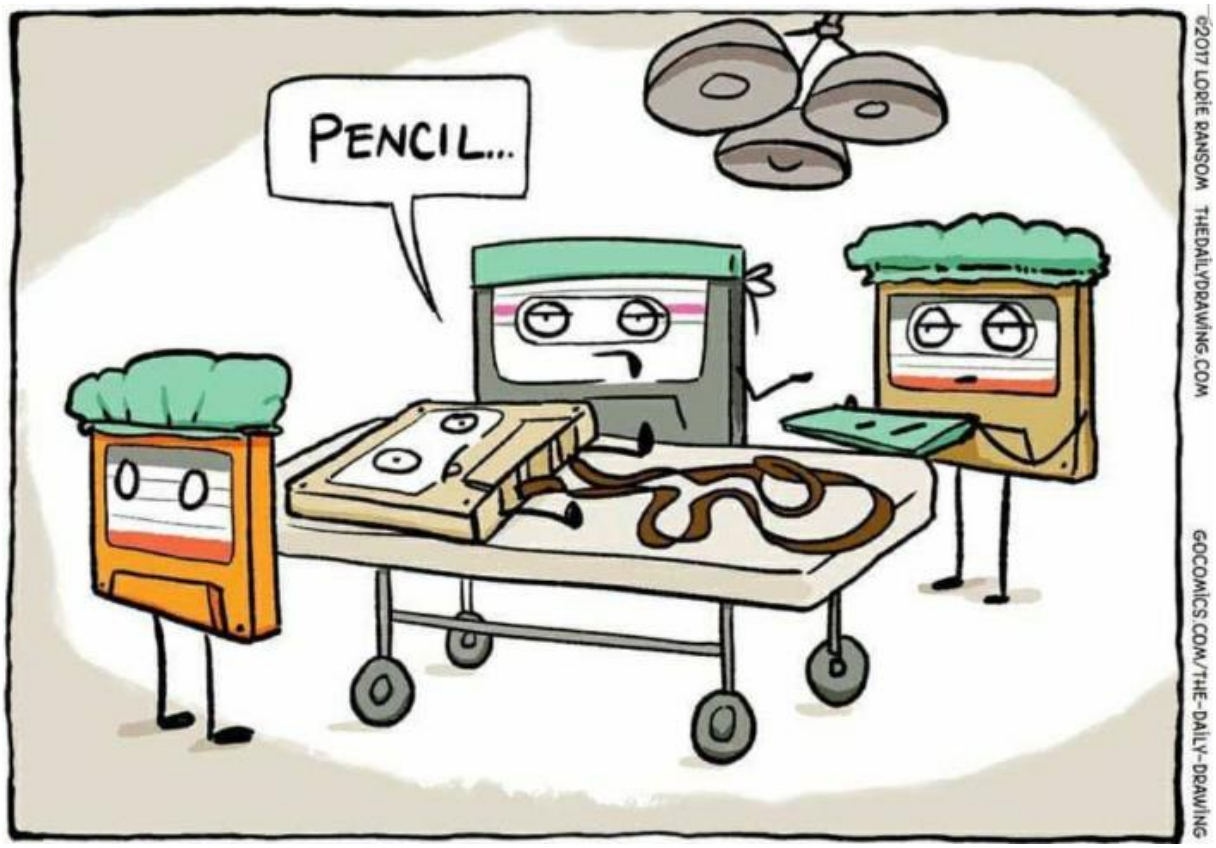
PAWI HS18

Datenarchäologie fürs KKL



Version	Datum	Autor	Änderungsgrund / Bemerkungen
0.1	18.09.2018	F. Rohrbach	Ersterstellung
0.2	22.09.2018	F. Rohrbach	Einleitung hinzugefügt
0.3	29.09.2018	F. Rohrbach	Projektplan & Anforderungskatalog hinzugefügt
0.4	01.10.2018	F. Rohrbach	Dateitypbestimmung & Konvertierung hinzugefügt
0.5	07.10.2018	F. Rohrbach	Volltextsuche & Indexierung hinzugefügt
0.6	07.12.2018	F. Rohrbach	Umstrukturierung
0.7	10.12.2018	F. Rohrbach	Korrekturen, Quellenverzeichnis, Anhang
0.8	23.12.2018	F. Rohrbach	Korrekturen, Zufriedenheit des Kundens

