Mateja Gyurica, BSc

Langzeitarchivierung von digitalen Geo - Dokumenten

MASTERARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieurin

Masterstudium Geomatics Science

eingereicht an der

Technischen Universität Graz

Betreuer

Dr. Ernst Primas

Institut für Geoinformation

Graz, Dezember 2015

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 entnommenen Stellen als solche kenntlich gemacht habe. Das in TUGRAZonline hochgeladene Textdokument ist mit der vorliegenden Masterarbeit identisch.

Datum

Unterschrift

Kurzfassung

Vorliegende Masterarbeit beschäftigt sich mit dem Prozess der Findung des geeignetsten Langzeitarchivformats von Geo - Dokumenten. Das Augenmerk wird dabei auf die Langzeitarchivierung mittels PDF/A (Portable Document Format/Archiving) gelegt. In Anlehnung daran wird in einem weiteren Teil der Arbeit ein Prozess für die Texterkennung mittels OCR (Optical Character Recognition) für gescannte Geo - Dokumente und deren abschließende Umwandlung in das Langzeitarchivformat PDF/A behandelt.

Dementsprechend bietet der erste Teil der Arbeit einen Einblick in die theoretischen Grundlagen der in dieser Masterarbeit behandelten Themenbereiche, während ein zweiter Teil den Aufbau, die Funktionsweise und die Durchführung des praktischen Teils beschreibt. Der praktische Teil umfasst dabei einerseits die Umsetzung eines PDF/A - Prozesses, bei dem mit Hilfe der gewählten Software, zwei verschiedene Varianten zur Findung der geeignetsten Archivvariante im Hinblick auf PDF/A untersucht werden und andererseits die Umsetzung eines OCR - Prozesses, bei dem nicht unmittelbar zugängliche Informationen aus den zuvor abgeleiteten PDF/A - Dokumenten herausgefiltert und diese Dokumente ebenfalls in ein für die Archivierung geeignetes Langzeitarchivformat mit Fokus auf PDF/A umgewandelt werden sollen. Ein letzter Teil der Arbeit geht auf die Anwendungen und die verschiedenen Einsatzgebiete von PDF/A ein, bei der vor allem klar gemacht werden soll, wie wichtig der Einsatz eines langzeitarchivfähigen Formates im Geo - Bereich ist bzw. auch das PDF/A - Format zunehmend an Bedeutung gewinnt.

Abstract

This master thesis describes the process of finding the most suitable format for geo - document long - term preservation with focus upon PDF/A (Portable Document Format/Archival). Based on this, a further part of the master thesis describes the process of OCR (Optical Character Recognition) for raster (scanned) geo - documents and their final conversion into the long - term preservation format PDF/A.

Therefore the first part of the master thesis provides an insight into the theoretical basics, whereas a second part deals with the practical questions referring the implementations, structures and functionalities. The main part of the practical section describes the realization of a PDF/A - process, with its two different ways of finding the most suitable long - term preservation format regarding PDF/A. Another part shows the implementation of the OCR - process, however not as detailed as the main process of PDF/A. With the OCR - process, document information that would not be immediately accessible from the resulting PDF/A - documents coming out of the main process, can be filtered out and made accessible. Furthermore these geo - documents can now be converted into the long - term preservation format PDF/A. A last part of the thesis deals with the use of PDF/A and shows different applications of PDF/A and also makes clear, how important the use of a long - term preservation format is in the geo - world as well as the PDF/A - format increasingly gains importance.

Inhaltsverzeichnis

1 Einleitung

Die Langzeitarchivierung von Geo - Daten und georelevanten Dokumenten gewinnt nachweislich an Bedeutung. Nicht nur in Österreich, sondern auch europaweit (z.B. Schweiz, Deutschland, Frankreich), sowie auch weltweit (z.B. USA) steht das Thema Langzeitarchivformate auf der Agenda, zumal der Zugriff auf Informationen nicht nur heute, sondern auch in Zukunft ermöglicht werden soll. Der Fokus bezieht sich dabei weitgehend darauf, dass Informationen gesammelt, gespeichert, archiviert und anschließend auch in 10, 50, 100 Jahren oder gar dauerhaft zugänglich sind (Drümmer et al., 2007; Oettler, 2013).

Die Frage, wieso ausgerechnet PDF/A (Portable Document Format/Archival, ISO - Norm 19005) als Archivformat verwendet werden soll, sei damit beantwortet, dass PDF/A sein visuelles Erscheinungsbild garantiert, denn alles was zur Anzeige des Dokuments notwendig ist, ist in diesem selbst erhalten und kann nicht durch die jeweilige Anzeigeplattform (Betriebssystem) verändert werden (Drümmer et al., 2007, S. 9; Oettler, 2013, S. 5). So können beispielsweise Textstellen, die eingebettete Schriftarten verwenden, diese wiederum zur Darstellung in einer anderen Betriebsumgebung heranziehen, ohne dass ein lokaler Ersatzfont gefunden werden muss (Drümmer et al., 2007, S. 46f).

Auch die Texterkennung und Volltextsuche spielen eine Rolle. Mit Hilfe von OCR (Optical Character Recognition) - Texterkennungsprogrammen werden gescannte Dokumente derart bearbeitet, sodass anschließend eine Volltextsuche in diesen ermöglicht werden kann. Vor allem bei Raster - Archivformaten wie z.B. TIFF (Tagged Image File Format) ist dies relevant, da es sich bei diesem ausschließlich um ein Dateiformat zur Speicherung von Bilddateien handelt, jedoch sollen auch in Rasterformaten nicht unmittelbar zugängliche Inhalte zugänglich gemacht werden (Drümmer et al., 2007, S. 7). Auch diese Dokumente sollen anschließend in die entsprechende, für die Langzeitarchivierung geeignetste Archivvariante mit Fokus auf PDF/A (ISO - Norm 19005) umgewandelt werden.

1.1 Über die Masterarbeit - Aufgabenstellung

Im Zuge der Masterarbeit sollen Fragen über die Langzeitarchivierung von Geo - Daten beantwortet werden. Dabei soll das Augenmerk auf die Langzeitarchivierung mittels PDF/A (Portable Document Format/Archival) gelegt und basierend auf die entsprechenden Normen und Standards (ISO - Norm 19005) die geeignetste Archivierungsvariante bzw. der geeignetste Langzeitarchivierungsprozess gefunden werden.

Der zweite Teil der Arbeit beschäftigt sich mit der OCR (Optical Character Recognition) - Texterkennung, bei der gescannte Dokumente mittels eines Texterkennungsprogramms durchsuchbar (z.B. Volltextsuche nach Text) gemacht werden sollen. Im Anschluss daran sind die Dokumente ebenfalls in die entsprechende, für die Langzeitarchivierung geeignetste Archivvariante mit Fokus auf PDF/A (ISO - Norm 19005) umzuwandeln.

1.2 Zielsetzung

Ziel der Masterarbeit ist es, anhand von verschiedenen, bereits eigenständig vorliegenden Geo - Dokumenten das geeignetste Format für die Langzeitarchivierung hinsichtlich PDF/A zu finden.

Dabei ist, abhängig vom Input - File (z.B. Raumordnungspläne, Katasterpläne, Bebauungspläne), eine Dateivorprüfung durchzuführen und anhand eines vordefinierten Prozesses gemäß ISO - Normen entweder mittels „Match & Fix“ - Variante (Variante mit höherem Analyse - Aufwand) zu analysieren oder mittels „Trial & Fix“ - Variante (Variante mit vorwiegend empirischem Ansatz) in die entsprechende, für die Langzeitarchivierung geeignetste Archivvariante mit Fokus auf PDF/A (ISO - Norm 19005) zu konvertieren.

Die Durchführung des Prozesses soll möglichst automatisiert erfolgen. Sollte im Zuge der Prozessdurchführung ein Eingriff in die Datei oder eine Konvertierung der Datei in eine der möglichen PDF/A - Varianten stattfinden, ist besonders darauf zu achten, dass dabei so wenig wie möglich an Information des jeweiligen Geo - Dokuments verloren geht bzw. auch die Originalität des Dokuments so gut wie möglich beibehalten werden kann.

In einem weiteren Schritt sollen gescannte Dokumente bzw. Ergebnis - Files aus dem Konvertierungsprozess mit Rasterinhalten mittels eines OCR (Optical Character Recognition) - Texterkennungsprogramms so bearbeitet werden, sodass im Rasterformat nicht unmittelbar zugängliche Inhalte zugänglich gemacht werden können (Volltextsuche). Auch diese Dokumente sollen anschließend in die entsprechende, für die Langzeitarchivierung geeignetste Archivvariante mit Fokus auf PDF/A (ISO - Norm 19005) umgewandelt werden.

1.3 Nicht - Ziele

Um die Masterarbeit überschaubarer zu machen, sei hier erwähnt, dass als Eingangsdaten im Prozess alle möglichen Datentypen (z.B. .pdf, .dxf, .jpg, .mdb) berücksichtigt, für die Findung der geeignetsten Archivierungsvariante jedoch ausschließlich PDF - und PDF/A - Dateien verwendet werden, die anhand der Prüfung auf deren Inhalte (vorgegebene Header bzw. Metadaten), aussortiert werden können. Der Schwerpunkt wird dabei auf Geo - Daten gerichtet, die als eigenständige Geo - Dokumente vorliegen. Im Zuge dessen werden daher Geo - Daten in Datenbanken, CAD - und GIS - Daten, sowie auch die Archivierung von Vektordaten im Rahmen der Masterarbeit nicht betrachtet.

Sollten im Zuge der Prüfung bereits PDF/A - Dateien vorliegen, werden diese im Rahmen der Durchführung aussortiert. Eine Umwandlung der bereits als PDF/A vorliegenden Datei in eine andere (höhere oder niedrigere) Version erfolgt in diesem Fall nicht, da sich je nach PDF/A - Variante und Konformitätsstufe andere Anforderungen an die jeweilige Version ergeben, was damit auch zu Informationsverlusten in der jeweiligen Datei führen könnte. Auch sollte an dieser Stelle erwähnt werden, dass jeglicher Eingriff oder Konvertierung einer Datei in eine andere (höhere oder niedrigere) Version eine Manipulation dieser bedeuten würde und damit jede Datei als „neue“ Datei behandelt werden müsste. Dies würde wiederum ei-

nem der Ziele der Masterarbeit widersprechen, welches besagt, dass einerseits so wenig wie möglich an Informationen im Zuge eines Eingriffs oder einer Konvertierung der Datei, in eine der möglichen PDF/A - Varianten, verloren gehen sollten, andererseits aber die Originalität möglichst beibehalten werden sollte, zumindest aber die Änderungen minimalst gehalten werden sollten. In den überwiegenden Fällen wird sich durch die Konvertierung das Dokument nicht ändern, dennoch ist eine Resultatprüfung gegenüber dem Original zum Zwecke der Qualitätssicherung sinnvoll.

Hingegen bei OCR (Optical Character Recognition) werden Dokumente bewusst manipuliert, mit dem Ziel, alle (verborgenen) Informationen in einer Datei auszuschöpfen (z.B. wäre ein Rastertext für eine Textsuche verborgener - also nicht unmittelbar zugänglicher - Text). Hier werden sowohl PDF - als auch PDF/A - Dokumente betrachtet. In diesem Fall wäre damit auch die Umwandlung dieser in eine (höhere oder niedrigere) Variante angemessen, da es hier mehr zu Informationsgewinnen als Verlusten kommen könnte. Dennoch sind sowohl Genauigkeitseinbußen als auch Fehler bei der Texterkennung nicht zu vermeiden (genaueres zur Funktionsweise von OCR kann Abschnitt 2.6 entnommen werden). Es sei hier erwähnt, dass OCR - Dokumente somit meist nicht den Kriterien der Langzeitarchivierung und Urkundenfähigkeit entsprechen, da durch den OCR - Prozess die Fehlermöglichkeiten durch falsche Zeicheninterpretationen wesentlich höher sind, als durch den PDF/A - Konvertierungsprozess, mit im Regelfall überhaupt keinen Änderungen.

2 Theoretische Grundlagen

Kapitel 2 beschäftigt sich mit den theoretischen Grundlagen der Langzeitarchivierung mittels PDF/A sowie der Texterkennung mittels OCR. Dem Leser soll damit ein Basiswissen über die in der Masterarbeit erarbeiteten Bereiche vermittelt werden.

2.1 Definition Geodaten

Als Geodaten seien all jene *„Daten mit direktem oder indirektem Bezug zu einem bestimmten Standort oder geografischem Gebiet“* (GeoDIG, 2010, §3 Abs. 1 Z 2 GeoDIG; EUR-Lex, 2007) bezeichnet, wobei mit einem direkten Bezug Objekte durch Koordinaten beschrieben werden, während sich ein indirekter Bezug auf ein administratives Gebiet (z.B. Stadt, Straße, PLZ) bezieht (GeoPortal Saarland, 2015).

Unter den Geodaten wird zwischen *Geobasisdaten* und *Geofachdaten* unterschieden. In Anlehnung an Stadt Graz (2015), Arbeitsgruppe der ARK AG ESys und des ARK IT-Ausschusses (2009), Land Steiermark - Amt der Steiermärkischen Landesregierung (2015) und GeoPortal Saarland (2015) seien diese wie folgt definiert:

- *Geobasisdaten* sind all jene Daten, die mittels Koordinaten beschrieben werden können und damit auch landschafts - und liegenschaftsbeschreibende Daten sind. Diese werden vorrangig im Vermessungs - und Katasterwesen eingesetzt. Darunter fallen beispielsweise amtliche Geodaten, Landschaftsmodelle, Höhenmodelle oder Bilder (Luft-/ Satelliten-/ Orthobilder).

- *Geofachdaten* sind alle Daten, die in verschiedenen Anwendungsgebieten mittels Raumbezug erhoben werden können. Darunter fallen z.B. Raumordnungsdaten, Flächennutzungspläne oder Verkehrs - und Klimatologiedaten.

Bartelme liefert eine allgemeinere Definition von Geodaten, wobei hier die Objekte der realen Welt (Natur) in einem räumlichen Bezugssystem betrachtet werden (Bartelme, 2005, S. 23; 212f; 337):

„Der Raumbezug stellt eines der wesentlichen Charakteristika von Geodaten dar. Fragen, in denen das Wo?, das Wie groß?, das Wie weit bis? eine entscheidende Rolle spielt, können nur dann fundiert beantwortet werden, wenn es ein für Fragende und Antwortende eindeutiges Bezugssystem gibt. Dies erlaubt es uns, Geodaten an einer bestimmten Stelle im Raum bzw. auf der Erdoberfläche zu fixieren. In Geoinformationssystemen heutiger Bauart übernehmen Koordinatensysteme diese Rolle“ (Bartelme, 2005, S. 212f).

Neben den Geodaten, seien auch *Metadaten*, als ein wichtiger Bestandteil dieser, genannt.

- *Metadaten* sind Daten über Daten, die Informationen über andere Daten enthalten. Es ist es auch möglich, Daten aufgrund vorgegebener Metadatenkriterien abzufragen (Bartelme, 2005, S. 37). Metadaten sind somit „*Informationen, die Geodatensätze*

und - dienste beschreiben und es ermöglichen, diese zu ermitteln, in Verzeichnisse aufzunehmen und zu nutzen“ (Österreichisches Parlament, 2015). Mit Verzeichnissen sind hier Archive (meist Langzeitarchive) der öffentlichen Verwaltung gemeint. Damit werden Metadaten zur Dokumentation, Verwaltung und Beschreibung der Geodaten, sowie auch für den Zugriff auf Geodaten benötigt (Wegner, 2000, S. 1).

Der Zusammenhang zwischen Metadaten und Geodaten ist in Abbildung 1 beispielhaft dargestellt. Dabei soll zum Vorschein gebracht werden, dass Geodaten durch Metadaten beschrieben werden können, sowie auch, dass Geodaten die räumliche Realität beschreiben können.

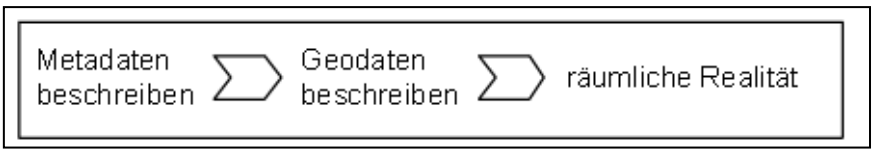

Abbildung 1: Zusammenhang zwischen Metadaten und Geodaten
(Quelle: Wegner, 2000, S. 2, zitiert nach Strobl, 1995, S. 276**)**

Eine zusätzliche Rolle unter den Geodaten spielen hier die sogenannten Modelltypen: *Vektordaten* und *Rasterdaten*. In Anlehnung an Bartelme (2005, S. 74ff und S. 127ff) seien diese wie folgt definiert:

- *Vektordaten* sind all jene Daten, die mittels punktförmiger, linienförmiger oder flächenförmiger (Polygon) Geometrie beschrieben werden können (z.B. zur Darstellung der Topologie eines Verkehrsnetzes).

- Bei *Rasterdaten* kann die Geometrie mittels Pixel (= Menge von Bildelementen, auch Rasterzelle oder Rastermasche genannt) beschrieben werden (so wie es z.B. in Digitalen Geländemodellen der Fall ist).

Die in dieser Masterarbeit verwendeten Geo - Daten seien all jene, die von verschiedenen Institutionen/Quellen übernommen wurden und bereits als eigenständige Geo - Dokumente vorliegen und nicht selbst (z.B. aus Datenbanken oder AutoCAD) generiert wurden. Das Augenmerk wird dabei vorwiegend auf eigenständige PDF - bzw. PDF/A - Dateien mit Geo - Inhalt gelegt.

Alle oben genannten Informationen wurden diversen Quellen entnommen. Aus Gründen der Übersichtlichkeit, wurden daher hier nur die wichtigsten Informationen herausgegriffen. Für nähere Erläuterungen sei der Leser an die angeführte Literatur bzw. Websites (siehe Abschnitt 2.1.1) verwiesen.

2.1.1 Weiterführende Literatur / Websites

Für detailliertere Informationen zum Begriff Geodaten, sei der Leser auf die nachstehende Literatur verwiesen, welche auch für diverse Recherchen im Zuge der Arbeit herangezogen wurden:

- Bartelme, N. (2005): *Geoinformatik - Modelle, Strukturen, Funktionen.*
- RIS - Rechtsinformationssystem des Bundes (www.ris.bka.gv.at, Stand: April 2015)
- EUR - Lex - Europäisches Recht (http://eur-lex.europa.eu, Stand: April 2015)
- GIS - Steiermark - Geoinformationssystem des Landes Steiermark (http://www.gis.steiermark.at/, Stand: April 2015)
- Österreichisches Parlament (http://www.parlament.gv.at, Stand: April 2015)
- AGIT - Angewandte Geoinformatik (http://www.agit.at, Stand: April 2015)
- GeoPortal Saarland (http://geoportal.saarland.de, Stand: April 2015)
- Geoportal Rheinland Pfalz (www.geoportal.rlp.de, Stand: April 2015)
- Bundesarchiv Deutschland (https://www.bundesarchiv.de, Stand: April 2015)

2.2 Definition Langzeitarchivierung

Die Langzeitarchivierung von Geo - Daten und georelevanten Dokumenten ist von großer Bedeutung, denn sie beschreibt die langfristige Aufbewahrung eines Dokuments und deren Inhalte. „*Kein Dokument darf verloren gehen, weder heute noch morgen und auch nicht in hundert Jahren*“ (Bundeskanzleramt Österreich, 2015).

Je nach Art des Dokuments werden vom Gesetzgeber bestimmte Aufbewahrungszyklen vorgeschrieben. Da die Langzeitarchivierung von Dokumenten eine Rolle spielt, seien folgend einige nationale (Tabelle 1) und internationale (Tabelle 2) Lösungen und deren ausgewählte gesetzlich vorgeschriebene Archivierungszeiträume aufgezählt.

Tabelle 1: Gesetzlich vorgeschriebene Archivierungszeiträume in Österreich

Österreich:	Aufbewahrungsdauer
Urkundenarchiv der Ziviltechniker:	
Urkundenarchivverordnung der Bundes-Architekten- und Ingenieurkonsulentenkammer, 2007 (www.arching.at/baik/) Gemäß §2 Abs.4 a-d, §5 Abs. 1 und §13 Abs. 1 und 3 Urkundenarchiv-verordnung: • Urkunden (wie z.B. Teilungspläne, Pläne zur Dokumentation örtlich begrenzter Dienstbarkeiten, Pläne für Vermessungswesen)	30 Jahre

sind für 30 Jahre aufzubewahren • Deren Speicherung erfolgt im Urkundenarchiv der bAIK • Urkunden und deren Beilagen sind in PDF/A - 1b zu archivieren	
Ziviltechnikergesetz (ZTG, 1993, www.ris.bka.gv.at): Gemäß §16 Abs. 1 ZTG (Ziviltechnikergesetz): • Urkunden sind (mit beigefügter Signatur) im Urkundenarchiv der Ziviltechniker (GOG - Gerichtsorganisationsgesetz) für zumindest 30 Jahre aufzubewahren	30 Jahre
Urkundenarchiv der Notare - cyberDOC:	
Urkundenarchivrichtlinien (UAR, 2007, www.notar.at): Gemäß Abs. 6 UAR (Urkundenarchivrichtlinie): • GOG - Urkunden und eigene Urkunden und Notariatsakten sind für die Dauer von 7 Jahren zu archivieren und freizugeben	7 Jahre
Urkundenarchiv der Rechtsanwälte - Archivium:	
Urkundenarchiv-RL, 2007 (http://www.rechtsanwaelte.at): Gemäß § 6 Abs. 1 und 2 Urkundenarchiv - RL: • Urkunden sind für die Dauer von 7 bzw. 30 Jahren im Archiv aufzubewahren	7 bzw. 30 Jahre
Geschäftsregister des BEV - Bundesamt für Eich - und Vermessungswesen (Digitales Katasterarchiv)	
BEV - Bundesamt für Eich- und Vermessungswesen **(Grundbuchs-Novelle, 2007, http://www.bev.gv.at):** Kommentare/Ministerialentwurf (§6 Abs. 1 und 2 Grundbuchs-Novelle, 2007): • Urkunden (wie z.B. Pläne, Handrisse, Bescheide, Grenzverhandlungsprotokolle, sonstige technische und schriftliche Unterlagen) sind auf unbegrenzte Dauer (im Sinne des Langzeitarchivs) aufzubewahren	Unbegrenzt

Tabelle 2: Gesetzlich vorgeschriebene Archivierungszeiträume in Deutschland

Deutschland:	Aufbewahrungsdauer
Notar- und Urkundenverzeichnis:	
Justizportal des Bundes und der Länder, 2015 (http://www.justiz.de/): • Notarielle Urkunden sind mindestens 100 Jahre aufzubewahren	100 Jahre
GBO-Grundbuchordnung, 2015 (http://www.rechtsportal.de): Gemäß §10 GBO (Grundbuchordnung): • Grundbücher und Urkunden sind vom Grundbuchamt dauernd aufzubewahren	Unbegrenzt

2.3 Archivdateiformate / Archivierungsformate

Folgende Abschnitte zeigen eine chronologische Entwicklung einiger gebräuchlicher Archivformate. Hierbei sollen, ausgehend von einer der ersten elektronischen Langzeitarchivvarianten wie etwa TIFF bis hin zu dem sich international durchsetzenden Standard des Langzeitarchivformats PDF/A, kurz vorgestellt werden.

2.3.1 TIFF (Tagged Image File Format)

Als eines der ersten elektronischen Langzeitarchivvarianten sei das TIFF (Tagged Image File Format) Format genannt. TIFF, als Format zur Speicherung von Bilddateien, wird bzw. wurde von vielen Institutionen für die Archivierung verwendet, zumal es die langfristige Reproduzierbarkeit garantiert. Da es sich aber ausschließlich um ein Rastergrafik - Format handelt, können Inhalte (mit Hilfe einer Volltextsuche) in einer TIFF - Datei nicht vorab durchsucht werden. Auch kann das TIFF - Format sehr viel Speicherplatz beanspruchen und ist überdies ein defacto (oder quasi bzw. Industrie) - Standard; also ein Standard der zwar weltweit genutzt wird, jedoch nicht offiziell von einer Normungsorganisation (z.B. ISO, CEN, DIN) beschlossen, genehmigt oder normiert wurde (Drümmer et al., 2007, S. 7; PDF Tools AG, 2009, S. 3; PDF Association, 2015a).

2.3.2 GeoTIFF (Geospatial Tagged Image File Format)

Die in der Geoinformation verwendete spezielle TIFF Variante, welche auch als Erweiterung dieser gilt, wird als GeoTIFF (Geospatial Tagged Image File Format) bezeichnet (Bartelme, 2005, S. 130). Diese erlaubt es, Rastergrafiken (wie z.B. Satelliten - oder Luftbilder) zusätzlich durch ein (Geo -) Referenzsystem zu beschreiben, um Positionen auf der Erdoberfläche eindeutig bestimmen zu können (Ruth, 2011). Auch in diesem Falle soll erwähnt werden, dass GeoTIFF ein Format zur Speicherung von Bilddateien ist und ebenso wie TIFF ein offengelegter Standard ist. Eine

Offenlegung selbst garantiert jedoch noch nicht, dass es sich um ein gesichertes Langzeitarchivformat handelt bzw. dass dies eine offiziell beschlossene, genehmigte oder normierte Norm ist, weshalb sich GeoTIFF (derzeit) nicht zur Langzeitarchivierung eignet.

2.3.3 PDF (Portable Document Format)

Um die Nachteile des TIFF Formats zu umgehen, wurde daher das Augenmerk auf das von Adobe Systems entwickelte Format PDF (Portable Document Format) gelegt. Dieses bietet zwar viele Vorteile gegenüber dem TIFF - Format (beispielsweise kann jegliche Information wie Video, Ton, Text in ein PDF Dokument verpackt werden) jedoch wird es aufgrund von Einschränkungen nicht als Langzeitarchivformat anerkannt, da Inhalte enthalten sein können, die das Dokument verändern können (Adobe Systems Incorporated, 2015; Oettler, 2013, S. 6). Sollte somit beispielsweise eine Schrift vorhanden sein, die der verwendete PDF - Reader in der Betriebsumgebung nicht führt (z.B. Microsoft Windows, Linux, Mac OS), wird automatisch von diesem ein Ersatzfont gefunden. Dies könnte unter anderem die Lesbarkeit eines Dokuments einschränken bzw. den Sinn des Inhalts ändern, in dem beispielsweise Zeichen oder Ziffern ausgelassen, anders dargestellt oder interpretiert werden (Drümmer et al., 2007, S. 46f).

Als Beispiel (Abbildung 2) sei hier das Eurozeichen, das Währungssymbol für den Euro, genannt, welches zeigen soll, dass die Anzeige eines Symbols durch die jeweilige Anzeigeplattform bzw. Betriebssystem verändert werden kann, falls die vorhandene Schriftart nicht im Dokument selbst eingebettet ist. Da in diesem Fall die Schriftart „FF Fago“ nicht den Anforderungen von PDF/A entspricht, ist damit auch hier keine exakte Darstellung des Zeichens möglich. An diesem Beispiel wird damit deutlich sichtbar, dass sich die beiden Zeichen, aufgrund der Findung eines Ersatzfonts, voneinander unterscheiden.

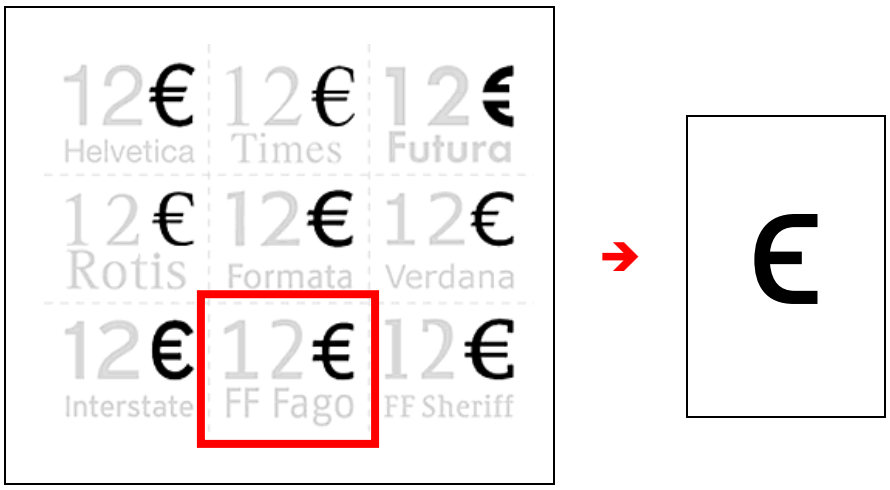

Abbildung 2: Beispiel zur Anzeige von Veränderungen eines Symbols aufgrund von Ersatzschriften (Quelle: http://de.wikipedia.org/wiki/Eurozeichen, Grafik bearbeitet im März 2015)

2.3.4 PDF/A (Portable Document Format/Archival)

Um die Einschränkungen der PDF - Variante zu überbrücken, wurde daher der auf PDF aufbauende Standard PDF/A (Portable Document Format/Archival) für die Archivierung entwickelt, mit welchem eine exakt reproduzierbare Darstellung von Inhalten und auch der Zugriff auf den gleichen Inhalt der Datei, wie zum Zeitpunkt der Erstellung gewährleistet wird. Damit garantiert PDF/A, dass ein Dokument das heute erstellt wird, auch in Zukunft und in allen Fällen genauso aussieht, wie zum Zeitpunkt der Erstellung (Drümmer et al., 2007, S. 8ff; Oettler, 2013, S. 5ff).

2.4 OAIS (Open Archival Information System)

Als Referenzmodell für die Archivierung sei hier OAIS (Open Archival Information System), ein in der ISO Norm 14721:2003 definierter Standard für die digitale Langzeitarchivierung genannt. OAIS bildet die Basis für all jene Standards, die sich mit der digitalen Archivierung beschäftigen, indem es das Grundgerüst für den Aufbau und die Funktionsweise eines digitalen Langzeitarchivs beschreibt (z.B. Revisionssicherheit, Langzeitarchivanforderungen, sicherer Input/Output). So können Langzeitarchivierungsstandards entwickelt werden, die dem Standard bzw. dem Referenzmodell OAIS genügen, unter anderem etwa das bereits zuvor erwähnte PDF/A - Format für die Langzeitarchivierung (Neuroth et al., 2010, S. 7).

Abbildung 3 zeigt den schematischen Aufbau dieses Modells, während Abbildung 4 die detailliertere Version der Komponenten dieses Modells darstellt. Um jedoch den Rahmen der Masterarbeit nicht zu sprengen, kann aus Gründen der Komplexität dieses Modells nicht näher darauf eingegangen werden. Für eine detailliertere Erläuterung sei der Leser auf die angeführte Literatur bzw. Websites (Abschnitt 2.4.1) verwiesen.

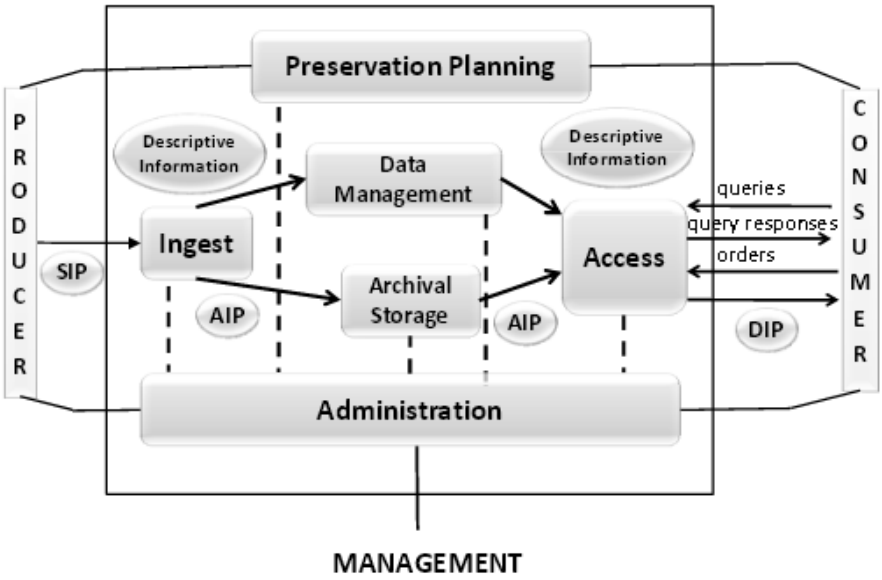

Abbildung 3: OAIS Überblicksmodell der Funktionseinheiten
(Quelle: http://public.ccsds.org/publications/archive/650x0m2.pdf, S. 4-1)

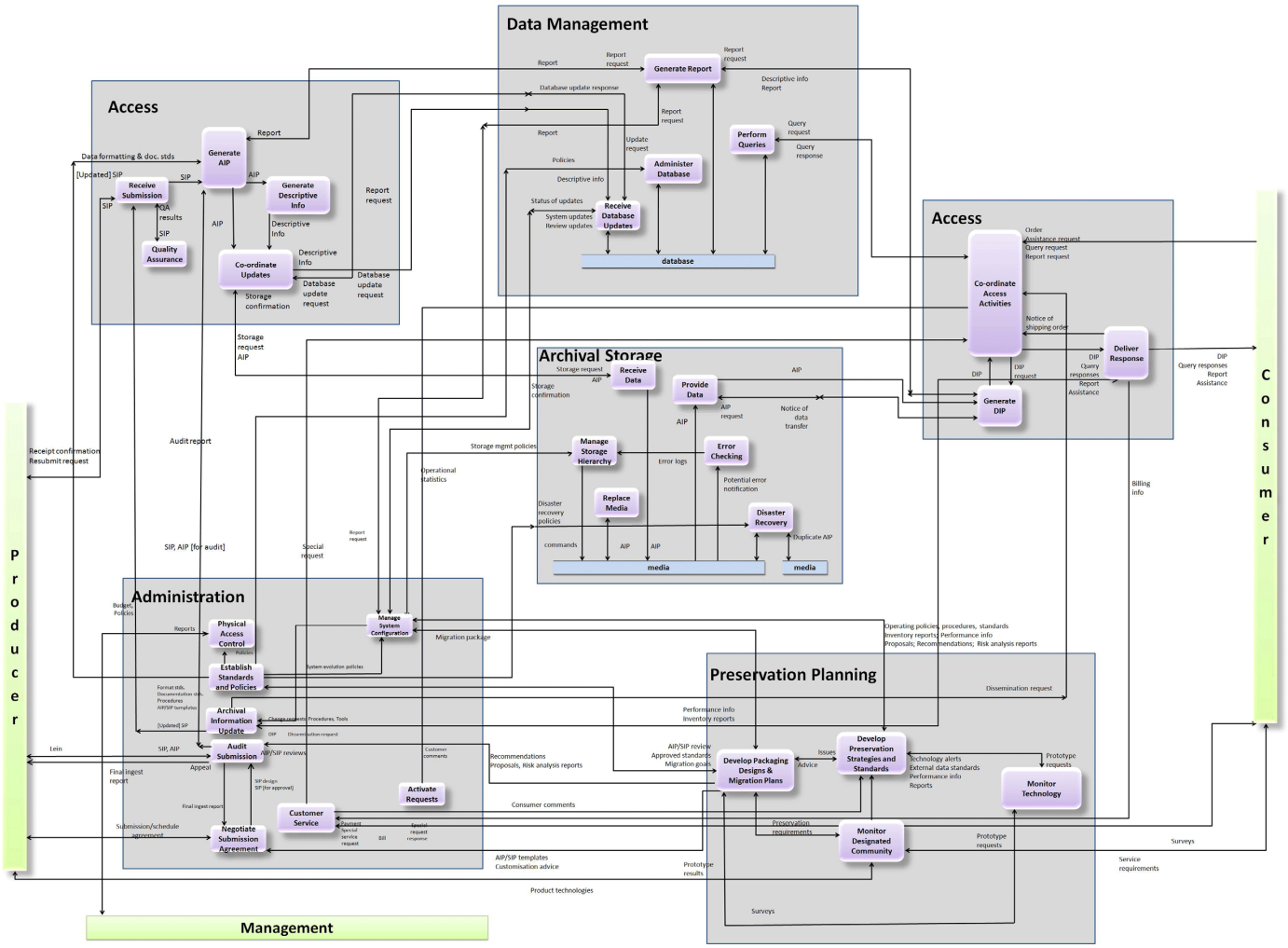

Abbildung 4: OAIS Detailmodell der Funktionseinheiten
(Quelle: https://archivengines.wordpress.com/tag/ccsds/)

2.4.1 Weiterführende Websites / Links

Für detailliertere Informationen zum OAIS Modell sei der Leser an dieser Stelle an die angeführte Literatur verwiesen:

- Nestor (http://www.langzeitarchivierung.de/Subsites/nestor/DE/Publikationen/Materialien/materialien_node.html, Stand: April 2015)
- The Consultative Committee for Space Data Systems - CCSDS (http://public.ccsds.org/publications/archive/650x0m2.pdf, Stand: April 2015)
- ArchivEngines (https://archivengines.wordpress.com/tag/ccsds/, Stand: April 2015)

2.5 Langzeitarchivierung mittels PDF/A

Dieser Abschnitt soll dem Leser einen Überblick über die Langzeitarchivierung mittels PDF/A liefern. Es sollen dabei wesentliche Aspekte über Definition, Zielen und verschiedene Versionen des Standards herausgegriffen werden.

2.5.1 Was ist PDF/A?

PDF/A ist ein in der ISO Norm 19005 definierter Standard für die Langzeitarchivierung von digitalen Dokumenten. Dabei handelt es sich um einen mehrteiligen Standard, eine Normreihe, mit dem Ziel, die Elemente und Inhalte eines PDF/A - Dokuments so zu beschreiben, sodass diese darauf aufbauend auch in Zukunft Informationen bewahren können und somit auch noch nach vielen Jahren eindeutig reproduzierbar sind. Zu beachten ist, dass die Normreihe lediglich die Eigenschaften der Inhalte eines solchen Dokuments beinhaltet, jedoch selbst keinen Ablaufmechanismus zur Archivierung beschreibt (Drümmer et al., 2007, S. 8ff; Oettler, 2013, S. 5).

PDF/A baut auf dem bekannten und weltweit verbreiteten PDF - Format von Adobe Systems auf (Oettler, 2013, S. 5). Bei PDF handelt es sich um einen offenen ISO - Standard (ISO 32000) für den Austausch von elektronischen Dokumenten. Neben den vielen Funktionen, die dieser Standard umfasst (z.B. einsetzbar unter nahezu allen gängigen Betriebssystemen, kostenlose Programme zur Anzeige, einfache Text - oder Metadatensuche innerhalb der Datei), basieren auf diesem auch speziell entwickelte PDF - Standards, unter anderem das eben erwähnte PDF/A - Format für die Langzeitarchivierung, sowie auch andere Standards wie etwa PDF/E (Portable Document Format/Engineering - technische Dokumente), PDF/X (Portable Document Format/Exchange) bzw. PDF/VT (Portable Document Format/Variable Data and Transactional Printing - Druckproduktion), PDF/H (Portable Document Format/Healthcare - Gesundheitswesen) und PDF/UA (Portable Document Format/Universal Accessibility - Barrierefreiheit) (Adobe Systems Incorporated, 2015; Oettler, 2013, S. 17; PDF Tools AG, 2015a).

