Atif Latif

Discovery, Triplification and Consumption of Pertinent Resources from Linked Open Data

Dissertation
submitted to the
Graz University of Technology, Austria

for the partial fulfillment of the requirements for the academic degree of
Doctor of Philosophy
(Ph.D)

at the Institute for Knowledge Management
Graz University of Technology, Austria

Supervisor & Assessor 1: Univ.-Prof.Dr. Klaus TOCHTERMANN
Assessor 2: Univ.-Doz.Ing.Mag.rer.nat.Mag.phil.Dr.phil. Andreas HOLZINGER

Graz, April 2011

Atif Latif

Discovery, Triplification and Consumption of Pertinent Resources from Linked Open Data

Dissertation

vorgelegt an der

Technischen Universität Graz, Austria

zur Erlangung des akademischen Grades
Doktor der Technischen Wissenschaften
(Dr.techn.)

durchgeführt am Institut für Wissensmanagement
Technische Universität Graz, Austria

Betreuer & Begutachter 1: Univ.-Prof.Dr. Klaus TOCHTERMANN
Begutachter 2: Univ.-Doz.Ing.Mag.rer.nat.Mag.phil.Dr.phil. Andreas HOLZINGER

Graz, im April 2011

Abstract

Linked Data provides a framework for the generation, publishing and sharing of information by use of semantic technologies. The Linking Open Data initiative plays a vital role in the realization of the Semantic Web at a global scale by publishing and interlinking diverse data sources on the Web. The access to a huge amount of Linked Data presents exciting opportunities for the next generation of Web-based applications. However, there are still too few use cases which exploit Linked Data to its full potential. This lack is mainly due to several open issues regarding Linked Data consumption, Linked Data publishing, and Linked Data applications, both from developers' as well as from users' points of view.

This PhD thesis aims to address the core issues of Linked Data consumption (e.g. searching and querying, simplified user interfaces, identity management and disambiguation) and to facilitate the publishing of Linked Data. It presents conceptual models that can help with understanding the Linked Data ecosystem. Moreover, it shows an innovative approach that employs the intelligent use of semantic technologies to develop an easy-to-use application for ordinary web users. This application can automatically acquire, process and present accurate and relevant information by consuming Linked Data and hiding all complex semantic mechanisms and querying structures.

Overall, this thesis makes five contributions:

- First, it identifies the need of conceptual constructs to better understand Linked Data and proposes a Linked Data value chain. Further, it highlights the potential pitfalls; one of them being the lack of user-friendly interfaces for exploring Linked Data resources.

- Second, to gain deeper insights and to simplify the search mechanisms for Linked Data, it develops and implements a technique for finding relevant URIs from different data sets of Linked Data.

- Third, it provides a user-friendly interface for exploring Linked Data called CAF-SIAL (Concept Aggregation Framework for Structuring Information Aspects of Linked Open Data). This framework is able to structure and present information in a user-friendly environment and dispenses with the need for learning how to query and explore Linked Data constructs. The evaluation results show that this fully automated system helps the users in the exploration of Linked Data resources.

- Fourth, a legacy HTML dataset of a digital journal was RDFized and auto-

matically interlinked with external Linked Data resources. It was then made openly available to the Linked Data community.

- Fifth, to show the added value from the Linked Data cloud, CAF-SIAL was applied in two use cases: In a digital journal (the Journal of Universal Computer Science), authors were linked with their profiles crafted from Linked Data in CAF-SIAL to facilitate the users of the journal in finding additional information for collaboration. In a second use case, an Expertise Mining System visualized potential experts in a hyperbolic tree. These experts were then linked with their profiles in CAF-SIAL, which proved very useful for the administration of the journal for identifying and assigning reviewers in a peer review setting.

Deutsche Kurzfassung der Dissertation

Linked Data bietet einen Rahmen für die Erstellung, Veröffentlichung und Weitergabe von Informationen durch den Einsatz semantischer Technologien. Die Linking Open Data Initiative spielt durch die Veröffentlichung und Vernetzung verschiedenster Datenquellen im Internet eine wichtige Rolle bei der Realisierung des Semantic Web auf globaler Ebene. Der Zugriff auf eine riesige Menge an Linked Data eröffnet spannende Möglichkeiten für die nächste Generation Web-basierter Anwendungen. Es gibt jedoch noch immer viel zu wenige Anwendungsfälle, die das volle Potential von Linked Data nutzen. Dies liegt hauptsächlich an einigen offenen Fragestellungen im Zusammenhang mit der Verwertung und der Veröffentlichung von Linked Data sowie mit Applikationen, die auf Linked Data aufbauen, sowohl aus Sicht der Entwickler als auch der Anwender.

Diese Dissertation soll die Kernthemen der Verwendung von Linked Data adressieren (z. B. Suchen und Abfragen, vereinfachte Benutzeroberflächen, Identity Management und Disambiguierung) sowie die Veröffentlichung von Linked Data erleichtern. Es werden konzeptionelle Modelle präsentiert, die dabei helfen können, das Linked Data Ökosystem besser zu verstehen. Darüber hinaus wird ein innovativer Ansatz vorgestellt, um durch den intelligenten Einsatz semantischer Technologien eine einfach zu bedienende Anwendung für normale Web-Nutzer zu entwickeln. Diese Anwendung kann durch den Einsatz von Linked Data relevante Informationen automatisch erfassen, verarbeiten und präsentieren und dabei komplexe semantische Mechanismen und Abfragestrukturen verbergen.

Insgesamt leistet diese Arbeit Beiträge in fünf Bereichen:

- Erstens identifiziert diese Arbeit die Notwendigkeit von konzeptionellen Konstrukten, um Linked Data besser zu verstehen, und stellt die Linked Data Wertschöpfungskette vor. Ferner werden mögliche Problembereiche aufgezeigt; einer davon der Mangel an benutzerfreundlichen Schnittstellen für die Nutzung von Linked Data Ressourcen.

- Zweitens wird eine Technik zum Auffinden relevanter URIs aus unterschiedlichen Datensätzen von Linked Data entwickelt und implementiert, um tiefere Einblicke zu gewinnen und die Suchmechanismen für verknüpfte Daten zu vereinfachen.

- Drittens wird eine benutzerfreundliche Schnittstelle für die Erkundung von Linked Data namens CAF-SIAL (Concept Aggregation Framework for Structuring Information Aspects of Linked Open Data) vorgestellt. Dieses Framework ist in der Lage, Informationen in einer benutzerfreundlichen Umgebung

zu strukturieren und darzustellen. Die Anwender müssen dabei nicht mehr lernen, wie Linked Data Konstrukte abgefragt und durchforstet werden. Die Ergebnisse der Evaluierung zeigen, dass dieses vollautomatische System die Nutzer bei der Exploration von Linked Data Ressourcen unterstützt.

- Viertens wurde ein HTML-Archiv eines digitalen Journals RDFiziert und automatisch mit externen Linked Data Ressourcen verknüpft. Anschlieend wurde es der Linked Data Community öffentlich zur Verfügung gestellt.
- Fünftens wurde, um den Mehrwert der Linked Data Cloud zu demonstrieren, CAF-SIAL in zwei Anwendungsfällen eingesetzt: In einer digitalen Zeitschrift (Journal of Universal Computer Science) wurden Autoren mit ihren Profilen verknüpft, die von CAF-SIAL basierend auf Linked Data erstellt wurden, um die Nutzer der Zeitschrift dabei zu unterstützen, zusätzliche Informationen für die Zusammenarbeit zu finden. In einem zweiten Anwendungsfall wurden potentielle Experten von einem Expertise Mining System in einem hyperbolischen Baum visualisiert. Diese Experten wurden dann mit ihren Profilen in CAF-SIAL verknüpft, was sich als sehr nützlich für die Verwalter der Zeitschrift bei der Identifizierung und Zuordnung von Rezensenten in einer Peer-Review-Umgebung erwies.

To my loving parents and family. Without their knowledge, wisdom, guidance, prayers and patience, I would not have the goals I have to strive and be the best to reach my dreams!

Acknowledgments

First of all, I would like to thank the Almighty Allah (S.W.T.), for His divine guidance and providence. His support, blessings, goodness, and kindness were always with me. He blessed me with motivation, passion, and hard work. It was his blessings which made me able to plan, visualize, and execute my dreams into the reality. I would like to dedicate this achievement to Him and to my inspirational father Muhammad Latif and to my loving mother Nasreen Akhter, whose prayers, motivations and belief in me got me that far. They are always a source of inspiration and driving force behind my studies. I would also like to share my achievements and joy with my brother Shahzad Latif, my sisters and their families. I find no more words to thank them for their continuous support.

Every PhD student wishes to conduct research under a visionary and inspirational supervisor, as well as with a researcher who is intelligent and motivating. I am very lucky to find this ideal supervisor in form of Professor Dr. Klaus Tochtermann. I am really thankful to him for giving me the exceptional opportunity of doing a PhD with him. His constant encouragement, help, and invaluable supervision helped me to groom my educational, research and writing capabilities which further excelled me to broaden my vision and brought best out of me for this research work. I owe my deepest gratitude to Professor Dr. Andreas Holzinger for being part of my dissertation evaluation committee and accepting the role of second reader despite his busy schedule. The discussions/comments of Prof. Holzinger were very useful for my research and thesis.

My dearest thanks go to my friend Dr. Muhammad Tanvir Afzal, who indeed acted as a mentor to me. His kind guidance and motivations always bring me out of the difficult patches throughout my research studies. I have learnt a lot from his company and discussions, which eventually led us to produce many publications. The time we passed together while in discussion, sports, travelling and leisurely, with no doubts is a very precious one in my life. I am also very obliged to him for introducing me with the renowned Professor Dr. Hermann Maurer of Institute for Information Systems and Computer Media (IICM) and with Journal of Universal Computer Science (J.UCS) administration. I want to extend my gratitude to Prof. Maurer on giving me chance to use J.UCS dataset for my research. The assistance provided by Prof. Maurer and J.UCS administration specially Dana Kaiser, Maria Luise Lample and Gabriele Leitner who helped me alot to innovate with new ideas and extend my research in different ways.

Many thanks to my colleagues at the Institute for Knowledge Management and Know Center, whose help, support and guidance has a big share in the success of my PhD studies, particularly to Patrick Höfler for helping hands in administration matters, with fruitful discussions, in introduction of new technologies and specially proof reading my thesis. I would also like to thank Dr. Alexander Stocker for his support and fruitful discussions throughout. I want to say special ”Thank You!” to: Anita Griesser for taking care of all my travel, departmental and administrative matters. She was always very kind and responsive in my each queried matter.

I was lucky enough to find myself among very loving and caring family of Pakistani scholars in Graz. They always fill in the gap of a family and made me happy while I was sad. They continuously motivated me to achieve my goals. Thanks to all who arranged and accompanied me in doing academics and non-academic activities in form of sports, parties and travelling. Especially, my gratitude goes out to Syed Khuram Shahzad, Muhammad Umer Saleem, Muhammad Farhan Hyder, Dr. Javed Ferzund, Dr. Anwar us Saeed and Dr. Muhammad Irfan. I am unable to mention names of all scholars yet all of you are special to me.

At the end my heartiest tribute to my country Pakistan and Higher Education Commission (HEC), Government of Pakistan, for partially funding my research as without their assistance I would not have a chance to pursue my PhD and fulfill my dreams. In the same spirit my tribute goes to the Austrian Agency for International Cooperation in Education and Research (OeAD-GmbH), Know Center, Knowledge Management Institute and Technology University of Graz for the moral and financial support to attend the conferences around the world.

Atif Latif
Graz, Austria, April 2011

Contents

List of Tables

List of Figures

Listings

Part I

Introduction and Background

Chapter 1

Introduction

"The Web was designed as an information space, with the goal that it should be useful not only for human-human communication, but also that machines would be able to participate and help"[Berners-Lee 1998]

A visionary statement given by Sir Tim Berners-Lee, who predicted the World Wide Web to be a dynamic and intelligent platform that also provides machine-readable data in order to help make better decisions. In the early era of the Web, these objectives were not really emphasized, and the main focus was on the technical capabilities, e.g. communication protocols, network structures and presentation of data. In the last decade, a tremendous development has been seen on the World Wide Web stage. This brought many revolutionary tools and techniques to the Web, making it more flexible, interactive, dynamic and business-oriented. Those important steps changed the way of information publishing and sharing, resulting in a flood of information and subsequently leading to information overload. Many disciplines like information search and retrieval, knowledge management and knowledge visualization evolved to process this information and provide a better user experience. These disciplines highlighted the importance and need of structured information to cope with the ever-increasing information on the web. Nowadays, a lot of research is going on in the area of structured and machine-processable information, and the vision of Sir Tim Berners-Lee could soon become reality.

The evolution of the web can be divided into three phases:

In the beginning, the web served to link documents and provide interfaces to access the information residing in different repositories. At this stage the provided data was mainly commercial or corporate content. The main focus here was limited to presenting this data to the users. Many data islands were created in this phase.

In the next stage, the social collaborative web coined as Web 2.0 [O'Reilly 2005] or social web brought the power of creating content to users. This social evolution provided applications (like blogs and wikis) and easy-to-use interfaces to express the personal knowledge on the web. A big mass of useful information was put on the web, but most of it in the form of unstructured data and with marginal capabilities of data organization based on basic metadata (tags) or crowd intelligence. Abrupt emergence of Social Media applications was witnessed in this phase, for example Wikipedia[1], Facebook[2], Flickr[3], Twitter[4]. All of these were hubs of social content creation.

The third phase of World Wide Web is known as *Semantic Web* (a.k.a. *Web 3.0* interchangeably. It is envisioned to cater the inherited issues of the conventional and social Web i.e. integration, assimilation and organization of data aggravated due to large amount influx of redundant unstructured data. It is also pinned as an intelligent web to act like personal assistant and can make intelligent decision mainly due to utility of structured data with semantic technologies. The overall vision of this phase is aligned with the Tim Berners-Lee 1998 semantic web vision.

"to make the Web machine-readable, allowing computers to integrate information from disparate sources to achieve the goals of end users"[Berners-Lee 1998]

Linked Data is one of the major initiatives in this regard. Linked Data is about employing the *Resource Description Framework (RDF)* and the *Hypertext Transfer Protocol (HTTP)* to publish structured data on the Web [Bizer et al. 2007]. It offers best practices to interlink data between different data sources, giving way to a single connected and interlinked Giant Global Graph. The principles of Linked Data were first outlined as a broad guideline by Berners-Lee in 2006 [Berners-Lee 2006]. Emergence of a *Web of Data* comes from the Linking Open Data project, a grass-root community effort founded in February 2007 and supported by the *W3C Semantic Web Education and Outreach Working Group.* The underlying goal of this community was to motivate people to expose, share, and connect Open Data using Linked Data principles. The community was able to attract people's attention by having a large amount of open data published. From its commencement in 2006, the project's ongoing efforts resulted in bootstrapping the Web of Data, which today

[1] http://www.wikipedia.com
[2] http://www.facebook.com
[3] http://www.flickr.com
[4] http://www.twitter.com

comprises billions of RDF triples including millions of links between data sources. The published datasets include data about books, movies, music, radio and television programs, reviews, scientific publications, genes, proteins, medicine, and clinical trials, geographic locations, people, companies, statistical and census data, etc., ranging from governmental to social community data. There is a huge potential for mashup applications to generate added value by reusing this pool of published data.

1.1 Motivation

The huge access to Linked Data presents exciting opportunities for the next generation of Web-based applications [Hausenblas 2009] [Hausenblas 2009a][Bizer et al. 2009]: data from different providers can be aggregated, and fragmented information from multiple sources can be integrated to achieve a more comprehensive view. While a few applications such as the BBC Music website [Kobilarov et al. 2009] have used Linked Data to significant benefit, so far the emphasis has been to exploit the data of interest from the Web of Data to create private and disconnected datasets for specific applications. There is few of these type of case studies, and there is potential need for applications that consume linked data in order to exploit the web of Linked Data to its full potential.

There are still several open issues from developer's and user's points of view, which makes the triplification / RDFization and development of Linked-Data-based applications challenging. These issues include the lack of conceptual frameworks for better understanding of Linked Data [Latif et al. 2009], lack of approaches for seamless integration of Linked Data from multiple sources for dynamic, on-the-fly discovery of available data [Volz et al. 2009] [Bizer et al. 2009], no well-supported information quality assessment measures[Jaffri et al. 2008] [Heath 2008a] and easy-to-use end user interfaces [Heath 2008]. Linked Data still has a long way to go to live up to its potential. A lot of effort is still needed to provide user-friendly interfaces built around the Linked Data cloud for browsing, searching and interacting with Linked Data, algorithms for triplification and organization of semantic data, and to automatically push relevant information to the users. Semantically enriched applications should be usable for the so-called normal web users so that they can consume Linked Data without knowing about the underlying complexities, like they do on the WWW nowadays.

Motivated by researching these open issues, this thesis describes conceptual frameworks on top of Linked Data, presents triplification techniques and applications specifically targeting normal web users that help them understand and explore the potentials of Linked Data in a better way.

1.2 Scope, Approach and Dissertation Foundation

The scope of this work is to investigate open issues in Linked Data as discussed in section 1.1. This thesis presents a framework which makes the data transition process of Linked Data more transparent, innovates with algorithms to triplify legacy data, and consumes linked data to build applications with easy-to-use features for the normal web users by hiding all the complexities. The developed application is also linked with a digital journal, an expertise mining system, and a dataset from the Academia Europaea. For the experimentation, Journal of Universal Computer Science, WWW'06, Academia Europaea and DBpedia are used as data sources.
The framework is divided into three main parts.

- Consumption of Linked Data resources.
- RDFization of legacy data.
- Application areas of Linked Data.

1.2.1 Consumption of Linked Data Resources

The first module of this framework focuses on the open challenges in Linked Data consumption, the Linked Open Data movement, and the requirements of conceptual models in the realm of Linked Data. To facilitate the understanding of the entities and processes involved in the transformation of Raw Data into Linked Data, the conceptual model of the Linked Data Value Chain is introduced. During the research for this model, the lack of intuitive user interfaces for Linked Data applications came to light and was subsequently investigated. For acquiring information from Linked Data, an intelligent URI locator technique was developed; for structuring information, a Concept Aggregation Framework was developed. To demonstrate the working of our techniques, a proof-of-concept application known as CAF-SIAL (Concept Aggregation Framework for Structuring Informational Aspects of Linked Open Data) was implemented and has been running since July 2009 at http://cafsial.opendatahub.org.

1.2.2 RDFization of Legacy Data

The second module of the proposed framework is about the RDFization of datasets for making contribution to the Linked Data Cloud. To achieve this task, firstly related datasets to digital journal were mined for the development of new features and further *"semantified"* to link them with the Linked Data Cloud. To demonstrate the working of this module, have selected three data sets, Journal of Universal Computer

Science (J.UCS) with related features, WWW'06 conference data and DBpedia as a case study. These data sets were then interlinked with other Linked Data knowledge bases DBpedia, DBLP and with conventional web CiteULike and Faceted DBLP and provided openly for discovery and interlinking purposes. Additionally to facilitate the users, an HTML interface for navigation between interlinked datasets and, to query these datasets, a SPARQL endpoint was provided.

1.2.3 Application Areas

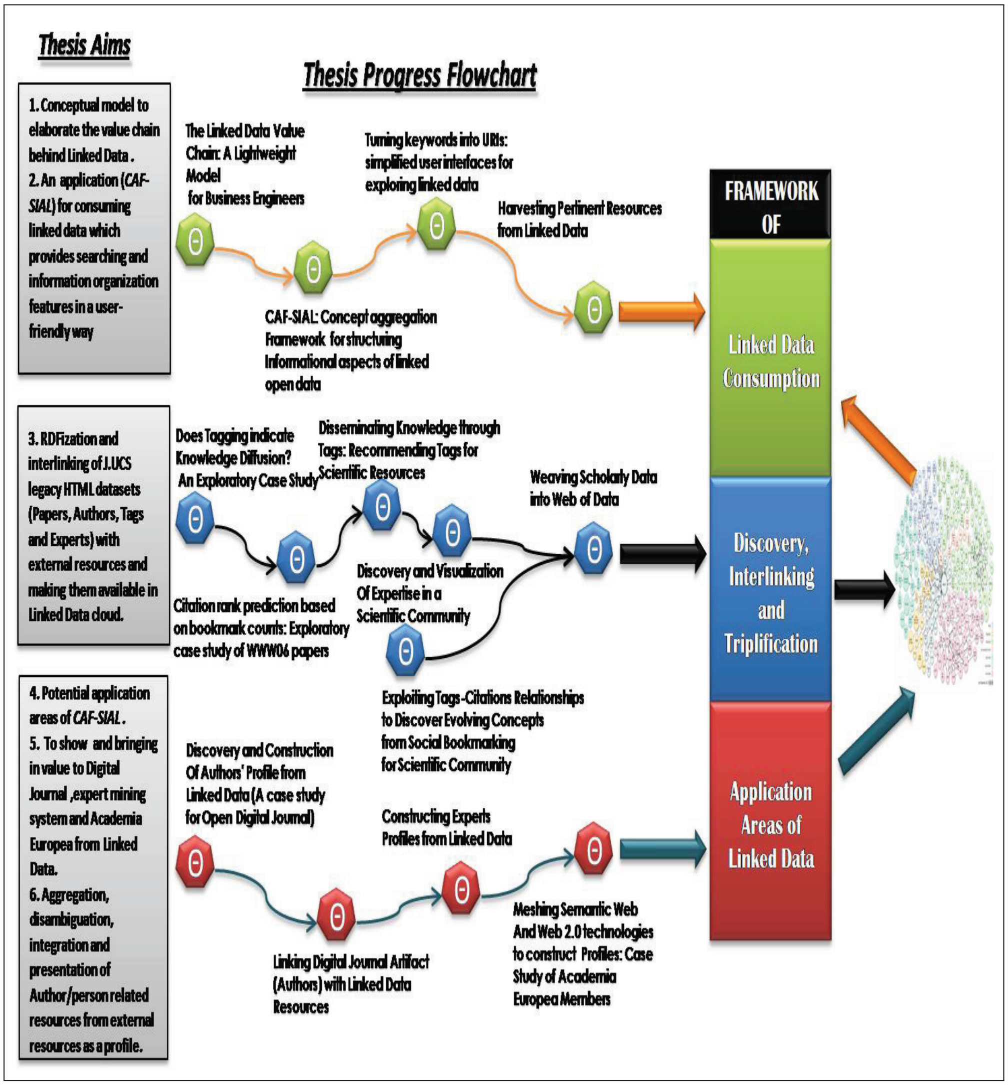
The last module of this framework was to study different application areas where Linked Data can generate value. In first module of the framework, an application was implemented which consumes Linked Data and uses the Concept Aggregation Framework for structuring the related aspects of a person into a profile. It was realized that application could be used as a third party tool to bring added value from the Linked Open Data sphere to different domains. It has been implemented in different areas such as:

- Digital Journal by providing journal users with the authors related information and their contribution in scientific community.
- On expert mining system, to assist journal administration to have a leverage of the additional information provided by the application to assign peer reviewing and editors to a particular paper.
- Academia Europea dataset, for constructing and presenting the profile of its members.

This dissertation is based on the various publications authored and co-authored in last three years. The orientation and relation of these publication according to set research objective is illustrated in figure 1.1. As discussed earlier, this dissertation makes contribution in 1) Linked Data Consumption, 2) RDFization of legacy data and 3) Application areas of Linked Data.

1.3 Research Questions

This section discusses the research questions investigated in this work. Wherever applicable, main research question are subdivided the into sub-research questions for simplification. In pursuit of these questions, two conceptual models and two proof-of-concept applications were developed.


Thesis Aims
Thesis Progress Flowchart
The Linked Data Value Chain: A Lightweight Model for Business Engineers
Turning keywords into URIs: simplified user interfaces for exploring linked data
Harvesting Pertinent Resources from Linked Data
CAF-SIAL: Concept aggregation Framework for structuring Informational aspects of linked open data
Does Tagging indicate Knowledge Diffusion? An Exploratory Case Study
Disseminating Knowledge through Tags: Recommending Tags for Scientific Resources
Weaving Scholarly Data into Web of Data
Citation rank prediction based on bookmark counts: Exploratory case study of WWW06 papers
Discovery and Visualization Of Expertise in a Scientific Community
Exploiting Tags-Citations Relationships to Discover Evolving Concepts from Social Bookmarking for Scientific Community
Discovery and Construction Of Authors' Profile from Linked Data (A case study for Open Digital Journal)
Constructing Experts Profiles from Linked Data
Linking Digital Journal Artifact (Authors) with Linked Data Resources
Meshing Semantic Web And Web 2.0 technologies to construct Profiles: Case Study of Academia Europea Members
FRAMEWORK OF
Linked Data Consumption
Discovery, Interlinking and Triplification
Application Areas of Linked Data

1. Conceptual model to elaborate the value chain behind Linked Data .
2. An application (*CAF-SIAL*) for consuming linked data which provides searching and information organization features in a user-friendly way

3. RDFization and interlinking of J.UCS legacy HTML datasets (Papers, Authors, Tags and Experts) with external resources and making them available in Linked Data cloud.

4. Potential application areas of *CAF-SIAL* .
5. To show and bringing in value to Digital Journal ,expert mining system and Academia Europea from Linked Data.
6. Aggregation, disambiguation, integration and presentation of Author/person related resources from external resources as a profile.

Figure 1.1: Dissertation Foundation

Following is the list of addressed research questions:
RQ1. How does the value chain behind Linked Data looks like?
RQ1.1. What entities and processes are involved in the transformation from raw data to Linked Data?
RQ1.2. What are the potential pitfalls that might occur?
RQ2. How should a system for consuming linked data look like which provides searching and information organization features in a user-friendly way? How can underlying semantic mechanics be concealed from end users?
RQ2.1. How can the URI of a resource be located from Linked Open Data?
RQ2.2. How can retrieved resources from Linked Data be structured and presented in more perceivable way?
RQ3. How can RDFization and interlinking of legacy HTML detasets (Authors, Papers, Experts and Tags) from Digital Journal, with external data sources be done. What are the benefits for making them available in Linked Data cloud?
RQ4. What are the potential application areas of CAF-SIAL Linked Data application?
RQ4.1 How digital journals can consume information from Linked Data resources and what values it can bring in?
RQ5. How aggregation, disambiguation, integration and presentation of Author/person related resources from external resources as a profile is performed and which external datasets are important?

Research question 1 and research question 2 are addressed in chapter 4, chapter 5 and chapter 6 respectively. Research questions 3 is investigated in chapter 7 and chapter 8, while research questions 4 and 5 are answered in chapter 9 and chapter 10.

1.4 Scientific Contributions

This Ph.D thesis aims at addressing core issues of the Linked Data (e.g. searching and querying, simplified user interfaces, identity management and disambiguation) identified by two conceptual models. More precisely, the objective of this thesis is to highlight how conceptual models can simplify the understanding of Linked Data and application development. How, innovative approaches mixed with intelligent use of semantic technologies employed to develop an easy to use application for normal web users is done, which can automatically acquires, process and present accurate and relevant information, by consuming Linked Data with hiding all complex mechanism of semantic and querying structures , as well as can bring benefits to other fields. The thesis first finds innovative ways for resource discovery from Linked Data and subsequently improves the state of the art for different publishing environments including digital journals. Overall, the thesis makes five contributions:

Firstly, it identifies the need of conceptual constructs for Linked Data understanding and proposes a Linked Data value chain. Further, it highlights potential pitfalls; one of them being the lack of user-friendly interfaces for exploring Linked Data resources. Secondly, to gain deeper insights and to simplify the search mechanisms for Linked Data, it develops and implements a technique for finding relevant URIs from different data sets of Linked Data. Thirdly, it provides a user-friendly interface for exploring Linked Data called CAF-SIAL (Concept Aggregation Frame- work for Structuring Information Aspects of Linked Open Data). This framework was able to structure and present information in a user-friendly environment and dispenses with the need for learning how to query and explore Linked Data constructs. The evaluation results show that this fully automated system worked satisfactorily. Furthermore, the system helped the users in exploring Linked Data resources. Fourthly, performed triplification, interlinking of mined and legacy HTML data set (Digital Journal, Recommended Tags and Identified experts) and made them available to the Linked data community for linking and discovery. Fifth, to judge and bring added value from Linked Data cloud, CAF-SIAL is applied in the area of Digital Journal (Journal of Universal Computer Science), Expertise Mining System and on Members of Academia Europea. In Digital Journal authors are linked with their profiles crafted from Linked Data in CAF-SIAL so to facilitate the users of the journal in finding additional information for collaboration etc. In Expertise Mining System, as second application area, the visualized potential experts in hyperbolic tree are linked with their profiles in CAF-SIAL. This expertise linking with CAF-SIAL proved very useful for the administration of journal in identifying and assigning reviewers in a peer review setting.

1.5 Published Work

Parts of the work presented in this thesis have been published in international journals and in the proceedings of international conferences and refereed workshops. The corresponding publications are listed below:

[Latif et al. 2009] Latif, A., Hoefler, P., Stocker, A., Ussaeed, A., Wagner, C. (2009). The Linked Data Value Chain: A Lightweight Model for Business Engineers. In: Proceedings of International Conference on Semantic Systems, pp. 568-576, Graz, Austria, 2-4, Sep. 2009.

[Latif et al. 2009a] Latif, A., Afzal, M.T., Ussaeed, A., Hoefler, P., Tochtermann, K. (2009). CAF-SIAL: Concept aggregation framework for structuring informational aspects of linked open data. In: Proceedings of International Conference on Networked Digital Technologies, pp. 100-105, Ostrava, Czech Republic, 28-31,

Jul. 2009.

[**Latif et al. 2009b**] Latif, A., Tanvir, M.T., Hoefler, P., UsSaeed, A., Tochtermann, K.(2009). Turning keywords into URIs: simplified user interfaces for exploring linked data. In: Proceedings of 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human, Seoul, Korea, 24-26 Nov. 2009.

[**Latif et al. 2010**] Latif, A., Afzal, M.T., Ussaeed, A., Hoefler, P., Tochtermann, K. (2009). Harvesting Pertinent Resources from Linked Data. In Journal of Digital Information Management (JDIM) 8 (3), pp. 205-212, June 2010.

[**Latif et al. 2010a**] Latif, A., Afzal, M.T., Helic, D., Tochtermann, K., Maurer, H. (2010). Discovery and Construction of Authors' Profile from Linked Data (A case study for Open Digital Journal), Proceedings of the WWW2010 Workshop Linked Data on the Web (LDOW 2010), CEUR Workshop Proceedings. CEUR-WS.org (2010)

[**Latif et al. 2010b**] Latif, A., Afzal, M.T., Tochtermann, K. (2010). Constructing Experts Profiles from Linked Data, In: Proceedings of 6th International Conference on Emerging Technologies (ICET), pp. 33-38 , Islamabad, Pakistan, 18-19, Oct. 2010.

[**Latif and Afzal 2011**] Latif, A., Afzal, M. T.(2011). Linking Digital Journal Artefact (Authors) with Linked Data Resources, accepted and to appear in Journal of Universal Computer Science, 2011.

[**Latif and Afzal 2011a**] Latif, A., Afzal, M. T.(2011). Weaving Scholarly Legacy Data into Web of Data. accepted and to appear in Journal of Universal Computer Science, 2011.

[**Korica-Pehserl and Latif 2011**] Korica-Pehserl, P., Latif, A.(2011). Meshing Semantic Web and Web 2.0 technologies to construct Profiles: Case Study of Academia Europea Members. Accepted and to appear in Third International Conference on Networked Digital Technologies (NDT 2011), Macau, China , 11-13 July 2011.

[**Afzal et al. 2009**] Afzal, M. T., Latif, A., Ussaeed, A., Sturm, P., Aslam, S., Andrews, K., Tochtermann, K., Maurer, H. (2009). Discovery and Visualization of Expertise in a Scientific Community. In: Proceeding of International Conference of Frontiers of Information Technology, Islamabad, Pakistan, 16-18, Dec. 2009.

[**Afzal and Latif 2011**] Afzal, M. T., Latif. (2011). Exploiting Tags-Citations Relationships to Discover Evolving Concepts from Social Bookmarking for Scientific Community. accepted and to appear in Journal of Universal Computer Science, 2011.

[**UsSaeed et al. 2008a**] Us Saeed, A., Afzal, A., Latif, A., Stocker, A., Tochtermann, K. (2008). Does Tagging indicate Knowledge Diffusion? An Exploratory Case Study. In: Proceedings of International Conference on Convergence and Hybrid Information Technology, pp. 605 - 610, Busan, Korea, Nov. 11-13, 2008.

[UsSaeed et al. 2008b] Us Saeed, A., Afzal, M. T., Latif, A., Tochtermann, K. (2008). Citation rank prediction based on bookmark counts: Exploratory case study of WWW06 papers. In: Proceedings of IEEE International Mutitopic Conference, pp. 392-397, Karachi, Pakistan, Dec. 23-24, 2008.

[UsSaeed et al. 2010] Us Saeed, A., Afzal, M. T., Latif, A., Tochtermann, K. (2010). Disseminating knowledge through Tags: Recommending Tags for scientific resources. In Journal of IT in Asia, Vol 3 (2010), pp. 25-36, Issue Date: Nov 2010, Print ISSN: 1823-5042.

1.6 Thesis Organization

This thesis is divided into five core parts. Part I explains the introduction and background of this conducted research with state of the art. Part II describes our proposed conceptual models i.e. Linked Data Value Chain and Concept Aggregation Framework, leads to the prototype Linked Data web application a.k.a. CAF-SIAL. Part III focuses on triplification of datasets. Part IV discusses the application areas of CAF-SIAL i.e. Expertise and authors Profile building in a Digital Journal environment. Part V describes the system evaluations, conclusions and future work. In the following, the thesis is outlined by summarising each chapter.

Part I: Introduction and Background

Chapter 1: Introduction. This chapter presents the motivation and scope of this research, gives an overview of the research questions. Furthermore describes the scientific contributions claimed by this research in the field of Linked Data and later on presents the structure of this thesis.

Chapter 2: Semantic Web. This chapter describes the history and basic concepts of Semantic Web. In addition provides insights of the core concepts and technologies used in Semantic Web Architecture.

Chapter 3: Linked Open Data Project and Related Work. This Chapter discusses the history of Linked Open Data project, building principles and the dataset distribution in the Linked Data cloud. Further on it presents demanding research challenges in the Linked Data community as a whole and more specifically state of the art systems and applications in the realm of Linked Data Consumption.

Part II: Linked Data Value Chain and Linked Data Consumption

Chapter 4: The Linked Data Value Chain This chapter outlines the need of conceptual models for the true realization of Linked Data value in Corporate and Business engineers. Moreover discusses the proposed Linked Data value chain, which simplifies and assists in identifying processes and participating entities in data transition as well as highlights the potential pitfalls which may appear in this process.

Chapter 5: Locating Intended URI's of Linked Data Resources. This chapter presents an intelligent technique for locating URIs from the huge repository of Linked Data. In further describes the Keyword-Uri Mapping Technique with detailed architectural design and concludes with the exploratory evaluations.

Chapter 6: Aggregation, Organization and Presentation of Information from Linked Data. This Chapter firstly, emphasizes on the need of simple user interfaces to explore linked data resources and discusses the proposed solution aka CAF-SIAL. Further it explains the Concept Aggregation Framework and discusses the implemented application architecture with case study.

Part III: Weaving Scholarly Legacy Data into Web of Data

Chapter 7: Preparing Data from Social Web and Digital Journal. This chapter first highlights the importance of converting legacy HTML data to Linked Data. In addition the related datasets of Journal (Tags and Expertise Calculation) and WWW'06 are explained with their mining processes. Furthermore it presents the semantic annotation strategy of papers with recommended tags on given author keywords. In the end of this chapter a case study of J.UCS and WWW'06 is discussed.

Chapter 8: RDFization and Interlinking of Legacy Data. This chapter presents approaches for modelling and RDFization of legacy data. Moreover discusses the interlinking of it with external datasets and describes the sample query example to highlight the conversion benefits of Linked Data.

Part IV: Application Areas

Chapter 9: Digital Journals – Discovering and Organizing Authors' Profiles from Linked Data. This chapter mentions Journal of Universal Computer Science (J.UCS), a Digital Open Journal as first application area of our proposed application CAF-SIAL. It highlights the added value which can be drawn from open datasets of Linked Data for the real world applications.It presents the extension of CAF-SIAL as a profiling system to construct the multi aspect authors profile as well as its linking with the J.UCS. In addition it gives insights about the methodology, architectural design, algorithm.

Chapter 10: Linking Experts Profiles in Linked Data. This chapter presents expertise mining system as a second application area of CAF-SIAL. A system visualize the current and potential experts in topical areas of a scientific discipline. Furthermore this chapter explains how locally mined and additional information of identified and visualized experts is pushed to the administration of a journal for selection and assignment of reviewing duties.

Part V: Results, Discussions and Future Work

Chapter 11: System Evaluations. This chapter outlines evaluation study of the CAF-SIAL web application to judge its effectiveness. Furthermore describes the design, methodology and procedures used in the evaluation study and discusses the results.

Chapter 12: Conclusion and Future Work. This chapter concludes the thesis with self assessment on the performed research and discusses the possible future directions and extensions of application.

Chapter 2

Semantic Web

This chapter focuses on the web technologies that appeared on the scene in the quest of the Tim Berners-Lee's vision on Semantic Web [Berners-Lee et al. 2001]. The novelty of this study is to use these technologies as a footstep to explore the development of new semantic solutions and applied them to materialize our proposed solution. The focal point of this chapter is to discuss the broad phenomenon of Semantic Web to understand the proposed framework in this thesis which is about Linked Data. It is important to discuss basis of Semantic Web, as these are the major constructs on which the whole Linked Data idea is based upon. In start of this chapter retrospect about Semantic Web vision is provided. In next sections of this chapter, the W3C Semantic Web architecture with respect to its layers and underlying technologies are discussed and explained with examples.

2.1 Introduction

The phenomenon of Semantic Web strongly emerged in the mid nineties when Vannevar Bush [Bush 1945] dreamt the idea of "Memex" in 1940s, an intelligent device which can help individuals to store their books, records and communications electronically and provide search facility in speedy and flexible manner. Over all "Memex" gives a sense of library with a searchable catalogue. The materialization of "Memex" idea took more than 20 years to see the real results in the form of Engelbart [Engelbart 1962] and Nelson [Nelson 1982] work.

The idea of Semantic Web again caught the attention for the second time Sir Tim Berners-Lee, Jim Hendler and Ora Lassila wrote an article about this concept in the Scientific American Journal in May 2001 [Berners-Lee et al. 2001]. In this piece of writing they emphasized on the powerful version of web, which can process the information in an intelligent way by understanding the users context and can perform the hard work by its own to facilitate them by doing daily routine works on web. For instance this daily routine work may vary from searching of a document to the scheduling of an appointment with doctor, literally working as a personalized agent. Simply, facilitating human's in big way by taking away all the laboriously probing efforts, which they need for getting through information on the web. [Berners-Lee et al. 2001] iterated on manifestation of this idea is possible by providing sufficient description of (resources, links between them) and the tools which can be used by machines to make the right decision in the right directions at the right time. As with the words of this article:

"The Semantic Web will bring structure to the meaningful content of Web pages, creating an environment where software agents roaming from page to page can readily carry out sophisticated tasks for users."

Essence of this above statement is to arm current web with well structures semantics which encapsulate resource description, its context and relationship with other resources. The adaptation of this new approach can make machines much more powerful to act as software agent. In this article Semantic web is defined as:

"...an extension of the current Web in which information is given well-defined meaning, better enabling computers and people to work in cooperation."

With the emergence of Semantic Web Roadmap [Berners-Lee 1998], a major effort went under way to re-engineer the Web according to the Vision of Semantic Web as a sole program. The Semantic Web is an initiative of the World Wide Web Consortium (W3C)[1], an international governing body who sets the standards and rules for the new and old technologies powering the World Wide Web. The Semantic Web initiative was started as a Web metadata Working Group in 1998 and later merged in to Semantic Web Activity group[2]. This group professed Semantic Web as:

"The Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries. It is a collaborative effort led by W3C with participation from a large number of researchers and industrial partners. It is based on the Resource Description Framework (RDF)."

Further, the activity of this group was focused at establishment of:

[1] http://www.w3.org

[2] http://www.w3.org/2001/sw/

- Common syntaxes for data representation,
- Common vocabularies for data description,
- Common languages for data interchange and exchange.

In the pursuit of these objectives by W3C, range of working groups in last decade emerged as a major contributor to the Semantic Web cause. Some of the important groups are:

- ***RDF Core Working Group***[3]***,*** produced a Resource Description Framework model, Syntax Specifications and resource description framework schema specification. It was also responsible for the updates and revision in proposed models.
- ***Web Ontology Working Group***[4]***,*** extended the RDF Core work and introduced a language for defining structured web based ontologies which will provide richer integration and interoperability of data among descriptive communities.
- ***Semantic Web Deployment Working Group***[5]***,*** to provide guidance in the form of W3C Technical Reports on issues of practical RDF development and deployment practices in the areas of publishing vocabularies, OWL usage, and integrating RDF with HTML documents.
- ***Rules Interchange Format Working Group***[6]***,*** to produce a core rule language plus extensions which together allow rules to be translated between rule languages and thus transferred between rule systems.
- ***SPARQL Working Group***[7]***,*** developed the SPARQL Query Language recommendation, also responsible for evaluation and updates in the specification of query language, as per recommended by users and implementers.

2.2 Semantic Web Stack

The architecture of Semantic Web also known as *W3C Semantic Web Stack* is heavily built around recommended components and technologies of said W3C activity

[3] http://www.w3.org/2001/sw/RDFCore/

[4] http://www.w3.org/2001/sw/WebOnt/

[5] http://www.w3.org/2006/07/SWD/

[6] http://www.w3.org/2005/rules/wiki/RIF$_{W}orking_{G}roup$

[7] http://www.w3.org/2009/sparql/wiki/Main$_{P}age$

groups. Semantic Web architecture has been developed in a layered fashion, as initially proposed by Sir Tim Berners-Lee [Berners-Lee 2000]. With many revisions from its inception, it is represented as shown in figure 2.1[8].

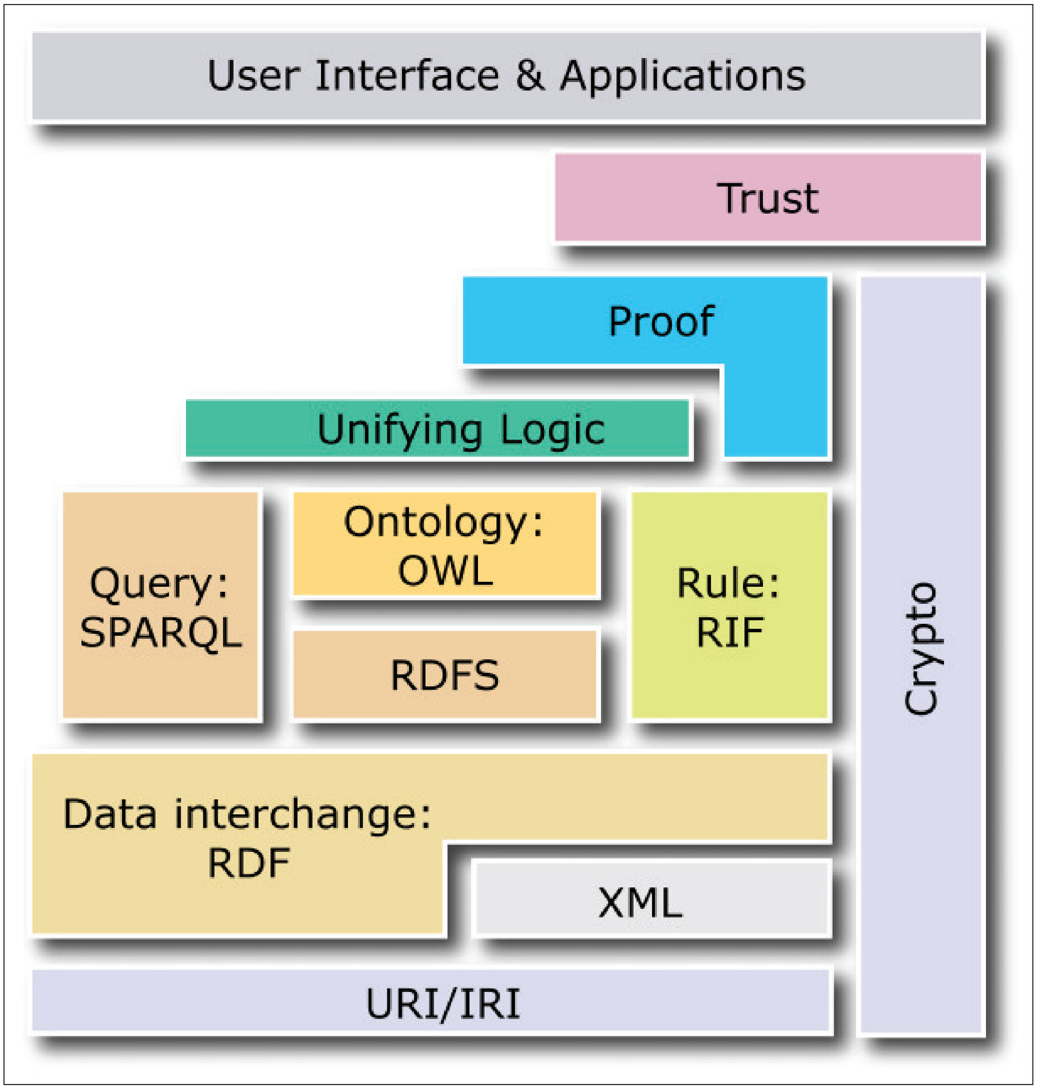


Figure 2.1: The W3C Semantic Web Stack

For simplification and better understanding, it is important to discuss the components of Semantic Web architecture in detail, as presented in coming sections of this chapter. The bottom two layers of architecture are well known for the Hypertext

[8] From W3C Semantic Web Activity: http://www.w3.org/2001/sw/layerCake.png

Web and provide the basis for the Semantic Web. Following sections have the details of the architectural layers and components:

2.2.1 Unicode and URI/IRI

Computers can only understand and process binary sequence like *011001* and other way around, we human only understand natural languages symbols like *S*, *8* numbers. To bridge a communication gap, Unicode encoding[9] effort set the standard for computer character representation by mapping symbols spoken in human languages to computer language. Uniform Resource Identifier(URIs) [Berners-Lee et al. 1998] provides a baseline to name, identify and to locate the unique resources e.g. pages on the Web. This process is also called as an addressing[10] and was a key factor in the success of Web. Internationalized Resource Identifier (IRIs)[Duerst and Suignard 2005] extends the URI pool of characters by adding up the Universal Character Set[11] to support Chinese, Korean and Arabic etc. languages symbols for naming resources. The first layer of Semantic Web stack is also known as Encoding and Addressing layer.