Inhaltsverzeichnis

Inhalt

1 Einleitung

Das Kongress- und Kulturzentrum Luzern (KKL) wurde zwischen 1992 und 1998 unter der Leitung von Dr. Thomas Held erbaut, der auch als Direktor der Stiftung „Avenir Suisse“ und als Kolumnist bekannt geworden ist. Während dieser Zeit entstand eine grosse Sammlung an digitalen Dokumenten des KKLs.
Etwa im Jahre 1997 wurde aufgrund von Diskussionen mit dem Totalunternehmer eine Dateistruktur mit dem Namen „Synopsis“ aufgebaut, in der alle gesammelte Dokumente abgelegt wurden. Diese enthält zum einen von der Trägerstiftung selbst produzierte Papiere (vor allem Korrespondenzen im WordPerfect Format) und zum anderen alle seit ca. 1994 empfangene Briefe und Unterlagen (eingescannt als TIFF Format). Insgesamt umfasst die Datenbank etwas mehr als 50‘000 Dokumente (52‘824 Dateien und 4‘666 Ordner).
Zudem wurde eine Access-Datenbank angelegt, in der alle Dokumente im TIFF Format mit Schlagworten ergänzt wurden. Die Datenbank enthält jeweils den Pfad zum Dokument in der Ordnerstruktur sowie die dazugehörigen Schlagwörter. Um ein gesuchtes Dokument in der Synopsis-Datenbank zu finden, muss bisher über die Suchfunktion von Access zurückgegriffen werden.
Die TIFF Dateien wurden mit kryptischen Namen versehen und in einer Ordnerstruktur nach Jahr, Monat und Tag abgelegt. Die WordPerfect Dateien wurden bisher nur in einer Ordnerstruktur nach Jahren geordnet, darunter wurden die Ordner meist mit Abkürzungen gekennzeichnet. Ausserdem wurden die Dateiendungen der WordPerfect Dateien für den Autorenkürzel sowie die Dokumentversion verwendet.
Die Dokumente in der „Synopsis“ Sammlung haben teilweise auch heute noch für Architekten, Ingenieure, Projektleiter und vor allem auch Baujuristen eine Relevanz. Ebenfalls aus lokalhistorischer und politikwissenschaftlicher Sicht ist dieser Datenbestand interessant, da er dokumentiert wie das KKL entstanden ist.
Das Kongress- und Kulturzentrum Luzern (KKL) wird 2018 20 Jahre alt. Aus diesem Anlass wird das lange weggelegte Daten-Aufbereitungsprojekt von Dr. Thomas Held (Auftraggeber dieses Projekts) nochmals in Angriff genommen.

1.1 Problemstellung

Digitale Daten sind neben verschiedensten Umwelteinflüssen auch noch dem immer schnelleren Wandel im Informatikbereich ausgesetzt. Datenträger und Formate verändern sich immer schneller. Werden die Daten nicht regelmässig gepflegt, so setzt man diese der Gefahr aus, nicht mehr lesbar zu sein.
In der Synopsis Datensammlung liegen viele Daten aus den 90er Jahren, welche heute nur noch schwer lesbar sind, da man erst die benötigten Programme aus der damaligen Zeit ausfindig machen muss (falls diese noch existieren). Zudem liegen die zu den Dokumenten gehörigen Metadaten an verschiedenen Orten (z.B. in der Ordnerstruktur oder in der Access Datenbank).

1.2 Ziel der Arbeit

Durch die Datenarchäologie soll ermöglicht werden, dass die digitalen Dokumente für die Nachwelt erhalten bleiben können. Dies soll mittels Umwandlung der TIFF- sowie WordPerfect-Dateien in das neuere PDF/A Format gewährleistet werden. Zusätzlich soll eine Web-Oberfläche zur Verfügung gestellt werden, welche berechtigten Anwendern die Suche nach Dateinamen, Schlagworten, Daten, sowie Volltextsuche in den Dateien im PDF/A Format ermöglicht.

2 Stand der Technik

2.1 Konvertierung

Um die Konvertierung der WordPerfect- sowie TIFF-Dateien anzugehen, musste sich zuerst mit den verschiedenen PDF Typen, Versionen und Formate befasst werden.

