordFlickr and FlickrClustr

The following table contains the clustered results from the 21 queries to Flickr and WordFlickr mentioned in chapter 8. Four of these results are also depicted in table 8.2.

For every of the 21 terms listed below, a query was submitted to both Flickr and WordFlickr. The parameters used were "query by tag" and "order by relevance". The number of results was limited to 200 photos per query. From these 200 results, the ten most frequently used tags are provided in the table below. The numbers in brackets are the tag frequencies.

The queries from "shoe" to "wine" were submitted on November 11th and the remaining queries were submitted on November 13th, 2006.

Term	shoe		garden		flower	
Query	**Flickr**	**WordFlickr**	**Flickr**	**WordFlickr**	**Flickr**	**WordFlickr**
10 Most Frequent Tags	shoe (168)	shoe (161)	garden (172)	garden (174)	flower (186)	flower (182)
	shoes (50)	pump (52)	flowers (38)	grove (63)	island (43)	chrysanthemum (83)
	foot (33)	shoes (38)	Garden (28)	california (56)	dublin (43)	macro (64)
	feet (31)	flipflop (31)	flower (22)	rose (54)	kerry (43)	nature (33)
	Shoe (22)	foot (21)	plants (20)	dave (49)	grave (43)	cornflower (32)
	legs (11)	feet (20)	Alaska (17)	beer (49)	ocean (43)	garden (30)
	leg (10)	heel (17)	UAF (17)	charleen (49)	derry (43)	yellow (30)
	toes (10)	flip-flop (14)	Fairbanks (17)	doreen (49)	landscape (43)	flowers (29)
	heel (9)	red (14)	gardens (17)	cocktails (49)	ruins (43)	blue (24)
	black (9)	sneakers (13)	Fairbanks, Alaska (17)	safley (49)	ireland (43)	Chrysanthemum (23)

Term	africa		europe		goose	
Query	**Flickr**	**WordFlickr**	**Flickr**	**WordFlickr**	**Flickr**	**WordFlickr**
10 Most Frequent Tags	Africa (133)	africa (118)	europe (127)	europe (105)	goose (144)	goose (148)
	africa (67)	Africa (82)	Europe (73)	Europe (87)	Goose (52)	bird (106)
	Kenya (42)	tanzania (81)	2002 (29)	travel (41)	geese (39)	Branta canadensis (87)
	Zoo (27)	family (60)	Italy (13)	June (37)	bird (36)	geese (85)
	Animals (27)	safari (53)	France (13)	2000 (37)	water (27)	Canada Goose (83)
	Plants (27)	vacation (41)	set:helsinki-tallinn_aug06 (13)	Trip (37)	birds (27)	canada goose (67)
	Tanzania (14)	Kenya (28)	Rome (12)	Eurotrip (36)	duck (24)	birds (61)
	Uganda (14)	Zoo (27)	france (12)	Backpacking (36)	park (23)	gosling (60)
	tanzania (14)	Animals (27)	travel (10)	adventure (34)	Setup (21)	nature (45)
	orphans (13)	Plants (27)	Paris (9)	hitchhiking (34)	Wild (21)	Goose (42)

Term	water		rock		pathway	
Query	**Flickr**	**WordFlickr**	**Flickr**	**WordFlickr**	**Flickr**	**WordFlickr**
10 Most Frequent Tags	water (178)	water (193)	rock (125)	rock (186)	pathway (172)	pathway (49)
	nature (40)	pool (169)	music (42)	limestone (172)	path (29)	path (9)
	tree (32)	james (107)	Rock (39)	ocean (28)	trees (28)	trees (8)
	river (31)	spring (97)	ROCK (36)	geology (28)	Pathway (26)	december (7)
	landscape (27)	fun (67)	live (29)	water (27)	nature (19)	november (7)
	night (26)	green (64)	metal (26)	cave (25)	trail (16)	garden (7)
	art (26)	boat (60)	show (23)	coast (23)	park (15)	trail (6)
	friends (26)	summer (59)	punk (22)	climb (22)	green (13)	2006 (6)
	me (26)	color (55)	musica (22)	landscape (22)	november (12)	nature (4)
	summer (25)	trees (53)	york (20)	sea (22)	december (12)	leaves (4)

Term	wine		red		frog	
Query	**Flickr**	**WordFlickr**	**Flickr**	**WordFlickr**	**Flickr**	**WordFlickr**
10 Most Frequent Tags	wine (161)	wine (174)	red (159)	red (183)	frog (158)	frog (182)
	Wine (39)	chardonnay (80)	Red (29)	cherry (116)	Frog (38)	tree frog (92)
	party (20)	rose (49)	green (24)	ruby (110)	nature (18)	treefrog (85)
	vino (16)	vin (37)	blue (20)	burgundy (85)	amphibian (15)	amphibian (53)
	glass (12)	pinot (28)	2006 (17)	rouge (68)	treefrog (14)	nature (48)
	2004 (11)	france (27)	macro (15)	scarlet (62)	2006 (13)	green (31)
	2005 (10)	white (27)	orange (13)	maroon (61)	Nisqually (11)	outdoors (24)
	yen (9)	bottle (22)	yellow (13)	cerise (58)	outdoors (11)	frogs (21)
	brokenwood (9)	cabernet (22)	black (13)	color (57)	london (11)	September (20)
	roadtrip (9)	Wine (22)	white (13)	pink (54)	September (11)	Nisqually (20)

Term	typhoon		railway		drama	
Query	**Flickr**	**WordFlickr**	**Flickr**	**WordFlickr**	**Flickr**	**WordFlickr**
10 Most Frequent Tags	typhoon (121)	typhoon (121)	railway (141)	subway (160)	drama (136)	drama (165)
	Typhoon (72)	Typhoon (72)	Railway (55)	metro (150)	Drama (54)	comedy (119)
	Eurofighter (17)	Eurofighter (17)	train (51)	railway (137)	theater (22)	theater (77)
	milenyo (16)	milenyo (16)	station (36)	station (81)	Sunset (18)	tragedy (73)
	2006 (15)	2006 (15)	steam (14)	railroad (76)	NDHS (18)	theatre (59)
	夕陽 (14)	夕陽 (14)	Train (14)	rail (73)	Scare (18)	greek (42)
	japan (13)	japan (13)	bridge (13)	urban (72)	Red (18)	movie (35)
	sunset (13)	sunset (13)	locomotive (13)	street (60)	Play (18)	celebrity (33)
	philippines (13)	philippines (13)	japan (11)	light (60)	tv (14)	lee (32)
	Milenyo (12)	Milenyo (12)	2006 (10)	city (57)	A Funny Thing Happened on the Way to the Forum (14)	stripped (32)

Term	college		visa		balcony	
Query	**Flickr**	**WordFlickr**	**Flickr**	**WordFlickr**	**Flickr**	**WordFlickr**
10 Most Frequent Tags	college (99)	college (161)	visa (124)	visa (123)	balcony (166)	amphitheatre (78)
	College (68)	dartmouth (131)	Visa (64)	Visa (65)	Balcony (24)	gallery (76)
	Graduation (35)	university (58)	canada (27)	canada (27)	building (16)	amphitheater (75)
	Claremont, CA (33)	jose (56)	calgary (25)	calgary (25)	blue (12)	balcony (52)
	Pomona College (33)	cal (56)	12/2/2005 (25)	banda (25)	green (12)	art (40)
	Claremont (33)	logitech (56)	banda (25)	12/2/2005 (25)	house (11)	architecture (33)
	Pomona (33)	intercollegiate (56)	Banff (21)	Banff (21)	night (10)	view (31)
	2005 (28)	figure (56)	motorhome (18)	Aussie (18)	2006 (10)	nature (31)
	May 2005 (28)	michigan (56)	Aussie (18)	australia (18)	architecture (10)	museum (31)
	Pomona College Graduation 2005 (28)	nationals (56)	Perth (18)	permanent residency (18)	view (10)	beautiful (30)