2.2.2 Data Exchange Layer: XML

After brief description of layer encoding and addressing from Semantic Web Stack, let's move to the next layer known as Data Exchange Layer. This layer consists of a simple tree structure models. It defines data structure standard and methods for interoperable exchange of the structured data. Proposed model for standard exchange of data on the Semantic Web Stack is the XML (Extensible Markup Language)[12]. XML is a common language used for structuring data on the Web. It was designed to improve the functionality of the Web by providing more flexible and adaptable information identification and exchanges. It allows embedding of tags constructs to annotate additional information within the text of a document. XML is a meta-language, gives the ability to design own customized mark-up languages for different types of documents. Further, it gives control to validate a new document according to the structural requirement defined in customization phase. For simplification an XML example document which describe title of a movie, its artist, release-data and director name, is shown in listing 2.1.

[9] http://www.unicode.org/

[10] http://www.w3.org/Addressing/

[11] http://anubis.dkuug.dk/JTC1/SC2/WG2/

[12] http://www.w3.org/XML/

```
<?xml version=”1.0” encoding=”UTF-8”?>
<movie id=”12”>
<title>The Terminator</title>
<release-date>26.10.1984</release-date>
<director>James Cameron</director>
<actors>
<actor>Arnold Schwarzenegger</actor>
<actor>Linda Hamilton</actor>
<actor>Michael Biehn</actor>
<actors>
</movie>
```

Listing 2.1: XML document example

The flexibility of XML is to represent any arbitrary information, however it is problematic for machine processing, mainly due to its inability in tags usage as explicit meaning. Many XML processing applications usually process well define set of tags which are pre-defined in natural languages. The lack of semantics is a limitation of the XML, making it difficult to integrate information coming from different XML documents. One way of tackling this limitation is to attach machine processable meaning with tags which are using the knowledge representation techniques like Resource Description Framework (RDF) and Ontologies.

2.2.3 RDF:

Semantic Web is typically about the interconnection of data rather than interconnection of documents as norm in Conventional Web. Resource Description Framework model is a step forward in this context. Resource Description framework (RDF), is also known as first layer of Semantic Web Core, provides a framework to identify resources by URIs and provides a graph base model to describe the relationship between resources. In this layer we will explain how formal description of data is given and even more important, how interlinking is done between these resources by help of examples.

In Resource Description framework model, statement is a basic unit for describing a resource and is formally known as triple. A triple, as name suggest consists of three components i.e. 1) Subject about whom we are putting a statement 2). predicate defines the typed relationship of the subject with object and 3). object provides cross-ponding value of subject in the triple. To illustrate the triple concept, let’s assume a statement:

”Arnold Schwarzenegger was born in Austria”

Simplified triple notation of this statement is. Subject ('Arnold Schwarzenegger'), Predicate ('was born') and Object ('Austria') and presented in figure 2.2.

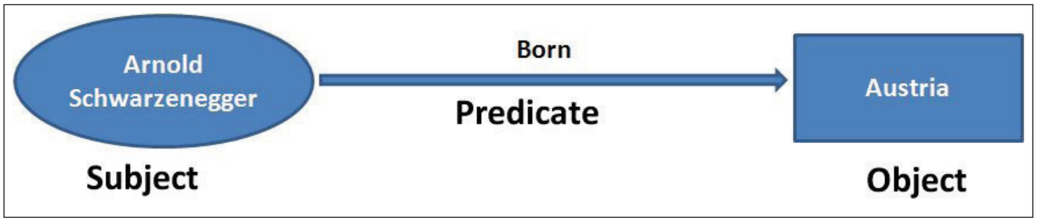


Figure 2.2: Triple (Statement) Visual Illustration

RDF has a simple data model that is easy for applications to process and manipulate [Klyne et al. 2004]. Before going into details of the RDF graph, we explain some important concepts related to RDF graph for better understanding.

URI Reference(URIref:) In RDF graph URIref is a unicode string that contains the valid character sequence to produce a valid URI. URIref is used to identify objects at metadata statements level [Hausenblas 2008]. For example, we rewrite our assumed example 'Arnold Schwarzenegger was born in Austria' by using URIrefs. Here we use the URIref `http://dbpedia.org/resource/Arnold_Schwarzenegger` instead of string *Arnold Schwarzenegger* to uniquely state about this object.

Blank node (bNode:) In RDF graph bNode is a type of node which cannot be identified by a URIref or literal. bNode is often used as anonymous identity where no global identity of a resource is required.

Literal: Literals in RDF graph are used to identify dates, values etc. as a lexical representation [Manola and Miller 2004]. Literals can be plain or typed. Anything represented by literals in RDF graph can also be identified by URIref.

With the elaboration, it is assumed that constructs of a triple can represented with these patterns:

- Subject, with an URIref or a bNode
- Predicate,with an URIref, showing typed relationship
- Object, with an URIref or a bNode or literal

We have applied said concepts on our example. Final statement is shown in figure 2.3.

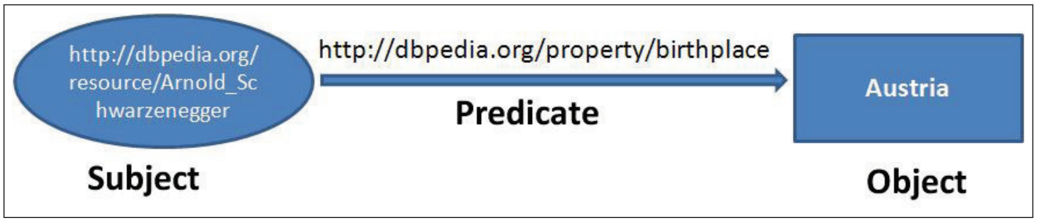

http://dbpedia.org/resource/Arnold_Schwarzenegger
http://dbpedia.org/property/birthplace
Predicate
Austria
Subject
Object

Figure 2.3: Triple Illustration

Herein, additional set of triples with our example i.e. Arnold Schwarzenegger, statement is attached as shown in figure 2.4. In figure, these added triples are further interpreted as a graph for every resource while arcs (predicate) cross-ponds to the relationships between them. By convention a directed link from subject to object is stressed in RDF data model .

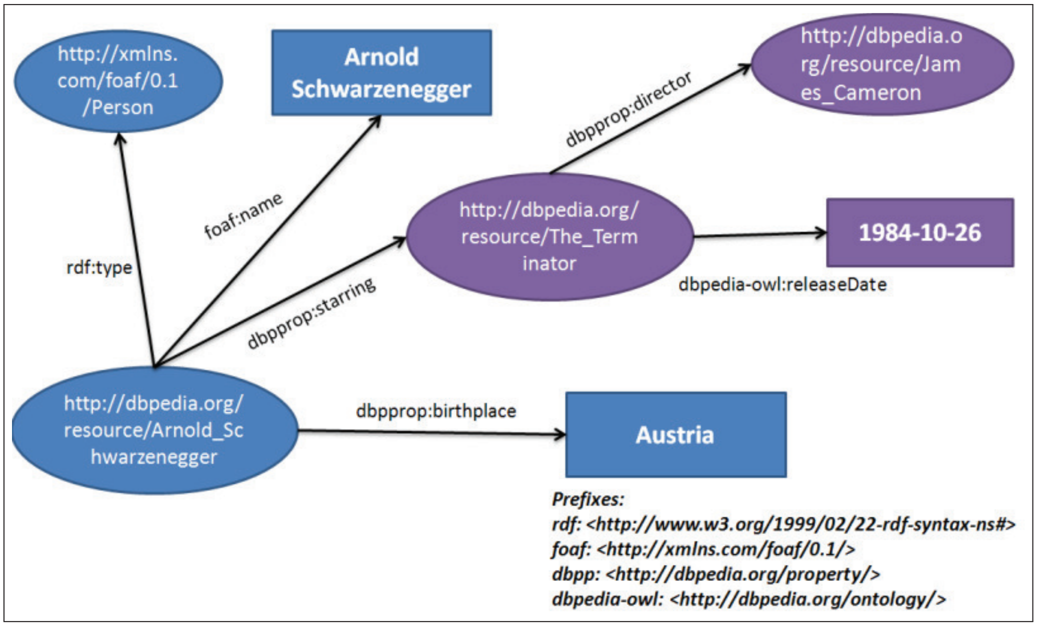

http://xmlns.com/foaf/0.1/Person
Arnold Schwarzenegger
http://dbpedia.org/resource/James_Cameron
dbpprop:director
foaf:name
http://dbpedia.org/resource/The_Terminator
1984-10-26
rdf:type
dbpedia-owl:releaseDate
dbpprop:starring
http://dbpedia.org/resource/Arnold_Schwarzenegger
dbpprop:birthplace
Austria
Prefixes:
rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
foaf: <http://xmlns.com/foaf/0.1/>
dbpp: <http://dbpedia.org/property/>
dbpedia-owl: <http://dbpedia.org/ontology/>

Figure 2.4: A RDF graph representing that Arnold is of rdf:type person, has a name, birthplace and is associated with film leading to additional information (From DBpedia Knowledge Base)

2.2.4 RDF Syntaxes

Several syntaxes are used to represent RDF. Here we elaborate three commonly used syntaxes.

RDF/XML

RDF/XML[13] is a W3C standard syntax to encode data into RDF representation.It is built upon the complaint RDF data model and is being used globally to represent the RDF data. Due to its complex syntax structures issues of understandability and readability are usually attached with its usage. For simplification, RDF/XML syntax is applied on Arnold Schwarzenegger example, as shown in listing 2.2.

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
      xmlns:dbpprop="http://dbpedia.org/property/"
      xmlns:dbpedia-owl="http://dbpedia.org/ontology/"
      xmlns:foaf="http://xmlns.com/foaf/0.1/">

<rdf:Description rdf:about="http://dbpedia.org/resource/Arnold_Schwarzenegger"><rdf:type rdf:resource="http://xmlns.com/foaf/0.1/Person"/></rdf:Description>
<rdf:Description rdf:about="http://dbpedia.org/resource/Arnold_Schwarzenegger"><foaf:name xml:lang="en">Arnold Schwarzenegger</foaf:name></rdf:Description>
<rdf:Description rdf:about="http://dbpedia.org/resource/Arnold_Schwarzenegger"><dbpprop:birthPlace xml:lang="en">Austria</dbpprop:birthPlace></rdf:Description>
<rdf:Description rdf:about="http://dbpedia.org/resource/Arnold_Schwarzenegger"><dbpedia-owl:starring rdf:resource="http://dbpedia.org/resource/The_Terminator"/></rdf:Description>
<rdf:Description rdf:about="http://dbpedia.org/resource/The_Terminator"><dbpedia-owl:releaseDate rdf:datatype="http://www.w3.org/2001/XMLSchema#date">1984-10-26</dbpedia-owl:releaseDate></rdf:Description>
<rdf:Description rdf:about="http://dbpedia.org/resource/The_Terminator"><dbpprop:director rdf:resource="http://dbpedia.org/resource/James_Cameron"/></rdf:Description>
</rdf:RDF>
```

Listing 2.2: RDF/XML

In an RDF/XML document there are two types of XML nodes: 1) resource XML nodes and 2) property XML nodes. Resource XML nodes are the subjects and objects of statements. These nodes are usually represented with rdf:Description tags

[13] http://www.w3.org/TR/rdf-syntax-grammar/

which further contain rdf:about attribute in them giving the URI of the resource they represent. In this example listing, the rdf:Description nodes are the resource nodes. Resource XML nodes contain only property XML nodes within them. Each property XML node represents a single statement which can either be a literal value or a URI. The subject of the statement is the outer resource XML node that contains the property. There are six statements in this example, the first four with the subject `http://dbpedia.org/resource/Arnold_Schwarzenegger` and the remaining with the subject `http://dbpedia.org/resource/The_Terminator`. The abbreviated URIs of the predicates in the six statements are rdf:type, foaf:name, dbpprop:birthPlace, dbpedia-owl:starring, dbpedia-owl:releaseDate and dbpprop:director.

N3

N3[14](Notation 3) is another syntax for RDF data representation. It was developed by Tim Berners-Lee and Dan Connolly [Beckett and Berners-Lee 2008] specifically keeping human readability in mind. N3 provides more compact, readable, extensive and expressive syntax to represent RDF data as compared with RDF/XML syntax. Terse RDF Triple Language (Turtle)[15] is a subset of N3 format and an official W3C standard. It covers only RDF data model and cannot go beyond that. It is commonly used and deployed by Semantic web developers for the discussions as it captures the abstract graph clearly. All the RDF statements written in Turtle Syntax are compatible with N3 syntax and can be used inside RDF query language(SPARQL).

```
@prefix xsd:  <http://www.w3.org/2001/XMLSchema#> .
@prefix rdf:  <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix dbpprop:  <http://dbpedia.org/property/> .
@prefix dbpedia:  <http://dbpedia.org/resource/> .
@prefix dbpedia-owl:  <http://dbpedia.org/ontology/> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .

dbpedia:Arnold_Schwarzenegger rdf:type  foaf:Person ;
foaf:name "Arnold Schwarzenegger"@en ;
dbpprop:birthPlace  "Thal, Austria"@en .
dbpedia:The_Terminator  dbpedia-owl:starring  dbpedia:
     Arnold_Schwarzenegger ;
dbpedia-owl:releaseDate "1984-10-26"^^xsd:date ;
dbpprop:director  dbpedia:James_Cameron ;
```

Listing 2.3: N3 Notation

[14] http://www.w3.org/DesignIssues/Notation3

[15] http://www.w3.org/TeamSubmission/turtle/

In N3/Turtle name-spaces are declared at the top with *@prefix* directive. RDF statements are just written out as the subject URI (in brackets or abbreviated with name-spaces) along with the predicate URI and followed by the object URI with period in the end.In listing 2.3, N3 syntax is applied on Arnold Schwarzenegger example, as shown above.

N3 also provides syntactical sugar options which supports in syntax abbreviations. For example in listing 2.3, a *semicolon(;)* is used five time to abbreviate repetition of subject URI. N3 also supports repetition of predicate URI abbreviation with *period(.)* sign and rdf:type predicate with *a* character.

Microformats:

Another syntax to publish RDF data on the web is known as Microformat[16] or RDFa [17]. Microformat usually comprises of a structure which is built around open standards i.e. XML, *XHTML*[18] and HTML tags. Some of the famous microformats are *hCard*[19], *hCalender*[20] and *hAudio*[21]. Microformats are embedded in a XHTML web document to semantically annotate the data which results in a rich metadata generation, useful for both machines and humans. Microformats cover limited domains and don't support RDF model completely. For complete RDF annotation support in an XHTML document, W3C built *RDFa (Resource Description Framework - in - attributes)* standard. RDFa and microformats can be used together in a single XHTML document, but for extensibility that is in more support for vocabularies, RDFa is preferred. For extraction of RDF statements from microformats and RDFa snippets *Gleaning Resource Descriptions from Dialects of Languages (GRDDL)*[22] mechanism is used. In following microformat snippet listing 2.4, Arnold Schwarzenegger is identified by foaf:person property and further described in starring relation with the Terminator *movie by* rel property.

[16] http://microformats.org/

[17] http://rdfa.info/

[18] http://www.w3.org/TR/xhtml11

[19] http://microformats.org/wiki/hcard

[20] http://microformats.org/wiki/hcalendar

[21] http://microformats.org/wiki/haudio

[22] http://www.w3.org/TR/grddl/

For illustration RDFa syntax is applied on Arnold Schwarzenegger example as shown below in Listing 2.4.