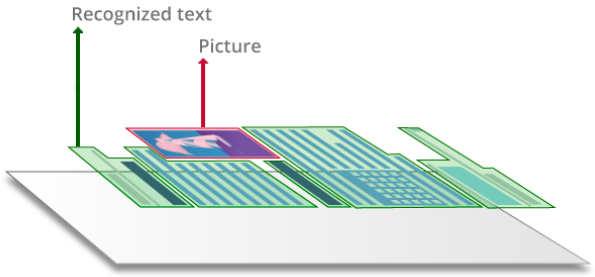
2.1.1 PDF Typen, Versionen und Standards

Typen

PDF („Portable Document Format“) Dokumente können grundsätzlich in drei unterschiedliche Typen [5] unterteilt werden:

- True PDF (Digital erstellte PDFs)

  Der Typ „True PDF“ enthält eine Textebene sowie eine Ebene für Grafiken. Dieser Typ entsteht meist durch Verwendung von virtuellen Druckern via der „Druckfunktion“ oder z.B. über die Microsoft Office Produkte. Der Unterschied zum Typ „Searchable PDF“ liegt darin, dass die Grafikebene komplett von der Textebene im PDF getrennt ist.

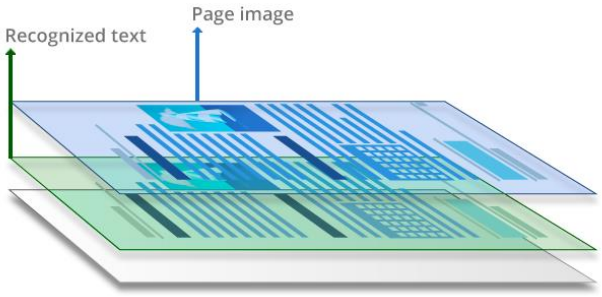


- Image-only PDF (Eingescannte PDFs):

  Der Typ „Image-only PDF“ enthält nur gerade eine Ebene, welche das Dokument als Grafik enthält. Dieser Typ PDF ist nicht durchsuchbar, kann jedoch mit Techniken wie OCR durchsuchbar gemacht werden.

- Searchable PDF (Durchsuchbare PDFs):

  Der Typ „Searchable PDF“ entsteht z.B., wenn ein PDF vom Typ „Image-only PDF“ mit der „Optical Character Recognition“ (OCR) nach Text durchsucht wird. Dabei wird neben der vorhandenen Grafikebene des „Image-only PDF“ eine weitere Ebene mit dem durch OCR erkannten Text hinzugefügt. Der Typ „Searchable PDF“ kann somit nach allen erkannten Texten durchsucht werden.



Versionen

Im Laufe der Zeit wurden dem PDF Format immer mehr Verbesserungen und Funktionen durch die Entwickler von „Adobe Systems“ hinzugefügt. Folgend ein kurzer Überblick der verschiedenen Versionen [6] mit den wichtigsten Neuerungen:

Version	Beschreibung
1.0	optisch attraktive Textdarstellung, Einbettung von Lesezeichen und dateiinternen Querverweisen, Reader heißt Carousel
1.1	externe Querverweise, Einbettung von Multimedia-Dateien in mittlerweile obsoleten Sound- und Video-Formaten, Dokumente können durchsucht werden, 40-Bit-Verschlüsselung
1.2	Möglichkeit der Verwendung des CMYK-Farbmodells, direktes Öffnen im Browserfenster durch Implementierung von Browser-Erweiterungen (sog. PlugIns), Zugänglichkeits-Plugin für Blinde für den Acrobat Reader 3, interaktive Elemente wie Checkboxen und Radiobuttons sind möglich
1.3	Unterstützung asiatischer Schriften, verbessertes Accessibility Plugin für den Acrobat Reader, weiterhin 40-Bit-Verschlüsselung, digitale Signaturen, JavaScript-Elemente möglich
1.4	Wegfall des Zugänglichkeits-Plugin, dafür Unterstützung von MSAA (Microsoft Active Accessibility) durch den Acrobat Reader, RC4-Verschlüsselung mit 40–128 Bit, Beschreibung des logischen Dokumentaufbaus mit Tags
1.5	Einbettung von Bildern im Format JPEG 2000, Filmen im Format MPEG und Audiodateien im Format MP3 möglich, Public Key-Verschlüsselung PKCS#7; Leseprogramm heißt ab jetzt Adobe Reader, keine Unterstützung mehr für DOS-basierte Windows-Versionen (Windows 95, 98, ME)
1.6	Unterstützung für das Universal-3D-Dateiformat, Einbettung von OpenType-Fonts, Unterstützung für XFA 2.2 Rich-Text-Elemente und Attribute, AES-Verschlüsselung, PKCS#7-Verschlüsselung mit SHA256, E-MAIL bis zu 4096 Bit, unzugängliche PDF-Dokumente können mit Tags versehen werden, bessere Unterstützung für mehrspaltige Dokumente, Formularfelder können mit Hilfe der Sprachausgabefunktionen vorgelesen werden, PDF-Creator mit Ausgabehilfeassistent: unterstützt beim Optimieren für Screenreader und Bildschirmvergrößerungsprogramme
1.7	weitere Verbesserung der 3D-Darstellungsoptionen, stärkere Verschlüsselungsalgorithmen (PKCS#7 mit SHA384, SHA512 und RIPEMD-160, 256-Bit AES), weitere Verbesserungen
2.0	Unterstützung für das PRC-Dateiformat, 3D-Messwerkzeuge, verbesserte Sicherheitsfunktionen