Term	cold		tiramisu		bamboo	
Query	**Flickr**	**WordFlickr**	**Flickr**	**WordFlickr**	**Flickr**	**WordFlickr**
10 Most Frequent Tags	cold (164)	cool (159)	tiramisu (153)	tiramisu (153)	bamboo (145)	bamboo (148)
	winter (40)	cold (155)	food (48)	food (48)	Bamboo (44)	Bamboo (41)
	snow (34)	chill (62)	dessert (44)	dessert (44)	Japan (10)	green (11)
	ice (29)	snow (51)	Tiramisu (28)	Tiramisu (28)	green (10)	Dendrocalamus Giganteus (10)
	Cold (26)	winter (47)	cake (28)	cake (28)	Canon (9)	Japan (10)
	night (17)	ice (44)	tiramisu (13)	tiramisu (13)	trees (8)	brazil (9)
	2006 (16)	nature (33)	chocolate (10)	party (10)	USA (8)	plant (9)
	water (13)	water (31)	party (10)	dinner (10)	EOS (8)	Canon (9)
	party (13)	blue (31)	dinner (10)	chocolate (10)	2006 (8)	bambu (9)
	Snow (12)	mountains (26)	recipes (9)	recipes (9)	water (8)	brasil (8)

References

[A9 2005] A9, http://a9.com/.
[ACM 2006] ACM Computing Classification Systems, http://www.acm.org/class/.
[Adamic et al. 2003] Adamic, L. A., Buyukkokten, O., and Adar, E., 2003, *A social network caught in the Web,* First Monday, Volume 8, Issue 6, June 2003. See also http://www.hpl.hp.com/research/idl/people/eytan/social.pdf.
[AIFF 1989] Apple Computer, Inc., 1989, *Audio Interchange File Format: „AIFF". A Standard for Sampled Sound Files Version 1.3,* http://www.tsp.ece.mcgill.ca/MMSP/Documents/AudioFormats/AIFF/Docs/AIFF-1.3.pdf, Accessed June 13th, 2005.
[Alexa 2006] Alexa Internet, Inc., 2006, *Related Info for: myspace.com,* http://www.alexa.com/data/details/traffic_details?q=&url=www.myspace.com, Accessed May 17th, 2006.
[Alexander 2006] Alexander, http://alexander.iicm.edu/.
[Anderson 2005] Anderson, T., 2005, *Distance learning – Social software's killer app?,* Proceedings of the 17th Biennial Conference of the Open and Distance Learning Association of Australia, Adelaide, Australia, November 2005. See also http://www.unisa.edu.au/odlaaconference/PPDF2s/13%20odlaa%20-%20Anderson.pdf.
[Andrews et al. 1995] Andrews, K., Maurer, H., and Kappe, F., 1995, *Hyper-G and Harmony: Towards the Next Generation of Networked Information Technology,* Proceedings of CHI'95, Denver, CO, U.S.A., May 1995, pp. 33-34.
[Andrus 2005] Andrus, D. C., 2005, *The Wiki and the Blog: Toward a Complex Adaptive Intelligence Community,* The Social Science Research Network (SSRN), September 2005. See also http://ssrn.com/abstract=755904.
[Annunziato et al. 2002] Annunziato, M., Lucchetti, M., and Pizzuti, S., 2002, *Adaptive Systems and Evolutionary Neural Networks: a Survey,* Proceedings of EUNITE02, Albufeira, Portugal, September 2002. See also http://home.tele2.it/stefanopizzuti/papers/HMAS2002p_13486.pdf.
[Anonymizer 2006] Anonymizer, http://www.anonymizer.com/.
[Answers 2006] Answers.com, http://www.answers.com/.
[Apache 2005] Apache HTTP Server, http://httpd.apache.org/.
[AskJeeves 2005] AskJeeves, http://www.askjeeves.com/.
[Autonomy 2005] Sybase, Inc., 2005, *A Better Search Engine with Sybase EAServer and Autonomy®,* http://www.sybase.com/content/1034309/EAS_CS_v2.pdf, Visited February 20th, 2005.
[Bahr 2002] Bahr, M., 2002, *The Wayback Machine und Google Cache – eine Verletzung deutschen Urheberrechts?,* http://www.jurpc.de/aufsatz/20020029.htm, Accessed May 11th, 2005.
[Balkin et al. 1999] Balkin, J. M., Noveck, B. S., and Roosevelt, K., 1999, *Filtering the Internet: A Best Practices Model,* http://islandia.law.yale.edu/isp/backup%2010-2003/papers/Filters0208.pdf, Accessed September 19th, 2005.
[BBC 2005] BBC News, 2005, *Berners-Lee on the read/write web,* http://news.bbc.co.uk/1/hi/technology/4132752.stm, Accessed May 10th, 2006.
[BBC 2006a] BBC iMP, http://www.bbc.co.uk/imp/.
[BBC 2006b] BBC Backstage, http://backstage.bbc.co.uk/.
[Benyon and Murray 1993] Benyon, D., and Murray, D., 1993, *Developing Adaptive Systems to Fit Individual Aptitudes,* Proceedings of the 1st International Conference on Intellligent User Interfaces, Orlando, FL, U.S.A., pp. 115-121.
[Berners-Lee 1994] Berners-Lee, T., 1994, *Universal Resource Identifiers in WWW. A Unifying Syntax for the Expression of Names and Addresses of Objects on the Network as Used in the World-Wide-Web,* Request for Comments 1630. See also http://www.w3.org/Addressing/rfc1630.txt.
[Berthold et al. 2000] Berthold, O., Federrath, H., and Köhntopp, M., 2000, *Project "Anonymity and Unobservability in the Internet",* Proceedings of the Tenth Conference on Computers, Freedom and Privacy: Challenging the Assumptions, Toronto, Canada, April 2000, pp. 57-65.
[Bharat et al. 2001] Bharat, K., Chang, B.-W., Henzinger, M., and Ruhl, M., 2001, *Who Links to Whom: Mining Linkage between Web Sites,* Proceedings of the First IEEE International Conference on Data Mining (ICDM'01), San Jose, CA, U.S.A., November 2001, pp. 51-58. See also http://toc.lcs.mit.edu/~ruhl/papers/2001-icdm.pdf.
[Biever 2005] Biever, C., 2005, *'Podcasters' deliver radio-on-demand,* New Scientist, Volume 185, Issue 2486, February 2005.
[Blogger 2005] Blogger, http://www.blogger.com/.

[Blood 2002] Blood, R., 2002, *The Weblog Handbook: Practical Advice on Creating and Maintaining Your Blog,* Perseus Publishing, Cambridge, MA.
[BMW 2006] Vodcast from BMW, http://vodcast.bmw.com/.
[Brandt 2005] Brandt, D., 2005, *Wikipedia Watch,* http://www.wikipedia-watch.org/, Accessed January 12th, 2006
[Britannica 2005] Encyclopædia Britannica, http://www.britannica.co.uk/.
[Britannica 2006] Encyclopædia Britannica, 2006, *Why try Britannica Online?,* http://www.britannica.com/premium/, Accessed January 9th, 2006.
[Brockhaus 2006] Brockhaus Enzyklopädie Digital, http://www.brockhaus-multimedial.de/.
[Bruns 2005a] Bruns, A., 2005, *Wikinews: The Next Generation of Alternative Online News?,* http://eprints.qut.edu.au/archive/00002288/01/Wikinews.pdf, Accessed January 14th, 2005.
[Bruns 2005b] Bruns, A., 2005, *Questions for Wikinewsians,* http://snurb.info/index.php?q=node/279, Accessed January 18th, 2006.
[Brusilovsky 1996] Brusilovsky, P., 1996, *Methods and techniques of adaptive hypermedia,* User Modeling and User Adapted Interaction, Volume 6, Issue 2-3, July 1996, pp. 87-129.
[Brusilovsky 2000] Brusilovsky, P., 2000, *Adaptive Hypermedia: From Intelligent Tutoring Systems to Web-Based Education.* In: Gauthier, G., Frasson, C., and VanLehn, Intelligent Tutoring Systems. Lecture Notes in Computer Science, Volume 1839, Springer-Verlag Berlin Heidelberg, 2000, pp. 1-7.
[Brusilovsky and Maybury 2002] Brusiklovsky, P., and Maybury, M. T., 2002, *From Adaptive Hypermedia to the Adaptive Web,* Communications of the ACM, Volume 45, Issue 5, May 2002, pp. 31-33.
[Buchanan et al. 2004] Buchanan, G., Blandford, A., Thimbleby, H., and Jones, M., 2004, *Integrating Information Seeking and Structuring: Exploring the Role of Spatial Hypertext in a Digital Library,* Proceedings of the 15th ACM Conference on Hypertext and Hypermedia, Santa Cruz, CA, U.S.A., August 2004, pp. 225-234.
[Burkhard 2004] Burkhard, R. A., 2004, *Learning from Architects: The Difference between Knowledge Visualization and Information Visualization,* Proceedings of the Eight International Conference on Information Visualisation (IV'04), London, U.K., July 2004, pp. 519-524.
[Burnett and Buerkle 2004] Burnett, G., and Buerkle, H., 2004, *Information Exchange in Virtual Communities: A Comparative Study,* Journal of Computer-Mediated Communication, Volume 9, Issue 2, January 2004. See also http://jcmc.indiana.edu/vol9/issue2/burnett.html.
[Bush 1945] Bush, V., 1945, *As We May Think,* Atlantic Monthly, July 1945, pp. 101-108. See also http://www.theatlantic.com/unbound/flashbks/computer/bushf.htm, Visited February 8th, 2005.
[Butterfield 2005] Butterfield, S., *The New New things,* FlickrBlog, http://blog.flickr.com/flickrblog/2005/08/the_new_new_thi.html, Accessed November 13th, 2006.
[Campbell 1999] Campbell, D., 1999, *Dublin Core Metadata and the Australian MetaWeb Project,* Proceedings of the Coordination Support Branch at the 10th National Library Technicians' Congerence, Fremantle, WA, Australia, September 1999. See also http://www.nla.gov.au/nla/staffpaper/dcampbell1.html.
[CC 2006] Creative Commons, http://creativecommons.org/.
[Chester and Gwynne 1998] Chester, A., and Gwynne, G., 1998, *Online Teaching: Encouraging Collaboration through Anonymity,* Journal of Computer-Mediated Communication, Volume 4, Issue 2, December 1998. See also http://jcmc.indiana.edu/vol4/issue2/chester.html.
[cIDf 2002] Content ID Forum, cIDf, 2002, *cIDf Specification 1.1,* September 11th, 2002. See also http://www.cidf.org/english/documents/cIDfSpecV11R11E.pdf.
[CiteULike 2006] CiteULike, http://www.citeulike.org/.
[Citizendium 2006] Citizendium, http://www.citizendium.org/.
[Clusty 2005] Clusty, http://www.clusty.com/.
[Columbia 2006] Columbia Encyclopedia, http://www.bartleby.com/65/.
[Davis 2005] Davis, I., 2005, *Talis, Web 2.0 and All That,* http://internetalchemy.org/2005/07/talis-web-20-and-all-that, Accessed January 27th, 2006.
[DC 2004] Dublin Core Metadata Initiative, 2004, *Dublin Core Metadata Element Set, Version 1.1: Reference Description,* http://dublincore.org/documents/dces/, Accessed April 29th, 2006.
[DC 2005] Dublin Core Metadata Initiative, 2005, *DCMI Metadata Terms,* http://dublincore.org/documents/dcmi-terms/, Accessed August 23rd, 2006.
[DC 2006a] Dublin Core Metadata Initiative, 2006, *DCMI Encoding Guidelines,* http://dublincore.org/resources/expressions/, Accessed April 27th, 2006.