```
<div xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
     xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
     xmlns="http://www.w3.org/1999/xhtml"
     xmlns:dbpprop="http://dbpedia.org/property/"
     xmlns:foaf="http://xmlns.com/foaf/0.1/"
     xmlns:dbpedia-owl="http://dbpedia.org/ontology/"
     xmlns:xsd="http://www.w3.org/2001/XMLSchema#">
  <div class="description" about="http://dbpedia.org/resource/Arnold_Schwarzenegger"
       typeof="foaf:Person">
    <div property="dbpprop:birthPlace" content="Austria" xml:lang="en"/>
    <div property="foaf:name" content="Arnold Schwarzenegger" xml:lang="en"/>
    <div rel="dbpedia-owl:starring">
      <div class="description" about="http://dbpedia.org/resource/The_Terminator">
        <div rel="dbpprop:director" resource="http://dbpedia.org/resource/James_Cameron"/>
        <div xmlns:d1e6="http://www.w3.org/2001/XMLSchema#" property="dbpedia-owl:releaseDate"
             content="1984-10-26"
             datatype="xsd:date"/>
      </div>
    </div>
  </div>
</div>
```

Listing 2.4: Microformat

2.2.5 Ontology, Rule and Query Layer

This layer is known as core of Semantic Web Cake and provides mechanism for knowledge representations. RDF itself serves as description of a graph made up of triples. To add more description to the data in RDF ontologies/vocabularies are used. Ontologies are meant to define classes, their attributes (properties) and their relationship backed by enough description with each other. This help in reusing and sharing of common understanding of a particular domain.
The general definition of ontology was presented by Tom Gruber [Gruber 1992] as:

”ontology to mean a specification of a conceptualization. That is, an ontology is a

description (like a formal specification of a program) of the concepts and relationships that can exist for an agent or a community of agents...."

In [Fensel 2000], Fensel discusses the importances of ontologies in creating Knowledge Bases and defined ontologies as:

"ontologies provide a shared and common understanding of a domain that can be communicated between people and application systems"

For standardized formal semantic definitions and descriptions of ontologies various knowledge representation languages are defined which follows as:

Resource Description Framework Schema (RDF-S): RDF-S[23] is an extension of RDF vocabulary. It provides a formal framework for describing the taxonomy of classes, attributes and their relationships. For example it sets the domain and range properties of the RDF classes and relate these classes and properties by using RDF-S vocabulary. RDF-S works very similar to *Object Oriented Language* e.g. creation of classes, subclasses and properties hierarchies. RDF-S is often used to create Lightweight Ontologies where inferences and rules are not supported.

OWL (Web Ontology Language): The Web Ontology Language (OWL)[24] is a W3C recommended standard to create ontologies. It is extended from RDF and RDF-S vocabulary. It provides more freedom and expressiveness about defining ontologies in a detailed way and provides greater machine interpret-ability of web content as compared with RDF and RDF-S. It is generally designed for applications/machines to digest content but not for humans. As derived from *Description Logic (DL)*, it also supports mapping on various reasoner to assist in:

- Checking consistency of knowledge and ontology
- Discovering hidden relationship between classes

There are three species of OWL which are ranked according to their expressiveness and reasoning support. *OWL Lite* for taxonomies and simple constrains, *OWL DL* for full description logic support, and *OWL Full* for maximum expressiveness and syntactic freedom of RDF. OWL is designed for the downward compatibility.

Rule Languages: RDF and OWL define formal semantics and these formal semantic can be used to define reasoning rules on top of ontologies and knowledge

[23] http://www.w3.org/TR/rdf-schema

[24] http://www.w3.org/TR/owl-guide

bases for inference purpose. To bring more variety and assistance in rule creation different rules languages are standardized. Two most popular W3C recommended rule languages are *RIF Basic Logic Dialect (RIF)*[25] and *Semantic Web Rule Language (SWRL)*[26].

Query: The knowledge represented in knowledge bases will not be of any use, if it is not made available as a network for querying. One of the powerful aspect provided by Semantic Web initiative is querying. A *Simple Protocol and RDF Query Language (SPARQL)*[27] is available for querying RDF data as well as RDFS and OWL ontologies with knowledge bases. SPARQL is SQL like language, but uses RDF triples and resources for both matching part of the query and for returning results of the query. SPARQL is not only query language, it is also a protocol for accessing RDF data.

2.2.6 Cryptography, Proof, Trust and User Interface layers

It is assumed that all logics and rules defined in lower layers will be performed correctly to produce results under proof layer and further on results will be passed to data proof layer; *cf.* [Obitko 2007] .The proof layer of Semantic Web Stack is responsible for providing the provenance information to the users for example: from which source data was taken, which procedures and algorithms were used to transform the data and in the end what personalization measures (any human involvement) were used for the construction of results; *cf.* [Sizov 2007]. Data Cryptography layer which runs all the way from base layer is responsible to check the validity of input data. It uses digital signatures or other verification measures to make sure that data comes from valid source. Trust can be ensured to the users on the basis of valid authentication signs (valid input and provenance) from these layers. In the end data is passed to User Interface layer for human consumption and interaction. For detail insights about layer architecture and layering issues these studies are recommended; *cf.* [Hausenblaus 2010] [Antoniou and van Harmelen 2004] [Obitko 2007].
Today, all of the semantic tool and technologies are built by taking consideration of Semantic Web architecture. One of the most popular project which is built around semantic technologies is Linked Open Data. Linked Open Data project provides a platform for structuring of data in massed into machine readable formats; *cf.* Chapter 3.

[25] http://www.w3.org/TR/rif-bld/
[26] http://www.w3.org/Submission/SWRL/
[27] http://www.w3.org/TR/rdf-sparql-query/

2.3 Summary

Semantic Web idea is discussed in details in this chapter. In the start of this chapter, a brief history of Semantic Web, from its inception to present state in realm of various developments is provided . Moreover in next parts of this chapter, Semantic Web architecture with all the layers and technologies are discussed and explained with the help of examples.

Chapter 3

Linked Open Data Project and Related Work

3.1 Introduction

Linked Data is considered as the descendent (sub topic) of the Semantic Web project. The main core of Linked Data principles are founded on the set of technologies offered by the Semantic Web that can be interlinked in a useful way. This chapter presents a retrospect on the Linked Open Data project with the detail description of its design issues and the phenomenon of linked Data cloud growth with the distribution of datasets. In the ending final part of this chapter, the state-of-the-art about Linked Data is narrated. These sections are arranged according to claim contribution of this thesis i.e. Linked Data Publishing, Linked Consumption and Linked Data application areas.

3.2 Open Data and Open Access Movement

The concept of *Open Access* is not new, it has been used quite frequently in the past to highlight the importance of *Open Data*. Open Data concept provides data to the users without any restriction which they can use, modify and create new data from it. There are various thoughts and concerns raised in the scientific community about how much open data is crucial in the progress of science and in the generation of new knowledge. As beautifully illustrated by Executive Director of Science Commons *Mr. John Wilbanks*, why open data is needed.

"Numerous scientists have pointed out the irony that right at the historical moment when we have the technologies to permit worldwide availability and distributed

process of scientific data, broadening collaboration and accelerating the pace and depth of discovery ...we are busy locking up that data and preventing the use of correspondingly advanced technologies on knowledge"

Since last decade, an advancement in digital Content creation, realization of Open Access among people and increased dependability upon digital content for research work lead to number of data movements, in which Open Access Movement(OA) [Knezo 2006] is one of the significant one. Open Access movement is a social movement which is dedicated to the academia for open access of scientific research materials. Along with these movements, effort for the copyright legalization, policies on usage, modification and re-creation of content has come up with many standards i.e. *Creative Commons*[1], *Science Commons*[2] and *Open Data Common*[3]. In the recent times, most of the projects for example: *Open Access digital journals*, *Digital Libraries*, *Linked Open Data*[4], *Wikipedia*[5] and *Freebase*[6] emerged which on the footprints of open access provides open access and data to its users and made a significant mark. The massive impact of this Open Access phenomenon is enormous in the fast and dynamic development of sciences and other disciplines.

3.3 Linking Open Data Project

The *W3C Semantic Web Education and Outreach*, Linking Open Data community project [SWEO Project 2007] is considered to be the biggest follower of Open Access , it extends the Web with Open Data Commons by publishing many interesting datasets in RDF. It also offers possibilities of interlinking between related data entities via typed links. Linked Data is about employing publishing and interlinking practices based on set of principles, given by TIM Berners-Lee [Berners-Lee 2006]. These Linked Data principles also known as Design Issues guides developer in opening and interlink their dataset.
Linked Data is based on four simple rules:

1. Use URIs as names for things
2. Use HTTP URIs so that people (and machines) can look up those names (see also [Sauermann et al. 2008])

[1] http://creativecommons.org/
[2] https://creativecommons.org/science
[3] http://www.opendatacommons.org/
[4] http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData
[5] http://www.wikipedia.org/
[6] http://www.freebase.com/

3. When someone looks up a URI, provide useful information (RDF Model and Vocabularies)

4. Include links to other URIs so that they can discover more things

In nutshell, Linked Data is not a new Web, it is an extension of Conventional Web also known as Web of Document in the Linked Data community. Before going in details of Linked Data principles, it is important to understand the basics of current Web architecture. It is built around set of simple technologies, URI for the global and unique identification [Berners-Lee et al. 1998], HTTP for universal content access mechanism [Fielding 1999] and HTML for the presentation on contents [Raggett et al. 1999] in human understandable way. Hyperlinks are used for connecting documents residing at different web servers to make Web connected as an information space. The use of these standard technologies completes the Web of Documents. Hyperlinks give a way to navigation between information from different web resources and facilitate search engines to crawl these connected documents for collecting information.
Linked data is directly built upon these standards and extends the Web architecture. Details of Linked Data Design issues is following:

1. First rule in Linked Data design issue is to assign URI to every tangible and describable things, this means, it is not restricted to only web documents or digital content. These tangible and describable things can be objects or abstract concepts like persons relationship to mass number of gas. For example assigning a URI to the relationship between objects, such as typed relationship URI of wheel to car and typed relationship URI of notebook to pencil. This will increase the encapsulation capacity of web with addition of concepts and their relationships which will bring in more variety to web.

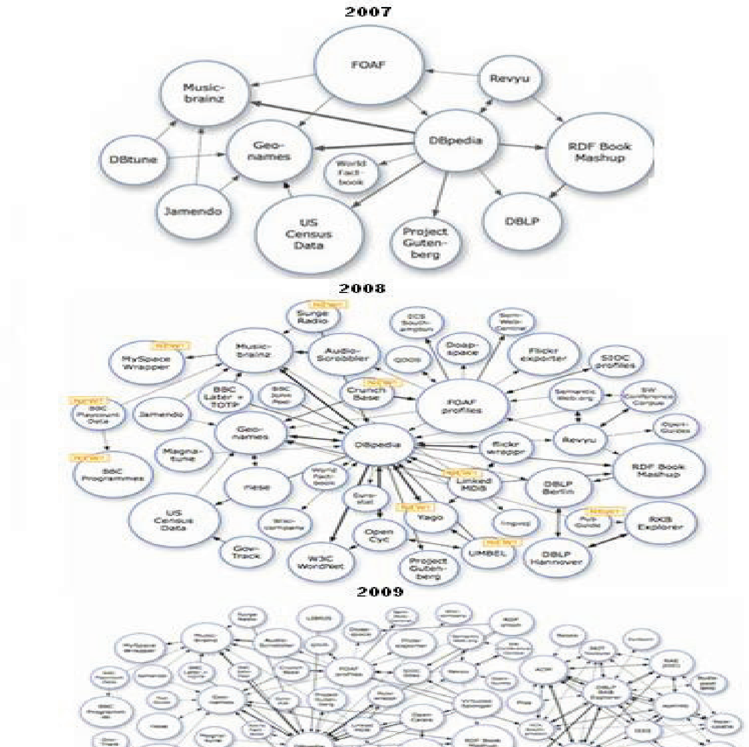
2. HTTP provides a universal content access mechanism and to keep things consistent, Linked Data second principle emphasizes on the assignment of HTTP URIs for naming things, which will work on HTTP protocols to give further description.

3. Third rule directs to use RDF data model for structuring and publishing of the data. This RDF data model is similar to graph model and defines the data as statements. RDF data model is explained in Chapter 2.

4. Fourth and last rule suggests to use typed links for connecting documents or any other concept. Typed links are empowered by vocabularies which gives sufficient information about relationship of two things. For example two persons

are connected with typed friend relationship, both these person have typed publication relationship with the research paper. This rule also iterate on having HTTP URIs on both ends of the typed relationship which if *dereferenced* provide additional information.

Many individuals and organizations have published variety of Linked Data datasets by following these principles. This publishing of datasets shaped into huge connected graph (Web of Data) and is usually illustrated by Linked Data Cloud. Bootstrapping of Linked Data went with Linking Open Data Project and few datasets get published and interlinked in the early stages as shown in figure 3.1. With valuable interest shown by new organization in recent and passed time, an exponential growth in Linked Data cloud has been seen as illustrated by figure 3.1.



2007

2008

2009

Figure 3.1: Linked Data Cloud Growth

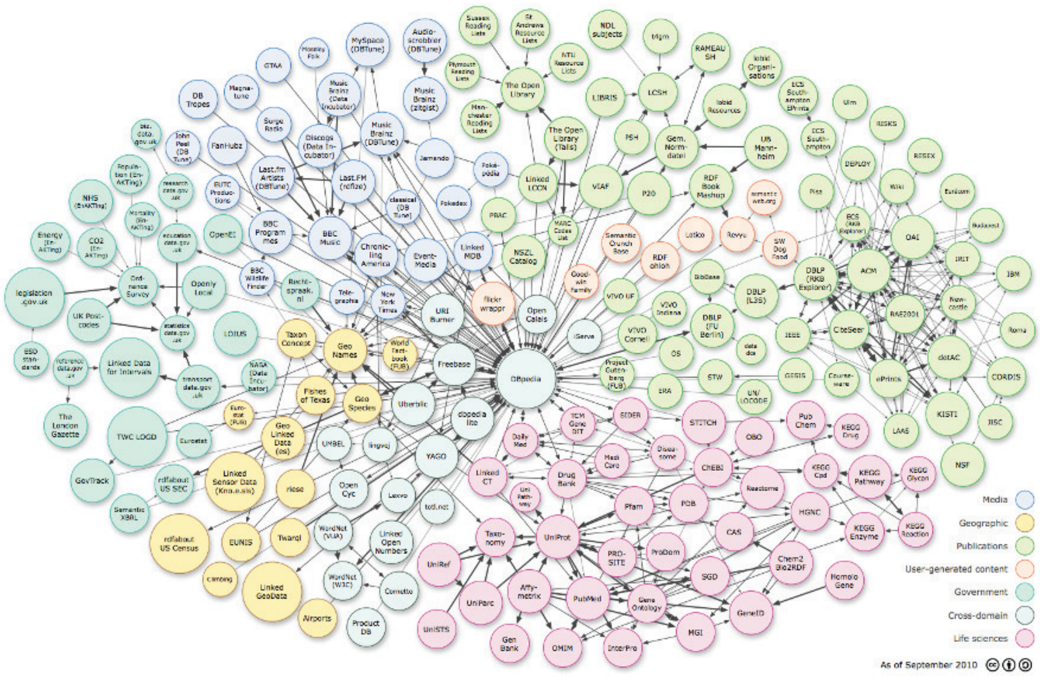


Figure 3.2: ”Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/”

Currently, Linked Data cloud consists of 203 datasets and contains billions of RDF statements interlinked and freely open to use. After four year of transitions, recent Linked Data Cloud gives an impression of giant connected graph, as illustrated by figure 3.2. Linked Data cloud is distributed in seven major disciplines i.e. Geographic, Media, Publications, Life Sciences, Government, Cross-domain and User-generated contents.

Detail about these categories is given below:

Geographic Datasets: Geographic category contains data about places, geographic coordinate and topical things which relates to this category. In total, 16 datasets appeared in this distribution and Geoname[7] considered as central hub for interlinking of geographic data.

[7] http://www.geonames.org/

Media Datasets: One of the major contributors in Linked Data publishing is the Media Industry that is covering data from newspaper, songs and movies. In total, 26 datasets are present in this category. BBC Programmes[8], Music Brainz[9] and Linked Movie DataBase[10] datasets are considered to be the important contributors amongst them.

Publications Datasets: On the other hand, publication category contains data from major digital libraries such as, digital Journal and conferences. Data included in these datasets contain information like papers, books, authors, taxonomies and venues information. In whole, 67 datasets present in this category and are highly interlinked to each other. Some of the major datasets of the publication category are DBLP [11], IEEE[12], ACM[13] and ePrints [14].

Life Science Datasets: Life Science category in Linked Data cloud holds data about genes, proteins, drugs and the facts researched on them. Over all 42 datasets are present in this class from which 20 datasets are published and interlinked by Bio2rdf project [Belleau et al. 2008]. Some major datasets of this category are Drug Bank [15] and Gene Ontology [16].

Government and Public Datasets: This category contains governmental and public sector datasets that covers statistics about populations, transport, education, crime, running development projects and legislations from various countries. This group contains 25 published and interlinked datasets. Eurostat [17] and `legislation.co.uk` [18] are regarded as important dataset of this class.

User Generated Content: Along with all possible important domain dataset, Linked Data cloud also encapsulated social data in it. This social data is generated from communities and published as user generated content category. In total, 7 datasets present in this group, worth mentioning are Reyu [19] and flickr RDF

[8] http://www.bbc.co.uk/programmes

[9] http://dbtune.org/musicbrainz/

[10] http://linkedmdb.org/

[11] http://www4.wiwiss.fu-berlin.de/dblp/

[12] http://ieee.rkbexplorer.com/

[13] http://acm.rkbexplorer.com/

[14] http://eprints.rkbexplorer.com/

[15] http://www4.wiwiss.fu-berlin.de/drugbank/

[16] http://go.bio2rdf.org/

[17] http://ec.europa.eu/eurostat/ramon/rdfdata/

[18] http://www.legislation.gov.uk/

[19] http://revyu.com/

wrapper[20].

Cross-Domain Datasets: Cross-Domain datasets are considered as heart of the Linked Data cloud. These datasets cover all possible topical concepts from multiple domains and are heavily interlinked by other categories. Thus, it provides a basis for single global information graph. Currently, 20 datasets are present in this category and the most populous cross domain datasets among Linked Data community is DBpedia and Freebase[21].

It is evident from above discussion that these distributions of datasets shows the diversity in the Linked Data cloud and access to this data offer various opportunities in Linked Data consumption, motivates for the new Linked Data publishing and awaits for the case studies constructed around this data to show its usability. As discussed in the introduction chapter, this thesis investigated these said areas and presents a framework for publishing, consumption and application areas of Linked Data. The related work is also sectioned according to the proposed framework. In the first section, Linked Data publishing approaches and tools are discussed. Second part elaborated different services and approaches for URI location and tools for Linked Data consumption are talked about. In the final section, some of the general areas are discussed which used Linked Data datasets for application development.

3.4 Linked Data Publishing

Linked Data publishing usually comprises of two steps i.e. 1) Publishing or conversion of raw data into RDF 2) Interlinking of this RDFizied data with external Linked Data resources. To facilitate the data conversion and interlinking many tools are available in Linked Data community. The conversation of raw data is usually represented in legacy of HTML format or along with some structures as spread sheets or tables. These tools server converted data in RDF store to present Linked Data on the web or generate a live view of Linked Data from relation data tables, depending upon the need of a publisher. Breed of these tools harbor data publishers from the technical aspects and publish data according to Linked Data principles. Some of the popular data conversion and interlinking tools are given below.

3.4.1 Triplify

Triplify [Auer et al. 2009] is a small plugin specially designed to bring out Linked Data out of web applications. It works on exploration of SQL queries used to dis-

[20] http://www4.wiwiss.fu-berlin.de/flickrwrappr/

[21] http://www.freebase.com/

play important information on a web page, but are not accessible to search engines. Triplify provides a configuration file, which is filled with important SQL queries, field names and their semantic property equivalent when embedded on a web server, start to publish Semantic Data for machine processing. Triplify produces Linked Data and JSON overview over web application relational database.

3.4.2 SparqPlug

SparqPlug [Coetzee et al. 2008]also known as *RDFizing service* converts the legacy HTML documents in RDF. It works on the serialization of *XHTML Document Object Model(DOM)* and SPARQL query model for data conversion. Further it facilitates users for issuing SPARQL query to get desired HTML content as RDF graph.

3.4.3 D2R Server

D2R server [Bizer and Cyganiak 2006] is a tool for publishing relational database as Linked data on the web. It presents a live view of Linked Data over the relational database and maintains the currency of data. D2R server provides a customizable declarative mapping file in which mapping of the relation table with targeted RDF vocabularies is defined. Based on this declarative mapping file, data publisher can create view of Linked Data over relational database as well as can create data dump in RDF/XML or N3 format. D2R server also provide an interface for this Linked Data presentation and navigation and description of individual resource is made available via HTTP protocol. In addition SPARQL interface enables applications to search and query the database using the SPARQL query language.

3.4.4 SILK Server

Silk framework [Volz et al. 2009] is a tool to perform interlinking, a second task of Linked Data publishing. Silk framework provides a set of services which are used to discover relationship of resource within different Linked dataset. By using *Silk - Link Specification Language data publisher* can specify the type of RDF links which should be present in Linked dataset, as well as the conditions which these links must fulfilled for interlinking. SILK framework works on the data sources which are interlinked with the SPARQL specification and provide SPARQL endpoint for access.

3.5 Linked Data Consumption

Usually, data from Linked Data is consumed in two steps i.e. 1) by locating URI of a resource 2) providing navigation and presentation of Linked Data. Commonly, in spotting URI of a resource SPARQL endpoint and set of services are used.

- In first scenario, a formulated SPARQL query along with HTTP request is sent to specific end point, which in response answers with results in various formats (N3, RDF/XML and JSON). However, for issuing this query an adequate knowledge about SPARQL query constructs and understanding of data is required. Vocabulary of Interlinked Datasets (VoID)[Alexander et al. 2009] is basically used to maintain the meta data description of RDF datasets. This description can be handful to get an overview of the datasets for constructing correct queries.

- Second option for location of URI is to use services. These set of services are provided as an *Application Protocol Interface (API)* by different data providers which take text or URI as an argument. Further it returns the results in form of URI's which further requires authentication and filtering on behalf of developers.

Some of the distinguished services in Linked Data community are introduced next.

3.5.1 SINDICE

Sindice [Tummarello et al. 2007] provides indexing and search services for RDF documents. Its public API allows forming a query with triple patterns that the requested RDF documents should contain. Sindice results often need to be analysed and refined before they can be directly used for a particular use case. Similar kinds of services are provided by semantic search engines like Falcon [Cheng et al. 2008] or Swoogle [Ding et al. 2004]. Sindice is used in developing of proof of concept application CAF-SIAL, mainly due to its larger indexing pool and the ease provided in use of public API.

3.5.2 SameAs

SameAs [22] from RKB explorer provides a service to find equivalent URIs annotated with sameas link in Linked Data datasets. It facilitate to find related data about a given resource URI from different sources in a fast track, but as downside one must have pre knowledge of the URI to be looked in this service. It also provide an API which takes URI as an argument and return result in multiple formats.

[22] http://www.sameas.org

3.5.3 DBpedia Lookup Service

DBpedia Lookup Service[23] provides a keyword search option in DBpedia datasets. It matches keyword with the label property of a resource and return cross-ponding URI of resources. For example resource `http://dbpedia.org/resource/Austria` can be searched with *Austria* keyword. It also provides an API which works on HTTP request and accepts keyword as a parameter to returns URI.

For the navigation and presentation of Linked Data to the end users different tools are around in Linked Data community, some of them are discussed below.

3.5.4 Linked Data Browsers

The current state of the art with respect to the consumption of Linked Open Data for end users is RDF browsers [Berners-Lee et al. 2006] [Kobilarov and Dickinson 2008]. Some tools such as Tabulator [Berners-Lee et al. 2006], Disco[24] , Zitgist data viewer[25], Marbles[26], Object Viewer[27] and Open link RDF Browser[28] can explore the Semantic Web directly. All these tools have implemented a similar exploration strategy, allowing the user to visualize an RDF sub-graph in a tabular fashion. The sub-graph is obtained by dereferencing [Berrueta and Phipps 2009] [Chimezie 2009] an URI, and each tool uses a distinct approach for this purpose. These tools provide useful navigational interfaces for the end users, but due to the abundance of data about a concept and the lack of filtering mechanisms, navigation becomes laborious and bothersome. In these applications, it is a tough task for a user to sort out important pieces of information without having the knowledge of underlying ontologies and basic RDF facts. Keeping in mind these issues, in this thesis a keyword search mechanism is suggested which work in user friendly way to locate and present information.

3.5.5 SPARQL Query Tool

Regarding the problem of searching and filtering in the Web of Data, a number of approaches and tools exist. One approach is to query a SPARQL endpoint that returns a set of RDF resources. There are a few tools that allow to explore a SPARQL endpoint. NITELIGHT [Russell et al. 2008], iSparql [Kiefer et al. 2007], Explorator

[23] http://wiki.dbpedia.org/Lookup?v=1e13

[24] http://www4.wiwiss.fu-berlin.de/bizer/ng4j/disco/

[25] http://dataviewer.zitgist.com/

[26] http://beckr.org/marbles

[27] http://objectviewer.semwebcentral.org/

[28] http://demo.openlinksw.com/rdfbrowser/index.html

[Samur and Daniel 2009] are Visual Query Systems (VQS) [Catarci et al. 2007] allow visual construction of SPARQL queries and differ mainly in the visual notation employed. However, in order to use these tools, the user must have comprehensive knowledge of the underlying RDF schemata and the semantic query languages (e.g. SPARQL). In summary, current tools allow users to manipulate the raw RDF data and do not provide user-friendly interfaces.

3.5.6 Faceted Search Tools

Contrary to VQS applications, Freebase Parallax [Hildebrand et al. 2006], the winner of Semantic Web challenge 2006, is based on the idea of faceted search. Freebase Parallax is a browser for exploring and presenting the structured data in a centralized infrastructure. Similar faceted search application YARS2 [Harth et al. 2006] explores distributed datasets using Subject-Predicate-Object constructs.

3.6 Linked Data Applications

Numerous applications has been built around Linked Data by meshing up data from various Linked datasets. These applications exploit benefits of linking from Linked Data and offer their users advantages over conventional web applications by providing them additional related information automatically. There are many domain specific Linked data applications, few of them are listed below.

3.6.1 dbrec

dbrec[29] [Passant 2010] is a music recommendation website build around Semantic Web technologies and Linked Data principles. It provides recommendation about singers from DBpedia having a relationship with other singers for example (playing the same genre, sharing a common record label or have performed in a same venue). To make things more explicit to the users, it also provides detail insights about every calculated recommendation. This application shows the potential of Linked Data in recommendation calculations, which in not discoverable due to disconnectedness of information in conventional web.

3.6.2 Revyu

Reyvu[30] [Heath and Motta 2008] is a review and rating site about any concept working on the principles of Linked Data, it gathers the review from users and find

[29] http://dbrec.net/

[30] http://revyu.com/

their matches in DBpedia for content enrichment. This information is presented to the users in HTML and for machine consumption in RDF. Reyvu contributes to Linked Data cause in dual way by interlinking of new content with DBpedia and in consuming Linked Data for better user experience.

3.6.3 DBPedia Mobile

DBpedia Mobile[31] [Becker and Bizer 2008] is a location aware Linked Data Browser designed for the I-Phone users. It is perfect use case for tourist, who by giving their current location can find description about the place from DBpedia, related reviews from Reyvu rating site and its photos from flickr wrapper.

3.7 Summary

This chapter introduces the Linked Open Data project in detail. In start, Linked Open Data project and its design issues are presented, elaborated and discussed. Further on various tools in context of Linked Data consumption, Linked Data publishing is listed and described. In the end of this chapter, application belonging to different domains but using Linked Data are presented.

[31] http://beckr.org/DBpediaMobile/

Part II

Linked Data Consumption: Linked Data Value Chain and Discovering Pertinent Resources from Linked Data

Linked Open Data idea claimed its recognition at World Wide Web stage by successfully attracting people attention. This idea motivates people to use Linked Data principles for opening up their wall gardened data which resulted in availability of huge amount of Linked data from various domains. However, applications that consume this data are not yet common. Reasons for this may include one or more of a number of open issues including lack of conceptual frameworks for better understanding of Linked Data, complex mechanism associated with the navigation and exploration of Linked Data resources (search in Web of Data). In addition, it also includes an appropriate end user interfaces for normal web users and lack of methods for seamless integration of Linked Data from multiple sources. Addressing these issues requires the development and investigation of concepts that can be applied in systems which consume Linked Data from the Linked Data datasets. This thesis addresses these research problems. In this regard, a framework *Linked Data Consumption* with innovative features is presented which firstly, identify the various stages in Linked Data fabrication and then propose and implement an application which can consume Linked Data. Main purpose of this application is to consume Linked Data and provides naive end-users with interface to interact with web of data.

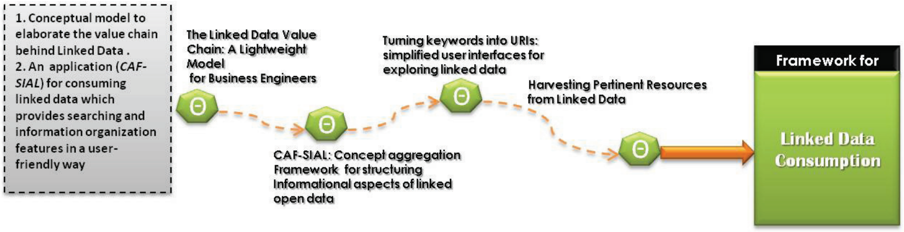

1. Conceptual model to elaborate the value chain behind Linked Data .
2. An application (CAF-SIAL) for consuming linked data which provides searching and information organization features in a user-friendly way
The Linked Data Value Chain: A Lightweight Model for Business Engineers
CAF-SIAL: Concept aggregation Framework for structuring Informational aspects of linked open data
Turning keywords into URIs: simplified user interfaces for exploring linked data
Harvesting Pertinent Resources from Linked Data
Θ
Θ
Θ
Θ
Framework for
Linked Data Consumption

Figure 3.3: Progress Flow Chart of Linked Data Consumption Framework

The figure 3.3, illustrated the research contributions with published work [Latif et al. 2009][Latif et al. 2009a][Latif et al. 2009b][Latif et al. 2010], shows the progress flow for the Linked Data Consumption framework. Firstly, a conceptual model *Linked Data Value Chain* was presented which identifies the different phases in Linked Data generation. It is directed to ease down the understanding process of Linked Data among users. In addition, it highlights the potential pitfalls which may occur in these phases; *cf.* Chapter 4. Investigation of these pitfalls led to an intelligent *URI Keyword mapping technique*; *cf.* Chapter 5 and subsequently a *Concept Aggregation Framework for Structuring Informational Aspects of Linked Open Data* to process the associated semantic information with identified URI was proposed; *cf.* Chapter 6. This framework aggregate, organize and consume information from different Linked Data datasets and present this worked upon information as a multi-aspect profile to the users.

Chapter 4

The Linked Data Value Chain

Linked Data is as essential for the Semantic Web as hypertext has been for the Web. For this reason, the W3C community project *Linking Open Data* has been facilitating the transformation of publicly available, open data into Linked Data since 2007. As of 2010, the vast majority of Linked Open Data is still generated by research communities and institutions while very less has been contributed from business community. There is still some nervousness in the business community for its adaption mainly due 1) lack of business cases on top of Linked Data 2) complicated procedures attached with data transformation 3) lack of consensus on ownership and access rights about legacy and third party data 4) how revenues will can be generated in the Linked Data Sphere. For a successful corporate uptake, we deem it important to have a strong conceptual groundwork, providing the foundation for the development of business cases revolving around the adoption of Linked Data. We therefore present the *Linked Data Value Chain*, a model that conceptualizes the current Linked Data sphere. The Linked Data Value Chain helps to identify and categorize potential pitfalls which have to be considered by business case engineers. We demonstrate this process within a concrete case study involving the BBC.

4.1 Introduction

For several years now, the *Semantic Web* [Berners-Lee et al. 2001] has been of great interest to the international research community. As a subtopic, the concept of *Linked Data* has gained much attention in the recent months.

Linked Data is based on four simple rules [Berners-Lee 2006]:

1. Use URIs as names for things

2. Use HTTP URIs so that people (and machines) can look up those names (see

also [Sauermann et al. 2008])

3. When someone looks up a URI, provide useful information

4. Include links to other URIs so that they can discover more things

To seed the Semantic Web with Linked Data and to promote its adoption, the W3C community project *Linking Open Data* [1] was founded in 2007 [Bizer et al. 2007]. The project helps to solve the causality dilemma (chicken-egg problem) between Semantic Web content and Semantic Web applications by providing RDF [2] data sets from existing open data repositories. To enable intelligent applications that generate a valuable output for the end user, a critical amount of high-quality interlinked datasets across different domains is a crucial precondition, as shown by [Jaffri et al. 2008] and [Raimond et al. 2008]. The vision of the scientific Linked Data community can therefore be described as follows: First, facilitate the generation of semantically enriched Linked Data, and as a result, semantic applications will be built on top of this data.

Linked Data incorporates a lot of potential for enterprises [Servant 2008]. However, there is a significant difference between the aims of a scientific community and the demands and requirements of enterprises, such as revenue flow and generated value. Furthermore, every successful commercial adoption requires the discussion of inherent technical, social and business risks connected to the Semantic Web and Linked Data.

We propose that limited commercial Semantic Web adoption is, among other reasons, caused by the lack of conceptual work supporting the development of business cases and the identification of associated risks. Our publication is motivated by these factors and intends to start a discussion which moves the Semantic Web and Linked Data closer to businesses.

In this we present the Linked Data Value Chain, a model of the Linked Data life cycle along with participating entities and involved roles and types of data. Further, we apply the Linked Data Value Chain to an existing business case from the BBC and use the aforementioned model to highlight potential pitfalls. We conclude our results and present an outlook to potential future research in last section of this chapter.

[1] http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

[2] http://www.w3.org/RDF

4.2 The Linked Data Value Chain

As a prerequisite for the development of successful business cases in the emerging context of Linked Data, three concepts have to be introduced first: *Participating Entities*, their assigned *Linked Data Roles* and processed *Types of Data*. Our contribution tries to support business case engineers with the process of assigning Linked Data Roles to Entities, modelling interactions and responsibilities of Linked Data Roles, and transforming data from Raw Data to Linked Data and Human-Readable Data, thereby increasing its value along the way.

We propose the Linked Data Value Chain as a lightweight model, which builds upon the concepts of Linked Data and Value Chains and makes the interdependencies of entities, roles and different types of data – as the output of the value creation process – explicit.

The value chain as introduced by [Porter 1985] is a concept from the business domain: In a nutshell, a value chain is a chain of activities producing outputs, each activity increasing the value of its particular output, finally shaping a highly valuable end product. In the case of Linked Data with respect to business cases, Human-Readable Data is the most valuable output for the targeted End User.

4.2.1 Participating Entities & Linked Data Roles

In the context of Linked Data, participating entities – both corporate and non-corporate, e.g. persons, enterprises, associations, and research institutes – can occupy one ore more of the following roles:

- A *Raw Data Provider* is a role that provides any kind of data in any non-RDF format.

- A *Linked Data Provider* is a role that provides any kind of data in a machine-readable Linked Data format. Such data is currently provided through dereference-able URIs, a SPARQL endpoint or an RDF dump.

- A *Linked Data Application Provider* is a role that processes Linked Data within an application and generates human-readable output for human end users.

- An *End User* is a human, consuming a human-readable presentation of Linked Data. He or she does not directly get in touch with Linked Data.

The Linked Data Value Chain allows a flexible assignment of roles to entities: In most cases, one entity just occupies one role, but it may – in extreme cases – also occupy all roles at once. For example, one enterprise could own the role of a

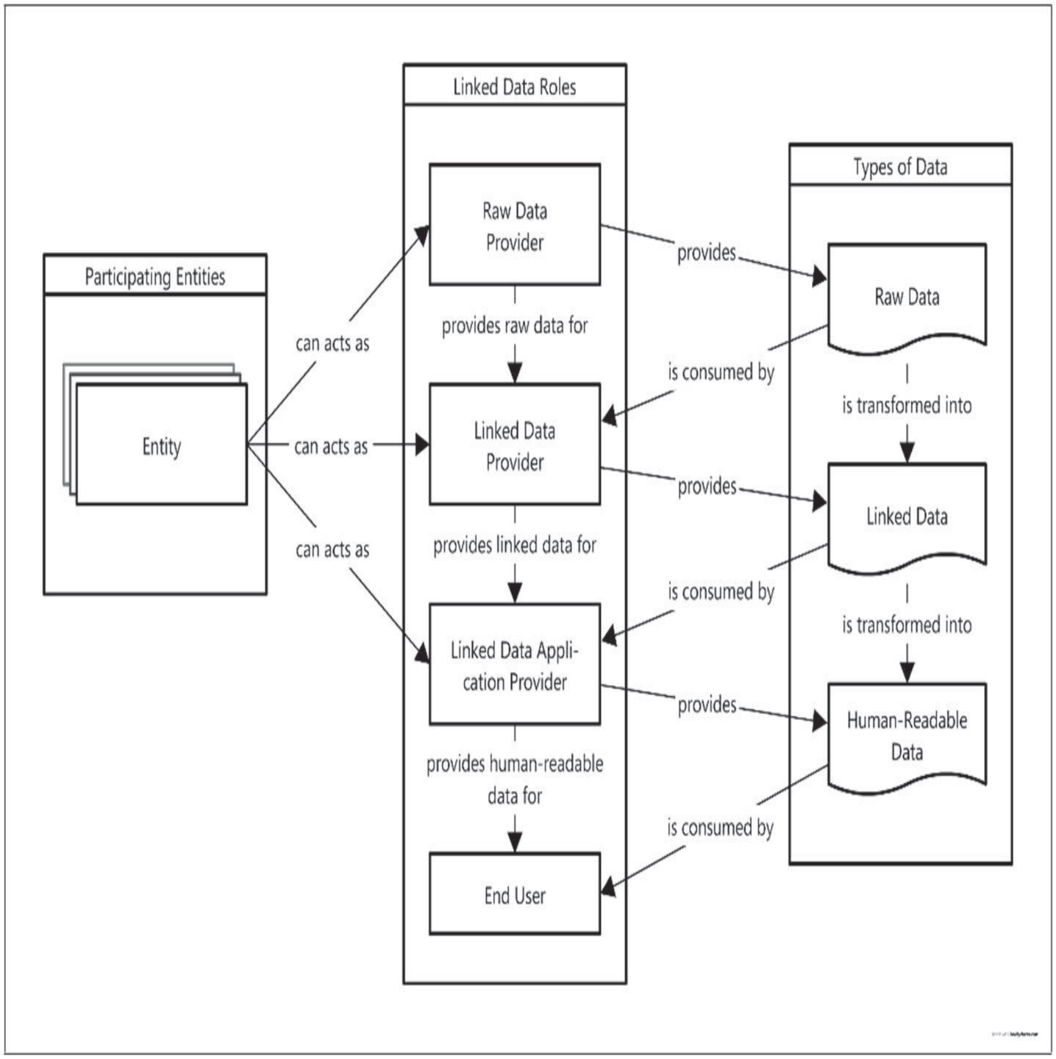


Figure 4.1: The Linked Data Value Chain

Data Provider, a Linked Data Provider, and a Linked Data Application Provider all at the same time. The Linked Data Value Chain also supports multiple sources of data: A Linked Data Provider may acquire Raw Data from more than one Raw Data Provider simultaneously, and will usually provide Linked Data to more than one Linked Data Application Provider.

4.2.2 Types of Data

Three different types of data can be identified within the Linked Data Value Chain:

- *Raw Data* is any kind of data (structured or unstructured) that has not yet been converted into Linked Data. Such data usually has some structure, but generally less structure than Linked Data, and is in most cases also not universally identifiable.

- *Linked Data* is data in a RDF format that uses dereference-able HTTP URIs to identify resources and is linked with other RDF data. This data can be generated by the Linked Data Provider itself, or data provided by a Raw Data Provider can be RDFized. Linked Data is intended to be consumed and processed by machines only.

- *Human-Readable Data* is any kind of data which is intended, arranged and formatted for consumption by humans. Consuming this data generates value for the human end user, which is crucial to the success of any Linked Data business case.

4.2.3 Interaction between Roles and Data

Entities can act in different roles. These roles are closely connected through three types of data, which they provide and/or consume: A Raw Data Provider provides Raw Data as the input for a Linked Data Provider, who turns it into Linked Data, increasing its value by semantically enriching it. Linked Data in turn serves as the basis for a Linked Data Application Provider, who generates Human-Readable Data as the most valuable output for a human End User.

Each combination of roles and entities as well as every transformation step of data holds inherent risks, some of which will be presented in the next section along with a concrete case study. Knowledge about such risks is crucial for the development of successful business cases. If not considered properly, these risks may become pitfalls to business success.

We identified two main areas where pitfalls may arise, grouping them into Role-Related Pitfalls and Data-Related Pitfalls. In a nutshell, Role-Related Pitfalls are either related to individual roles or to the interaction of different roles. Data-Related Pitfalls are either related to the data itself or the data transformation process. We will explain selected pitfalls emerging in the following BBC case study.

4.3 Applying the Linked Data Value Chain

4.3.1 BBC Case Study

As the idea of Linked Data is still young, there are not many appealing Web interfaces for human end users yet. One enterprise that is on the cutting edge, both in regard to deployed Semantic Web technologies and the end-user interface, is the BBC[3]. Furthermore, the BBC is a pioneer when it comes to adopting Linked Data within a business case. Their system utilizes Linked Data technologies to interconnect distributed micro-sites within the BBC network, e.g. BBC News[4] and BBC Music[5], and reuses external data from DBpedia and MusicBrainz [Kobilarov et al. 2009]. By doing so, the BBC generates additional value for the human end users, while allowing them to immediately consume contextually and semantically relevant content from third party sites as well as interconnected BBC sites.

Linked Data Roles	**Participating Entities**
Raw Data Provider	BBC Wikipedia
Linked Data Provider	BBC MusicBrainz DBpedia
Linked Data Application Provider	BBC

Table 4.1: Linked Data Roles and Participating Entities in BBC case study

We apply the Linked Data Value Chain to examine role assignments along with their interactions and data transformation within the BBC business case. As summarized in Table 4.1, BBC acts as Raw Data Provider and Linked Data Provider for their own data as well as Linked Data Application Provider for all data, including external data from Wikipedia (Raw Data Provider) via DBpedia (Linked Data Provider) and from MusicBrainz (Linked Data Provider).

[3] http://www.bbc.co.uk/

[4] http://news.bbc.co.uk/

[5] http://www.bbc.co.uk/music/

4.3.2 Discussion of Potential Pitfalls

In the BBC case study, BBC micro-sites utilize data from DBpedia as an important input for the business case. Unfortunately, *transforming Raw Data* (from Wikipedia) *to Linked Data* (via DBpedia) is a very *time-consuming* and, at best, *semi-automated* effort, which is currently undertaken by a team of researchers [Auer et al. 2007]. Linked Data generated this way is therefore hardly ever *complete, correct and up-to-data* [Jaffri et al. 2008]. There are no service level agreements or similar contracts between BBC and DBpedia or DBpedia and Wikipedia securing all these issues. If not considered well, such performance and data quality risks may become pitfalls.

Second, BBC end users may edit content on BBC Websites which is provided by DBpedia / Wikipedia. Unfortunately, BBC does not provide an *automated feedback loop leading back to the Linked or Raw Data Providers*, in this case DBpedia and Wikipedia. Such feedback loops are not implemented and, most of the time, neither conceptualized yet. Currently, users are requested by BBC to directly edit the respective articles in Wikipedia. Unfortunately, a synchronization of data between Wikipedia and DBpedia will take a very long time, depending on Wikipedia's *data dumping* and DBpedia's *transformation intervals*, possibly annoying a human end user, who certainly is not interested in such technological issues, but wants to see his or her changes promptly, if not in real time.

Third, BBC provides related *links to third party sites*. Such a procedure is a pitfall to successful commercialization, because users may leave the site of the Linked Application Provider (BBC). Reusing data should be based on *widgets and embedded content*, thereby making users stay on the site longer, but still having the benefits from consuming third party contents.

Fourth, users need information about the *provenance of Linked Data* in order to be able to *assess if they trust the third party content* at hand. Therefore, Linked Data Application Providers should state from which Linked Data Providers and Raw Data Providers they present data. Correctness and actuality of data strongly depends on the involved Raw Data Provider, the underlying algorithms and techniques which convert Raw Data into Linked Data as well as on the quality of service provided by Linked Data Provider.

At the moment, the incentives to BBC are clear as they, among other things, want to connect their micro-sites, but when other commercial players phase will start participating with different responsibilities, various pitfalls like obvious consensus on benefits, licensing, real-time interactions between data providers and validation mechanism to judge quality of datasets will arise. Use cases and conceptual designs, which illustrate and exemplify possible benefits, are needed as still in the Linking Open Data paradigm benefits to expose data are not always obvious.

By applying the Linked Data Value Chain, it is apparent that BBC is focusing on limited datasets (DBpedia, MusicBrainz) for cross-linking. We also foresee that in Linked Data cloud there are potential dataset candidate which can be integrated in the BBC system. But this scalability might have inherent risks which also needed to be considered. As a consequence, with inclusion of more roles, range of potential sources of queried data will be immense and to assemble and presenting it to the user according to the context and in a single interface coherently way will be a tough ask [Heath 2008].

Up till now, Linked Open Data has exhibited most of the applications which are at proof of concept level .No real fully linked commercial application has evolved yet. The pitfalls and risks mentioned here may be one of the reason that hinders the true commercialization of Linked Open Data. Clear use cases describing value chain mechanism for commercial players will also be helpful to bridge the gap. To avoid these pitfalls a standardization effort is strongly needed to provide clear and fundamental interactions and licensing framework for the involved stake holders.

4.4 Summary

In this chapter, a Linked Data Value Chain as a lightweight model for business case engineers to support the conceptualization of successful business cases is presented. Moreover, three main concepts identified by Linked Data value chain are introduced which are: Different *Entities* acting in different *Roles* both consuming and providing different *Types of Data*. It was also demonstrated that the assignment of roles to entities, the combination and involvement of roles, the data selected as well as the data transformation process itself hold inherent business risks. In the end of this chapter, Linked Data Value Chain within a concrete case study from BBC to extract pitfalls was applied and validated.

Chapter 5

Locating Intended URI's of Linked Data Resources

The Semantic Web strives to add structure and meaning to the Web, thereby providing better results and easier interfaces for its users. One important foundation of the Semantic Web is Linked Data, the concept of interconnected data, describing resources by use of RDF and URIs. Linked Open Data provides the opportunity to explore and combine datasets on a global scale – something which has never been possible before. However, at its current stage, the Linked Data cloud yields little benefit for end users who know nothing of ontologies, triples and SPARQL. An intelligent KeywordURI mapping technique has been introduced in this chapter to address this concern. The proposed technique has been applied in a simplified end user interface for Linked Open Data. Users don't need to remember a URI any more to find resources from Linked Data. User enters a keyword and the system discovers the most relevant resources from Linked Data. The system employs a two-layered approach. In the first layer users are auto-suggested with resources (URIs) matched to the entered keywords from the locally maintained triple store. In the second layer user's entered keywords is matched with meta-data of resources indexed by a semantic search engine (Sindice). The exploratory evaluations have shown that the system can reduce user's workload in finding required URIs from Linked Data.

5.1 Introduction

Semantic Web has been gaining interest in the international research community for the last decade. In the recent years, the concept of Linked Data has become more and more important. Linked Data, as defined by Tim Berners-Lee, is based on four simple rules [Berners-Lee 2006]:

1. Use URIs as names for things
2. Use HTTP URIs so that people can look up those names
3. When someone looks up a URI, provide useful information
4. Include links to other URIs so that they can discover more things

The core of these guidelines is about assigning unique URI to each entity (Information web resources[1] and Non Information web resources[2])[TAG 2005] and linking it to the other resources (see also [Sauermann et al. 2008]).

To seed the Semantic Web with publicly available Linked Data, the W3C community project Linking Open Data [SWEO Project 2007] was founded in 2007 [Bizer et al. 2007]. The project tries to overcome the causality dilemma (chicken-egg problem) between Semantic Web content and Semantic Web applications by providing RDF data sets from existing open data repositories. As of May 2009, the Linked Open Data cloud consists of over 4.7 billion RDF triples, which are interlinked by around 142 million RDF/OWL links.

DBpedia[3] [Auer et al. 2007], a *semantified* version of Wikipedia, currently plays a central role in the Linked Data sphere. The wide variety of resources described by Wikipedia, and therefore by DBpedia, makes it an ideal candidate for providing URIs for interlinking diverse datasets, describing many different kinds of things, such as people, places, songs, books and many more. Each resource that is described by Linked Data can be uniquely identified by its URI. Relations and attributes of this URI can then be queried by use of SPARQL. However, regular Web users have never even heard of URIs or SPARQL. Therefore, when non-expert users interact with the Semantic Web, the first step is to translate their queries into URIs. For example, when a user wants to know something about *Arnold Schwarzenegger*, it is necessary to find a URI that is the Semantic Web equivalent of this person, e.g. `http://dbpedia.org/resource/Arnold_Schwarzenegger`

Currently, if users want to query the Web of Linked Data, they need to know the URI of the object of interest as well as the vocabularies (or ontologies) that describe it. Factors like little analysis of datasets for interlinking [Jaffri et al. 2008], lack of conceptual model emphasizing on roles at data transformation level [Latif et al.

[1] Digital content with which we interact usually on the traditional web i.e. documents, images, spreadsheets and other multimedia files is known as information resource.

[2] Things other than Digital Content, this can be a concept or tangible thing about which we want to share data i.e. people, places, physical products, chemical compounds, scientific concepts and many more is considered as non-information resources. More precisely all "real-world objects" that exist outside of the Web are non-information resources.

[3] http://dbpedia.org/

2009], and slow adoption of standard vocabularies gives rise to problems of wrong interlinking, duplication, and co-reference of URIs [Hugh et al. 2009]. These factors complicate the task of finding a suitable URI. Finding and identifying the ”right” URI in the ever-growing Linked Data cloud is an ongoing point of discussion (see for example [W3C Discussion 2009]).

There are currently two general practices of getting Linked Data from the Semantic Web:

1. Using SPARQL endpoint to query about resources [SPARQL endpoint].
2. Searching (crawling / dereferencing) RDF documents on the Web [Oren et al. 2008][Ding et al. 2004].

From a developer standpoint, these methods pose interesting challenges. However, when it comes to end users, they should not be bothered with these technical complexities. For users, the Semantic Web should be a black box: Some input, for example keywords, returns some output that is qualitatively better than it would have been before the Semantic Web. Complying with these facts and ongoing discussions, there is a need to devise strategies which can help naive web users to locate a desired URI.

5.2 Test-Sets and Services

This section discusses the state of the art on semantically riched repositories and utilities powering Linked Data. We used following semantic repositories and services as our experiment test-bed for locating URIs.

5.2.1 DBpedia

DBpedia is currently one of the most promising knowledge bases, having a complete ontology along with Yago [Suchanek et al. 2007] classification. It currently describes more than 2.6 million things, including at least 213,000 persons, 328,000 places, 57,000 music albums, 36,000 films, and 20,000 companies [Auer et al. 2007]. The knowledge base consists of 274 million pieces of information (RDF triples). The openly available RDF dumps make DBpedia an interesting subject of study. There has been valuable work done on studying the reliability of Wikipedia URIs [Hepp et al. 2007] that are being used by DBpedia. This study suggests that the meaning of a URI stays stable approximately 93% of the time. Its heavy interlinking within the Linked Open Data cloud makes it a perfect resource to search URIs. To implement our proposed technique, we concentrated on the part of DBpedia that encompasses data about people.

5.2.2 SINDICE

The Sindice [Oren et al. 2008] provides indexing and search services for RDF documents. Its public API allows in forming a query with triple patterns that the requested RDF documents should contain. Sindice results very often need to be analysed and refined before they can be directly used for a particular use case. Similar kinds of services are provided by semantic search engines like Falcon [Cheng et al. 2008] or Swoogle [Ding et al. 2004].

We used Sindice in our work due to its larger indexing pool and the ease provided in use of public API.

5.2.3 SameAS

SameAs[4] from RKB explorer [Glaser and Millard, 2007] provides a service to find equivalent URIs. It thereby makes it easier to find related data about a given resource from different sources. The intelligent use of these datasets and application can help to find a URI in a systematic way.

We proposed a Keyword-URI mapping technique which systematically uses DBpedia personal dataset and Sindice public API service to provide users with seamless mapping of their queried keywords to URIs. We have implemented our technique on to a running application called as CAF-SIAL: Concept Aggregation Framework for Structuring Informational Aspects of Linked Open Data [5].

5.3 Keyword-Uri Mapping Technique

The design of the Keyword-URI mapping is depicted in figure 5.1. The proposed technique is divided into three parts called Triple Construction, Auto-Suggestion, and Semantic Search Service. The triple construction technique discusses the data acquisition and the process of converting it to triples. The auto-suggestion technique discusses how the suggestions are derived from the local data store and highlights the added value of providing seamless URI mapping. In the semantic search service, the querying and filtering of retrieved results is discussed.

5.3.1 Triple Store Construction

The DBpedia data is maintained locally for guaranteed response and to avoid the negative consequences of a sudden downtime of DBpedia. The Persondata dump[6]

[4] http://sameas.org/

[5] http://cafsial.opendatahub.org/

[6] http://downloads.dbpedia.org/3.3/en/persondata_en.nt.bz2

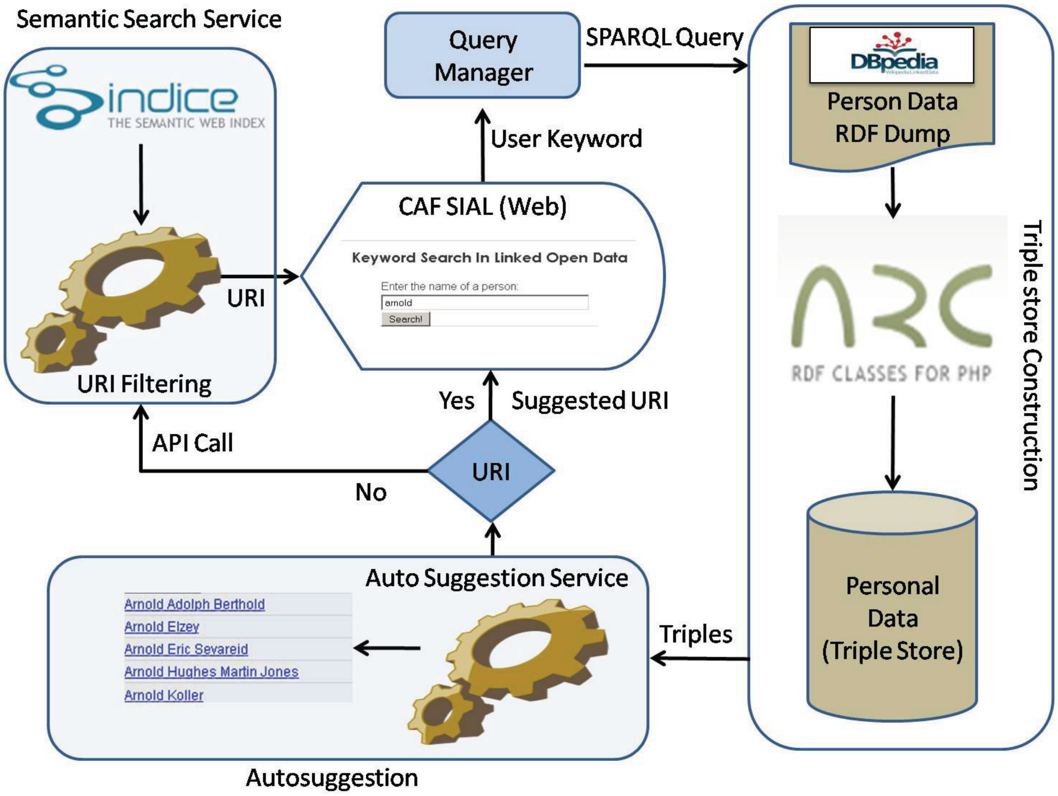


Figure 5.1: Keyword-URI mapping design

was downloaded from DBpedia. This dump contains information about persons extracted from the English and German Wikipedia, represented using the FOAF vocabulary ARC[7], a flexible system for RDF data, is then used to import this dump into a local triple store. This triple store provides an interface for the SPARQL queries. At the moment, there are 62,313 URIs of persons stored in the CAF-SIAL triple store.

5.3.2 Auto-Suggestion

When a user starts entering a keyword for a search, a SPARQL query is constructed on the fly on every key press and an AJAX-enabled auto-suggestion module is invoked. The auto-suggestion module is responsible for finding all possible occurrences of an entered person name in the local triple store and returning a list of suggestions. These suggestions help users in the following aspects:

- With auto-complete, users need to type less
- Give user leverage about searching possibilities within dataset
- On-the-fly disambiguation of concepts having similar or the same names
- Selecting the correct concept

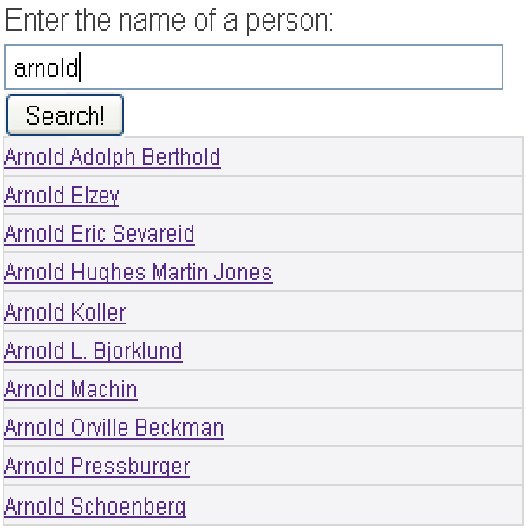
On user selection from any of the suggested option, the underlying URI of the selected keyword is passed on for further processing. The presentation of the list of suggestions is shown in figure 5.2.

5.3.3 Semantic Search Service

In case the keyword is not mapped to any concept in the local triple store, the semantic search service is invoked. The public API of Sindice is used for this operation. It returns an RDF file containing number of URIs belonging to different data sources. This file is then parsed into the local triple store by using ARC. Further on, a URI filtering service is called, and the URI matching our set description, i.e. a DBpedia person type resource, is filtered out. If more than one URI belonging to DBpedia person type exist, the first one in the list is picked.

The SINDICE service provides a faster crawling procedure as compared to other semantic search engines. DBpedia releases new data dumps approximately every six months. Hence the newly entered or updated resources will not be part of the older DBpedia dump as well as our local triple store. Meanwhile Sindice, due to

[7] `http://arc.semsol.org/`



Enter the name of a person:

arnold

Figure 5.2: Auto-suggestions

its fast crawling procedure, will be having an index of these newly added resources, which may be very useful to locate new resources. This will increase our system's performance and ensure up-to-date URI supply to users.

5.4 Exploratory Evaluations

Based on exploratory case studies, we have shown a numeric evaluation for the proposed technique. The evaluation shows that the system is able to find the intended resource with less human effort and help in minimizing workload load of users in semantic search paradigm. For example, if a user searches for the name *Arnold*, our auto-suggestion service, based on DBpedia disambiguation and person data archives, suggests 52 persons having *Arnold* in their first name, last name or as a substring. Users can choose the intended person from the auto-suggested list to view details. This will lessen the effort of the user and will help to disambiguate similar concepts on the fly. The first four entries in Table 5.1 describe the cases when auto-suggestion was successful.

When auto-suggestion module does not return any entry, the semantic search service (Sindice) is initialized. The last 3 entries in Table 5.1 are based on these cases. For example, if a user queries for *Saeed Anwar*, a famous Pakistani cricketer. Sindice will return 29 URIs belonging to different Web resources like DBpedia, DBLP, Mindswap etc. Our URI filtering service will filter out and locate the URI which is referring to the DBpedia person type. By these means, the system is able to find

Keyword	**Search Mechanism**	**Matched Resources**	**Picked URI**
Arnold	Auto-suggestion	52	1
Bush	Auto-suggestion	17	1
Clinton	Auto-suggestion	25	1
Tim	Auto-suggestion	63	1
Parvez Musharraf	Sindice	255	1
Saeed Anwar	Sindice	29	1
Shane Warne	Sindice	107	1

Table 5.1: Exploratory Evaluations

only one URI for the search term from DBPedia. The findings from an exploratory study show that this service will save user from exploring all results and in filtering out the intended result. The findings are given in Table 5.1.

5.5 Summary

In this chapter we investigated two popular resources DBpedia and Sindice (Semantic Indexer) and discussed how intelligent manipulation of current available semantic technologies can be done to locate the URI. We assume that this approach will help users to search in Linked Data huge repository without compulsion of learning Semantic Web mechanics i.e. query languages. Presentation technique like auto-suggestion can also help user to discover new resources and search in more contextualized way.

Chapter 6

Aggregation, Organization and Presentation of Information from Linked Data

Possibly the biggest benefit of Linked Data from the end user viewpoint is the integrated access to data from wide range of multiple distributed and heterogeneous data sources. Simply, this process involves the integration of data from sources which are not clearly selected by users, as explicitly doing so acquire unacceptable working overhead mainly due to association of complex mechanisms and extra learning cycles. The successful Linked Open Data movement has amassed large quantities of structured data from diverse, openly available data sources; there is still a lack of user-friendly interfaces and mechanisms for exploring this huge resource. The main challenge for the development of user interfaces for developers is to provide:

1. Interface option which can give ease to end users in exploration and navigation of the linked data seamlessly as users are accustomed to

2. Consolidated presentation view of the information located from multiple heterogeneous data sources which are understandable to the common end-users, who arguably don't know much about linked data mechanisms.

In this chapter, we describe a methodology for harvesting relevant information from the gigantic Linked Data cloud. The methodology is based on combination of information: identification, extraction, integration and presentation. Relevant information is identified by using a set of heuristics. The identified information resource is extracted by employing an intelligent URI discovery technique. The extracted information is further integrated with the help of a Concept Aggregation Framework. Then the information is presented to end users in logical informational aspects. Thereby, the proposed system is capable of hiding complex underlying semantic mechanics from end users and help users in locating relevant information. In this chapter, we describe the methodology and its implementation in the running application.

6.1 Introduction

World Wide Web can be seen as a huge repository of networked resources. Due to its exponential growth, it is a challenging task for search engines to locate meaningful pieces of information from heavily redundant and unstructured resources. The semantic paradigm of information processing suggests a solution to the above problem: Semantic resources are structured, and related semantic meta-data can be used to query and search the required piece of information in a very precise manner. On the other hand, the bulk of the data currently residing on the Web is unstructured or semi-structured at best.

Therefore, the W3C launched the Linking Open Data[1] (LOD) movement, a community effort that motivates people to publish their information in a structured way (RDF)[2]. Linked Data not only ”semantifies” different kinds of open data sets, but it also provides a framework for interlinking. This framework is based on the rules described by Tim Berners-Lee [Berner-Lee 2006]. As of May 2009, the Linked Open Data cloud consists of over 4.7 billion RDF triples, which are interlinked by around 142 million RDF/OWL links [Auer et al. 2009]. Although Linked Open Data has created huge volumes of data and has attracted the attention of many researchers, it still lacks broad recognition, especially in commercial domains. This is, amongst other reasons, because of complex semantic search and end user applications [Latif et al. 2009a].

In the absence of official standards, DBpedia[3] and Yago[4], amongst others, are considered de facto standards for classification. DBpedia is also a central interlinking hub for Linked Data. Facts about specific resources, extracted from the infoboxes of Wikipedia, are structured in the form of properties as defined by DBpedia's ontology [Auer et al. 2007]. This ontology is associated with Yago's classification to identify the type (person, place, organization, etc.) of the resource. For instance, a query about Arnold Schwarzenegger returns about 260 distinct properties, encapsulating nearly 900 triples in the raw RDF form. Such semantic data is not (easily) graspable by end users. Representing this bulk of structured information in a simple and

[1] http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

[2] http://www.w3.org/RDF/

[3] http://dbpedia.org/

[4] http://www.mpi-inf.mpg.de/yago-naga/yago/

concise way is still a challenge.

Recently, a few applications have emerged, which provide user interfaces to explore Linked Data datasets [Berners-Lee et al. 2006a] [Kobilarov and Dickinson 2008]. These applications use SPARQL endpoints to query Linked Open Data with Subject-Predicate-Object (SPO) logic. SPO logic represents a triple, which is a building block of RDF. A triple establishes a relationship between two resource types. One resource is called subject and the other one object. The relationship between subject and object is called predicate. For example, Arnold Schwarzenegger (subject) is governor of (predicate) California (object). Now, in order to exploit Linked Data resources using SPARQL endpoint with interfaces of recent applications, users have to understand the underlying semantic structures (triples, ontologies, properties). The same gap between semantic search and end user applications has also been identified by [Chakrabarti 2004].

Each resource that is described by Linked Data can be uniquely identified by its URI [Sauermann et al. 2008]. Relations and attributes of this URI can then be queried by use of SPARQL. However, regular Web users have never even heard of URIs or SPARQL. Therefore, when non-expert users interact with the Semantic Web, the first step is to translate their queries into URIs. For example, when a user wants to know something about ”Arnold Schwarzenegger”, it is necessary to find a URI that represents this person in the Semantic Web e.g. `http://dbpedia.org/resource/Arnold_Schwarzenegger`.

To overcome the URI discovery, an intelligent Keyword-URI mapping technique has been introduced as described in Chapter 5. When the system has identified a correct resource URI, then it pro-actively picks up a set of properties related to the selected resource. The most relevant set of properties is grouped together by using the Concept Aggregation Framework. This property set is pre-computed for each resource type. This approach conceptualizes the most relevant information of a resource in an easily perceivable construct.

We also propose a two-step keyword search process in order to hide the underlying Subject-predicate-Object (SPO) logic. In the first step, users search for a keyword, and the system auto-suggests related entries to exactly specify the subject. Then, information related to that subject is structured using the aggregation framework. Furthermore, to avoid searching a specific property (predicate) of the selected subject by its name, a keyword based ’search within’ facility is provided where the specified keyword is mapped to a certain property or set of properties.

6.2 Linked Data Consumption State of the Art

6.2.1 Linked Data Browsers

The current state of the art with respect to the consumption of Linked Open Data for end users is RDF browsers [Berners-Lee et al. 2006] [Kobilarov and Dickinson 2008]. Some tools such as Tabulator [Berners-Lee et al. 2006], Disco[5] , Zitgist data viewer[6], Marbles[7], Object Viewer[8] and Open link RDF Browser[9] can explore the Semantic Web directly. All these tools have implemented a similar exploration strategy, allowing the user to visualize an RDF sub-graph in a tabular fashion. The sub-graph is obtained by dereferencing [Berrueta and Phipps 2009] [Chimezie 2009] an URI, and each tool uses a distinct approach for this purpose. These tools provide useful navigational interfaces for the end users, but due to the abundance of data about a concept and the lack of filtering mechanisms, navigation becomes laborious and bothersome. In these applications, it is a tough task for a user to sort out important pieces of information without having the knowledge of underlying ontologies and basic RDF facts. Keeping in mind these issues, we suggest a keyword search mechanism to reduce the workload of the users.

6.2.2 SPARQL Query Tool

Regarding the problem of searching and filtering in the Web of Data, a number of approaches and tools exist. One approach is to query a SPARQL endpoint that returns a set of RDF resources. There are a few tools that allow to explore a SPARQL Endpoint. NITELIGHT [Russell et al. 2008], iSparql [Kiefer et al. 2007], Explorator [Samur and Daniel 2009] are Visual Query Systems (VQS) [Catarci et al. 2007] allow visual construction of SPARQL queries and differ mainly in the visual notation employed. However, in order to use these tools, the user must have comprehensive knowledge of the underlying RDF schemata and the semantic query languages (e.g. SPARQL). In summary, current tools allow users to manipulate the raw RDF data and do not provide user-friendly interfaces.

6.2.3 Faceted Search Tools

Contrary to VQS applications, Freebase Parallax [Hildebrand et al. 2006], the winner of Semantic Web challenge 2006, is based on the idea of faceted search. Freebase

[5] http://www4.wiwiss.fu-berlin.de/bizer/ng4j/disco/

[6] http://dataviewer.zitgist.com/

[7] http://beckr.org/marbles

[8] http://objectviewer.semwebcentral.org/

[9] http://demo.openlinksw.com/rdfbrowser/index.html

Parallax is a browser for exploring and presenting the structured data in a centralized infrastructure. Similar faceted search application YARS [Harth et al. 2006] explores distributed datasets using Subject Predicate Object constructs.

6.3 Concept Aggregation Framework

The Concept Aggregation Framework aggregates relevant concepts from DBpedia and organizes the most important informational aspects related to a resource.

The scope of this application is limited to DBpedia and Yago. DBpedia covers 23 types of resources (places, people, organizations, etc), initially, we selected the resource type person for the experimentations.

The Concept Aggregation Framework is shown in figure 6.1. The aggregation classification layer is responsible for aggregating the most relevant information related to the person in question. This information is collected based on the list of related properties compiled at the property aggregation layer. The properties are extracted from knowledge bases shown in the aggregation knowledge bases layer.

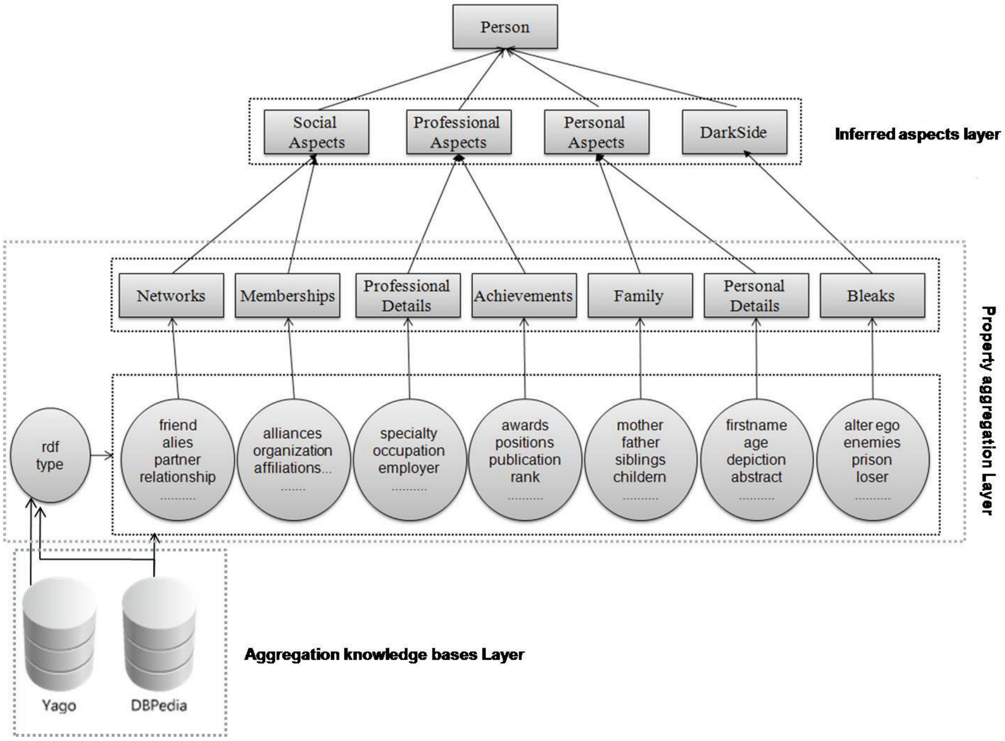
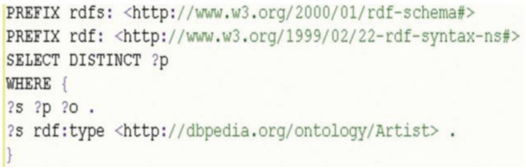
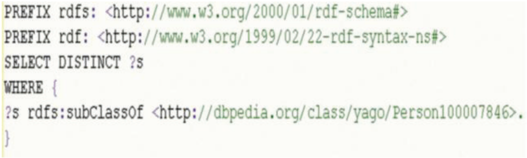

Person
Social Aspects
Professional Aspects
Personal Aspects
DarkSide
Inferred aspects layer
Networks
Memberships
Professional Details
Achievements
Family
Personal Details
Bleaks
Property aggregation Layer
rdf type
friend alies partner relationship ....
alliances organization affiliations... ....
specialty occupation employer ....
awards positions publication rank ....
mother father siblings childern ....
firstname age depiction abstract ....
alter ego enemies prison loser ....
Yago
DBPedia
Aggregation knowledge bases Layer

Figure 6.1: Concept Aggregation Framework