Standards (Formate)

Neben den PDF Versionen kamen im Laufe der Zeit auch das Bedürfnis nach genormten PDF Standards (nach ISO Standard) hinzu. Folgend eine kurze Übersicht über die verschiedenen PDF Standards (Formate) [7][8][9] und deren wichtigsten Eigenschaften:

Standard	Beschreibung
PDF/A	ISO genormtes Format für Langzeitarchivierung
PDF/A-1	Langzeitarchivierung basierend auf Version 1.4 mit ISO Norm ISO 19005-1:2005
PDF/A-1a	Eindeutige visuelle Reproduzierbarkeit als auch inhaltliche Strukturierung
PDF/A-1b	Eindeutige visuelle Reproduzierbarkeit
PDF/A-2	Langzeitarchivierung basierend auf Version 1.7 mit ISO Norm ISO 19005-2:2011
PDF/A-2a	Realisierung aller Anforderungen der ISO 19005-2, insbesondere alle strukturellen und semantischen Eigenschaften.
PDF/A-2b	Mindestanforderung an PDF/A-2 Dokument
PDF/A-2u	Vereinfachung der Durchsuchbarkeit von Texten in Unicode-Format
PDF/A-3	Langzeitarchivierung basierend auf Version 1.7 mit ISO Norm ISO 19005-3:2012 und mit Container
PDF/A-3a	Barrierefreiheit
PDF/A-3b	Visuelle Integrität
PDF/A-3u	Vereinfachung der Durchsuchbarkeit von Texten in Unicode-Format
PDF/E	Format für Ingenieurwesen mit Fähigkeit zur interaktiven 3D-Darstellung. Genormt auf ISO 24517
PDF/UA	Format für barrierefreie PDF-Dokumente
PDF/VT	Format für Variablen Datendruck (VDP)
PDF/X	Format für Verwendung von Druckvorlage

2.1.2 PDF/A Konvertierung – WordPerfect und TIFF

Um die WordPerfect- und TIFF-Dateien ins PDF/A Format umzuwandeln, wurden verschiedene Softwarelösungen getestet. Hierbei ist es für die spätere Indexierung (siehe Punkt 3.8) notwendig, dass nach der Konvertierung das erstellte PDF eine Textebene besitzt. Bei Dokumenten mit Text-Formatierung wie z.B. WordPerfect-Dateien wird die Textebene meist ohne weiteren Aufwand automatisch beim generieren des PDFs miterstellt. Bei Dokumenten ohne Text-Formatierung wie z.B. die TIFF-Dateien (eingescannte Dokumente) wird die Textebene nicht ohne Verwendung von OCR Software erstellt („Image-only PDF"). Mit Hilfe einer OCR Software kann ein „Seachable PDF" von einer TIFF-Datei erstellt werden (siehe Punkt 2.2.1)".

OfficeToPDF

„OfficeToPDF" [10] von „CogniDox" ist eine Open-Source Software, die es über die Kommandozeile erlaubt, verschiedene Dokumenttypen ins PDF Format zu konvertieren.

FoxPDF WordPerfect to PDF Converter

Die kommerzielle Software „FoxPDF WordPerfect to PDF Converter" [11] der Firma „FoxPDF" wurde spezielle für die Konvertierung von WordPerfect Dateien ins PDF Format ausgelegt.

ABCpdf .NET Converter

„ABCpdf .NET Converter" [12] ist eine Bibliothek von „webSupergoo", welche es ermöglicht über eine Schnittstelle verschiedenste Dateitypen ins PDF Format zu konvertieren.