[DC 2006b] Dublin Core Metadata Initiative, 2006, *Dublin Core Projects – Alphabetical,* http://dublincore.org/projects/, Accessed August 22nd, 2006.
[DCMI 2006] Dublin Core Metadata Initiative, http://dublincore.org/.
[DDC 2006] Dewey Decimal Classification, http://www.oclc.org/dewey/.
[del.icio.us 2006] del.icio.us, http://del.icio.us/.
[Denning et al. 2005] Denning, P., Horning, J., Parnas, D., and Weinstein, L., 2005, *Wikipedia Risks,* Communications of the ACM, Volume 48, Issue 12, December 2005, p. 152.
[DeRoure et al. 2001] DeRoure, D., Hall, W., Reich, S., et al., 2001, *Memoir – An Open Distributed Framework for Enhanced Navigation of Distributed Information,* Information Processing and Management, Volume 37, Issue 1, 2001, pp. 53-74. See also http://eprints.ecs.soton.ac.uk/4240/01/2001-4240.pdf.
[Di Iorio and Vitali 2003] Di Iorio, A., and Vitali, F., 2003, *A Xanalogical Collaborative Editing Environment,* Proceedings of the Second International Workshop of Web Document Analysis (WDA2003), Edinburgh, UK, August 2003. See also http://www.csc.liv.ac.uk/~wda2003/Papers/Section_III/Paper_11.pdf.
[DOI 2003] International Digital Object Identifier Foundation, 2003, *The DOI® Handbook Version 2.6.0,* January 2003. See also http://dx.doi.org/10.1000/186.
[DOM 1998] Wood, L., Apparao, V., Byrne, S., et al., 1998, *Document Object Model (DOM) Level 1 Specification Version 1.0,* http://www.w3.org/TR/1998/REC-DOM-Level-1-19981001/, Accessed March 29th, 2005.
[DOM 2005] Le Hégaret, P., Whitmer, R., and Wood, L., 2005, *W3C Document Object Model (DOM),* http://www.w3.org/DOM/, Accessed March 29th, 2005.
[Downes 2005] Downes, S., 2005, *E-learning 2.0,* ACM eLearn Magazine, Volume 2005, Issue 10, October 2005.
[Dreamweaver 2006] Dreamweaver, http://www.adobe.com/products/dreamweaver/.
[DSP 2005] Apple DVD Studio Pro, http://www.apple.com/finalcutstudio/dvdstudiopro/.
[DukeCast 2006] DukeCast, http://dukecast.oit.duke.edu/.
[Dunbar 1993] Dunbar, R. I. M., 1993, *Co-Evolution of Neocortex Size, Group Size and Language in Humans,* Behavioral and Brain Sciences, Volume 16, Issue 4, 1993, pp. 186-735. See also http://www.bbsonline.org/documents/a/00/00/05/65/bbs00000565-00/bbs.dunbar.html.
[Dunning 2000] Dunning, A., 2000, *Designing Shakespeare,* http://ahds.ac.uk/performingarts/creating/case-studies/shakespeare/index.htm, Accessed September 24th, 2006.
[eBay 2005] eBay, http://www.ebay.com/.
[Ebner and Zechner 2006] Ebner, M., and Zechner, J., 2006, *Das BauWiki zum Thema: Konstruktiver Stahlbetonbau,* http://bauwiki.tugraz.at/, Accessed January 9th, 2006.
[ECMA 1999] ECMA, 1999, *ECMAScript Language Specification, 3rd Edition,* Standard ECMA-262. See also http://www.ecma-international.org/publications/files/ECMA-ST/Ecma-262.pdf.
[Efimova and de Moor 2004] Efimova, L., and de Moor, A., 2004, *Beyond Personal Webpublishing: An Exploratory Study of Conversational Blogging Practices,* Proceedings of the 37th Annual Hawaii Internation Conference on System Sciences (HICSS'04), Big Island, HI, U.S.A. See also http://doi.ieeecomputersociety.org/10.1109/HICSS.2005.118.
[eInk 2005] LG.Philips LCD, E Ink Corporation, 2005, *Tablet-Sized Felxible Display,* http://www.eink.com/press/images/highres_downloads/LG.Philips_LCD_E_E_Ink_Flex_Tablet_Display.jpg, Accessed May 7th, 2006.
[eInk 2006] E Ink, http://www.eink.com/.
[Elderez 1997] Elderez, S., 1997, *Information encountering: a conceptual framework for accidental information discovery,* Proceedings of an Internation Conference on Information Seeking in Context (ISIC), Tampere, Finland, 1997, pp. 412-421.
[Elgg 2006] Elgg, http://elgg.org/.
[Emigh and Herring 2005] Emigh, W., and Herring, S. C., 2005, *Collaborative Authoring on the Web: A Genre Analysis of Online Encyclopedias,* Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS-38), Big Island, HI, USA, Track 4, p. 99a.
[Eurekster 2006] Eurekster, http://www.eurekster.com/.
[Eventful 2006] Eventful, http://eventful.com/.
[Everything2 2006] Everything2, http://www.everything2.org/.
[Findlater and McGrenere 2004] Findlater, L., and McGrenere, J., 2004, *A Comparison of Static, Adaptive, and Adaptable Menus,* Proceedings of the 2004 SIGCHI Conference on Human Factors in Computing Systems (CHI 2004), Vienna, Austria, April 2004, pp. 89-96.