```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT DISTINCT ?p
WHERE {
?s ?p ?o .
?s rdf:type <http://dbpedia.org/ontology/Artist> .
}
```

Figure 6.2: Building DBPedia property dump



```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT DISTINCT ?s
WHERE {
?s rdfs:subClassOf <http://dbpedia.org/class/yago/Person100007846>.
}
```

Figure 6.3: Building Yago classification dump

6.3.1 Aggregation Knowledge Bases Layer

DBpedia, Yago and Umbel ontologies mainly contribute in the identification and classification of the resources. Two of them (DBpedia and Yago) are considered complete knowledge bases [Suchanek et al. 2007]. The underlying mechanism in our system is as follows:

We have generated two knowledge bases, a DBpedia property dump and a Yago classification dump. The DBpedia property dump is built by querying each type of a person (Artist, Journalist, etc.) from SNORQL query explorer[10] (SPARQL endpoint of DBpedia). Then we aggregate all the distinct property sets for each person. Out of 21 queried person types in total, we were able to collect distinct properties of 18, which are presented in Table 6.1. It shows the number of distinct properties in total that we collected for a specific person as well as the number of properties picked by a set of experts, which will be mapped to defined aspects. The formulated query for this operation is given in figure 6.2.

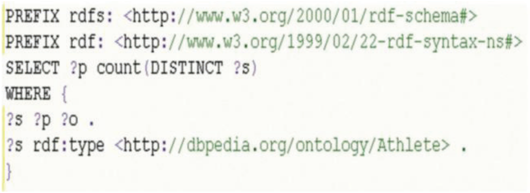
The Yago Classification Dump is built by querying subclasses of Person class from SNORQL query explorer. The query is shown in figure 6.3.

To decide which of these properties should be presented to the user, a query is formulated to get the count of every distinct property used for person type. After getting the count, the rank is assigned to each property. The higher the rank, the more prominently the property will be displayed. For example, some of the properties of person type Athlete like *"Position"* (70939 times), *"clubs"* (46101 times) and *"debutyear"* (9247 times) provide interesting statistics to organize properties in a

[10] http://dbpedia.org/snorql/

Person Type	Total Properties	Picked Properties
Artist	2111	409
Journalist	186	55
Cleric	419	76
BritishRoyalty	252	47
Athlete	2064	496
Monarch	337	50
Scientist	421	126
Architect	132	41
PlayboyPlaymate	125	37
Politician	36	18
MilitaryPerson	725	158
FictionalCharacter	599	273
Criminal	287	74
CollegeCoach	282	124
OfficeHolder	1460	634
Philosopher	226	71
Astronaut	168	62
Model	211	99

Table 6.1: Selection of persons' properties from DBPedia



```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT ?p count(DISTINCT ?s)
WHERE {
?s ?p ?o .
?s rdf:type <http://dbpedia.org/ontology/Athlete> .
}
```

Figure 6.4: Computing property rank

more conceivable fashion. The formulated query to get the count of each distinct property is shown in figure 6.4.

6.3.2 Property Aggregation Layer

This layer first identifies the profession type. This works in two steps. In the first step, the resource type (RDF type) is identified by using DBpedia. In the case where in the retrieved set of properties, there is no property mapped within DBpedia knowledge base, the system tries to map the retrieved property to a Yago class. For example if the retrieved property is *"AustrianComputerScientist"* which is not listed in DBpedia knowledge base, then the system maps it to the Yago hierarchy and can infer that the person belongs to the profession of *"Scientist"* because *"AustrianComputerScientist"* is a subclass of *"Scientist"*.

Based on a resource type, we have extracted all the possible properties from the DBpedia Property Dump. We then have manually identified sets of properties indicating an informational concept (networks, memberships, family, achievements etc.) related to a person. These concepts are aggregated and mapped to the related informational aspect identified in the inferred aspects layer. More than one concept may be mapped to a single informational concept defined at the inferred aspects layer.

6.3.3 Inferred Aspects Layer

The information for a resource such as person may be organized and viewed in different informational aspects like personal, professional, social etc. The most popular search engine like Google also tries to present such informational aspects related to a topic in its top results. It has been shown in [Brin and Page 2008] that how Google rank its results to provide the most relevant contents. For example, in a response to a user query of *"Bill Clinton"*, Google top ten results are based, amongst other things, on personal information (biography) and his professional career (president, writer). These results, however, depend on the complex link analysis of Web pages (citations to Web pages from different sources) along with weight mechanisms assigned to different factors [Feldstein 2009] [Boykin 2005]. Google is considered as the most popular search engine having 64.2% share in U.S search market [Lipsman 2009]. Inspired from Google's success in calculating and presenting the results in diverse and important informational aspects related to a query, we developed a concept aggregation framework where diverse yet important aspects of a person are represented in inferred aspect layer.

6.4 System Architecture

The system architecture is depicted in figure 6.5. The implemented system is divided into four modules called query manager, auto-suggestion module, information retrieval module and search within property module. The query manager is a controlling module of the application. It is responsible in translating the keyword search query into SPARQL queries. The auto-suggestion module helps users to disambiguate entered search term. The information retrieval module is responsible for locating the URIs and extracting related information. The search within property module provides the facility of searching within all retrieved properties of a resource.

6.4.1 Auto-Suggestion Module

The query manager triggers the auto suggestion module by converting the searched keyword of a user into a SPARQL query. This module interacts with the DBpedia person and the DBpedia disambiguation triple store to auto-suggest persons with names that match the entered keyword. This module has been discussed in detail in section 6.3. If the user does not select any of the suggested terms, or in case of a distinct query (no auto-suggestions yielded), the searched term is passed on to the information retrieval module for further processing.

6.4.2 Information Retrieval Module

This module is further divided into four processes:

1. URI locator

2. LOD retrieval

3. Parser

4. Concept aggregation

The searched term is passed to the URI locator process which will query the locally maintained data sets (i.e. DBpedia Title TS, DBpedia Person Data TS, and

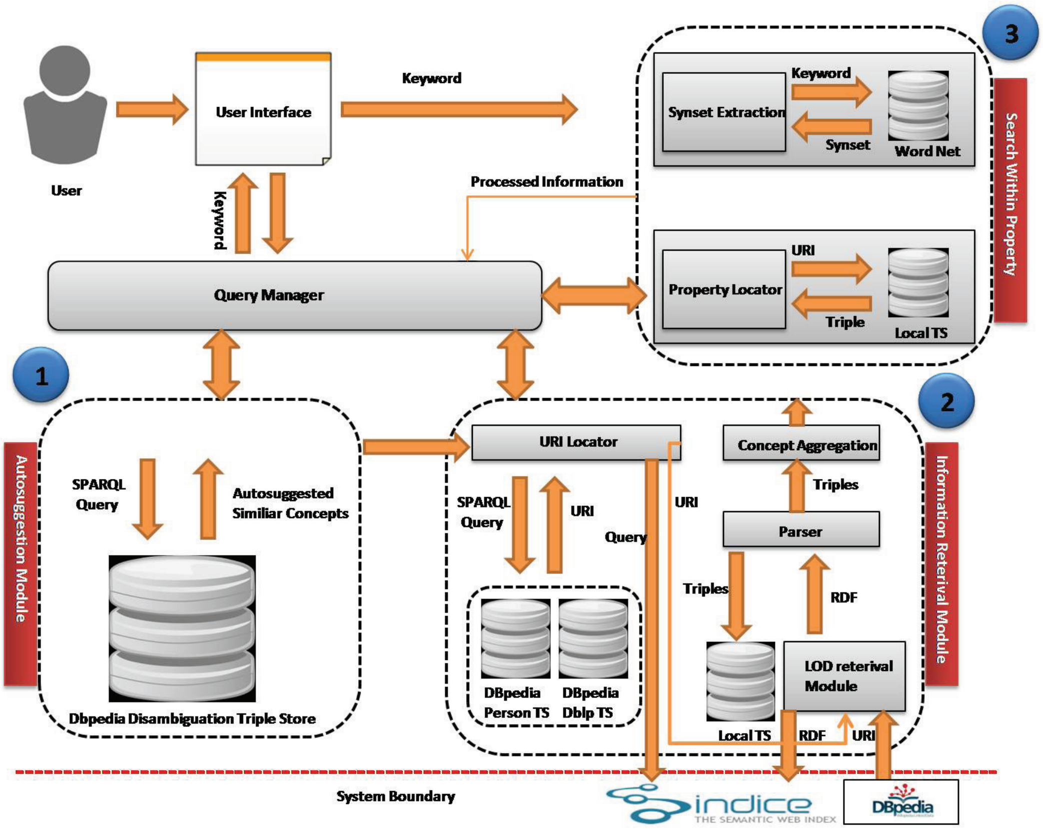


Figure 6.5: System architecture for CAF-SIAL

DBLP TS) to get a URI. If this fails, a new query is formulated for the SINDICE[11] Web service to locate the URI. After locating the URI of a resource, the Linked Data retrieval process dereferences that URI at the DBpedia server to get the respective resource RDF description. This RDF description is further passed to the Parser process. This process parses RDF description into triples and stored them locally. Then, the concept aggregation process is called to sort out the most important information aspect of the resource and in the end; the output is presented to the user.

6.4.3 Search Within Property Module

This module lets the user search within all properties of a resource retrieved from the information retrieval module. When a user enters a keyword to search some information about a resource, the synset extraction process queries wordnet[12] to retrieve the synset of searched keyword. This synset is passed to the query manager and for each word in the synset, it query the local triple store through the property locator process. The property locator process matches the keyword as substring in the retrieved property set. All matched properties are then extracted and presented to the user.

6.5 Use Case Scenario

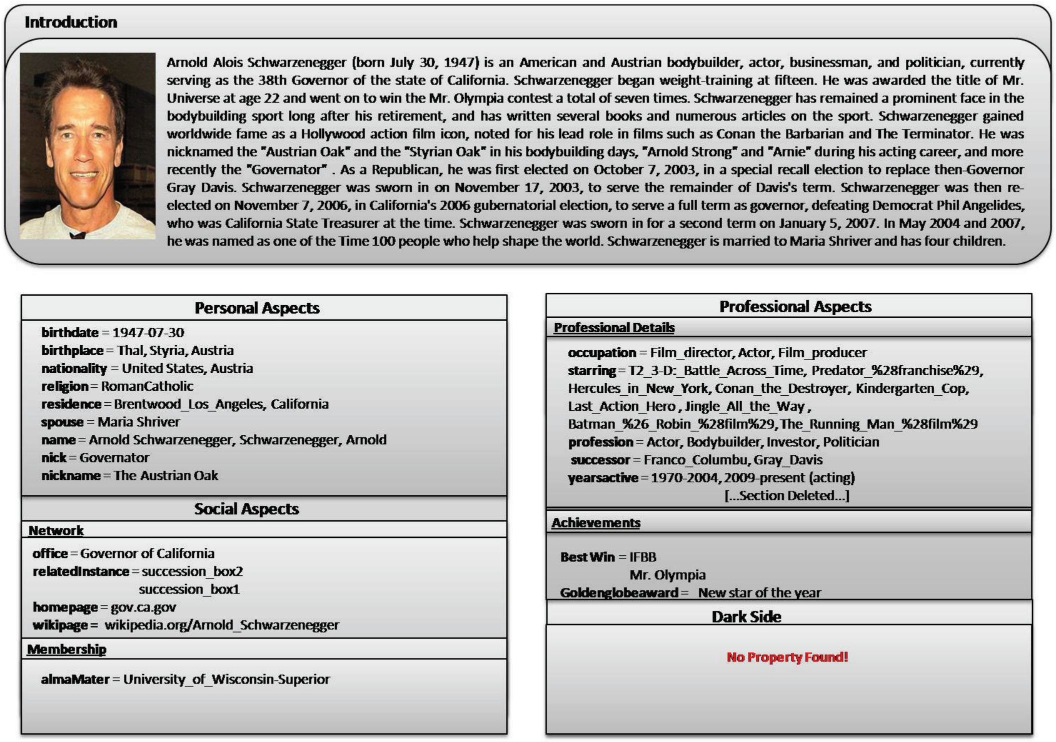
The working of the system is described with the help of a use case scenario. We have selected *"Arnold Schwarzenegger"* for this example. The reason of this selection is that the selected person is affiliated with four interesting and diverse professions along with multiple awards and achievements. This will help in understanding the overall working of the system.

These capabilities make him a distinct person and a suitable choice for the use case. The application flow is explained as follows: User starts typing the search term *"Arnold"*. The persons' names starting with the keyword *"Arnold"* are auto-suggested. For example *"Arnold Bax"*, *"Arnold Bennett"*, *"Arnold Schwarzenegger"* etc. as depicted in figure 6.6.

The user selects *"Arnold Schwarzenegger"* to see his details as shown in figure 6.6. The output is comprised of different informational aspects such as social, personal, and professional. Important properties are shown on the top for each informational aspect. The important property list was prepared manually and the weight to each

[11] http://sindice.com/

[12] http://wordnet.princeton.edu/wordnet/



Introduction

Arnold Alois Schwarzenegger (born July 30, 1947) is an American and Austrian bodybuilder, actor, businessman, and politician, currently serving as the 38th Governor of the state of California. Schwarzenegger began weight-training at fifteen. He was awarded the title of Mr. Universe at age 22 and went on to win the Mr. Olympia contest a total of seven times. Schwarzenegger has remained a prominent face in the bodybuilding sport long after his retirement, and has written several books and numerous articles on the sport. Schwarzenegger gained worldwide fame as a Hollywood action film icon, noted for his lead role in films such as Conan the Barbarian and The Terminator. He was nicknamed the "Austrian Oak" and the "Styrian Oak" in his bodybuilding days, "Arnold Strong" and "Arnie" during his acting career, and more recently the "Governator" . As a Republican, he was first elected on October 7, 2003, in a special recall election to replace then-Governor Gray Davis. Schwarzenegger was sworn in on November 17, 2003, to serve the remainder of Davis's term. Schwarzenegger was then re-elected on November 7, 2006, in California's 2006 gubernatorial election, to serve a full term as governor, defeating Democrat Phil Angelides, who was California State Treasurer at the time. Schwarzenegger was sworn in for a second term on January 5, 2007. In May 2004 and 2007, he was named as one of the Time 100 people who help shape the world. Schwarzenegger is married to Maria Shriver and has four children.

Personal Aspects

birthdate = 1947-07-30
birthplace = Thal, Styria, Austria
nationality = United States, Austria
religion = RomanCatholic
residence = Brentwood_Los_Angeles, California
spouse = Maria Shriver
name = Arnold Schwarzenegger, Schwarzenegger, Arnold
nick = Governator
nickname = The Austrian Oak

Social Aspects

Network

office = Governor of California
relatedInstance = succession_box2
succession_box1
homepage = gov.ca.gov
wikipage = wikipedia.org/Arnold_Schwarzenegger

Membership

almaMater = University_of_Wisconsin-Superior

Professional Aspects

Professional Details

occupation = Film_director, Actor, Film_producer
starring = T2_3-D:_Battle_Across_Time, Predator_%28franchise%29, Hercules_in_New_York, Conan_the_Destroyer, Kindergarten_Cop, Last_Action_Hero , Jingle_All_the_Way , Batman_%26_Robin_%28film%29,The_Running_Man_%28film%29
profession = Actor, Bodybuilder, Investor, Politician
successor = Franco_Columbu, Gray_Davis
yearsactive = 1970-2004, 2009-present (acting)
[...Section Deleted...]

Achievements

Best Win = IFBB
Mr. Olympia
Goldenglobeaward = New star of the year

Dark Side

No Property Found!

Figure 6.6: Informational aspects of Arnold Schwarzenegger

property is assigned on the basis of its count automatically. The screen shot shows his important professional details concisely and in easily graspable manner.

6.6 Summary

To reap the real benefits from Linked Data multiple heterogeneous resources for the end users, we have proposed a Concept Aggregation Framework to facilitate users in searching with in Linked Data sphere and presented them with a consolidated view of queried resource (Persons) in as simple as possible way. This work tries to bridge the gap between semantic search and the end user. The proposed keyword-based search mechanism has simplified the process of finding information from Linked Data by hiding underlying semantic logic. With the help of Concept Aggregation Framework, the information related to a resource (consisting of hundreds of properties) was structured in major and most relevant categories of informational aspects. This reduced the user effort to find the required information. Keyword-URI locating technique was very helpful to identifying a particular resource from huge Linked Data repository. The evaluation of the systems has shown promising results. This application is online and accessible at `http://cafsial.opendatahub.org`. We assume that this application is a step forward for the future applications development which can be built on top of the Linked Data by intelligent manipulation of underlying semantic mechanisms.

Part III

Weaving Scholarly Legacy Data into Web of Data

One of the popular medium for codified information dissemination is Electronic Journals. Electronic journals publish manuscripts online and offer indexing, searching, interactive visualizations to users, and a number of functionalities. With the advancement in science and Open access movement [Roberts et al. 2001], the need of open electronic publishing has increased exponentially. Different approaches have been defined to make this huge repository accessible to all scientific community [Marchionini and Maurer 1995]. With the acceptance of open access movement, many open access journals have subsequently emerged these days. The Journal of Universal Computer Science [J.UCS 1994] is one of them. The Journal of Universal Computer Science (J.UCS) is a high-quality electronic publication that deals with all aspects of Computer Science [Calude et al. 1994]. J.UCS has incorporated many innovative features from its commencement, such as the enabling of semantic and extended search, Experts, Tag recommendations and Links into Future. But all this features are accessible in legacy HTML format and due to unstructured nature of its format, very little is offered to the machine processing domain.
The Linked Data Project supports the Open data movement in a way, to open information with more structured metadata, which is machine understandable, easily locatable, identifiable and linkable for better searching and discovery processes. To reap the real benefits offered by Linked Data paradigm and taking J.UCS to Linked Data sphere, in this thesis a framework for *Linked Data publishing* is provided to convert Legacy HTML data of Open Digital Journal (J.UCS). This framework first model all the artefacts of Digital Journal (paper, authors, recommended tags and experts) by using various ontologies and then find their relevant bits in Linked Data datasets for interlinking with external resources. Further on, all this data is RDFized as Linked Data and provided openly in the Linked Data cloud. An HTML interface to navigate between these linked datasets and SPARQL endpoint for querying is also setup for the users.

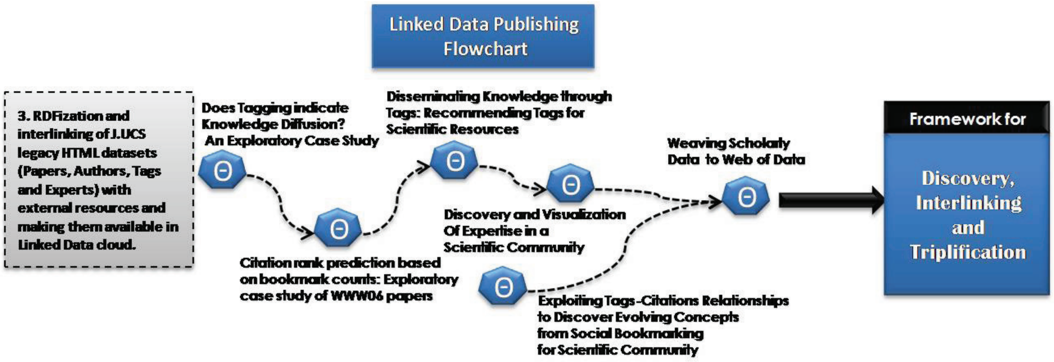

Linked Data Publishing Flowchart
3. RDFization and interlinking of J.UCS legacy HTML datasets (Papers, Authors, Tags and Experts) with external resources and making them available in Linked Data cloud.
Does Tagging indicate Knowledge Diffusion? An Exploratory Case Study
Citation rank prediction based on bookmark counts: Exploratory case study of WWW06 papers
Disseminating Knowledge through Tags: Recommending Tags for Scientific Resources
Discovery and Visualization Of Expertise in a Scientific Community
Exploiting Tags-Citations Relationships to Discover Evolving Concepts from Social Bookmarking for Scientific Community
Weaving Scholarly Data to Web of Data
Framework for
Discovery, Interlinking and Triplification

Figure 6.7: Flow Chart of Linked Data Publishing

The figure 6.7, illustrated the research contributions with published work [UsSaeed et al. 2008] [UsSaeed et al. 2008a] [UsSaeed et al. 2010] [Afzal et al. 2009] [Afzal and Latif 2011] [Latif and Afzal 2011a], shows the progress flow for the Linked Data publishing framework. At the initial stage, importance of tags and expertise calculation in a digital Journal Environment were proved. These studies led to the implementation of Tag Recommendations system from CiteULike and Experts in J.UCS web server; *cf.* Chapter 7. In addition to convert all this Legacy, HTML data to Linked Data for machine processing and to make it available in Linked Data community, a framework for Linked Data publishing which encapsulates the innovative approaches for modeling, interlinking and RDFization is proposed in the end; *cf.* Chapter 8.

Chapter 7

Preparing Data from Social Web and Digital Journal

7.1 Introduction

The Linking Open Data project provides a new publishing paradigm for creating machine processable structured data on the Web. It also offers best practices for interlinking related data lying isolated wall gardened by using typed linking. The inclusion of different domain oriented datasets in Linked Data cloud shows its success and acceptance in communities from different walks of life. At present structured data from domain of Government, Geo-locations, Medicine, Wikipedia and Social communities e.g. is available in Linked Data cloud. Many big companies in World Wide Web e.g. BBC, Freebase, Music Brainz have explored this data to make interesting applications. The heavy presence of scientific publications datasets (e.g. DBLP, IEEE, ACM and Citeseer.) in Linked Data cloud also underpins the importance of Linked Data in the Scientific Community.

With the advancement in science and Open access movement [Roberts et al. 2001], the need of open electronic publishing increased exponentially. Different approaches have been defined to make this huge repository accessible to all scientific community [Marchionini and Maurer 1995]. With the acceptance of open access movement, many open access journals have subsequently emerged. The resources provided by the open journals are presented to users in the legacy HTML format. Due to unstructured content generation, lot of the important data stayed meaningless and unlinked. In the HTML data presentation paradigm, more emphasis is put on the metadata generation at document level, the important concepts present in form of data remained unfocused and sustainment of data context and description becomes difficult due to untyped linking within and outside the dataset. For example to spot scientific papers published by an author in different venues which are indexed by various digital libraries will be bothersome, one has to rely on heuristics. Further on the integration and disambiguation will be difficult due to inability of HTML providing enough semantics.

Linked Data provides solutions to above problems, in where not only documents but also data is considered as superior citizen. Every concept is assigned a unique identity (URI) and defined typed relationship makes the data more explicit. Data concepts are supported by enough semantics, the disambiguation and integration of data became easy. As data is structured in RDF graph model, variety of queries also become possible.

Taking in account the innovative features presented by Linked Data, we have RDFized the legacy HTML data of digital journal. The journal selected for this study is Journal of Universal Computer Science [J. UCS 1994]. The journal provides number of innovative features and many important datasets i.e. Links into future [Afzal 2009], experts [Afzal et al. 2008] and recommended tags for papers from social Bookmarking System [Afzal et al. 2010]. We are of view that RDFizing these datasets can prove good for the scientific community to use these resources for knowledge creation and discovery.

For the RDFization of J.UCS datasets we have employed an innovative modeling and interlinking strategy. To model the typed relationship with in and outside dataset, we reused well known ontologies like Dublin Core, SIOC etc. For the interlinking we choose DBpedia, DBLP. In DBpedia we search for the author's URI and verify them on devised heuristics. While in DBLP we search for the contributions / publications published by authors in different venues. After RDFization process, J.UCS dataset has been interlinked within the Linked Data cloud and also made available as RDF dump. For human consumption an HTML interface over RDFized data is provided for navigation through J.UCS resources. For machine consumption and querying a SPARQL endpoint is setup and made available.

In following section, the preparation of tags and recommendation of tags for WWW'06 and J.UCS datasets from social bookmarking system CiteULike is presented and discussed.

7.2 Preparing data from Social Bookmarking System

Social bookmarking and tagging has become a very successful phenomenon in the Web and getting more popular day by day. Systems adhering to these principles transform the way in which users manage and disseminate content in the conventional web environment. These systems enable the users to add keywords (tags) to web resources (web-pages, images, documents, papers) without having to rely on a controlled vocabulary [Marlow et al. 2006]. It's potential to improve the search on the web, resulted in new forms of social communication and generated new opportunities for data mining.

In previous studies [UsSaeed et al. 2008], we have investigated the importance of tags in a scientific domain. It has shown that there exist a positive correlation between bookmark counts and citations. Based on these potential uses of citations in the research community it is argued that bookmark counts of research papers can be used in a similar way as an alternative popularity indicator. Building on that, a tag based resource recommender system was purposed for scientific papers which can provide links to the most related resources from social bookmarking system by providing contextual tags.

The tag based scientific paper recommender system exploits author keywords of scientific publications to link these resources with tags in CiteULike which is a social bookmarking and tagging application. For a focused resource, the tags extracted from CiteULike based on author keywords were compared with the corresponding tag cloud of CiteULike. The result shows that system extends the authors keyword set with social tags providing links to rich and focused resources in CiteULike. This also enhances the serendipitous discovery of emerging concepts related to that resource. Such a system may enhance the discovery of related and popular resources for researchers hence furthering the dissemination of knowledge.

7.3 Annotation of Papers with Tags

For this exercise, we selected two datasets WWW'06 and J.UCS. The social bookmarking system used in these experiments was CiteULike. CiteULike is a social bookmarking system where a huge number of users share scientific papers and tag them accordingly. Major task is to find the most relevant resources from CiteULike for all papers published within WWW'06 and J.UCS. On the WWW'06 side and J.UCS, every paper is assigned with suitable keywords by the authors of the paper, while on CiteULike side, papers are tagged with some keywords by the users of the CiteULike. To find relevant resources for WWW'06 papers and J.UCS, authors' assigned keywords were used to mine the tags from CiteULike . The papers at WWW'06 and J.UCS are further annotated with the matched tags. In the following sections we will discuss details about datasets, processes and results in this study.

7.3.1 WWW'06 dataset

This dataset is comprised of all published full papers in the conference World Wide Web 2006. The statistics are shown below:

Total papers published in WWW'06 = 84
Total Keywords for all papers = 5129
Unique Keywords = 107

7.3.2 J.UCS Dataset

The dataset for J.UCS was acquired until volume 15, issue 7. The statistics[1] are shown below:

J.UCS total papers = 1460 (until volume 15, issue 7)
J.UCS papers other than managing editor column = 1271
J.UCS papers having one or more author's keywords = 1187
Total keywords for 1271 papers were 5397
Unique Keywords were 3935

7.3.3 CiteULike dataset

The dataset of CiteULike was acquired in August, 2009. The statistics for tags and papers is shown below.

Total tag assignments in CiteULike = 6.5 million
Total Papers in CiteULike = about 2 million
Unique tags = 348420

7.3.4 Matching Author's Keywords with CiteULike Tags

To match papers' keywords of WWW'06 and J.UCS with CiteULike tags, a two-tier approach was adopted. First, the tag extraction tries to find an exact match between papers' keywords and CiteULike tags. Subsequently, a partial match between both datasets was checked. The partial match enhanced discovery of relevant tags but also introduced some noise. Afterwards, some heuristics were used to clean the noise and the discovered tags were used to annotate the corresponding papers.

Direct Match of WWW'06 Author's Keywords

WWW'06 papers for which at least one keyword is matched= 52/84 = 62%
Unique Keywords of WWW'06 matched = 102/107 = 95%

[1] J.UCS tags related statistics was appeared in PhD Thesis "Context Aware Information Discovery for Scholarly e-Community" of Muhammad Tanvir Afzal at TUGraz and to appear in my Co-authored Paper [Afzal and Latif 2011].

Partial Match of WWW'06 Author's Keywords

WWW'06 Papers for which at least one tag is matched = 52/84 = 62%
Total results of WWW'06 Keywords matched with CiteULike = 5129
Total CiteULike unique tags matched = 4228/348420

Direct Match of J.UCS Author's Keywords

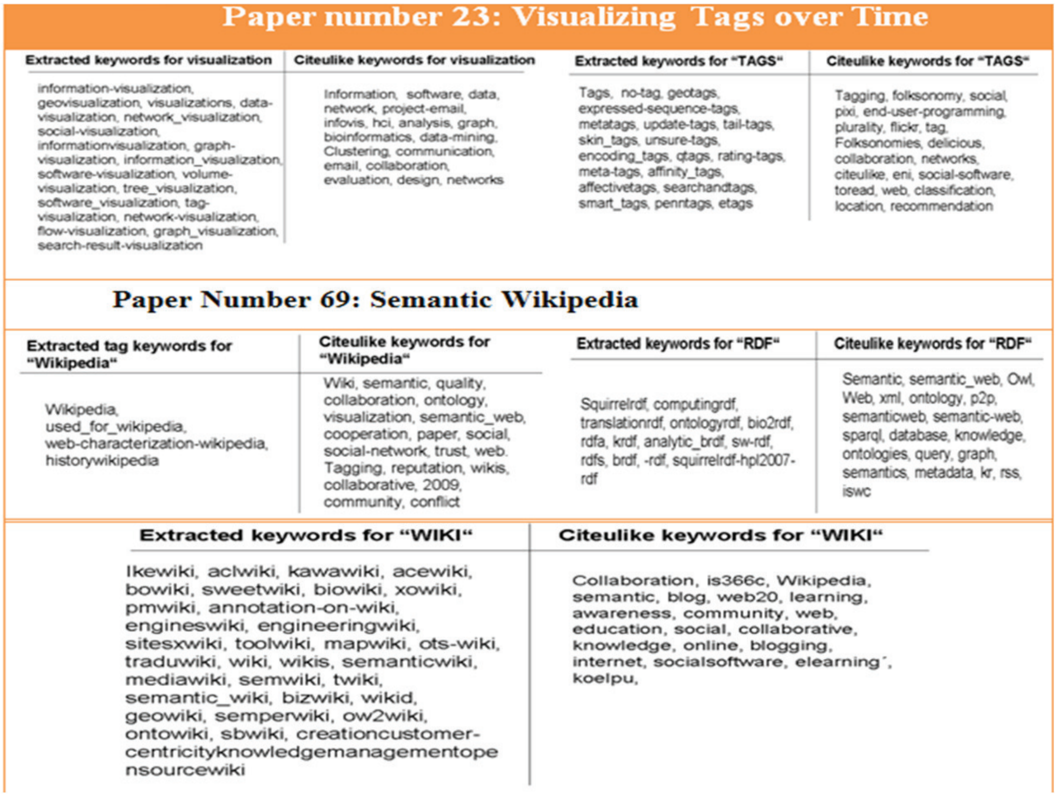
J.UCS Papers for which at least one keyword is matched= 665/1187 = 56%
All J.UCS Keywords matched = 760/3935 = 19%

Partial Match of J.UCS Author's Keywords

J.UCS Papers for which at least one tag is matched = 683/1187 = 58%
Total J.UCS Keywords matched =797/3935= 20%
Total CiteULike unique tags matched = 91766/348420
The knowledge discoveries are significantly enhanced by employing partial match as compared to direct match which brought more variety in the end results.

7.4 WWW'06 and J.UCS Case studies

The goal of this study was to discover and annotate the scientific resources papers (WWW'06 and J.UCS) with set of relevant and focused tags by using social bookmarking source. For simplification an orientation of tags in CiteULike and extracted tags for two sample papers are given in figure 7.1. A heading "CiteULike keywords for visualization" in figure 7.1 provides a list of tag terms on a user search keyword 'visualization' (only top 20 are shown). These terms are computed from the tag co-occurrence frequency, for example, terms related to 'visualization' search keyword are the terms which same users assigned to resources along with tag term 'visualization'. This list definitely brings diversity in tags but remain distant from focused resources putting an extra burden on user to find focused resources. However in our case, system extracts the tag terms from CiteULike tags based on direct and partial match of authors' keywords of a particular research paper. In this way the highly relevant discovered tags are linked with the paper. These tags were then compared in CiteULike tags by using direct and partial match. The extracted tag terms for 'visualization' and 'tags' are shown in figure 7.1. The extracted tags for visualization remains in the same focus and will link the resources in CiteULike which will often be related to the scope of visualization.



Paper number 23: Visualizing Tags over Time

Extracted keywords for visualization	Citeulike keywords for visualization	Extracted keywords for "TAGS"	Citeulike keywords for "TAGS"
information-visualization, geovisualization, visualizations, data-visualization, network_visualization, social-visualization, informationvisualization, graph-visualization, information_visualization, software-visualization, volume-visualization, tree_visualization, software_visualization, tag-visualization, network-visualization, flow-visualization, graph_visualization, search-result-visualization	Information, software, data, network, project-email, infovis, hci, analysis, graph, bioinformatics, data-mining, Clustering, communication, email, collaboration, evaluation, design, networks	Tags, no-tag, geotags, expressed-sequence-tags, metatags, update-tags, tail-tags, skin_tags, unsure-tags, encoding_tags, qtags, rating-tags, meta-tags, affinity_tags, affectivetags, searchandtags, smart_tags, penntags, etags	Tagging, folksonomy, social, pixi, end-user-programming, plurality, flickr, tag, Folksonomies, delicious, collaboration, networks, citeulike, eni, social-software, toread, web, classification, location, recommendation

Paper Number 69: Semantic Wikipedia

Extracted tag keywords for "Wikipedia"	Citeulike keywords for "Wikipedia"	Extracted keywords for "RDF"	Citeulike keywords for "RDF"
Wikipedia, used_for_wikipedia, web-characterization-wikipedia, historywikipedia	Wiki, semantic, quality, collaboration, ontology, visualization, semantic_web, cooperation, paper, social, social-network, trust, web. Tagging, reputation, wikis, collaborative, 2009, community, conflict	Squirrelrdf, computingrdf, translationrdf, ontologyrdf, bio2rdf, rdfa, krdf, analytic_brdf, sw-rdf, rdfs, brdf, -rdf, squirrelrdf-hpl2007-rdf	Semantic, semantic_web, Owl, Web, xml, ontology, p2p, semanticweb, semantic-web, sparql, database, knowledge, ontologies, query, graph, semantics, metadata, kr, rss, iswc

Extracted keywords for "WIKI"	Citeulike keywords for "WIKI"
Ikewiki, aclwiki, kawawiki, acewiki, bowiki, sweetwiki, biowiki, xowiki, pmwiki, annotation-on-wiki, engineswiki, engineeringwiki, sitesxwiki, toolwiki, mapwiki, ots-wiki, traduwiki, wiki, wikis, semanticwiki, mediawiki, semwiki, twiki, semantic_wiki, bizwiki, wikid, geowiki, semperwiki, ow2wiki, ontowiki, sbwiki, creationcustomer-centricityknowledgemanagementopensourcewiki	Collaboration, is366c, Wikipedia, semantic, blog, web20, learning, awareness, community, web, education, social, collaborative, knowledge, online, blogging, internet, socialsoftware, elearning´, koelpu,

Figure 7.1: Tag Recommendation for Research Papers

The second dataset which we prepare for the RDFization process is Expert datasets, accumulated from Expertise Mining system which was proposed in our last study [Afzal et al. 2009]. This expertise mining system was developed over J.UCS dataset to figure out the potential experts from pool of authors and editors. These experts was provided at J.UCS web server, hence is considered as an integral part of J.UCS dataset. This all makes expert dataset as a perfect candidate for RDFization and valuable input for other scientific community. The procedure how this expertise mining was done is discussed in the following section.

7.5 Expertise Calculation

The discovery of expertise is crucial in supporting a number of tasks, particularly for reviewing duites in Digital Journal environment. For supporting this, an automated technique which incorporates multiple facets in providing a more representative assessment of expertise is proposed. The combination of both tangible and intangible metrics to shed deeper insights into the intensity of expertise is presented. The system mines multiple facets for an electronic journal and then calculates expertise weights. The measures provided are, however, not absolute indicators of expertise as the discoveries are limited by the coverage of the database of publications and expert profiles used. In order to enhance the knowledge discoveries, a hyperbolic tree visualization technique was implemented to visualize experts.

7.5.1 Multi-Aspect Expert Profile

In exploring a comprehensive characterization of expertise, A multifaceted approach of mining the expertise for a digital journal was proposed by [Afzal et al. 2008]. The multiple facets are represented by the following measurements: number of publications, number of citations received, extent and proportion of citations within a particular area, expert profile records, and experience. We have thus incorporated the use of user-defined profiles, ”experience atom” (as proposed by [Mockus and Herbsleb 2002] to indicate fundamental experiential units), reference mining results and a characterization of expert participation as facets of an expert profile. Combining all these factors provides a better indication of expertise with regards to a particular topic. There are two main sources of information used to construct an expert profile 1) user inputs and 2) system discoveries.

User inputs are taken from reviewers of the journal J.UCS. J.UCS has over 300 reviewers on its editorial board. The expertise of these reviewers are specified and maintained according to the ACM classification scheme [ACM-CCS 1998]. This information was extracted from J.UCS and used to populate the expert profile database.

The second source for constructing expert profiles is computed by the system. The computation considers the number of publications of an individual, number of citations a person receives, and the person’s duration of publication in the respective area.

7.5.2 Data Extraction

Within J.UCS, ACM topics, editors, and every individual paper are represented in an XML notation, which needs to be parsed to extract metadata. The metadata (paper title, authors, ACM topic, etc.) related to a paper is represented inside the XML file. The extracted data was used to populate a relational database. The database presents a coherent view of all data with relationships (category, paper, authors, and citations). For citation extraction, a technique called Template-based Information Extraction using Rule-based Learning (TIERL) was developed by [Afzal et al. 2009b]. The data from this database was then used to calculate and visualize experts within the J.UCS environment.

7.5.3 Weight Assignment

In this system, experts are grouped into one of two categories: 1) reviewers (persons currently manually assigned as reviewers for a particular ACM topic category) and 2) high-profile authors (persons flagged automatically as experts in a particular topic). Reviewers are selected by the editor-in-chief based on their expertise in the respective ACM topical area. Reviewers for a particular ACM category are visualized without any further calculation. High profile authors are calculated based on weights assigned to them. Three weights called publication weight, citation weight, and reviewer weight are calculated as follows:

Publication Weight = No. of publications / duration (No. of years).
Citation Weight = No. of citations received by an author / total citations in an ACM topic.
Reviewer Weight = No. of reviewers of J.UCS / Total no. of J.UCS Authors.

The total weight is defined as the sum of the three previous weights:

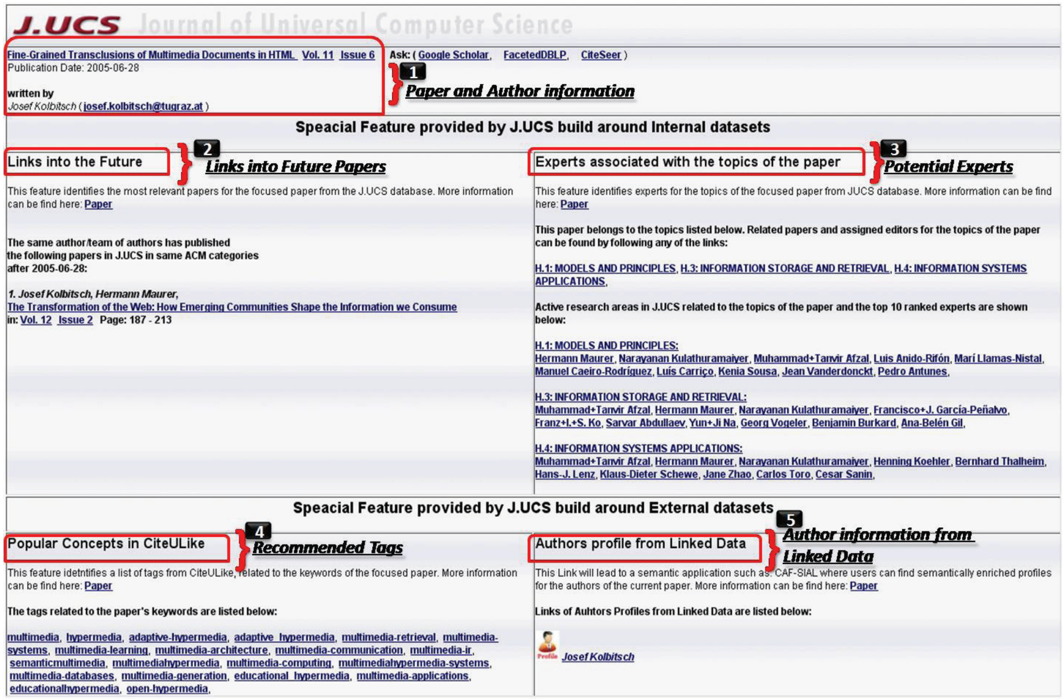
Total weight = publication weight + citation weight + reviewer weight
High-profile authors are then ranked according to their total weight.

After calculation of experts the potential reviewers and editors are mapped to the topical classification of J.UCS and subsequently visualized in the Hyperbolic tree.

Recommended tags , experts, author and paper related information is shown to the journal users at J.UCS web server according to their context. A screen shot of a page containing all this information is shown in figure 7.2. As apparent by the figure, all of the information is presented and displayed in HTML format and no machine understandable data is generated from that. In the next chapter a modeling and RDFization of these dataset is performed and discussed in detail.

7.6 Summary

This chapter presents and retrospect two systems 1. Tag recommendation from Social Bookmarking system and 2. Expert dataset calculated by expertise mining system. In addition to it the preparation of datasets from above two system is revisited to prepare them for the RDFization process. These systems are used in the context of a computer science journal (J.UCS) and is provided at J.UCS webserver in HTML format. At the end of this study, three datasets WWW'06, J.UCS semantically annotated papers and experts are prepared for RDFization. In the context of Linked


J.UCS Journal of Universal Computer Science
1
Paper and Author information
2
Links into Future Papers
3
Potential Experts
4
Recommended Tags
5
Author information from Linked Data
Profile

Fine-Grained Transclusions of Multimedia Documents in HTML Vol. 11 Issue 6
Publication Date: 2005-06-28

written by
Josef Kolbitsch (josef.kolbitsch@tugraz.at)

Ask: (Google Scholar, FacetedDBLP, CiteSeer)

Speacial Feature provided by J.UCS build around Internal datasets

Links into the Future

This feature identifies the most relevant papers for the focused paper from the J.UCS database. More information can be find here: Paper

The same author/team of authors has published the following papers in J.UCS in same ACM categories after 2005-06-28:

1. Josef Kolbitsch, Hermann Maurer,
The Transformation of the Web: How Emerging Communities Shape the Information we Consume
in: Vol. 12 Issue 2 Page: 187 - 213

Experts associated with the topics of the paper

This feature identifies experts for the topics of the focused paper from JUCS database. More information can be find here: Paper

This paper belongs to the topics listed below. Related papers and assigned editors for the topics of the paper can be found by following any of the links:

H.1: MODELS AND PRINCIPLES, H.3: INFORMATION STORAGE AND RETRIEVAL, H.4: INFORMATION SYSTEMS APPLICATIONS,

Active research areas in J.UCS related to the topics of the paper and the top 10 ranked experts are shown below:

H.1: MODELS AND PRINCIPLES:
Hermann Maurer, Narayanan Kulathuramaiyer, Muhammad+Tanvir Afzal, Luis Anido-Rifón, Marí Llamas-Nistal, Manuel Caeiro-Rodríguez, Luís Carriço, Kenia Sousa, Jean Vanderdonckt, Pedro Antunes,

H.3: INFORMATION STORAGE AND RETRIEVAL:
Muhammad+Tanvir Afzal, Hermann Maurer, Narayanan Kulathuramaiyer, Francisco+J. García-Peñalvo, Franz+I.+S. Ko, Sarvar Abdullaev, Yun+Ji Na, Georg Vogeler, Benjamin Burkard, Ana-Belén Gil,

H.4: INFORMATION SYSTEMS APPLICATIONS:
Muhammad+Tanvir Afzal, Hermann Maurer, Narayanan Kulathuramaiyer, Henning Koehler, Bernhard Thalheim, Hans-J. Lenz, Klaus-Dieter Schewe, Jane Zhao, Carlos Toro, Cesar Sanin,

Speacial Feature provided by J.UCS build around External datasets

Popular Concepts in CiteULike

This feature idetntifies a list of tags from CiteULike, related to the keywords of the focused paper. More information can be find here: Paper

The tags related to the paper's keywords are listed below:

multimedia, hypermedia, adaptive-hypermedia, adaptive_hypermedia, multimedia-retrieval, multimedia-systems, multimedia-learning, multimedia-architecture, multimedia-communication, multimedia-ir, semanticmultimedia, multimediahypermedia, multimedia-computing, multimediahypermedia-systems, multimedia-databases, multimedia-generation, educational_hypermedia, multimedia-applications, educationalhypermedia, open-hypermedia,

Authors profile from Linked Data

This Link will lead to a semantic application such as: CAF-SIAL where users can find semantically enriched profiles for the authors of the current paper. More information can be find here: Paper

Links of Auhtors Profiles from Linked Data are listed below:

Josef Kolbitsch

Figure 7.2: J.UCS paper and related dataset HTML representation

Data, these datasets if RDFized and make available in Linked Data Cloud, can offer various benefits for example:

- Open access, sharing these findings with scientific community will help in emergence of new knowledge and studies.
- Similar topical papers from scientific community can link into these datasets to use recommended tags and cluster down the similar papers.
- Interlinking of other datasets will open new knowledge discoveries.
- New applications which can recommend papers on the basis of similar tags are possible too.

Chapter 8

RDFization and Interlinking of Legacy Data

8.1 Introduction

Linked Open Data strives to add structural dimension to data with RDF alliance. Design Issues of Linked Data as discussed earlier; *cf.* Chapter 3 emphasizes on the publishing of data in RDF by giving each data chunk a unique URI, which is further deferenceable to present more meaningful information. The most anticipated potential of exposing data into Linked Data are.

- Every concept has a unique identity (URI) in the document which are further discoverable, reusable and linkable.
- It remove the data silos and minimize the authoritative access to the data making Web as connected Global Graph.
- All concepts are modelled by a single Resource Description Framework RDF, bringing in consistency to structured data representation and lead to interoperability.
- Leverage of asking complex question in the form of SQL type SPARQL queries from different interconnected datasets which will lead to the discovery of hidden patterns and relationships.
- Increase in value and visibility of data by interlinking with external data resources.

A growing number of linked datasets as well as supporting tools and techniques are emerging rapidly. The heavy presence of scientific publications datasets highlights the importance and acceptance of Linked Data in a Scientific Community. To reap benefits offered by Linked Data and to make contribution for the Linked Data community and cloud, a methodology is employed to expose and interlink Digital Journal (Journal of Universal Computer Science) legacy HTML data as RDF.

8.2 Dataset

The Journal of Universal Computer Science (J.UCS) is a high quality electronic publisher that deals with all aspects of Computer Science [Calude et al. 1994] and is thus one of the oldest electronic journals. J.UCS has been appearing monthly since 1995 with uninterrupted publications. From years a lot of interest has been shown by international authors for publications in this Journal. The statistics of hits, page views and download of papers from J.UCS website, further consolidates the fact of Journal as a popular and trusted medium in dissemination of quality scientific information. At present all journal related information is provided in HTML format online. Such information must therefore be exposed in structured format for machine processing and to get benefits from Linked Data.
For simplicity, we have categorized J.UCS dataset into two parts:

8.2.1 Papers

Published papers data is maintained at J.UCS server in Volumes and further classified in issues by published dates. Related information of the papers i.e. paper title, keywords, volume number, issue number, links into future, submission date, acceptance data, published data and categories is maintained in various data tables. Recommended tags associated with papers by author keywords from CiteULike is also maintained separately; *cf.* Chapter 7. In total 1478 papers till volume 15 were considered in this study.

8.2.2 Authors

Information about the paper's authors is maintained separately at J.UCS server. This information mainly consists of the number of papers by an author, name, email address, affiliations and expertise if any; *cf.* Chapter 7 . In total 2597 authors and co-authors related with papers were considered in this study.

8.3 Weaving Approach to Web of Data

For weaving Journal data to Web of Data two approaches 1). modeling and RDFization 2). Interlinking were used.

8.3.1 Modeling and RDFization

First step in the RDFization of J.UCS Data is to look around the ontologies for modeling the characteristics of J.UCS datasets i.e. persons and papers. By sticking to Linked Data Community norms, it was decided to reuse already available ontologies instead of creating new one. We selected *FOAF*[1], *SIOC*[2], *Dublin Core*[3], *SWC*[4], *swrc*[5], *VCard*[6] and *SKOS*[7] ontologies for this process. To model person related properties we used FOAF and SIOC ontologies while Dublin Core, iswc, VCard and SKOS ontologies are used to define publication related properties. In Figure 8.1 , the overview of modeled properties for paper and person is given. Paper and Person classes are connected to each other by paper authorship properties.

The next step in this process is to convert legacy data into RDF. Currently different tools i.e. D2R[Bizer and Cyganiak 2006], triplify [Auer et al. 2009] are available for conversion. D2R server is preferred due to its performance, scalability and SPARQL endpoint feature. D2R server application was installed at local machine and a mapping file on top of J.UCS data tables was created carefully. In writing mapping file a lot of emphasis was put to add rich description for interoperability between ontologies and discovery of data. For example we used various indicators like. swc:Paper, dcmi:Text, sioc:Item, swrc:Article, foaf:Documentto to describe the paper. Similar approach was used to describe the publication of an author. D2R mapping file written for J.UCS dataset conversion is provided in Appendix A.

8.3.2 Interlinking

For the interlinking of J.UCS data with in Linked Data cloud, set of services are developed which search for the relevant information from Linked Data datasets. DBPedia and DBLP in particular for interlinking are focused. DBpedia is considered as heavily interconnected dataset, which makes it a perfect candidate for interlinking with in and with other datasets. DBLP is considered one of the biggest indexed digital library, which presents pointers to the additional papers author may have published else where.

As shown in figure 8.2, a SINDICE API (A semantic Web indexer) is used to query for the intended resource URIs. Next, filtration of the DBLP and DBpedia URI's out of the returned results is performed. Further on a validation service,

[1] http://xmlns.com/foaf/0.1/

[2] http://rdfs.org/sioc/ns

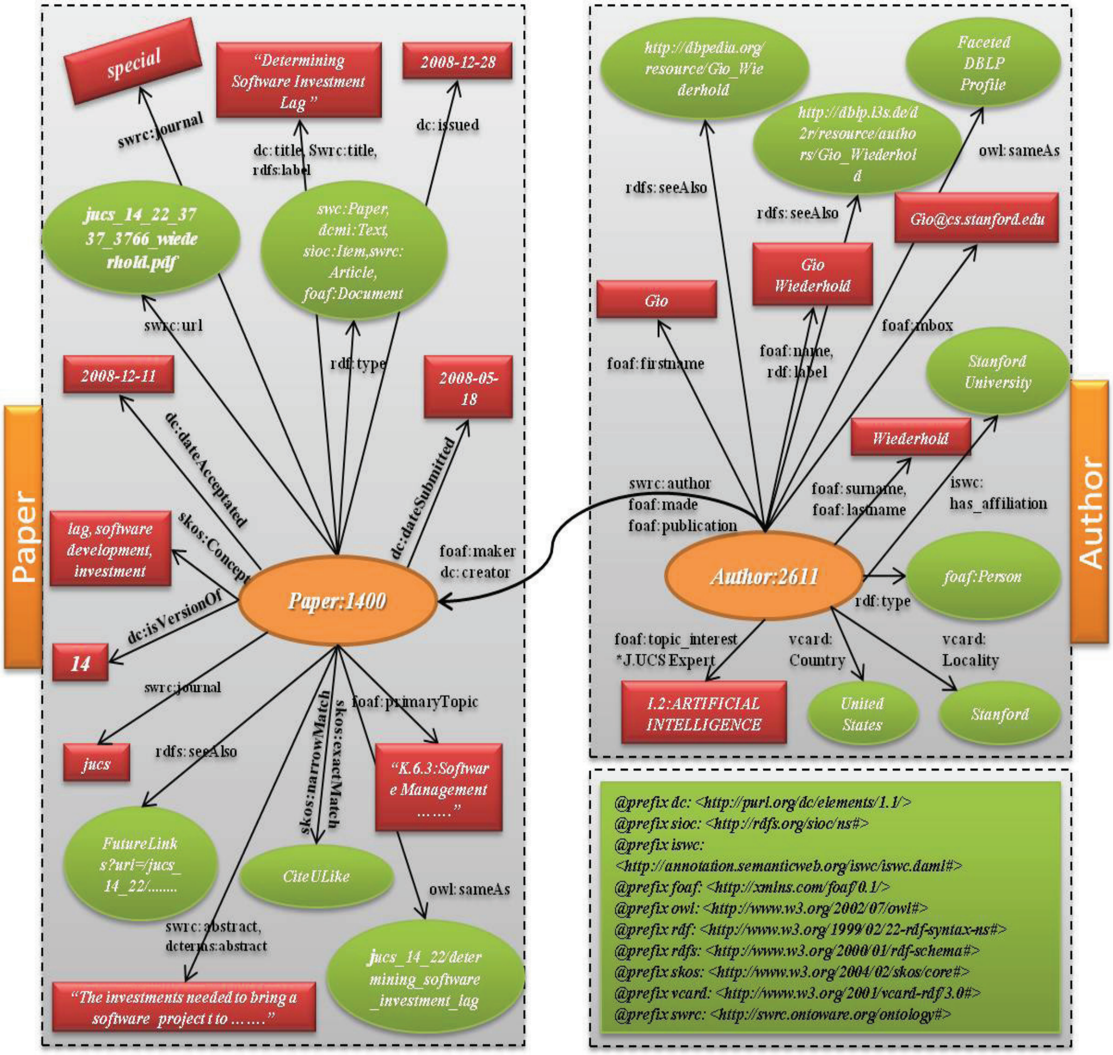
[3] http://purl.org/dc/elements/1.1/

[4] http://data.semanticweb.org/ns/swc/ontology

[5] http://swrc.ontoware.org/ontology

[6] http://www.w3.org/2001/vcard-rdf/3.0

[7] http://www.w3.org/2004/02/skos/core


Paper
special
swrc:journal
"Determining Software Investment Lag"
dc:title, Swrc:title, rdfs:label
2008-12-28
dc:issued
jucs_14_22_3737_3766_wiederhold.pdf
swc:Paper, dcmi:Text, sioc:Item,swrc:Article, foaf:Document
swrc:url
rdf:type
2008-12-11
dc:dateAcceptated
2008-05-18
dc:dateSubmitted
lag, software development, investment
skos:Concept
dc:isVersionOf
14
Paper:1400
foaf:maker
dc:creator
swrc:journal
jucs
rdfs:seeAlso
skos:narrowMatch
skos:exactMatch
foaf:primaryTopic
"K.6.3:Software Management ……."
FutureLinks?url=/jucs_14_22/........
CiteULike
owl:sameAs
swrc:abstract, dcterms:abstract
jucs_14_22/determining_software_investment_lag
"The investments needed to bring a software project to ……."
Author
http://dbpedia.org/resource/Gio_Wiederhold
Faceted DBLP Profile
http://dblp.l3s.de/d2r/resource/authors/Gio_Wiederhold
owl:sameAs
rdfs:seeAlso
rdfs:seeAlso
Gio@cs.stanford.edu
Gio Wiederhold
Gio
foaf:mbox
foaf:firstname
foaf:name, rdf:label
Stanford University
Wiederhold
swrc:author
foaf:made
foaf:publication
foaf:surname, foaf:lastname
iswc:has_affiliation
Author:2611
foaf:Person
rdf:type
foaf:topic_interest *J.UCS Expert
vcard:Country
vcard:Locality
I.2:ARTIFICIAL INTELLIGENCE
United States
Stanford