Corel WordPerfect X9

„WordPerfect" [13] der Firma „Corel" ist Software mit der die ursprünglichen „WPD"-Dokumente erstellt wurden. Seit der Version 12 unterstützt es die Funktion „Publish to PDF" welche es ermöglicht, das WordPerfect-Dokument im PDF Format abzuspeichern.

PDF-XChange Standard

Die kommerzielle Software „PDF-XChange Standard" [14] von „1 for All Software" ist ein virtueller Drucktreiber. Damit lassen sich über jegliche Programme die eine Druckfunktion anbieten, den zu druckenden Inhalt als PDF abzuspeichern.

PDF24 Creator

„geek Software" bietet mit der Freeware „PDF24 Creator" [15] einen weiteren virtuellen Drucktreiber.

2.2 Texterkennung (OCR)

Die WordPerfect-Dateien werden beim Konvertieren ins PDF/A Format meist automatisch mit einer Textebene ausgestatten („Searchable PDF“).

Die ins PDF/A konvertierte TIFF-Dateien besitzen jedoch nach der Konvertierung meist nicht automatisch eine Textebene („Image-only PDF“), weshalb der Text mittels Texterkennungssoftware bzw. „Optical Character Recognition“ (OCR) erkannt und danach als Textebene in die PDF/A-Datei zusammengefügt werden muss. Dadurch kann ein „Searchable PDF“ erstellt werden. Folgend wurden verschiedene Softwarelösungen dazu geprüft.

Tesseract-OCR

Die Open-Source Software „Tesseract-OCR“ [16] von „Apache“ bietet eine solide Lösung für Texterkennung, welche nach Wunsch auch mittels Parameter angepasst bzw. optimiert werden kann. „Tesseract-OCR“ gilt als Vorreiter im Bezug auf die Texterkennung, weshalb auch viele kommerzielle Software die diese Funktionalität benötigen, „Tesseract-OCR“ verwenden.

IronOcr

„IronOcr“ [18] der Firma „Iron“ ist eine Freeware C# Biblothek für die Texterkennung. Zu erwähnen ist ihre einfache Ansteuerung über die Schnittstellen in der .NET Entwicklung.

2.2.1 „Image-only PDF“ zu „Searchable PDF“

Um den aus den TIFF-Dateien erstellten „Image-only PDF“ eine Textebene mit dem durch die Texterkennung erkanntem Text hinzuzufügen und somit daraus ein „Searchable PDF“ zu generieren, wurden verschiedene Softwarelösungen geprüft.

Acrobat Pro DC

Die kommerzielle Software „Acrobat Standard DC“ [59] von „Adobe“ ist eine der bekanntesten Software im Bereich des PDF Formats. Sie enthält auch eine Funktionalität um ein „Image-only PDF“ in ein „Searchable PDF“ umzuwandeln. Dabei wird die integrierte Texterkennung verwendet.

PDF-XChange + Ghostscript + iTextSharp + Tesseract-OCR + hOcr2Pdf.Net

Eine weitere Möglichkeit ein „Image-only PDF“ in ein „Searchable PDF“ zu konvertieren, bietet die Kombination von verschiedener Freeware (PDF-XChange + Ghostscript + iTextSharp + Tesseract-OCR + hOcr2Pdf.Net [19]). Die verschiedenen Freeware Produkte tragen alle einen Beitrag zum Endprodukt „Searchable PDF“ bei.

2.2.2 Auslesen der Textebene

Um für die Volltextsuche (siehe Punkt 3.8) an den in der Textebene der „Searchable PDF“ liegenden Text zu gelangen, wurden verschiedene Softwarelösungen geprüft.

PDFBox.NET

Die Freeware Bibliothek „PDFBox In .NET“ [24] ist eine für die .NET Entwicklung Portierung Software „PDFBBox“ der Firma „Square PDF .NET“.

TikaOnDotnet

„TikaOnDotnet“ [17] ist eine Freeware Bibliothek die auf der „Tika“ Bibliothek von „Apache“ aufbaut.