[Flickr 2006] Flickr, http://www.flickr.com/.
[Flinn and Maurer 1995] Flinn, B., and Maurer, H., 1995, *Levels of Anonymity,* Journal of Universal Computer Science, Volume 1, Issue 1, January 1995, pp. 35-47. See also http://www.jucs.org/jucs_1_1/levels_of_anonymity.
[Foley et al. 1997] Foley, J. D., van Dam, A., et al., 1997, *Computer Graphics: Principles and Practice,* Second Edition in C, Addison Wesley.
[Fountain et al. 1990] Fountain, A. M., Hall, W., Heath, I., and Davis, H. C., 1990, *MICROCOSM: An Open Model for Hypermedia with Dynamic Linking,* Proceedings of Hypertext: Concepts, Systems and Applications, Paris, France, pp. 298-311.
[Foust 1994] Foust, J. A., 1994, *Welcome to the Interpedia!,* Posting in the comp.infosystems.interpedia newsgroup. Retrieved from http://groups.google.com/group/comp.infosystems.interpedia/browse_thread/thread/6ec8bc1f8bece5a1/590f92a599e06589#590f92a599e06589.
[Freyne and Smyth 2004] Freyne, J., and Smyth, B., 2004, *An Experiment in Social Search,* Proceedings of the 3rd International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems (AH2004), Eindhoven, The Netherlands, August 2004.
[Friendster 2006] Friendster, http://www.friendster.com/.
[FrontPage 2006] Microsoft FrontPage, http://www.microsoft.com/frontpage/.
[Fu et al. 2001] Fu, X., Shi, W., Akkerman, A., et al., 2001, *CANS: Composable, Adaptive Network Services Infrastructure,* Proceedings of the 3rd USENIX Symposium on Internet Technologies and Systems (USITS01), San Francisco, CA, U.S.A., March 2001. See also http://www.usenix.org/events/usits01/full_papers/fu/fu.pdf.
[Furl 2006] LookSmart Furl, http://furl.net/.
[Furnas et al. 1987] Furnas, G. W., Landauer, T. K., Gomez, L. M., and Dumais, S. T., 1987, *The Vocabulary Problem in Human-System Communication,* Communications of the ACM, Volume 30, Issue 11, November 1987, pp. 964-971. See also http://www.si.umich.edu/~furnas/Papers/vocab.paper.pdf.
[Furnas et al. 2006] Furnas, G. W., Fake, C., von Ahn, L., et al., 2006, *Why Do Tagging Systems Work,* Proceedings and Extended Abstracts of the Conference on Human Factors in Computing Systems (CHI 2006), Montréal, QC, Canada, April 2006, pp. 36-39.
[Furuta et al. 1997] Furuta, R., Shipman, F. M., Marshall, C. C., et al., 1997, *Hypertext Paths and the World Wide Web: Experiences with Walden's Paths,* Proceedings of the 8th ACM Conference on Hypertext and Hypermedia, Southampton, UK, April 1997, pp. 167-176.
[Getty 2004] The Getty Museum, http://www.getty.edu/art/.
[Getty 2006a] Getty Thesaurus of Geographic Names, http://www.getty.edu/research/tools/vocabulary/tgn/
[Getty 2006b] Art & Architecture Thesaurus, http://www.getty.edu/research/conducting_research/vocabularies/aat/.
[GFDL 2005] Free Software Foundation, Inc., 2005, *GNU Free Documentation License,* http://www.gnu.org/copyleft/fdl.html, Visited February 8th, 2005.
[Giles 2005] Giles, J., 2005, *Internet encyclopaedias go head to head,* Nature, Volume 438, December 2005, pp. 900-901. See also http://www.nature.com/nature/journal/v438/n7070/full/438900a.html.
[Gillmor 2004] Gillmor, D., 2004, *We the Media,* O'Reilly Media, Sebastopol, CA, U.S.A.
[Glushko et al. 1988] Glushko, R. J., Weaver, M. D., Coonan, T. A., and Lincoln, J. E., 1988, *"Hypertext Engineering": Practical Methods for Creating a Compact Disc Encyclopedia,* Proceedings of the ACM Conference on Document Processing Systems, Santa Fe, NM, U.S.A., December 1988, pp. 11-19.
[Golder and Huberman 2005] Golder, S., and Huberman, B. A., 2005, *The Structure of Collaborative Tagging Systems,* http://www.hpl.hp.com/research/idl/papers/tags/tags.pdf, Accessed October 11th, 2006.
[Google 2005] Google Answers, http://answers.google.com/.
[Google 2006] Google Earth, http://earth.google.com/.
[Gordon 1999] Gordon, D. M., 1999, *Close Encounters,* The Sciences, September/October 1999. See also http://www.shaav.com/professional/shad/ants.pdf.
[Gordon 2000] Gordon, D. M., 2000, *Ants at Work: How an Insect Society is Organized,* W. W. Norton & Company, New York, 2000.
[Guinchard 2002] Guinchard, C., 2002, *Dublin Core use in libraries: a survey,* OCLC Systems & Services, Volume 18, Issue 1, pp. 40-50.
[GuruNet 2005] GuruNet, http://www.gurunet.com/.
[Guy and Tonkin 2006] Guy, M., and Tonkin, E., 2006, *Folksonomies: Tidying Up Tags?,* D-Lib Magazine, Volume 12, Issue 1, January 2006. See also http://www.dlib.org/dlib/january06/guy/01guy.html.

[Haake and Hicks 1996] Haake, A., and Hicks, D., 1996, *VerSE: Towards Hypertext Versioning Styles,* Proceedings of the Seventh ACM Conference on Hypertext, Washington DC, U.S.A., March 1996. See also http://www.dilip.info/HT96/P45/HT96HaakeHicks.html.

[Ham 2005] Ham, P., 2005, *Applying Research to Design Online Learning Environments,* International Journal of the Computer, the Internet and Management, Special Issue August 2005. See also http://www.journal.au.edu/ijcim/2005/specialaug05/PP26.pdf.

[Hammond et al. 2005] Hammond, T., Hannay, T., Lund, B., and Scott, J., 2005, *Social Bookmarking Tools (I): A General Review,* D-Lib Magazine, Volume 11, Issue 4, April 2005. See also http://www.dlib.org//dlib/april05/hammond/04hammond.html.

[Heinrich and Maurer 2000] Heinrich, E., and Maurer, H., 2000, *Active Documents: Concept, Implementation and Applications,* Journal of Universal Computer Science, Volume 6, Issue 12, December 2000, pp. 1197-1202. See also http://www.jucs.org/jucs_6_12/active_documents_concept_implementation/.

[Heinrich et al. 2001] Heinrich, E., Johnson, R., Luo, D., Maurer, H., and Sapper, M., 2003, *Learner-Formulated Questions in Technology-Supported Learning Applications,* Proceedings of ED-MEDIA 2001, Tampere, Finland, pp. 720-725. See also http://www.iicm.edu/iicm_papers/learner_formulated_questions.pdf.

[Herring et al. 2004] Herring, S. C., Scheidt, L. A., Bonus, S., and Wright, E., 2004, *Bridging the Gap: A Genre Analysis of Weblogs,* Proceedings of the 37th Annual Hawaii Internation Conference on System Sciences (HICSS'04), Big Island, HI, U.S.A. See also http://doi.ieeecomputersociety.org/10.1109/HICSS.2004.1265271.

[Heymann and Garcia-Molina 2006] Heymann, P., and Garcia-Molina, H., 2006, *Collaborative Creation of Communial Hierarchical Taxonomies in Social Tagging Systems,* Stanford InfoLab Technical Report 2006-10, http://arxiv.org/abs/cs.DL/0508082, Accessed November 10th, 2006.

[Holland 1962] Holland, J. H., 1962, *Outline for a Logical Theory of Adaptive Systems,* Journal of the ACM (JACM), Volume 9, Issue 3, July 1962, pp. 297-314.

[HotI 2006] Hot Interconnects, http://www.hoti.org/.

[HP 2004] Firth, S., 2004, *Photographic memories: Always-on camera captures life's fleeting moments,* http://www.hpl.hp.com/news/2004/jan-mar/casualcapture.html, Accessed May 21st, 2006.

[HTML 1999] Raggett, D., Le Hors, A., and Jacobs, I., 1999, *HTML 4.01 Specification,* http://www.w3.org/TR/html4/, Accessed April 28th, 2005.

[HTTP 1999] Fielding, R., Gettys, J., Mogul, J., et al., 1999, *Hypertext Transfer Protocol – HTTP/1.1,* Request for Comments 2616. See also ftp://ftp.isi.edu/in-notes/rfc2616.txt.

[Hutchins 2001] Hutchins, W. J., 2001, *Machine translation over fifty years,* Histoire, Epistemologie, Langage, Volume 23, Issue 1, pp. 7-31. See also http://ourworld.compuserve.com/homepages/wjhutchins/HEL.pdf.

[Hyperwave 2005] Hyperwave, http://www.hyperwave.com/.