Da PDF selbst keine (zukünftige) Software - und Hardwareunabhängigkeit und auch nicht die eindeutige Lesbarkeit in (zukünftigen) IT - Umgebungen gewährleisten kann, (PDF Tools AG, 2009, S. 5), jedoch die Langzeitarchivierung von Geodaten und georelevanten Dokumenten zunehmend an Bedeutung gewinnt, gibt es daher die Möglichkeit auf den von ISO entwickelten Standard PDF/A zuzugreifen, welcher insbesondere dazu entwickelt wurde, um die exakt reproduzierbare Darstellung von Inhalten und auch den Zugriff auf den gleichen Inhalt der Datei, wie zum Zeitpunkt der Erstellung, zu garantieren (Drümmer et al., 2007, S. 9).

Der derzeitige Stand der Technik gibt drei verschiedene Versionen (genaueres siehe Abschnitt 2.5.4) dieses Standards an (Oettler, 2013; ISO, 2005; 2011; 2012):

- PDF/A - 1 (ISO 19005 - 1:2005)
- PDF/A - 2 (ISO 19005 - 2:2011)

- PDF/A - 3 (ISO 19005 - 3:2012)

Diese unterteilen sich wiederum in verschiedene Untergruppen (genaueres siehe Abschnitt 2.5.4.4). Werden nun weitere Anforderungen an den Standard gestellt, wird der bereits bestehende grundsätzlich nicht widerrufen, geändert oder korrigiert, sondern es werden neue Versionen dieses Standard entwickelt, zumal verschiedene Workflows auch verschiedene Anforderungen haben (PDF Tools AG, 2014). Korrigenda werden nur in speziellen Fällen (z.B. bei Interpretationsspielräumen hinsichtlich technischer Umsetzung) zu einzelnen Normpunkten zwecks Präzisierung veröffentlicht. Hinsichtlich der Kompabilität ist zu erwähnen, dass die niedrigeren Versionen immer auch den Anforderungen der höheren Versionen genügen. Ein PDF/A - 1b - Dokument genügt beispielsweise den Anforderungen von PDF/A - 2b, hingegen kann ein PDF/A - 2 - Dokument, wenn es die zugelassene Funktionalität von PDF/A - 2 voll ausschöpft (z.B. Transparenzen), die Normpunkte von PDF/A - 1 überschreiten.

2.5.2 Warum PDF/A?

Neben den vielen bereits erwähnten Aspekten, sei hier basierend auf Drümmer et al. (2007, S. 9) die wesentliche Zielsetzung des ISO Komitees wie folgt definiert:

„Ziel des PDF/A-Standards ist, dass PDF-Dokumente erstellt werden können, deren visuelles Erscheinungsbild über die Zeit erhalten bleibt. Dabei sollen diese Dateien unabhängig sein von Software und Systemen zu Herstellung, Speicherung und Reproduktion.“

Weiters soll hier ein Überblick über die wesentlichen Vorteile für die Nutzung von PDF/A gegeben sein. Es soll darauf hingewiesen werden, dass das behandelte Thema einen sehr großen Bereich umfasst und daher folgend nur die wichtigsten Punkte herausgegriffen wurden. Für die Zusammenfassung dieser wurden Drümmer et al. (2007), Oettler (2013), PDF Tools AG (2014) und PDF Association (2014; 2015b) herangezogen.

- Internationaler ISO - Standard:
  - PDF/A (ISO 19005) ist ein Standard für die Langzeitarchivierung von digitalen Dokumenten, mit dem Ziel, dass der Inhalt von Dokumenten (z.B. Darstellung, Struktur, eingebettete Objekte, Metadaten) gleich dargestellt werden soll, wie zum Erstellungszeitpunkt.
  - Die international gültige PDF/A - Normreihe (PDF/A - 1 (ISO 19005 - 1:2005), PDF/A - 2 (ISO 19005 - 2:2011) und PDF/A - 3 (ISO 19005 - 3:2012)) beschreibt die Eigenschaften der Inhalte eines PDF/A - Dokuments, die erfüllt werden müssen, um Dokumente auch noch nach vielen Jahren eindeutig reproduzieren zu können, jedoch keinen Ablaufmechanismus zur Archivierung.

- Metadaten:
  - Unter Metadaten versteht man Daten über Daten (z.B. Informationen über ein Referenzellipsoid, welches Grundlage für eine kartographische Abbildung ist; Autoreninformation zu einem archivierten Dokument; nähere Signaturinformationen zu einer signierten Urkunde).
  - Mit Metadaten ist es möglich, die Inhalte eines Dokuments zu beschreiben.
  - Diese können direkt in ein PDF/A - Dokument integriert und auch mit dem Metadaten - Standard XMP (Extensible Metadata Platform), das Metadaten verschiedenster Art vereint, festgehalten werden.
  - Ein (einziges) Metadatenfeld, welches automatisch bei der Erstellung eines PDF/A - Dokuments durch die verwendete Konvertierungssoftware hinzugefügt wird, ist Pflicht. Bei diesem Feld handelt es sich um die PDF/A - Kennung, welche angibt, in welcher PDF/A - Version und Konformitätsstufe das Dokument vorliegt.

- Volltextsuche:
  - Eine Volltextsuche ermöglicht dem Anwender verschiedenste Informationen (beispielsweise einen Text nach einem bestimmten Wort) in einem Dokument zu suchen.
  - Auch PDF/A bietet die Möglichkeit einer Volltextsuche innerhalb eines Dokuments an, wobei Rasterinformationen im Dokument nicht durchsucht werden können.

- Speicherplatz:
  - Da beispielsweise eine einfache Bilddatei sehr viel Speicherplatz einnehmen kann, wird das Speicherplatzproblem bei PDF/A durch diverse Bildkompressionen (z.B. JPEG, JBIG2, seit PDF/A - 2 auch JPEG2000) und Farbtrennungsverfahren gelöst.
  - Dabei spielt nicht nur die Kompressionsart, sondern auch die Kompressionsstufe eine wesentliche Rolle: je größer die Kompressionsstufe, desto unschärfer wird das Bild bzw. der Text. Dies könnte vor allem bei OCR - Verfahren problematisch sein, denn hier könnten beispielsweise Zeichen verschmelzen (i wird zu l)

- Digitale Signaturen:
  - PDF/A - Dokumente lassen sich mit einer digitalen Signatur versehen, welche der Zertifizierung des Dokuments dient.
  - Digitale Signaturen dienen dementsprechend auch als Beweis, dass das Dokument nach dem Signieren nicht verändert bzw. manipuliert wurde.
  - Aus kryptographischen Sicherheitsgründen sollten Signaturen jedoch durch regelmäßiges Nach - oder Resignieren aufgefrischt werden.

- Gültigkeit von PDF/A - Dokumenten:
  - Um die Gültigkeit von Dokumenten zu gewährleisten, werden in absehbaren Abständen die Inhalte des PDF/A - Standards vom ISO Komitee weiterentwickelt.
  - Bestehende Standards werden dabei nicht geändert, korrigiert oder widerrufen, sondern in weiteren Fassungen des Standards erweitert.

- Barrierefreiheit:
  - PDF/A soll für Menschen mit Behinderung, wie etwa Sehbehinderung oder Einschränkung im motorischen Bereich, zugänglich sein, indem Dokumente barrierefrei gemacht werden.
  - Dabei sollen Struktur und Lesereihenfolge mittels Tagged PDF (= strukturiertes PDF) eines Dokuments eindeutig sein, sodass beispielsweise ein Screenreader (= Vorlesesoftware, wie z.B. kostenlose Adobe Reader) einen Text im Dokument korrekt vorlesen kann, sowie auch Inhalte von Grafiken und Bildern, die mittels Tagged PDF realisiert wurden, in Worten beschrieben werden können.

- Breite Akzeptanz:
  - Die Langzeitarchivierung mittels PDF/A findet bereits in vielen Bereichen und Einsatzgebieten Anwendung (z.B. Gesundheitswesen, Gerichtswesen, Behörden, Verwaltung, Regierung) und setzt sich nicht nur auf nationaler, sondern auch internationaler Ebene immer mehr durch, sei es, um Dokumente langfristig aufzubewahren oder um einen Austausch in einem einheitlichen Format zu ermöglichen.
  - Zudem gibt es auch zahlreiche Organisationen, Anwender und Entwickler, die mit ihrem Wissen einen Beitrag zum Thema PDF/A leisten. Als Beispiel sei hier die Organisation PDF/A Competence Center (http://www.pdfa.org/) genannt, welche an der Entwicklung und Umsetzung rund um PDF/A stark beteiligt ist.

- Plattformunabhängigkeit:
  - PDF/A ist plattformunabhängig und damit auch kompatibel mit unterschiedlicher Hard - und Software.
  - Somit kann ein PDF/A - Dokument auf verschiedenen Medien (z.B. unterschiedliche Betriebssysteme wie Microsoft Windows, Mac OS, Linux) originalgetreu dargestellt werden. Denn alles was zur Anzeige des Dokuments notwendig ist, ist in diesem selbst enthalten (z.B. eingebettete Schriftarten).
  - Da digitale Dokumente langfristig aufbewahrt werden sollen, sich der Stand der Technik jedoch immer weiterentwickeln kann, ist die Platt-

formunabhängigkeit von besonderer Bedeutung, zumal das Lesen von PDF/A - Dateien auch in Zukunft ermöglicht werden soll.

2.5.3 Validierung und Konvertierung

Um sicher zu gehen, dass es sich bei der vorliegenden Datei (auch wenn ein PDF/A - Flag gesetzt ist) wirklich um eine der ISO - Norm entsprechende, gültige PDF/A - Datei handelt, ist es notwendig diese zu validieren bzw. auf PDF/A - Konformität gemäß dem Standard zu überprüfen, bevor diese im Archiv abgelegt werden kann. Dabei sind grundlegende Anforderungen, sowie auch verbotene Inhalte bzw. Eigenschaften einer solchen Datei für die jeweilige PDF/A - Version zu überprüfen. Validierung ist insbesondere deshalb notwendig, weil ein gesetzter PDF/A - Flag aufgrund der unterschiedlichen Qualität der Erstellungssoftware (von Shareware bis hin zu renommierten Herstellern) noch nicht für die 100% - ige Einhaltung der ISO - Norm bürgt. Validierung ist aber auch deswegen notwendig, weil PDF/A - Dateien nicht gegen Manipulation geschützt werden können bzw. dürfen (in PDF/A ist keine Verschlüsselung oder Passwortvergabe erlaubt) (Drümmer et al., 2007, S. 36).

2.5.4 PDF/A - Versionen

PDF/A - Versionen bezeichnen verschiedene Versionen des PDF/A - Standards, welche einen Einblick geben, wie Inhalte in einem Dokument beschaffen sein müssen. Je nach Anforderung an den Workflow, werden neue Versionen des Standards entwickelt, wobei ältere Versionen keinem Update unterliegen, sondern in ihrer Art und Weise weiterbestehen und damit keinem Verfallsdatum unterliegen (PDF Tools AG, 2014).

Der derzeitige Stand der Technik umfasst drei Versionen, die in den folgenden Abschnitten beschrieben werden. Ferner soll Tabelle 3 die wesentlichen Unterschiede dieser Versionen zeigen.

Alle folgenden Informationen wurden Drümmer et al. (2007), Oettler (2013) und PDF Tools AG (2011; 2012) entnommen. Aus Gründen der Übersichtlichkeit, seien hier nur die wichtigsten Informationen daraus genannt. Für nähere Erläuterungen sei der Leser an die angeführte Literatur bzw. Websites (siehe Abschnitt 2.5.5) verwiesen.

2.5.4.1 PDF/A - 1

Die PDF/A - 1 Version, als erste Version des PDF/A - Standards, wurde mit der PDF - Version 1.4 eingeführt und unterteilt sich in zwei Untergruppen (PDF/A - 1a und - 1b, wobei nähere Details zu ihrer Bedeutung Abschnitt 2.5.4.4 entnommen werden können).

In dieser Version werden grundlegende Anforderungen, sowie auch verbotene Inhalte bzw. Eigenschaften hinsichtlich einer PDF/A - Datei dieser Version beschrieben. So müssen hier beispielsweise alle verwendeten Bilder, Grafiken und Schrift-

zeichen im Dokument eingebettet sein oder auch XMP (Extensible Metadata Platform) für die Integration von Metadaten im Dokument verwendet werden, während z.B. Transparenzen, gewisse Kompressionsarten oder PDF - Ebenen untersagt sind.

PDF/A - 1 bildet das Grundgerüst des PDF/A - Standards, da sich der Stand der Technik jedoch ständig weiterentwickelt, zeigt diese Version damit Einschränkungen auf, die in weiteren Versionen zu überwinden versucht werden.

2.5.4.2 PDF/A - 2

Die PDF/A - 2 Version stellt eine Weiterentwicklung der PDF/A - 1 Version dar und ist darauf gerichtet, Einschränkungen der ersten Version zu überbrücken. Somit erlaubt diese Version z.B. die Verwendung von verschiedenen Kompressionsarten wie JPEG2000, Transparenzen oder auch PDF - Ebenen. Darüber hinaus erlaubt PDF/A - 2 andere Dokumente einzubetten, wobei hier darauf geachtet werden muss, dass es sich beim eingebetteten Dokument um eine gültige PDF/A - Datei handelt.

Ebenso wie die erste Version, unterteilt sich auch die zweite Version in verschiedene Untergruppen (PDF/A - 2a, - 2b und - 2u), und basiert auf der PDF Version 1.7. Nähere Informationen bezüglich Untergruppen können Abschnitt 2.5.4.4 entnommen werden.

2.5.4.3 PDF/A - 3

Die auf PDF/A - 2 aufbauende dritte Version, PDF/A - 3, unterscheidet sich von dieser nicht all zu viel, jedoch gibt es bezüglich Einbettung von Dokumenten eine erhebliche Änderung. So können bei dieser dritten Version beliebige Dokumentenformate (wie z.B. Microsoft Word oder Excel) eingebettet werden, die nicht notwendigerweise ein gültiges PDF/A - Format darstellen müssen, jedoch zur Sicherheit ein solches ebenfalls eingebettet werden sollte. Sollte der Benutzer in ferner Zukunft mit einem eingebetteten Dokument arbeiten wollen und sich der Stand der Technik nicht geändert haben, so ist es möglich mit dem Originaldokument zu arbeiten. Sollte sich der Stand der Technik jedoch geändert haben, kann der Benutzer das zur Sicherheit zusätzlich eingebettete, in PDF/A konvertierte Dokument zum Arbeiten verwenden.

Auch an dieser Stelle soll erwähnt werden, dass PDF/A - 3 auf der PDF Version 1.7 basiert und sich ebenfalls in verschiedene Untergruppen (PDF/A - 3a, - 3b und - 3u) unterteilt, welche im folgenden Abschnitt näher beschrieben werden.

2.5.4.4 Konformitätsstufen: a, b, u

Die oben genannten PDF/A - Versionen gliedern sich in verschiedene Untergruppen, welche als Konformitätsstufen bezeichnet und in drei Bereiche eingeteilt werden: a, b und u.

Je nach Aufgabenstellung und Verwendungszweck können Dokumente in einer der verfügbaren Version, sowie auch Konformitätsstufe im Archiv abgelegt werden.

Basierend auf Drümmer et al. (2007), Oettler 2013 und PDF Association (2011) seien die Konformitätsstufen wie folgt definiert, wobei hier mit der am einfachsten zu erstellenden Stufe begonnen werden soll:

- *Stufe b* (basic):
  - Die Basic Variante garantiert die eindeutige visuelle Reproduzierbarkeit der Inhalte.
  - Dies bedeutet, dass eine mit dieser Stufe erstellte PDF/A - Datei, auch in ferner Zukunft genauso aussehen würde, wie zum jetzigen Zeitpunkt der Erstellung.
  - Allerdings sagt diese nichts über die inhaltliche, logische Struktur eines Dokuments aus und ermöglicht auch nicht die 100% - Textextraktion oder - durchsuchbarkeit in einem Dokument.
  - Stufe b lässt sich in allen drei Versionen des PDF/A - Standards finden.

- *Stufe a* (accessible oder advanced):
  - Stufe a, als eine in der Regel mit höherem Aufwand zu erstellende Stufe, beinhaltet jegliche Anforderungen des Standards.
  - Darunter fallen alle Informationen über die eindeutige, inhaltliche und logische Struktur eines Dokuments. So müssen hier alle Informationen bzw. Elemente, die sich in einem Dokument befinden (wie z.B. Dokumenten - Titel, Paragraphen oder Bilder), mittels Tagged PDF (= strukturiertes PDF) integriert werden. Dieses gibt dabei die Reihenfolge der Anordnung der Inhalte in einem Dokument an (Abbildung 5).
  - Solch eine Realisierung ist vor allem dann wichtig, wenn z.B. Informationen aus einem Dokument abgeleitet werden sollen, sodass beispielsweise die Struktur und Lesereihenfolge eines Dokuments durch einen Screenreader (= Vorlesesoftware, wie z.B. kostenlose Adobe Reader) korrekt wiedergegeben werden kann bzw. bei Bildern anhand des Tags erkenntlich gemacht werden kann, wann oder wo diese aufgenommen wurden bzw. was sich auf diesen befindet.
  - Ferner ist der erweiterte Zugriff auf alle Inhalte (z.B. Text, Bild), eindeutige visuelle Reproduzierbarkeit, sowie auch die Abbildbarkeit von Text im Schriftzeichenstandard „Unicode“ einer in dieser Stufe erstellten PDF/A - Datei von Nöten.
  - Diese Variante wurde bereits mit PDF/A - 1 eingeführt und ist in allen drei Versionen des PDF/A - Standards zu finden.

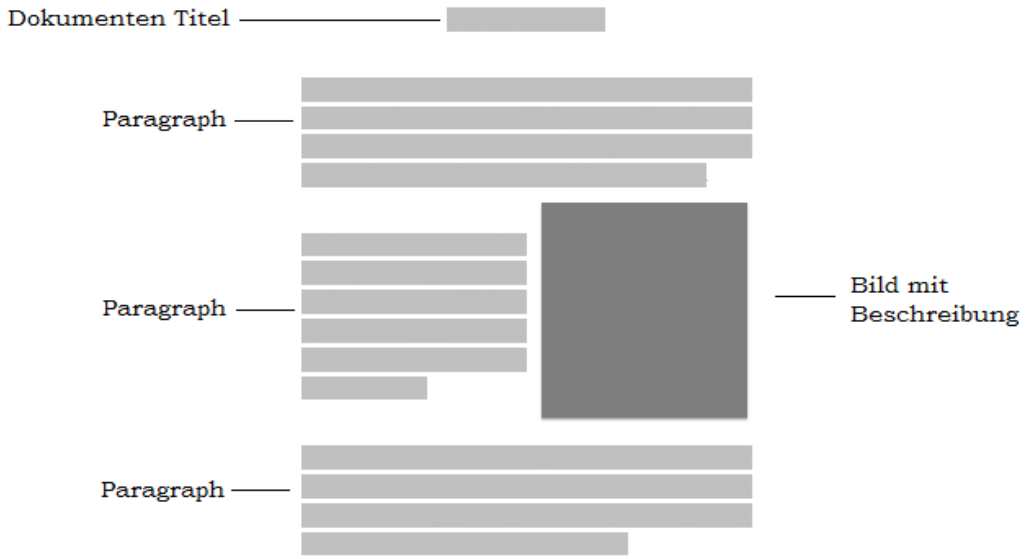

Abbildung 5: Beispiel eines strukturierten Inhalts der Stufe a

- *Stufe u* (unicode):
  - Bei dieser Variante handelt es sich lediglich um eine Erweiterung der Stufe b, wobei der gesamte Text und damit alle Fonts eines Dokuments im Schriftzeichenstandard „Unicode“ korrekt kodiert sein müssen, um hier das Durchsuchen und Extrahieren (z.B. von Text) zu ermöglichen.
  - Der Unicode selbst ist ein internationaler Standard der besagt, dass sich Zeichen eindeutig einer Unicode - ID zuordnen lassen müssen, um den weltweiten Austausch, Verarbeitung und Anzeige von Texten zu ermöglichen (Unicode Inc, 2014).
  - Die Unicode - Variante wurde erst mit PDF/A - 2 eingeführt und ist auch in PDF/A - 3 integriert.

2.5.4.5 Tabellenübersicht - Vergleich verschiedener Versionen

Tabelle 3 soll eine Übersicht über die wesentlichen Unterschiede der erforderlichen und untersagten Anforderungen der verschiedenen PDF/A - Versionen und deren Konformitätsstufen geben. Da das behandelte Thema einen sehr großen Bereich umfasst, werden hier nur die wichtigsten Unterschiede herausgegriffen. Alle folgenden Informationen wurden Drümmer et al. (2007), Oettler (2013) und PDF Tools AG (2009; 2011; 2012) entnommen, sowie auch mit dem Internationalen ISO - Standards (ISO, 2005; 2011; 2012) verglichen.

Tabelle 3: Vergleich verschiedener PDF/A - Varianten

	PDF/A Version:	- 1a	- 1b	- 2a	- 2b	- 2u	- 3a	- 3b	- 3u
Features:									
Transparenz		✗	✗	✓	✓	✓	✓	✓	✓
PDF – Ebenen		✗	✗	✓	✓	✓	✓	✓	✓
Einbettung von:									
Signaturen		✓	✓	✓	✓	✓	✓	✓	✓
Dokumenten (PDF/A)		✗	✗	✓	✓	✓	✓	✓	✓
Dokumenten (sonstige, nicht notwendigerweise PDF/A)		✗	✗	✗	✗	✗	✓	✓	✓
Video - und Audiodaten		✗	✗	✗	✗	✗	✗	✗	✗
Bilder, Grafiken, Schriften		✓	✓	✓	✓	✓	✓	✓	✓
Text:									
Durchsuchbarkeit		✓	~	✓	~	✓	✓	~	✓
Extraktion		✓	~	✓	~	✓	✓	~	✓
Kompressionen (Bild):									
JPEG		✓	✓	✓	✓	✓	✓	✓	✓
JPEG2000		✗	✗	✓	✓	✓	✓	✓	✓
LZW		✗	✗	✓	✓	✓	✓	✓	✓
Verweise (Hyperlinks)		✓	✓	✓	✓	✓	✓	✓	✓
Metadaten *(Verwendung von XMP für Dokumenten - Management)*		✓	✓	✓	✓	✓	✓	✓	✓
Aktionen & JavaScript		✗	✗	✗	✗	✗	✗	✗	✗
Verschlüsselung mit einem Passwort		✗	✗	✗	✗	✗	✗	✗	✗
Alternative Bilder *(für niedriger aufgelöste Bildschirmdarstellung)*		✗	✗	✗	✗	✗	✗	✗	✗

✓ = erlaubt ✗ = untersagt, ~ = nicht zu 100% möglich, da Schriftzeichen nicht eindeutig einer Unicode - ID zugeordnet werden müssen

2.5.5 Weiterführende Websites / Links

Detailliertere Informationen zum PDF/A - Standard und den verschiedenen Versionen und Konformitätsstufen können auch unter nachstehender Literatur bzw. Websites, welche auch für diverse Recherchen im Zuge der Arbeit herangezogen wurden, entnommen werden:

- PDF Association (http://www.pdfa.org/, Stand: April 2015)
- PDF Tools AG (http://www.pdf-tools.com/, Stand: April 2015)
- callas software GmbH (http://www.callassoftware.com, Stand: April 2015)
- Four Pees NV (http://support.fourpees.com, Stand: April 2015)
- ISO - International Organization for Standardization (http://www.iso.org, Stand: Juni 2015)

2.6 Texterkennung mittels OCR

Hauptziel dieser Masterarbeit ist die Findung des geeignetsten Formats für die Langzeitarchivierung von Geo - Dokumenten, unter der Voraussetzung, möglichst wenig Information im Zuge einer Konvertierung in dieses Format (z.B. durch Elimination verbotener Inhalte für ein entsprechendes Langzeitarchivformat) zu verlieren. Dabei sollte auch der Originalinhalt möglichst erhalten bleiben. Es stellt sich aber hier die Frage, ob nicht unmittelbar zugängliche Informationen (z.B. Rastertext) eines Langzeitarchiv - Dokuments durch entsprechende Nachbearbeitung zugänglich gemacht werden können und dieses Dokument sozusagen als Zusatz - Dokument mit Mehrwert, neben dem Original - Archivdokument, welches bestehen bleibt, behandelt werden und im Archiv abgelegt werden kann.

Der zweite Teil der Arbeit beschäftigt sich hierzu mit der Texterkennung mittels OCR (Optical Character Recognition). Auch hier soll dem Leser ein Überblick über Definition und Funktion dieser speziellen Variante gegeben werden. Da die Texterkennung mittels OCR nicht Kernthema der Masterarbeit ist, wird hier nur bis zu einem gewissen Grad darauf eingegangen. Vielmehr soll hier das grundlegende Prinzip dieses Verfahrens wiedergegeben werden. Abschließend sollen jene Schritte dokumentiert werden, um eine Umwandlung in das Langzeitarchivformat PDF/A zu ermöglichen. Damit sollen die aus den langzeitarchivierten Originaldokumenten abgeleiteten OCR - Dokumente, neben den Originaldokumenten, ebenfalls im Archiv abgelegt werden können.

2.6.1 Grundlagen

OCR ist ein Verfahren für die automatische bzw. optische Text - und Zeichenerkennung bei gescannten Dokumenten. Darunter fallen beispielsweise alle gescannten Dokumente, die in verschiedenen Formaten wie z.B. PDF, PDF/A, Microsoft Word oder als Grafikformate (Bilddateien) wie z.B. JPEG oder TIFF, vorliegen (Solid Documents, 2015a; 2015b; intarsys consulting GmbH, 2015a). Im Rahmen der Masterarbeit wird das Augenmerk dabei ausschließlich auf PDF - bzw. PDF/A - Dateien gelegt, wobei je nach

Norm auch andere eingebettete Dokumente (z.B. JPEG, TIFF, BMP, PNG) in diesen erlaubt werden.

Beim OCR - Verfahren werden vorwiegend gescannte Dokumente oder Bilddateien auf Zeichen untersucht und anschließend wiederum in solche konvertiert, um alle darin nicht unmittelbar zugänglichen Inhalte zugänglich machen zu können. Das grundlegende Prinzip der Texterkennung beruht dabei darauf, dass alle sich auf dem (gescannten) Dokument befindenden Elemente zunächst in einzelne Ebenen aufgespalten werden. Die verwendete Software erkennt dabei, in welcher Ebene sich „Textelemente“ und in welcher sich „grafische Elemente“ befinden. Innerhalb der Ebene „Textelemente“ werden einzelne Absätze, Zeilen, Wörter und Zeichen von der Software erkannt. Diese Ebene wird mittels der optischen Zeichenerkennung (OCR) so bearbeitet, dass die darin befindlichen Zeichen erkannt und interpretiert werden, um somit beispielsweise das Verschmelzen von Zeichen (z.B. i wird zu l) weitgehend zu vermeiden. Hierbei kommen auch verschiedene Methoden und Algorithmen zum Einsatz, wobei an dieser Stelle nicht näher auf diese eingegangen wird. Die beiden Ebenen werden in einem letzten Schritt wieder vereinigt bzw. komprimiert. Abschließend wird, je nach Wahl des Ausgabeformats (z.B. .txt, .doc, .rtf, .pdf), ein digitales Dokument erstellt, das im Anschluss daran weiterbearbeitet werden kann (wie z.B. Durchsuchen oder Extrahieren von Text) (Tanner, 2004; Booth & Jeremy, 2006; Vorbach, 2014; Computer Bild, 2009a).

Im Rahmen der Masterarbeit, sollen die zuvor genannten Dokumente (im PDF - bzw. PDF/A - Format) mittels dem OCR - Texterkennungsverfahren derart bearbeitet werden, sodass auch bei vorhandenen Rasterformaten nicht unmittelbar zugängliche Inhalte zugänglich gemacht werden können und damit eine Volltextsuche in diesen ermöglicht wird. Diese Dokumente sollen anschließend in die entsprechende, für die Langzeitarchivierung geeignetste Archivvariante mit Fokus auf PDF/A (ISO - Norm 19005) umgewandelt werden.

Die wesentlichen Vorteile bzw. möglichen Probleme, die sich bei der Durchführung der Texterkennung ergeben könnten, seien in Anlehnung an Solid Documents (2015a; 2015b), intarsys consulting GmbH (2015a) und Cognitive Technologies (2015) wie folgt definiert:

a) *Vorteile*:

- Je nach Art der Eingangsdaten, können mit Hilfe von OCR gescannte Dokumente so bearbeitet werden, sodass anschließend eine Volltextsuche in diesen ermöglicht werden kann. Es können auch bei vorhandenen Rasterformaten im Dokument, alle nicht unmittelbar zugänglichen bzw. unstrukturierten Inhalte zugänglich gemacht werden.
- Für die Erstellung eines digitalen Dokuments können je nach Bedarf und Aufgabenstellung verschiedene Ausgabeformate (z.B. .txt, .rtf, .pdf, .doc) gewählt werden.

- Die Wahl einer Sprache, in welcher das Dokument bearbeitet werden soll, ermöglicht die gezielte Bearbeitung des Dokuments hinsichtlich automatischer Buchstaben - bzw. Worterkennung.

b) Mögliche Probleme:

- Je nach Wahl der Sprache, in welcher das Dokument bearbeitet werden soll, werden verschiedene Zeichen oder Ziffern richtig interpretiert, andere wiederum nicht.
  - Als Beispiel seien hier Umlaute (ä, ö, ü) im deutschsprachigen Raum genannt.
  - Sollte das Dokument in einer nicht - deutschsprachigen Sprache übersetzt werden, ergeben sich diesbezüglich Probleme.
  - Damit kann die Lesbarkeit eines Dokuments stark eingeschränkt werden, indem Zeichen und Ziffern anders dargestellt oder fehlerhaft interpretiert werden bzw. könnten auch Lücken in einem Text entstehen, wenn Zeichen nicht korrekt interpretiert oder sogar ausgelassen werden.
- Zur Erstellung eines digitalen Dokuments stehen verschiedene Ausgabeformate (z.B. .txt, .rtf, .doc, .pdf) zur Verfügung.
  - Eine direkte Umwandlung in eine entsprechende Langzeitarchivvariante mit Fokus auf PDF/A ist jedoch nicht ohne weiteres möglich, da bereits in einem ersten Schritt bei diesem Verfahrens (Spaltung der Inhalte des Dokuments in verschiedene Ebenen) ein Eingriff in die Datei stattfindet und damit auch zu Änderungen der Inhalte in der Datei führt.

2.6.2 Weiterführende Websites / Links

Für detailliertere Informationen zur Texterkennung mittels OCR sei der Leser an dieser Stelle an die nachstehende Literatur bzw. Websites, welche auch für Recherchen zur Hilfe herangezogen wurden, verwiesen:

- intarsys consulting GmbH (http://www.intarsys.de/, Stand: April 2015)
- Solid Documents (http://www.soliddocuments.com/, Stand: April 2015)
- Cognitive Technologies (http://www.cognitiveforms.com/, Stand: April 2015)
- PDF Association (http://www.pdfa.org/, Stand: April 2015)
- PDF Tools AG (http://www.pdf-tools.com/, Stand: April 2015)

3 Prozessbeschreibungen

In diesem Kapitel sollen dem Leser die Prozesse zur Durchführung der PDF/A - Konvertierung bzw. OCR - Texterkennung näher gebracht werden. Weiters wird hier auf die verwendeten Applikationen eingegangen.

Bei der Durchführung der Prozesse wird außerdem besonderes Augenmerk darauf gelegt, dass diese möglichst automatisiert ausgeführt werden können und somit so wenig wie möglich manuell eingegriffen werden muss. Es sei hier allerdings erwähnt, dass im Rahmen der Durchführung dieser Prozesse keine Programmierung erfolgt, sondern eine bestmögliche Einbindung der zur Verfügung stehenden Applikationen (gegebenenfalls Adaptierung der Softwareparameter). Zwischenschritte, die nicht automatisiert erfolgen können (da sie die Applikation nicht bietet), werden somit manuell umgesetzt, mit dem Hinweis, wie diese optimiert (= automatisiert) werden könnten.

3.1 Ablaufdiagramm - PDF/A

Um den Aufbau und die Funktionsweise des PDF/A - Prozesses näher zu bringen, sei hier einleitend ein Konzept vorgestellt, welches die wesentlichen Aspekte zur Durchführung darstellen soll.

Da das Ziel der Masterarbeit die Findung des geeignetsten Formats für die Langzeitarchivierung hinsichtlich PDF/A ist, stellt sich die Umsetzung dieses Prozesses als wichtig heraus, denn dieser soll zudem die Grundlage für einen Automatisierungsprozess sein, bei dem im Zuge der Durchführung so wenig wie möglich in eine Datei eingegriffen werden soll, sowie auch so wenig wie möglich an Informationen im Zuge der Konvertierung der Datei in eine PDF/A - Variante verloren gehen sollte, da die Beibehaltung der Originalität der Datei im Vordergrund steht. Sollte beispielsweise eine Datei einer höheren PDF/A - Version (z.B. PDF/A - 2) in eine niedrigere PDF/A - Version (z.B. PDF/A - 1) konvertiert werden, könnten in diesem Fall bestimmte Eigenschaften (z.B. das Löschen von Transparenzinformationen mittels Autokorrektur der Konvertierungssoftware, da diese in PDF/A - 1 nicht erlaubt sind) verloren gehen, die es allerdings hier zu vermeiden gilt.

Die folgenden Abbildungen (Abbildung 6, Abbildung 7 und Abbildung 8) zeigen den schematischen Aufbau dieses Konzepts, während im folgendem Abschnitt (Abschnitt 3.1.1) näher auf die einzelnen Schritte eingegangen wird.