2.3 Volltextsuche und Indexierung

Die Volltextsuche ermöglicht im Gegensatz zur normalen Suche, bei der nur nach Metadaten eines Dokuments gesucht werden kann, die Suche auch innerhalb von Dokumenten. Um diese der späteren Web-Anwendung zur Verfügung zu stellen, muss der Text der ins PDF-Format konvertierten Dokumente ausgelesen und durchsuchbar gemacht werden. Bei grösseren Datenmengen kann das Durchsuchen von allen Texten der Dokumente in einer Datenbank oft zeitaufwendig sein. Abhilfe schafft dabei die „Indexierung“, welche die Texte der Dokumente nach möglichen Suchbergriffen durchsucht und diese für den späteren Gebrauch speichert.

Zur Realisierung der Volltextsuche und Indexierung gibt es mehrere Lösungsansätze:

- Verwendung der Datenbank-Volltextsuche
- Verwendung von Volltextsuche-Software

2.3.1 Datenbank-Volltextsuche (DBMS)

Einige Datenbanken unterstützen von Haus aus eine Volltextsuche (meist inklusive Indexierung). Diese erkennen den Text der unterstützten Dokumente, die in der Datenbank liegen von selbst aus (z.B. bei den PDF Dateien der Text der Textebene). Ausserdem übernehmen diese meist auch gleich die Indexierung. Vorteil ist hier sicher die einfachere Implementierung sowie die hohe Geschwindigkeit durch die Indexierung. Nachteil ist die Abhängigkeit zum verwendeten Datenbank-Typ, sowie Notwendigkeit sogenannter „iFilter“ [25][26][27] (zur Unterstützung von verschiedener Dateiformate) für das zu indexierende Format. Die in der Datenbank liegenden PDF Dateien müssen zwingend eine Textebene besitzen, um den Inhalt über den „iFilter“ auszulesen.

Microsoft SQL-Volltextsuche

Die Volltextsuche inklusive Indexierung der Microsoft SQL Datenbank ermöglicht eine einfache Volltextsuche über Dokumente, die in der SQL Datenbank liegen. Die einfache Erstellung des Indexes („Indexierung“) mit Hilfe von „iFilter“ zählt zu den Vorteilen. Jedoch bietet die Volltextsuche wenig Funktionalitäten, so muss man z.B. mit viel Aufwand SQL-Abfragen bauen, um mehrere Suchwörter zu berücksichtigen. Bei komplexeren Suchabfragen ist die Microsoft SQL-Volltextsuche im Vergleich zu „Lucene“ (siehe Punkt 3.8) langsamer.

2.3.2 Volltextsuche-Software (Full Text Search Engine)

Ein weiterer Ansatz bieten Software wie z.B. „Lucene“ an. Dabei wird die Volltextsuche von einer Software übernommen. Die Software liest ebenfalls wie die Datenbank-Volltextsuche die Texte der Dokumente aus und legt diese im „Index“ ab. Bei „Lucene“ wird dieser „Index“ lokal abgelegt. Vorteil ist hier die Unabhängigkeit zum Datenbank-Typ, die einfache Implementierung sowie die hohe Geschwindigkeit durch die Indexierung. Nachteil ist „Overhead“ der entsteht, da zwischen Datenbank und Web-Anwendung noch die Volltextsuche-Software steht.

Lucene

„Lucene“ [28] ist eine Open-Source Softwarelösung von „Apache Software Foundation“ welche sich vor allem durch die Leistungsfähigkeit, Skalierbarkeit und Anpassbarkeit auszeichnet. Ein Vorteil gegenüber der Datenbank-Volltextsuche ist der Umfang der Funktionalitäten, die ohne grossen Entwicklungsaufwand genutzt werden kann wie z.B. die „Score“ mit der „Lucene“ automatisch die gefundenen Resultate nach Relevanz ordnet. Auch die Suchabfrage nach mehreren Eigenschaften sind in „Lucene“ einfacher zu realisieren als bei einer Datenbank-Volltextsuche.