[Ide and Véronis 1998] Ide, N., and Véronis, H., 1998, *Word Sense Disambiguation: The State of the Art,* Computational Linguistics, Volume 24, Issue 1, March 1998, pp. 1-40. See also http://www.up.univ-mrs.fr/~veronis/pdf/1998wsd.pdf.

[IEEE 2006] IEEE Spectrum Radio, http://spectrum.ieee.org/radio/.

[Indymedia 2006] Indymedia (Independent Media Center), http://www.indymedia.org/.

[Ingrassia 2005] Ingrassia, L., 2005, *wiki-whatdia?,* Accessed through http://poynter.org/forum/view_post.asp?id=10748, Accessed January 8th, 2006.

[iPhoto 2006] iPhoto, http://www.apple.com/ilife/iphoto/features/photocasting.html.

[iTunes 2005] iTunes Music Store, http://www.apple.com/itunes/store/.

[Ivanova 2004] Ivanova, N., 2004, *Metadata and Australian Universities: an Environmental Scan,* Proceedings of the Tenth Australian World Wide Web Conference, Gold Coast, QLD, Australia, July 2004. See also http://ausweb.scu.edu.au/aw04/papers/refereed/ivanova/paper.html.

[JAP 2006] JAP, http://anon.inf.tu-dresden.de/.

[JISC-CETIS 2005] JISC/CETIS Conference Podcast, http://www.e-framework.org/events/conference/conference_audio.xml.

[Johnson 2001] Johnson, S., 2001, *Emergence. The Connected Lives of Ants, Brains, Cities and Software,* Allen Lane The Penguin Press, 2001.

[Joinson 2001] Joinson, A. N., 2001, *Self-disclosure in computer-mediated communication: The role of self-awareness and visual anonymity,* European Journal of Social Psychology, Volume 31, Issue 2, March/April 2001, pp. 177-192.

[JUCS 2005] Journal of Universal Computer Science (J.UCS), http://www.jucs.org/.

[Kappe 1995] Kappe, F., 1995, *Maintaining Link Consistency in Distributed Hyperwebs,* Proceedings of the INET'95 Conference, Honolulu, HI, U.S.A., June 1995. See also http://www.isoc.org/HMP/PAPER/073/html/paper.html.

[Kautz et al. 1997] Kautz, H., Selman, B., and Shah, M., 1997, *Referral Web: Combining Social Networks and Collaborative Filtering,* Communications of the ACM, Volume 40, Issue 3, March 1997, pp 63-65.

[Koch 2004] Koch, P.-P., 2004, *JavaScript – Get selection,* http://www.quirksmode.org/js/selected.html, Accessed May 10th, 2005.

[Kolbitsch 2005a] Transclusions in HTML-Based Environments, http://www.kolbitsch.org/research/transclusions/.

[Kolbitsch 2005b] Fine-Grained Transclusions of Multimedia Documents in HTML, http://www.kolbitsch.org/research/mm_transclusions/.

[Kolbitsch 2006a] Analysis: Anonymous Users in Wikipedia, http://www.kolbitsch.org/research/wikipedia_anonymous_users_analysis/.

[Kolbitsch 2006b] WordFlickr, http://www.kolbitsch.org/research/wordflickr/.

[Kolbitsch 2006c] FlickrClustr, http://www.kolbitsch.org/research/flickrclustr/.

[Koman 2002] Koman, R., 2002, *How the Wayback Machine Works,* http://webservices.xml.com/lpt/a/ws/2002/01/18/brewster.html, Accessed May 10th, 2005.

[Korica et al. 2006] Korica, P., Maurer, H., and Schinagl, W., 2006, *The Growing Importance of e-Communities on the Web,* Proceedings of the IADIS International Conference on Web Based Communities, San Sebasitan, Spain, February 2006.

[Krottmaier 2002] Krottmaier, H., 2002, *Transcluded Documents: Advantages of Reusing Document Fragments,* Proceedings of the 6th International ICCC/IFIP Conference on Electronic Publishing (ELPUB2002), Karlovy Vary, Czech Republic, November 2002, pp. 359-367. See also http://hkrott.iicm.edu/docs/publications/elpub-2002.pdf.

[Krottmaier 2003] Krottmaier, H., 2003, *Enhanced Annotations,* Proceedings of the International Conference on Society for Information Technology and Teacher Education (SITE 2003), pp. 991-993. See also http://dl.aace.org/11854.

[Krottmaier and Helic 2002] Krottmaier, H., and Helic, D., 2002, *Issues of Transclusions,* Proceedings of the World Conference on E-Learning in Corporate, Government, Healthcare, & Higher Education (E-Learn 2002), Montreal, Canada, October 2002, pp. 1730-1733. See also http://coronet.iicm.edu/denis/pubs/elearn2002b.pdf.

[Krottmaier and Maurer 2001] Krottmaier, H., and Maurer, H., 2001, *Transclusions in the 21st Century,* Journal of Universal Computer Science, Volume 7, Issue 12, December 2001, pp. 1125-1136. See also http://www.jucs.org/jucs_7_12/transclusions_in_the_21st/.

[Kunze 1999] Kunze, J., 1999, *Encoding Dublin Core Metadata in HTML,* IETF RFC 2731, http://www.ietf.org/rfc/rfc2731.txt, Accessed April 29th, 2006.

[Langley 1997] Langley, P., 1997, *Machine Learning for Adaptive User Interfaces,* Proceedings of the 12th German Annual Conference on Artificial Intelligence, Freiburg, Germany, September 1997, pp. 53-62. See also http://lyonesse.stanford.edu/~langley/papers/adapt.ki97.ps.

[Langley and Fehling 1998] Langley, P., and Fehling, M., 1998, *The Experimental Study of Adaptive User Interfaces,* Technical Report, Institute for the Study of Learning and Expertise, March 1998, Palo Alto, CA, U.S.A. See also http://lyonesse.stanford.edu/~langley/papers/adapt.exper.ps.

[Lenhart et al. 2004] Lenhart, A., Horrigan, J., and Fallows, D., 2004, *Content Creation Online: 44 % of U.S. Internet users have contributed their thoughts and their files to the online world,* http://www.pewinternet.org/pdfs/PIP_Content_Creation_Report.pdf, Accessed January 27th, 2005.

[Lennon and Maurer 2003] Lennon, J., and Maurer, H., 2003, *Why it is Difficult to Introduce e-Learning into Schools and Some New Solutions,* Journal of Universal Computer Science, Volume 9, Issue 10, October 2003, pp. 1244-1257. See also http://www.jucs.org/jucs_9_10/why_it_is_difficult/.

[Lenya 2006] Apache Lenya, http://lenya.apache.org/.

[Leuf and Cunningham 2001] Leuf, B., and Cunningham, W., 2001, *The Wiki Way. Quick Collaboration on the Web,* Addison-Wesley, 2001.

[Lih 2004] Lih, A., 2004, *Wikipedia as Participatory Journalism: Reliable Sources? Metrics for evaluating collaborative media as a news resource,* Proceedings of the 5th International Symposium on Online Journalism, Austin, TX, USA, April 2004. See also http://jmsc.hku.hk/faculty/alih/publications/utaustin-2004-wikipedia-rc2.pdf.