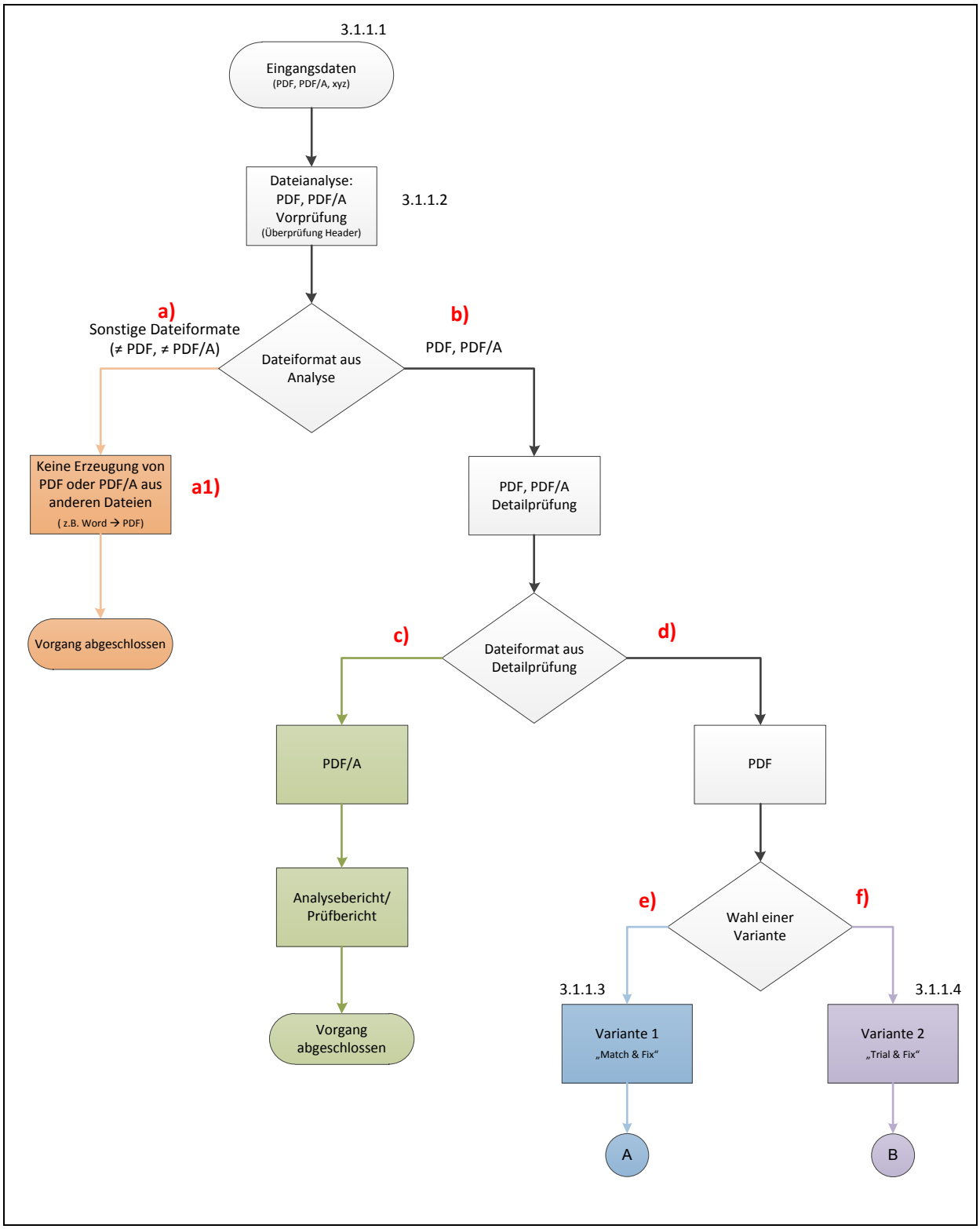

Abbildung 6: Konzept PDF/A - Prozess

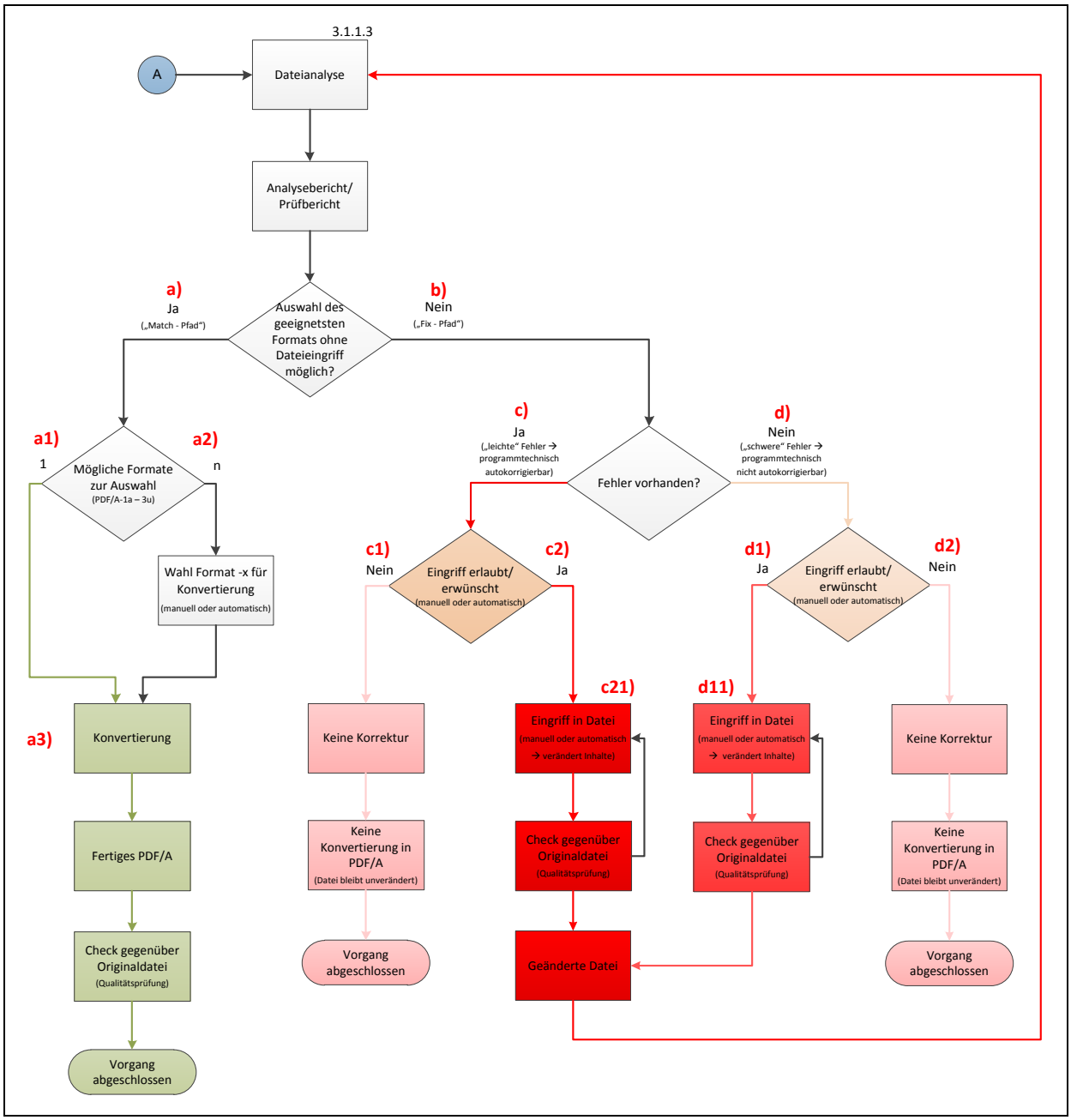

Abbildung 7: Variante 1 - „Match & Fix“

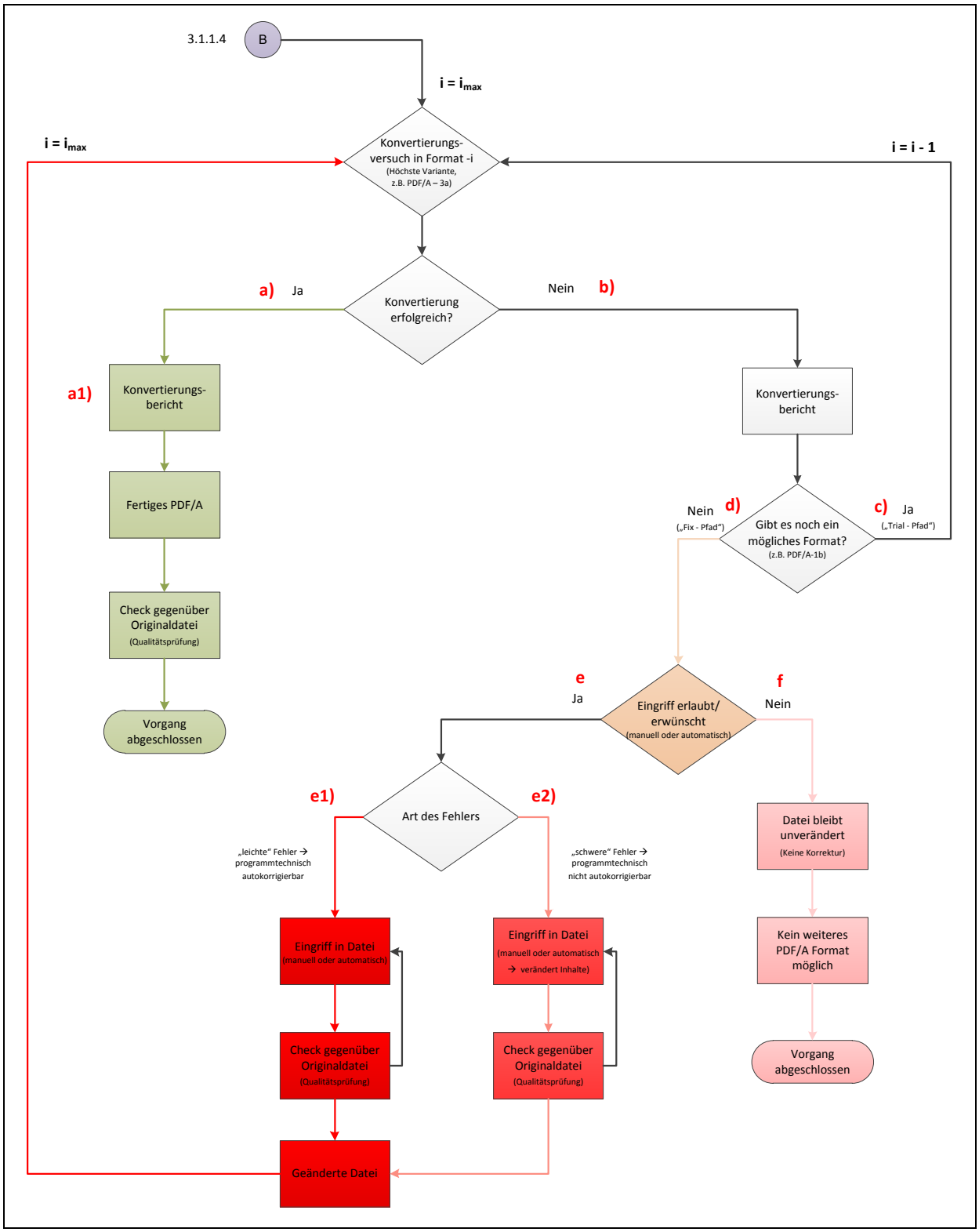

Abbildung 8: Variante 2 - „Trial & Fix"

3.1.1 Beschreibung - Ablaufdiagramm PDF/A

Für die Umsetzung des Prozesses sei hier näher auf die einzelnen Schritte eingegangen. Um dem Prozessablauf sowohl bildlich als auch textlich folgen zu können, werden die in Abbildung 6, Abbildung 7 und Abbildung 8 gezeigten Abschnitte punktweise aufgezählt und beschrieben.

Ferner sollen, für einen besseren Überblick, in Abschnitt 3.1.1.5 die wesentlichen Unterschiede, sowie auch die Gemeinsamkeiten der für diesen Prozess gewählten Varianten gegeben sein.

3.1.1.1 Eingangsdaten

Als Eingangsdaten seien all jene Geo - Daten definiert, die als eigenständige Geo - Dokumente vorliegen und von verschiedenen Institutionen/Quellen übernommen wurden (siehe Abschnitt 4.1). Darunter fallen alle möglichen Dateitypen (z.B. .pdf, .dxf, .jpg, .mdb), wobei das Augenmerk auf eigenständige PDF - und PDF/A - Dateien für die Findung der geeignetsten Archivierungsvariante gelegt wird. Eine Aussortierung der Dateien nach der Dateinamenserweiterung wird hier nicht angestrebt, da beispielsweise aufgrund von manuellen Nachbearbeitungsfehlern eine PDF - Datei versehentlich *„.pdg“* anstatt *„.pdf“* heißen könnte. Vielmehr soll eine Prüfung der Datei auf dessen Inhalte erfolgen, bei der anhand vorgegebener Header bzw. Metadaten, PDF - und PDF/A - Dateien aussortiert werden können. Die Überprüfung selbst erfolgt im Rahmen der Masterarbeit weder programm - noch softwaretechnisch, sondern rein konzeptionell.

3.1.1.2 Dateianalyse

Im Zuge der Dateianalyse sollen anhand des Analyseberichts Dateien mit Hilfe der verwendeten Software analysiert werden. Dabei sollen vorrangig die Inhalte der vorliegenden Dateien überprüft werden (ob diese beispielsweise JPEG - Dateien, Signaturen, Fonts und der gleichen beinhalten). Die Prüfung der Datei auf dessen Inhalt erfolgt dabei nicht nach der Dateinamenserweiterung, sondern anhand vorgegebener Header bzw. Metadaten, die anzeigen, ob das vorliegende Dokument eine PDF - oder eine PDF/A - Datei ist.

a) Anhand der Prüfung der Inhalte der Datei werden alle fälschlich als PDF deklarierte Dateien (z.B. Dateien wie Microsoft Word, Excel oder sonstige Dateien mit falscher Dateinamenserweiterung) im Zuge des Prozesses nicht in PDF - Dateien umgewandelt.
    a1) Sollte sich herausstellen, dass das vorliegende Dokument keine PDF - oder PDF/A - Datei ist, wird diese im Zuge des Prozesses nicht weiter verarbeitet, sondern herausgefiltert.
b) Bei Vorliegen einer PDF - bzw. PDF/A - Datei wird anhand der Detailprüfung entschieden, ob die vorliegende Datei eine PDF/A - Datei ist, welche herausgefiltert werden kann oder eine PDF - Datei ist, die mittels zweier möglicher Varianten weiterbehandelt werden kann.

c) Anhand der Detailprüfung wird erkannt, dass die vorliegende Datei eine PDF/A - Datei ist, welche herausgefiltert werden kann, da damit bereits ein korrektes PDF/A - Langzeitarchivformat vorliegt, welches nicht erneut untersucht werden muss.

   An dieser Stelle sei erwähnt, dass damit auch die Zielsetzung der Masterarbeit erfüllt ist, bei der einerseits die Durchführung des Prozesses möglichst automatisiert (mit so wenig wie möglich Eingriffen in die Datei) stattfinden und andererseits so viel wie möglich an Informationen im Zuge einer Konvertierung (und damit auch die Originalität einer Datei) beibehalten werden soll.

   Zusätzlich sei erwähnt, dass die Konvertierung in eine andere, als die bereits vorliegende PDF/A - Variante, hier nicht im Vordergrund steht, da jegliche Konvertierung einen Eingriff in die Datei darstellen würde, was damit auch zu Änderungen der Inhalte einer Datei führen könnte (z.B. wird die aufgebrachte Signatur zerstört).

d) Anhand der Detailprüfung wird erkannt, dass die vorliegende Datei eine PDF - Datei ist, die mittels einer der beiden Varianten weiterbearbeitet werden kann.

   e) Liegt eine PDF - Datei vor, so kann diese mittels *Variante 1* (siehe Abschnitt 3.1.1.3) weiter behandelt werden.

   f) Liegt eine PDF - Datei vor, so kann diese mittels *Variante 2* (siehe Abschnitt 3.1.1.4) weiter bearbeitet wird.

3.1.1.3 Variante 1 - „Match & Fix“

Die „Match & Fix“ - Variante beschäftigt sich mit der Dateianalyse im Hinblick auf die erlaubten, vorgegebenen bzw. normierten Inhalte der verschiedenen PDF/A - Formate. Dabei wird anhand von Normen und Standards versucht, alle erlaubten und untersagten Inhalte einer PDF/A - Datei zu definieren und diese entsprechend dem Standard der geeignetsten Archivvariante zuzuordnen.

Anhand des Analyse - oder Prüfberichts soll im Zuge der „Match & Fix“ - Variante entschieden werden:

a) Die vorliegende Datei beinhaltet alle erlaubten, notwendigen bzw. normierten Inhalte der jeweiligen PDF/A - Version. Damit ist die Auswahl des geeignetsten Formats ohne Eingriff in die Datei möglich.

   a1) Anhand der erlaubten Inhalte der PDF/A - Datei ist zu überprüfen, ob *ein* („1“) mögliches PDF/A - Format gewählt und damit eine abschließende Konvertierung in dieses Format durchgeführt werden kann.

   a2) Anhand der erlaubten Inhalte der PDF/A - Datei ist zu überprüfen, ob *mehrere* („n“) mögliche PDF/A - Formate gewählt werden können, um eine abschließende Konvertierung in ein gewähltes PDF/A - Format durchführen zu können.

Die Entscheidung, welches der möglichen Formate gewählt werden soll, ist abhängig vom Benutzer. Dies kann sowohl manuell, als auch programmtechnisch erfolgen.

a3) Nach Überprüfung der erlaubten Inhalte der PDF/A - Datei, ob *ein* („1“) oder *mehrere* („n“) mögliche Formate für die Konvertierung zur Wahl stehen, ist die Datei in das entsprechende PDF/A - Format zu konvertieren.

An dieser Stelle sei erwähnt, dass jegliche Konvertierung in eine gewählte PDF/A - Variante einen Eingriff in die Datei darstellt, weshalb hier nach der Konvertierung ein zusätzlicher Vergleich der konvertierten Datei mit der Originaldatei notwendig ist, um sicher zu gehen, dass weder rechtliche (z.B. aufgebrachte Signatur), noch visuelle (z.B. Fonts oder Transparenzen) Inhalte geändert wurden. Diese Qualitätsprüfung kann entweder automatisiert (z.B. mittels Binärvergleich) oder manuell (z.B. durch visuelle Prüfung) erfolgen.

b) Die vorliegende Datei beinhaltet nicht alle erlaubten, notwendigen bzw. normierten Inhalte der jeweiligen PDF/A - Version und damit ist die Auswahl des geeignetsten Formats nicht ohne Eingriff in die Datei möglich. Diesbezüglich ist auch zu überprüfen, ob „leichte“ oder „schwere“ Fehler (siehe Abschnitt 3.1.1.5 b) in der vorliegenden Datei vorhanden sind.

c) Sollten „leichte“ Fehler (Definition siehe Abschnitt 3.1.1.5 b), die programmtechnisch mit Hilfe der Software mittels Autokorrektur behoben werden könnten, auftreten, ist zu entscheiden, ob ein Eingriff in die Datei stattfinden soll oder nicht.

c1) Soll kein Eingriff in die Datei stattfinden, bleibt diese unverändert und damit ist auch keine Konvertierung in ein PDF/A - Format möglich.

c2) Soll ein Eingriff in die Datei stattfinden, ist darauf zu achten, dass bei Eingriff in die Datei die Inhalte dieser verändert werden könnten, was zur erneuten Durchführung des Prozesses und erneuten Entscheidungsmaßnahmen bezüglich der erlaubten Inhalte der verschiedenen PDF/A - Formate führt (Beginn wieder bei Abschnitt 3.1.1.3 a).

c21) Ein Eingriff in die Datei führt aufgrund der Änderungen der Inhalte der Datei zur erneuten Durchführung des Prozesses und erneuten Entscheidungsmaßnahmen bezüglich der erlaubten Inhalte der verschiedenen PDF/A - Formate (Beginn wieder bei Abschnitt 3.1.1.3 a).

Grundsätzlich ist jeglicher Eingriff in die Datei eine Manipulation dieser, da aufgrund des Eingriffs grundlegende Details geändert werden könnten. Damit ist

jede Datei als „neue“ Datei zu behandeln und mit der Originaldatei zu vergleichen (Qualitätsprüfung).

d) Sollten „schwere“ Fehler (Definition siehe Abschnitt 3.1.1.5 b) vorhanden sein und im Zuge der Analyse zu viele Fehler auftreten, die programmtechnisch mit Hilfe der Software nicht mittels Autokorrektur korrigierbar sind bzw. nicht entscheidbar ist, in welche Richtung korrigiert werden sollte (da z.B. weder für das eine noch das andere Format eine eindeutige Tendenz vorherrscht; dies wäre beispielsweise bei einer fehlerhaften Transparenz möglich, welche entweder zu reparieren wäre, um PDF/A - 2 erreichen zu können (allerdings mit der Gefahr der Änderung der Transparenz und möglicherweise Unleserlichmachung darunterliegender Informationen) oder es wäre die Transparenz durch Rasterisierung (Verschmelzung der Layer) zu eliminieren, um etwa PDF/A - 1 erreichen zu können), ist zu entscheiden, ob ein Eingriff in die Datei stattfindet oder nicht.

   d1) Soll ein Eingriff in die Datei stattfinden, ist darauf zu achten, dass bei Eingriff in die Datei die Inhalte dieser verändert werden, was zur erneuten Durchführung des Prozesses und erneuten Entscheidungsmaßnahmen bezüglich der erlaubten Inhalte der verschiedenen PDF/A - Formate führt (Beginn wieder bei Abschnitt 3.1.1.3).

      d11) Ein Eingriff in die Datei führt aufgrund der Änderungen der Inhalte der Datei zur erneuten Durchführung des Prozesses und erneuten Entscheidungsmaßnahmen bezüglich der erlaubten Inhalte der verschiedenen PDF/A - Formate (Beginn wieder bei Abschnitt 3.1.1.3).

      Auch hier sei darauf hingewiesen, dass jeglicher Eingriff in die Datei eine Manipulation dieser darstellt, da grundlegende Details geändert werden könnten. Auf Grund dessen ist jede Datei als „neue“ Datei zu behandeln und mit der Originaldatei zu vergleichen (Qualitätsprüfung).

   d2) Soll kein Eingriff in die Datei stattfinden, bleibt diese unverändert und damit findet auch keine Korrektur der PDF - Datei und folglich auch keine Konvertierung in eine PDF/A - Datei statt.

3.1.1.4 Variante 2 - „Trial & Fix“

Variante 2, in weiterer Folge als „Trial & Fix“ - Variante bezeichnet, beschäftigt sich mit dem sukzessiven Konvertierungsversuch (empirischer Ansatz) in die verschiedenen PDF/A - Varianten, um das geeignetste Format für die Archivierung für das vorliegende Dokument zu finden. Dabei wird versucht, die vorliegende Datei zunächst in

die höchste Variante (erstes Trial) zu konvertieren, da diese grundsätzlich die meisten Funktionen bzw. Features zulässt und somit die Chance besteht, die meisten (im Idealfall auch alle) Inhalte bzw. Informationen der Ausgangsdatei in die Langzeitarchivierungsdatei mitzunehmen. Im Falle, dass die Datei nicht den Normen und Standards der gewählten Variante entspricht, ist anschließend in die nächst niedrigere Variante zu konvertieren (zweites Trial). Diese Konvertierungsversuche bzw. diese Trials sind vorrangig zu wiederholen, bis das gewünschte Resultat erreicht ist. Dies bedeutet, dass durch die Konvertierungsversuche bzw. Trials (also ohne Dateieingriff) in erster Linie versucht wird, eine passende PDF/A - Variante zu finden. Erst wenn dies nicht gelingt, wird zum ersten Mal ein Eingriff in die Datei zugelassen und anschließend mit der veränderten Datei erneut eine Konvertierungs - bzw. Trial - Schleife durchlaufen.

Damit sei die Reihenfolge der Konvertierung wie folgt definiert:

1. PDF/A - 3a → - 3b → - 3u;

   Sollte kein Format zutreffen, ist in die nächst niedrigere Version zu konvertieren
2. PDF/A - 2a → - 2b → - 2u;

   Sollte kein Format zutreffen, ist in die nächst niedrigere Version zu konvertieren
3. PDF/A - 1a → - 1b;

   Sollte kein Format zutreffen, ist zu entscheiden, ob ein Eingriff in die Datei stattfinden soll oder nicht

Sollte die sukzessive Konvertierung kein Ergebnis liefern, muss der Benutzer entscheiden, ob ein Eingriff in die Datei stattfinden soll und damit alle notwendigen Inhalte gemäß der Norm in die Datei eingefügt bzw. korrigiert werden sollen. Ein Eingriff in die Datei sollte allerdings nur dann stattfinden, wenn keine weitere PDF/A - Variante mehr zur Konvertierung zur Auswahl steht.

Anhand des ersten (oder i - ten) Konvertierungsversuchs soll im Zuge der „Trial & Fix“ - Variante entschieden werden:

a) Die vorliegende Datei entspricht der im Zuge des aktuellen Konvertierungsversuchs gewählten PDF/A - Variante und kann als fertige PDF/A - Datei aussortiert werden.

   Damit liegt bereits ein geeignetes PDF/A - Langzeitarchivformat vor, welches nicht in ein weiteres PDF/A - Format konvertiert werden muss, da die Durchführung des Prozesses einerseits möglichst automatisiert (mit so wenig wie möglich Eingriffen in die Datei) stattfinden soll, und andererseits so viel wie möglich an Informationen (und damit auch die Originalität einer Datei) im Zuge einer Konvertierung beibehalten werden sollen.

   Zudem würde auch jegliche Konvertierung in eine PDF/A - Variante, einen Eingriff in die Datei darstellen, was damit auch zu Änderungen der Inhalte

einer Datei führen könnte, weshalb hier nach der Konvertierung ein zusätzlicher Vergleich der konvertierten Datei mit der Originaldatei notwendig ist, um sicher zu gehen, dass weder rechtliche (z.B. aufgebrachte Signatur), noch visuelle (z.B. Fonts oder Transparenzen) Inhalte geändert wurden.

a1) Konvertierung erfolgreich:

Anhand des Konvertierungsberichts wird bestätigt, dass die vorliegende Datei der gewählten PDF/A - Variante entspricht und somit als fertige PDF/A - Datei aussortiert werden kann. Damit liegt bereits ein geeignetes PDF/A - Langzeitarchivformat vor, welches nicht in ein weiteres PDF/A - Format konvertiert werden muss, da die Durchführung des Prozesses einerseits möglichst automatisiert (mit so wenig wie möglich Eingriffen in die Datei) stattfinden soll, und andererseits so viel wie möglich an Informationen (und damit auch die Originalität einer Datei) im Zuge einer Konvertierung beibehalten werden sollen.

Des weiteren würde jegliche Konvertierung in eine gewählte PDF/A - Variante einen Eingriff in die Datei darstellen. Diesbezüglich ist hier nach der Konvertierung ein zusätzlicher Vergleich der konvertierten Datei mit der Originaldatei notwendig, um zu überprüfen, dass weder rechtliche (z.B. aufgebrachte Signatur), noch visuelle (z.B. Fonts oder Transparenzen) Inhalte geändert wurden.

b) Die vorliegende Datei entspricht nicht der im Zuge des aktuellen Konvertierungsversuchs gewählten PDF/A - Variante.

c) Konvertierung fehlgeschlagen:

Anhand des Konvertierungsberichts lässt sich darauf schließen, dass die vorliegende Datei nicht der gewünschten oder gewählten Variante entspricht und daher ein erneuter Konvertierungsversuch in die nächst niedrigere Variante stattfinden soll.

Die Notwendigkeit, in die nächst niedrigere Version zu konvertieren ergibt sich aus der Zielsetzung der Masterarbeit, bei der die Durchführung des Prozesses möglichst automatisiert erfolgen (mit so wenig wie möglichen Eingriffen in die Datei) bzw. so viel wie möglich an Informationen im Zuge einer Konvertierung (und damit auch die Originalität einer Datei) beibehalten werden sollen.

d) Konvertierung fehlgeschlagen:

Anhand des Konvertierungsberichts lässt sich darauf schließen, dass die vorliegende Datei nicht der gewünschten oder gewählten Variante entspricht und auch kein weiteres Format für die Konvertierung mehr zur Auswahl steht, weshalb hier entschieden werden muss, ob ein Eingriff in die Datei stattfinden soll, um die vorab gewünschte oder gewählte PDF/A - Variante zu erhalten, oder nicht. Bei einem Eingriff ist jedenfalls zu überprüfen, ob „leichte“

oder „schwere“ Fehler (Definition siehe Abschnitt 3.1.1.5 b) in der vorliegenden Datei vorhanden sind.

e) Sollte ein Eingriff in die Datei stattfinden, ist zunächst zu überprüfen, ob „leichte“ oder „schwere“ Fehler (Definition siehe Abschnitt 3.1.1.5 b) vorhanden sind. Zusätzlich ist darauf zu achten, dass bei einem Eingriff in die Datei die Inhalte dieser verändert werden könnten, was zur erneuten Durchführung des Konvertierungsprozesses und erneuten Entscheidungsmaßnahmen bezüglich der gewählten PDF/A - Variante führt (Beginn wieder bei Abschnitt 3.1.1.4), um die vorab gewünschte oder gewählte PDF/A - Variante zu erhalten.

   e1) Sollte ein Eingriff in die Datei stattfinden und „leichte“ Fehler (Definition siehe Abschnitt 3.1.1.5 b) vorhanden sein, die programmtechnisch mit Hilfe der Software mittels Autokorrektur behoben werden könnten, ist darauf zu achten, dass es aufgrund des Eingriffs zu Änderungen der Inhalte in der Datei kommen könnte, was zur erneuten Durchführung des Prozesses und erneuten Entscheidungsmaßnahmen bezüglich der gewählten PDF/A - Variante führt (Beginn wieder bei Abschnitt 3.1.1.4).

   Grundsätzlich ist jeglicher Eingriff in die Datei eine Manipulation dieser, da aufgrund des Eingriffs grundlegende Details geändert werden könnten. Damit ist jede Datei als „neue“ Datei zu behandeln und mit der Originaldatei zu vergleichen (Qualitätsprüfung).

   e2) Sollte ein Eingriff in die Datei stattfinden und „schwere“ Fehler (Definition siehe Abschnitt 3.1.1.5 b) vorhanden sein, die programmtechnisch nicht mit Hilfe der Software mittels Autokorrektur behoben werden könnten, ist darauf zu achten, dass es aufgrund des Eingriffs zu Änderungen der Inhalte in der Datei kommt, was zur erneuten Durchführung des Prozesses und erneuten Entscheidungsmaßnahmen bezüglich der gewählten PDF/A - Variante führt (Beginn wieder bei Abschnitt 3.1.1.4).

   Da jeglicher Eingriff in die Datei eine Manipulation darstellt, bei der aufgrund des Eingriffs grundlegende Details geändert werden könnten, ist jede Datei als „neue“ Datei zu behandeln und mit der Originaldatei zu vergleichen (Qualitätsprüfung).

f) Sollte kein Eingriff in die Datei stattfinden, so bleibt diese unverändert und damit ist auch die Wahl für ein weiteres PDF/A - Format nicht mehr möglich bzw. findet auch hier kein erneuter Konvertierungsversuch in eine andere Variante statt.

3.1.1.5 Grundsätzliches zu den beiden Varianten

Für einen besseren Überblick, seien folgend die wesentlichen Unterschiede, sowie auch die Gemeinsamkeiten der beiden zuvor beschriebenen Varianten genannt.

a) Wesentliche Unterschiede der beiden Varianten

- Mittels *Variante 1 („Match & Fix")* wird versucht, unter Berücksichtigung der erlaubten Inhalte gemäß den Normen und Standards der verschiedenen PDF/A - Formate, die geeignetste, gewünschte oder gewählte Variante zu erlangen.
  - *Variante 1* stellt sich damit, im Vergleich zu Variante 2, nicht nur aufgrund der äußerst detaillierten Datei - Analyse als umfangreicher heraus, sondern ist auch in ihrer Art und Weise viel komplexer umzusetzen, denn alles was für die jeweilig gewählte PDF/A - Variante notwendig ist, ist aus den Normen selbst herauszulesen und umzusetzen. Die Ausgangsdatei ist somit soweit zu analysieren, um deren Bestandteile den Anforderungen der verschiedenen PDF/A - Versionen gegenüberzustellen bzw. „Matchen" zu können. Gegebenenfalls sind auch gewisse Korrekturen („Fixes") in Richtung einer möglichen PDF/A - Variante vorzunehmen, wenn diesbezüglich eine Tendenz erkennbar ist. Eine Tendenz wäre beispielsweise erkennbar, wenn hinsichtlich einer der möglichen PDF/A - Varianten (z.B. PDF/A - 2) weniger Fehler auftreten (d.h. weniger Korrekturen (Fixes) notwendig sind, um die Normanforderungen einzuhalten), als für eine andere PDF/A - Variante (z.B. PDF/A - 1).

- *Variante 2 („Trial & Fix")* beschäftigt sich mit der sukzessiven Findung des geeignetsten Formats durch fortlaufende Konvertierungsversuche („Trials"), ohne dabei vorab zu analysieren, welche PDF/A - Variante für die vorliegende Datei die geeignetste wäre.
  - *Variante 2* stellt sich damit als weniger komplex heraus, da hier lediglich sukzessive Konvertierungsversuche („Trials") stattfinden, bei der die vorliegende Datei zunächst versucht wird in die höchste PDF/A - Variante (PDF/A - 3a) zu konvertieren. Sollten die Inhalte der vorliegenden Datei für die gewählte Variante nicht den Normen entsprechen, ist hier die nächst niedrigere Variante zu wählen. Dieser Vorgang ist vorrangig (vor einem Eingriff in die Datei) zu wiederholen, bis eine PDF/A - Variante gefunden wird. Sollte im Zuge der sukzessiven Konvertierung kein Resultat erreicht werden können, besteht die Möglichkeit eines Eingriffs in die Datei,

um die notwendigen Inhalte der Datei gemäß Normen und Standards festzulegen bzw. zu korrigieren. Ein Eingriff in die Datei ist allerdings nur dann vorgesehen, wenn die Wahl für ein weiteres PDF/A - Format nicht mehr möglich ist. Dies bedeutet, dass hier nicht bereits bei einem ersten Durchlauf („Trial“) mögliche Fehler im Hinblick auf diese PDF/A - Version korrigiert werden, sondern zunächst versucht wird, die Datei in eine andere bzw. niedrigere PDF/A - Version (mit geringeren Anforderungen) zu konvertieren.

- Die Wahl, von der höchsten Variante in die nächst niedrigere zu konvertieren („Top - Down“), ergibt sich hier als vorteilhaft, denn die höchste Variante stellt die meisten Vorgaben, Funktionen bzw. Features zur Verfügung, womit auch hier am meisten Informationen in einer Datei beibehalten werden können. Allerdings stellt sich diese Methode als komplexer heraus, als der umgekehrte Fall, bei der von der niedrigsten Variante in die nächst höhere („Bottom - Up“) konvertiert werden soll, da hier die meisten Normvorgaben erfüllt werden müssen.
- Würde stattdessen von der niedrigsten Variante in die nächst höhere („Bottom - Up“) konvertiert werden, würde diese zwar als weniger komplex gelten, als der umgekehrte Fall („Top - Down“), da hier die wenigsten Normpunkte erfüllt werden müssen, jedoch könnten bei dieser Methode, im Falle eines notwendigen Eingriffes, all jene Inhalte korrigiert werden, die bei einer höheren Variante erlaubt gewesen wären.
  - Diese Methode würde daher der Zielsetzung der Masterarbeit widersprechen, bei der die Durchführung des Prozesses möglichst automatisiert (mit so wenig wie möglich Eingriffen in die Datei) erfolgen soll, sowie auch so wenig wie möglich an Informationen (im Zuge der Konvertierung der Datei für die Findung des geeignetsten Formats für die Langzeitarchivierung hinsichtlich PDF/A) verloren gehen sollte.

b) Beschreibung Fehlerarten und Eingriffe in die Datei

Beide Varianten haben gemeinsam, dass im Zuge der Prozessdurchführung ein manueller oder automatischer Eingriff in die Datei stattfinden könnte, bei welchem es zu Veränderungen der Inhalte der Datei kommen kann. Ein

Eingriff in die Datei findet vor allem dann statt, wenn „leichte“ oder „schwere“ Fehler im Zuge der Prozessdurchführung auftreten können. Dies ist dann der Fall, wenn Informationen in der Datei vorhanden sind, die in der entsprechenden Variante nicht erlaubt sind (beispielsweise sind Transparenzen verboten in PDF/A - 1 und müssen daher entfernt werden, falls dies die geeignetste, gewünschte oder gewählte PDF/A - Variante ist bzw. könnten auch Metadaten, wie z.B. eine nicht korrekte Definition eines Fonts oder Farbraums, nicht vollständig vorhanden sein, was in diesem Fall alle PDF/A - Formate betreffen könnte). Die Definition aller „leichten“ oder „schweren“ Fehler, wurden im Rahmen der Masterarbeit selbst gewählt, stehen aber in Anlehnung an die gewählte Software, welche selbst Korrekturvorschläge für gewisse Fehler angeben kann.

- Als „leichte“ Fehler werden all jene bezeichnet, welche mit Hilfe der Software selbst behoben werden können.
  - Dies könnten z.B. Definitionsfehler in Schriftarten, nicht korrekte Metadateninformationen, nicht korrekte Farbdefinitionen, fehlerhafte Fontdefinitionen und der gleichen sein.
  - Hierbei kann es sich auch um einen *automatischen Eingriff* in die Datei handeln, bei der die Software einen Korrekturvorschlag zur Verbesserung dieser leichten Fehler angeben kann. Beispielsweise erkennt die Software anhand der Fontdefinitionen, um welchen Font es sich bei der vorliegenden Datei handelt. Da aber nach wie vor Fehler vorhanden sein könnten, kann die Software zur Verbesserung dieser einen Korrekturvorschlag angegeben. Damit wäre eine Autokorrektur (und damit auch ein automatischer Eingriff in die Datei) möglich, welche jedoch nicht zwingend durchgeführt werden muss.
  - Dem Benutzer steht auch die Möglichkeit eines *manuellen Eingriffs* in die Datei zur Verfügung, bei der Korrekturen ohne Hilfe der Software durchgeführt werden können.
  - Um sicher zu gehen, dass tatsächlich alle vorhandenen Fehler behoben wurden, ist die Datei bei jedem Eingriff wieder als „neue“ Datei zu behandeln, bei der der Prozess erneut durchlaufen und die Datei erneut analysiert bzw. mit der Originaldatei verglichen werden muss (Qualitätsprüfung).

- „Schwere“ Fehler sind all jene Fehler, die durch den Benutzer selbst (z.B. durch Löschen, Verändern oder Hinzufügen) zu korrigieren sind (sogenannte massive Eingriffe).
  - Dies könnten z.B. fehlende Schriftarten, fehlerhafte Strukturinformationen und der gleichen sein.

- Hierbei wird es sich hauptsächlich um einen *manuellen Eingriff* in die Datei handeln, bei der die Software selbst keine Korrekturvorschläge zur Verbesserung dieser schweren Fehler angeben kann, sondern der Benutzer selbst in die Datei eingreifen muss.
- Dem Benutzer steht somit die Möglichkeit z.B. das Löschen, Verändern oder Hinzufügen von fehlerhaften Inhalten in der Datei zur Verfügung.
- Um sicher zu gehen, dass tatsächlich alle vorhandenen Fehler behoben wurden, ist die Datei bei jedem Eingriff wieder als „neue“ Datei zu behandeln, bei der der Prozess erneut durchlaufen und die Datei erneut analysiert bzw. mit der Originaldatei verglichen werden muss (Qualitätsprüfung).

Als Eingriff in die Datei können generell das Löschen, das Verändern oder Hinzufügen von Inhalten bzw. Metadaten in einer Datei verstanden werden.

- *Löschen*: z.B. das Entfernen von verbotenen Inhalten bezüglich der geeignetsten, gewünschten oder gewählten PDF/A - Variante (z.B. Transparenzen)
- *Verändern*: z.B. das Verändern von z.B. anderen Schriftarten, falls der aktuelle Font nicht verwendbar/fehlerhaft ist
- *Hinzufügen*: z.B. das Hinzutun von Metadaten, falls im Zuge der Durchführung des Prozesses unklare Definitionen auftreten (z.B. Farbraumdefinition)

c) Problematik/Folgen des Eingriffs in die Datei

Ein Eingriff in die Datei kann sowohl rechtliche, als auch inhaltsrelevante (visuelle) Folgen haben - denn überall dort, wo externe Dateien (z.B. Urkunden) in einem Langzeitarchiv archiviert werden sollen, dürfen die Inhalte nachträglich nicht geändert werden.

Um die Auswirkungen bei einem Eingriff in die Datei näher zu bringen, seien hier zwei Beispiele angeführt:

Als *erstes Beispiel* sei hier eine auf der Datei aufgebrachte Signatur (= Versiegelung) genannt. Wird beispielsweise eine PDF - Datei (vom Urkundenersteller) mit einer Signatur versehen und diese Datei anschließend (vom Empfänger) in eine PDF/A - Version konvertiert, ergäbe dies jedenfalls rechtliche Folgen, während visuell gesehen unter Umständen gar keine Unterschiede auftreten könnten:

- *Rechtliche Folgen:*
  - Findet ein Eingriff in die Datei statt (z.B. durch Schriftartenänderung im Zuge der PDF/A - Konvertierung), wird damit der Siegel gebrochen, was die Signatur nicht nur ungültig, sondern auch die Rückführung bzw. Verifikation auf seinen Ursprung oder Hersteller nicht mehr zu 100% nachvollziehbar macht.

- *Inhaltsrelevante (visuelle) Folgen:*
  - Betrachtet man ein und dieselbe Datei vor bzw. nach der Bearbeitung, wird visuell gesehen kaum ein Unterschied gegenüber der Ausgangsdatei zu erkennen sein.
  - An dieser Stelle sei jedoch vermerkt, dass jede Datei, in die ein Eingriff stattfindet, verändert wird.
  - Wird somit die im Anschluss generierte PDF/A - Datei gegenüber der Ausgangsdatei überprüft bzw. verglichen, handelt es sich dabei nicht um die identische Datei.
  - Um dies besser zum Ausdruck bringen zu können, sei hier auf das in Abschnitt 2.3.3 genannte Beispiel des Eurozeichens, dem Währungssymbol für den Euro, verwiesen, bei dem deutlich zu erkennen ist, dass ein Symbol durch die jeweilige Anzeigeplattform bzw. Betriebssystem verändert bzw. (visuell gesehen) anders dargestellt werden kann.

Als *zweites Beispiel* sei eine Mail mit beigefügtem PDF - Antrag (z.B. ein Antrag auf Durchführung einer Amtshandlung mit Bürgerkartensignatur, vgl. BEV - Bundesamt für Eich- und Vermessungswesen, 2012), genannt. Fände hier ein Eingriff in die Datei im Zuge einer Konvertierung statt, könnte dies ebenfalls rechtliche, sowie auch visuelle Auswirkungen haben:

- *Rechtliche Folgen:*
  - Obwohl auch hier mit einem Eingriff in die PDF - Datei die Signatur ungültig werden würde, würde jedoch rechtlich gesehen eine Antragsstellung vorliegen, wenngleich auch die Authentizität nicht mehr gegeben wäre. Es müsste somit von der Behörde erhoben bzw. verifiziert werden, ob der genannte Antragssteller auch tatsächlich der jenige ist, der angeführt ist. Bei intakter Signatur wäre dies mittels Signaturprüfung möglich.