```
@prefix dc: <http://purl.org/dc/elements/1.1/>
@prefix sioc: <http://rdfs.org/sioc/ns#>
@prefix iswc:
<http://annotation.semanticweb.org/iswc/iswc.daml#>
@prefix foaf: <http://xmlns.com/foaf/0.1/>
@prefix owl: <http://www.w3.org/2002/07/owl#>
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
@prefix skos: <http://www.w3.org/2004/02/skos/core#>
@prefix vcard: <http://www.w3.org/2001/vcard-rdf/3.0#>
@prefix swrc: <http://swrc.ontoware.org/ontology#>
```

Figure 8.1: Modeling of J.UCS datasets

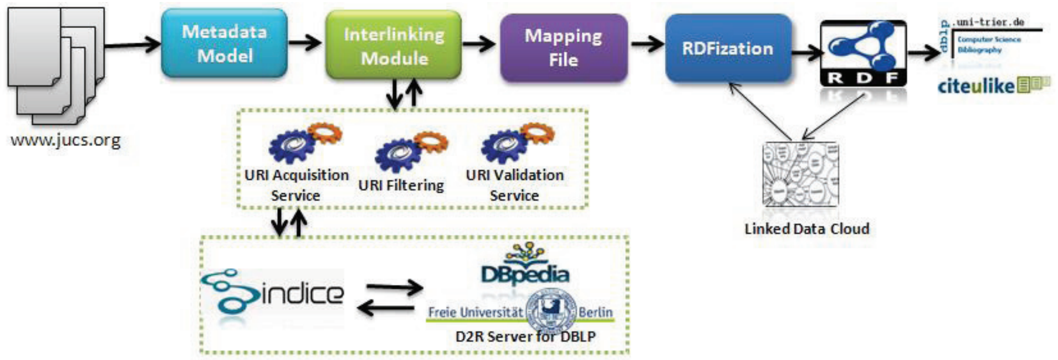


Figure 8.2: Framework for Rdfization of J.UCS Dataset

integral part of interlinking module is called to verify the collected URIs. This Validation service works on the set of heuristics. Based on the output of verification module, discovered URI's were linked with owl:sameAS property. Detail about these processing steps is also discussed in; *cf.* chapter 9 – section 9.4. For bring in the conventional web flare in this RDFization, CiteULike tag and Faceted DBLP links also used for interlinking. Overview of weaving approach in steps is illustrated in Figure. 8.2.

8.3.3 Design of URIs

URIs are required in order to identify entities, types and their relationship. According to Linked Data principles URI's of the resources must be persistent and unique. The URI's used for the J.UCS dataset followed the pattern of:

http://www.jucs.org/d2rq/(resource or page)/(type)/(id or title)

where *(type)* represent the type of resource and *(id or title)* crossponds to the unique identification of resource maintained in the dataset. URI part *(resource or page)* is subjective to content negotiation process i.e in HTML description (if supported by Web-Browser) *page* literal is used while for direct RDF description / dereferencing *resource* literal *resource* is used. Name space of J.UCS website as the part of URI to preserve information about the origin and owner of this datasets is used. For example, the following URI:

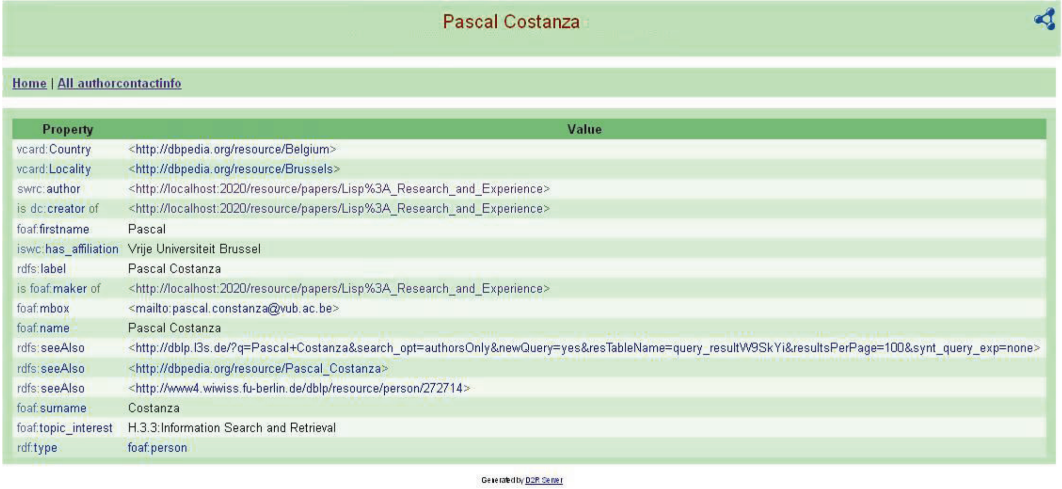
```
http://www.jucs.org/d2rq/resource/authors/Klaus_Tochtermann
```

identifies author *Klaus Tochtermann* while for paper representation following URI is used.

`http://www.jucs.org/d2rq/resource/papers/The_Dortmund_Family_of_Hypermedia_Models_-_Concepts_and_their_Application`

8.4 Data Access

By converting J.UCS HTML data into machine readable (RDF) format, all information is provided by semantically typed relationships. It allows information to be visualized as a single information space like a graph, which can be queried and dereferenced to get additional unapparent information. By following design principles it is tried to provide URIs which are deferenceable so that additional meaningful information is presented. In figure 8.3 and figure 8.4 semantic representation of an author *"Pascal Costanza"* and of a paper titled as *"Software Is More Than Code* after RDFization in local webserver environment is presented respectively . Both of these figure illustrated the semantic properties, which are further dereference-able and are interlinked with other external resources by *sameAs* and *seeAlso* properties.



Pascal Costanza

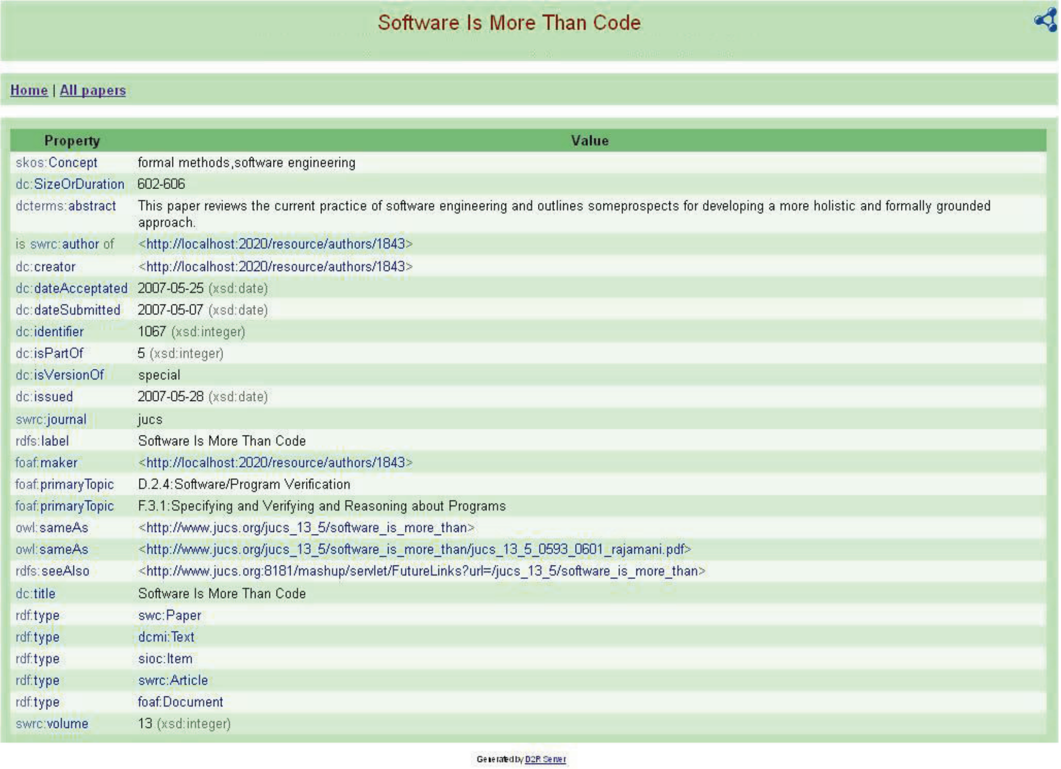
Home | All authorcontactinfo

Property	Value
vcard:Country	<http://dbpedia.org/resource/Belgium>
vcard:Locality	<http://dbpedia.org/resource/Brussels>
swrc:author	<http://localhost:2020/resource/papers/Lisp%3A_Research_and_Experience>
is dc:creator of	<http://localhost:2020/resource/papers/Lisp%3A_Research_and_Experience>
foaf:firstname	Pascal
iswc:has_affiliation	Vrije Universiteit Brussel
rdfs:label	Pascal Costanza
is foaf:maker of	<http://localhost:2020/resource/papers/Lisp%3A_Research_and_Experience>
foaf:mbox	<mailto:pascal.constanza@vub.ac.be>
foaf:name	Pascal Costanza
rdfs:seeAlso	<http://dblp.l3s.de/?q=Pascal+Costanza&search_opt=authorsOnly&newQuery=yes&resTableName=query_resultW9SkYi&resultsPerPage=100&synt_query_exp=none>
rdfs:seeAlso	<http://dbpedia.org/resource/Pascal_Costanza>
rdfs:seeAlso	<http://www4.wiwiss.fu-berlin.de/dblp/resource/person/272714>
foaf:surname	Costanza
foaf:topic_interest	H.3.3:Information Search and Retrieval
rdf:type	foaf:person

Generated by D2R Server

Figure 8.3: Semantic Representation of an Author

A SPARQL endpoint of RDFized J.UCS data is provided for querying. We also have provided the RDF/XML dump of J.UCS dataset for downloading at this link `http://www.jucs.org/download/JUCS_rdfdump`.



Software Is More Than Code

Home | All papers

Property	Value
skos:Concept	formal methods,software engineering
dc:SizeOrDuration	602-606
dcterms:abstract	This paper reviews the current practice of software engineering and outlines someprospects for developing a more holistic and formally grounded approach.
is swrc:author of	<http://localhost:2020/resource/authors/1843>
dc:creator	<http://localhost:2020/resource/authors/1843>
dc:dateAccepted	2007-05-25 (xsd:date)
dc:dateSubmitted	2007-05-07 (xsd:date)
dc:identifier	1067 (xsd:integer)
dc:isPartOf	5 (xsd:integer)
dc:isVersionOf	special
dc:issued	2007-05-28 (xsd:date)
swrc:journal	jucs
rdfs:label	Software Is More Than Code
foaf:maker	<http://localhost:2020/resource/authors/1843>
foaf:primaryTopic	D.2.4:Software/Program Verification
foaf:primaryTopic	F.3.1:Specifying and Verifying and Reasoning about Programs
owl:sameAs	<http://www.jucs.org/jucs_13_5/software_is_more_than>
owl:sameAs	<http://www.jucs.org/jucs_13_5/software_is_more_than/jucs_13_5_0593_0601_rajamani.pdf>
rdfs:seeAlso	<http://www.jucs.org:8181/mashup/servlet/FutureLinks?url=/jucs_13_5/software_is_more_than>
dc:title	Software Is More Than Code
rdf:type	swc:Paper
rdf:type	dcmi:Text
rdf:type	sioc:Item
rdf:type	swrc:Article
rdf:type	foaf:Document
swrc:volume	13 (xsd:integer)

Generated by D2R Server

Figure 8.4: Semantic Representation of a Paper

8.4.1 Query Example

As mentioned earlier the conversion of J.UCS Data into single information space gives us the leverage to ask complex question in the form of queries. For example, we want to find an information about an author who has published a paper in J.UCS Issue after year 2005 and is a ranked expert in the area of ”Information Search and Retrieval” with in J.UCS. As well as author has a DBpedia page. By issuing this query we discovered that ”Pascal Costanza” affiliated with ”Vrije Universiteit Brussel” fulfilled our query criteria. The formalization of this query is given in listing 8.1.

With the RDFization of data, variety of other queries are also possible e.g.

- Finding the name of authors who has publication in J.UCS special issue.
- Finding the paper which belong to Computer Graphics category and having more than one authors.
- Finding papers which have CiteULike Tags links.

```
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix dc: <http://purl.org/dc/elements/1.1/> .
@prefix sioc: <http://rdfs.org/sioc/ns#>.
@prefix iswc: <http://annotation.semanticweb.org/iswc/iswc.daml#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .

SELECT * WHERE
{
?Publication sioc:title ?Title;
dc:issued ?PublicationDate.
?Author foaf:maker ?Publication;
foaf:topic_interest "H.3.3:Information Search and Retrieval";
foaf:name ?Name;
foaf:mbox ?Email;
iswc:has_affiliation ?Affiliation;
rdfs:seeAlso ?ExternalLinks.

FILTER(?PublicationDate > "2005-12-31"^^xsd:date).
FILTER regex(?ExternalLinks, "http://dbpedia.org" )
}
```

Listing 8.1: Sample SPARQL Query on J.UCS Dataset

8.5 Summary

In this chapter an innovative approach is presented to convert J.UCS legacy HTML data into machine readable format. Furthermore the interlinking of authors and papers with the DBpedia, DBLP, Faceted DBLP and CiteULike datasets is presented and elaborated. In the end an example which highlights the benefits and power of structured and Linked Data in making complex queries is narrated. The contribution made by this chapter is following:

- Converted the Legacy HTML Journal Data into RDF and made it available to Linked Data cloud for scientific community to reuse and interlink.
- Each artefact of J.UCS was annotated with the deference-able unqiue URI.
- Interlinked Journal Data with DBpedia, DBLP Linked Data resource and with Faceted DBLP, CiteULike conventional Web resources.
- Provided an HTML interface on top of converted semantic data for navigation between interlinked connected resources.

- A SPARQL endpoint was setup on top of RDFzied J.UCS data which can listen and answer internal and external SPARQL queries with results.

The generated semantic data can be accessed at `http://www.jucs.org/d2rq/`.

Part IV

Application Areas

Linked Data, with its valuable interlinked datasets offer many benefits, but lack of application case studies in exploiting these benefits in different domains is hindering Linked Data potentials. In previous chapters (4 and 5), a framework of URI identification and Concept Aggregation Framework for structuring the related aspects of a person (in to a profile) was discussed. This framework presented the innovative approaches to consume Linked Data which produced an application CAF-SIAL, that is built around DBpeida (Linked Data dataset) as a proof of concept. But for the generalizability of this application, it was required to use it as a third party application in different domains. To achieve this, a framework for application areas of Linked Data is proposed in this thesis. In this framework three application areas of Linked Data (Digital Journal, Expertise Mining System and Academia Europea Members) were identified and connected with Linked Data. The successful linking have shown effectiveness and value in to these fields by automatically identifying, disambiguating, organizing and presenting persons related information from Linked Data. This efficiency is hard to find in conventional web searching.
Identified application areas by the this frameworks are:

1. Digital Open Journal, for getting the authors related information and their contribution in scientific community.
2. Expert mining system, to assist Journal administration to have a leverage of the additional information provided by Concept aggregation framework to assign peer reviewing and editors to a particular paper.
3. Academia Europea Members, for getting their related information from Linked Data.

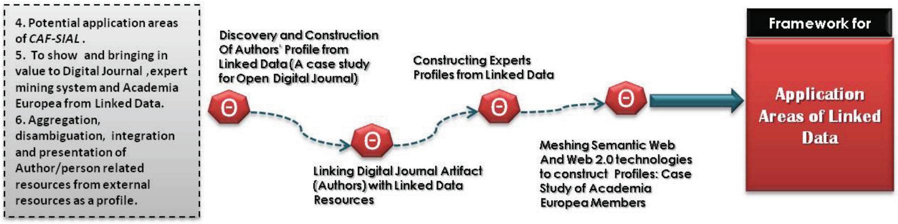

4. Potential application areas of CAF-SIAL .
5. To show and bringing in value to Digital Journal ,expert mining system and Academia Europea from Linked Data.
6. Aggregation, disambiguation, integration and presentation of Author/person related resources from external resources as a profile.
Discovery and Construction Of Authors' Profile from Linked Data (A case study for Open Digital Journal)
Linking Digital Journal Artifact (Authors) with Linked Data Resources
Constructing Experts Profiles from Linked Data
Meshing Semantic Web And Web 2.0 technologies to construct Profiles: Case Study of Academia Europea Members
Framework for
Application Areas of Linked Data

Figure 8.5: Flow Chart of Linked Data Application Areas

The figure 8.3, illustrated the research contributions with published work [Latif et al. 2010a] [Latif et al. 2010b] [Latif and Afzal 2011] [Korica-Pehserl and Latif 2011], shows the progress flow for Applications Areas of Linked Data framework. at the first stage, Authors of digital journal are linked with Linked Data application as Linked Data application area case study; *cf.* Chapter 9. In addition to it, another application on Expertise Mining for Digital Journal administration is linked to push expert's additional information. This additional information helped the administration to assign reviewing duties to different experts; *cf.* Chapter 10. Academia Europea Members dataset was explored as the third application area of Linked Data; *cf.* [Korica-Pehserl and Latif 2011].

Chapter 9

Digital Journals – Discovering and Organizing Authors' Profiles from Linked Data –

Profiling systems tend to find information about persons on some particulars; an established profiling system plays a vital role in the success of multidisciplinary person data management systems e.g. digital open journals. In digital journal's environment, a well-linked collection of electronic resources is of great importance especially in creating opportunities for collaborations between organization, institutions, and persons. Finding information about authors in a digital journals environment is crucial to increase the overall visibility, efficiency and unprecedented success. Traditional profiling system exploit structured data from closed system (e.g. emails and spread sheets) or unstructured data from open system (e.g. the web) gaining very limited and specific results. The emergence of many semantically rich and structured datasets powered by Linked Open Data movement can assist in more controlled search and productive results. Inspired from Linked Open Data initiative, we have developed a tool which can establish links between authors of digital journal with relevant semantic resources available in Linked Open Data. The proposed system is able to disambiguate authors and can: 1) locate, 2) retrieve, and 3) structure the relevant semantic resources. Furthermore, the system constructs comprehensive aspect oriented authors profiles from heterogeneous datasets of Linked Open Data on the fly. This approach was implemented on a running digital journal known as Journal of Universal Computer Science (J.UCS). This study showed a lot of potentials for the profiling systems in Linked Open Data. It is our strong belief that this study can motivate researchers and developers to investigate different application areas where Linked Open Data can contribute, bring added value, and can take the idea of open access further.

9.1 Digital Publishing Systems - Introduction

Recently, for the universal access of published scientific knowledge, the scientific community has consensus on the requirements for the systems that provide access to published research papers, i.e. to a comprehensive collection that can be indexed, searched and linked efficiently [Hitchcock 2002]. The open access movement drives the scientific community to *texdynamic digital archive* [Roberts et al. 2001] which reflects the visionary ideas of Roberts et al.: *"it will enable researchers to take on the challenge of integrating and interconnecting the fantastically rich, but extremely fragmented and chaotic, scientific literature"* [Roberts et al. 2001].

In an Open digital journal environment, a well establish profiling system and a well-linked collection of electronic resources is of great importance in supporting number of task: 1) providing instant access to related resources 2) increasing knowledge visibility 3) supporting forthcoming research which is usually innovated based on exiting knowledge 4) creating opportunities for collaborations between organization, institutions, and persons. Modern digital journals provide such functionalities by intelligent linking of relevant resources with heterogeneous repositories [Afzal, 2009] [Afzal et al. 2009].

Profiling systems aims to find and filter information about a person on the basis of some evidence. The information found by these systems can be used to rank, classify and clustered down experts and similar interest group people. Usually, profiling system operates on the structured data gathered from closed system (e.g. email, parsing spreadsheets and document entries) and unstructured data from open system (e.g. the web) to construct a profile. These profiling systems have important implications in the success of digital journal; and can be enhanced with more intelligent information search and retrieval techniques to bring more variety in results, as shown in this study.

In the context of current research, finding information about authors (author's profiles) in a digital journal's environment is crucial to increase the overall productivity and unprecedented success. The discovery of author's profiles helps in accomplishing the following task: 1) users of digital journals need to search the research collaborators, 2) users need to search experts to seek guidance, and 3) journal administration want to explore new reviewers. All of this is not possible only by looking author's publication list alone, but one also needs other information of the authors. This information may include a short biography, research areas, co-author network, research projects, academic records, achievements, and geographical position of an author. Although much of this information may be acquired from the web by exploring existing tools such as: search engines, citation indexes, and social networks. Finding task oriented information is bit of a challenge, simply due to 1) availability of the huge amounts of unstructured data and 2) wall-gardened data repositories. Thus, users are often frustrated. Hence, there is a need to have a system which can retrieve, aggregate, structure information from diverse sources, and can present a coherent view of authors profiles at one place. Furthermore, this information can be supplied in users local focus and context.

In the past, various techniques/heuristics have been applied to find relevant resources using number of ways: 1) by exploiting metadata of resources [Giles et al. 1998], 2) by performing natural language processing on unstructured data, 3) by training the system on some machine learning algorithms [Kennedy and Shepherd 2005], by computing text similarity, and bibliographic analysis [Giles et al. 1998]. However, linking relevant resources with the help of all available metadata produce satisfactory results but most of the time, the results produced from unstructured data (whatever technique is used) is not up to the mark. In this scenario, The Semantic Web tries to structure data which can be processed intelligently by machines.

One of the most successful projects of the Semantic Web is Linked Open Data (Linked Open Data) [Bizer et al. 2009] which provides semantically rich and interlinked datasets. These datasets can be explored for knowledge discovery and creating cross-references between relevant resources. The ever growing Linked Open Data cloud in its own narrates the success story of this movement. In approximation, there are about 13.1 billion RDF triples (November 2009) which came cross from different practical, social, business and research domains. At present research community is changing its gear from techniques of opening authoritative data repositories to build applications which can make use/reuse of currently semantified datasets. The utmost goal is to make realization of the power of Linked Open Data in business, public and government as well as making the web cleaner and connected. The semantified data about persons present in different dataset of Linked Open Data e.g. DBpedia [Auer et al. 2007], FOAF [Brickley and Miller 2004], SIOC [Breslin et al. 2005], DBLP [1] opens new horizon to built applications that can discover and link persons information on the fly.

Motivated from the success of Linked Open Data and new publishing paradigm [Berners-Lee and Hendler 2001], we have established links for authors of a digital journal with the semantically rich data sets of Linked Open Data. The process of creating cross-referencing consist of following sub-process 1) disambiguation of indi-

[1] `http://dblp.l3s.de/d2r/`

viduals, 2) retrieval of information resources 3) structuring the retrieved information and 4) presentation of structured resources. Although the data sets represented by Linked Open Data are semantically rich, but there are certain problems e.g. locating exact resource URI, lack of quality interlinking [Jaffri et al. 2008], heterogeneity in ontologies, and the lack of user interfaces for Linked Open Data consumption in scientific communities as well as in business enterprises [Heath, 2008][Latif et al. 2009]. To locate intended resources URI, author of this paper has already presented an intelligent Keyword-URI mapping technique [Latif et al. 2009a]. To retrieve and structure information resources, an innovative Concept Aggregation Framework [Latif et al. 2009b] was employed. To present a coherent view of information, a user interface named as CAF-SIAL was developed[2].

Going further from this application, we have applied our past research experiences to establish a link between authors of a digital journal with their DBpedia mined profiles [Latif et al. 2010b] and their contributions (published work) crafted from Linked Open Data cloud. We are confident that these types of applications can help in interconnecting the conventional web applications (e.g. digital journals) with third generation semantic web applications. This will bring added value for a digital journal from semantic domain and will increase visibility of these applications.

9.2 Concept Aggregation Framework (Revised)

As explained in previous chapter, a concept aggregation framework [Latif et al. 2010a] was introduced for structuring informational aspects of a resource type *person*. This framework was able to organize various semantic resources of Linked Data into broad logical informational aspects i.e. personal, professional, social, and dark side, representing an overall picture of a person. The Linked Data resources allowed us to explore semantic representations of resource properties and relationships to come up with such a framework.

The implemented algorithm was able to link a semantic resource/property/relationship with the most suitable informational aspect. To translate resources from Linked Open Data into informational aspects, a layered approach was used as explained in next paragraph and can be seen in figure 9.1. The framework has been applied in CAF-SIAL system which is up and running since 2009. The proposed system has also bridged a gap between semantic search and end users by hiding complex mechanics of the Semantic Web. The evaluations of the system showed a certain wavier of effort from users in searching and presentation of information [Latif et al. 2010a].

[2] http://cafsial.opendatahub.org/

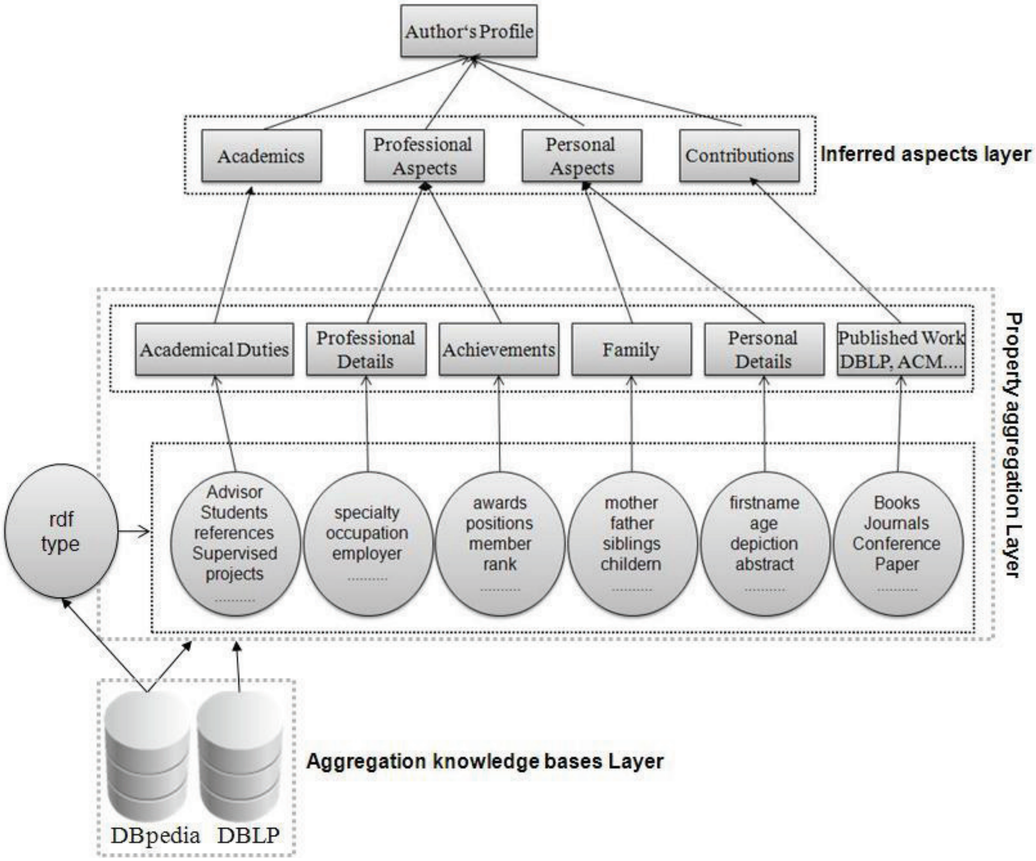

Author's Profile
Academics
Professional Aspects
Personal Aspects
Contributions
Inferred aspects layer
Academical Duties
Professional Details
Achievements
Family
Personal Details
Published Work DBLP, ACM....
Property aggregation Layer
rdf type
Advisor Students references Supervised projects ....
specialty occupation employer ....
awards positions member rank ....
mother father siblings childern ....
firstname age depiction abstract ....
Books Journals Conference Paper ....
Aggregation knowledge bases Layer
DBpedia
DBLP

Figure 9.1: Concept Aggregation Framework (digital publishing domain)

This paper is continuity to our previous work by exploring a digital journal as an application area. Authors of a digital journal have been linked with Linked Open Data resources using concept aggregation framework. In a scientific domain, the authors can be represented in four informational aspects like: 1) personal, 2) professional, 3) academics, and (4) their contributions (Published work). The figure 7.1 depicts an overall working of the framework. The framework is represented in three layers: 1) aggregation knowledge bases layer, 2) property aggregation layer, and 3) inferred aspect layer.

In the aggregation knowledge bases layer, all properties of a person (author) are extracted and aggregated from DBpedia. In property aggregation layer, retrieved properties are processed using a semi-automated technique. The relevant properties are filtered and linked with sub-informational aspects of an author as shown in figure 9.1. In previous implementation, a set of experts manually tagged and assigned relevant properties to sub-informational aspects. However, in the current application, a semi-automated technique based on set of developed heuristics worked fine. In the inferred aspect layer, the sub-informational aspects are further linked to main informational aspects to get a comprehensive view of authors. More information about the concept aggregation framework can be found in [Latif et al. 2009a] [Latif et al. 2009b] [Latif et al. 2010a].

9.3 Datasets

9.3.1 Journal of Universal Computer Science

The Journal of Universal Computer Science (J.UCS) is a high-quality electronic publisher that deals with all aspects of Computer Science [Calude et al. 1994]. J.UCS has been appearing monthly since 1995 with uninterrupted publications. According to the survey paper on electronic journals [Liew and Foo 2001], J.UCS has incorporated innovative features such as the enabling of semantic and extended search and its annotative and collaborative features. It was one of the first electronic published journals to have implemented features such as personal and public-annotations, multi-format publications, multi categorization, etc. These features have made J.UCS a rather unique electronic journal. Readers of such high-quality electronic journals expect and anticipate highly sophisticated features, such as automatic reference analysis, similarity search between documents and other features using knowledge management technology [Krottmaier 2003].

The J.UCS dataset provides the list of the authors who have published their work in any of the journal issues. The author ID maintained at J.UCS server along with first, middle and last name is tabulated in this dataset. In total 2593 authors from

J.UCS were used for this experiment.

9.3.2 DBpedia

DBpedia, a semantic flip of Wikipedia is one of the biggest examples of Social Semantic Web. DBpedia is considered one of the most promising knowledge bases, having a complete ontology along with Yago classification [Suchanek et al. 2007]. It currently describes more than 3.4 million things, including at least 312,000 persons, 413,000 places etc [Auer et al. 2007]. The knowledge base consists of 479 million pieces of information (RDF triples). The openly available RDF dumps make DBpedia an interesting subject of study. There has been valuable work done on studying the reliability of Wikipedia URIs [Hepp et al. 2007] that is a backbone of the DBpedia. This study suggests that the meaning of a URI stays stable approximately 93% of the time. Its heavy interlinking within the Linked Open Data cloud makes it a perfect resource to search URIs. For current experiments, we concentrated on the part of DBpedia that encompasses data about persons.

Two RDF dumps about personal information (Persondata and Links to DBLP) were selected to find relevant information of the J.UCS authors. These datasets are freely available in RDF dumps[3] .

Persondata

This dataset includes the information about persons extracted from the English and German Wikipedia, represented using the FOAF vocabulary.

Links to DBLP

Links between computer scientists in DBpedia and their contributions in the DBLP database are enlisted in a same:as relationship in this dataset. To follow the DBLP links, The D2R Server, a semantified version of DBLP bibliography was accessed from Berlin and Hanover SPARQL endpoints[4]. This D2R Server is based on the XML dump of the DBLP database. The DBLP database provides bibliographic information on major computer science journals and conference proceedings. The database contains more than 800.000 articles and 400.000 authors [Michael 2009].

[3] http://wiki.dbpedia.org/Downloads34

[4] http://dblp.l3s.de/d2r/snorql/

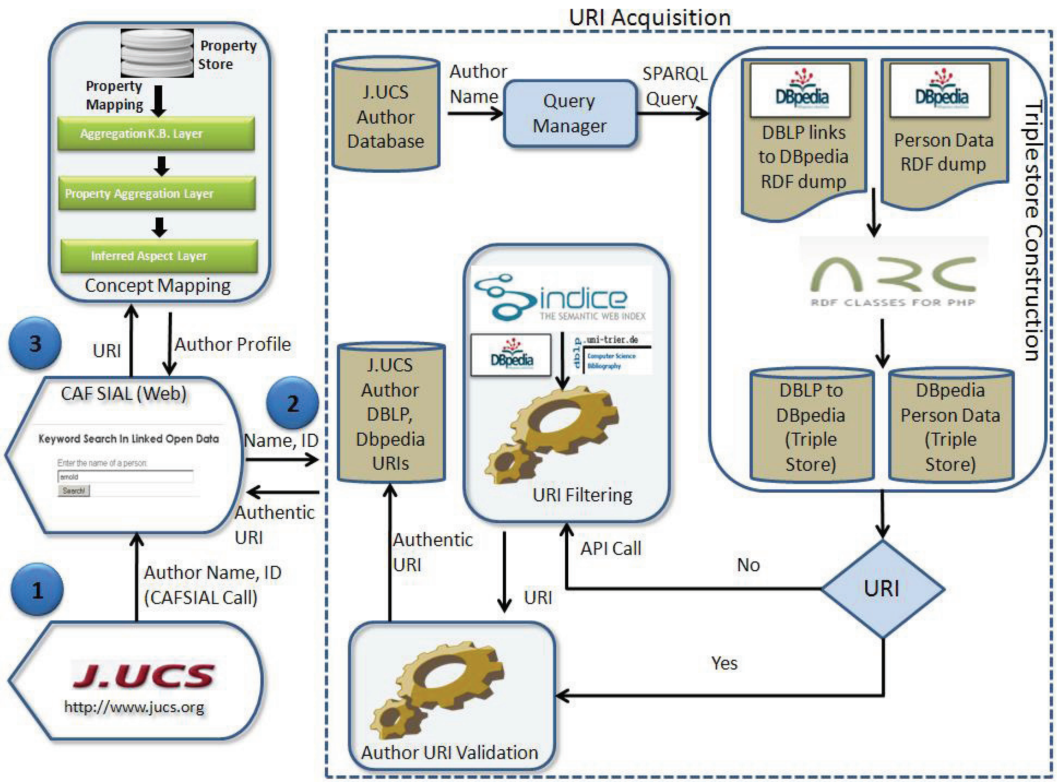


Figure 9.2: System Design

9.4 System Design

The architecture design of the application is depicted in figure 9.2. The proposed system is divided into four modules named such as: 1) Database and Triple Store Construction, 2) URI Acquisition, 3) Author URI Validation, and 4) Concept Mapping.

The database and triple store construction part discusses the data acquisition, and manipulation of J.UCS data set along with the conversion process of RDF person dataset into local triple store. The URI acquisition module describes how the URI of a J.UCS author is acquired from local triple stores and from remote semantic search services. Author's URI validation module encompasses the developed heuristic to validate the URI. Last module discusses the concept aggregation and presentation of the results.

9.4.1 Database & Triple Store Construction

For this application three datasets one from J.UCS and other two from DBpedia were used. Along with these datasets, web service of Sindice [Oren et al. 2008] was utilized when search in local triple store fails. The downloaded RDF dumps were converted into a local triple store by using ARC2[5]. ARC2 star triple store configuration gives a facility for querying at statement level. In DBpedia person triple store total 29,498 URIs are stored. The DBLP to DBpedia dataset contains information of 196 computer scientist.

9.4.2 Uri Acquisition

A layered approach was employed to acquire the intended URI from local triple store and from Linked Open Data Cloud [Latif et al. 2009a]. In total four processes were involved in this approach each process resulting as an input for the subsequent layer ending up with the retrieval of desired resource. Following are the names and details of the processes.

1. J.UCS dataset Pre-Processing
2. Direct matching of J.UCS authors with DBpedia Persondata dataset
3. Direct matching of J.UCS authors with Links to DBLP dataset
4. Querying and Filtering of URI from Sindice

9.4.3 J.UCS Dataset Pre-Processing

Sometimes, the authors' names contain *umlaute* character and anomalies (irregularity in name) which need to be processed before matching. An automated script was written to remove such inconsistencies. Subsequently first, middle, and last names were concatenated to construct a full name for the matching process.

DBpedia Person Data Direct Matching With J.UCS Author Dataset

A complete J.UCS author name, acquired from the previous step, was considered for matching in the *DBpedia Persondata* triple store. This operation yielded in very low success rate.

[5] `http://arc.semsol.org/`

Process Name	**Considered Authors**	**DBpedia URI**	**DBLP URI**
DBpediaPerson direct match	2593	7	NA
DBpediatoDBLP direct match	2586	8	8
Sindice(Semantic Search Service)	2578	322	2285
Total Computed Results	2593	337	2293
Uri-Authentication	337	66	2293

Table 9.1: J.UCS Authors URI processing

DBpedia Links To DBLP Direct Matching

In this process, authors were matched with DBLP local triple store. The result of this matching was also not satisfactory.

To increase the discovery of URI, different online semantic Search engines were analyzed like Falcon [Cheng et al. 2008], Swoogle [Ding et al. 2004] and Sindice. Based on the up to date-ness, large indexing corpus, and easy API access, the Sindice was selected for further matching of URIs.

Sindice Search Service

A web service was written to call the API of Sindice with the formulated query. This API call was executed iteratively for every unfound J.UCS author name. In response, Sindice provided the list of the URIs which were further filtered out on the basis of DBpedia provenance. In the end direct matching of author full name with the DBpedia filtered out URIs was performed to select the exact URI.

After this processing, 337 DBpedia URIs out of entire J.UCS author list were found giving substantial improvement to the results. The details about the author found in each step are mentioned in the Table 9.1. These results were stored in a data table for further processing.

Uri Validation

Our past research of CAF-SIAL system helped us in developing set of heuristics to validate the acquired resource URI. Validation and disambiguation of URI is an important part of this application. By manual de-referencing and inspection of the acquired URI's, we discovered some inconsistencies:

1. URI of the respective author exists (wrongly indexed by Sindice) but with no information making it useless.

2. Many ambiguous URI's which matched with exact name of the intended J.UCS author leading to wrong person.

To disambiguate authors, a set of heuristics were written as described below:

After inspection it was noted that there are certain kind of properties for a person type which can be exploit to disambiguate individuals. These properties are dbpedia:Abstract / dbpedia:Comment and SKOS categories. For example, SKOS categories and keywords, being used to represent the persons belonging to education profession are: *"computer science, computer scientist, professor, informatics, researcher ...."* etc. All of these constructing properties represent a person belonging to scientific community. Thus the persons having same names and belonging to different professions can easily be filtered out.

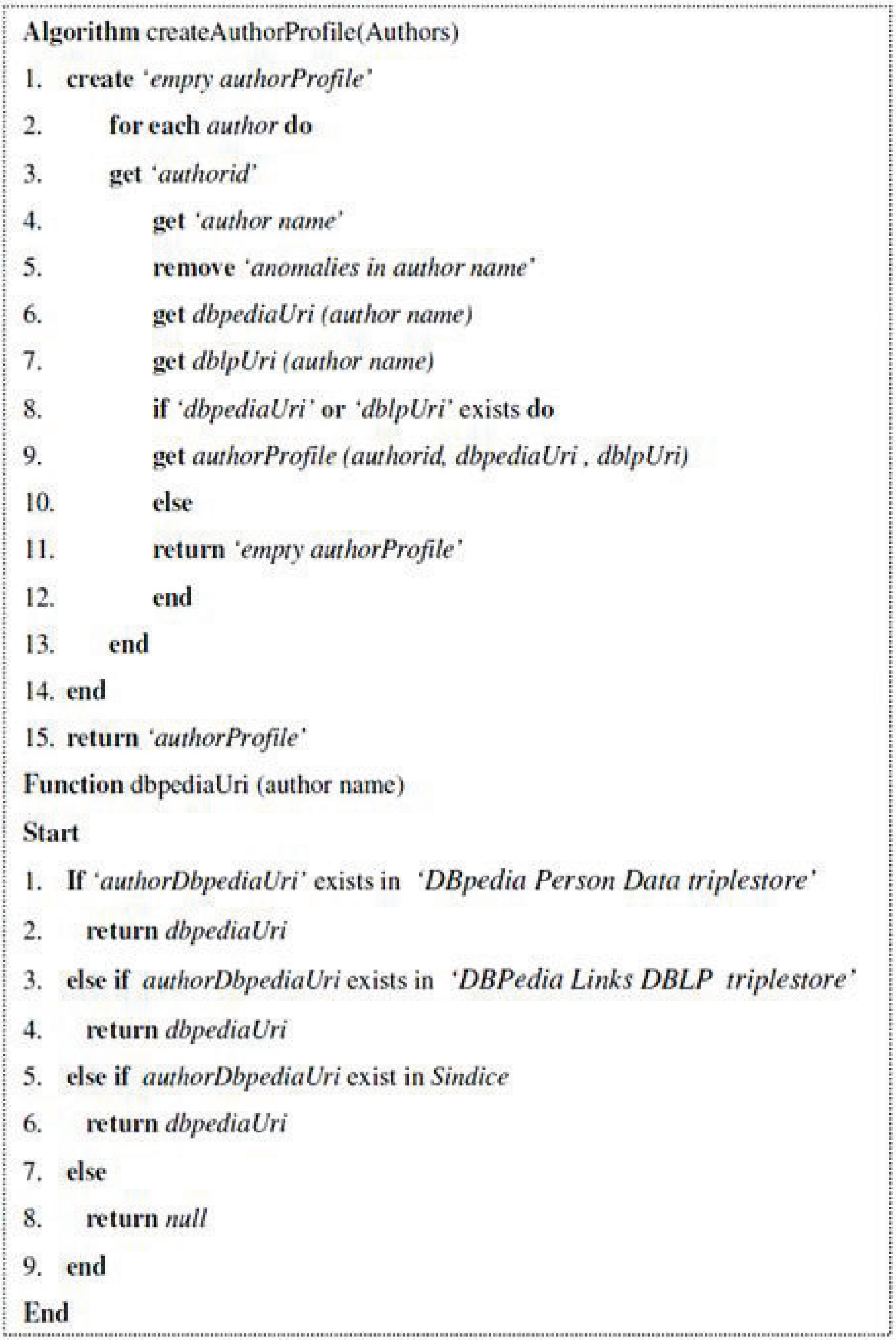
An automated script was written to check the keywords in the abstract property and SKOS categories of the URI. After applying this script on 337 authors, 66 URIs were left. The remaining URI's were either bad links (showing no data) or representing non-scientific persons. In our future implementations, we will investigate disambiguation of authors having same full name and belonging to scientific community.

9.4.4 Concept Mapping

The concept mapping module is responsible for mapping the retrieved set of properties onto concept aggregation framework. The concept mapping module receives an authenticated URI from URI validation module. This URI is used to retrieve available set of properties from Linked Open Data. These properties are passed to concept aggregation framework which organizes these properties in informational aspects of an author. The concept mapping module returns a comprehensive author profile represented in logically organized informational aspects. These aspects are further visualized to users.

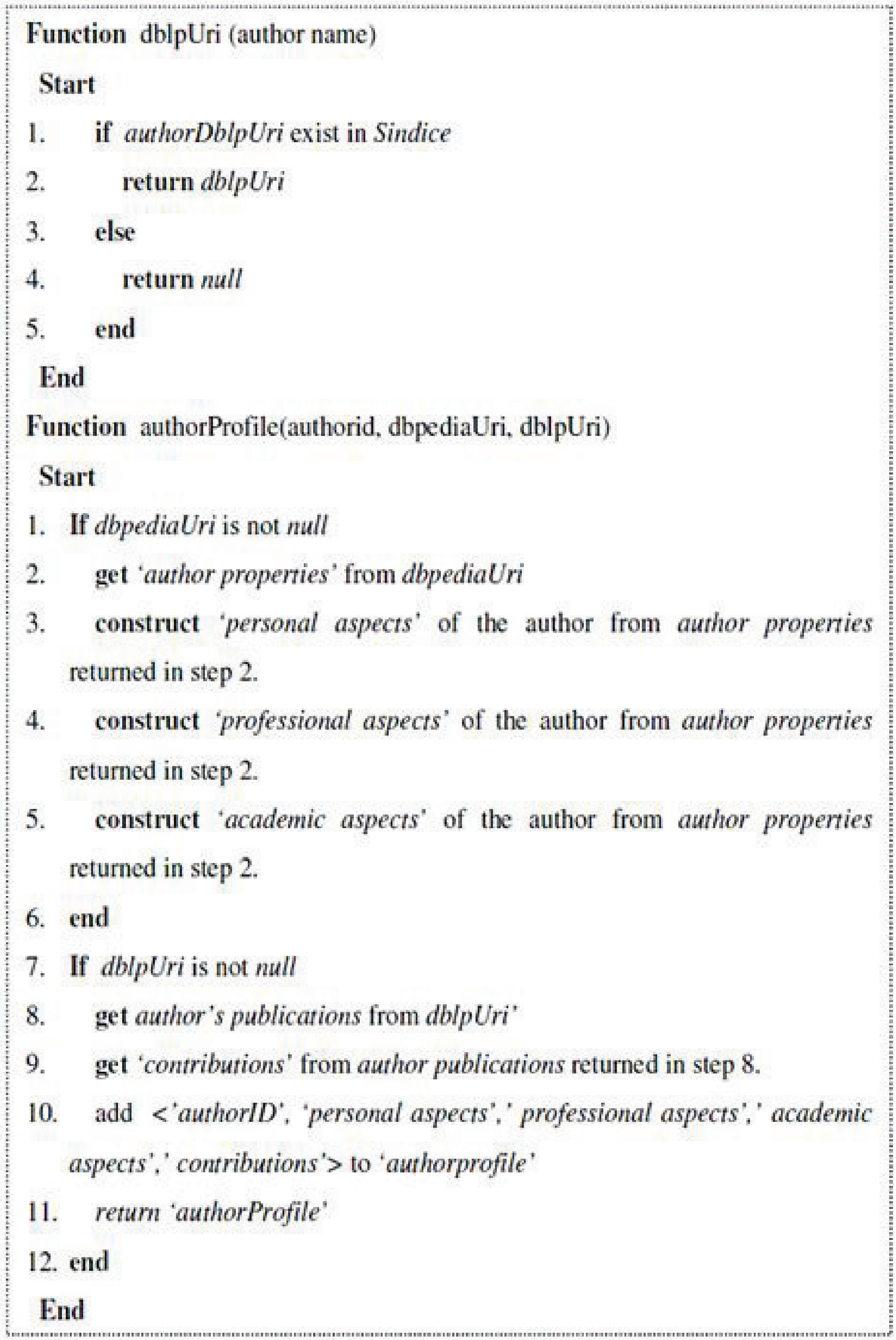
9.5 Algorithm

The algorithm for constructing author profile is shown in figure 9.3 and 9.4 . This algorithm takes authors, author id, author name, dbpediaUri and dblpUri as an input for processing and ends with a return of constructed author profile.



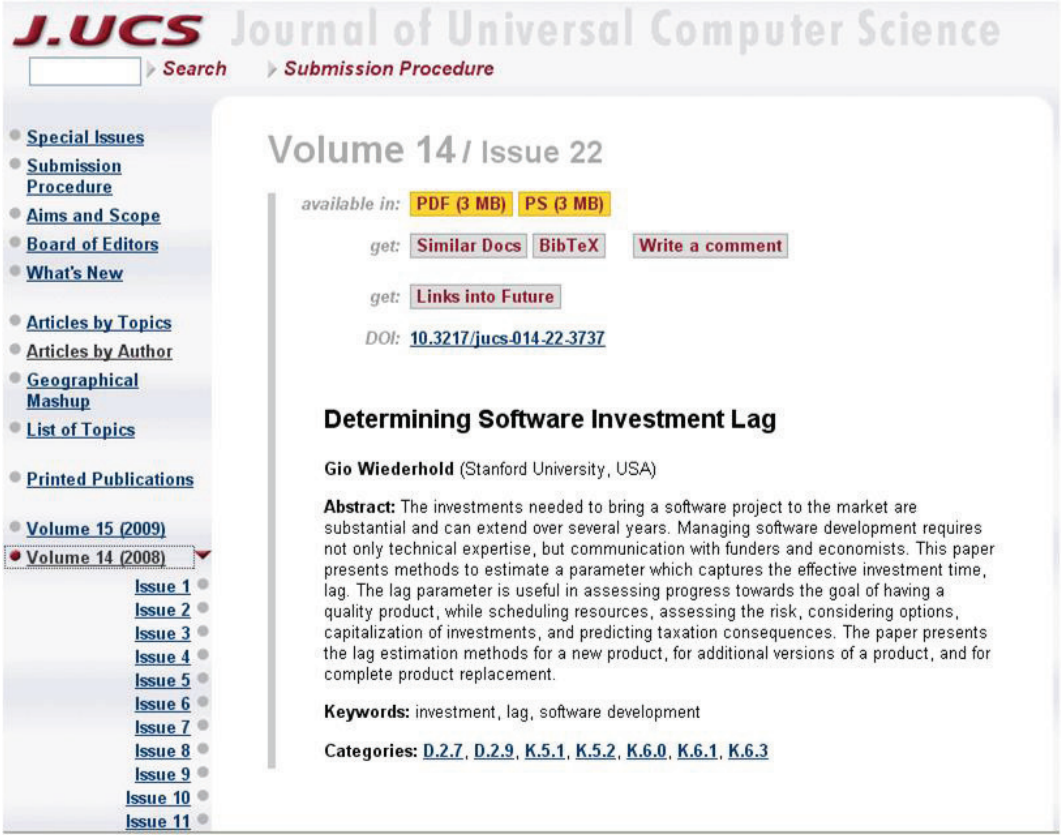
```
Algorithm createAuthorProfile(Authors)
create 'empty authorProfile'
    for each author do
    get 'authorid'
        get 'author name'
        remove 'anomalies in author name'
        get dbpediaUri (author name)
        get dblpUri (author name)
        if 'dbpediaUri' or 'dblpUri' exists do
        get authorProfile (authorid, dbpediaUri , dblpUri)
        else
        return 'empty authorProfile'
        end
    end
end
return 'authorProfile'
Function dbpediaUri (author name)
Start
If 'authorDbpediaUri' exists in 'DBpedia Person Data triplestore'
  return dbpediaUri
else if authorDbpediaUri exists in 'DBPedia Links DBLP triplestore'
  return dbpediaUri
else if authorDbpediaUri exist in Sindice
  return dbpediaUri
else
  return null
end
End
```

Figure 9.3: Algorithm to construct an author profile