[Lindahl and Blount 2003] Lindahl, C., and Blount, E., 2003, *Weblogs: Simplifying Web Publishing,* Computer, Volume 36, Issue 11, November 2003, pp. 114-116.
[Liu and Silverman 2001] Liu, S., and Silverman, M, 2001, *A Practical Guide to Biometric Security Technology,* IT Professional, Volume 3, Issue 1, January 2001, pp. 27-32.
[Lum and Lau 2002] Lum, W. Y., and Lau, F. C. M., 2002, *A Context-Aware Decision Engine for Content Adaptation,* Pervasice Computing, Volume 1, Issue 3, July 2002, pp. 41-49. See also http://doi.ieeecomputersociety.org/10.1109/MPRV.2002.1037721.
[Ma et al. 2003] Ma, L., Tan, T., Wang, Y., and Zhang, D., 2003, *Personal Identification Based on Iris Texture Analysis,* IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 25, Issue 12, pp. 1519-1533.
[Macromedia 2005] Macromedia, Inc., 2005, *Macromedia Flash File Format (SWF) Specification License,* http://www.macromedia.com/software/flash/open/licensing/fileformat/, Accessed June 16th, 2005.
[Marlow et al. 2006] Marlow, C., Naaman, M., Davis, M., and boyd, d. (sic), 2006, *Taggin Paper, Taxonomy, Flickr, Academic Article, ToRead* (sic), Proceedings of the Seventeenth Conference on Hypertext and Hypermedia (HT06), Odense, Denmark, August 2006, pp. 31-40. See also http://alumni.media.mit.edu/~cameron/cv/pubs/2006-ht06-tagging-paper.
[Marshall 1997] Marshall, C. C., 1997, *Annotation: From Paper Books to Digital Library,* Proceedings of the ACM Conference on Digital Libraries, Philadelpha, PA, U.S.A., July 1997, pp. 131-140.
[Marshall and Bly 2004] Marshall, C. C., and Bly, S., 2004, *Sharing Encountered Information: Digital Libraries Get a Social Life,* Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries, Tucson, AZ, U.S.A., June 2004, pp. 218-227.
[Marshall and Shipman 1995] Marshall, C. C., and Shipman, F. M., 1995, *Spatial Hypertext: Designing for Change,* Communications of the ACM, Volume 38, Issue 8, August 1995, pp. 88-97. See also http://www.csdl.tamu.edu/~shipman/papers/cacm-1995.pdf.
[Mathes 2004] Mathes, A., 2004, *Folksonomies – Cooperative Classification and Communication Through Shared Metadata,* http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.pdf, Accessed January 25th, 2006.
[Maurer 1988] Maurer, H., 1988, *Privacy and Security on Videotex Systems,* Communication Technologies for the 1990s, 1988, pp. 49-55.
[Maurer 1996] Maurer, H., 1996, *Hyperwave: The Next Generation Web Solution,* Addison-Wesley Longman, London, UK, April 1996.
[Maurer 1998] Maurer, H., 1998, *Web-Based Knowledge Management,* Computer, Volume 31, Issue 3, March 1998, pp. 122-123.
[Maurer 2001] Maurer, H., 2001, *Beyond Classical Digital Libraries,* Proceedings of the NIT Conference, Global Digital Library Development in the New Millenium, Beijing, Tsinghua University Press, 2001, pp. 165-173. See also http://www.iicm.edu/iicm_papers/beyond_digital_libraries.doc.
[Maurer and Lennon 1996] Maurer, H., and Lennon, J. A., 1996, *Flexible Link Architecture in Hypermedia Systems,* Proceedings of ED-MEDIA'96, Boston, MA, U.S.A., pp. 384-389.
[Maurer and Tochtermann 2002] Maurer, H., and Tochtermann, K., 2002, *On a New Powerful Model for Knowledge Management and its Applications,* Journal of Universal Computer Science, Volume 8, Issue 1, January 2002, pp. 85-96. See also http://www.jucs.org/jucs_8_1/on_a_new_powerful/.
[Maurer et al. 2006] Maurer, H., Krottmaier, H., and Dreher, H., 2006, *Important Aspects of Modern Digital Libraries,* Proceedings of the Second International Conference on Digital Libraries (ICDL 2006), New Delhi, India, December 2006, pp. 843-855.
[McHenry 2004] McHenry, R., 2004, *The Faith-Based Encyclopaedia,* http://www.techcentralstation.com/111504A.html, Accessed February 25th, 2006.
[Microsoft 2005a] Microsoft Corporation, 2005, *Adaptive Systems and Interaction,* http://research.microsoft.com/adapt/, Accessed August 7th, 2005.
[Microsoft 2005b] Microsoft Corporation, 2005, *About Cross-Frame Scripting and Security,* http://msdn.microsoft.com/workshop/author/om/xframe_scripting_security.asp, Accessed May 10th, 2005.
[Milgram 1967] Milgram, S., 1967, *The Small World Problem,* Psychology Today, Issue 2, May 1967, pp 60-67.
[Millen et al. 2005] Millen, D., Feinberg, J., Kerr, B., 2005, *Social Bookmarking in the Enterprise,* ACM Queue, Volume 3, Issue 9, November 2005, pp. 28-35.
[Miller 2005] Miller, P., 2005, *Web 2.0: Building the New Library,* Ariadne, Issue 45, October 2005. See also http://www.ariadne.ac.uk/issue45/miller/.

[MissIzzy 2006] Miss Izzy, http://www.missizzy.org/.
[Mödritscher et al. 2004] Mödritscher, F., García-Barrios, V. M., and Gütl, C., 2004, *The Past, the Present and the Future of adaptive E-Learning. An Approach within the Scope of the Reaserch Project AdeLE,* Proceedings of the International Conference on Interactive Computer Aided Learning (ICL2004), Villach, Austria, September 2004. See also http://www.iicm.edu/iicm_papers/icl2004/adaptive_e-learning/adaptiv_e-learning.pdf.
[Moore et al. 2001] Moore, A., Brailsford, T. J., and Stewart, C. D., 2001, *Personally tailored teaching in WHURLE using conditional transclusion,* Proceedings of the Twelfth ACM Conference on Hypertext and Hypermedia, Århus, Denmark, August 2001, pp. 163-164. See also http://whurle.sourceforge.net/whurle-ht01-short-paper-final.pdf.
[MPEG-7 2002] International Organisation for Standardisation ISO/IEC JTC1/SC29/WG11, 2002, *MPEG-7 Overview (version 8),* http://mpeg.telecomitalialab.com/standard/mpeg-7/mpeg-7.htm, besucht am 7. November 2002.
[Mülner 2001] Mülner, H., 2001, *Software für ein multimediales Lexikon,* Personal Communication.
[MySpace 2006] MySpace, http://www.myspace.com/.
[Narendra and Parthasarathy 1989] Narendra, K. S., and Parthasarathy, K., 1989, *Adaptive Identification and Control of Dynamical Systems using Neural Networks,* Proceedings of the 28th Conference on Decision and Control, Tampa, FL, U.S.A., December 1989, pp. 1737-1738.
[Nature 2006] Nature Podcast, http://www.nature.com/nature/podcast/.
[Nelson 1965] Nelson, T. H., 1965, *A File Structure for the Complex, the Changing and the Indeterminate,* Proceedings of the ACM 20th National Conference, Cleveland, OH, U.S.A., August 1965, pp. 84-100.
[Nelson 1981] Nelson, T. H., 1981, *Literary Machines,* Mindful Press, 1981.
[Nelson 1995] Nelson, T. H., 1995, *The Heart of Connection: Hypermedia Unified by Transclusion,* Communications of the ACM, Volume 38, Issue 8, August 1995, pp. 31-33.
[Nelson 1996] Nelson, T. H., 1996, *Generalized Links, Micropayment and Transcopyright,* http://www.almaden.ibm.com/almaden/npuc97/1996/tnelson.htm, Accessed May 1st, 2003.
[Nelson 1998] Nelson, T. H., 1998, *Transcopyright: Pre-Permission for Virtual Republishing*, http://www.aus.xanadu.com/xanadu/transcopy.html, Accessed May 3rd, 2005.
[Nelson 1999] Nelson, T. H., 1999, *Xanalogical Structure, Needed Now More than Ever: Parallel Documents, Deep Links to Content, Deep Versioning and Deep Re-Use,* ACM Computing Surveys, Volume 31, Issue 4es, December 1999. See also http://xanadu.com.au/ted/XUsurvey/xuDation.html.
[Netscape 2004] Netscape Communications, 2004, *JavaScript Developer Central,* http://developer.netscape.com/tech/javascript/, Accessed February 3rd, 2004.
[Novak and Wurst 2004] Novak, J., and Wurst, M., 2004, *Supporting Knowledge Creation and Sharing in Communities Based on Mapping Implicit Knowledge,* Journal of Universal Computer Science, Volume 10, Issue 3, pp. 235-251. See also http://www.jucs.org/jucs_10_3/supporting_knowledge_creation_and.
[OAI 2006] Open Archives Initiative, http://www.openarchives.org/.
[OhmyNews 2006] Ohmy News, http://english.ohmynews.com/.
[O'Murchu et al. 2004] O'Murchu, I., Breslin, J. G., and Decker, S., 2004, *Online Social and Business Networking Communities,* Technical Report, August 2004. See also http://www.nextwebgeneration.org/publications/techpapers/documents/DERI-TR-2004-08-11.pdf.
[OpenBC 2006] Open Business Club, http://www.openbc.com/.
[O'Reilly 2005] O'Reilly, T., 2005, *What Is Web 2.0. Design Patterns and Business Models for the Next Generation of Software,* http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html, Accessed January 26th, 2006.
[Orkut 2006] Orkut, http://www.orkut.com/.
[Orlowski 2004] Orlowski, A., 2004, *Avoid Friendster and its clones, warnes security expert,* The Register, February 10th, 2004, http://www.theregister.co.uk/2004/02/10/avoid_friendster_and_its_clones/, Accessed February 3rd, 2006.
[Pam 1995] Pam, A., 1995, *Where World Wide Web Went Wrong,* Proceedings of the AUUG'95 & Asia-Pacific World Wide Web '95 Conference & Exhibition, Sydney, Australia, September 1995. See also http://www.csu.edu.au/special/conference/apwww95/papers95/apam/apam.html.
[Pam 1997] Pam, A., 1997, *Fine-Grained Transclusion in the Hypertext Markup Language,* Internet Draft. See also http://xanadu.com.au/archive/draft-pam-html-fine-trans-00.txt.
[Phan et al. 2002] Phan, T., Zorpas, G., and Bagrodia, R., 2002, *An Extensible and Scalable Content Adaptation Pipeline Architecture to Support Heterogeneous Clients,* Proceedings of the 22nd International Confer-