- *Inhaltsrelevante (visuelle) Folgen:*
  - Visuell gesehen, könnten sich bei der Betrachtung des Antrags Probleme ergeben.
  - Der Antrag selbst müsste zum gegenwärtigen Zeitpunkt nicht zwingend im PDF/A - Format vorliegen, jedoch müsste die Authentizität des Dokuments nachprüfbar sein (z.B. durch eine digitale Signatur).
  - Allerdings ist, aufgrund des fehlenden PDF/A - Formats, nicht gewährleistet, dass sämtliche Fonts im Dokument eingebettet sein müssen. Somit könnte die Schriftart durch die jeweilige Anzeigeplattform bzw. Betriebssystem unterschiedlich interpretiert bzw. (visuell gesehen) anders dargestellt werden, da bei der Betrachtung des Dokuments eine Ersatzschrift (Ersatzfont) für diese gefunden und damit der Antragsinhalt missinterpretiert werden könnte.
  - Um dies zu veranschaulichen, sei auch an dieser Stelle auf das in Abschnitt 2.3.3 genannte Beispiel des Eurozeichens verwiesen.
  - Als Folge dieser Visualisierungsprobleme müsste die Behörde hier zur Klärung gegebenenfalls ein Ermittlungsverfahren einleiten.

d) Konvertierung einer Datei in ein PDF/A - Format

Im Zuge der Prozessdurchführungen könnte es, anstatt zu einem Eingriff, zu einer Konvertierung der Datei in ein mögliches PDF/A - Format kommen. Dies geschieht dann, wenn anhand des Analyse - oder Prüfberichts bzw. Konvertierungsberichts hervorgeht, dass bereits alle Normpunkte erfüllt sind und damit ein gültiges PDF/A - Format vorliegt.

Auch an dieser Stelle sei vermerkt, dass nicht nur der Eingriff, sondern auch jegliche Konvertierung einer Datei in eine der möglichen PDF/A - Varianten einen Eingriff in diese darstellt, die Zielsetzung der Masterarbeit jedoch vorsieht, dass einerseits die Durchführung der Prozesse möglichst automatisiert (mit so wenig wie möglich Eingriffen in die Datei) erfolgt und andererseits so viel wie möglich an Informationen im Zuge einer Konvertierung (und damit auch die Originalität der Datei) beibehalten werden sollen.

3.1.2 Verwendete Software

Einleitend soll hier ein Überblick über einige PDF/A - Softwareprogramme und Hersteller geboten werden. Da bereits eine große Auswahl an Programmen und Herstellern besteht, die ihre Produkte als Trial oder als Freeware Version für die PDF/A - Umwandlung und - Validierung zur Verfügung stellen, seien hier, ohne näher ins Detail zu gehen, einige der federführenden Hersteller von Softwareprodukten genannt, die nicht nur wegen ihrer Mitgliedschaft und Zusammenarbeit im ISO Komitee zu einer der wichtigen Herstellern von Softwareprogrammen gehören, sondern auch wegen ihrer gemeinsamen Zielsetzung rund um PDF/A zu den Experten im Bereich Langzeitarchivierung mittels PDF/A zählen:

- PDF Tools AG (http://www.pdf-tools.com, Stand: März 2015)
- intarsys consulting GmbH (http://www.intarsys.de, Stand: März 2015)
- Solid Documents (http://www.soliddocuments.com, Stand: März 2015)
- callas software GmbH (http://www.callassoftware.com, Stand: März 2015)

Alle oben genannten Hersteller sind Mitglieder des PDF/A Competence Centers (http://www.pdfa.org), deren Ziel *„die Förderung des Informations- und Erfahrungsaustausches auf dem Gebiet Langzeitarchivierung gemäß ISO 19005: PDF/A“* (PDF Association, 2015c) ist.

Folgend soll hier auf die von callas software GmbH verwendete Software näher eingegangen werden, welche nicht nur für die Archivierung von Dokumenten entwickelt wurde, sondern auch die Eigenschaft besitzt, schnell und effizient PDF/A - Dokumente zu erstellen und zu überprüfen.

3.1.2.1 Callas - Über die Software

callas software GmbH ist eine deutsche Firma, gegründet im Jahre 1995 von Olaf Drümmer mit Sitz in Berlin. Seit der Gründung beschäftigt sich die Firma mit PDF - und PDF/A - Technologien und gehört auch zu einem der ersten Entwickler im Bereich PDF/A (callas software GmbH, 2015a; PDF Association, 2015d).

Olaf Drümmer als Gründer und Geschäftsführer dieser Firma, gehört zu den Experten im Bereich PDF/A. Neben den zahlreichen Artikeln und Werken die er verfasst hat, sei hier das auf PDF/A aufbauende Standardwerk *„PDF/A Kompakt. Digitale Langzeitarchivierung mit PDF“* (Drümmer et al., 2007) verwiesen. Nicht nur seine Mitarbeit im ISO Komitee, sondern auch seine Mitgliedschaft in internationalen Institutionen und Verbänden, wie etwa DIN oder das PDF/A Competence Center, sowie auch sein Expertenwissen rund um PDF/A, waren ausschlaggebend für die Wahl dieser Software (Drümmer et al., 2007, S. 86), während ein weiterer Grund der verstärkte Einsatz und die Integration dieser Software von Adobe selbst ist (callas software GmbH, 2015f).

Neben den verschiedenen von callas software GmbH entwickelten Produkten, seien hier zwei wesentliche genannt, die sowohl als Desktop - als auch als serverbasierte Produkte erhältlich sind (callas software GmbH, 2015b). Beide zeigen große Ähnlichkeit, unterscheiden sich jedoch im Folgendem (callas software GmbH, 2015c; 2015d):

1) *callas pdfToolbox* richtet sich hauptsächlich an Benutzer, die das äußere Erscheinungsbild und damit auch bestimmte Eigenschaften einer Datei verändern (wie z.B. Schriften oder Transparenzen) oder Dateien in das PDF - Format umwandeln wollen.
2) *callas pdfaPilot* wurde primär für die Archivierung entwickelt, wobei auch Überprüfungen auf den PDF/A - Standard und Konvertierungen von Dokumenten in PDF/A - Dokumente ermöglicht werden. Weiters können hier auch alle Punkte in der oben genannten Software *callas pdfToolbox* angewendet werden.

Ausgehend von diesen verschiedenen Integrationsmöglichkeiten der Callas Software und ausreichender Recherche, wurde die *pdfaPilot Desktop* Version (Version: 5.1.211) zur Umsetzung des Prozesses gewählt, da diese einerseits für die Archivierung entwickelt wurde, andererseits auch mit Hilfe der Desktop Version sehr schnell und effizient PDF/A - Dokumente erstellt und überprüft werden können.

Es handelt sich dabei um ein Produkt von callas software GmbH, welches als Gratis - Vollversion per E - Mail für fünfzehn Tage für verschiedene Betriebssysteme (z.B. Linux, Mac OS, Microsoft Windows) erworben werden kann. Für die Nutzung der Software ist weiters ein erneuter Kontakt mit der callas software GmbH notwendig, um die jeweilige Version durch einen Lizenzkey (Schlüssel) zu aktivieren (callas software GmbH, 2011).

Basierend auf (callas software GmbH, 2015e) sollen folgend auch einige wichtige Funktionen der verwendeten Software aufgelistet werden:

- PDF Dateien können auf den PDF/A - Standard (ISO - 19005) überprüft und konvertiert werden.
- Die Software kann Vorschläge zur Korrektur hinsichtlich gefundener Fehler machen.
- Es werden alle PDF/A - Versionen (PDF/A - 1, PDF/A - 2, PDF/A - 3) und deren Konformitätsstufen (a, b, u) von der Software unterstützt.
- Durchgeführte Arbeiten können mit bereitgestellten Reports dokumentiert und nachvollzogen werden.

3.1.2.2 Umgang mit der Software

Die Software *pdfaPilot Desktop* ist ein einfach zu bedienendes GUI (grafische Benutzeroberfläche). Nach Start der Software öffnet sich ein Anzeigefenster (Abbildung 9), das verschiedene Wahlmöglichkeiten bietet, Dokumente verschiedenster Art zu laden (callas software GmbH, 2011).

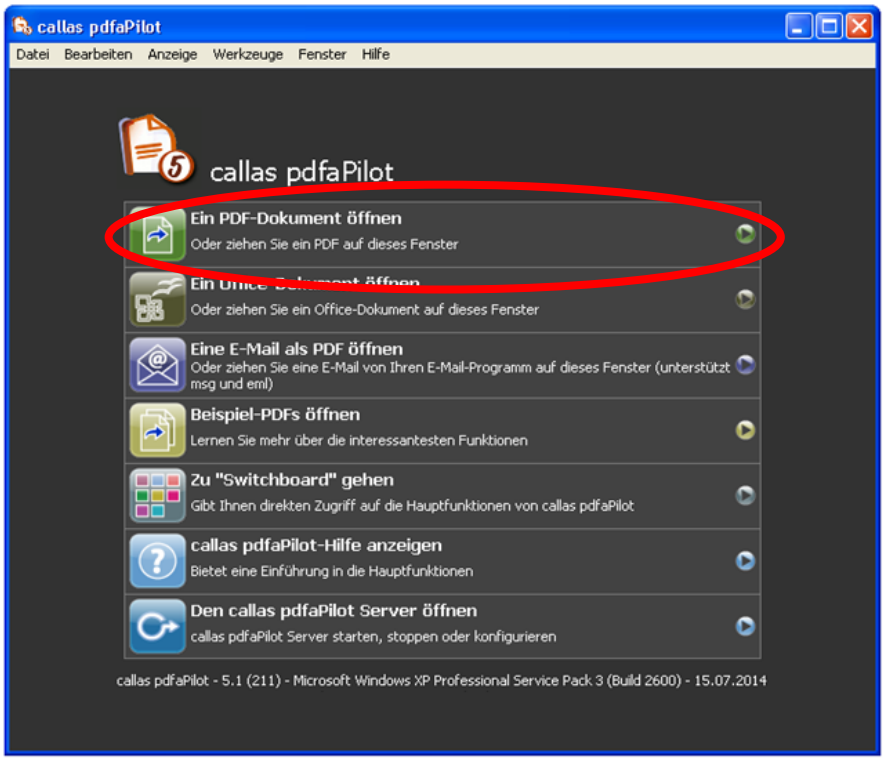

Abbildung 9: Screenshot callas pdfaPilot Hauptfenster
(Screenshot bearbeitet im März 2015)

Das dazugehörige Arbeitsfenster (Abbildung 10) zeigt verschiedene Elemente und Möglichkeiten dieses GUI zu bedienen. Möchte der Benutzer beispielsweise ein Dokument nach PDF/A konvertieren oder auf PDF/A überprüfen, genügt ein einfacher Klick. Neben der Überprüfung und Generierung verschiedenster PDF/A - Versionen, können auch Dateien eingebettet und extrahiert werden, sowie Metadaten einer Datei betrachtet oder entfernt werden. Weiters können Eigenschaften eines Dokuments (z.B. S/W - Ausdruck, Konvertierung von Bildern nach PDF), Seitenlayouts (z.B. Skalierung, Anpassung des Seitenbereichs) und verschiedene Anordnungen eines Dokuments für den Druck (z.B. beidseitigen Druck, mehrere Seiten pro Blatt) einer Datei bearbeitet werden (callas software GmbH, 2011).

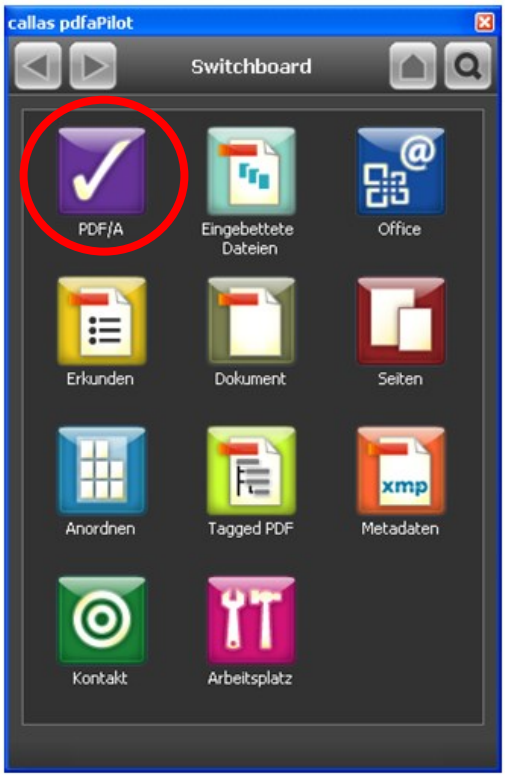

Abbildung 10: Screenshot callas pdfaPilot Switchboard (Screenshot bearbeitet im März 2015)

Die Software selbst bietet auch vordefinierte Profile, Prüfungen und Korrekturen (Abbildung 11, Abbildung 12 und Abbildung 13) an, um verschiedene Aktionen (z.B. Überprüfung auf PDF/A - Version; Überprüfen, ob Schriften eingebettet sind; Anzeige der Bildauflösung der vorhandenen Bilder) durchführen zu können. Entsprechende Reports im PDF - , XML - oder TXT - Format liefern jeweils einen Bericht über die durchgeführte Arbeit und etwaiger aufgetretener Fehler (z.B. über Schriften, Bilder, Farben), die wiederum durch die Software bzw. dem Benutzer behoben werden können (callas software GmbH, 2011).

Für detailliertere Informationen möchte der Leser hier auf das Manual der verwendeten Software von callas software GmbH (callas software GmbH, 2011) weiter verwiesen werden.

Abbildung 11: Vordefinierte Profile
(Screenshot bearbeitet im März 2015)

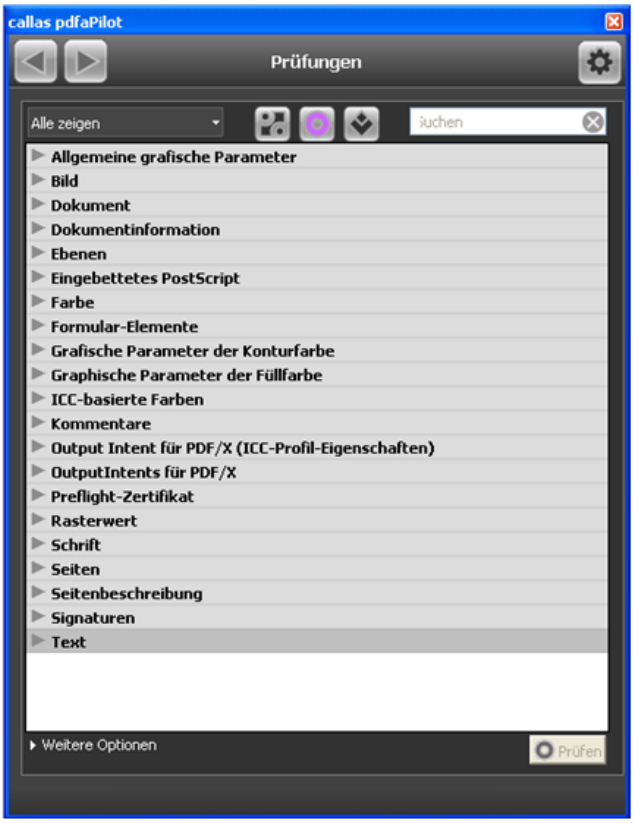

Abbildung 12: Vordefinierte Prüfungen
(Screenshot bearbeitet im März 2015)

Abbildung 13: Vordefinierte Korrekturen
(Screenshot bearbeitet im März 2015)

3.2 Ablaufdiagramm - OCR

An dieser Stelle soll ein Konzept über den Aufbau und die Funktionsweise des Texterkennungsverfahrens näher gebracht werden. Da die Texterkennung mittels OCR nicht Kernthema der Masterarbeit ist, wird hier nicht ins Detail auf die einzelnen Schritte eingegangen. Vielmehr soll hier das grundlegende Prinzip der Texterkennung wiedergegeben und jene Schritte dokumentiert werden, mit denen eine Umwandlung in das Langzeitarchivformat PDF/A ermöglicht werden kann, sodass die aus den langzeitarchivierten Originaldokumenten abgeleiteten OCR - Dokumente, neben den Originaldokumenten, ebenfalls im Langzeitarchiv abgelegt werden können.

Abbildung 14 zeigt den schematischen Aufbau dieses Konzepts, während im folgendem Abschnitt (Abschnitt 3.2.1) näher auf die einzelnen Schritte eingegangen wird.

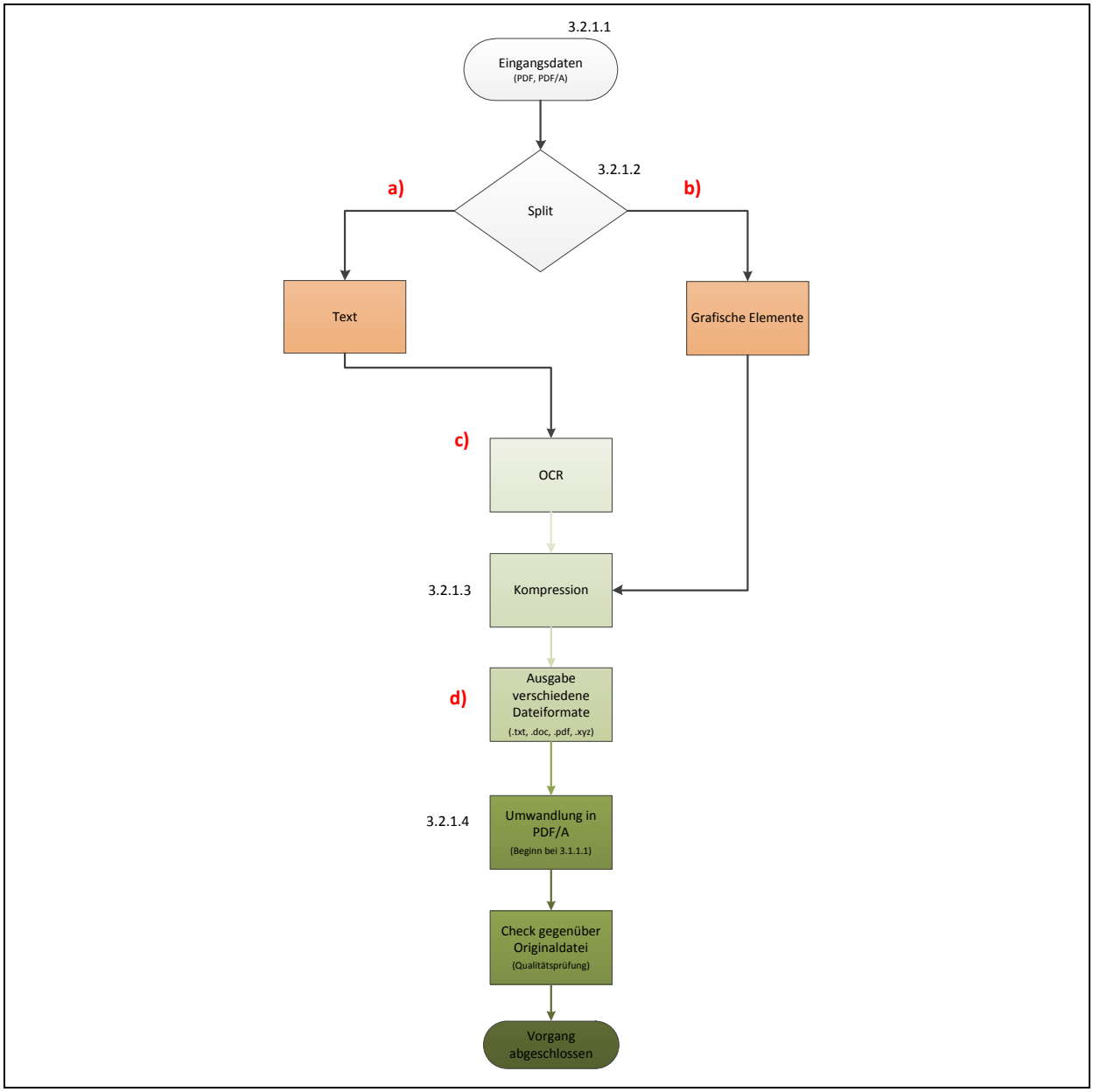

Abbildung 14: Konzept OCR - Prozess

3.2.1 Beschreibung - Ablaufdiagramm OCR

Für die Umsetzung des Prozesses sei hier näher auf die einzelnen Schritte eingegangen. Um dem Prozessablauf sowohl bildlich als auch textlich folgen zu können, werden die in Abbildung 14 gezeigten Abschnitte punktweise aufgezählt und beschrieben.

3.2.1.1 Eingangsdaten

Als Eingangsdaten seien all jene Dokumente definiert, bei denen das Durchsuchen oder Extrahieren von Text - Informationen mit Hilfe einer Volltextsuche nicht bzw. nur teilweise möglich ist.

Das Augenmerk wird dabei ausschließlich auf PDF - bzw. PDF/A - Dateien gelegt, wobei je nach Norm auch andere eingebettete Dokumente (z.B. JPEG, TIFF, BMP, PNG) in diesen erlaubt werden.

Mit Hilfe des OCR - Texterkennungsverfahrens sollen diese Dokumente derart bearbeitet werden, sodass anschließend eine Volltextsuche in diesen ermöglicht werden kann bzw. auch nicht unmittelbar zugängliche Informationen (z.B. Rastertext) zugänglich gemacht werden können. Diese, aus den langzeitarchivierten Originaldokumenten abgeleiteten OCR - Dokumente, sollen abschließend in die entsprechende, für die Langzeitarchivierung geeignetste Archivvariante mit Fokus auf PDF/A (ISO - Norm 19005) umgewandelt werden.

3.2.1.2 Split

Mit Hilfe des Split - Funktion, welche im Programm integriert ist, werden zunächst alle sich auf dem Dokument befindenden Elemente in einzelne Ebenen aufgespalten. Die Software erkennt dabei, in welcher Ebene sich „Textelemente“ und in welcher sich „grafische Elemente“ befinden. Ohne näher ins Detail zu gehen, sei hier der grundlegende Ablauf der weiteren Schritte in Anlehnung an Hermann (2008), Vorbach (2014) und Computer Bild (2009a; 2009b; 2009c) dokumentiert:

a) Die Ebene „Textelemente“ wird von der Software derart bearbeitet, sodass hier einzelne Absätze, Zeilen, Wörter und Zeichen in dieser Ebene erkannt und interpretiert werden.

b) In der Ebene „grafische Elemente“ befinden sich alle Grafiken und Bilder. Diese werden von der Software üblicherweise nicht weiter bearbeitet. Sollte sich jedoch ein Text auf Bildern befinden, bietet die Software die Möglichkeit an, Text von Bild zu unterscheiden. An dieser Stelle kommen verschiedene Methoden und Algorithmen, welche im Programm integriert sind, zum Einsatz, um eine Texterkennung in Bildern zu ermöglichen.

c) Bei der Texterkennung mittels OCR kommen verschiedene Methoden und Algorithmen zum Einsatz, um eine optimale Texterkennung zu ermöglichen. Eine der wichtigsten Methoden ist die Merkmals - und Mustererkennung, bei der das Programm versucht, die vorhandenen Zeichen zu erkennen und zu interpretieren. Da in der jeweiligen Software sowohl Zei-

chen als auch Wörterbücher integriert sind, können damit beispielsweise das Verschmelzen von Zeichen (z.B. i wird zu l) oder falsche Interpretationen von Zeichen (z.B. ä wird zu a) und Wörtern weitgehend vermieden werden.

3.2.1.3 Kompression

In diesem Schritt werden die beiden Ebenen „Textelemente“ und „grafische Elemente“ wieder vereinigt bzw. komprimiert.

d) Je nach Wahl des Ausgabeformates (z.B. .txt, .rtf, .doc, .pdf), kann ein digitales Dokument erstellt werden, das im Anschluss daran weiterbearbeitet werden kann (z.B. Durchsuchen oder Extrahieren von Text, Umwandlung in ein PDF/A - Dokument).

3.2.1.4 Umwandlung in PDF/A

Da es sich bei der Texterkennung mittels OCR um einen Eingriff in die Datei bzw. eine Konvertierung der Datei handelt, ist an dieser Stelle zu beachten, dass die Inhalte dieser verändert werden (auch wenn die Umwandlung der Datei nach dem OCR - Prozess im PDF/A - Format ohne weiteren Eingriff möglich ist). Im Vergleich zum minimalen Eingriff bei der Konvertierung einer PDF - Datei in eine PDF/A - Datei, wie es beim PDF/A - Prozess der Fall war, ist die Texterkennung ein weitaus umfangreicherer Eingriff. Um daher ein akzeptables, für die Langzeitarchivierung geeignetes PDF/A - Format zu erhalten, ist jede aus dem OCR - Prozess erhaltene Datei mit dem zuvor definierten Prozess für die Langzeitarchivierung mittels PDF/A (Abschnitt 3.1) erneut zu analysieren. Dabei sollten vorrangig die Inhalte der Datei überprüft werden, bei der anhand vorgegebener Header bzw. Metadaten überprüft werden soll, ob bei der vorliegenden Datei tatsächlich alle Normpunkte erfüllt werden, um diese in ein PDF/A - Dokument umwandeln und abschließend im Archiv ablegen zu können.

3.2.2 Verwendete Software

Da bereits eine große Auswahl an Programmen und Herstellern besteht, die ihre Produkte als Trial oder als Freeware Version für die OCR - Umwandlung zur Verfügung stellen, seien hier, ohne näher ins Detail zu gehen, einige der federführenden Hersteller von Softwareprodukten genannt.

- intarsys consulting GmbH (http://www.intarsys.de/, Stand: Juli 2015)
- Solid Documents (http://www.soliddocuments.com/, Stand: Juli 2015)
- PDF Tools AG (http://www.pdf-tools.com/, Stand: Juli 2015)
- LuraTech Solutions GmbH (https://www.luratech.com/, Stand: Juli 2015)
- Nuance Communications (http://www.nuance.de, Stand: Juli 2015)

Folgend soll hier auf die von Nuance Communications verwendete Software *Nuance OmniPage Ultimate* (Version: 19.0) näher eingegangen werden, welche primär für die Texterkennung, sowie auch die Umwandlung von Dokumenten in ein digitales Format entwickelt wurde (Nuance Communications, 2015a).

An dieser Stelle sei außerdem angemerkt, dass die Texterkennung mittels OCR nicht Kernthema der Masterarbeit ist, weshalb hier auch keine intensiveren Recherchen bezüglich Textverarbeitungsprogrammen durchgeführt wurden und daher auch die Wahl der Software von Nuance Communications aufgrund ausgewählter Kriterien (z.B. Gratis - Testversion, keine Einschränkungen der Software durch eine Lizenz für den Testzeitraum, kein Wasserzeichen auf den bearbeiteten Dokumenten, Konvertierung der Dateien in ein PDF/A - Format) erfolgte.

3.2.2.1 OmniPage - Über die Software

Nuance Communications ist eine amerikanische Firma, gegründet im Jahre 1992, mit Sitz in den USA (Burlington, Massachusetts). Seit Gründung beschäftigt sich die Firma im Bereich der Sprach - und Bildverarbeitungstechnologie (Nuance Communications, 2015b; 2015c) und ist auch Mitglied des PDF/A Competence Centers (PDF Association, 2015f).

Neben den verschiedenen von Nuance Communications entwickelten Produkten, sei hier auf die in dieser Masterarbeit verwendeten Software *Nuance OmniPage Ultimate* eingegangen, welche primär für die Bearbeitung (z.B. Texterkennungsfunktion) verschiedenster Eingangsdaten entwickelt wurde. Die Software ermöglicht damit beispielsweise Bilder oder PDF - Dateien durchsuchbar zu machen und je nach Anwendung, diese in verschiedenen Formaten, wie etwa als PDF -, XML - oder als Word - Datei, abzuspeichern. Auch das Langzeitarchivformat PDF/A wird in den Versionen PDF/A - 1b, - 2b und - 2u unterstützt (Nuance Communications, 2015a).

Bei der vorliegenden Software handelt sich um ein Produkt, welches als Gratis - Testversion per E - Mail für 15 Tage für verschiedene Betriebssysteme (z.B. Windows XP oder Windows 8) erworben werden kann. Für die Nutzung der Software ist lediglich ein Download dieser in dem mitgelieferten Link notwendig.

Basierend auf Nuance Communications (2015a) sollen folgend auch einige wichtige Funktionen der verwendeten Software (Version: 19.0) aufgelistet werden:

- Gescannte Dateien können mit Hilfe der Software durchsuchbar gemacht und wiederum als digitale Dokumente abgespeichert werden.
- Verschiedene Ein - und Ausgabeformate, wie etwa PDF, Bilddateien oder Microsoft Office Produkte, werden unterstützt.
- Die Software kann Dokumente in über 120 Sprachen bearbeiten und auch Inhalte, mit Hilfe der integrierten Sprachtechnologie der Software, im Dokument vorlesen.
- Das Langzeitarchivformat PDF/A wird von der Software in verschiedenen Versionen (PDF/A - 1b, - 2b und - 2u) unterstützt.

3.2.2.2 Umgang mit der Software

Nuance OmniPage Ultimate ist eine einfach zu bedienende Software. Nach Start der Software öffnet sich zunächst ein Anzeigefenster (Abbildung 15), welches einerseits anzeigt, wie lange die jeweilige Testversion noch genutzt werden kann und andererseits darauf hinweist, dass die Eingabe einer Seriennummer möglich ist, falls die Software gekauft wird.

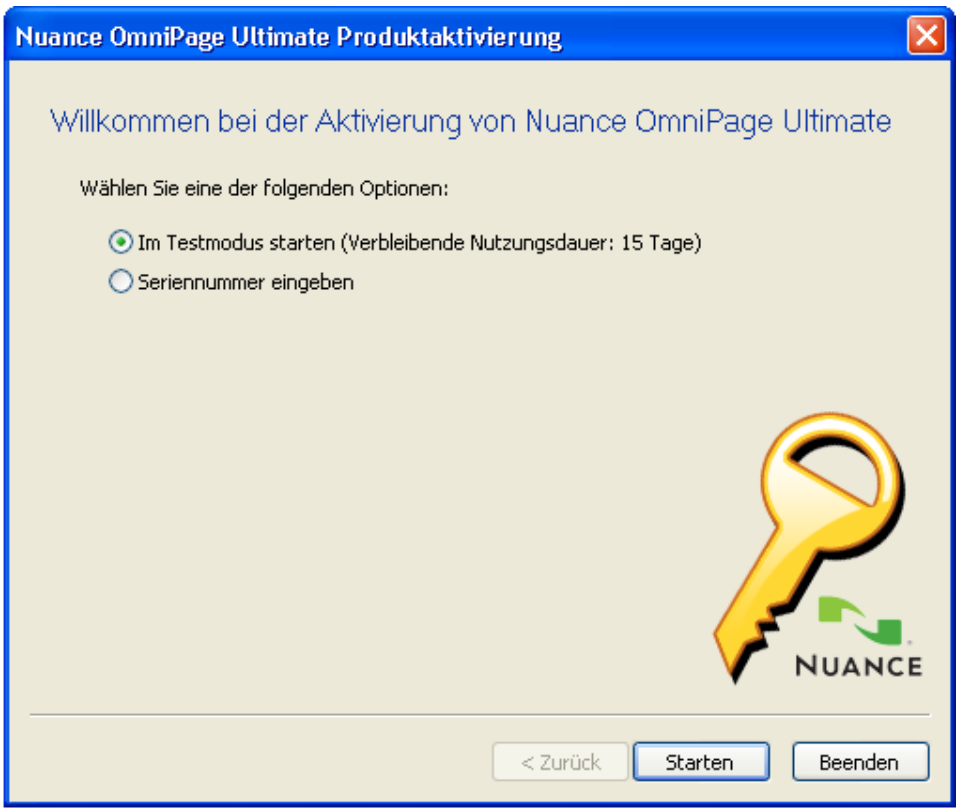

Abbildung 15: Screenshot Nuance OmniPage Ultimate Produktaktivierung

Nach Bestätigung, öffnet sich das dazugehörige Arbeitsfenster, welches verschiedene Wahlmöglichkeiten bietet, Dokumente verschiedenster Art zu laden und zu bearbeiten. Möchte der Benutzer beispielsweise gescannte Dokumente durchsuchbar machen, reicht ein einfacher Klick auf die bereits in der Software definierten Arbeitsprozesse (Abbildung 16). Auch die Erstellung eines neuen, selbstdefinierten Prozesses, sowie auch das Abändern eines bereits bestehenden Prozesses ist ebenfalls möglich (Abbildung 17).

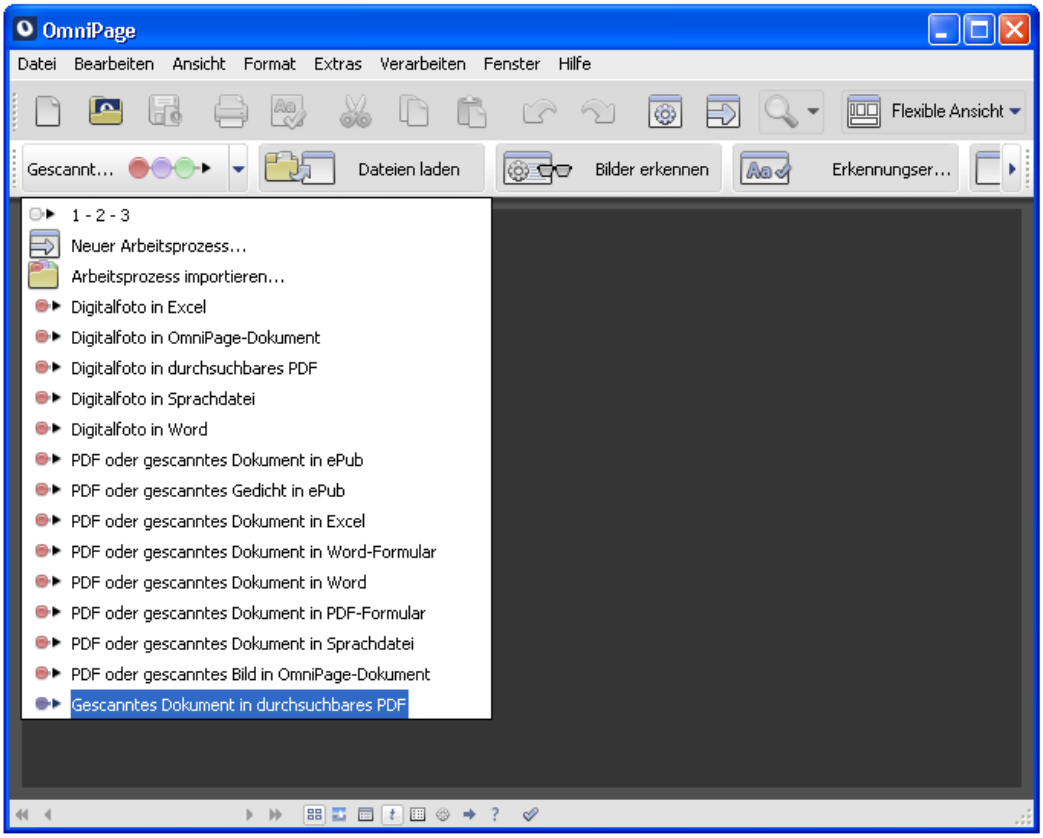

Abbildung 16: Vordefinierte Arbeitsprozesse

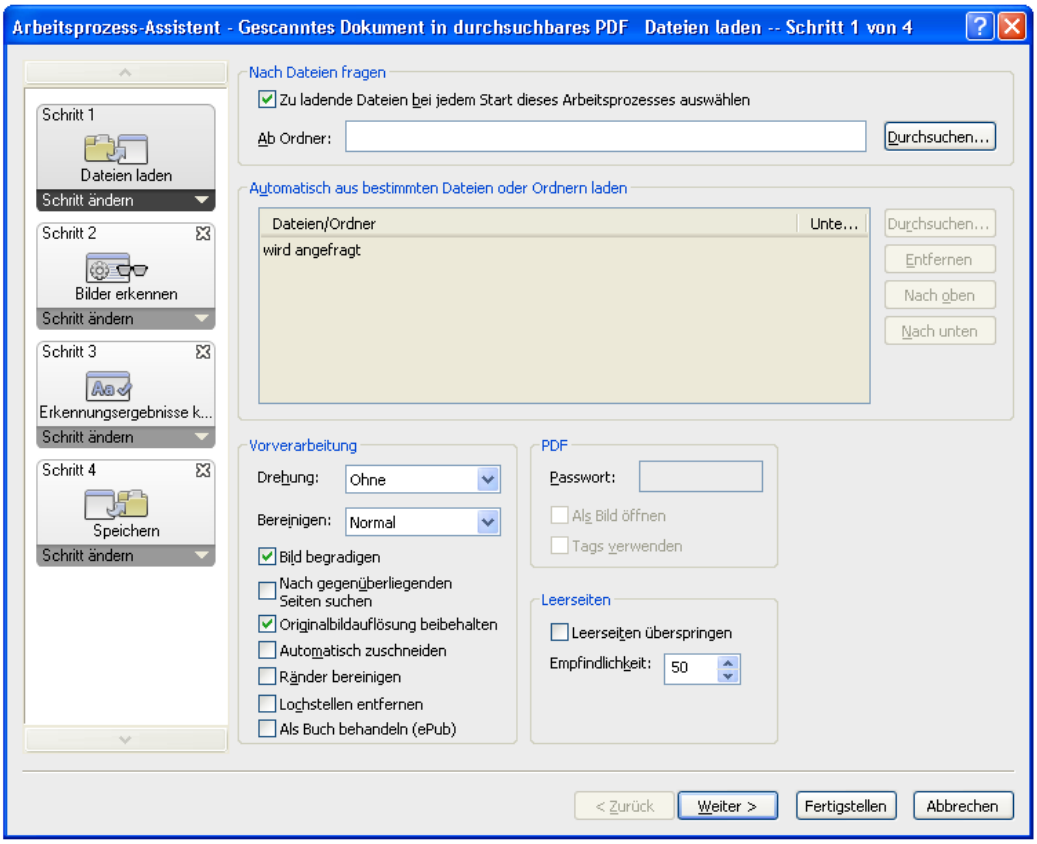

Abbildung 17: Ändern eines vordefinierten Arbeitsprozesses

Bevor die OCR - Prüfung abgeschlossen werden kann, bietet die Software Korrekturvorschläge an, um Zeichen bzw. Wörter richtig zu erkennen und zu interpretieren (Abbildung 18). Die angezeigten Korrekturvorschläge können zudem durch den Benutzer selbst manuell abgeändert werden. Soll jedoch keine Korrektur von Zeichen bzw. Wörtern vorgenommen werden, können die vorgeschlagenen Korrekturen ignoriert und damit der Text so beibehalten werden, wie ursprünglich im Dokument detektiert.

Abbildung 18: Korrekturvorschläge zur richtigen Zeicheninterpretation

Da die Langzeitarchivierung mittels PDF/A im Vordergrund steht, besteht hier die Möglichkeit, nach Durchführung des OCR - Prozesses, dieses Format automatisch als Ausgabeformat auszuwählen (Abbildung 19).