```
Function dblpUri (author name)
 Start
1.     if authorDblpUri exist in Sindice
2.        return dblpUri
3.     else
4.        return null
5.     end
 End
Function authorProfile(authorid, dbpediaUri, dblpUri)
 Start
1.  If dbpediaUri is not null
2.     get 'author properties' from dbpediaUri
3.     construct 'personal aspects' of the author from author properties
    returned in step 2.
4.     construct 'professional aspects' of the author from author properties
    returned in step 2.
5.     construct 'academic aspects' of the author from author properties
    returned in step 2.
6.  end
7.  If dblpUri is not null
8.     get author's publications from dblpUri'
9.     get 'contributions' from author publications returned in step 8.
10.    add <'authorID', 'personal aspects',' professional aspects',' academic
    aspects',' contributions'> to 'authorprofile'
11.    return 'authorProfile'
12. end
 End
```

Figure 9.4: Algorithm to construct an author profile(cont.)



Volume 14 / Issue 22

DOI: 10.3217/jucs-014-22-3737

Determining Software Investment Lag

Gio Wiederhold (Stanford University, USA)

Abstract: The investments needed to bring a software project to the market are substantial and can extend over several years. Managing software development requires not only technical expertise, but communication with funders and economists. This paper presents methods to estimate a parameter which captures the effective investment time, lag. The lag parameter is useful in assessing progress towards the goal of having a quality product, while scheduling resources, assessing the risk, considering options, capitalization of investments, and predicting taxation consequences. The paper presents the lag estimation methods for a new product, for additional versions of a product, and for complete product replacement.

Keywords: investment, lag, software development

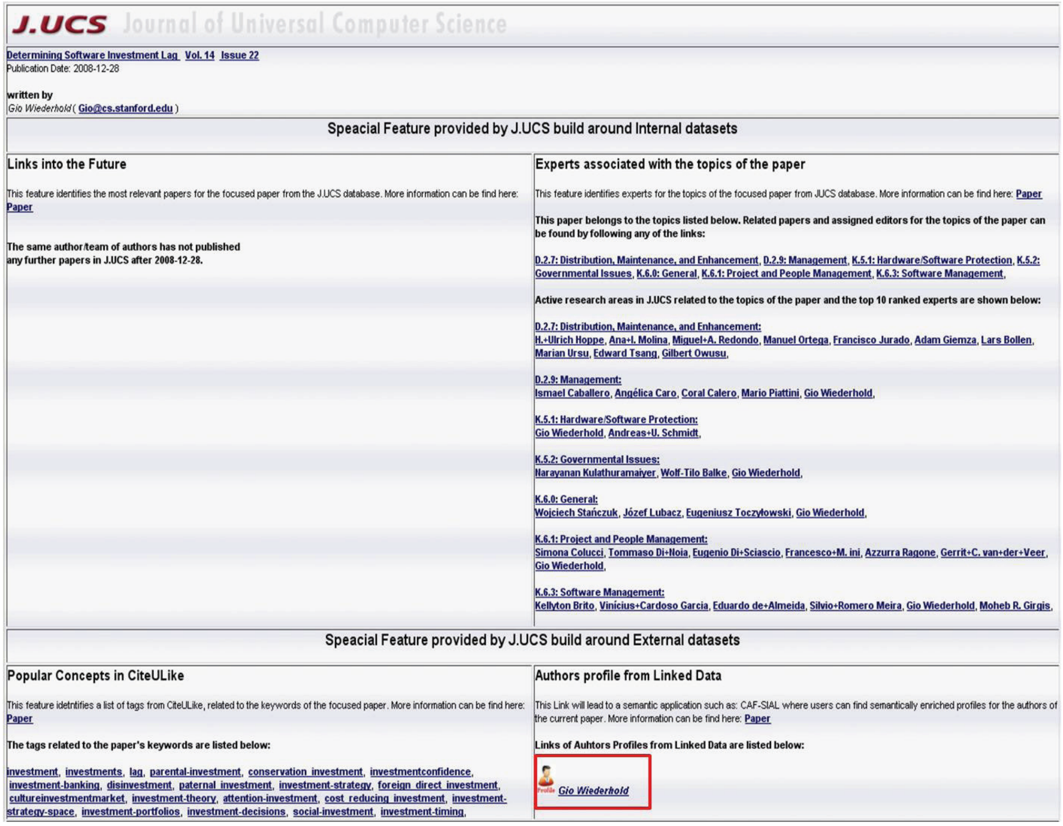
Categories: D.2.7, D.2.9, K.5.1, K.5.2, K.6.0, K.6.1, K.6.3

Figure 9.5: J.UCS interface for viewing a paper

9.6 Visualization

In Journal of Universal Computer Science management system, papers are classified into the volumes and further in to the issues according to their published date. Users can browse any article from the entire body of papers. For example, in figure 3, a user is viewing a paper titled *'Determining Software Investment Lag'* published in the *Vol. 14 / Issue 22*. When the user clicks on the *'Links into Future'* button, he is directed to java servlet screen as shown in figure 9.5. The servlet receives a reference of the viewing paper as a parameter. It subsequently fetches data from different database tables to visualize the *'Author Metadata'*, *'Links into the Future'*, *'Experts'*, *'Popular Concepts (Tags)'* and *'Author Profile Link'*.

On a successful query the author name and author id is hyperlinked with the URL of CAF-SIAL application as a query string. For the convenience of the users the icon representing profile sign is visualized along with the hyperlinked author name marked as rectangle in figure 9.6. When user clicks on the on *Gio Widerhold*



J.UCS Journal of Universal Computer Science

Determining Software Investment Lag Vol. 14 Issue 22
Publication Date: 2008-12-28

written by
Gio Wiederhold (Gio@cs.stanford.edu)

Speacial Feature provided by J.UCS build around Internal datasets

Links into the Future

This feature identifies the most relevant papers for the focused paper from the J.UCS database. More information can be find here: Paper

The same author/team of authors has not published any further papers in J.UCS after 2008-12-28.

Experts associated with the topics of the paper

This feature identifies experts for the topics of the focused paper from JUCS database. More information can be find here: Paper

This paper belongs to the topics listed below. Related papers and assigned editors for the topics of the paper can be found by following any of the links:

D.2.7: Distribution, Maintenance, and Enhancement, D.2.9: Management, K.5.1: Hardware/Software Protection, K.5.2: Governmental Issues, K.6.0: General, K.6.1: Project and People Management, K.6.3: Software Management,

Active research areas in J.UCS related to the topics of the paper and the top 10 ranked experts are shown below:

D.2.7: Distribution, Maintenance, and Enhancement:
H.+Ulrich Hoppe, Ana+I. Molina, Miguel+A. Redondo, Manuel Ortega, Francisco Jurado, Adam Giemza, Lars Bollen, Marian Ursu, Edward Tsang, Gilbert Owusu,

D.2.9: Management:
Ismael Caballero, Angélica Caro, Coral Calero, Mario Piattini, Gio Wiederhold,

K.5.1: Hardware/Software Protection:
Gio Wiederhold, Andreas+U. Schmidt,

K.5.2: Governmental Issues:
Narayanan Kulathuramaiyer, Wolf-Tilo Balke, Gio Wiederhold,

K.6.0: General:
Wojciech Stańczuk, Józef Lubacz, Eugeniusz Toczyłowski, Gio Wiederhold,

K.6.1: Project and People Management:
Simona Colucci, Tommaso Di+Noia, Eugenio Di+Sciascio, Francesco+M. ini, Azzurra Ragone, Gerrit+C. van+der+Veer, Gio Wiederhold,

K.6.3: Software Management:
Kellyton Brito, Vinícius+Cardoso Garcia, Eduardo de+Almeida, Silvio+Romero Meira, Gio Wiederhold, Moheb R. Girgis,

Speacial Feature provided by J.UCS build around External datasets

Popular Concepts in CiteULike

This feature idetntifies a list of tags from CiteULike, related to the keywords of the focused paper. More information can be find here: Paper

The tags related to the paper's keywords are listed below:

investment, investments, lag, parental-investment, conservation investment, investmentconfidence, investment-banking, disinvestment, paternal investment, investment-strategy, foreign direct investment, cultureinvestmentmarket, investment-theory, attention-investment, cost reducing investment, investment-strategy-space, investment-portfolios, investment-decisions, social-investment, investment-timing,

Authors profile from Linked Data

This Link will lead to a semantic application such as: CAF-SIAL where users can find semantically enriched profiles for the authors of the current paper. More information can be find here: Paper

Links of Auhtors Profiles from Linked Data are listed below:

Profile *Gio Wiederhold*

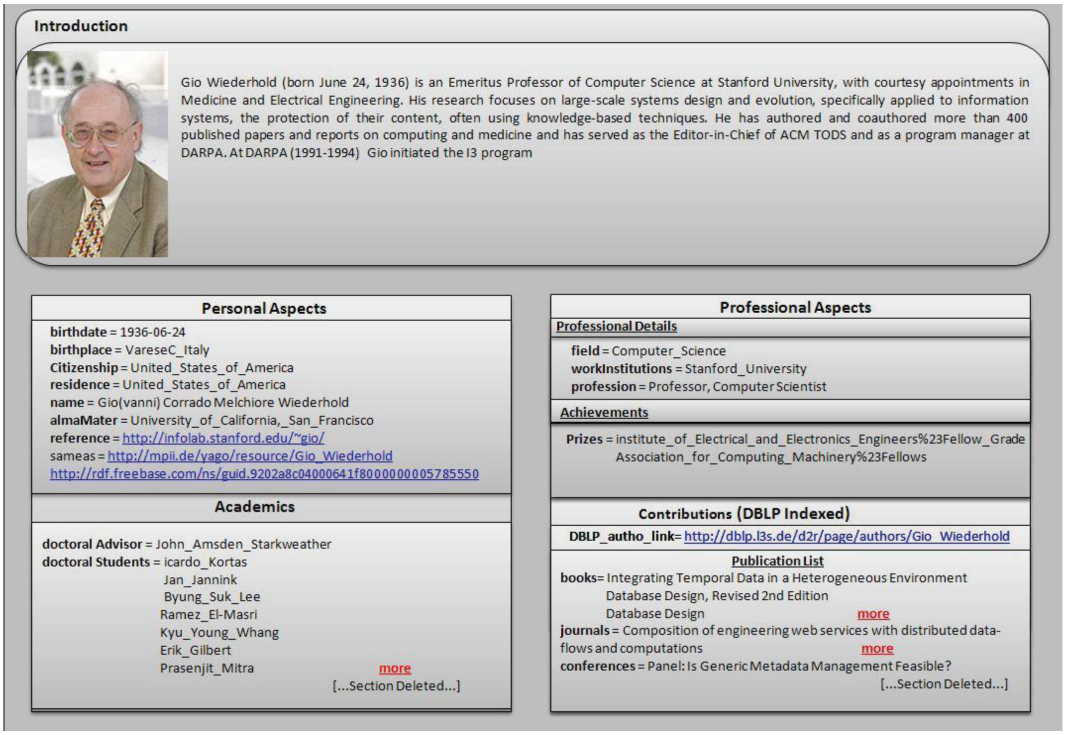
Figure 9.6: Links into Future with Authors' Profile notation

author; it leads user to CAF-SIAL application where concept aggregation framework makes a visualization of the author profile as shown in figure 9.7.

From the figure 9.7, it is obvious that a user will get instant information about the person. A brief introduction of a person along with the picture is shown. Furthermore different aspects like personal, professional, academics, and contributions are shown to the user. From this coherent view, user gets a first overall impression about a person and can follow any hyperlink to see further details.

9.7 Limitations

This application is dependent on readily availability of data from DBpedia and DBLP. The down time or maintenance issue is considered as limitations of this application.



Introduction

Gio Wiederhold (born June 24, 1936) is an Emeritus Professor of Computer Science at Stanford University, with courtesy appointments in Medicine and Electrical Engineering. His research focuses on large-scale systems design and evolution, specifically applied to information systems, the protection of their content, often using knowledge-based techniques. He has authored and coauthored more than 400 published papers and reports on computing and medicine and has served as the Editor-in-Chief of ACM TODS and as a program manager at DARPA. At DARPA (1991-1994) Gio initiated the I3 program

Personal Aspects

birthdate = 1936-06-24
birthplace = VareseC_Italy
Citizenship = United_States_of_America
residence = United_States_of_America
name = Gio(vanni) Corrado Melchiore Wiederhold
almaMater = University_of_California,_San_Francisco
reference = http://infolab.stanford.edu/~gio/
sameas = http://mpii.de/yago/resource/Gio_Wiederhold
http://rdf.freebase.com/ns/guid.9202a8c04000641f8000000005785550

Academics

doctoral Advisor = John_Amsden_Starkweather
doctoral Students = icardo_Kortas
Jan_Jannink
Byung_Suk_Lee
Ramez_El-Masri
Kyu_Young_Whang
Erik_Gilbert
Prasenjit_Mitra more
[...Section Deleted...]

Professional Aspects

Professional Details

field = Computer_Science
workInstitutions = Stanford_University
profession = Professor, Computer Scientist

Achievements

Prizes = institute_of_Electrical_and_Electronics_Engineers%23Fellow_Grade
Association_for_Computing_Machinery%23Fellows

Contributions (DBLP Indexed)

DBLP_autho_link= http://dblp.l3s.de/d2r/page/authors/Gio_Wiederhold

Publication List

books= Integrating Temporal Data in a Heterogeneous Environment
Database Design, Revised 2nd Edition
Database Design more
journals = Composition of engineering web services with distributed data-flows and computations more
conferences = Panel: Is Generic Metadata Management Feasible?
[...Section Deleted...]

Figure 9.7: Visualization of Author Profilen

9.8 Findings

After this study, we figured out some recommendations which can help in making improvements in datasets and can benefit the ever growing data openness idea.

1. By following trend of more openness and information sharing, users / authors must participate in the social semantic web (Wikipedia) to build up their professionals profiles for the sake of collaborations and visibility.

2. Semantic Search engines Sindice and Social semantic web repository DBpedia to clean the dataset from non existing URI's.

3. There are many profiles which are represented ambiguously with the similar name but representing other persons. Additional methods need to be introduced by DBpedia to represent ambiguous persons pages.

At the moment in DBpedia, a unique ID is assigned to person over title text of a page which is considered a good practice to uniquely identify a resource unless persons having similar name, he/she cannot be identified correctly. This can clearly lead to the name ambiguity and searching problem. Assigning a URI on the basis of profession or certain characteristic can be an option.

9.9 Summary

In this chapter we highlighted the added value which can be drawn from rich semantic metadata repository of Linked Data for the real world application like profiling systems in Digital Journals. This chapter investigates and highlights the potentials for digital publishing system (e.g. J.UCS) by intelligent manipulation of the semantic search services and datasets. This linking is helpful for different scenarios e.g.: for users who are searching research collaborators, for journal administration who want to assign new reviewers and for users who want to explore experts to seek guidance. A comprehensive profile of an author is structured and visualized at one place providing various opportunities for collaborations. This is helpful in getting deep insights of authors work, personal and professional life.

Chapter 10

Linking Experts Profiles in Linked Data

Finding and assigning experts to a subjective field in academics, enterprises and in almost every practical field is an important and challenging task. Chapter 7 proposes and implements an automated approach for measuring expertise profile in academia. The proposed approach incorporates multiple metrics for measuring an overall expertise level. To visualize a rank list of experts, an extended hyperbolic visualization technique is proposed and implemented for Journal of Universal Computer science. In further, to facilitate the administration in assignment of experts for peer reviews duties etc. additional related information of an expert is pushed. In scientific domain, the identification of experts is normally based on number of factors like: number of publications, citation record, and experience etc. However, the discovered experts cannot be assigned reviewing duties immediately. One also need further information about expert like the country, university, service record, contributions, honors, and name of conferences/journals where the discovered expert is already serving as editor/reviewer. To some extent, this information can be found from search engines using heuristics, by applying *Natural Language Processing* and *Machine Learning techniques*. However, the emergence of many semantically rich and structured datasets from Linked Open Data movement can facilitate in more controlled search and fruitful results.

This chapter further comprised of an automatic technique to find the required information about experts using Linked Open Data dataset. The expert profile is discovered, aggregated, clustered, structured, and visualized to the administration of peer-review system.This highlights Expertise Mining as an other area where the proposed system CAF-SIAL in this thesis holds potential.

10.1 Introduction

The expertise mining is an important task in different scenarios and settings such as enterprises, medicine, software engineering, on-line forums, and peer-review systems etc. Particularly, in peer-review systems, this task can be divided into two sub-tasks such as: the discovery of expertise, and the assignment of reviewing duties to the discovered experts.

Once the experts have been discovered for the focused paper or topic, subsequently, the peer-review administration carefully assigns reviewing duties to the discovered experts. For the assignment task, the administration further looks on the information about discovered experts. For example, the university name, country information, and service record of an expert is required to avoid conflict of interest. The awards/honors and reviewing services of an expert is searched to see whether the discovered expert is qualified enough for the reviewing duties. This information can be acquired from search engines using SOAP APIs. The information can further be filtered by applying Natural Language Processing and Machine Learning techniques .This further requires community effort to correct and edit the discovered information. However, the quality of results can be enhanced if we have semantically rich meta-data. In this regard, one of the successful projects of the Semantic Web, the Linked Open Data (Linked Open Data)[Bizer et al. 2009] provides semantically rich datasets. These datasets can be exploited to discover relevant information and creating cross-references.

Usually, the construction of expert profile from Linked Data consists of four tasks: 1) disambiguation of experts, 2) retrieval of information resources 3) structuring the retrieved information and 4) presentation of structured resources. Although the data sets provided by Linked Open Data are semantically rich, but there are some problems e.g. locating exact resource URI, lack of quality interlinking [Jaffri et al. 2008], and not much of user interfaces for Linked Data consumption [Heath 2008]. To locate intended resources URI, we have already presented an intelligent Keyword-URI mapping technique [Latif et al. 2009b]. To retrieve and structure information resources, an innovative Concept Aggregation Framework [Latif et al. 2009a] was developed. To present a comprehensive view of information, a user interface named as CAF-SIAL was developed.

From our previous research experience in this domain, we have employed a set of rigid heuristics to find expert profiles from Linked Open Data. The experts are disambiguated using semantic metadata and intended information about experts is acquired and structured by using Concept Aggregation Framework. This approach is implemented for an online journal such as Journal of Universal Computer Science. In J.UCS, ACM categories are used to categorize the paper and experts are visualized using hyperbolic tree on these categories. On clicking an ACM topic, a ranked list of experts is superimposed in spiral visualization. When the J.UCS administration click on any discovered expert, a comprehensive profile of an expert is visualized further clustered down into informational aspects (personal, professional, academics and contributions). The presented expert profile facilitates J.UCS administration to look for required information about experts immediately. This helps the administration to assign reviewing duties more efficiently.

10.2 Related Work

Expertise finding and linking them to their respective resources can be divided into two phases 1) approaches to find experts and 2) discovery and construction of an expert profiles.

Expertise finder systems in the past have been innovatively applied in helping Ph.D applicants in finding relevant supervisors [Liu and Dew 2004] and also in identifying peer-reviewers for a conference [Rodriguez and Bollen 2008]. There are also expertise detection systems that were based entirely on an analysis of user activity and behavior while being engaged in an electronic environment. Krulwich et al., [Krulwich and Burkey 1995] have analyzed the number of interactions of an individual within a discussion forum as a means of constructing an experts profile. Mockus et al., [Mockus and Herbsleb 2002] employed data from a software projects change management records to locate people with desired expertise in a large organization. Yimam [Yimam 1999] have further shown that a decentralized approach can be applied for information gathering in the construction of expertise profiles. Tho et al. [Tho et al. 2007] employed a citation mining retrieval technique where a cross mapping between author clusters and topic clusters was applied to assign areas of expertise to serve as an additional layer of search results organization.

In scientific domain, Cameron [Cameron et al. 2007] employed a manually crafted taxonomy of 100 topics in DBLP [DBLP 2009] covering the research areas of a small sample of researchers appearing in DBLP. The number of publication is not always a true reflective of expertise of an individual. Cameron [Cameron et al. 2007] employed publication impact as an additional measure for expertise mining. In our previous work [Afzal et al. 2009], we employed number of factors as facets for expertise mining. We used multi-faceted approach where one can rank expert more comprehensively.

The other important task besides identification of experts is to assign reviewing duties to the discovered experts. For the assignment duties, the peer-review system administration needs further information about experts like expert's country, university, service record etc. This information can be acquired from recently emerged semantically enriched and interlinked datasets curate on principle of Linked Open Data. Various studies have been conducted to highlight the potential benefits in using these dataset. Freebase [Bollacker et al. 2008], a general collaborative knowledge base, identifies and aggregate facts about resources (persons, movies, places etc.) from Linked Open Data. Further this system presents a detail and consolidated view of this aggregated facts to the user. Another system RKB explorer [Glaser and Millard 2007], functions specifically as a person's data identifier from Linked Open Data. This system provides its user of an organization (project members and other researchers) with a unified view of data gathered from various Linked Open Data data sources related to a selected domain. Recently, Stankovic [Stankovic et al. 2010] proposed set of heuristics (employed on related information of a subjective person present on web) to identify experts from Linked Open Data datasets. This study has shown that much of the information needed in this regard i.e. personal information, publication list, authored books, and blogs etc. can be located from Linked Open Data. All these studies showcased Linked Open Data as a rich test bed for exploration of data in context of various domains.

In our previous work, we proposed and implemented a system CAF-SIAL (Concept Aggregation Framework for Structuring Informational Aspects of Linked Open Data) [Latif et al. 2009a] which uses Semantic technologies and Linked Open Data datasets to search persons data. CAF-SIAL is a proof of concept application to discover and present informational aspect of resource (Person) from Linked Open Data. CAF-SIAL is based on a methodology for harvesting person's relevant information from the gigantic Linked Open Data cloud. The methodology is based on combination of information: identification, extraction, integration and presentation. Relevant information is identified by using a set of heuristics. The identified information resource is extracted by employing an intelligent URI discovery technique. The extracted information is further integrated with the help of a Concept Aggregation Framework. Then the information is presented to end users in logical informational aspects. This automatic system scaled in accuracy and efficiency. Our system can retrieve relevant information for a person and can structure and visualize it comprehensively. Keeping in mind related work in this field and our problem definition, we visualized the calculated experts in hyperbolic tree and then interlinked it with CAF-SIAL to facilitate the administration of J.UCS for getting information of a person.

10.3 Application Architecture

This section discusses system design for expertise mining, visualization, construction and linking of experts profiles. System design is divided in to three layers which

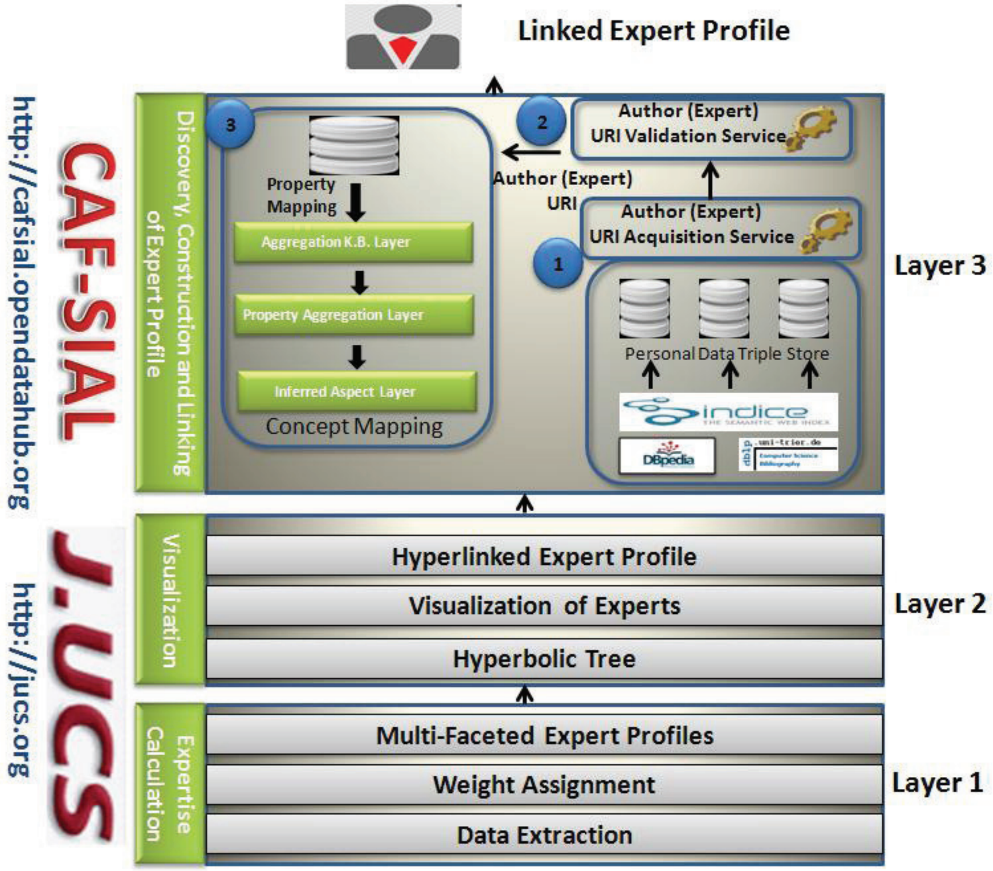


Figure 10.1: Application Architecture

interact with each other to make system operational as depicted in figure 10.1.

- Expertise Calculation.
- Information Visualization.
- Locating and Construction of Expert Profile in Linked Open Data.

10.4 Expertise Calculation

In our previous research work, we used a multi-faceted approach for ranking experts for computer science topics. Three weights i.e. publication weight, citation weight, and reviewer weight were used for this task. Subsequently, all weights were added to get an overall ranking of experts in the area; *cf.* [Afzal et al. 2009] and Chapter 8.

10.5 Information Visualization

Experts are essentially attached to a node within the ACM classification hierarchy. A hyperbolic browser [Afzal et al. 2009] is used to provide intuitive navigation within the ACM classification hierarchy. For any selected node in the ACM hierarchy, a spiral is used to visualize the ranked list of high-profile experts for that node. The spiral is simply superimposed upon and around the selected node where as the top rank expert remains in the center.

10.6 Locating and Construction of Expert Profile in Linked Open Data

Third layer of application architecture is divided into three sub parts named such as: 1) expert URI Acquisition Service, 2) expert URI Validation Service, and 3) concept mapping service as shown above in figure 10.1.

10.6.1 Expert URI Acquisition Service

Expert URI acquisition Service is responsible to acquire URI of an expert from local triple stores and from remote semantic search services.

ARC2 utility was used to construct local triple store for better RDF and SPARQL query management. Three datasets were used, one from J.UCS and other two from DBpedia. Along with these datasets, web service of Sindice [Oren et al. 2008] (a semantic search engine) was also utilized when the local search failed. Three processing steps were performed to acquire the intended URI.

1. Direct matching of J.UCS authors with DBpedia Persondata dataset.
2. Direct matching of J.UCS authors with Links to DBLP dataset.
3. Querying and Filtering of URI from Sindice.

Few experts found their match in the first two processes. To increase the discovery of URI, different online semantic Search engines were analyzed like Sindice [Oren et al. 2008], Falcon [Cheng et al. 2008], Swoogle [Ding et al. 2004]. Based on the up-to-date version, vast indexing corpus, and easy API access, the Sindice was selected for further matching of URIs. Sindice provided the list of the URIs for focused expert which was further filtered out on the basis of DBpedia provenance as explained in next section.

Process Name	**Considered Experts**	**DBpedia URI**	**DBLP URI**
DBpediaPerson direct match	1073	2	NA
DBpediatoDBLP direct math	1071	5	5
Sindice(Semantic Search Service)	1065	124	944
Uri-Authentication	131	28	977
DBLP & DBpedia URIs	1073	28	28
DBLP URI	1045	NA	949

Table 10.1: J.UCS Experts URI processing

10.6.2 Expert URI Validation Service

Expert URI validation service performs identity management to find correct URI of the expert. By manual de-referencing and inspection of the acquired URI's, we found two major inconsistencies in acquired URI data: 1) URI of the respective expert exists (wrongly indexed by Sindice), containing no RDF statements making it useless 2) name ambiguity: in many cases, there were persons having same name but belonging to different domains of science and arts, leading to wrong person.

To find the correct URI for an expert, set of heuristics were devised i.e.: 1) URI must contain some RDF statements, 2) presence of certain keywords in the abstract or SKOS categories such as: "computer science, computer scientist, professor, informatics, researcher ...." etc. This set of keywords helped us in successfully disambiguated computer science individuals from the persons belonging to other domains.

Statistics regarding number of results (queried experts, retrieved URIS of DBLP and DBpedia and filtered URI) is mentioned in Table 10.1. After processing, we were able to find 28 experts having both DBLP and DBpedia URIs and 949 experts having only DBLP URI. In total 977 experts were linked with CAF-SIAL application.

10.6.3 Concept Mapping Service

The concept mapping service maps the retrieved set of properties from experts URI onto concept aggregation framework. The concept mapping service receives a valid URI from URI Authentication Service. This URI is used to retrieve available set of properties from Linked Open Data (DBpedia Server). These properties were passed to concept aggregation framework which organized these properties in informational aspects of an author. The concept mapping module returns a comprehensive author profile represented in logically organized informational aspects. These aspects are further visualized to users.

10.7 Case Study Visualization

The experts of J.UCS are visualized using hyperbolic visualization [Hyperbolic Package 2009]. In the current work, hyperlink for all the discovered experts is generated. These hyperlinks connect experts discovered in this study with CAF-SIAL application. For Example, when a user navigate the hyperbolic tree and click on the Node "G. Mathematics of Computing" (ACM category), tree is redrawn on the canvas focusing the Node "G." to center along with a Spiral showing ranked list of experts. In this case one expert named as "Cliff Jones" is displayed as shown in the figure 10.2.

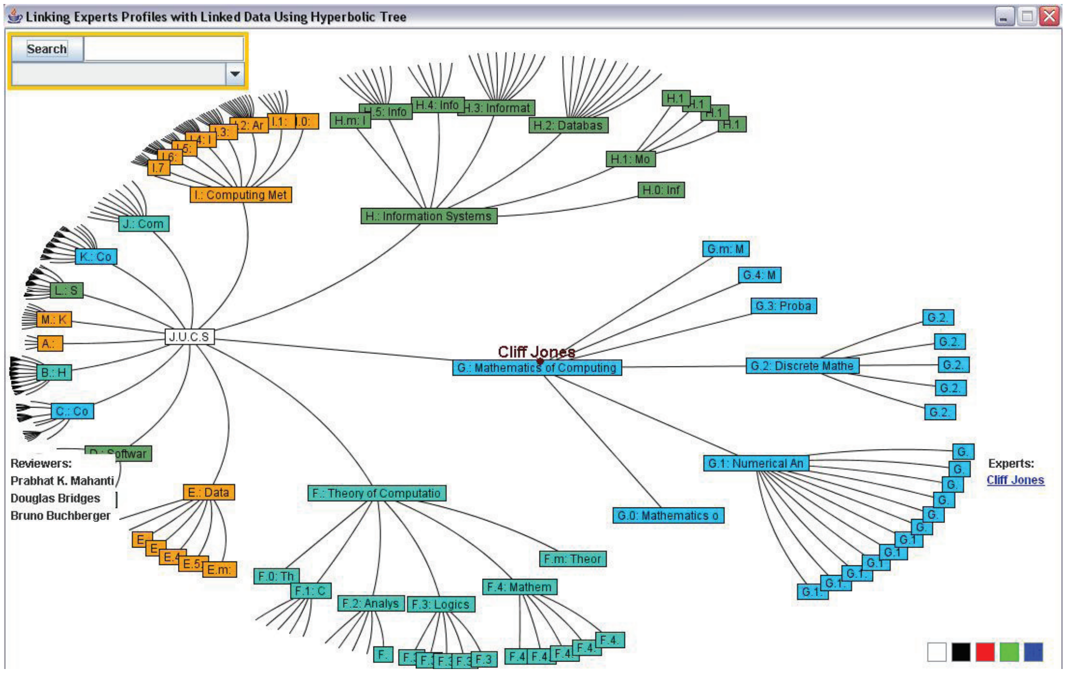
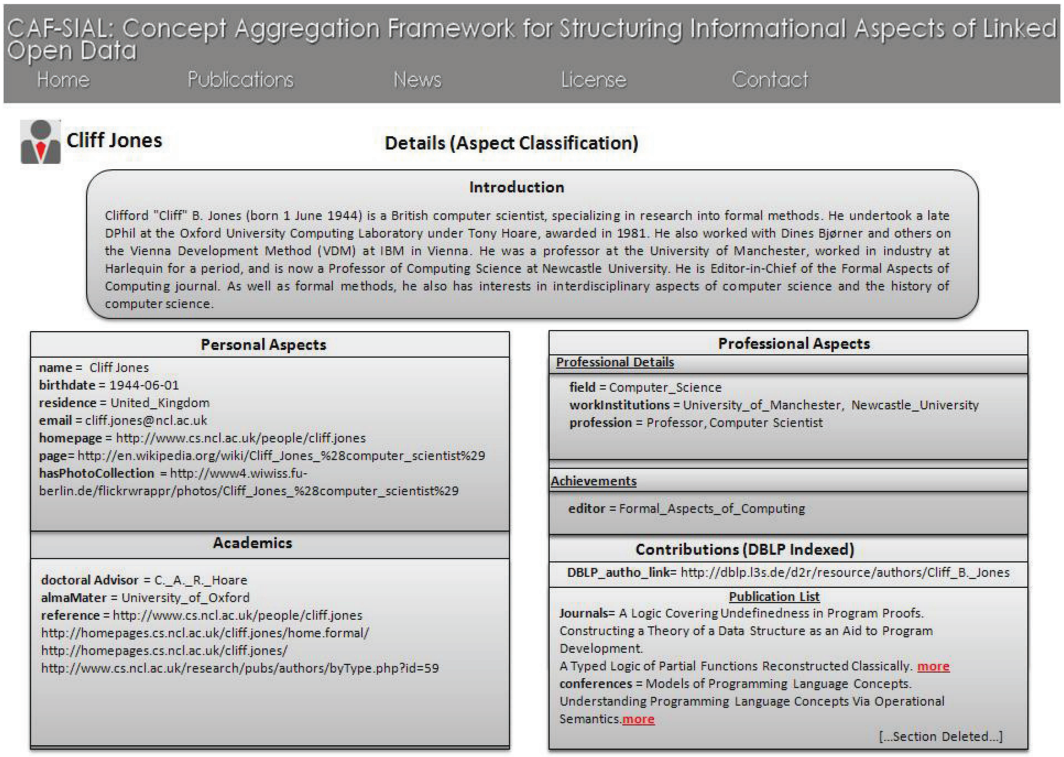


Figure 10.2: Extended Hyperbolic Tree Visualization

When the user clicks on the expert name, it leads user to CAF-SIAL application where concept aggregation framework makes a visualization of the expert profile as shown in figure 10.3. From visualized expert profile, it is obvious that an administration will get instant information about the concerned person. A brief description of a person is presented in an introduction section. Furthermore different aspects like personal, professional, academics, and contributions are shown to the administration. From this coherent view, user gets a first overall impression about a person and can follow any hyperlink to see further details. This comprehensive expert profile assists J.UCS administration to assign reviewing duties.


CAF-SIAL: Concept Aggregation Framework for Structuring Informational Aspects of Linked Open Data
Home
Publications
News
License
Contact

Cliff Jones

Details (Aspect Classification)

Introduction

Clifford "Cliff" B. Jones (born 1 June 1944) is a British computer scientist, specializing in research into formal methods. He undertook a late DPhil at the Oxford University Computing Laboratory under Tony Hoare, awarded in 1981. He also worked with Dines Bjørner and others on the Vienna Development Method (VDM) at IBM in Vienna. He was a professor at the University of Manchester, worked in industry at Harlequin for a period, and is now a Professor of Computing Science at Newcastle University. He is Editor-in-Chief of the Formal Aspects of Computing journal. As well as formal methods, he also has interests in interdisciplinary aspects of computer science and the history of computer science.

Personal Aspects

name = Cliff Jones
birthdate = 1944-06-01
residence = United_Kingdom
email = cliff.jones@ncl.ac.uk
homepage = http://www.cs.ncl.ac.uk/people/cliff.jones
page= http://en.wikipedia.org/wiki/Cliff_Jones_%28computer_scientist%29
hasPhotoCollection = http://www4.wiwiss.fu-berlin.de/flickrwrappr/photos/Cliff_Jones_%28computer_scientist%29

Academics

doctoral Advisor = C._A._R._Hoare
almaMater = University_of_Oxford
reference = http://www.cs.ncl.ac.uk/people/cliff.jones
http://homepages.cs.ncl.ac.uk/cliff.jones/home.formal/
http://homepages.cs.ncl.ac.uk/cliff.jones/
http://www.cs.ncl.ac.uk/research/pubs/authors/byType.php?id=59

Professional Aspects

Professional Details

field = Computer_Science
workInstitutions = University_of_Manchester, Newcastle_University
profession = Professor, Computer Scientist

Achievements

editor = Formal_Aspects_of_Computing

Contributions (DBLP Indexed)

DBLP_autho_link= http://dblp.l3s.de/d2r/resource/authors/Cliff_B._Jones

Publication List

Journals= A Logic Covering Undefinedness in Program Proofs.
Constructing a Theory of a Data Structure as an Aid to Program Development.
A Typed Logic of Partial Functions Reconstructed Classically. more
conferences = Models of Programming Language Concepts.
Understanding Programming Language Concepts Via Operational Semantics.more

[...Section Deleted...]

Figure 10.3: Visualized Expert Profile

10.8 Summary

This chapter presented a new system to identify, visualize and link potential experts with the CAF-SIAL Linked Data application. It is used in the context of a computer science journal to identify and assign reviewers to areas of computer science. Further to facilitate the administration responsible to select and assignment of reviewing duties in a peer review setting, additional information of identified experts in a well crafted profile is pushed and linked in Linked Open Data. This system can easily be generalized to other scientific communities.

The main contributions of this chapter are:

1. Linking of visualized potential experts with Linked Open Data by using concept aggregation framework.

2. Pushing experts related information to administration of journal in a well maintained profile, which they can use for taking peer review assignment decisions.

Part V

Results and Discussions

Chapter 11

System Evaluations

11.1 Introduction