ence on Distributed Computing Systems (ICDCS'02), Vienna, Austria, April 2002, pp. 507-516. See also http://pcl.cs.ucla.edu/papers/files/icdcs2002-cap.pdf.
[Pivec and Baumann 2004] Pivec, M., and Baumann, K., 2004, *The Role of Adaptation and Personalisation in Calssroom-Based Learning and in e-Learning,* Journal of Universal Computer Science, Volume 10, Issue 1, January 2004, pp. 73-89. See also http://www.jucs.org/jucs_10_1/the_role_of_adaptation/.
[Pod411 2006] podCast 411, 2006, *podcast directory and information,* http://www.podcast411.com/, Accessed February 3rd, 2006.
[PodStats 2006] N.N., 2006, *Podcasting Statistics,* http://www.podcastingstats.com/, Accessed February 3rd, 2006.
[Pollock 2001] Pollock, H., 2001, *Who Let the Blogs Out?,* Yahoo Internet Life, May 2001. See also http://web.archive.org/web/20010813193029/http://yil.com/features/feature.asp?volume=07&issue=05&keyword=blogs.
[Powell 2003] Powell, A., 2003, *Expressing Dublin Core in HTML/XHTML meta and link elements,* http://dublincore.org/documents/dcq-html/index.shtml, Accessed April 29th, 2006.
[Prelinger 2005] Prelinger Archives, http://www.archive.org/details/prelinger/.
[PURL 2005] Shafer, K., Weibel, S., Jul, E., and Fausey, J., 2005, *Introduction to Persistent Uniform Resource Locators,* http://purl.oclc.org/docs/inet96.html, Visited February 24th, 2005.
[QTVR 2005] Apple Computer, Inc., 2005, *QuickTime VR,* http://developer.apple.com/documentation/QuickTime/InsideQT_QTVR/insideqt_qtvr.pdf, Accessed June 13th, 2005.
[RDF 2004] Miller, E., Swick, R., and Brickley, D., 2004, *Resource Description Framework (RDF),* http://www.w3.org/RDF/, Visited February 3rd, 2005.
[Rechenberg 1973] Rechenberg, I., 1973, *Evolutionsstrategie. Optimierung technischer Systeme nach Prinzipien der biologischen Evolution,* Frommann-Holzoog, Stuttgart, 1973.
[Reiter and Rubin 1997] Reiter, M. K., and Rubin, A. D., 1997, *Crowds: Anonymity for Web Transactions,* ACM Transactions on Information System Security, DIMACS Technical Report 97-15, April 1997. See also ftp://cag.lcs.mit.edu/pub/dm/papers/reiter:crowds.ps.gz.
[Rodriguez et al. 2005] Rodriguez, P., Tan, S.-M., and Gkantsidis, C., 2005, *On the feasibility of Commercial, Legal P2P Content Distribution,* Proceedings of the IEEE 10th International Workshop on Web Content Caching and Distribution (WCW2005), Sophia Antipolis, France, September 2005. See also http://research.microsoft.com/~pablo/papers/CCR.pdf.
[RSS 1999] Libby, D., 1999, *NETSCAPE COMMUNICATIONS RSS 0.91 Spec, revision 3,* http://my.netscape.com/publish/formats/rss-spec-0.91.html, Accessed February 3rd, 2005.
[RSS 2001] Swartz, A., 2001, *RDF Site Summary (RSS) 1.0,* http://web.resource.org/rss/1.0/, Accessed February 3rd, 2005.
[Rubart et al. 2001] Rubart, J., Haake, J. M., Tietze, D. A., and Wang, W., 2001, *Organizing Shared Enterprise Workspaces Using Component-Based Cooperative Hypermedia,* Proceedings of the Twelfth ACM Conference on Hypertext and Hypermedia, Aarhus, Denmark, 2001, pp. 73-82.
[Schlimmer and Hermens 1993] Schlimmer, J. C., and Hermens, L. A., 1993, *Software Agents: Completing Patterns and Constructing User Interfaces,* Journal of Artificial Intelligence Research, Volume 1, November 1993, pp. 61-89. See also http://www.cs.cmu.edu/afs/cs/project/jair/pub/volume1/schlimmer93a.pdf.
[Schmidt and Bannon 1992] Schmidt, K., and Bannon, L., 1992, *Taking CSCW Seriously: Supporting Articulation Work,* Computer Supported Cooperative Work (CSCW): An International Journal, Volume 1, Issue 1, 1992, pp. 7-40. See also http://www.it-c.dk/people/schmidt/papers/cscw_seriously.pdf.
[Seigenthaler 2005] Seigenthaler, J., 2005, *A false Wikipedia 'biography',* USA Today, http://www.usatoday.com/news/opinion/editorials/2005-11-29-wikipedia-edit_x.htm, Accessed January 12th, 2006.
[Shirky 2005] Shirky, C., 2005, *Ontology is Overrated: Links, Tags, and Post-Hoc Metadata,* O'Reilly Emerging Technologies Conference, San Diego, CA, U.S.A., March 2005. Available online at http://www.shirky.com/writings/ontology_overrated.html, Accessed November 11th, 2006. See also http://www.itconversations.com/audio/download/itconversations-470.mp3.
[Skype 2006] Skype, http://www.skype.com/.
[Slashdot 2006] Slashdot, http://www.slashdot.org/.
[Stotts and Furuta 1991] Stotts, P. D., and Furuta, R., 1991, *Dynamic Adaptation of Hypertext Structure,* Proceedings of the Third Annual Conference on Hypertext, San Antonio, TX, U.S.A., December 1991, pp. 219-231.
[Su 2005] Su, C.-C., 2005, *An Open Source Portal for Educators,* Teaching English as a Second or Foreign Language, Volume 9, Issue 1, June 2005. See also http://writing.berkeley.edu/TESL-EJ/ej33/int.html.