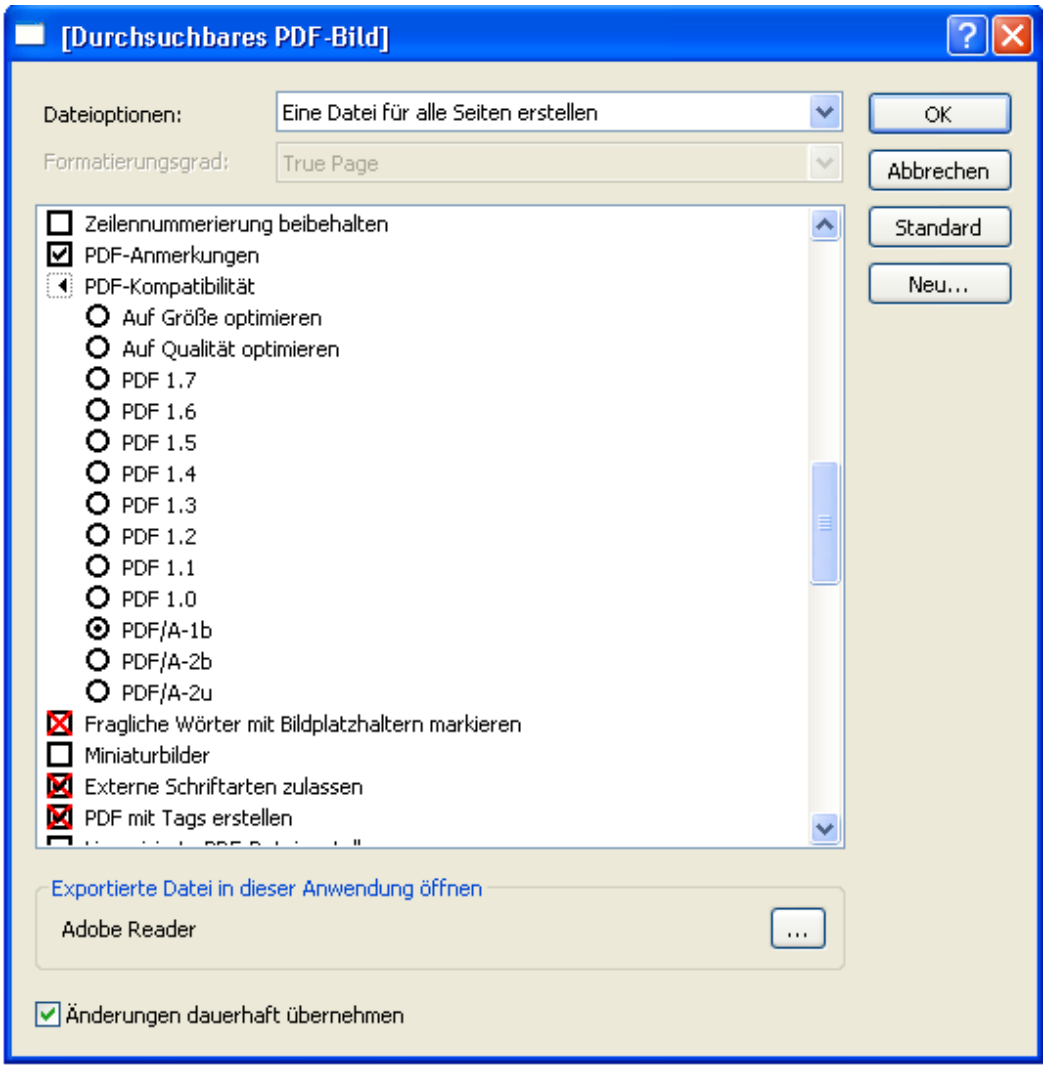

Abbildung 19: Wahl des Ausgabeformats PDF/A

Für detaillierte Informationen möchte der Leser hier auf das Manual der verwendeten Software von Nuance Communications (Nuance Communications, 2013) weiter verwiesen werden.

4 Praktische Umsetzung

Folgendes Kapitel beschreibt die vorhandenen und verwendeten Datenbestände, die im Zuge der Durchführung herangezogen wurden.

Weiters soll ein Überblick über relevante Zwischenergebnisse der Durchführung (z.B. Screenshots zur PDF/A - Erzeugung, Fehler -/ und Analysereports) zur durchgeführten Arbeit gegeben werden.

4.1 Rohausgangsdaten und Datenaufbereitung

Vor der Durchführung der Arbeit, mussten die dazu benötigten Geo - Dokumente eingeholt und auch geeignet vorbereitet werden. Bei den vorliegenden Dokumenten handelt es sich um unterschiedliche Dateitypen (z.B. .pdf, .dxf, .jpg, .mdb), die als eigenständige Dokumente vorliegen und nicht selbst (z.B. aus Datenbanken oder AutoCAD) generiert wurden. Im Zuge der Durchführung des praktischen Teils der Masterarbeit, wurde zwar das Augenmerk vorwiegend auf PDF - bzw. PDF/A - Dateien gelegt, jedoch erfolgte hier keine Aussortierung der Dateien nach der Dateinamenserweiterung, da diese aufgrund von manuellen Nachbearbeitungsfehlern eine fehlerhafte Dateiendung haben könnten (z.B. könnte die Datei fälschlicherweise *„.pdg“* anstatt *„.pdf“* heißen). Dementsprechend erfolgte daher bei der Durchführung des praktischen Teils eine Prüfung der Datei auf dessen Inhalte, bei der anhand vorgegebener Header bzw. Metadaten, alle PDF - bzw. PDF/A - Dateien aussortiert werden konnten.

Alle vorliegenden Dokumente wurden von verschiedenen Autoren, Herstellern, Planverfassern, Diensten, Institutionen und Eigentümern übernommen, deren Zustimmung (Urheberrecht) zur Verwendung im Rahmen der Masterarbeit vorliegt. Auch Muster - bzw. Demodaten, die frei verfügbar und über das Internet abrufbar sind, wurden hier berücksichtigt, sowie auch einzelne Testdaten selbst generiert. Insgesamt standen somit 390 Geo - Datensätze zur Verfügung, welche folgend überblicksmäßig aufgezählt sind:

- **Stadt Graz - Stadtplanungsamt**
  (http://www.graz.at/cms/ziel/1345767/DE/, **Stand: 10.06.2015)**
  - Bebauungspläne
  - Flächenwidmungspläne
  - Musterbilder (Gesamtbilder, Luftbilder, Orthophotos)

- **BEV - Bundesamt für Eich - und Vermessungswesen, Gratis - Musterdaten (**http://www.bev.gv.at/portal/page?_pageid=713,2031039&_dad=portal&_schema=PORTAL, **Stand 11.06.2015)**
  - Bodenschätzungsergebnisse
  - Digitale Geländehöhenmodelle

  - Digitale Landschaftsmodelle
  - Grundlagenvermessungen (Festpunkte)
  - Kataster und Verzeichnisse
  - Landkarten
  - Luftbildprodukte
  - Österreichisches Adressregister

- **Stadtvermessungsamt (Stand: 24.02.2015)**
  - Vermessungsurkunden
  - Teilungspläne
  - Qualitätsverbesserungen
  - Mappenberichtigungen

- **Private Quelle: Vermessungspläne VHW (3/93, 46/61, 9/62), KG 63282 (Stand: 20.01.2015)**
  - Bestandspläne (Grundstück 509/2)
  - Gebäudepläne
  - Umbaupläne
  - Teilungspläne
  - Mappenberichtigungen

4.2 Screenshots relevanter Zwischenergebnisse

Folgend sei, jeweils für den PDF/A - Prozess sowie auch für den OCR - Prozess, ein aus den verfügbaren Datenbeständen ausgewähltes Beispiel gezeigt, welches mit Hilfe von Screenshots die durchgeführte Arbeit bzw. die Prozessabläufe dokumentieren soll. Ferner seien in Kapitel 5 die wesentlichen Ergebnisse der Durchführung gezeigt. Alle verfügbaren Daten, Protokolle und Ergebnisse über die Durchführung des praktisches Teils der Arbeit, sind in der beiliegenden DVD zu finden.

4.2.1 PDF/A - Prozess

Folgend sei die praktische Umsetzung des PDF/A - Prozesses anhand eines beliebig gewählten Dokuments aus den verfügbaren Daten gezeigt. Um den Prozessablauf sowohl bildlich als auch textlich folgen zu können, seien hier die einzelnen Schritte der in Kapitel 3.1 genannten Ablaufdiagramme beschrieben.

Schritt 3.1.1.1: Eingangsdaten

- Vermessungsurkunde - Qualitätsverbesserung
  - 010222-2014_Q-Plan.dxf
  - 010222-2014_Q-Plan.pdf
  - 010222-2014_Q-Plan_pdfa_asig.pdf

Schritt 3.1.1.2: Dateianalyse (PDF, PDF/A Vorprüfung)

- Laden der Dateien
  - Die oben genannten PDF - Dateien wurden erfolgreich geladen (Abbildung 20 und Abbildung 21).
  - Das Öffnen der DXF - Datei ergab Probleme (Abbildung 22).
  - Um zu Überprüfen, ob die Software anhand vorgegebener Header bzw. Metadaten überprüft, ob die vorliegende Datei eine PDF - Datei oder eine sonstige Datei ist, wurde die Dateinamenserweiterung einer der PDF - Dateien von „*.pdf*“ in „*.xyz*“ umbenannt. Zusätzlich wurde die Dateinamenserweiterung der DXF - Datei von „*.dxf*“ in „*.pdf*“ umbenannt. Die Software erkennt jedoch in beiden Fällen, welche Datei eine PDF - Datei und welche eine sonstige Datei ist.

Abbildung 20: Laden der Dateien (010222-2014_Q-Plan.pdf)

Abbildung 21: Laden der Dateien (010222-2014_Q-Plan_pdfa_asig.pdf)

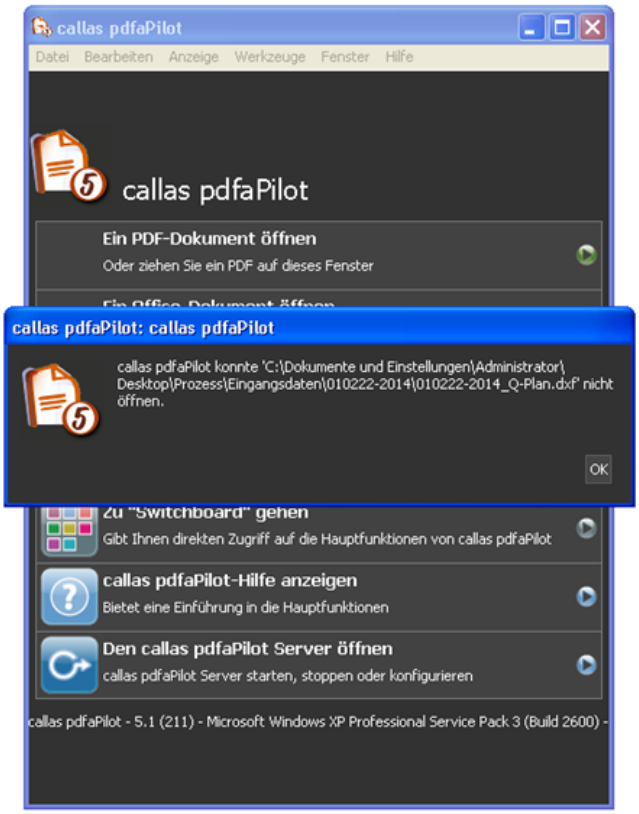

Abbildung 22: Laden der Dateien (010222-2014_Q-Plan.dxf)

Schritt a: Dateianalyse (PDF, PDF/A Vorprüfung)

- Sonstige Dateiformate (≠ PDF, ≠ PDF/A)
  - Da das Öffnen der DXF - Datei Probleme ergab (Abbildung 22), wird diese Datei nicht weiter behandelt, da es sich hierbei **nicht um eine PDF - bzw. PDF/A - Datei** handelt (Schritt a).
  - Im Zuge dessen erfolgt hier keine Erzeugung einer PDF - bzw. PDF/A - Datei (Schritt a1).
  - Damit ist der Vorgang an dieser Stelle abgeschlossen.

Schritt b: Dateianalyse (PDF, PDF/A Vorprüfung)

- Dateien auf PDF - bzw. PDF/A - Format überprüfen
  - Mit Hilfe der Software (Abbildung 23) kann anhand der Detailprüfung entschieden werden, ob die vorliegende Datei eine PDF - bzw. PDF/A - Datei ist.

Abbildung 23: Dateianalyse (PDF - , PDF/A - Vorprüfung)
(Screenshot bearbeitet im März 2015)

Schritt c und d: Dateiformat aus Detailprüfung

- An dieser Stelle werden PDF/A - Dateien herausgefiltert, während PDF - Dateien mittels einer der beiden Varianten weiter behandelt werden.

Schritt c: 010222-2014_Q-Plan_pdfa_asig.pdf

- Hier wird erkannt, dass die vorliegende Datei **eine gültige PDF/A - Datei** (PDF/A - 1b) ist, welche nicht weiter untersucht werden muss (Abbildung 24).
- Der dazugehörige Analyse -/Prüfbericht befindet sich im Anzeigefenster der Software.
- Damit ist der Vorgang an dieser Stelle abgeschlossen.

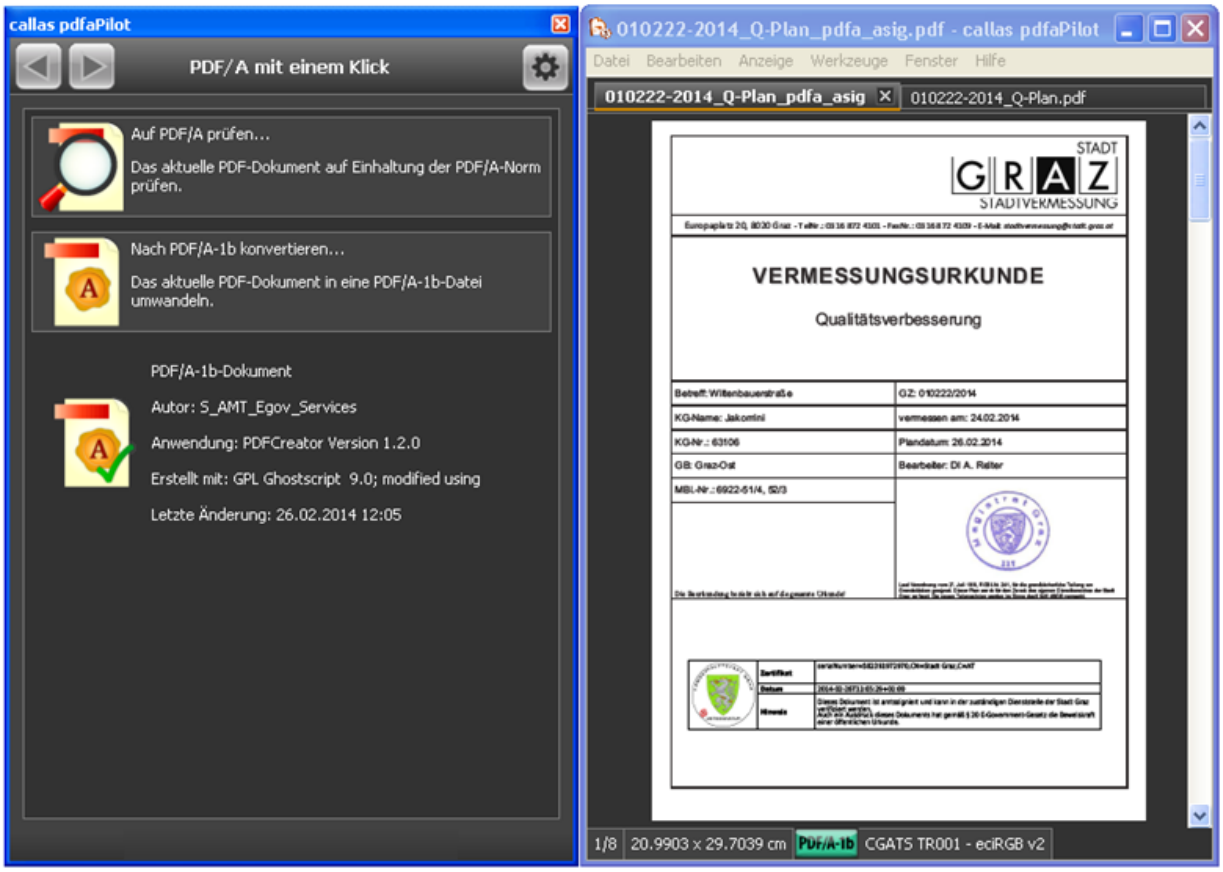

Abbildung 24: Überprüfung auf PDF/A - Version
(Screenshot bearbeitet im März 2015)

Schritt d: 010222-2014_Q-Plan.pdf

- Hier wird erkannt, dass die vorliegende Datei **keine gültige PDF/A - Datei**, aber eine gültige PDF - Datei, ist (Abbildung 25), welche im Zuge des Prozesses mittels einer der beiden Varianten weiter untersucht werden muss (Schritte e und f).
- Mit Hilfe des Anzeigefensters der Software, können alle für die jeweils gewählte PDF/A - Version fehlenden Informationen bzw. Fehler angezeigt werden, sowie auch ein Report über alle Fehlermeldungen erzeugt werden.

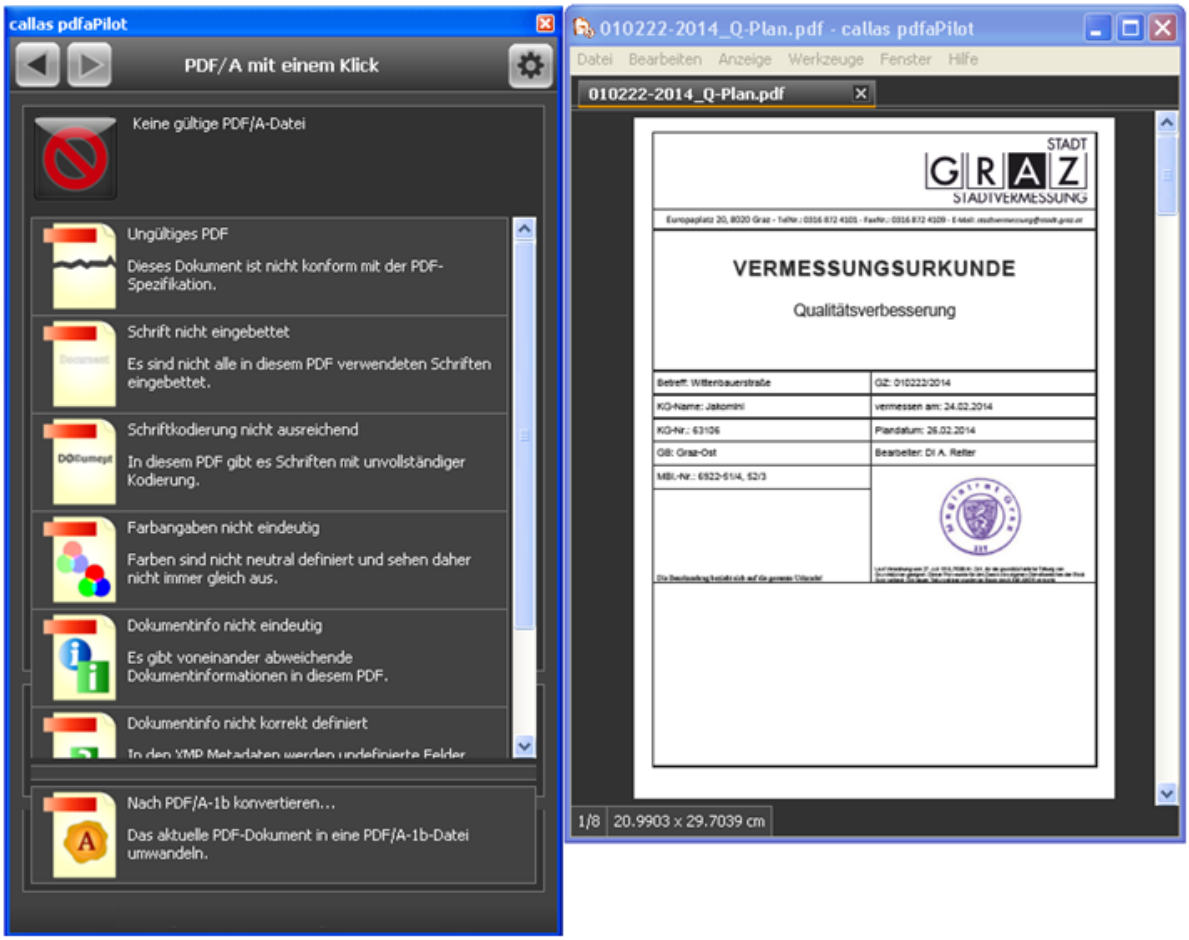

Abbildung 25: Überprüfung auf PDF/A - Version
(Screenshot bearbeitet im März 2015)

4.2.1.1 Variante 1 - „Match & Fix“

Schritt 3.1.1.3: Dateianalyse

- Anhand der Dateianalyse des zuvor durchgeführten Prozesses und des vorliegenden Analyse -/Prüfberichts wird erkannt, dass die vorliegende Datei eine PDF - Datei ist, die hier mittels Variante 1 weiter behandelt wird.
- An dieser Stelle ist zu überprüfen, ob die Auswahl des geeignetsten Formats ohne Dateieingriff möglich ist.

Schritt a: Match - Pfad

- Anhand der erlaubten Inhalte der PDF/A - Datei ist zu überprüfen, ob *ein* („1“) oder *mehrere* („n“) mögliche PDF/A - Formate vorliegen, um eine abschließende Konvertierung in ein gewähltes PDF/A - Format ohne Eingriff in die Datei durchführen zu können.
- Da die Software das eigentliche „Matching“ (Analyse der Ausgangsdatei und Gegenüberstellung der verschiedenen Normanforderungen) nicht unterstützt, kann hier nur versuchsweise ermittelt werden, welche PDF/A - Varianten möglich sind bzw. wo unter Umständen Fehler zu korrigieren wären (gegebenenfalls durch Autokorrektur). Dazu ist mit Hilfe der Software an dieser Stelle in alle mögliche Formate (beginnend bei der höchst wählbaren Variante) zu konvertieren (Schritt a1 bzw. a2).
- Beispielhaft sei hier die Konvertierung der PDF/A - 3a zur PDF/A - 3u gezeigt (Abbildung 26, Abbildung 27 und Abbildung 28).

PDF/A - 3a:

- Eine Konvertierung in die gewählte Version ist hier nicht möglich, da hier die Strukturinformation fehlt und die Schriftkodierung nicht ausreichend ist.
- Mit Hilfe des Anzeigefensters der Software, können alle für die jeweils gewählte PDF/A - Version fehlenden Informationen bzw. Fehler angezeigt werden, sowie auch ein Report über alle Fehlermeldungen erstellt werden (Abbildung 26).

= Autokorrektur möglich, = keine Autokorrektur möglich,
= nicht erfolgreich durchgeführte Autokorrektur

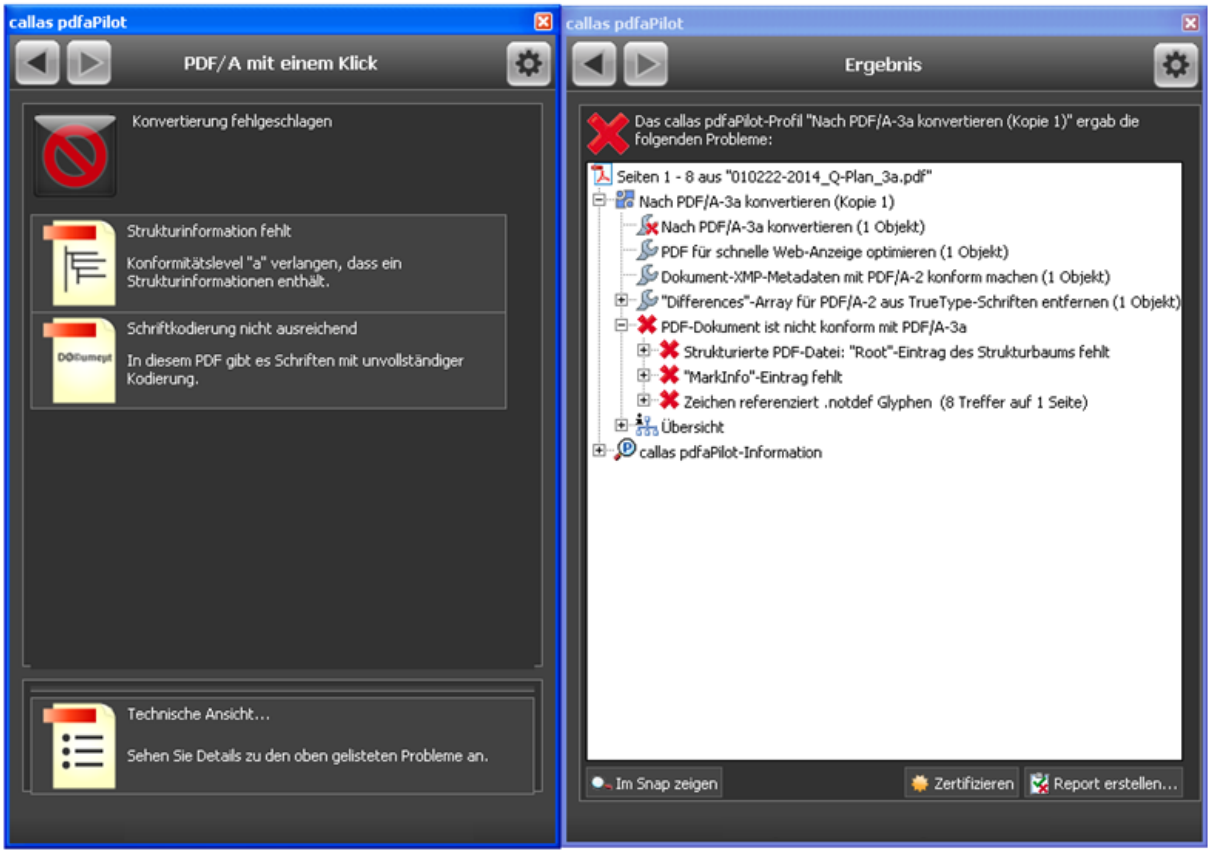

Abbildung 26: Konvertierung in gewählte Version nicht erfolgreich (Screenshot bearbeitet im März 2015)

PDF/A - 3b:

- Eine Konvertierung in die gewählte Version ist hier möglich.
- Der dazugehörige Analyse -/ Prüfbericht bzw. der Report über die erfolgreiche Konvertierung befindet sich im Anzeigefenster der Software (Abbildung 27).

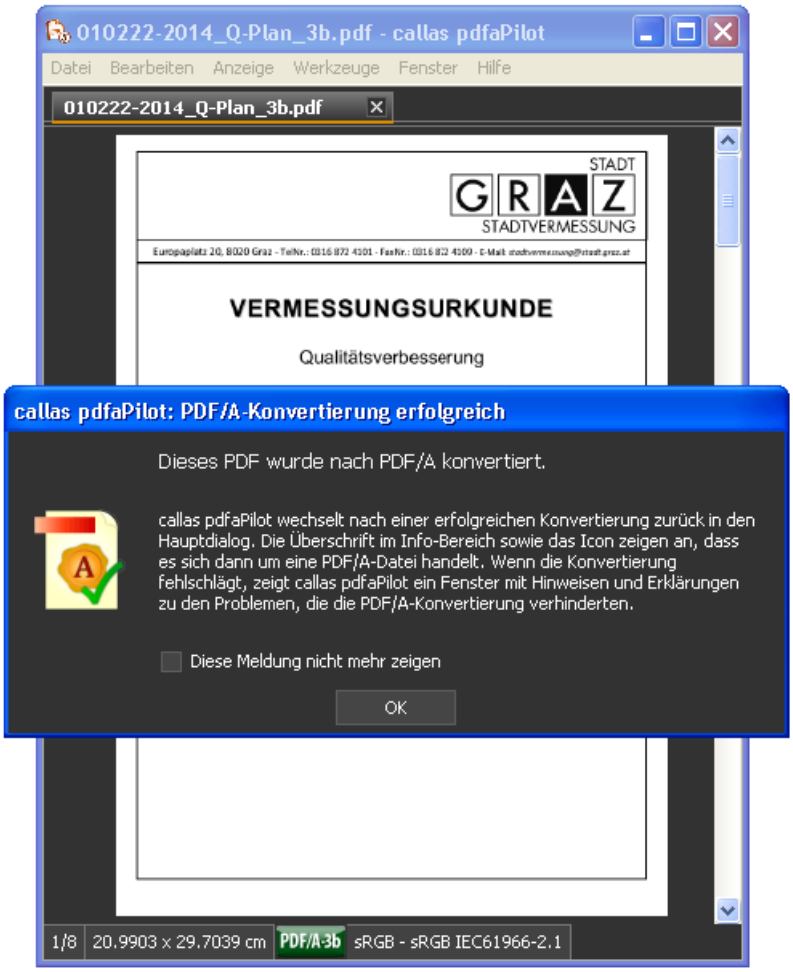

Abbildung 27: Konvertierung in gewählte Version erfolgreich

PDF/A - 3u:

- Eine Konvertierung in die gewählte Version ist hier möglich.
- Der dazugehörige Analyse -/ Prüfbericht bzw. der Report über die erfolgreiche Konvertierung befindet sich im Anzeigefenster der Software (Abbildung 28).

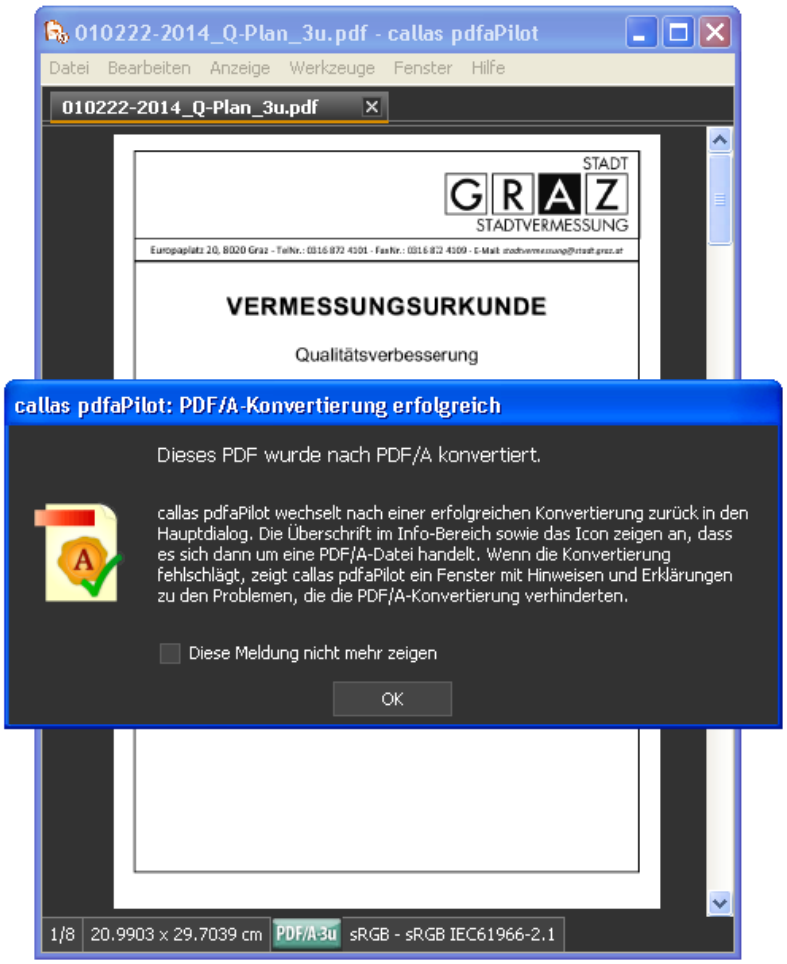

Abbildung 28: Konvertierung in gewählte Version erfolgreich

Ergebnis (Schritt a3): Konvertierung

- Anhand der erlaubten Inhalte der Datei wurde überprüft, wie viele mögliche PDF/A - Formate für die Konvertierung gewählt werden können.
  - Nach Durchführung der Konvertierungen wird ersichtlich, dass insgesamt fünf mögliche Formate (PDF/A - 1b, - 2b, - 2u, - 3b, - 3u) für eine abschließende Konvertierung in ein gewähltes PDF/A - Format gewählt werden können.
  - Die Entscheidung, welches der möglichen Formate gewählt werden soll, ist abhängig vom Benutzer.

Check gegenüber Originaldatei (Qualitätsprüfung):

- Da jegliche Konvertierung einen Eingriff in die Datei darstellt, ist jede konvertierte Datei als „neue“ Datei zu behandeln und mit der Originaldatei zu vergleichen (Qualitätsprüfung).
- Ein Vergleich (Pixelvergleich) der Dateien kann mit Hilfe der Software automatisch ausgeführt werden, jedoch ist dies hier nur bedingt möglich, da die Software (laut *Dietrich von Seggern* von der *callas software GmbH* (http://www.callas

software.com/de) und gemäß der erzielten Ergebnisse, die sich auf der beiliegenden DVD befinden) für einen Vergleich von konvertierten Dateien nicht gedacht bzw. geeignet ist. Die Empfindlichkeit beim Vergleich der Dateien ist hier derart groß, dass bereits geringste Unterschiede (z.B. Farb - oder Transparenz-unterschiede) von der Software ausgegeben werden.

- An dieser Stelle wäre eine geeignetere Software für den Vergleich von Originaldatei und konvertierter Datei notwendig, wobei im Rahmen der Masterarbeit dazu keine Nachforschungen betrieben wurden.

Schritt b: Fix - Pfad

- Anhand des Analyse -/Prüfberichts wird ersichtlich, dass die vorliegende Datei Fehler und damit nicht alle erlaubten, notwendigen bzw. normierten Inhalte der jeweiligen PDF/A - Version beinhaltet. Damit ist auch die Konvertierung in das gewählte PDF/A - Format nicht ohne Eingriff in die Datei möglich ist.
- An dieser Stelle ist außerdem zu überprüfen, ob „*leichte*“ (Schritt c) oder „*schwere*“ Fehler (Schritt d) vorhanden sind, unter der Annahme, dass die PDF/A - 1a Version die geeignetste, gewünschte oder gewählte Variante ist.

Schritt c: Fehler vorhanden („leichte“ Fehler)

- Als „*leichte*“ Fehler werden all jene bezeichnet, die mit Hilfe der Software selbst korrigiert werden können.
- An dieser Stelle ist daher zu entscheiden, ob ein Eingriff in die Datei stattfinden soll oder nicht.

Schritt c2: Eingriff in die Datei

- Sollte ein Eingriff in die Datei stattfinden, ist darauf zu achten, dass die Inhalte dieser verändert werden könnten, was zur erneuten Durchführung des Prozesses und erneuten Entscheidungsmaßnahmen bezüglich der erlaubten Inhalte der verschiedenen PDF/A - Formate führt.

Schritt c21: Eingriff in die Datei

- Mit Klick auf die jeweilige Fehlermeldung wird ersichtlich, was korrigiert werden müsste, um die jeweils gewählte PDF/A - Version zu erhalten. Beispielhaft sei hier die Fehlermeldung „*Schrift nicht eingebettet*“ gezeigt (Abbildung 30).
- Die Software selbst bietet vordefinierte Profile, Prüfungen und Korrekturen an, um etwaige Fehlermeldungen verbessern zu können.
- Zudem kann nach jeder durchgeführten Korrektur ein dazugehöriger Report über das erhaltene Ergebnis errichtet werden.
- Da jeglicher Eingriff in die Datei eine Manipulation dieser darstellt, ist nach jeder Korrektur die Datei als „neue“ Datei zu behandeln und mit der Originaldatei zu vergleichen (Qualitätsprüfung).

- Es wird versucht, alle hier auftretenden Fehlermeldungen (Abbildung 29) mit Hilfe der Software zu korrigieren.
- Nach jedem Schritt wird die geänderte Datei mit der Originaldatei verglichen und anschließend der Prozess erneut durchgeführt (Überprüfung auf PDF/A - Version), um zu überprüfen, ob und welche Fehler bereits korrigiert wurden bzw. ob die gewählte PDF/A - Version mit der jeweiligen Korrektur erreicht wurde.
- Ein Ergebnis einer Korrektur kann Abbildung 31 entnommen werden.