CAF-SIAL: Concept Aggregation Framework for Structuring Informational Aspects of Linked Open Data, described in Chapter 5 and Chapter 6 , was implemented as a test of the algorithms and approaches proposed in this research work. After implementation of the system, an evaluation was carried out to achieve three things.

1. Assess the ability of the CAF-SIAL system as a whole (algorithms) to locate and aggregate the data from Linked Open Data cloud in a way which is easily perceivable by the end users.

2. To determine whether meaningful results could be produced with a minimal level of data input.

3. To determine whether providing a user interface, which is powered by keyword search mechanism with auto-suggestion facility and have backend processing of semantic meta-data (Triples) for constructing aspect oriented profiles, reduce the effort of a naive end-user while he interacts with Linked Open Data in search and exploration purposes as compared with other systems like Marble, Snorql and Freebase.

Jenny Kitzinger [Kitzinger 1995] discusses methods for evaluating systems and he introduced the new dimension in system evaluation by presenting focus groups methodology. Focus groups are a type of group interview that brinks on communication between research participants to produce data. Group interviews are often used simply as a quick and handy way to collect data from several people concurrently. This means that instead of the researcher asking each person to respond to a question in turn, participants are encouraged to communicate and interact with each other in asking questions, sharing their experiences and thoughts to answer. The method is particularly useful for exploring people's knowledge and experiences and can be used to examine not only what people think but how they think and why they think that way.

CAF-SAIL system was evaluated with the help of user interviews. We collected data with the help of combining focus groups and post-search interviews.

11.2 Design

This user study is designed to evaluate system CAF-SIAL .Intended objective of this study is to find out, the score of aspects (Presentation, Performance, and Ease in use) in different use case scenario in comparison with existing operational system. In the end this evaluations will be use to validate our proposed techniques and to improve our system in line of suggested directions. The evaluation was designed to address Research Question 2.1 and Research Question 2.2.

- How can the URI of a resource be located from Linked Open Data?
- How can retrieved resources from Linked Data be structured and presented in more perceivable way?

These research questions can be rephrased into more operational terms:

Linked open data is one of the driving forces in true realization of Semantic web by providing diverse kind of semantified (structured) datasets. With success of this project there are some convolutions which also peeped up along the way. The one factor to some extent is ignoring user perspective (usability) by emphasizing more on complex practices to make data more structure. As a result the normal operation like searching, filtering and organization of information is not more a trivial job in semantic web paradigm. We have proposed an idea of simple user interface which at maximum tries to hide the underlying complex mechanics associated with operations like searching, filtering. This system contribute in devising a technique to locate a respective URI from huge Linked data cloud and then filter out informational aspect more meaningful to user, which at the moment present is in abundance and non-conceivable way in the Linked Open Data pool.

11.3 Method

11.3.1 Participants

The sample used in the evaluation consisted of in total 10 participants mainly the web users. To get the diversified opinions on our proposed system, we selected two focus groups on the basis of these given characteristics.

1. All participants of both groups were Web users and had knowledge of basic Web search.

2. Participant of first focus group had knowledge of Semantic technologies (SPARQL, interacted with Linked Open applications for searching and browsing).

3. Participants of second group were naive web users. They had neither used Semantic technologies nor had interacted with Linked Open Data applications.

We held two focus group sessions having 4-6 participants each (10 participants in total). Each group session was conducted by a skilled representative.

11.3.2 Procedure

All participant were asked to use CAF-SIAL, Marble, Snorql, and Freebase to search for the persons they liked belonging to different field of professions with respect to the DBpedia person type ontology. For example users were asked to query for persons belonging to this profession list in the CAF-SIAL, Marble, Snorql, and Freebase:

- Politician.
- Athlete
- Actor
- Monarch
- Scientist
- Journalist
- Architect
- Football manager
- Criminal
- Philosopher

In the end we conducted semi-structured interviews. Users were asked about feedbacks in the form of scores rating (1 to 10) to judge performance, complexity and overall satisfaction in using these systems. In summary, the questionnaire prepared for the evaluation consisted of three parts and was used accordingly by the interviewer.

1. In start the overall understanding of the user to semantic web was checked.

2. User was given various searching tasks in Semantic search system.

3. User was asked about their overall satisfaction on different systems according to given tasks completion.

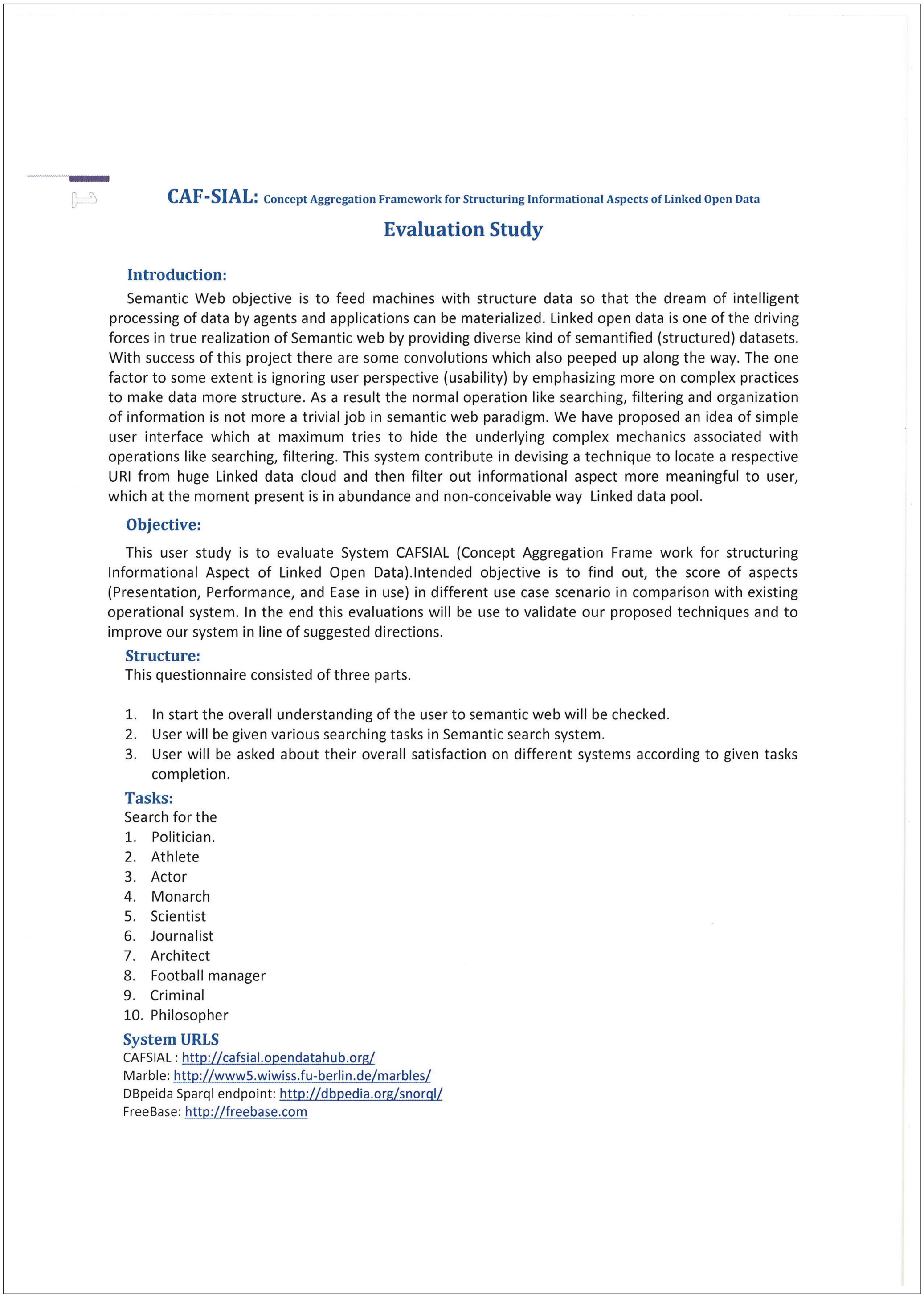
The screen shots of the complete questionnaire are shown in figure...

11.4 Resutls and Discussion

The interviews showed that it was very difficult to get the sought-after information from Marble and Snorql, because the users did not know the exact URI of the resource, and due to the difficulty of formulating Subject Predicate Object logic. On the other hand, Freebase and CAF-SIAL systems were easier to use. Although Freebase was comparatively better in terms of providing rich information and content organization, CAF-SIAL was useful most of the time to get the sought-after information. One negative aspect that was mentioned about CAF-SIAL was the fact that when the users searched for a person and then clicked on a particular property within the retrieved result set, they were again redirected to complex systems like DBpedia. Then again it became difficult to find required information from a long list of properties. When comparing CAF-SIAL to Freebase, there are some noteworthy differences:

1. DBpedia, which provides that basis for CAF-SIAL, is built around a controlled vocabulary (an ontology, actually), whereas freebase adopts the folksonomy approach in which people can add new categories much like tags [OReilly 2007].

2. Along with the semi-automatic approach of Freebase to collect and organize data in to their knowledge base, a group of editors is responsible to pre-check the organization and add new knowledge in a structured way manually. On the other hand, CAF-SIAL works on a set of heuristics. These heuristics were defined manually once by experts and then applied on the system to organize knowledge in an entirely automated way. The system makes a transition to exploit Linked Data resources in an autonomous way, which could provide significant help in navigating the ever-growing Linked Data cloud. The system has been made publicly available at `http://cafsial.opendatahub.org/`. For continuous evaluations, users can give feedback at any time on-line. The submitted information is saved and we plan to extend the system by incorporating user's comments and feedback.



CAF-SIAL: Concept Aggregation Framework for Structuring Informational Aspects of Linked Open Data

Evaluation Study

Introduction:

Semantic Web objective is to feed machines with structure data so that the dream of intelligent processing of data by agents and applications can be materialized. Linked open data is one of the driving forces in true realization of Semantic web by providing diverse kind of semantified (structured) datasets. With success of this project there are some convolutions which also peeped up along the way. The one factor to some extent is ignoring user perspective (usability) by emphasizing more on complex practices to make data more structure. As a result the normal operation like searching, filtering and organization of information is not more a trivial job in semantic web paradigm. We have proposed an idea of simple user interface which at maximum tries to hide the underlying complex mechanics associated with operations like searching, filtering. This system contribute in devising a technique to locate a respective URI from huge Linked data cloud and then filter out informational aspect more meaningful to user, which at the moment present is in abundance and non-conceivable way Linked data pool.

Objective:

This user study is to evaluate System CAFSIAL (Concept Aggregation Frame work for structuring Informational Aspect of Linked Open Data).Intended objective is to find out, the score of aspects (Presentation, Performance, and Ease in use) in different use case scenario in comparison with existing operational system. In the end this evaluations will be use to validate our proposed techniques and to improve our system in line of suggested directions.

Structure:

This questionnaire consisted of three parts.

1. In start the overall understanding of the user to semantic web will be checked.
2. User will be given various searching tasks in Semantic search system.
3. User will be asked about their overall satisfaction on different systems according to given tasks completion.

Tasks:

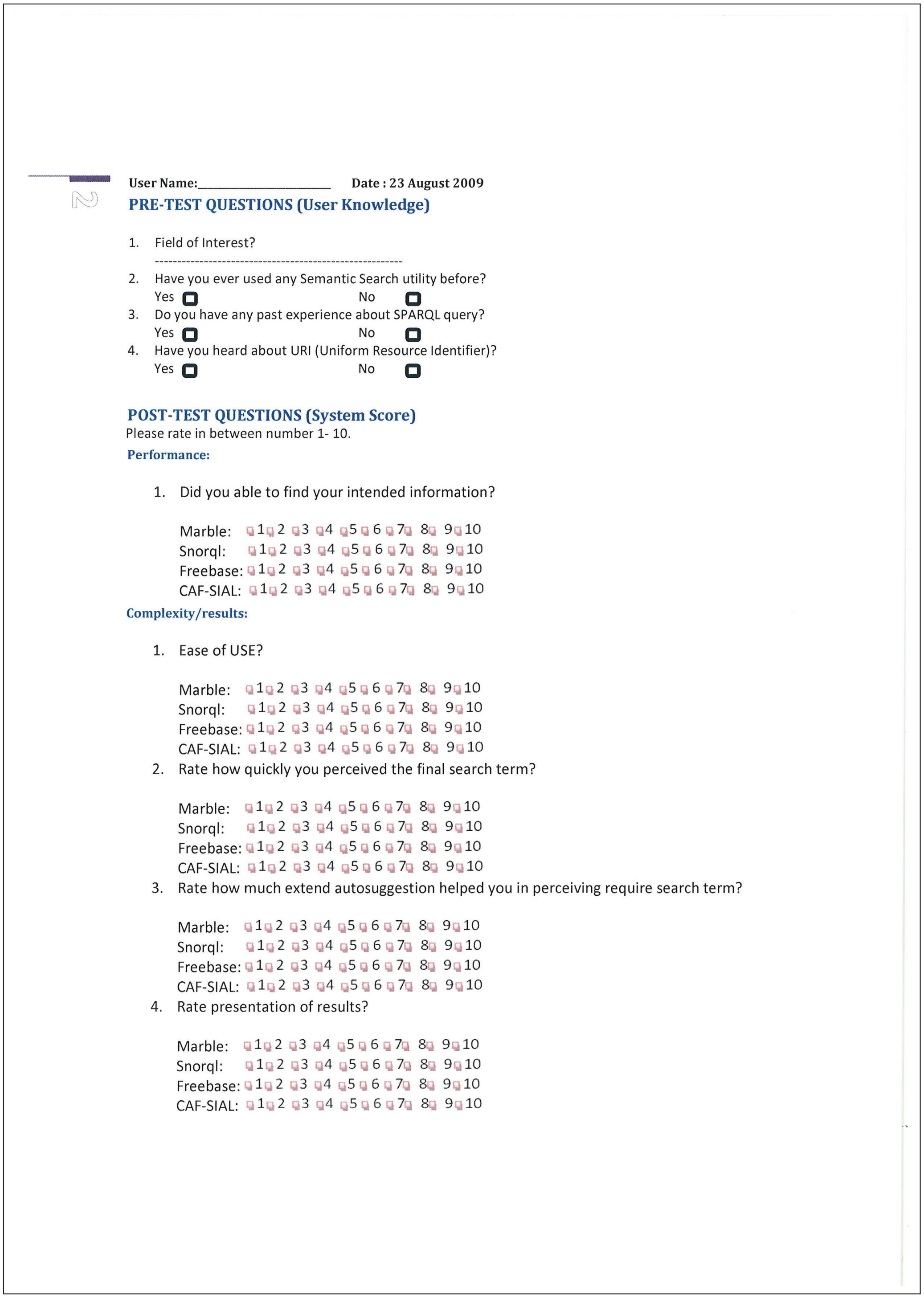
Search for the

1. Politician.
2. Athlete
3. Actor
4. Monarch
5. Scientist
6. Journalist
7. Architect
8. Football manager
9. Criminal
10. Philosopher

System URLS

CAFSIAL : http://cafsial.opendatahub.org/
Marble: http://www5.wiwiss.fu-berlin.de/marbles/
DBpeida Sparql endpoint: http://dbpedia.org/snorql/
FreeBase: http://freebase.com

Figure 11.1: System Evaluations Questionnaire - Page1-



User Name:________________________ **Date : 23 August 2009**

PRE-TEST QUESTIONS (User Knowledge)

1. Field of Interest?

   ----
2. Have you ever used any Semantic Search utility before?

   Yes □ No □
3. Do you have any past experience about SPARQL query?

   Yes □ No □
4. Have you heard about URI (Uniform Resource Identifier)?

   Yes □ No □

POST-TEST QUESTIONS (System Score)

Please rate in between number 1- 10.

Performance:

1. Did you able to find your intended information?

   Marble: □1□2 □3 □4 □5□6□7□ 8□ 9□10
   Snorql: □1□2 □3 □4 □5□6□7□ 8□ 9□10
   Freebase: □1□2 □3 □4 □5□6□7□ 8□ 9□10
   CAF-SIAL: □1□2 □3 □4 □5□6□7□ 8□ 9□10

Complexity/results:

1. Ease of USE?

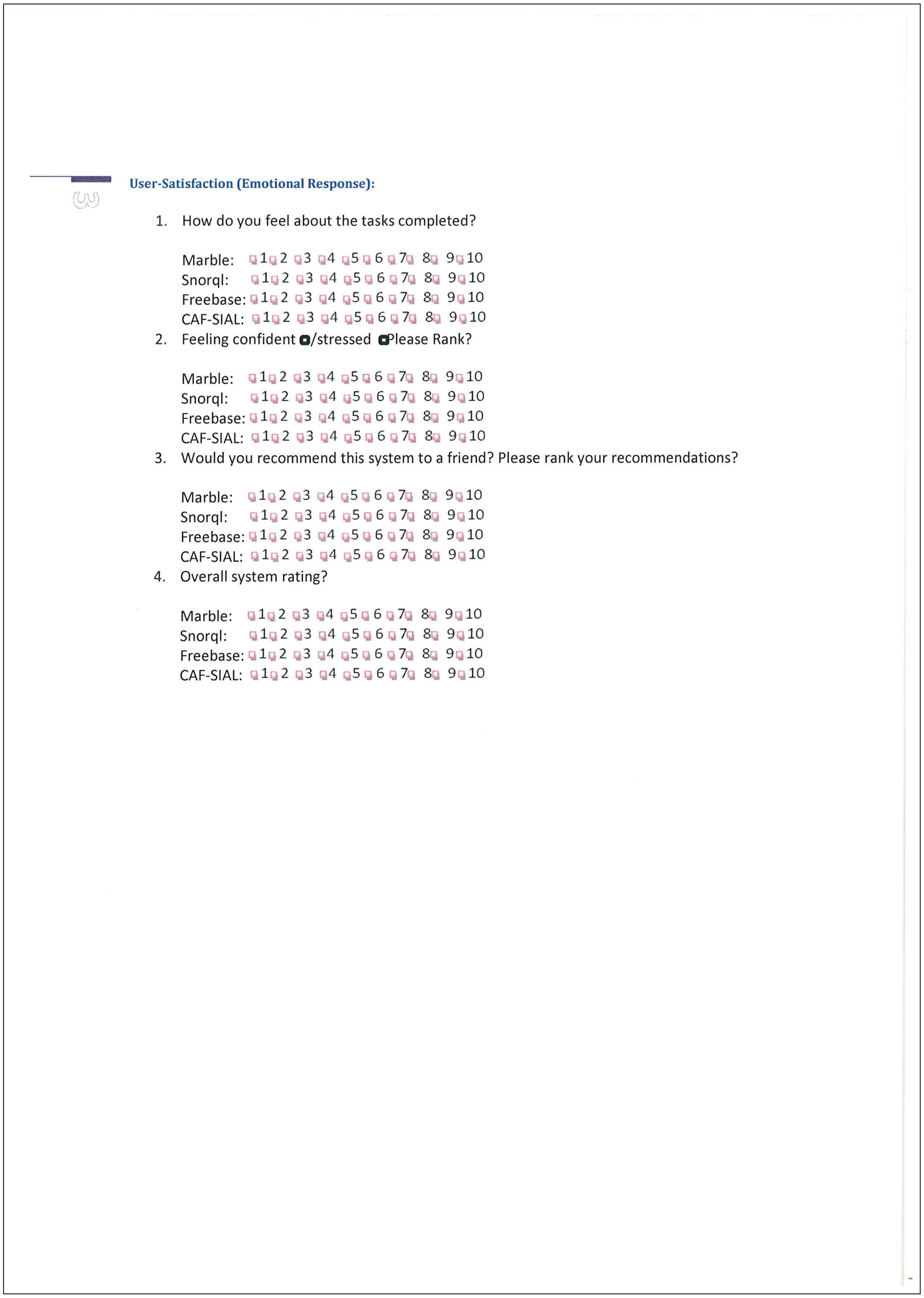
   Marble: □1□2 □3 □4 □5□6□7□ 8□ 9□10
   Snorql: □1□2 □3 □4 □5□6□7□ 8□ 9□10
   Freebase: □1□2 □3 □4 □5□6□7□ 8□ 9□10
   CAF-SIAL: □1□2 □3 □4 □5□6□7□ 8□ 9□10
2. Rate how quickly you perceived the final search term?

   Marble: □1□2 □3 □4 □5□6□7□ 8□ 9□10
   Snorql: □1□2 □3 □4 □5□6□7□ 8□ 9□10
   Freebase: □1□2 □3 □4 □5□6□7□ 8□ 9□10
   CAF-SIAL: □1□2 □3 □4 □5□6□7□ 8□ 9□10
3. Rate how much extend autosuggestion helped you in perceiving require search term?

   Marble: □1□2 □3 □4 □5□6□7□ 8□ 9□10
   Snorql: □1□2 □3 □4 □5□6□7□ 8□ 9□10
   Freebase: □1□2 □3 □4 □5□6□7□ 8□ 9□10
   CAF-SIAL: □1□2 □3 □4 □5□6□7□ 8□ 9□10
4. Rate presentation of results?

   Marble: □1□2 □3 □4 □5□6□7□ 8□ 9□10
   Snorql: □1□2 □3 □4 □5□6□7□ 8□ 9□10
   Freebase: □1□2 □3 □4 □5□6□7□ 8□ 9□10
   CAF-SIAL: □1□2 □3 □4 □5□6□7□ 8□ 9□10

Figure 11.2: System Evaluations Questionnaire - Page2-



(3)

User-Satisfaction (Emotional Response):

1. How do you feel about the tasks completed?

   Marble: □1□2 □3 □4 □5□6 □7□ 8□ 9□10
   Snorql: □1□2 □3 □4 □5□6 □7□ 8□ 9□10
   Freebase: □1□2 □3 □4 □5□6 □7□ 8□ 9□10
   CAF-SIAL: □1□2 □3 □4 □5□6 □7□ 8□ 9□10

2. Feeling confident ▣/stressed ▣Please Rank?

   Marble: □1□2 □3 □4 □5□6 □7□ 8□ 9□10
   Snorql: □1□2 □3 □4 □5□6 □7□ 8□ 9□10
   Freebase: □1□2 □3 □4 □5□6 □7□ 8□ 9□10
   CAF-SIAL: □1□2 □3 □4 □5□6 □7□ 8□ 9□10

3. Would you recommend this system to a friend? Please rank your recommendations?

   Marble: □1□2 □3 □4 □5□6 □7□ 8□ 9□10
   Snorql: □1□2 □3 □4 □5□6 □7□ 8□ 9□10
   Freebase: □1□2 □3 □4 □5□6 □7□ 8□ 9□10
   CAF-SIAL: □1□2 □3 □4 □5□6 □7□ 8□ 9□10

4. Overall system rating?

   Marble: □1□2 □3 □4 □5□6 □7□ 8□ 9□10
   Snorql: □1□2 □3 □4 □5□6 □7□ 8□ 9□10
   Freebase: □1□2 □3 □4 □5□6 □7□ 8□ 9□10
   CAF-SIAL: □1□2 □3 □4 □5□6 □7□ 8□ 9□10

Figure 11.3: System Evaluations Questionnaire - Page3-

Chapter 12

Conclusion and Future Work

The huge access to Linked Data presents exciting opportunities for the next generation of Web-based applications but still there are several open issues from developer's and user's points of view, which makes the publishing and development of Linked-Data-based applications challenging and need attention. This thesis explores the various aspects of Linked Data and investigates some of the open issues and has proposed a framework, in the form of conceptual models and applications as a solution.

This final chapter reflects on the goals set at the beginning of this research in form of an assessment regarding the main research question and the sub-research questions.

12.1 Assessment

This section serves as a self-assessment part of this thesis to highlight the achievements with respect to the asked research questions as listed in Section 1.3. In the following, the investigated research questions are revisited and discussed from a retrospective viewpoint.
The main research question of this thesis is to investigate:

> *How conceptual models can simplify the understanding of Linked Data and application development. How, innovative approaches mixed with intelligent use of semantic technologies employed to develop an easy to use application for normal web users is done, which can automatically acquires, process and present accurate and relevant information, by consuming Linked Data with hiding all complex mechanism of semantic and querying structures , as well as can bring benefits to other fields.*

In this thesis 5 research questions in total are investigated, which further on the basis of contributions are categorized into three common aspects of Linked Data as Linked Data consumption, Linked Data publishing and Linked Data application areas. These contributions are discussed in the following sections:

12.1.1 Linked Data Consumption

Linking Open Data has been facilitating the transformation of publicly available, open data into Linked Data since 2007. As of 2010, the vast majority of Linked Open Data is still generated by research communities and institutions while very less has been contributed from business community. There is still some nervousness in the business community for its adaption mainly due 1) lack of business cases on top of Linked Data 2) complicated procedures attached with data transformation 3) lack of consensus on ownership and access rights about legacy and third party data 4) how revenues generated in the Linked Data Sphere. For a successful corporate uptake, it is important to have a strong conceptual groundwork, providing the foundation for the development of business cases revolving around the adoption of Linked Data which in particular simplifying the Linked Data generation mechanism by explicating the phases and stake holders. For this, these listed research questions are investigated.

How does the value chain behind Linked Data looks like? What entities and processes are involved in the transformation from raw data to Linked Data? What are the potential pitfalls that might occur?

In pursuit of these research questions Chapter 4, contributed to facilitate the uptake of the Linked Data vision in commercial and non commercial level. A contribution was made by presenting the Linked Data Value Chain as a lightweight model for business case engineers to support the conceptualization of successful commercial Linked Data use cases. Thereby, identifying three main concepts: Different *Entities* acting in different *Roles* both consuming and providing different *Types of Data*. It also demonstrated that the assignment of roles to entities, the combination and involvement of roles, the data selected as well as the data transformation process itself hold inherent risks which must be taken in account before Linked Data uptake in commercial and non commercial disciplines. To validate the proposed Linked Data Value Chain, it was applied to British Broad Casting (BBC) case study; this successfully identified the entities, their roles and interactions which were involved in the generation of Linked Data.

It is assumed here that Linked Data Value Chain can be a good starting point for commercial and non- commercial parties in understanding Linked Data generation as a whole in a simple conceptual way. It will assist interested in identifying various roles, interactions and phases involved in linked data generation, as well as help them to conceptualize the value flow on the basis of data conversion and pitfalls which may harm the process in the Linked Data application process. This Linked Data value chain helped me in understanding Linked Data sphere in a better way and also highlighted some of the open issues, most important lack of easy to user interface specifically directed for the normal web users for exploring Linked Data. It was investigated that Linked Data cannot penetrate much until the similar standard functionalities in searching, navigation and organizations of information is provided to the users which they have enjoy in conventional web.

Possibly the biggest benefit of Linked Data from the end user viewpoint is the integrated access to data from wide range of multiple distributed and heterogeneous data sources. Simply, this process involves the integration of data from sources which are not clearly selected by users, as explicitly doing so acquire unacceptable working overhead mainly due to association of complex mechanisms and extra learning cycles. But doing search first to locate the resource and then process it for meaningful information from Linked Data is entirely different and complex for a normal web users as compared to the conventional web. As Linked Data is envisioned to server machines rather than humans, the importance and need of easy to use interfaces which provide linked Data in human perceivable way is essential. These interfaces must work in such an intelligent way, which makes a life of normal web user easy by hiding all the complex mechanism of searching, information consolidation and most importantly make use of Linked Data for human consumption, present in abundance. So in chapter 4 and 5 this formulated research questions are investigated.

How should a system for consuming linked data look like which provides searching and information organization features in a user-friendly way? How can underlying semantic mechanics be concealed from end users?

In pursuit of these research question two layered approach as Locating URI of the resource from Linked Data and the presentation of URI related information in a conceivable way is employed.

In first layer, contribution was made by introducing keyword-URI mapping technique which provides users with the keyword search facility on top of the Linked Data and automatically mapped keyword with the URI of Linked Data resources, elevating need of remembering URI from users. Furthermore an auto-suggestion service is also provided which facilitates users more in searching by providing users on the fly matched concepts with entered keywords. The exploratory study inducted; *cf.* Chapter 5 – section 5.5 – showed that auto-suggestion and Keyword-URI mapping technique proved beneficial to the users in searching over Linked Data.

In second layer, contribution was made by proposing and implementing a Concept Aggregation Framework. This framework on the basis of a novel algorithm gathers, processes, integrates and structures the related information of a resource (Person) from multiple data sources into categories of informational aspects e.g. Description, Personal, Professional, Social and Dark side.

By combining both of these techniques i.e. Keyword-URI mapping and Concept Aggregation Framework, an application named as CAF-SIAL is implemented and made accessible at `http://cafsial.opendatahub.org` for use. The evaluation of the systems has shown promising results; *cf.* Chapter 11. The foresee assessment of contributions made in quest of the research question let us to assume that, this kind of application is a step forward for the:

- Consumption of Linked Data
- Organization and presentation approaches for making Linked Data explicit and understandable
- Future applications development which can be built on top of the Linked Data by intelligent manipulation of technologies and hiding underlying semantic mechanisms.

12.1.2 Linked Data Publishing

Digital open Journals plays a vital role in the dissemination of codified knowledge. The Journal of Universal Computer Science (J.UCS) is a high-quality electronic publication journal that deals with all aspects of Computer Science [Calude et al. 1994]. J.UCS has incorporated many innovative features from its commencement, such as Experts, Tag recommendations. But all this features are accessible in legacy HTML format and due to unstructured nature of its format, very little is offered to the machine processing domain. The Linked Data Project supports the Open data movement in a way, to open information with more structured meta-data, which is machine understandable, easily locatable, identifiable and linkable for better searching and discovery processes. To reap the real benefits offered by Linked Data paradigm and taking J.UCS to Linked Data sphere, in this thesis a framework for *Linked Data publishing* is provided to convert Legacy HTML data of Open Digital Journal (J.UCS). The formulated research question is from the Linked Data publishing domain and is formulated as:

How can RDFization and interlinking of legacy HTML detasets (Authors, Papers, Experts and Tags) from Digital Journal, with external data sources be done. What are the benefits for making them available in Linked Data cloud?

This research questions was answered in two steps in this thesis. Firstly, a contribution is made by linking papers of a digital journal on given authors keywords with resources available in social bookmarking system. In addition to this contribution, an expertise mining system is implemented and calculated experts through this system are visualized in a hyperbolic tree for Journal administration which proved very helpful. Moreover, consolidation of these contributions in dataset is performed and these features are made available in Digital Journal Web site for its users in Legacy HTML format.

Secondly, contribution is made, by converting all these datasets (Authors, Papers, calculated Experts and Tags) into machine readable format (RDF) and by interlinking them with external Linked Data resources. At first, all these datasets are modeled into semantic representation by reusing various popular ontoltogies; *cf.* Chapter 8 – subsection 8.3.1 –. To interlink these datasets with the external datasets, an interlinking approach is employed this approach find and validate the related resources URIs from Linked Data external datasets. Further on, a D2R server mapping file is created for the RDFization of these relational data tables. In the end RDFzied dump of J.UCS datasets, interlinked with external Linked Data resources is offered openly for the knowledge discovery and further interlinking. For the human interaction and navigation with this RDFzied J.UCS data, an HTML interface with the SPARQL query endpoint is provided. Assessing the results after investigation of the research question, it is assumed that a handful contribution is made for the scientific community, Linked Data community and for Digital Journal itself by RDFizing this J.UCS dataset. Some of the foreseen assessment of the claimed contributions are:

- J.UCS became one of the few Digital Open Journal publishing their data as Linked Data and is frequently contributing to the Linked Data cloud.
- J.UCS auhtor, papers, tags and experts are interlinked with DBpedia, DBLP, Faceted DBLP and CiteUlike, which can server as resource for additional knowledge discovery and further interlinking by external resources.
- Simple and complex SPARQL queries are possible on J.UCS datasets which will lead to discovery of hidden concepts.
- Third party mashup applications can be build by minting J.UCS and other scientific publishing Linked Data resources.

12.1.3 Linked Data Application Area

Concept Aggregation Framework for Structuring Informational Aspects of Linked Open Data (CAF-SIAL) was proposed and developed to provide an easy to use interface for normal web users. Taking in account, the lack of application case studies exploiting the benefits of Linked Data in different domains, it was decided to apply the CAF-SIAL in different domains to bring Linked Data value in as well as introducing new potential case studies for Linked Data. Following research questions are investigated in the last part of Linked Data application area and made contribution in.

What are the potential application areas of CAF-SIAL Linked Data application? How aggregation, disambiguation, integration and presentation of Author/person related resources from external resources as a profile is performed and which external datasets are important?

In investigation of these research questions two application areas of CAF-SIAL are explored i.e. Digital Journal and Expertise Mining System.

Firstly, in Digital Journal domain, the authors are linked with the CAF-SIAL application, which in turn collect all the possible related information of authors from Linked Data resources and present that information in a profile. This contributed in highlighting the potentials for digital publishing system (e.g. J.UCS) by intelligent manipulation of the semantic search services and Linked datasets. This linking is provided at J.UCS website for journal users.

Secondly, Expertise Mining system is explored as another application area of CAF-SIAL. Calculated experts which are visualized in hyperbolic are further linked with CAF-SIAL to provide J.UCS administration with additional information in well structured profile. This linking contribution helped administration of journal in finding potential reviewers and experts, as well as provide leverage of expert profile from CAF-SIAL as an additional information source to help in assigning reviewing duties. For aggregation of person related information two linked datasets i.e. DBpeida and DBLP are used. For the disambiguation of ambiguous experts entities, heuristics are employed which function in very accurate way to yield with correct information. The mentioned research is now in prototype use for Journal of Universal Computer Science and proving very useful.

This linking is assessed to be helpful in digital journal environment for different scenarios:

- A comprehensive profile of an author is structured and visualized at one place. This is helpful in getting deep insights of authors work, personal and professional life.
- For users who are searching for research collaborators.

- For users who want to explore experts to seek guidance.
- For journal administration, in finding experts and assignment of reviewing duties to them.

12.2 Future Work

This section highlights future directions of the worked covered in this thesis. In future a major work is dedicated to the improvement in interface and introduction of new features in CAF-SIAL (Concept Aggregation Framework for Structuring Informational Aspects of Linked Open Data) application.
The future improvements are following:

First, to bring more interactivity in the CAF-SIAL application, the editing option with drag and drop is planned where users can move the specific property from one aspect to the other. This will help in built up of crowd-sourced profile's property database. This will subsequently bring users more close to the application and the quality of information will also improve. In addition to it users will be provided with option of 'Add information' where they can add any of their known facts to the profile.

Second, to cope with the limitation of CAF-SIAL which is reliability on the availability of data from DBpedia and DBLP, a caching mechanism is planned which will cache the retrieved data from above datasets and will provide users with content uninterrupted in case DBpedia or DBlP services are not running.

Third, a service is planned which can pre-emptively look for the rdf type person concept in the presented profile. This service will highlight the person type concepts to help visiting users in exploring other person profile in CAF-SIAL without need of search keyword.

Fourth, to build a trained dataset of properties which are generally used to represent a person type concept is planned. For this all the retrieved properties related to person from DBpedia will be trained with machine learning algorithm and in the end a generalized trained dataset will be established and will be make available to the community for use.

Fifth, to locate the missing personal information and to bring more variety of data into person, author and expert profile other datasets like FOAF (Friend of a Friend) and SIOC (Semantically Interconnected Online Communities) will be included into CAf-SIAL searching corpus. In addition for the Digital Journal Author and Experts publication record, other paper indexing services like IEEE, ACM and Semantic Web Conference Corpus present in Linked Data cloud will be used to get more published work by them.

12.3 Summary

This chapter presented a self-assessment view of the contributions made in answering the research questions asked in Section 1.3 and to discussed the set of goals during this research work. Finally, interesting future work and research directions were suggested in this chapter. Concluding, the goal of this research work was to investigate and get know the Linked Data sphere better by introducing new conceptual frameworks and to research the various aspects of Linked Data i.e. Linked Data publishing, Linked Data consumption and Linked Data application areas. This dissertation research allows the development of new applications and conceptual models to exploit Linked Data potentials as (e.g, Linked Data value chain, Linked Data consumption with creation of new simple user interface over Linked Data resources by using innovative URI location, aggregation and presentation technique and in publishing scientific community data as Linked Data), as well as introducing new domains (Digital journal, Expertise mining system) as Linked Data case studies.

Appendix A

D2R Server Mapping File for J.UCS dataset RDFization