[Surowiecki 2005] Surowiecki, J., 2005, *The Wisdom of Crowds. Why the Many Are Smarter Than the Few,* Little, Brown Book Group, UK, 2005.
[SVG 2003] Ferraiolo, J., Fujisawa, J., and Jackson, D. (Eds.), 2003, *Scalable Vector Graphics (SVG) 1.1 Specification,* http://www.w3.org/TR/SVG/, Accessed June 12th, 2005.
[Technorati 2006] Technorati, http://www.technorati.com/.
[Thau 2003] Thau, R. S., 2003, *Apache API notes,* http://modules.apache.org/doc/API.html, Accessed May 21st, 2005.
[Tomlin 2005] Tomlin, S., 2005, *Science in the web age: The expanding electronic universe,* Nature, Volume 438, December 2005, p. 547. See also http://www.nature.com/nature/journal/v438/n7068/full/438547a.html.
[Tonkin 2005] Tonkin, E., 2005, *Making the Case for a Wiki,* Ariadne, Issue 42, January 2005, http://www.ariadne.ac.uk/issue42/tonkin/, Accessed December 12th, 2006.
[Tonkin 2006] Tonkin, E., 2006, *Folksonomies: The Fall and Rise of Plain-text Tagging,* Ariadne, Issue 47, April 2006, http://www.ariadne.ac.uk/issue47/tonkin/, Accessed November 7th, 2006.
[Treese 2004] Treese, W., 2004, *Open Systems for Collaboration,* netWorker, Volume 8, Issue 1, March 2004, pp. 13-16.
[Typo3 2006] TYPO3, http://typo3.org/.
[UDCC 2006] Universal Decimal Classification, http://www.udcc.org/.
[van Veen 2006] van Veen, T., 2006, *Serving Services in Web 2.0,* Ariadne, Issue 47, April 2006. See http://www.ariadne.ac.uk/issue47/vanveen/.
[Vester 1978] Vester, F., 1978, *Denken, Lernen, Vergessen,* dtv (Deutscher Taschenbuch Verlag), 1978.
[Viégas and Donath 2004] Viégas, F. B., and Donath, J., 2004, *Social Network Visualization: Can We Go Beyond the Graph?,* CSCW'04 Workshop on Social Networks for Design an Analysis: Using Network Information in CSCW, 2004 ACM Conference on Computer Supported Cooperative Work (CSCW 2004), Chicago, IL, U.S.A., November 2004. See also http://www.ischool.washington.edu/mcdonald/cscw04/papers/viegas-cscw04.pdf.
[Viégas et al. 2004] Viégas, F. B., Wattenberg, M., and Kushal, D., 2004, *Studying cooperation and conflict between authors with history flow visualizations,* Proceedings of the 2004 Conference on Human Factors in Computing Systems (CHI 2004), Vienna, Austria, pp. 575-582.
[Voß 2005] Voß, J., 2005, *Measuring Wikipedia,* Proceedings of the 10th International Conference of the International Society for Scientometrics and Informetrics (ISSI 2005), Stockholm, Sweden, July 2005. See also http://eprints.rclis.org/archive/00003610/01/MeasuringWikipedia2005.pdf.
[Vossen 1999] Vossen, P. (ed.), 1999, *The Final Wordnets for Dutch, Spanish, Italian and English,* EuroWordNet Technical Report LE2-4003, http://www.illc.uva.nl/EuroWordNet/docs/D032D033PS.zip, Accessed November 15th, 2006.
[Vuorikari 2005] Vuorikari, R., 2005, *Can personal digital knowledge artefacts' management and social networks enhance learning?,* http://elgg.net/riina/files/-1/1444/social_networks_learning_vuorikari_22_9_2005.pdf, Accessed March 9th, 2006.
[W3C 2006] W3C Markup Validation Service, http://validator.w3.org/.
[Wales 2005] Wales, J., 2005, *Re: [Wiki-research-l] Re: Comparison of Wikipedia with Brittannica,* Posting on the Wiki-research-l mailing list, December 17th, 2005. See also http://mail.wikipedia.org/pipermail/wiki-research-l/2005-December/000105.html.
[Ward 2003] Ward, J., 2003, *A Quantitative Analysis of Unqualified Dublin Core Metadata Element Set Usage within Data Providers Registered with the Open Archives Initiative,* Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'03), Houston, TX, U.S.A., May 2003, pp. 315-317.
[WCL 2006] Washington College of Law Podcast, http://www.wcl.american.edu/podcast/podcast.cfm.
[WDG 2006] WDG HTML Validator, http://www.htmlhelp.com/tools/validator/.
[Weber and Specht 1997] Weber, G., and Specht, M., 1997, *User Modeling and Adaptive Navigation Support in WWW-Based Tutoring Systems,* Proceedings of the Sixth Internation Conference on User Modeling (UM97), Chia Laguna, Italy, June 1997, pp. 289-300. See also http://www.cs.uni-sb.de/UM97/ps/WeberG.ps.
[Weiss 2003] Weiss, A., 2003, *Ending Spam's Free Ride,* netWorker, Volume 7, Issue 2, June 2003, pp. 19-24.
[Weiss 2005] Weiss, A., 2005, *The Power of Collective Intelligence,* netWorker, Volume 9, Issue 3, September 2005, pp. 17-23.
[Wikimedia 2006a] Wikimedia Meta-Wiki, 2006, *List of largest wikis,* http://meta.wikimedia.org/w/index.php?title=List_of_largest_wikis&oldid=270365, Accessed January 13th, 2006.

[Wikimedia 2006b] Wikimedia Meta-Wiki, 2006, *Article validation feature,* http://meta.wikimedia.org/w/index.php?title=Article_validation_feature&oldid=254904, Accessed January 15th, 2006.
[Wikinews 2006a] Wikinews, http://www.wikinews.org/.
[Wikinews 2006b] Wikinews, 2006, *Wikinews:Mission statement,* http://en.wikinews.org/wiki/Wikinews, Accessed January 18th, 2006.
[Wikinews 2006c] Wikinews, 2006, *Statistics – Wikinews,* http://en.wikinews.org/wiki/Special:Statistics, Accessed October 29th, 2006.
[Wikipedia 2006a] Wikipedia, http://www.wikipedia.org/.
[Wikipedia 2006b] Wikipedia, *Wikipedia – The free encyclopaedia,* http://en.wikipedia.org/wiki/Wikipedia, Accessed October 29th, 2006.
[Wikipedia 2006c] Wikipedia, 2006, *Wikipedia:Neutral point of view,* http://en.wikipedia.org/w/index.php?title=Wikipedia:Neutral_point_of_view&oldid=34875574, Accessed January 13th, 2006.
[Wikipedia 2006d] Wikipedia, 2006, *Wikipedia:Version 1.0 Editorial Team/Wiki Sort,* http://en.wikipedia.org/w/index.php?title=Wikipedia:Version_1.0_Editorial_Team/Wiki_Sort&oldid=34857842, Accessed January 15th, 2006.
[Wikitravel 2006] Wikitravel, http://www.wikitravel.org/
[WikiWikiWeb 2006] Wiki Wiki Web, http://c2.com/cgi/wiki?WikiWikiWeb.
[Wilde and Lowe 2002] Wilde, E., and Lowe, D., 2002, *XML Linking Language.* In: Wilde, E., and Lowe, D., *XPath, XLink, XPointer, and XML: A Practical Guide to Web Hyperlinking and Transclusion,* Addison-Wesley Professional, pp. 169-198. See also http://searchwebservices.techtarget.com/searchWebServices/downloads/wilde-lowe_07.pdf.
[WordNet 2006] WordNet, http://wordnet.princeton.edu/.
[WPrandom 2006] Wikipedia, 2006, http://en.wikipedia.org/wiki/Special:Random, Accessed February 16th, 2006.
[WPspecial 2006] Wikipedia, 2006, *Special pages,* http://en.wikipedia.org/wiki/Wikipedia:Specialpages, Accessed February 16th, 2006.
[WPstats 2006] Wikipedia, 2006, *Statistics,* http://en.wikipedia.org/wiki/Special:Statistics, Accessed February 22nd, 2006.
[X3D 2005] Web3D Consortium, 2005, *Information technology – Computer graphics and image processing – Extensible 3D (X3D),* International Standard ISO/IEC FDIS 19775:200x, http://www.web3d.org/x3d/specifications/ISO-IEC-19775-IS-X3DAbstractSpecification/, Accessed June 12th, 2005.
[Xiao et al. 2004] Xiao, J., Stasko, J., and Catrambone, R., 2004, *An Empirical Study of the Effect of Agent Competence on User Performance and Perception,* Proceedings of the Third Internation Joint Conference on Autonomous Agents & Multi Agent Systems (AAMAS'04), New York, NY, U.S.A., July 2004, pp. 178-185. See also http://www.aamas2004.org/proceedings/024_xiaoj_competence.pdf.
[Xipolis 2005] Xipolis, http://www.xipolis.net/.
[XLink 2001] DeRose, S., Maler, E., and Orchard, D., 2001, *XML Linking Language (XLink) Version 1.0,* http://www.w3.org/TR/xlink/, Accessed April 28th, 2005.
[XML 2003] W3C, 2003, *Extensible Markup Language (XML),* http://www.w3.org/XML/, Accessed April 28th, 2005.
[XPointer 2002] DeRose, S., Daniel, R. Jr., Grosso, P., et al., 2002, *XML Pointer Language (XPointer),* http://www.w3.org/TR/xptr/, Accessed May 2nd, 2005.
[Yahoo 2006] Yahoo! Directory, http://search.yahoo.com/dir/.
[Yarowsky 1992] Yarkowsky, D., 1992, *Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora,* Proceedings ofthe 14th Conference on Computational Linguistics (COMLING-92), Nantes, France, August 1992, pp. 454-460.
[Yocom 2004] Yocom, M., 2004, *Mac OS History,* http://www.macos.utah.edu/Documentation/MacOSX-Classes/macosxone/macintosh.html, Accessed May 3rd, 2005.
[YouTube 2006] YouTube, http://www.youtube.com/.
[Zachte 2006] Zachte, E., 2006, *Wikipedia Statistics English,* http://stats.wikimedia.org/EN/TablesWikipediaEN.htm, Accessed February 22nd, 2006.
[Zakon 2005] Zakon, R., 2005, *Hobbes' Internet Timeline v8.1,* http://www.zakon.org/robert/internet/timeline/#Growth, Accessed September 19th, 2005.