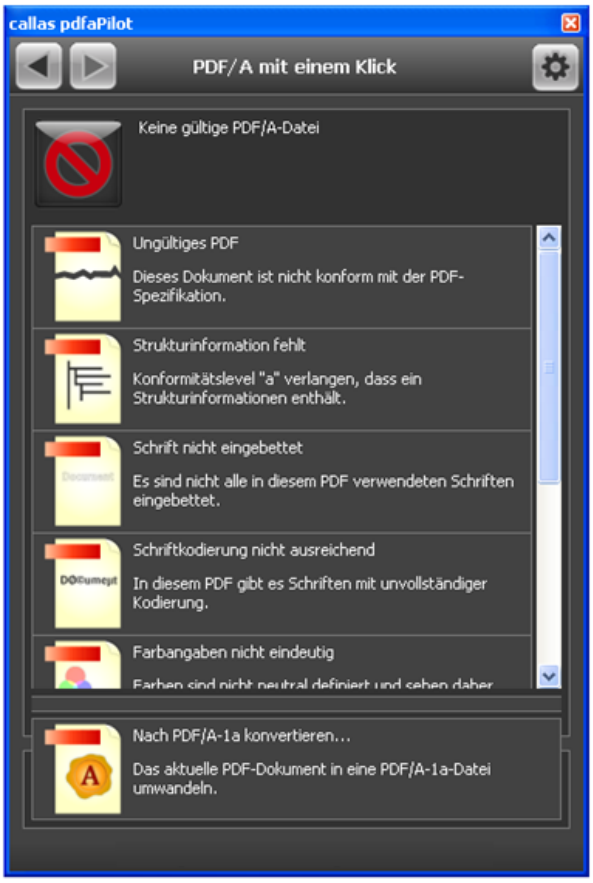

Abbildung 29: Fehlermeldungen (Screenshot bearbeitet im März 2015)

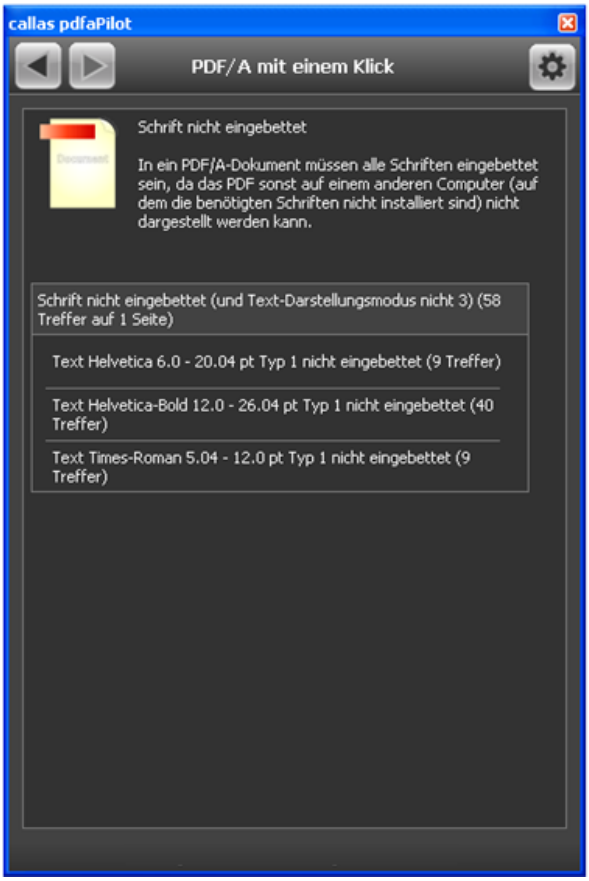

Abbildung 30: Anzeige einer Fehlermeldung (Screenshot bearbeitet im März 2015)

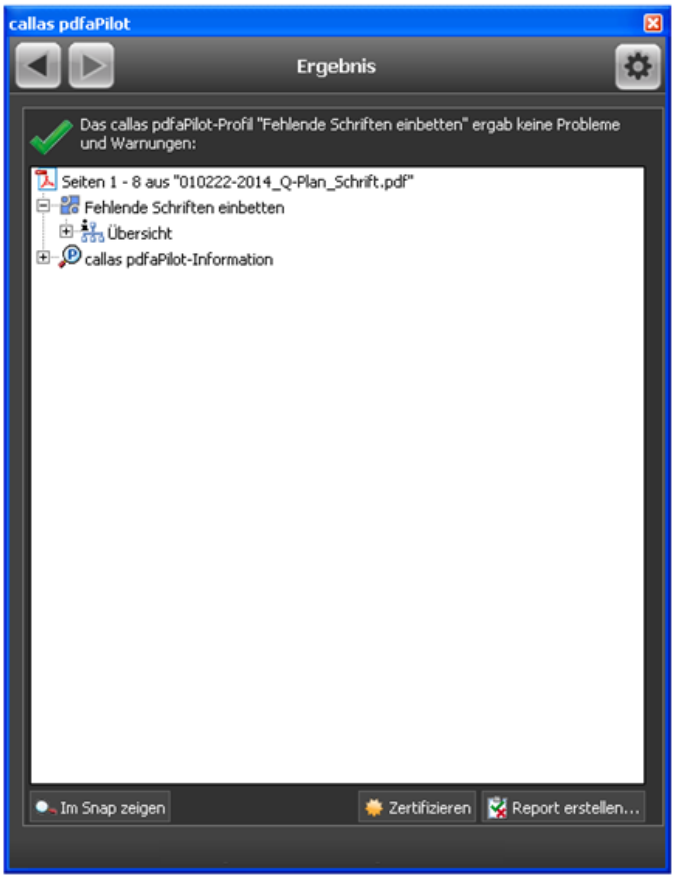

Abbildung 31: Beispielhaftes Ergebnis einer Korrektur (Screenshot bearbeitet im März 2015)

Ergebnis:

- Nach Durchführung aller möglicher Korrekturen wird ersichtlich, dass alle „leichten“ Fehler mit Hilfe der Software korrigiert werden können.
- Da die Software Strukturinformationen nicht korrigieren kann, handelt es sich hierbei um einen „schweren“ Fehler (Schritt d), der programmtechnisch nicht mit Hilfe der Software korrigiert werden kann.

Check gegenüber Originaldatei (Qualitätsprüfung):

- Da jegliche Korrektur einen Eingriff in die Datei darstellt, ist jede korrigierte Datei als „neue“ Datei zu behandeln und mit der Originaldatei zu vergleichen (Qualitätsprüfung).
- Ein Vergleich der Dateien erfolgt mit Hilfe der Software (Abbildung 32).
- Abbildung 33 zeigt einen Vergleich zwischen der konvertierten Datei und der Originaldatei.

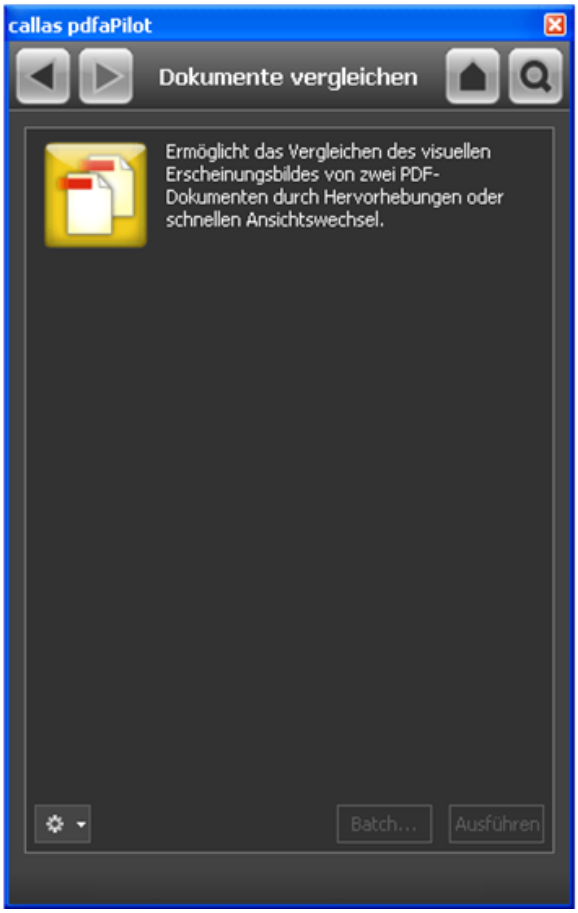

Abbildung 32: Check gegenüber PDF/A - Datei (Qualitätsprüfung, Screenshot bearbeitet im März 2015)

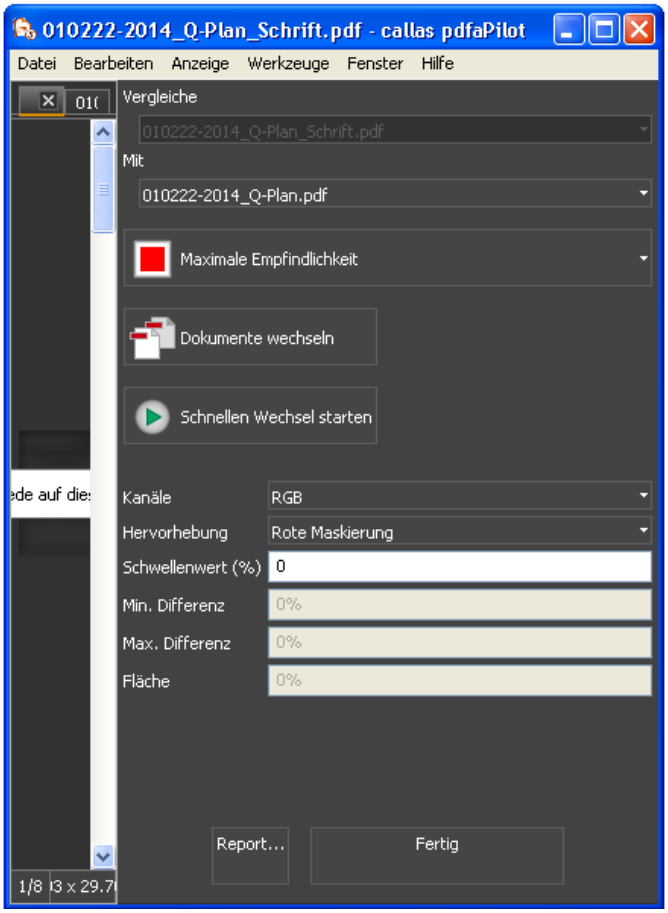

Abbildung 33: Vergleich (Korrektur Farbangaben mit Originaldatei)

Schritt d: Fehler vorhanden („schwere“ Fehler)

- Als „schwere“ Fehler werden all jene bezeichnet, die nicht mit Hilfe der Software selbst korrigiert werden können.
- An dieser Stelle ist daher zu entscheiden, ob ein Eingriff in die Datei stattfinden soll oder nicht.
- Da die Software Strukturinformationen nicht korrigieren kann, handelt es sich hierbei um einen „schweren“ Fehler, der programmtechnisch nicht mit Hilfe der Software korrigiert werden kann (Abbildung 34).
  - An dieser Stelle wird entschieden, nicht in die Datei einzugreifen (Schritt d2), da für Strukturinformationen die Originaldatei (und nicht die vorliegende PDF - Datei) bearbeitet werden müsste.

Abbildung 34: Ergebnis der durchgeführten Arbeit (Screenshot bearbeitet im März 2015)

4.2.1.2 Variante 2 - „Trial & Fix“

Schritt 3.1.1.4: Konvertierungsversuch

- Anhand der Dateianalyse des zuvor durchgeführten Prozesses und des vorliegenden Analyse -/Prüfberichts wird erkannt, dass die vorliegende Datei eine PDF - Datei ist, die hier mittels Variante 2 weiter behandelt wird.
- An dieser Stelle ist anhand des Konvertierungsversuchs zu überprüfen, ob die Auswahl des geeignetsten Formats mit oder ohne Dateieingriff möglich ist.
- Dazu ist mit Hilfe der Software an dieser Stelle sukzessive (bei jedem Prozess - bzw. Trial - Schritt) in alle möglichen Formate (beginnend bei der höchst wählbaren Variante) zu konvertieren.

PDF/A - 3a:

- Eine Konvertierung in die gewählte Version ist hier nicht möglich, da hier die Strukturinformation fehlt und die Schriftkodierung nicht ausreichend ist.
- Mit Hilfe des Anzeigefensters der Software, können alle für die jeweils gewählte PDF/A - Version fehlenden Informationen bzw. Fehler angezeigt werden, sowie auch ein Report über alle Fehlermeldungen erstellt werden (Abbildung 35).

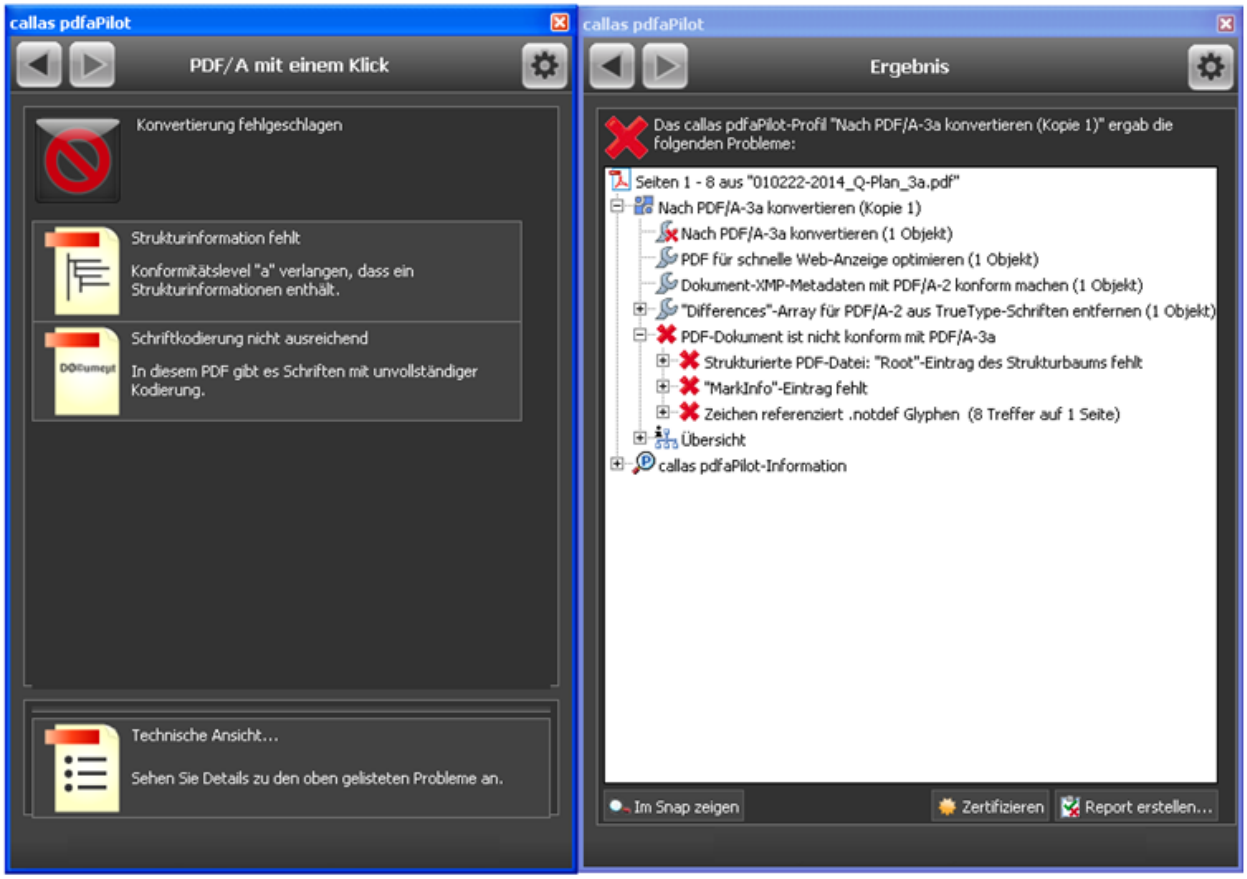

Abbildung 35: Konvertierung in gewählte Version nicht erfolgreich (Screenshot bearbeitet im März 2015)

PDF/A - 3b:

- Eine Konvertierung in die nächst niedrigere (gewählte) Version ist hier laut Prozessdefinition möglich.
- Der dazugehörige Analyse -/Prüfbericht bzw. der Report über die erfolgreiche Konvertierung befindet sich im Anzeigefenster der Software (Abbildung 36).

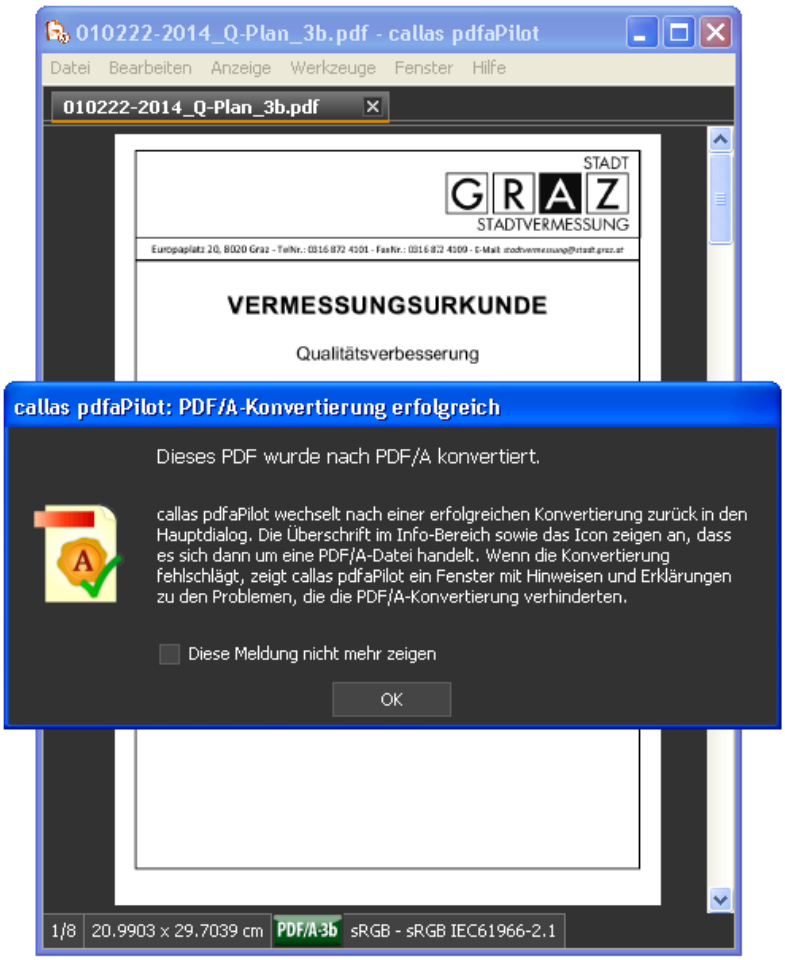

Abbildung 36: Konvertierung in gewählte Version erfolgreich

Check gegenüber Originaldatei (Qualitätsprüfung):

- Da jegliche Korrektur einen Eingriff in die Datei darstellt, ist jede korrigierte Datei als „neue“ Datei zu behandeln und mit der Originaldatei zu vergleichen (Qualitätsprüfung).
- Ein Vergleich (Pixelvergleich) der Dateien kann mit Hilfe der Software automatisch ausgeführt werden, jedoch ist dies hier nur bedingt möglich, da die Software (laut *Dietrich von Seggern* von der *callas software GmbH* (http://www.callassoftware.com/de) und gemäß der erzielten Ergebnisse, die sich auf der beiliegenden DVD befinden) für einen Vergleich von konvertierten Dateien nicht gedacht bzw. geeignet ist. Die Empfindlichkeit beim Vergleich der Dateien ist hier derart groß, dass bereits geringste Unterschiede (z.B. Farb - oder Transparenzunterschiede) von der Software ausgegeben werden.
- An dieser Stelle wäre eine geeignetere Software für den Vergleich von Originaldatei und konvertierter Datei notwendig, wobei im Rahmen der Masterarbeit dazu keine Nachforschungen betrieben wurden.

Ergebnis:

- Hier erfolgt eine Konvertierung in PDF/A - 3b.
- Der Vorgang wäre an dieser Stelle abgeschlossen.
- Sollte dennoch PDF/A - 3a erreicht werden wollen, muss dies anhand eines manuellen Eingriffs in die Originaldatei erfolgen, da es sich hierbei um einen „schweren“ Fehler handelt (Schritt e2), der programmtechnisch nicht mit Hilfe der Software korrigiert werden kann.

4.2.2 OCR - Prozess

Folgend sei die praktische Umsetzung des OCR - Prozesses anhand eines beliebig gewählten Dokuments aus den verfügbaren Daten gezeigt. Um den Prozessablauf sowohl bildlich als auch textlich folgen zu können, seien hier die einzelnen Schritte des in Kapitel 3.2 genannten Ablaufdiagramms beschrieben.

Schritt 3.2.1.1: Eingangsdaten

- Mappenberichtigung
  - VHW 9_62_1b.pdf

- Laden der Dateien
  - Die oben genannte PDF/A - Datei wurde erfolgreich in der Software *Nuance Omnipage Ultimate* geladen (Abbildung 37)
  - Mit Hilfe des OCR - Texterkennungsverfahrens soll das Dokument derart bearbeitet werden, sodass anschließend eine Volltextsuche in diesem ermöglicht werden kann bzw. auch nicht unmittelbar zugängliche Informationen (z.B. Rastertext) zugänglich gemacht werden können.
  - Das aus dem langzeitarchivierten Originaldokument abgeleitete OCR - Dokument, soll abschließend in das Langzeitarchivformat PDF/A umgewandelt und in der Software *callas pdfaPilot* mit der Eingangs - PDF/A - Datei verglichen werden.

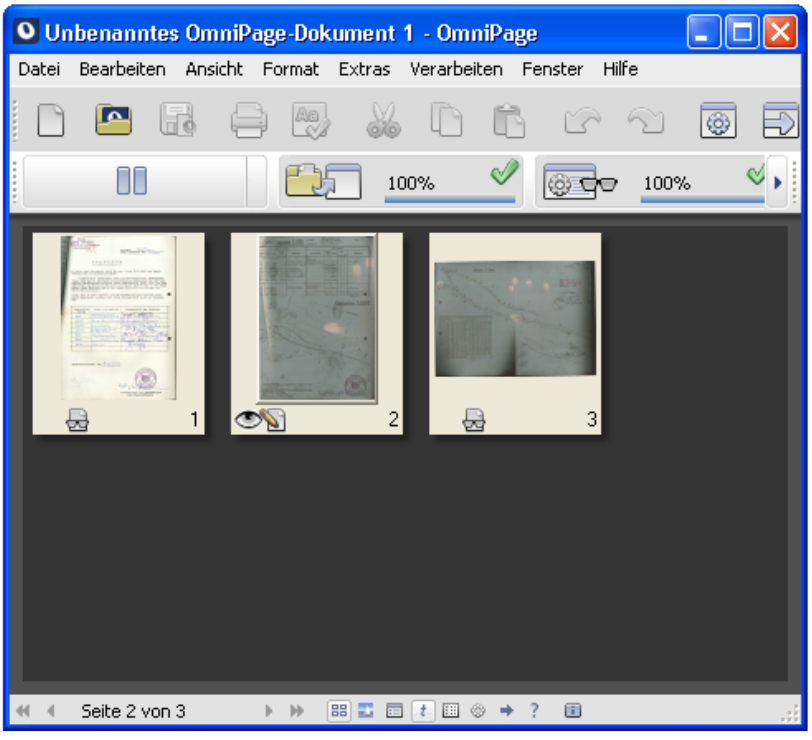

Abbildung 37: Laden der Dateien

Schritt 3.2.1.2: Split

- Mit Hilfe des Split - Funktion, welche im Programm integriert ist, werden zunächst alle sich auf dem Dokument befindenden Elemente in einzelne Ebenen aufgespalten.
- Die Software erkennt dabei, welche Ebene „Textelemente“ und welche die „grafische Elemente“ ist.

Schritt a: Ebene „Textelemente“

- Die Ebene „Textelemente“ wird von der Software derart bearbeitet, sodass hier einzelne Absätze, Zeilen, Wörter und Zeichen in dieser Ebene erkannt und richtig interpretiert werden.

Schritt b: Ebene „grafische Elemente“

- In der Ebene „grafische Elemente“ befinden sich alle Grafiken und Bilder.
- Befindet sich Text auf diesen, bietet die Software die Möglichkeit an, verschiedene Methoden und Algorithmen einzusetzen, um eine Texterkennung in Grafiken und Bildern zu ermöglichen.

Schritt c: OCR

- Beim OCR - Verfahren kommen verschiedene Methoden und Algorithmen zum Einsatz (z.B. Merkmals - und Mustererkennung), um eine optimale Texterkennung zu ermöglichen.
- Dabei werden vorwiegend gescannte Dokumente oder Bilddateien auf Zeichen untersucht, um alle darin nicht unmittelbar zugänglichen Inhalte zugänglich machen zu können.
- Die verwendete Software bietet dazu eine Erkennungsprüfung an, um alle in dem Dokument befindlichen Zeichen richtig zu erkennen und zu interpretieren, um somit auch beispielsweise das Verschmelzen von Zeichen (z.B. i wird zu l) oder falsche Interpretationen von Zeichen (z.B. ä wird zu a) und Wörtern zu vermeiden.
- Beispielhaft sei dies für das verwendete Dokument in Abbildung 38 bis Abbildung 41 gezeigt.

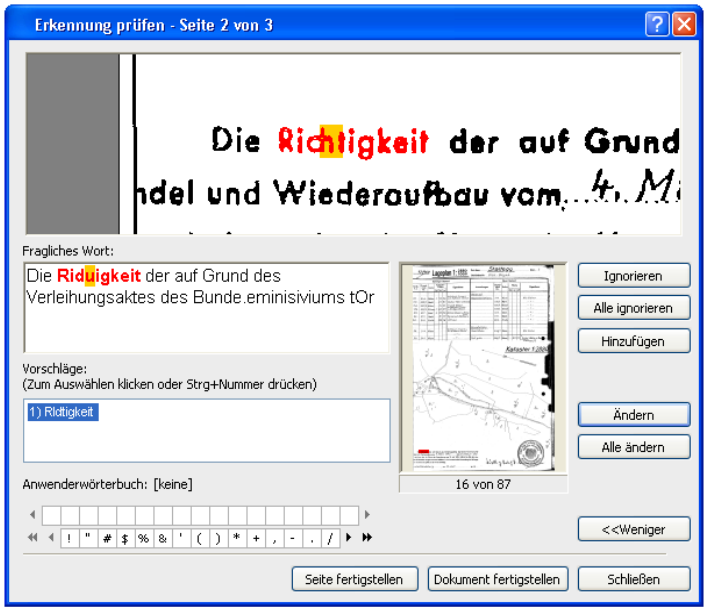

Abbildung 38: Beispiel einer Erkennungsprüfung

Abbildung 39: Erkennungsprüfung (Erkennung eines handgeschriebenen Wortes)

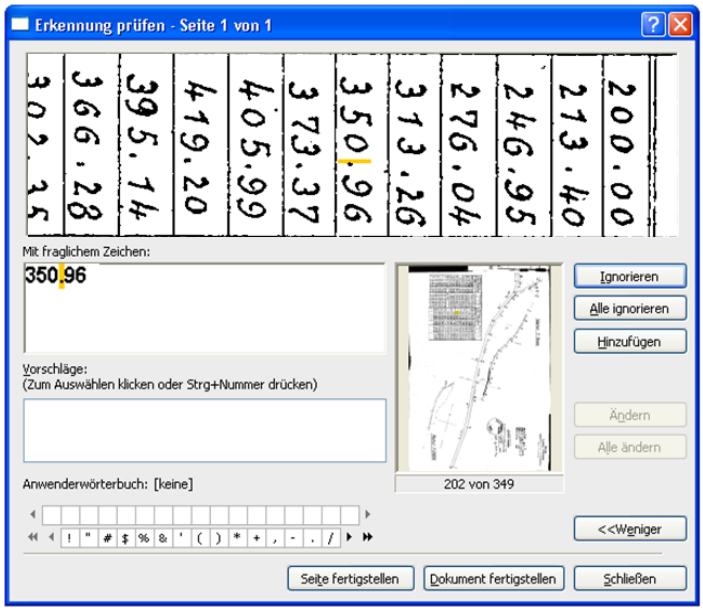

Abbildung 40: Erkennungsprüfung (Erkennung schräg geschriebener Koordinaten)

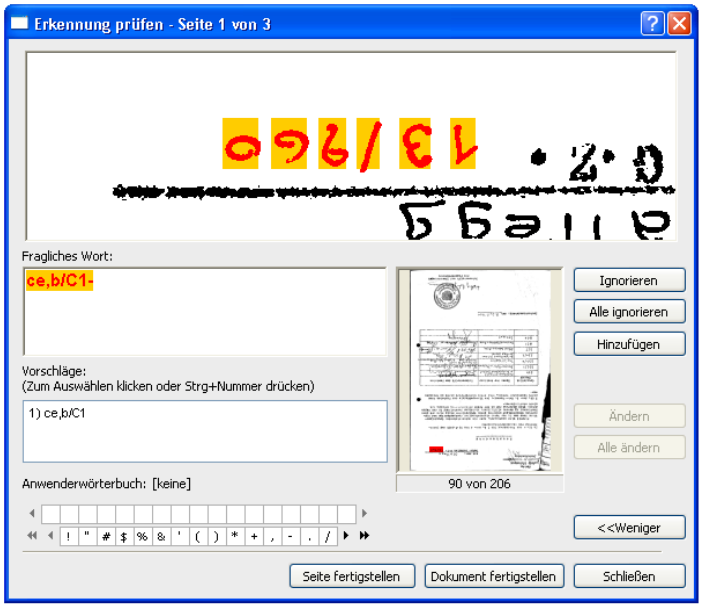

Abbildung 41: Erkennungsprüfung (Erkennung einer verkehrt geschriebenen Geschäftszahl)

Schritt 3.2.1.3: Kompression

- In diesem Schritt werden die beiden Ebenen „Textelemente“ und „grafische Elemente“ wieder vereinigt bzw. komprimiert.

Schritt d: Ausgabe verschiedene Dateiformate

- Je nach Wahl des Ausgabeformates (z.B. .txt, .rtf, .doc, .pdf), kann ein digitales Dokument erstellt werden, das im Anschluss daran weiterbearbeitet werden kann (z.B. Durchsuchen oder Extrahieren von Text, Umwandlung in ein PDF/A - Dokument).
- In diesem Fall wird ein durchsuchbares PDF - Dokument erstellt, welches anschließend in ein PDF/A - Dokument konvertiert wird Abbildung 42.

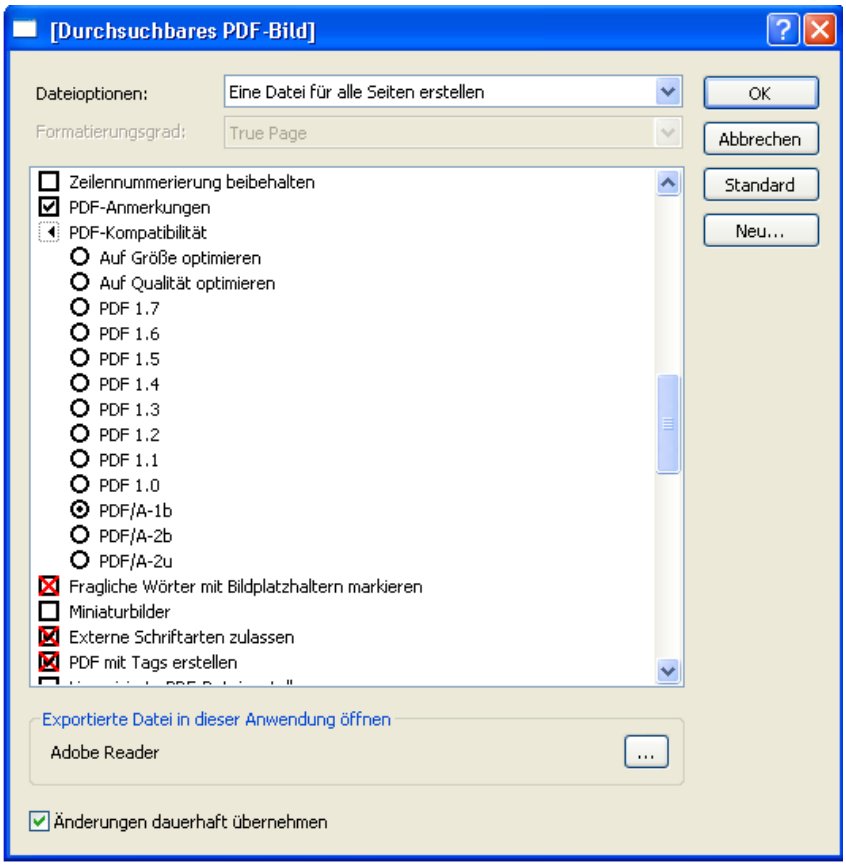

Abbildung 42: Wahl des Ausgabeformats

Schritt 3.2.1.4: Umwandlung in PDF/A

- Da es sich bei der Texterkennung mittels OCR um einen Eingriff in die Datei bzw. eine Konvertierung der Datei handelt, ist an dieser Stelle zu beachten, dass die Inhalte dieser verändert werden.
- Um daher ein akzeptables, für die Langzeitarchivierung geeignetes PDF/A - Format zu erhalten, ist jede aus dem OCR - Prozess erhaltene Datei, mit dem für die Langzeitarchivierung mittels PDF/A definierten Prozess, erneut zu analysieren.
- Dabei wird die aus dem OCR - Prozess erhaltene Datei in ein PDF/A - Dokument umgewandelt (Abbildung 42) und zunächst mit Hilfe der Software *callas pdfaPilot* auf das PDF/A - Format geprüft (Abbildung 43). Mit Hilfe der Software kann somit entschieden werden, ob die vorliegende Datei tatsächlich eine gültige PDF/A - Datei ist.
- Sollte sich herausstellen, dass die vorliegende Datei eine gültige PDF/A - Datei ist, wird diese anschließend mit der Eingangs - PDF/A - Datei verglichen (Abbildung 43).

Abbildung 43: Überprüfung auf PDF/A – Format
(Screenshot bearbeitet im März 2015)

Hier wird erkannt, dass die vorliegende Datei eine gültige PDF/A - Datei (PDF/A - 1b) ist, welche anschließend mit der Eingangs - PDF/A - Datei verglichen werden kann (Abbildung 44).

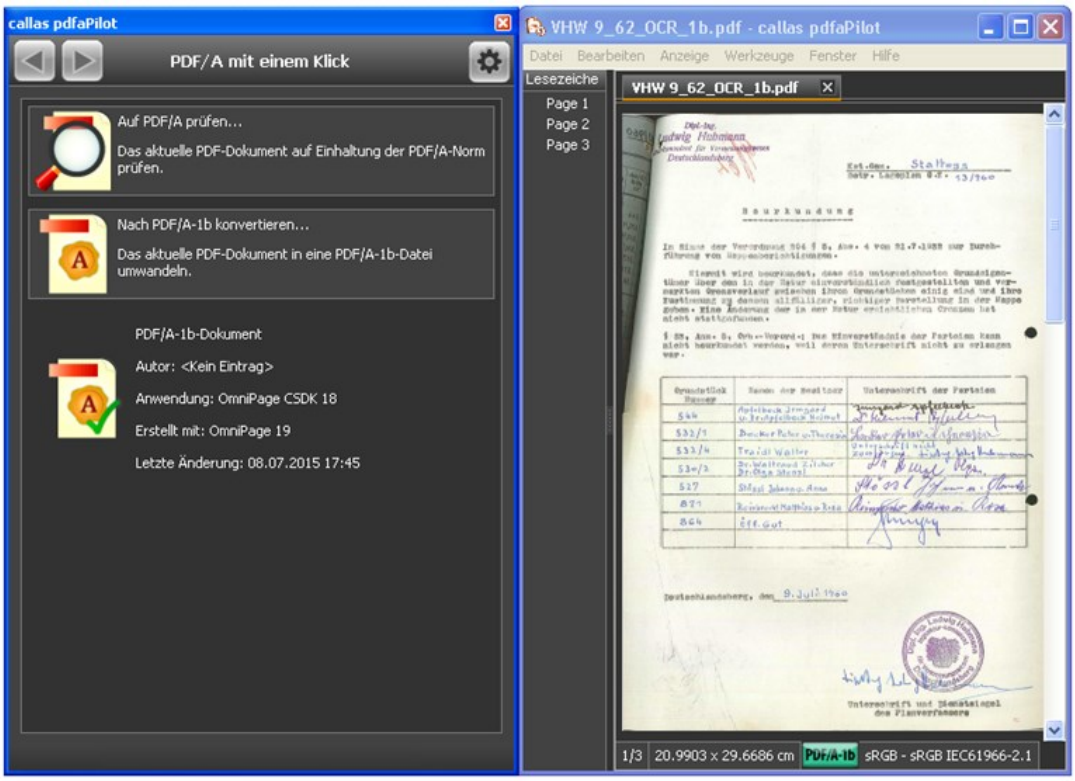

Abbildung 44: Überprüfung auf PDF/A - Version
(Screenshot bearbeitet im März 2015)

Check gegenüber Originaldatei (Qualitätsprüfung):

- Da es sich bei der Texterkennung mittels OCR um einen Eingriff in die Datei handelt, ist jede Datei als „neue“ Datei zu behandeln und mit der oben genannten Eingangs - PDF/A - Datei zu vergleichen (Qualitätsprüfung).
- Ein Vergleich der beiden Dateien kann mit Hilfe der Software *callas pdfaPilot* (Abbildung 45) automatisch ausgeführt werden. Für den Vergleich der beiden Dateien ist die Software an dieser Stelle geeignet, da hierbei ein Pixelvergleich (Text/Raster) stattfindet.
- Folgend sei ein Vergleich zwischen der mittels OCR bearbeiteten Datei und der oben genannten Eingangs - PDF/A - Datei gegeben (Abbildung 46).
- Zusätzlich sei beispielhaft ein vergrößerter Ausschnitt eines Vergleichs in Abbildung 47 gezeigt.

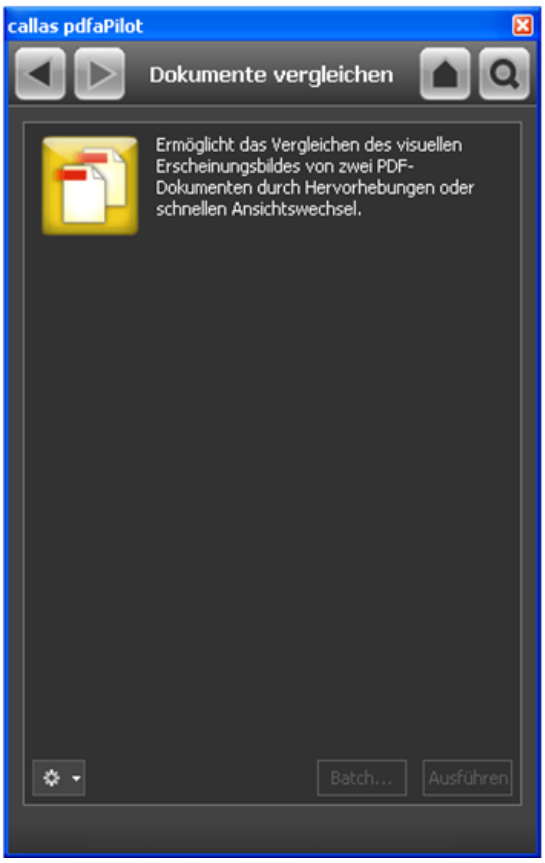

Abbildung 45: Check gegenüber PDF/A - Datei (Qualitätsprüfung, Screenshot bearbeitet im März 2015)

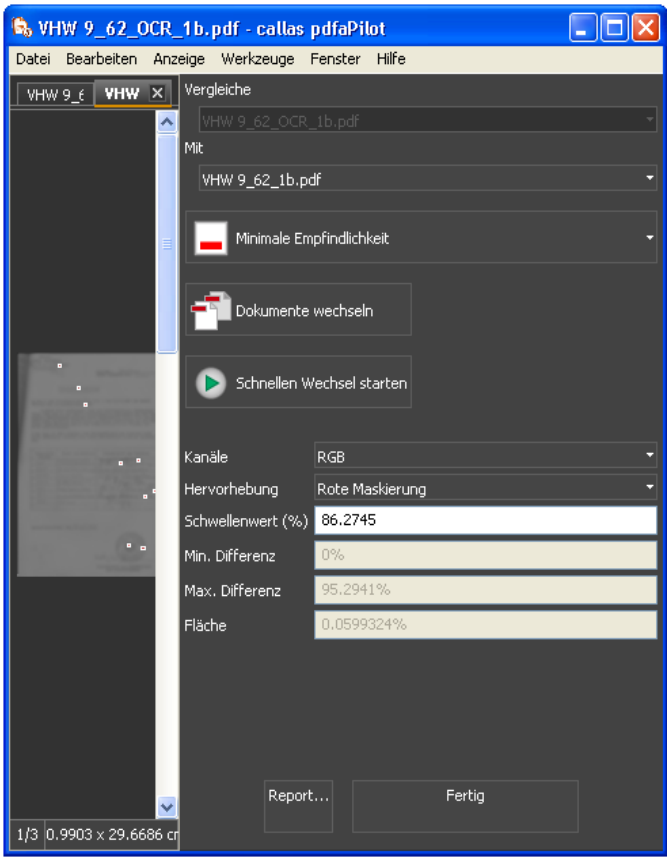

Abbildung 46: Vergleich (OCR - PDF/A - 1b mit Eingangs - PDF/A - 1b)

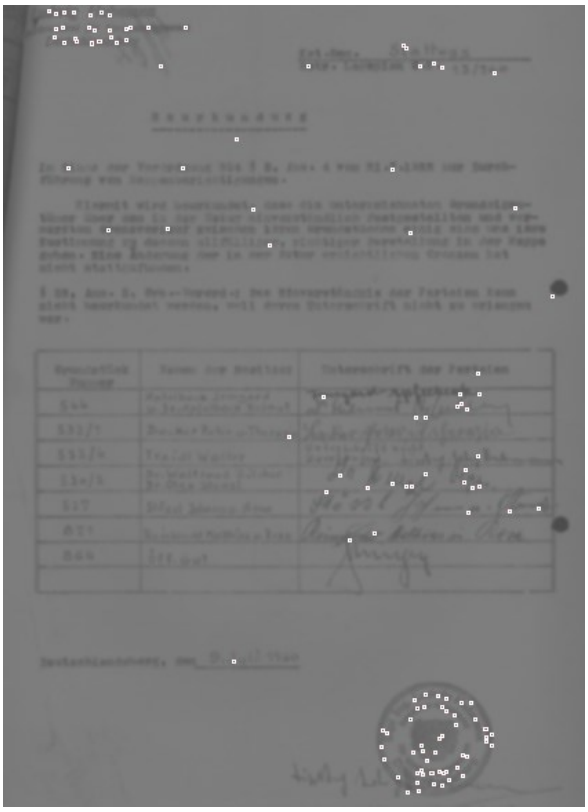

Abbildung 47: Vergrößerter Ausschnitt (Vergleich PDF/A - 3b mit Originaldatei)

Ergebnis:

- Nach Durchführung der OCR - Texterkennung wird ersichtlich, dass sich die aus dem OCR abgeleitete Datei mit der oben genannten Eingangs - PDF/A - Datei unterscheidet.
- Der Vorgang wäre an dieser Stelle abgeschlossen.