```
@prefix map: <file:/c:/jdk1.5/bin/mapping_jucs.n3#>.
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix d2rq: <http://www.wiwiss.fu-berlin.de/suhl/bizer/D2RQ/0.1#> .
@prefix d2r: <http://sites.wiwiss.fu-berlin.de/suhl/bizer/d2r-server/
    config.rdf#> .
@prefix dc: <http://purl.org/dc/elements/1.1/> .
@prefix dcmi: <http://purl.org/dc/dcmitype/> .
@prefix dcterms: <http://purl.org/dc/terms/>.
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix swc: <http://data.semanticweb.org/ns/swc/ontology#>.
@prefix swrc: <http://swrc.ontoware.org/ontology#>.
@prefix vcard: <http://www.w3.org/2001/vcard-rdf/3.0#> .
@prefix wgs84_pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>.
@prefix geo: <http://www.geonames.org/ontology#>.
@prefix rev: <http://purl.org/stuff/rev#>.
@prefix sioc: <http://rdfs.org/sioc/ns#>.

map:db_liferay a d2rq:Database;
  d2rq:jdbcDriver ”com.mysql.jdbc.Driver”;
  d2rq:jdbcDSN ”jdbc:mysql://localhost/mashup_dbo1?autoReconnect=true”;
  d2rq:username ”root”;
  .
  # Table authorcontactinfo
map:authorcontactinfo a d2rq:ClassMap;
  d2rq:dataStorage map:db_liferay;
  d2rq:uriPattern "person/@@authorcontactinfo.ID@@";
  d2rq:class foaf:person;
  .
map:authorcontactinfo_lable a d2rq:PropertyBridge;
  d2rq:belongsToClassMap map:authorcontactinfo;
  d2rq:join "authors.ID=authorcontactinfo.AuthorID";
  d2rq:pattern "@@authors.Firstname@@ @@authors.Middlename@@ @@authors.
      Lastname@@";
  d2rq:property rdfs:label;
  d2rq:datatype xsd:string;
  .
map:authorcontactinfo_mbox a d2rq:PropertyBridge;
  d2rq:belongsToClassMap map:authorcontactinfo;
  d2rq:column "authorcontactinfo.Email";
  d2rq:property foaf:mbox;
    d2rq:datatype xsd:string;
  .
map:authorcontactinfo_firstname a d2rq:PropertyBridge;
  d2rq:belongsToClassMap map:authorcontactinfo;
  d2rq:column "authors.Firstname";
  d2rq:join "authors.ID=authorcontactinfo.AuthorID";
  d2rq:property foaf:firstname;
  d2rq:datatype xsd:string;
  .
map:authorcontactinfo_surname a d2rq:PropertyBridge;
  d2rq:belongsToClassMap map:authorcontactinfo;
  d2rq:join "authors.ID=authorcontactinfo.AuthorID";
  d2rq:column "authors.Lastname";
  d2rq:property foaf:surname;
  d2rq:datatype xsd:string;
  .
map:authorcontactinfo_paper a d2rq:PropertyBridge;
  d2rq:belongsToClassMap map:authorcontactinfo;
  d2rq:refersToClassMap map:papers;
  #d2rq:column "authorcontactinfo.ID";
  #d2rq:condition "authors.ID=authorcontactinfo.AuthorID";
  d2rq:join "authorcontactinfo.ID=papers_authors.AuthorContactID";
  d2rq:join "papers_authors.PaperID = papers.ID";
  d2rq:property foaf:made;
  .
    #Table keywords
map:keywords a d2rq:ClassMap;
  d2rq:dataStorage map:db_liferay;
  d2rq:uriPattern "keywords/@@keywords.ID|urlify@@";
  d2rq:class swrc:topic;
.
map:keywords_name a d2rq:PropertyBridge;
  d2rq:belongsToClassMap map:keywords;
  d2rq:column "keywords.Keyword";
  d2rq:property rdfs:label;
  d2rq:datatype xsd:string;
.
    # Table papers
map:papers a d2rq:ClassMap;
  d2rq:dataStorage map:db_liferay;
  d2rq:uriPattern "papers/@@papers.ID@@";
  d2rq:class swrc:Article;
  d2rq:class swc:Paper;
    d2rq:class foaf:Document;
  d2rq:class dcmi:Text;
  d2rq:class sioc:Item;
  .
map:papers_ID a d2rq:PropertyBridge;
  d2rq:belongsToClassMap map:papers;
  d2rq:column "papers.ID";
  d2rq:property dc:identifier;
  d2rq:datatype xsd:integer;
  .
map:papers_Title a d2rq:PropertyBridge;
  d2rq:belongsToClassMap map:papers;
  d2rq:column "papers.Title";
  d2rq:property dc:title;
  d2rq:property sioc:title;
  d2rq:property rdfs:label;
  d2rq:datatype xsd:string;
  .
map:papers_Keywords a d2rq:PropertyBridge;
    d2rq:belongsToClassMap map:papers;
  d2rq:refersToClassMap map:keywords;
    d2rq:join "papers.ID = papers_keywords.KeywordID";
  d2rq:join "papers_keywords.KeywordID = keywords.ID";
  d2rq:property dc:subject;
  d2rq:property foaf:topic;
  d2rq:property sioc:topic;
    .
map:papers_authors a d2rq:PropertyBridge;
    d2rq:belongsToClassMap map:papers;
    d2rq:refersToClassMap map:authorcontactinfo;
    d2rq:join "authorcontactinfo.ID=papers_authors.AuthorContactID";
    d2rq:join "papers_authors.PaperID=papers.ID";
  d2rq:property dc:creator;
  d2rq:property foaf:maker;
  d2rq:property swrc:author;
  d2rq:property sioc:has_creator;
  .
map:papers_rating a d2rq:PropertyBridge;
```

```
    d2rq:belongsToClassMap map:papers;
    d2rq:refersToClassMap map:ratingsentry;
    d2rq:join "papers.ID=ratingsentry.classPK";
  d2rq:property rev:hasReview;
  .
map:papers_SubmissionDate a d2rq:PropertyBridge;
  d2rq:belongsToClassMap map:papers;
  d2rq:column "papers.SubmissionDate";
  d2rq:property dc:dateSubmitted;
  d2rq:datatype xsd:date;
  .
map:papers_AcceptanceDate a d2rq:PropertyBridge;
  d2rq:belongsToClassMap map:papers;
  d2rq:column "papers.AcceptanceDate";
  d2rq:property dc:dateAcceptated;
  d2rq:datatype xsd:date;
  .
map:papers_JournalName a d2rq:PropertyBridge;
  d2rq:belongsToClassMap map:papers;
  d2rq:column "papers.JournalName";
  d2rq:property swrc:journal;
  d2rq:datatype xsd:string;
  .
map:papers_Issue a d2rq:PropertyBridge;
  d2rq:belongsToClassMap map:papers;
  d2rq:column "papers.Issue";
  d2rq:datatype xsd:integer;
  d2rq:property dc:issued;
  .
map:papers_Volume a d2rq:PropertyBridge;
  d2rq:belongsToClassMap map:papers;
  d2rq:column "papers.Volume";
  d2rq:datatype xsd:integer;
  d2rq:property swrc:volume;
  .
map:papers_PublishDate a d2rq:PropertyBridge;
  d2rq:belongsToClassMap map:papers;
  d2rq:column "papers.PublishDate";
  d2rq:property dc:date;
  d2rq:datatype xsd:date;
  .
map:papers_Number a d2rq:PropertyBridge;
  d2rq:belongsToClassMap map:papers;
  d2rq:column "papers.Number";
  d2rq:datatype xsd:integer;
  d2rq:property dc:SizeOrDuration;
  .
map:papers_URL a d2rq:PropertyBridge;
```

```
  d2rq:belongsToClassMap map:papers;
  d2rq:column ”papers.URL”;
  d2rq:property rdfs:seeAlso;
  d2rq:property owl:sameAs;
  .
map:papers_PDFURL a d2rq:PropertyBridge;
  d2rq:column ”papers.PDFURL”;
  d2rq:belongsToClassMap map:papers;
  d2rq:property rdfs:seeAlso;
  d2rq:property owl:sameAs;
  .
```

Listing A.1: D2r Mapping File

Appendix B

List of Publications

The work covered by this thesis led to following publications:
[**Latif et al. 2009**] Latif, A., Hoefler, P., Stocker, A., Ussaeed, A., Wagner, C. (2009). The Linked Data Value Chain: A Lightweight Model for Business Engineers. In: Proceedings of International Conference on Semantic Systems, pp. 568-576, Graz, Austria, 2-4, Sep. 2009.
[**Latif et al. 2009a**] Latif, A., Afzal, M.T., Ussaeed, A., Hoefler, P., Tochtermann, K. (2009). CAF-SIAL: Concept aggregation framework for structuring informational aspects of linked open data. In: Proceedings of International Conference on Networked Digital Technologies, pp. 100-105, Ostrava, Czech Republic, 28-31, Jul. 2009.
[**Latif et al. 2009b**] Latif, A., Tanvir, M.T., Hoefler, P., UsSaeed, A., Tochtermann, K.(2009). Turning keywords into URIs: simplified user interfaces for exploring linked data. In: Proceedings of 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human, Seoul, Korea, 24-26 Nov. 2009.
[**Latif et al. 2010**] Latif, A., Afzal, M.T., Ussaeed, A., Hoefler, P., Tochtermann, K. (2009). Harvesting Pertinent Resources from Linked Data. In Journal of Digital Information Management (JDIM) 8 (3), pp. 205-212, June 2010.
[**Latif et al. 2010a**] Latif, A., Afzal, M.T., Helic, D., Tochtermann, K., Maurer, H. (2010). Discovery and Construction of Authors' Profile from Linked Data (A case study for Open Digital Journal), Proceedings of the WWW2010 Workshop Linked Data on the Web (LDOW 2010), CEUR Workshop Proceedings. CEUR-WS.org (2010)
[**Latif et al. 2010b**] Latif, A., Afzal, M.T., Tochtermann, K. (2010). Constructing Experts Profiles from Linked Data, In: Proceedings of 6th IEEE International Conference on Emerging Technologies (ICET), pp. 33-38 , Islamabad, Pakistan, 18-19, Oct. 2010.

[**Latif and Afzal 2011**] Latif, A., Afzal, M. T.(2011). Linking Digital Journal Artefact (Authors) with Linked Data Resources, accepted and to appear in Journal of Universal Computer Science, 2011.

[**Latif and Afzal 2011a**] Latif, A., Afzal, M. T.(2011). Weaving Scholarly Legacy Data into Web of Data. accepted and to appear in Journal of Universal Computer Science, 2011.

[**Korica-Pehserl and Latif 2011**] Korica-Pehserl, P., Latif, A.(2011). Meshing Semantic Web and Web 2.0 technologies to construct Profiles: Case Study of Academia Europea Members. Accepted and to appear in Third International Conference on Networked Digital Technologies (NDT 2011), Macau, China , 11-13 July 2011.

[**Afzal and Latif 2011**] Afzal, M. T., Latif.(2011). Exploiting Tags-Citations Relationships to Discover Evolving Concepts from Social Bookmarking for Scientific Community. accepted and to appear in Journal of Universal Computer Science, 2011.

[**Afzal et al. 2009**] Afzal, M. T., Latif, A., Ussaeed, A., Sturm, P., Aslam, S., Andrews, K., Tochtermann, K., Maurer, H. (2009). Discovery and Visualization of Expertise in a Scientific Community. In: Proceeding of International Conference of Frontiers of Information Technology, Islamabad, Pakistan, 16-18, Dec. 2009.

[**UsSaeed et al. 2008a**] Us Saeed, A., Afzal, A., Latif, A., Stocker, A., Tochtermann, K. (2008). Does Tagging indicate Knowledge Diffusion? An Exploratory Case Study. In: Proceedings of International Conference on Convergence and Hybrid Information Technology, pp. 605 - 610, Busan, Korea, Nov. 11-13, 2008.

[**UsSaeed et al. 2008b**] Us Saeed, A., Afzal, M. T., Latif, A., Tochtermann, K. (2008). Citation rank prediction based on bookmark counts: Exploratory case study of WWW06 papers. In: Proceedings of IEEE International Mutitopic Conference, pp. 392-397, Karachi, Pakistan, Dec. 23-24, 2008.

[**UsSaeed et al. 2010**] Us Saeed, A., Afzal, M. T., Latif, A., Tochtermann, K. (2010). Disseminating knowledge through Tags: Recommending Tags for scientific resources. In Journal of IT in Asia, Vol 3 (2010), pp. 25-36, Issue Date: Nov 2010, Print ISSN: 1823-5042.

Bibliography

[**ACM-CCS 1998**] ACM (1998). ACM Computing Classification System.

[**About Google Scholar 2009**] About Google Scholar 2009 About Google Scholar. http://scholar.google.at/intl/en/scholar/about.html

[**Afzal et al. 2007**] Afzal, M. T., Kulathuramaiyer, N., Maurer, H. (2007). Creating Links into the Future, Journal of Universal Computer Science, 13 (9), pp. 1234-1245, 2007.

[**Afzal et al. 2008**] Afzal, M. T., Kulathuramaiyer, N., Maurer, H. (2008). Expertise Finding for an Electronic Journal, In: Proceedings of International Conference on Knowledge Management and Knowledge Technologies, pp. 436-440, Graz, Austria, 3-5, Sep. 2008.

[**Afzal 2009**] Afzal, M. T. (2009). Discovering Links into the Future on the Web, In: Proceedings of Fifth International Conference on Web Information Systems and Technologies, pp. 123-129, Lisbon, Portugal, 23-26, Mar. 2009.

[**Afzal et al. 2009**] Afzal, M. T., Latif, A., Ussaeed, A., Sturm, P., Aslam, S., Andrews, K., Tochtermann, K., Maurer, H. (2009). Discovery and Visualization of Expertise in a Scientific Community, In: Proceedings of International Conference of Frontiers of Information Technology, Islamabad, Pakistan, 16-18, Dec. 2009.

[**Afzal et al. 2009a**] Afzal, M. T., Maurer, H., Balke, W. T., Kulathuramaiyer, N. (2009). Improving Citation Mining, In: Proceedings of International Conference on Networked Digital Technologies, pp. 116-121, Ostrava, Czech Republic, 28-31, Jul. 2009.

[**Afzal et al. 2009b**] Afzal, M. T., Balke, W. T., Kulathuramaiyer, N., Maurer, H. (2009). Rule based Autonomous Citation Mining with TIERL, In Journal of Digital Information Management (JDIM) 8 (3), pp. 196-204, June 2010.

[**Afzal and Latif 2011**] Afzal, M. T., Latif. (2011). Exploiting Tags-Citations Relationships to Discover Evolving Concepts from Social Bookmarking for Scientific Community. accepted and to appear in Journal of Universal Computer Science, 2011.

[**Alexander et al. 2009**] Alexander, K., Cyganiak, R., Hausenblas, M., Zhao, J. (2009) Describing Linked Datasets. In: Proceedings of the Second Workshop on Linked Data on the Web (LDOW2009) at WWW2009, 2009.

[**Antoniou and van Harmelen 2004**] Antoniou, G., van Harmelen, F. (2004). A Semantic Web Primer, Cambridge MA: MIT Press, 2004.

[**Auer et al. 2007**] Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z. (2007). DBpedia: A Nucleus for a Web of Open Data, 6th International Semantic Web Conference, Busan, Korea (2007), `http://richard.cyganiak.de/2008/papers/dbpedia-iswc2007.pdf`

[**Auer et al. 2009**] Auer, S., Dietzold, S., Lehmann, J., Hellmann, S., Aumueller, D. (2009). Triplify - Lightweight Linked Data Publication from Relational Databases, In: Proceedings of International World Wide Web Conference (WWW 09), Madrid, Spain, pp. 621-630, 2009.

[**Becker and Bizer 2008**] Becker, C., Bizer, C. (2008). DBpedia Mobile - A Location-Aware Semantic Web Client. In: Proceedings of the Semantic Web Challenge at ISWC, 2008.

[**Beckett and Berners-Lee 2008**] Beckett, D., Berners-Lee, T. (2008). Turtle - Terse RDF Triple Language, W3C, 2008. Available at: `http://www.w3.org/TeamSubmission/turtle/`

[**Belleau et al. 2008**] Belleau, F., Nolin, M., Tourigny, N., Rigault, P., Morissette, J. (2008). Bio2RDF: Towards a mashup to build bioinformatics knowledge systems. Journal of Biomedical Informatics, Volume 41(5), pp. 706-16, 2008.

[**Berners-Lee et al. 1994**] Berners-Lee, T., Cailliau, R., Luotonen, A., Nielsen, H. F., Secret, A. (1994). The World-Wide Web. Communications of the ACM, 37(8): pp - 7682, 1994.

[**Berners-Lee 1998**] Berners-Lee, T.(1998). Semantic Web Road Map. W3C, 1998. Available at: `http://www.w3.org/DesignIssues/Semantic.html`

[**Berners-Lee et al. 1998**] Berners-Lee, T., Fielding, R., Irvine, U., L. Masinter. (1998). Uniform Resource Identifiers (URI): Generic Syntax. RFC 2396, IETFNetwork Working Group, 1998.

[**Berners-Lee 2000**] Berners-Lee, T.(2000). Semantic Web on XML - slide 10, W3C, 2000. Available at: http://www.w3.org/2000/Talks/1206-xml2k-tbl

[**Berners-Lee et al. 2001**] Berners-Lee, T., Hendler, J., Lassila, O. (2001). The Semantic Web, In Scientific American, pp- 2831, May 2001. http://www.sciam.com/2001/0501issue/0501berners-lee.html

[**Berners-Lee and Hendler 2001**] Berners-Lee, T., Hendler, J. (2001). Scientific publishing on the Semantic Web. Nature 410, 1023-1024, 26 Apr. 2001.

[**Berners-Lee 2006**] Berners-Lee, T. (2006). Linked Data – Design Issues, July 2006. http://www.w3.org/DesignIssues/LinkedData.html

[**Bizer and Cyganiak 2006**] Bizer, C., Cyganiak, R. (2008). D2R Server Publishing Relational Databases on the Semantic Web. Poster at the 5th International Semantic Web Conference(ISWC), Nov. 2006.

[**Bizer et al. 2007**] Bizer, C., Heath, T., Ayers, D., Raimond, Y. (2007). Interlinking Open Data on the Web, Demonstrations Track at the 4th European Semantic Web Conference, Innsbruck, Austria, May 2007. http://www.eswc2007.org/pdf/demo-pdf/LinkingOpenData.pdf

[**Bizer et al. 2009**] Bizer, C., Heath, T., Berners-Lee, T. (2009). Linked data the story so far, International Journal on Semantic Web and Information Systems (IJSWIS), 2009.

[**Bollacker et al. 2008**] Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J. (2008). Freebase: a collaboratively created graph database for structuring human knowledge, In: Proceedings of ACM SIGMOD international conference on Management of data, pp. 12471250, 2008.

[**Breslin et al. 2005**] Breslin, J. G., Harth, A., Bojars, U., Decker, S. (2005). Towards Semantically-Interlinked Online Communities. In: Proceedings of the Second European Semantic Web Conference, ESWC 2005, May 29- June 1, 2005, Heraklion, Crete, Greece, 2005.

[**Brickley and Miller 2004**] Brickley, D., Miller, L. (2004). FOAF Vocabulary Specification. Namespace Document 2 Sept. 2004, FOAF Project, 2004. http://xmlns.com/foaf/0.1/

[**Bush 1945**] Bush, V. (1945). As We May Think. The Atlantic Monthly, 176(1):101108, 1945.

[**Chen et al. 2007**] Chen, C., Maceachren, A., Tomaszewski, B., MacEachren, A. (2007). Tracing conceptual and geospatial diffusion of knowledge, Lecture Notes in Computer Science, 4564, pp.265-274, 2007.

[**Cheng et al. 2008**] Cheng, G., Ge, W., Qu, Y. (2008). Falcons: Searching and Browsing Entities on the Semantic Web. In: Proceedings of 17th International World Wide Web Conference, pp. 1101-1102, Beijing, China, 21-25, Apr. 2008.

[**Calude et al. 1994**] Calude, C., Maure, H., Salomaa, A. (1994). Journal of Universal Computer Science, In Journal of Universal Computer Science 0 (0), pp. 109-116, 1994.

[**Cameron et al. 2007**] Cameron, D., Aleman-Meza, B., Decker, S. L., Arpinar, I. B. (2007). SEMEF: A Taxonomy based Discovery of Experts, Expertise and Collaboration Networks. University of Georgia, LSDIS Lab, Technical Report, July 2007.

[**Candela et al. 2009**] Candela, L., Castelli, D., Fuhr, N., Ioannidis, Y., Klas, C.-P., Pagano, P., Ross, S., Saidis, C., Schek, H.-J., Schuldt, H., Springmann, M. (2006). Current Digital Library Systems: User Requirements vs Provided Functionality, Deliverable D1.4.1, Mar. 2006.

[**Catarci et al. 2007**] Catarci, T., F, Levialdi., M, S., Batini, C. (2007). Visual Query Systems for Databases: A Survey. Journal of Visual Languages and Computing, 8(2), pp. 215-260, 2007.

[**Coetzee et al. 2008**] Coetzee, P., Heath, T., Motta, E. (2008). Sparqplug: Generating linked data from legacy html, sparql and the dom. In: Proceedings of CEUR-WS Vol-369 of Linked Data on the Web (LDOW2008), Beijing, China, 2008. `http://data.semanticweb.org/workshop/LDOW/2008/paper/13`

[**Cowan et al. 2000**] Cowan, R., Paul, A. D., Foray, D. (2000). The Explicit Economics of Knowledge Codification and Tacitness, Industrial and Corporate Change, 9(2), pp.211-253, 2000.

[**DBLP 2009**] Digital Bibliography and Library Project, 2009. `http://www.informatik.uni-trier.de/~ley/db/`

[**Ding et al. 2004**] Ding, L., Finin, T., Joshi, A., Pan, R., S. Cost, R., Peng, Y., Reddivari, P., C. Doshi, V., Sachs, J. (2004). Swoogle: A Search and Metadata Engine for the Semantic Web. In: Proc. Thirteenth ACM Conference on Information and Knowledge Management, pp. 652 - 659, Washington, D.C., USA, 8-13, Nov. 2004.

[**Duerst and Suignard 2005**] Duerst, M., Suignard, M. (2005). Internationalized Resource Identifiers (IRIs). RFC 3987, IETFNetwork Working Group, 2005.

[**Engelbart 1962**] Engelbart, D, C. (1962). Augmenting human intellect: A conceptual framework. Technical report, SRI, 1962.

[**Fielding 1999**] Fielding, R. (1999). Hypertext transfer protocol http/1.1. request for comments: 2616, 1999. http://www.w3.org/Protocols/rfc2616/rfc2616.html

[**Fensel 2000**] Fensel, D. (2000). Ontologies: Silver Bullet for Knowledge Management and Electronic Commerce, 2000. http://citeseer.ist.psu.edu/413498.html

[**Giles et al. 1998**] Giles, C.L., Bollacker, K.D., Lawrence, S. (1998). CiteSeer: An Automatic Citation Indexing System. In: Proceedings of 3rd ACM Conference on Digital Libraries, ACM Press pp. 8998, Pittsburgh, 1998.

[**Gruber 1992**] Gruber, Thomas R. (1993). A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition, 5(2):199-220, 1993.

[**Glaser and Millard 2007**] Glaser, H. Millard, I. C. (2007). RKB explorer: Application and infrastructure. In: Proceedings of Semantic Web Challenge, 2007.

[**Harth et al. 2007**] Harth, A., Umbrich, J., Hogan, A., Decker. S. (2007). YARS2: A Federated Repository for Querying Graph Structured Data from the Web. In: Proceedings of the 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference, pp. 211-224, Busan, Korea, 11-15, Nov. 2007.

[**Hausenblas 2008**] Hausenblas, M. (2008). Building Scalable and Smart Multimedia Applications on the Semantic Web. PhD thesis, Graz University of Technology, 2008.

[**Hausenblas 2009**] Hausenblas, M. (2009). Exploiting Linked Data For Building Web Applications. IEEE Internet Computing, 2009.

[**Hausenblas 2009a**] Hausenblas, M. (2009a). Linked Data Applications. Technical Report, DERI, 2009.

[**Heath 2008**] Heath, T. (2008). How Will We Interact with the Web of Data?, IEEE Internet Com-puting, vol. 12, no. 5, pp. 88-91, 2008. http://tinyurl.com/ct3gx2

[**Heath 2008a**] Heath, T. (2008a). Information-seeking on the Web with Trusted Social Networks from Theory to Systems. PhD Thesis, The Open University, 2008.

[**Heath and Motta 2008**] Heath, T., Motta, E. (2008). Revyu: Linking reviews and ratings into the Web of Data. Journal of Web Semantics, 6(4): pp. 266-273, 2008.

[**Hepp et al. 2007**] Hepp, M., Siorpaes, K. and Bachlechner, D. (2007). Harvesting Wiki Consensus Using Wikipedia Entries as Vocabulary for Knowledge Management, IEEE Internet Computing, 11(5), pp.54-65, Sep. 2007.

[**Hildebrand et al. 2006**] Hildebrand, M., Ossenbruggen, V., Hardman, J. (2006). Facet: A Browser for Heterogeneous Semantic Web Repositories. In: Proceedings of International Semantic Web Conference, pp. 272-285, Athens, Georgia, USA, 5-9, Nov. 2006.

[**Hitchcock 2002**] Hitchcock, S. M. (2002). Perspectives in Electronic Publishing: Experiments with a New Electronic Journal Model. PhD thesis, University of Southampton. 18 Feb, 2002.

[**Holzinger et al. 2008**] Holzinger, A., Geierhofer, R., Modritscher, F. Tatzl, R. (2008). Semantic Information in Medical Information Systems: Utilization of Text Mining Techniques to Analyze Medical Diagnoses. Journal of Universal Computer Science, 14, 22, 3781-3795, 2008.

[**Hotho et al. 2006**] Hotho, A., Jaschke, R., Schmitz1, C., Stumme, G. (2006). Information Reterival in Folksonomies: Search and Ranking, LECTURE NOTES IN COMPUTER SCIENCE, 4011, pp.411-426, 2006.

[**Huang et al. 2008**] Huang, Y.C., Hung, C.C., Hsu, J.Y.: You Are What You Tag, in AAAI, 2008.

[**Hugh et al. 2009**] Hugh, G., Jaffri, A., Ian, M.(2009). Managing Co-reference on the Semantic Web, In WWW 2009 Workshop: Linked Data on the Web (LDOW2009), Madrid, Spain, 2009. http://eprints.ecs.soton.ac.uk/17587/

[**Hyperbolic Package 2009**] Hyperbolic Tree Library, 2009. `http://hypertree.cvs.sourceforge.net/viewvc/hypertree/hypertree/`

[**Jaffri et al. 2008**] Jaffri, A., Glaser, H., Millard, I. (2008). URI Disambiguation in the Context of Linked Data, Linked Data on the Web Workshop at the 17th International World Wide Web Conference, Beijing, China, 2008. http://events.linkeddata.org/ldow2008/papers/19-jaffri-glaser-uri-disambiguation.pdf

[**Kennedy and Shepherd 2005**] Kennedy, A., Shepherd, M. (2005). Automatic Identification of Home Pages on the Web, In: Proceedings of 38th Hawaii International Conference on System Sciences, 2005.

[**Kiefer et al. 2007**] Kiefer, C., Bernstein, A., Stocker, M. (2007). The fundamentals of iSparql a virtual triple approach for similarity-based Semantic Web tasks. In: Proceedings of the 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference, pp. 295-308, Busan, Korea, 11-15, Nov. 2007.

[**Kitzinger 1995**] Kitzinger, J. (1995). Qualitative research. Introducing focus groups. British Medical Journal, 311, pages 299-302, 1995.

[**Klyne et al. 2004**] Klyne, G., Carroll, J. J., McBride, B. (2004). RDF/XML Syntax Specification (Revised). W3C recommendation, World Wide Web Consortium, 2004.

[**Knezo 2006**] Knezo, G. J. (2006). Open Access Publishing and Citation Archives: Background and Controversy, 2006. www.ipmall.info/hosted_resources/crs/RL33023-061010.pdf

[**Kobilarov et al. 2009**] Kobilarov, G., Scott, T., Raimond, Y., Oliver, S., Sizemore, C., Smethurst, M., Lee, R.(2009). Media meets Semantic Web - How the BBC uses DBpedia and Linked Data to make Connections, In European Semantic Web Conference (ESWC 2009), Heraklion, Greece, 2009. http://www.georgikobilarov.com/publications/2009/eswc2009-bbc-dbpedia.pdf

[**Krottmaier 2003**] Krottmaier, H. (2003). Links to the Future, Journal of Digital Information Management, In Journal of Universal Computer Science 1 (1), pp. 3-8, 2003.

[**Krulwich and Burkey 1995**] Krulwich, B., Burkey, C. (1995). ContactFinder: Extracting Indications of Expertise and Answering Questions with Referrals, Technical Report. In the Working Notes of the Symposium on Intelligent Knowledge Navigation and Retrieval, AAAI Press, pp. 85-91. 1995.

[**Latif et al. 2009**] Latif, A., Hoefler, P., Stocker, A., Ussaeed, A., Wagner, C. (2009). The Linked Data Value Chain: A Lightweight Model for Business Engineers, In: Proceedings of International Conference on Semantic Systems, pp. 568-576, Graz, Austria, 2-4, Sep. 2009.

[**Latif et al. 2009a**] Latif, A., Afzal, M.T., Hoefler, P., UsSaeed, A., Tochtermann,K. (2009). Translating Keywords into URIS, In: Proceedings of 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human, ACM, Seoul, Korea, 24-26 Nov. 2009.

[**Latif et al. 2009b**] Latif, A., Afzal, M.T., Ussaeed, A., Hoefler, P., Tochtermann,K.: CAF-SIAL: Concept aggregation framework for structuring informational aspects of linked open data, In: Proceedings of International Conference on Networked Digital Technologies, pp. 100-105, Ostrava, Czech Republic, 28-31, Jul. 2009.

[**Latif et al. 2010**] Latif, A., Afzal, M.T., Ussaeed, A., Hoefler, P., Tochtermann, K. (2010). Harvesting Pertinent Resources from Linked Data, In Journal of Digital Information Management (JDIM) 8 (3), pp. 205-212, June 2010.

[**Latif et al. 2010a**] Latif, A., Afzal, M.T., Helic, D., Tochtermann, K., Maurer, H. (2010). Discovery and Construction of Authors' Profile from Linked Data (A case study for Open Digital Journal), Proceedings of the WWW2010 Workshop Linked Data on the Web (LDOW 2010), CEUR Workshop Proceedings. CEUR-WS.org, 2010.

[**Latif et al. 2010b**] Latif, A., Afzal, M.T., Tochtermann, K. (2010). Constructing Experts Profiles from Linked Data, In: Proceedings of 6th International Conference on Emerging Technologies (ICET), pp. 33-38 , Islamabad, Pakistan, 18-19, Oct. 2010.

[**Latif and Afzal 2011**] Latif, A., Afzal, M. T.(2011). Linking Digital Journal Artefact (Authors) with Linked Data Resources, accepted and to appear in Journal of Universal Computer Science, 2011.

[**Latif and Afzal 2011a**] Latif, A., Afzal, M. T.(2011). Weaving Scholarly Legacy Data into Web of Data. accepted and to appear in Journal of Universal Computer Science, 2011.

[**Korica-Pehserl and Latif 2011**] Korica-Pehserl, P., Latif, A.(2011). Meshing Semantic Web and Web 2.0 technologies to construct Profiles: Case Study of Academia Europea Members. Accepted and to appear in Third International Conference on Networked Digital Technologies (NDT 2011), Macau, China , 11-13 July 2011.

[**Liew and Foo 2001**] Liew, C.L., Foo, S. (2001). Electronic Documents: What Lies Ahead?, In: Proceedings of 4th International Conference on Asian Digital Libraries, pp 88-105, Banglore, India, 10-12, Dec. 2001.

[**Liu and Dew 2004**] Liu, P., Dew, P. (2004). Using Semantic Web Technologies to Improve Expertise Matching within Academia. In: Proceedings of the 2nd International Conference on Knowledge Management, pp. 370-378, Graz, Austria, June 30- July 2, 2004.

[**Manola and Miller 2004**] Miller, F., Miller, E. (2004). RDF Primer, W3C Recommendation, 2004. Available at: `http://www.w3.org/TR/rdf-primer/`

[**Marchionini and Maurer 1995**] Marchionini, G., Maurer, H. (1995). The roles of digital libraries in teaching and learning, Communication of the ACM, vol. 38, No. 4, pp. 67-75, 1995.

[**Marlow et al. 2006**] Marlow, C., Naaman, M., Boyd, M., Davis, M. (2006). HT06, tagging paper, taxonomy, Flickr, academic article, to read. In: Proceedings of the seventeenth conference on Hypertext and hypermedia, pp.31-40, Odense, Denmark, pp. 22-25, Aug. 2006.

[**Michael 2009**] Michael, L. (2009). DBLP - Some Lessons Learned. PVLDB 2(2), pp. 1493-1500, 2009.

[**Michlmayr and Cayzer 2007**] Michlmayr, E., Cayzer, S. (2007). Learning User Profiles from Tagging Data and Leveraging them for Personalized Information Access, In: Proceedings of 16th International World Wide Web Conference, Banff, Alberta, Canada, 8-12, May. 2007.

[**Mika 2005**] Mika, P. (2005). Ontologies Are Us: A Unified Model of Social Networks and Semantics. In: Proceedings of 4th International Semantic Web Conference, pp. 5-15, Galway, Ireland, 6-10, Nov. 2005.

[**Mockus and Herbsleb 2002**] Mockus, A., Herbsleb, J. A. (2002). Expertise Browser: A Quantitative Approach to Identifying Expertise. In: Proceedings of International Conference on Software Engineering, pp. 503-512, Orlando, Florida, 19-25, May. 2002.

[**Nelson 1982**] Nelson, T. (1982). Literary Machines. Eastgate Systems, 1982.

[**Obitko 2007**] Obitko, Marek. (2007). Translations between Ontologies in Multi-Agent Systems. Ph.D Thesis, Department of Cybernetics, Faculty of Electrical Engineering, Czech Technical University, 2007. `http://www.obitko.com/tutorials/ontologies-semantic-web/introduction.html`

[**Oren et al. 2008**] Oren, E., Delbru, R., Catasta, M., Cyganiak, R., Stenzhorn, H., Tummarello, G. (2008). Sindice.com: A Document-oriented Lookup Index for Open Linked Data. International Journal of Metadata, Semantics and Ontologies, 3(1), pp. 3752, 2008.

[**O'Reilly 2005**] O'Reilly, T. (2005). What is Web2.0, 2005. http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html

[**O'Reilly 2007**] O'Reilly, T. (2007). Freebase Will Prove Addictive. O'Reilly Radar, 2007. http://radar.oreilly.com/archives/2007/03/freebase-will-p-1.html.

[**Passant 2010**] Passant, A. (2010). dbrec - Music Recommendations Using DBpedia. In Book Series of The Semantic Web ISWC 2010. Lecture Notes in Computer Science, Volume: 6497, pp. 209-224, 2010.

[**Pipek et al. 2002**] Pipek, V., Hinrichs, J., Wulf, V. (2002). Sharing Expertise Challenges for Technical Support. In Ackerman, M./Pipek, V./Wulf, V. (eds): Beyond Knowledge Management: Sharing Expertise, pp. 111-136, MIT-Press, Cambridge MA, 2002.

[**Porter 1985**] Porter, M. E. (1985). Competitive advantage: Creating and sustaining superior performance, Free Press, 1985.

[**Postellon 2008**] Postellon, D. C. (2008). Hall and Keynes join Arbor in the citation indices. Nature, 452, pp. 282, 2008.

[**Raimond et al. 2008**] Raimond, Y., Sutton, C., Sandler, M. (2008). Automatic Interlinking of Music Datasets on the Semantic Web, In WWW 2008 Workshop: Linked Data on the Web (LDOW2008), Beijing, China, 2008. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.123.9753&rep=rep1&type=pdf

[**Raggett et al. 1999**] Raggett, D., Le Hors, A., Jacobs, I. (1999). Html 4.01 specification - w3c recommendation,1999. http://www.w3.org/TR/html401/

[**Roberts et al. 2001**] Roberts, R.J., Varmus, H.E., Ashburner, M., Brown, P.O., Eisen, M.B., Khosla, C., Kirschner, M., Nusse, R., Scott, M., Wold, B. (2001). Building A GenBank of the Published Literature. Science, 291 (5512), 2318-2319, 2001.

[**Rodriguez and Bollen 2008**] Rodriguez, M.A., Bollen, J. (2008). An Algorithm to Determine Peer- Reviewers. In: Proceedings of the 17th ACM conference on Information and Knowledge Management, pp. 319-328, Napa Valley, California, USA, 26-30, Oct. 2008.

[**Russell et al. 2008**] Russell, A., Smart, P. R., Braines, D. and Shadbolt, N. R. (2008). NITELIGHT: A Graphical Tool for Semantic Query Construction. In: Proceedings of Semantic Web User Interaction Workshop, Florence, Italy, 5, Apr. 2008.

[**Samur and Daniel 2009**] Samur, C. F., Daniel, S. (2009). Explorator: a tool for exploring RDF data through direct manipulation. In: Proceedings of Linked Data on the Web Workshop (LDOW). Madrid, Spain. 20, Apr. 2008.

[**Sauermann et al. 2008**] Sauermann, L., Cyganiak, R., Ayers, D., Vlkel, M. (2008). Cool URIs for the Semantic Web, W3C Interest Group Note, 2008. `http://www.w3.org/TR/2008/NOTE-cooluris-20081203/`

[**Servant 2008**] Servant, F.P. (2008). Linking Enterprise Data, Linked Data on the Web Workshop at the 17th International World Wide Web Conference, Beijing, China, 2008. `http://events.linkeddata.org/ldow2008/papers/21-servant-linking-enterprise-data.pdf`

[**Sizov 2007**] Sizov, S. (2007). What Makes You Think That? The Semantic Web's Proof Layer. Intelligent Systems, IEEE , vol.22, no.6, pp.94-99, Nov.-Dec. 2007

[**SPARQL endpoint**] Sparql Endpoint Description. http://esw.w3.org/SparqlEndpointDescription

[**Spoerri 2004**] Spoerri, A. (2004). RankSpiral: Toward Enhancing Search Results Visualizations. In: Posters Compendium, IEEE Symposium on Information Visualization, pp. 39-40, Austin, Texas, USA, 10-12, Oct. 2004.

[**Stankovic et al. 2010**] Stankovic, M., Wagner, C., Jovanovic, J., Laublet, P. (2010). Looking for Experts? What can Linked Data do for You?. In CEUR-WS Vol-628 Proceedings of the Linked Data on the Web Workshop (LDOW 2010), Raleigh, North Carolina, USA. April 27, 2010.

[**Suchanek et al. 2007**] Suchanek, F. M., Kasneci, G., Weikum, G. (2007). Yago: A Core of Semantic Knowledge - Unifying WordNet and Wikipedia. In: Proceedings of 16th International World Wide Web Conference, pp. 697-706, Banff, Alberta, Canada, 8-12, May. 2007.

[**SWEO Project 2007**] Semantic Web Education and Outreach Linked Open Data Project, 2007. `http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData`

[**TAG 2005**] W3C Technical Architecture Group TAG.httpRange-14: What is the range of the HTTP dereference function?, 2005. http://www.w3.org/2001/tag/issues.html

[**Tho et al. 2007**] Tho, Q.T., Hui, S.C., Fong, A.C.M. (2007). A Citation Based Document Retrieval System for Finding Research Expertise, Elsevier: Information Processing and Management, Issue 43, pp. 248-264, 1, Jan. 2007.

[**Tsai 2001**] Tsai, W. (2001). Knowledge Transfer in Intra-Organizational Networks: Effects of Network Position and Absorptive Capacity on Business Unit Innovation and Performance, Academy of Management Journal, 44(5), pp. 996-1004, 2001.

[**Turtle Graphics 2009**] Turtle Graphics, 2009. `http://www.gkrueger.com/java/aufgaben/loesung/TurtleGraphics.java`

[**Upstill et al. 2003**] Upstill, T., Craswell, N., Hawking, D. (2003). Query-Independent Evidence in Home Page Finding, ACM Transactions on Information Systems 21(3), pp. 286313, July 2003.

[**UsSaeed et al. 2008**] Us Saeed, A., Afzal, A., Latif, A., Stocker, A., Tochtermann, K. (2008). Does Tagging indicate Knowledge Diffusion? An Exploratory Case Study. In: Proceedings of International Conference on Convergence and Hybrid Information Technology, pp. 605 - 610, Busan, Korea, Nov. 11-13, 2008.

[**UsSaeed et al. 2008a**] Us Saeed, A., Afzal, M. T., Latif, A., Tochtermann, K. (2008). Citation rank prediction based on bookmark counts: Exploratory case study of WWW06 papers, In: Proceedings of IEEE International Mutitopic Conference, pp. 392-397, Karachi, Pakistan, Dec. 23-24, 2008.

[**UsSaeed et al. 2010**] Us Saeed, A., Afzal, M. T., Latif, A., Tochtermann, K. (2010). Disseminating knowledge through Tags: Recommending Tags for scientific resources. In Journal of IT in Asia, Vol 3 (2010), pp. 25-36, Issue Date: Nov 2010, Print ISSN: 1823-5042, 2010.

[**Volz et al. 2009**] Volz, J., Bizer, C., Gaedke, M., Kobilarov, G. (2009). SilkA Link Discovery Framework for the Web of Data. In: Proceedings of CEUR-WS Vol-538 of 2nd Linked Data on the Web Workshop (LDOW2009), Madrid, Spain, 2009. `http://events.linkeddata.org/ldow2009/papers/ldow2009_paper13.pdf`

[**Wu et al. 2006**] Wu, H., Zubair, M., Maly, K. (2006). Harvesting Social Knowledge from Folksonomies. In: Proceedings of the seventeenth conference on Hypertext and hypermedia, pp.111-114, Odense, Denmark, 22-25, Aug. 2006.

[**W3C Discussion 2009**] W3C Public Linked Data Discussion, 2009. `http://www.mailarchive.com/public-lod@w3.org/msg02032.html`

[**Yimam 1999**] Yimam, D. (1999). Expert Finding Systems for Organizations: Domain Analysis and the DEMOIR Approach. In: Proceedings of ECSCW 99 Beyond Knowledge Management: Management Expertise Workshop, pp. 276-283, Copenhagen, Denmark, 12-16, Sep. 1999.

TU Graz
Graz University of Technology
Senat

Deutsche Fassung:
Beschluss der Curricula-Kommission für Bachelor-, Master- und Diplomstudien vom 10.11.2008
Genehmigung des Senates am 1.12.2008

EIDESSTATTLICHE ERKLÄRUNG

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommene Stellen als solche kenntlich gemacht habe.

Graz, am …………………………… ………………………………………………..
(Unterschrift)

Englische Fassung:

STATUTORY DECLARATION

I declare that I have authored this thesis independently, that I have not used other than the declared sources / resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

…………………………… ………………………………………………..
date (signature)