5 Ergebnisse

An dieser Stelle seien die wesentlichen Ergebnisse der Durchführung der beiden Prozesse aufgelistet.

5.1 Langzeitarchivierung mittels PDF/A

Hauptziel der Masterarbeit war die Findung des geeignetsten Langzeitarchivformats mit Fokus auf PDF/A. Hierfür wurden alle vorliegenden Dokumente verschiedenster Dateitypen analysiert und jeweils mittels des in Kapitel 3.1 gezeigten PDF/A - Prozesses verarbeitet. Das Augenmerk wurde dabei ausschließlich auf eigenständige PDF - bzw. PDF/A - Dateien mit Geo - Inhalt gelegt. Besonders zu berücksichtigen war, dass bei der Konvertierung in ein Langzeitarchivformat einerseits möglichst wenig Information verloren gehen sollte (z.B. Elimination verbotener Inhalte, wie etwa Transparenzen in PDF/A - 1b), andererseits sollte aber auch der Konvertierungsprozess möglichst automatisiert ablaufen. Hierbei mussten aber entweder gewisse Entscheidungen schon vorab definiert oder manuell durch den Benutzer gesetzt werden (z.B. hinsichtlich des Eingriffs in eine Datei, wenn weder für die eine noch die andere PDF/A - Variante eine eindeutige Tendenz vorherrscht und Änderungen für die Erreichung beider Varianten möglich wären).

Folgend seien die wichtigsten Ergebnisse der beiden in Kapitel 3.1 vorgestellten Varianten, sowie auch ein allgemeinerer Überblick über den PDF/A - Prozessablauf und über die Software gegeben.

5.1.1 Ergebnisse zu Variante 1 - „Match & Fix“

Variante 1 (*„Match & Fix“*) beschäftigte sich mit der Findung des geeignetsten Langzeitarchivformats im Hinblick auf PDF/A, unter Berücksichtigung aller erlaubten, vorgegebenen bzw. normierten Inhalte der verschiedenen PDF/A - Varianten.

Da sich ein erster Schritt mit der Auswahl des geeignetsten Formats ohne Dateieingriff beschäftigte (Schritt a), musste in diesem Fall eine Konvertierung mit Hilfe der Software *callas pdfaPilot* in alle möglichen Formate erfolgen. Da die Software jedoch das eigentliche „Matching“ (Analyse der Ausgangsdatei und Gegenüberstellung der verschiedenen Normanforderungen) nicht unterstützte, konnte hier nur versuchsweise ermittelt werden, welche PDF/A - Varianten möglich bzw. wo unter Umständen Fehler zu korrigieren gewesen wären (gegebenenfalls durch Autokorrektur).

Bei der Umsetzung des Prozesses ergab sich, dass in den meisten Fällen fünf mögliche Formate (PDF/A - 3b, - 3u, - 2b, - 2u, - 1b) zur Auswahl standen. Jede Konvertierung erforderte jedoch, dass die Ausgangsdatei erneut geladen werden musste, da sonst mit der bereits konvertierten Datei weiter gearbeitet worden wäre. Da im Rahmen der Durchführung jedoch einerseits der Prozessablauf möglichst automatisiert (mit so wenig wie möglich an Eingriffen in die Datei) stattfinden sollte, andererseits aber bei der Durchführung des Prozesses keine Programmierung erfolgte, ergab sich die Konvertierung in alle möglichen Formate als äußerst umfangreich und zeitaufwendig. In

diesem Fall wäre es wünschenswert, einen Vorschlag der Software für alle möglichen bzw. auswählbaren PDF/A - Formate zu bekommen, in die ohne (oder mit vorgeschlagenen) Autokorrekturen konvertiert werden könnte und bei der der Benutzer selbst eines der zur Verfügung stehenden Formate auswählen könnte.

In einem weiteren Schritt wurde die Auswahl des geeignetsten Formats mit Eingriff in die Datei behandelt (Schritt b), wobei hier zwischen „leichten“ und „schweren“ Fehlern zu unterscheiden war, welche entweder mittels der Software oder mittels Benutzer behoben werden konnten bzw. mussten, um die geeignetste, gewählte oder gewünschte PDF/A - Variante zu erhalten. Bei der Umsetzung des Prozesses ergab sich, dass in den meisten Fällen alle von der Software gefundenen Fehlerarten korrigiert werden konnten, jedoch ergaben sich in den überwiegenden Fällen bei der Konformitätsstufe a (Accessible oder Advanced) Probleme. Diese Stufe benötigte Strukturinformationen der Originaldatei und diesbezüglich auch des Originaldokument und nicht das vorliegende PDF - Dokument.

5.1.2 Ergebnisse zu Variante 2 - „Trial & Fix“

Variante 2 (*„Trial & Fix“*) beschäftigte sich mit dem sukzessiven Konvertierungsversuch der als PDF deklarierten Dateien, zur Findung des geeignetsten Langzeitarchivformats hinsichtlich PDF/A. In einem ersten Schritt wurde daher versucht, die vorliegenden PDF - Dateien zunächst in die höchste Variante (PDF/A - 3a) zu konvertieren, da diese die meisten Vorgaben, Funktionen und auch Features zur Verfügung stellte und damit potentiell auch am meisten Informationen in einer Datei beibehalten werden könnten. Stellte sich heraus, dass diese Variante nicht erreicht werden konnte, wurde in einem weiteren Schritt sukzessive in die nächst niedrigeren Varianten konvertiert (Schritt c), mit der Präferenz, vorerst noch keine Eingriffe vornehmen zu müssen.

Bei der Umsetzung des Prozesses ergab sich, dass in den meisten Fällen eine Konvertierung in die höchste Variante (PDF/A - 3a) nicht möglich war, da hierfür Strukturinformationen des Dokuments verlangt worden wären und diesbezüglich auch die Originaldatei und nicht das vorliegende PDF - Dokument benötigt worden wäre. Währenddessen konnte die PDF/A - 3b Variante in den überwiegenden Fällen erreicht werden, womit auch der Prozessdurchlauf an dieser Stelle abgeschlossen werden konnte. Damit ergab sich auch Variante 2 als weniger zeitaufwendig als Variante 1, da binnen kürzester Zeit ein Ergebnis erreicht werden konnte.

5.1.3 Erläuterungen zu den beiden Varianten (zum Prozessablauf)

Ein Vergleich der beiden Varianten liefert, dass sich die „Match & Fix“ - Variante (Variante 1) als komplexer und auch zeitintensiver herausstellte, da hier bereits in einem ersten Schritt eine Konvertierung in alle möglichen Formate erfolgen musste. Eine Konvertierung durch die Software *callas pdfaPilot* in gleichzeitig alle Formate, war hier jedoch nicht möglich, da diese, wie bereits in Abschnitt 5.1.1 erwähnt, das eigentliche „Matching“ nicht unterstützte. Aus diesem Grund konnte hier nur versuchsweise ermittelt werden, welche PDF/A - Varianten möglich bzw. wo unter Umständen Fehler zu

korrigieren gewesen wären. Gegebenenfalls wäre hier eine Erweiterung der Software durch entsprechende Programmierung möglich.

Die „Trial & Fix“ - Variante (Variante 2) hingegen, stellte sich im Vergleich zur „Match & Fix“ - Variante (Variante 1) als einfacher, schneller und effizienter heraus, da bei der sukzessiven Konvertierung binnen kürzester Zeit ein Ergebnis erreicht werden konnte, **weshalb diese Variante hier auch zu bevorzugen war**. Um dies besser zu veranschaulichen, sei in Abbildung 48 und Abbildung 49 der bereits in Kapitel 3.1 gezeigte PDF/A - Prozess in vereinfachter Weise dargestellt.

Sowohl bei Variante 1 als auch bei Variante 2, erfolgte ein Check gegenüber der Originaldatei (Qualitätsprüfung), bei der eine konvertierte Datei als „neue“ Datei zu behandeln und mit der Originaldatei zu vergleichen war, da jegliche Konvertierung in eine PDF/A - Variante einen Eingriff in die Datei darstellte. Der visuelle Vergleich zwischen den Dateien lieferte dabei keine Unterschiede. Diesbezüglich sei als Beispiel ein solcher Vergleich zwischen Originaldatei und konvertierter Datei anhand eines Bebauungsplans in Abbildung 50 gezeigt. Weitere Beispiele seien anhand eines Teilungsplans (Abbildung 51) und einer Landkartendarstellung (Abbildung 52) gezeigt.

Ein Vergleich (Pixelvergleich) der Dateien, der mit Hilfe der Software automatisch ausgeführt werden konnte, lieferte jedoch unzureichende Ergebnisse, da die Software (laut *Dietrich von Seggern* von der *callas software GmbH* (http://www.callas software.com/de) und gemäß der erzielten Ergebnisse, die sich auf der beiliegenden DVD befinden) für einen Vergleich von konvertierten Dateien nicht gedacht bzw. geeignet war. Die Empfindlichkeit beim Vergleich der Dateien war hier derart groß, dass bereits geringste Unterschiede (z.B. Farb - oder Transparenzunterschiede) von der Software ausgegeben wurden. An dieser Stelle wäre eine geeignetere Software für den Vergleich von Originaldatei und konvertierter Datei notwendig, wobei im Rahmen der Masterarbeit dazu keine Nachforschungen betrieben wurden. Es sei hier zudem darauf hingewiesen, dass es hinsichtlich der Präsentation erfolgreich konvertierter Geo - Dokumente problematisch ist, entsprechende Ergebnisse zu zeigen, da sich Originaldatei und konvertierte Datei im Regelfall nicht voneinander unterscheiden dürften. Im Idealfall sollte es daher für den Betrachter nicht erkennbar sein, dass eine Ausgangsdatei in ein Langzeitarchivformat konvertiert wurde, zumal die Konvertierung ohne Veränderungen die Voraussetzung bzw. das Fundament der PDF/A - Norm ist und somit ein „Erfolg“ erst dann vorliegt, wenn keine Unterschiede zur Ausgangsdatei nach der Konvertierung zu erkennen sind. Um aber dennoch einen Überblick über die unterschiedlichen Geo - Dokumente und die erfolgreich konvertierten Dateien zu zeigen, wurden zur besseren Ersichtlichmachung die Ergebnisdaten exemplarisch mit „PDF/A - 1b“ gekennzeichnet (Abbildung 50, Abbildung 51, Abbildung 52). Um die tatsächlichen Ergebnisse betrachten und verifizieren zu können, sei der Leser an dieser Stelle auf die beiliegende Daten - DVD verwiesen.

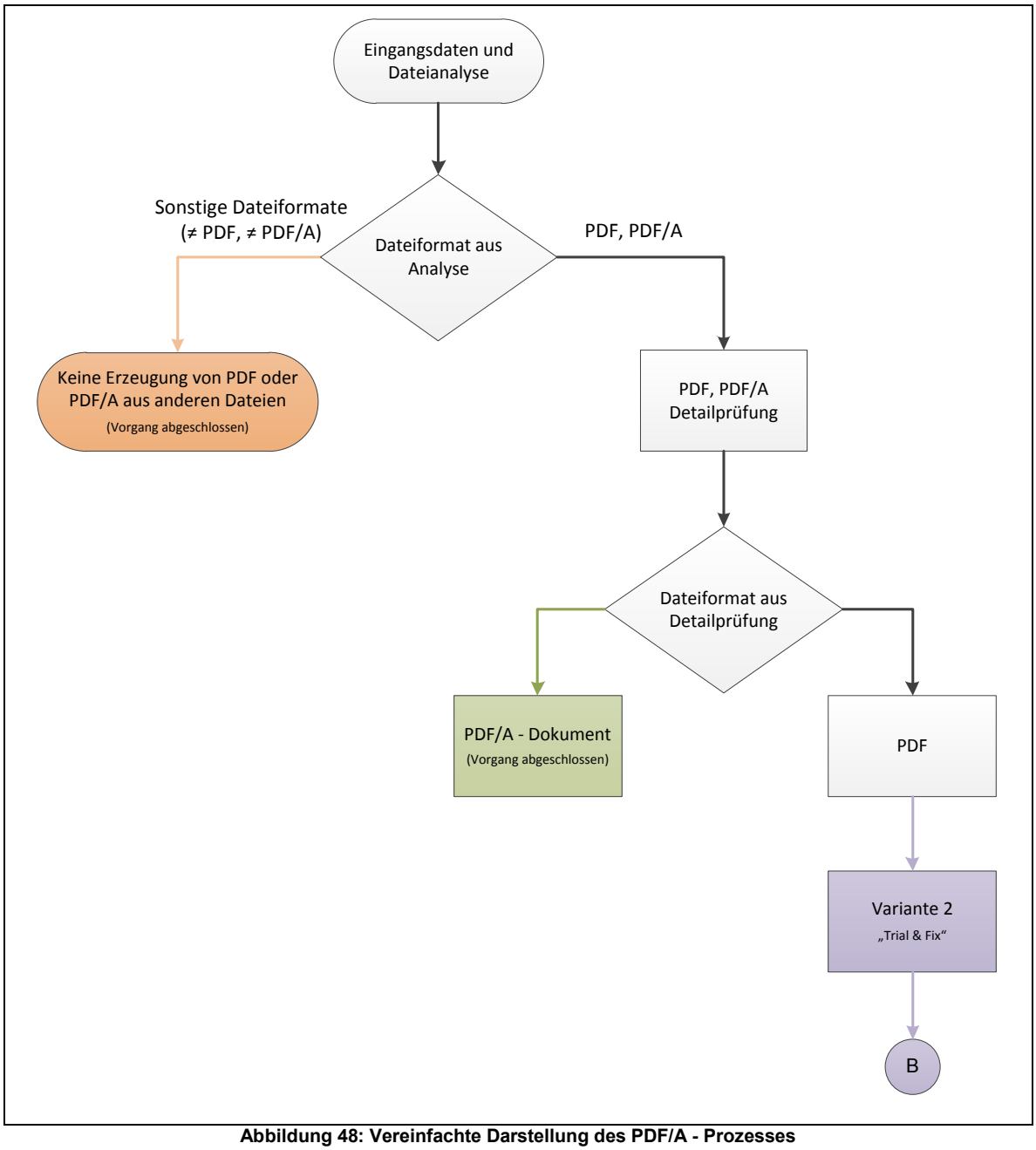

Abbildung 48: Vereinfachte Darstellung des PDF/A - Prozesses

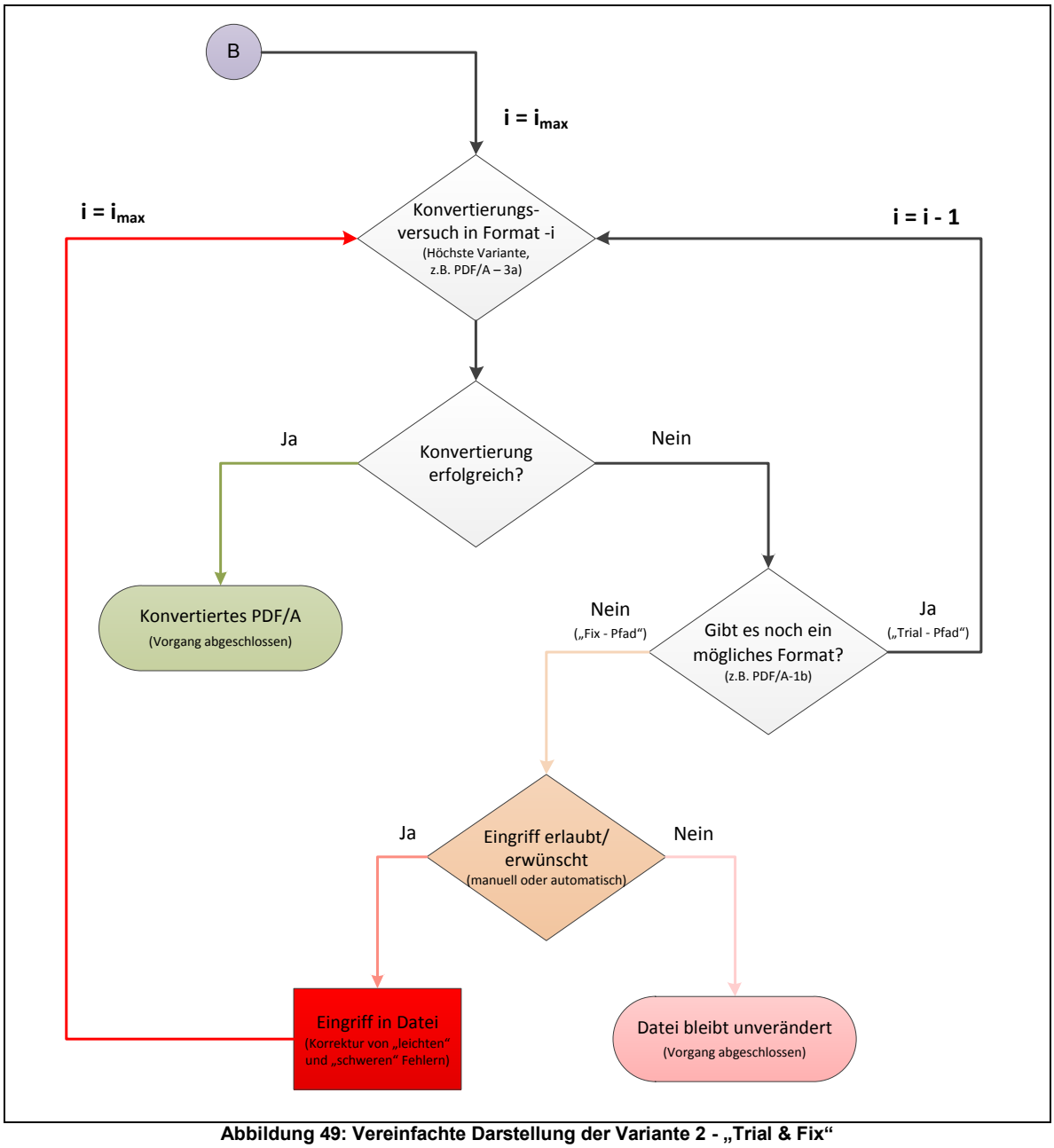

Abbildung 49: Vereinfachte Darstellung der Variante 2 - „Trial & Fix“

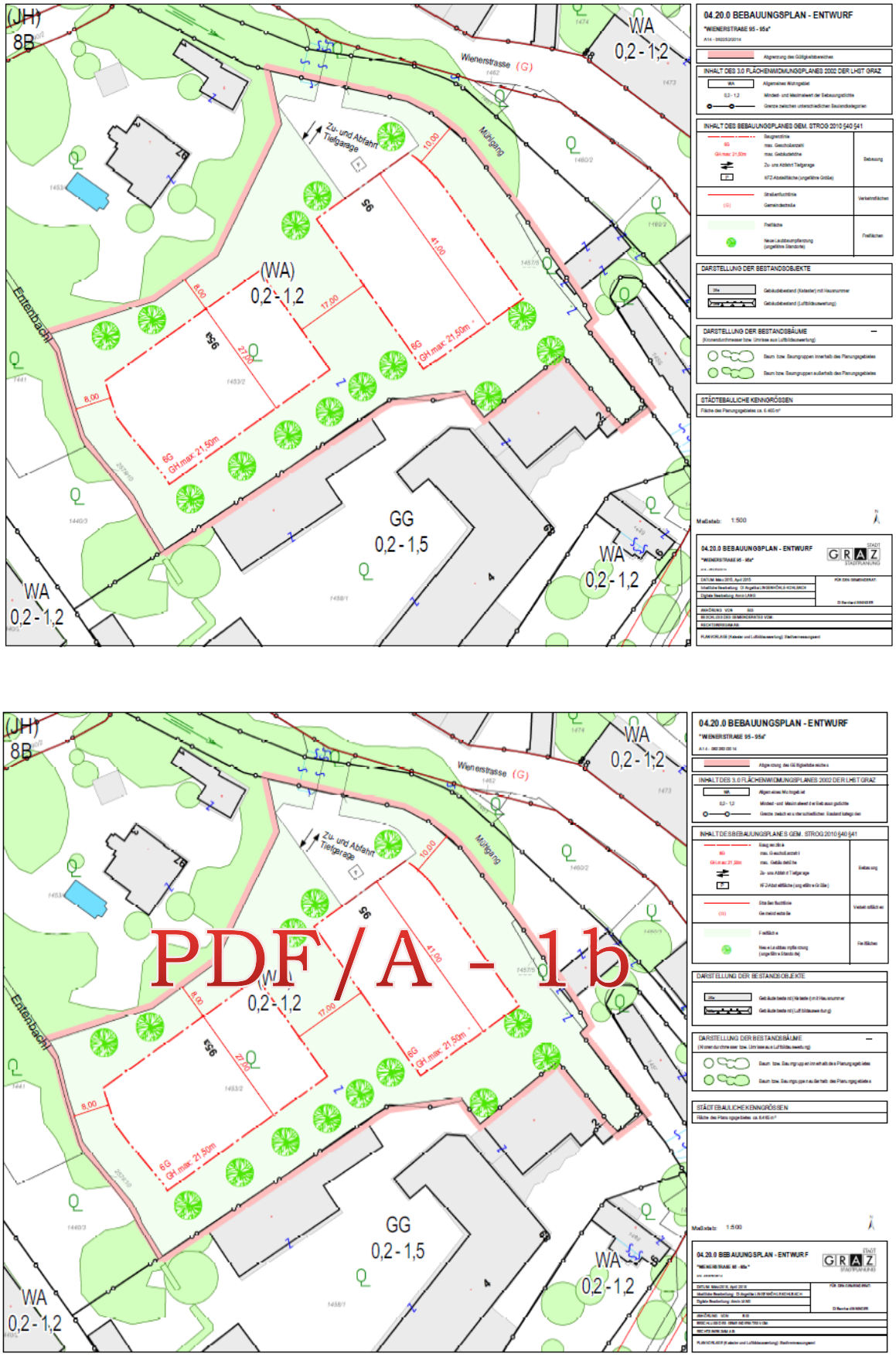

Abbildung 50: Gegenüberstellung der Originaldatei (oben) mit der aus dem PDF/A - Prozess erhaltenen PDF/A - Datei (unten) anhand eines Bebauungsplans

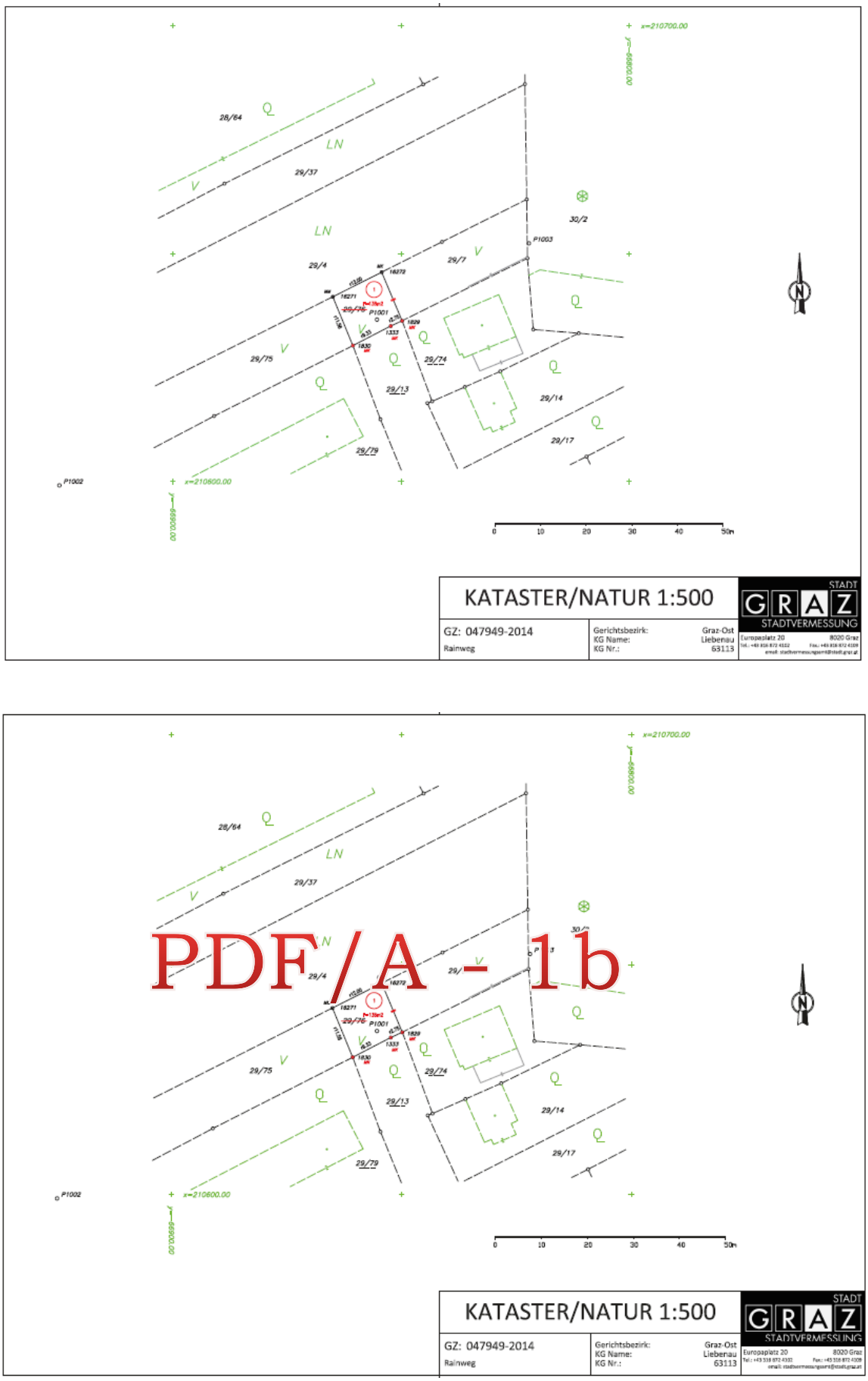

Abbildung 51: Gegenüberstellung der Originaldatei (oben) mit der aus dem PDF/A - Prozess erhaltenen PDF/A - Datei (unten) anhand eines Teilungsplans

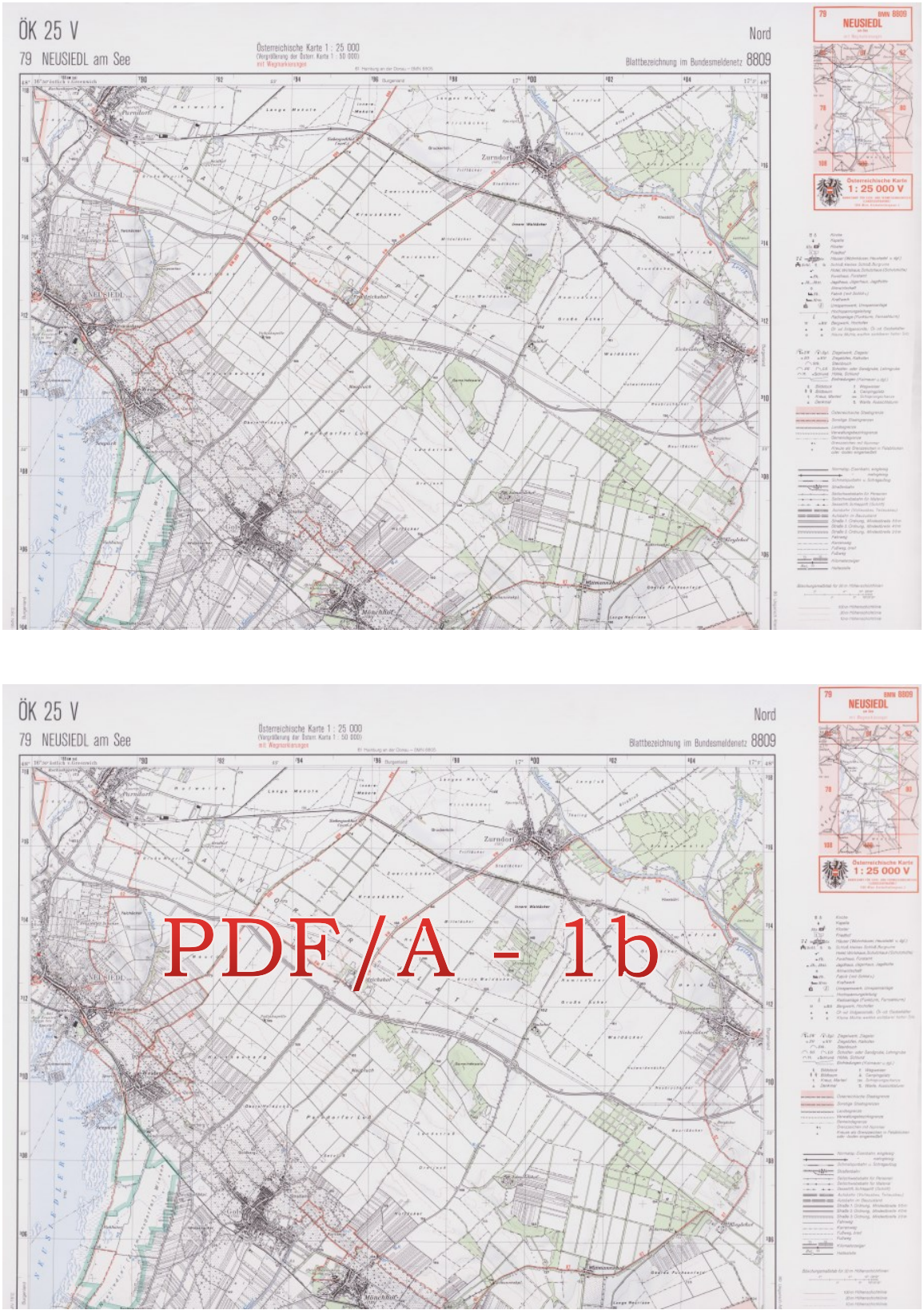

Abbildung 52: Gegenüberstellung der Originaldatei (oben) mit der aus dem PDF/A - Prozess erhaltenen PDF/A - Datei (unten) anhand einer Landkartendarstellung

Um einen Überblick über den PDF/A - Prozessablauf und auch über die verwendete Software *callas pdfaPilot* zu geben, seien folgend in Tabelle 4 die wesentlichen Erläuterungen, die sich im Zuge der Durchführung ergaben, aufgelistet.

Tabelle 4: Erläuterungen zum PDF/A - Prozess unter Verwendung der Software *callas pdfaPilot*

• Eine Konvertierung der Datei in die Basic - und - Unicode - Variante ist in den überwiegenden Fällen möglich	• Eine Konvertierung in die Accessible (oder Advanced) - Variante ist in den meisten Fällen nicht möglich, da hierfür Strukturinformationen des Originaldokuments benötigt werden
• Eine Konvertierung in eine gewählte PDF/A - Variante ist mit Hilfe der Software mit nur einem Klick möglich	• Für jede Konvertierung in eine gewählte PDF/A - Variante ist die Ausgangsdatei erneut zu laden, da sonst mit der zuvor konvertierten Datei gearbeitet wird
• In jedem Konvertierungsschritt kann die Datei gesondert abgespeichert werden; dementsprechend wird die Ausgangsdatei nicht überschrieben Ein dazugehöriger Report über eine erfolgreiche bzw. fehlgeschlagene Konvertierung ist entweder im Anzeigefenster der Software sichtbar oder kann mit Hilfe der Software erstellt und separat abgespeichert werden	• Die Größe der Datei bestimmt die Dauer einer Konvertierung in mögliche PDF/A - Formate (je größer die Datei, desto länger dauert eine Konvertierung) Die Größe der Datei hängt in diesem Fall mit ihrem Inhalt zusammen. Sollten beispielsweise Rasterinhalte vorhanden sein, ist die Datei größer und damit wird auch die Konvertierung der Datei durch die Software verlangsamt
• Die sukzessive Konvertierung (Variante 2) ergibt sich als weniger zeitaufwendig, da binnen kürzester Zeit ein Ergebnis erreicht werden kann	• Die Konvertierung in alle möglichen Formate (Variante 1) ist zeitaufwendig, weswegen ein Vorschlag der Software für alle möglichen Formate wünschenswert wäre
• Die Software kann Vorschläge zur Korrektur hinsichtlich gefundener Fehler machen	

5.1.4 Fazit bzw. wichtigstes Ergebnis zum PDF/A - Prozess

Anhand der Testfiles konnte, im Zusammenspiel mit der verwendeten Software, ein Prozess entwickelt werden, der die Findung des am besten geeignetsten PDF/A - Formats realisiert, unter der Berücksichtigung, dass so wenig wie möglich an (ursprünglicher) Information des jeweiligen Geo - Dokuments verloren gehen darf bzw. auch die Originalität des Geo - Dokuments so gut wie möglich beibehalten werden kann. Diese Prozessvariante wurde als **„Trial & Fix“ - Variante** deklariert.

5.2 Texterkennung mittels OCR

Die Texterkennung mittels OCR beschäftigt sich damit, nicht unmittelbar zugängliche Informationen aus den vorliegenden Langzeitarchiv - Dokumenten abzuleiten. Ziel ist es daher, aus den vorhandenen Dateien beispielsweise einen Raster - Text aus einer Grafik bzw. Bilddatei zu extrahieren, wobei das Langzeitarchiv - Dokument dabei nicht zerstört werden soll. Die aus dem in Kapitel 3.2 gezeigten OCR - Prozess abgeleiteten Dateien sind im Anschluss daran wieder in ein geeignetes PDF/A - Format umzuwandeln und mit den PDF/A - Eingangsdaten zu vergleichen bzw. ebenfalls als Archivdokumente abzulegen. Diesbezüglich wurde im Rahmen der Masterarbeit die Texterkennung mittels OCR bzw. der OCR - Prozess nur auf PDF/A - Dokumente angewandt, wobei es sich bei diesen um PDF/A - 1b Dokumente handelte. Eine Untersuchung des OCR - Prozesses mit PDF - Dokumenten wäre hier ebenfalls denkbar gewesen, jedoch waren in diesem Falle PDF/A - Dokumente (PDF/A - 1b) ausreichend, da bereits aus diesen Zusatz - Dokumente mit Mehrwert erstellt werden konnten. Auf die Untersuchungen mittels PDF wurde daher nicht näher eingegangen.

Folgend seien die wichtigsten Ergebnisse des in Kapitel 3.2 gezeigten OCR - Prozesses gegeben.

5.2.1 Ergebnisse zur Texterkennung mittels OCR

Bei der Texterkennung mittels OCR wurden ausschließlich die aus dem PDF/A - Prozess hervorgehenden PDF/A - 1b Dokumente betrachtet, aus denen es galt, mit Hilfe der Software *Nuance OmniPage Ultimate*, nicht unmittelbar zugängliche Informationen (z.B. Raster - Text) abzuleiten bzw. diese Dokumente durchsuchbar zu machen. Eine vereinfachte Darstellung des bereits in Kapitel 3.2 gezeigten OCR - Prozesses, sei zu diesem Zwecke in Abbildung 53 dargestellt.

Bei der Umsetzung des OCR - Prozesses ergab sich, dass gescannte Dokumente von der Software durchsuchbar gemacht werden konnten. Sowohl ursprünglich elektronisch verfasste, als auch händisch verfasste Dokumente (Schriften, Pläne, Skizzen) die im Rasterformat vorlagen, wurden von der Software erkannt. Diesbezüglich konnte auch die Software Korrekturvorschläge zur manuellen Zeichenbearbeitung anbieten, falls Zeichen bzw. Wörter nicht im Wörterbuch der Software integriert waren (Abbildung 38 bzw. Abbildung 39). Weitere Vorteile ergaben sich außerdem darin, dass das Aussehen eines Dokuments bei der OCR - Durchführung nicht zerstört wurde und somit das Erscheinungsbild des Originaleingangsdokuments beibehalten werden konnte (Abbildung 54), obwohl hierbei nun auch nicht unmittelbar zugängliche Informationen zugänglich gemacht wurden (Abbildung 55). Problematisch erwies sich jedoch, dass je nach Seitenlayout des Eingangsdokuments, einzelne Zeichen bzw. Wörter von der Software erkannt, andere wiederum nicht erkannt werden konnten. So wurden Zeichen bzw. Wörter, die beispielsweise schräg (Abbildung 40) oder verkehrt geschrieben (Abbildung 41) in einem Dokument vorhanden waren, falsch interpretiert oder erst gar nicht als Zeichenfolgen bzw. Wörter erkannt. Eine weitere Schwierigkeit lag darin, dass nicht alle vorhandenen Dokumente, aufgrund einer vorgegebenen Mindestbildgröße

einer Bilddatei von 16 Pixel (Nuance Communications, 2013), von der Software eingelesen und somit bearbeitet bzw. im folgenden Prozessschritt (Qualitätsprüfung) gegenüber der Eingangsdatei (PDF/A - Ergebnisdatei aus dem PDF/A - Prozess) werden konnten. Da die Texterkennung mittels OCR jedoch nicht Kernthema der Masterarbeit ist, wurde an dieser Stelle nicht näher auf dieses Problem eingegangen.

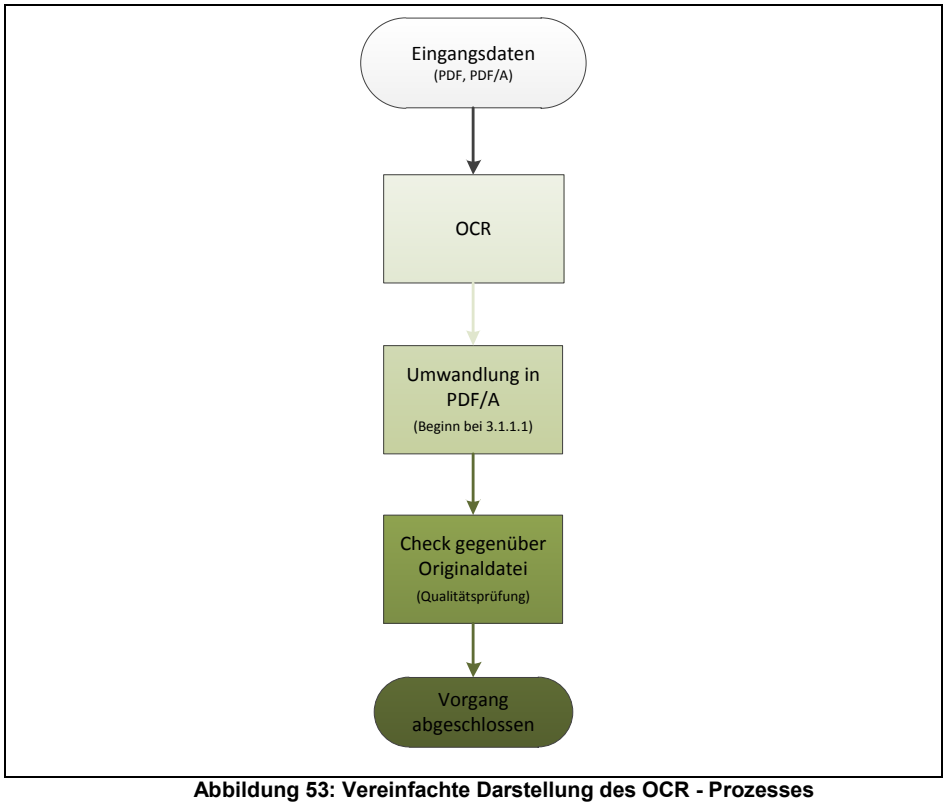

Abbildung 53: Vereinfachte Darstellung des OCR - Prozesses

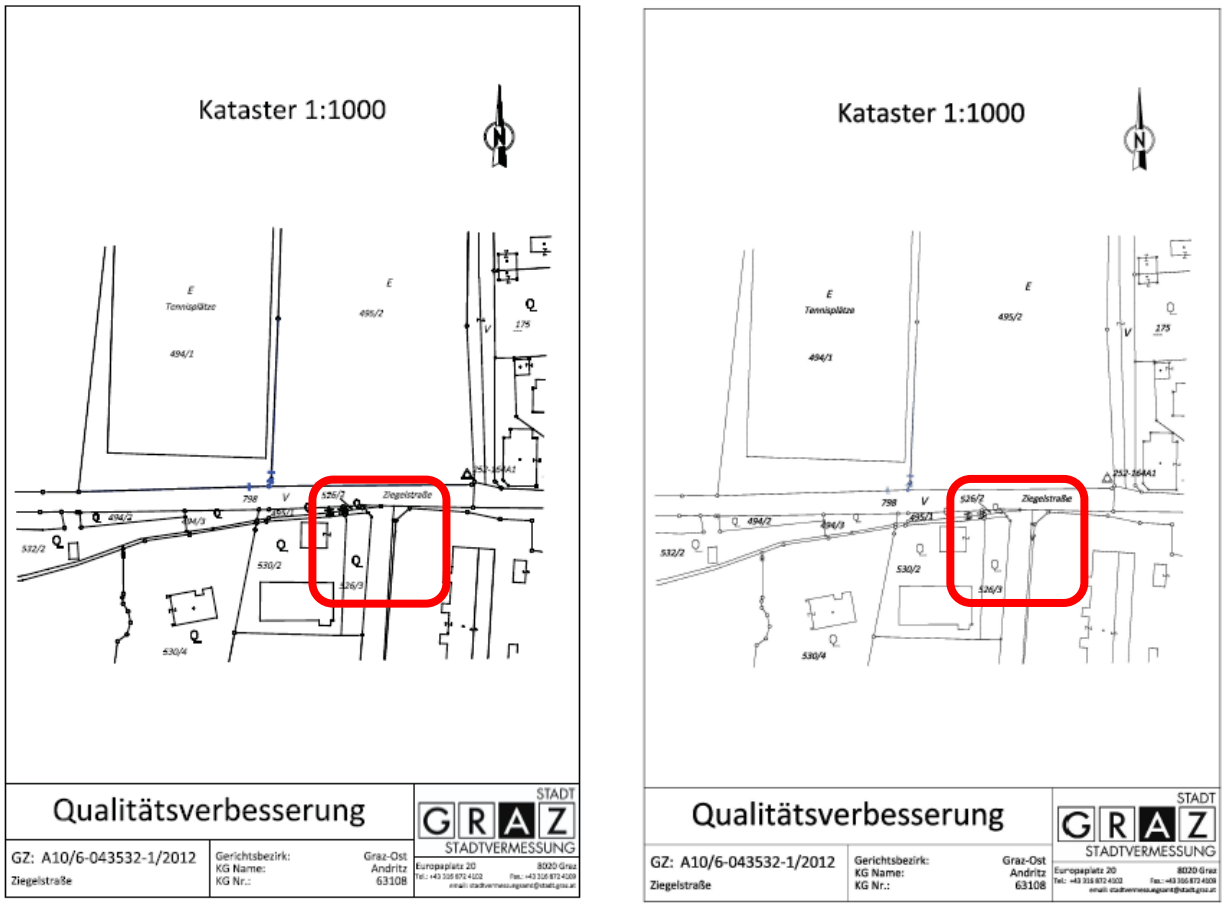

Abbildung 54: Gegenüberstellung eines Original - Raster - PDF/A - Eingangsdokuments (links) mit dem aus dem OCR - Prozess erhaltenen Text - PDF/A - Dokuments (rechts)

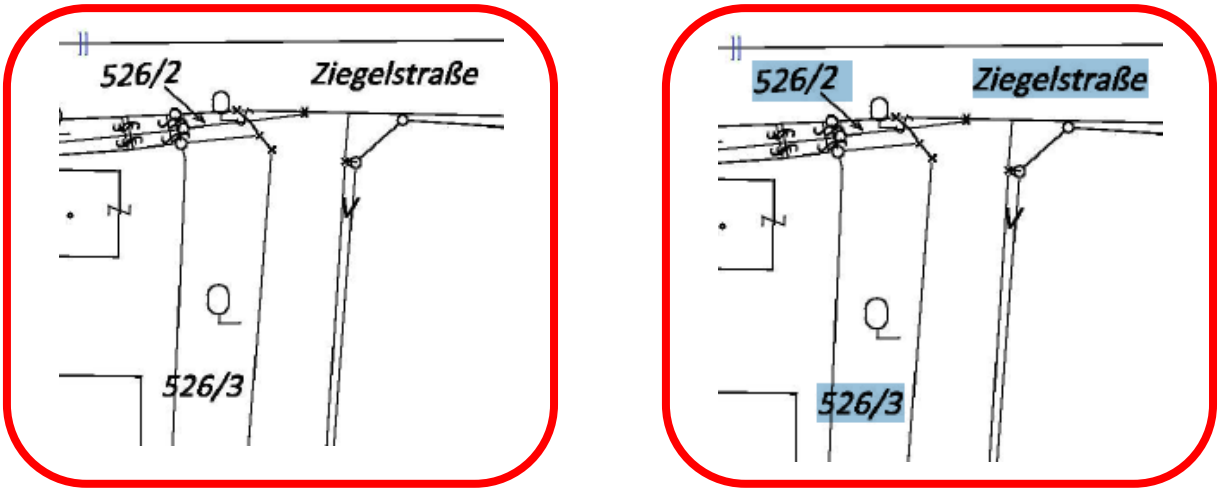

Abbildung 55: Vergrößerter Ausschnitt des Original - Raster - PDF/A - Eingangsdokuments (links) und dem aus dem OCR - Prozess erhaltenen Text - PDF/A - Dokuments (rechts)

Um einen Überblick über den OCR - Prozessablauf und auch über die verwendete Software *Nuance OmniPage Ultimate* zu geben, seien in Tabelle 5 die wesentlichen Erläuterungen, die sich im Zuge der Durchführung ergaben, aufgelistet.

Tabelle 5: Erläuterungen zum OCR - Prozess unter Verwendung der Software *Nuance OmniPage Ultimate*

• Die Durchführung des OCR - Prozesses erfolgt schnell und automatisch durch die Software	• Die Größe der Ergebnisdatei ändert sich je nach Inhalt (wird in den meisten Fällen größer als die Eingangsdatei)
• Die Software bietet Korrekturvorschläge zur manuellen Zeichenbearbeitung an, falls verwendete Zeichen bzw. Wörter nicht im Wörterbuch der Software integriert sind	• Je nach Seitenlayout des Eingangsdokuments, werden einzelne Zeichen bzw. Wörter richtig erkannt, andere wiederum nicht
• Ein dazugehöriger Report über eine erfolgreiche OCR - Prüfung ist im Anzeigefenster der Software sichtbar	
• Das Seitenlayout und auch das Aussehen des Originaldokuments kann beibehalten werden	

5.2.2 Fazit bzw. wichtigstes Ergebnis zum OCR - Prozess

Anhand der Testfiles konnte, im Zusammenspiel mit der verwendeten Software, ein Prozess entwickelt werden, mit welchem nicht unmittelbar zugängliche bzw. suchbare (Raster -) Informationen, in den aus dem PDF/A - Prozess abgeleiteten Langzeitarchivdokumenten, zugänglich gemacht und somit aus diesen Dokumenten, schrifterkannte OCR - PDF/A - Dokumente (sogenannte Zusatz - Dokumente mit Mehrwert), erzeugt werden konnten. Der Mehrwert dieser Dokumente liegt dabei in der Zugänglichkeit (Durchsuchbarkeit) hinsichtlich ursprünglichem Rastertext, in nunmehr verfügbaren Informationen mit Augenmerk auf Geo - Informationen (z.B. Ried - (Gebietsbezeichnungen), Grundstücksnummern, Koordinaten). Eine Untersuchung der Fehlerwahrscheinlichkeit bzw. Robustheit hinsichtlich Zeichenerkennung wurde im Rahmen der Masterarbeit nicht durchgeführt.

5.3 Resümee

Im Rahmen der Masterarbeit wurden mit Hilfe der in Kapitel 3.1 und 3.2 vorgestellten Prozessabläufe PDF - bzw. PDF/A - Dateien mit Geo - Inhalt untersucht. Unter den vorliegenden Dateien befanden sich hierbei sowohl neuere, elektronisch verfasste, als auch ältere, teilweise händisch verfasste und eingescannte Dokumente.

Im Zuge der PDF/A - Prozessdurchführung ergab sich, dass mit Hilfe der gewählten Software *callas pdfaPilot* und den beiden vorgestellten Varianten alle vorliegenden Dateien in PDF/A - Dateien konvertiert werden konnten. Es stellte sich jedoch heraus, dass in vielen

Fällen Geo - Informationen nur als Rasterbild zur Verfügung standen (z.B. Naturdarstellung in einem Teilungsplan). Um diese Informationen bzw. Geo - Informationen zugänglich zu machen, wurde im Rahmen der Masterarbeit zusätzlich eine Untersuchung dieser mittels eines OCR - Prozesses durchgeführt. Hierbei ergab sich, dass mit Hilfe der gewählten Software *Nuance OmniPage Ultimate*, aus nahezu allen vorliegenden Dokumenten, Zusatz - Dokumente mit Mehrwert erstellt werden konnten. Vor allem bei neueren, elektronisch verfassten Dokumenten erwies sich die Texterkennung mittels OCR als robuster (hinsichtlich Zeichenerkennungsfehlern), im Vergleich zu älteren, teilweise händisch verfassten und eingescannten Dokumenten. Die Problematik der in Abschnitt 5.2.1 erwähnten Schwierigkeiten von schräg oder verkehrt vorliegender (oder sogar mangelhafter) Schrift konnten mit Hilfe von OCR zumindest teilweise gelöst werden. Genauigkeitseinbußen, sowie auch Fehler durch falsche Zeicheninterpretationen bei der Texterkennung sind jedoch auch hier nicht vorab zu vermeiden. Dennoch könnte das Zusammenspiel von PDF/A und OCR auch in Zukunft die meisten der vorhandenen Probleme beheben, um somit optimal für die Langzeitarchivierung geeignete PDF/A - Dokumente zu erhalten.

6 Einsatzgebiete und Ausblick

In diesem Kapitel wird die Anwendung von PDF/A in verschiedenen Bereichen und die Verwendung auf nationaler und internationaler Ebene beschrieben. Zusätzlich wird in einem weiteren Abschnitt auf die Zukunft und weitere Entwicklungen des PDF/A - Standards eingegangen.

6.1 PDF/A - Einsatz rund um die Welt

Viele Gründe sprechen für den Einsatz des PDF/A - Formats. Einer davon sei die Möglichkeit einer langfristigen Archivierung von Dokumenten in diesem Format. Ein anderer sei die Findung eines einheitlichen Formats, um mit Dokumenten rund um die Welt arbeiten zu können. Darauf aufbauend bezieht sich ein weiterer Grund auf die Plattformunabhängigkeit, was damit das Format kompatibel mit unterschiedlicher Hard - und Software macht und somit ein konfliktfreier Austausch und Anzeige von Dokumenten ermöglicht werden kann. Das Format PDF/A für die Langzeitarchivierung findet bereits in vielen Bereichen und Einsatzgebieten Anwendung. Denn, PDF/A wird nicht nur im privaten, sondern auch im öffentlichen Bereich genutzt. So findet der Standard beispielsweise Anwendung im Gesundheitswesen, Rechnungswesen, Gerichtswesen, aber auch unter den Behörden, in der Regierung und Verwaltung und in Unternehmen findet die Langzeitarchivierung mittels PDF/A Anerkennung. Sei es um Dokumente für längere Zeit aufzubewahren oder nur um einen Austausch in einem einheitlichen Format zu ermöglichen (Oettler, 2013).

National und auch international setzt sich der PDF/A - Standard immer mehr durch und es wird kontinuierlich an neuen Projekten in diesem Bereich gearbeitet. In Anlehnung an Drümmer et al. (2007, S. 11f), Oettler (2013, S. 13ff) und Primas (2013/14), seien hier einige ausgewählte, aus der Praxis bekannte Bereiche und Anwendungen, in denen die Langzeitarchivierung mittels PDF/A Verwendung findet, zusammengefasst.

- PDF/A wird in der **öffentlichen Verwaltung** verwendet. Das *Amt für Veröffentlichungen der Europäischen Union (Publications Office of the European Union)* hat alle Veröffentlichungen (wie z.B. Gesetze oder Bekanntmachungen) im PDF/A - Format zu publizieren. Dies ist vor allem wegen der vereinfachten Suche oder dem Eintragen von Informationen mittels (XMP -) Metadaten vorteilhaft.
- Auch im *Europäisches Patentamt* werden Patente im PDF - sowie auch im PDF/A - Format veröffentlicht; wiederum wegen der vereinfachten Suche und dem Eintragen von (XMP -) Metadateninformationen.
- Länderweit (z.B. *Niederlanden, Brasilien, Dänemark, Frankreich, Schweiz, Deutschland*) wird in der öffentlichen Verwaltung, wie z.B. in Regierungen, Behörden oder Standesämtern, der PDF/A - Standard eingesetzt. Hier steht vor allem die Garantie der Erhaltung des visuellen Erscheinungsbildes eines Dokuments im Vordergrund.
- Auch *Österreich* ist in der öffentlichen Verwaltung vertreten. So verwendet beispielsweise das Österreichische Staatsarchiv (ÖStA) das auf dem Referenzmodell OAIS

aufbauende Langzeitarchivformat PDF/A für die langfristige Aufbewahrung von Dokumenten. Auch das BEV (Bundesamt für Eich - und Vermessungswesen) betreibt sein Geschäftsregister mit dem Digitalen Kartenarchiv, einerseits als revisionssicheres Archiv, mit der gesetzlichen Vorgabe, Teilungspläne und zugehörige Beilagen als PDF/A - 1b vorzuhalten und andererseits von den zur elektronischen Einbringung verpflichteten Berufsgruppen dieses Format eingebracht zu erhalten.

- Ferner wird der PDF/A - Standard in **Wirtschaft und Industrie** (z.B. um Dokumente für Kernkraftwerke langfristig aufzubewahren), im **Gesundheitswesen** (z.B. um Befunde oder Laborberichte für 30 Jahre und länger aufzubewahren) oder auch im **Bankwesen oder bei Versicherungen** (z.B. um Kreditunterlagen für 50 Jahre oder länger zu archivieren) eingesetzt. Auch das Archiv der Architekten und Ingenieurkonsulenten verwendet PDF/A - 1b zur Langzeitarchivierung derer Urkunden.
- Zudem wird PDF/A auch in **Gesetzgebung und Justiz** eingesetzt, um beispielsweise Meldungen oder Urkunden auf elektronischem Wege zu archivieren.
- Des Weiteren gibt es zahlreiche andere Anwendungen von PDF/A **in verschiedenen Bereichen**. So können beispielsweise auch E-Mails, Broschüren, Handbücher, Informationsblätter, Konstruktionszeichnungen, Pläne, Verträge und vieles mehr bereits im PDF/A - Format abgespeichert werden.
- Damit ist der PDF/A - Standard auch im **privaten Bereich** stark vertreten. Diverse Kunden und Unternehmen nutzen zudem auch die in dieser Masterarbeit verwendete Software *callas pdfaPilot*. So ist eines der Unternehmen die *Mercedes - Benz Schweiz AG*, welche die Software für die elektronische Archivierung von Dokumenten (z.B. Formulare, Rechnungen oder Bauzeichnungen) verwendet. Ein weiterer Bereich in dem die Software verwendet wird, ist die italienische Handelskammer *InfoCamere*, welche vor allem verschiedene Unternehmen dabei unterstützt, das PDF/A - Format für die langfristige Aufbewahrung von Dokumenten einzusetzen callas software GmbH (2015g).
- Darüber hinaus gibt es auch zahlreiche andere Hersteller von Softwareprodukten, die sich mit der Archivierung mittels PDF/A beschäftigen. Als einer davon sei *intarsys consulting GmbH* genannt, welche mit ihrem Produkt die Herstellung von elektronischen Dokumenten im PDF/A - Standard ermöglichen. *intarsys consulting GmbH* besitzt überdies auch viele verschiedene Kunden, die ihre Software im privatwirtschaftlichen Bereich für die Langzeitarchivierung verwenden. Unter anderem seien hier z.B. Deutsche Post DHL Group, Lufthansa Technik oder Ricola AG genannt. Als weitere Hersteller von Softwareprodukten seien *PDF Tools AG* oder *LuraTech Solutions GmbH* genannt, welche ebenfalls Kunden in vielen Unternehmen und Organisationen bzw. Ländern besitzen, die ihr Produkt sowohl im privaten als auch im öffentlichen Bereich einsetzen (intarsys consulting GmbH (2015b); PDF Tools AG (2015b); PDF Association (2015e)).

6.2 Zukunft von PDF/A

In vielen Bereichen findet die Archivierung mittels PDF/A bereits Anwendung (siehe Abschnitt 6.1). Dennoch stellt sich die Frage, ob Dokumente derart archiviert werden können, sodass sie tatsächlich in 10, 50, 100 Jahren oder sogar dauerhaft zugänglich sein werden.

Jeder, der Dokumente dauerhaft archivieren will, könnte in PDF/A Vorteile finden. Nicht nur, weil PDF/A ein in der ISO Norm definierter Standard ist, der gegebenenfalls vom ISO Komitee stets weiterentwickelt wird, sondern auch, weil PDF/A sein visuelles Erscheinungsbild garantiert und damit auch, dass Inhalte in einem Dokument und auch seine Darstellung auf jeglicher Art von Anzeigeplattform nicht verändert werden, sondern jederzeit gleich aussehen, wie zum Zeitpunkt der Erstellung des Dokuments. Auch die Wahl einer der bereits existierenden PDF/A - Varianten, die sich je nach Aufgabenstellung und Anforderungen an den Workflow unterscheiden, ist möglich, was damit auch die Flexibilität dieses Archivformats zeigt bzw. auch dazu beiträgt, die richtige Variante für die jeweilige Aufgabenstellung zu finden.

Der derzeitige Stand der Technik gibt drei Versionen dieses Standards vor. Da PDF/A - 3 bereits zu einer der neuesten Version des Standards gehört (erschienen im Jahre 2012), steht laut *Thomas Zellmann* von der *PDF Association* (http://www.pdfa.org/) und *Dietrich von Seggern* von der *callas software GmbH* (http://www.callassoftware.com/de) zurzeit nicht die Entwicklung von weiteren PDF/A - Versionen im Vordergrund. Dennoch arbeitet ISO derzeit daran, die bereits normierte PDF - Version 1.7 (ISO 32000 - 1) auf die PDF - Version 2.0 (ISO 32000 - 2) zu erweitern, sodass, falls es in der Zukunft weitere PDF/A - Versionen geben sollte (beispielsweise PDF/A - 4), diese auf PDF 2.0 basieren dürften (Stand: April 2015).

Abkürzungsverzeichnis

BMP	Bitmap Image File
CEN	European Committee for Standardization
DIN	Deutsches Institut für Normung
GBO	Grundbuchordnung (Deutschland)
GeoTIFF	Geospatial Tagged Image File Format
GIF	Graphics Interchange Format
ISO	International Organization for Standardization
JBIG2	Joint Bi - level Image Experts Group
JPEG	Joint Photographic Experts Group
LZW	Lempel - Ziv – Welch
OAIS	Open Archival Information System
OCR	Optical Character Recognition
ÖStA	Österreichisches Staatsarchiv
PNG	Portable Network Graphics
PDF	Portable Document Format
PDF/A	Portable Document Format/Archival
PDF/E	Portable Document Format/Engineering
PDF/H	Portable Document Format/Healthcare
PDF/UA	Portable Document Format/Universal Accessibility
PDF/VT	Portable Document Format/ Variable Data and Transactional Printing
PDF/X	Portable Document Format/Exchange
RIS	Rechtsinformationssystem des Bundes
TIFF	Tagged Image File Format
UAR	Urkundenarchivrichtlinie
VHW	Veränderungshinweis
XMP	Extensible Metadata Platform
ZTG	Ziviltechnikergesetz

Abbildungsverzeichnis

Tabellenverzeichnis

Literaturverzeichnis

Adobe Systems Incorporated. (2015). *Was ist Adobe PDF?* https://acrobat.adobe.com/at/de/products/about-adobe-pdf.html. Zuletzt besucht am 16.04.2015

Arbeitsgruppe der ARK AG ESys und des ARK IT-Ausschusses. (2009). *Handreichung zur Archivierung elektronisch vorliegender Geodaten*. https://www.bundesarchiv.de/imperia/md/content/bundesarchiv_de/fachinformation/ark/handreichung_geodaten_20090928.pdf

Bartelme, N. (2005). *Geoinformatik - Modelle, Strukturen, Funktionen.* 4. Auflage. Springer - Verlag Berlin Heidelberg.

BEV - Bundesamt für Eich- und Vermessungswesen. (2012). *Antrag auf Durchführung einer Amtshandlung*. http://www.bev.gv.at/pls/portal/docs/PAGE/BEV_PORTAL_CONTENT_ALLGEMEIN/0200_PRODUKTE/BESTELLFORMULARE/ANTRAG_AUF_DURCHFUEHRUNG_EINER_AMTSHANDLUNG.PDF. Zuletzt besucht am 13.05.2015

Booth, J. M., & Jeremy, G. (2006). *Optimizing OCR Accuracy on Older Documents: A Study of Scan Mode, File Enhancement, and Software Products.* Version 2.0. U.S. Government Printing Office, Washington, DC. http://www.gpo.gov/pdfs/fdsys-info/documents/WhitePaper-OptimizingOCRAccuracy.pdf.

Bundeskanzleramt Österreich. (2015). *Digitales Archiv Österreich*. http://www.bundeskanzleramt.at/site/cob__41804/7440/default.aspx. Zuletzt besucht am 19.04.2015

callas software GmbH. (2011). *pdfaPilot Handbuch*. http://www.callassoftware.com/files/attachments/.321/callas_pdfaPilot_Handbuch_DE.pdf. Zuletzt besucht am 03.06.2015

callas software GmbH. (2015a). *Über callas software*. http://www.callassoftware.com/de/kontakt. Zuletzt besucht am 14.04.2015

callas software GmbH. (2015b). *Produkte*. http://www.callassoftware.com/de/produkte. Zuletzt besucht am 06.05.2015

callas software GmbH. (2015c). *pdfToolbox*. http://www.callassoftware.com/de/produkte/pdftoolbox. Zuletzt besucht am 06.05.2015

callas software GmbH. (2015d). *pdfaPilot*. http://www.callassoftware.com/de/produkte/pdfapilot. Zuletzt besucht am 16.05.2015

callas software GmbH. (2015e). *pdfapilot - Funktionen*. http://www.callassoftware.com/de/produkte/pdfapilot/?type=product&product=pdfapilotdesktop&tab=key-features. Zuletzt besucht am 06.05.2015

callas software GmbH. (2015f). *Adobe setzt verstärkt auf PDF-Kompetenz von callas software bei Acrobat DC*. http://www.callassoftware.com/de/news/2015/05/adobe-setzt-verstaerkt-auf-pdf-kompetenz-von-callas-software-bei-acrobat-dc. Zuletzt besucht am 13.05.2015

callas software GmbH. (2015g). *Anwenderberichte*. https://www.callassoftware.com/de/Zitate. Zuletzt besucht am 30.07.2015

Cognitive Technologies. (2015). *Cognitive PDF/A.* http://www.cognitiveforms.com/files/white_paper_PDF_A_en.pdf. Zuletzt besucht am 29.04.2015

Computer Bild. (2009a). *Wissen: Alles über Texterkennung*. http://www.computerbild.de/artikel/cb-Ratgeber-So-funktioniert-Texterkennung-4514615.html. Zuletzt besucht am 11.06.2015

Computer Bild. (2009b). *Zeichenerkennung im Detail*. http://www.computerbild.de/artikel/cb-Ratgeber-So-funktioniert-Texterkennung-4514908.html. Zuletzt besucht am 11.06.2015

Computer Bild. (2009c). *So aufwendig ist Texterkennung*. http://www.computerbild.de/artikel/cb-Ratgeber-So-funktioniert-Texterkennung-4515036.html. Zuletzt besucht am 11.06.2015

Drümmer, O., Oettler, A., & Seggern, D. v. (2007). *PDF/A kompakt. Digitale Langzeitarchivierung mit PDF.* Version 1. callas software GmbH, Berlin. http://www.pdfa.org/wp-content/uploads/2011/08/PDFA_kompakt_pdfa1b.pdf.

EUR-Lex. (2007). *Richtlinie 2007/2/EG des Europäischen Parlaments und des Rates vom 14. März 2007 zur Schaffung einer Geodateninfrastruktur in der Europäischen Gemeinschaft (INSPIRE)*. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2007:108:0001:0014:de:PDF. Zuletzt besucht am 16.04.2015

GBO-Grundbuchordnung. (2015). *Aufbewahrung von Grundbüchern und Urkunden beim Grundbuchamt*. http://www.rechtsportal.de/Gesetze/Gesetze/Allgemeines-Zivilrecht/Grundbuchordnung/ERSTER-ABSCHNITT-Allgemeine-Vorschriften/10-Aufbewahrung-von-Grundbuechern-und-Urkunden-beim-Grundbuchamt. Zuletzt besucht am 03.06.2015

GeoDIG. (2010). *Geodateninfrastrukturgesetz 14/2010*. In der zum Zeitpunkt der Erstellung dieser Masterarbeit gültigen Fassung vom 14.12.2012 (BGBl 109/2012). https://www.ris.bka.gv.at/Dokumente/BgblAuth/BGBLA_2010_I_14/BGBLA_2010_I_14.html. Zuletzt besucht am 16.04.2015

GeoPortal Saarland. (2015). *Geodaten und Metadaten*. http://geoportal.saarland.de/portal/de/informationen-de/geodaten-und-metadaten.html. Zuletzt besucht am 19.04.2015

Grundbuchs-Novelle. (2007). *155/ME XXIII.GP - Ministerialentwurf - Materialien*. http://www.parlament.gv.at/PAKT/VHG/XXIII/ME/ME_00155/imfname_096675.pdf. Zuletzt besucht am 02.06.2015

Hermann, T. (2008). *Evaluierung von schnellen Verfahren zur Zeichenerkennung mit Kameras.* Diplomarbeit. http://daiw.de/Diplomarbeit/Diplomarbeit_Tobias_Hermann.pdf.

intarsys consulting GmbH. (2015a). *PDF/A Live! image extension*. https://www.intarsys.de/pdf-produkte/pdfa-extended. Zuletzt besucht am 07.05.2015

intarsys consulting GmbH. (2015b). *Unsere Kunden*. https://www.intarsys.de/referenzen. Zuletzt besucht am 30.07.2015

ISO. (2005). *Document management - Electronic document file format for long-term preservation - Part 1: Use of PDF 1.4 (PDF/A-1).*

ISO. (2011). *Document management - Electronic document file format for long-term preservation - Part 2: Use of ISO 32000-1 (PDF/A-2).*

ISO. (2012). *Document management - Electronic document file format for long-term preservation - Part 3: Use of ISO 32000-1 with support for embedded files (PDF/A-3).*

Justizportal des Bundes und der Länder. (2015). *Verzeichnisse*. http://www.justiz.de/verzeichnis/index.php. Zuletzt besucht am 05.05.2015

Land Steiermark - Amt der Steiermärkischen Landesregierung. (2015). *GIS-Steiermark GeoDaten*. http://www.gis.steiermark.at/cms/ziel/74005/de/. Zuletzt besucht am 19.04.2015

Neuroth, H., Oßwald, A., Scheffel, R., Strathmann, S., & Huth, K. (2010). *nestor Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung.* Version 2.3. http://nestor.sub.uni-goettingen.de/handbuch/nestor-handbuch_23.pdf.

Nuance Communications. (2013). *Omnipage Ultimate - User's Guide*. http://www.nuance.de/ucmprod/groups/imaging/@web-enuk/documents/collateral/nc_028956.pdf. Zuletzt besucht am 04.06.2015

Nuance Communications. (2015). *Mit OmniPage Dokumente nicht nur konvertieren, sondern auch transformieren*. http://www.nuance.de/for-business/by-product/omnipage/ultimate/index.htm. Zuletzt besucht am 02.07.2015

Nuance Communications. (2015a). *OmniPage Ultimate*. http://www.nuance.de/for-business/by-product/omnipage/ultimate/index.htm. Zuletzt besucht am 02.06.2015

Nuance Communications. (2015b). *Nuance Fakten*. http://www.nuance.de/company/company-overview/fast-facts/index.htm. Zuletzt besucht am 04.06.2015

Nuance Communications. (2015c). *Nuance - Wir über uns*. http://www.nuance.de/company/index.htm. Zuletzt besucht am 04.06.2015

Oettler, A. (2013). *PDF/A kompakt 2.0: PDF für die Langzeitarchivierung. Der ISO-Standard – von PDF/A-1 bis PDF/A-3.* 1. Auflage. Association for Digital Document Standards e.V., Berlin. http://www.pdfa.org/wp-content/uploads/2013/03/PDFA-kompakt-2_0_screen.pdf.

Österreichisches Parlament. (2015). *Entwurf. Bundesgesetz, zur Schaffung einer Geodateninfrastruktur des Bundes (Geodateninfrastrukturgesetz – GeoDIG)*. http://www.parlament.gv.at/PAKT/VHG/XXIV/ME/ME_00055/fnameorig_156209.html. Zuletzt besucht am 19.04.2015

PDF Association. (2011). *PDF/A – ein Blick auf die technische Seite*. http://www.pdfa.org/2011/09/pdfa-%E2%80%93-ein-blick-auf-die-technische-seite/?lang=de. Zuletzt besucht am 28.04.2015

PDF Association. (2014). *PDF/A (Flyer, deutsch)*. http://www.pdfa.org/publication/pdfa-flyer-deutsch/?lang=de. Zuletzt besucht am 06.05.2015

PDF Association. (2015a). *Die zehn am weitesten verbreiteten Märchen über PDF/A*. http://www.pdfa.org/competence-center/pdfa-competence-center/die-zehn-am-weitesten-verbreiteten-marchen-uber-pdfa/?lang=de. Zuletzt besucht am 28.04.2015

PDF Association. (2015b). *Antworten auf häufig gestellte Fragen zu PDF/A*. http://www.pdfa.org/competence-center/pdfa-competence-center/antworten-auf-haufig-gestellte-fragen-zu-pdfa/?lang=de. Zuletzt besucht am 28.04.2015

PDF Association. (2015c). *PDF/A Competence Center*. http://www.pdfa.org/competence-center/pdfa-competence-center/?lang=de. Zuletzt besucht am 07.05.2015

PDF Association. (2015d). *Olaf Drümmer*. http://www.pdfa.org/author/olafdruemmer/?lang=de. Zuletzt besucht am 06.05.2015

PDF Association. (2015e). *LuraTech-Kunde DAK-Gesundheit gewinnt den ersten PDF/A User Award*. http://www.pdfa.org/press_release/luratech-kunde-dak-gesundheit-gewinnt-den-ersten-pdfa-user-award/?lang=de. Zuletzt besucht am 30.07.2015

PDF Association. (2015f). *Nuance Communications GmbH*. http://www.pdfa.org/organization/nuance-communications-gmbh/?lang=de. Zuletzt besucht am 07.08.2015

PDF Tools AG. (2009). *PDF/A - der Standard für die Langzeitarchivierung*. Version 2.4. https://www.pdf-tools.com/public/downloads/whitepapers/whitepaper-pdfa-de.pdf. Zuletzt besucht am 16.04.2015

PDF Tools AG. (2011). *PDF/A-2 Übersicht*. http://www.pdf-tools.com/public/downloads/flyers/Flyer-PDFA2-Uebersicht-DE.pdf. Zuletzt besucht am 16.04.2015

PDF Tools AG. (2012). *PDF/A-3 Übersicht*. http://www.pdf-tools.com/public/downloads/flyers/Flyer-PDFA2-Uebersicht-DE.pdf. Zuletzt besucht am 16.04.2015

PDF Tools AG. (2014). *Die 10 wichtigsten Dinge, die Sie über PDF/A wissen sollten*. http://www.pdf-tools.com/public/downloads/flyers/Flyer-PDFA-10-Dinge-DE.pdf. Zuletzt besucht am 28.04.2015

PDF Tools AG. (2015a). *PDF ISO-Standards*. http://www.pdf-tools.com/pdf/pdf-iso-standard-pdfa-pdfx.aspx. Zuletzt besucht am 28.04.2015

PDF Tools AG. (2015b). *Kundenreferenzen*. http://www.pdf-tools.com/pdf/Kundenreferenzen.aspx. Zuletzt besucht am 30.07.2015

Primas, E. (2013/14). *E-Geo-Government.* Vorlesungsskriptum zu Selected Topics A (LV-Nr. 510.320 und 510.321).

Ruth, M. (2011). *GeoTIFF FAQ Version 2.4*. Version 2.4. http://www.remotesensing.org/geotiff/faq.html. Zuletzt besucht am 28.04.2015

Solid Documents. (2015a). *Solid OCR*. http://www.soliddocuments.com/de/solid-ocr.htm. Zuletzt besucht am 29.04.2015

Solid Documents. (2015b). *Solid OCR*. http://downloads.soliddocuments.com/pdfs/solid_ocr.pdf. Zuletzt besucht am 29.04.2015

Stadt Graz. (2015). *Geodaten*. http://www.geoportal.graz.at/cms/ziel/4751551/DE/. Zuletzt besucht am 19.04.2015

Tanner, S. (2004). *Deciding Whether Optical Character Recognition is Feasible.* King's Digital Consultancy Services. http://www.odl.ox.ac.uk/papers/OCRFeasibility_final.pdf.

UAR. (2007). *Urkundenarchivrichtlinie der Österreichischen Notariatskammer*. In der zum Zeitpunkt der Erstellung dieser Masterarbeit gültigen Fassung vom 18.04.2013. http://www.notar.at/download_file/force/101/270/. Zuletzt besucht am 28.04.2015

Unicode Inc. (2014). *About the Unicode Standard*. http://www.unicode.org/standard/standard.html. Zuletzt besucht am 29.04.2015

Urkundenarchiv-RL. (2007). *Richtlinie gemäß § 37 Abs 1 Z 7 RAO über die Errichtung und Führung eines anwaltlichen Urkundenarchivs (Urkundenarchiv-RL). Österreichischer Rechtsanwaltskammertag*. In der zum Zeitpunkt der Erstellung dieser Masterarbeit gültigen Fassung vom 01.07.2007. http://www.rechtsanwaelte.at/index.php?eID=tx_nawsecuredl&u=0&g=0&t=1413437249&hash=7f54af0137125f977ce330dd42f251

c8c8e22d37&file=uploads/tx_templavoila/urkundenarchiv_rl01072007.pdf. Zuletzt besucht am 30.05.2015

Urkundenarchivverordnung der Bundes-Architekten- und Ingenieurkonsulentenkammer. (2007). *Amtliche Nachrichten der Architekten und Ingenieurkonsulenten 219/07*. In der zum Zeitpunkt der Erstellung dieser Masterarbeit gültigen Fassung vom I/2013. http://www.arching.at/baik/upload/pdf/amtliche%20nachrichten/48_197.vo_194_198_signaturkarten_vo.pdf. Zuletzt besucht am 28.04.2015

Vorbach, P. (2014). *Analysen und Heuristiken zur Verbesserung von OCR-Ergebnissen bei Frakturtexten.* Masterarbeit. https://opus.bibliothek.uni-wuerzburg.de/files/10652/Vorbach_Paul_Fraktur-OCR.pdf.

Wegner, H. (2000). *Metadaten als Mittel zur Organisation von Geodaten in der Stadtplanung*. http://www.agit.at/php_files/myagit/papers/2000/wegner_EA_6.pdf

ZTG. (1993). *Ziviltechnikergesetz 1993-ZTG 156/1994*. In der zum Zeitpunkt der Erstellung dieser Masterarbeit gültigen Fassung vom 21.01.2015 (BGBl 4/2013). https://www.ris.bka.gv.at/GeltendeFassung.wxe?Abfrage=Bundesnormen&Gesetzesnummer=10012368. Zuletzt besucht am 28.04.2015