Solr

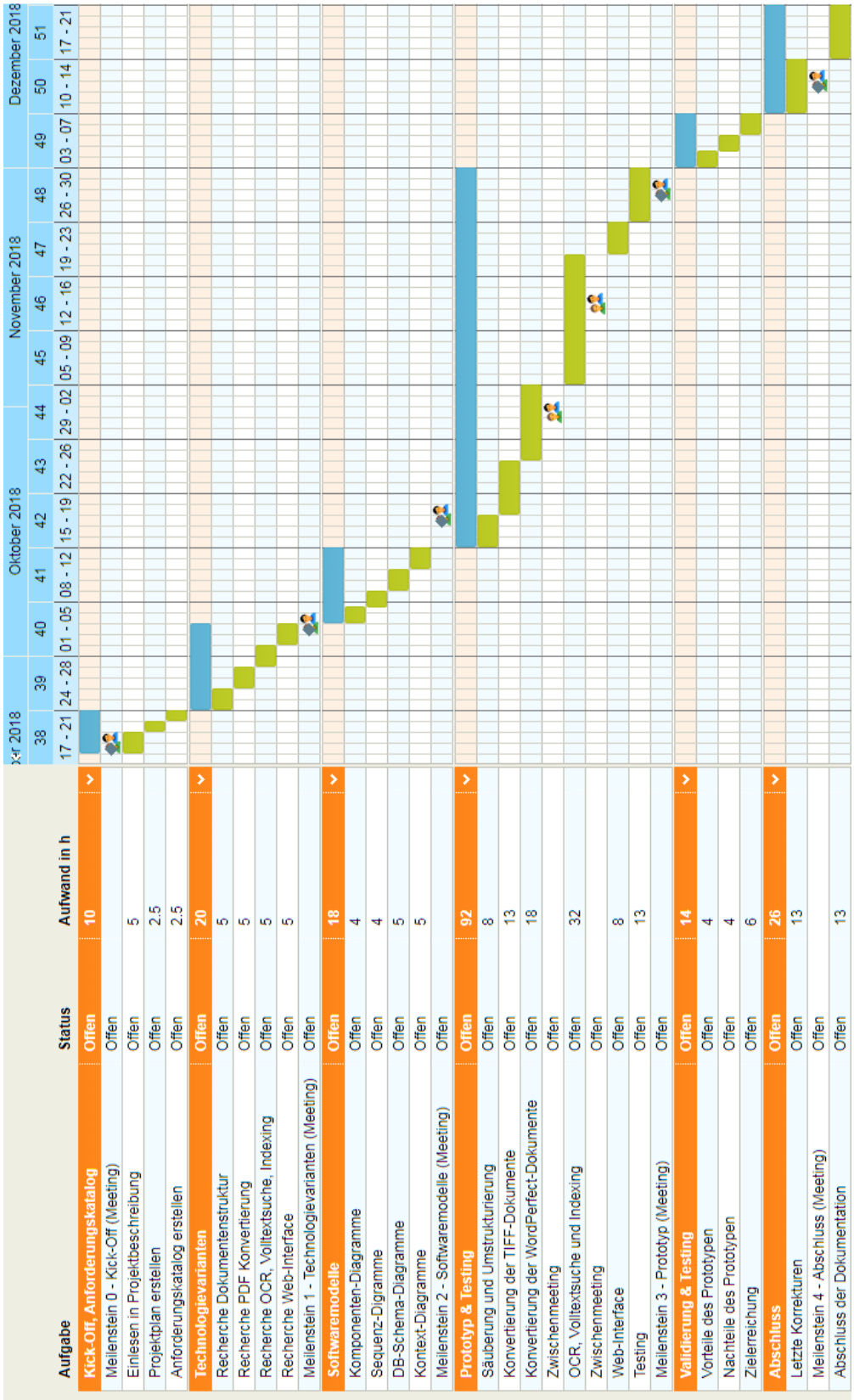
„Solr“ [29] ist eine Open-Source Softwarelösung von „Apache Software Foundation“ welches auf „Lucene“ aufbaut. „Solr“ läuft als separater Server und wird mittels Schnittstelle (Web-API) angesprochen. Dabei lässt sich Solr sehr tief konfigurieren, ohne eine Zeile Code programmieren zu müssen. Auch die Indexierung übernimmt „Solr“. Zu den grossen Vorteilen gehört die einfache Implementierung und die einfach zu verwendenden Schnittstellen. Zu den Nachteilen zählt der „Overhead“ der dadurch entsteht, dass die Suchbegriffe und Suchergebnisse über eine Schnittstelle ausgetauscht werden müssen.

Elasticsearch

„Elasticsearch“ [30] ist eine Open-Source Softwarelösung von „elastic“, welches ebenfalls auf „Lucene“ aufbaut und sich nicht gross in den Eigenschaften von „Solr“ unterscheidet. Der grösse Unterschied zu „Solr“ besteht darin, dass sich „Elasticsearch“ mehr Funktionalitäten im Bereich der Analysen bietet, während sich „Solr“ stärker auf die Textsuche konzentriert. Ausserdem ist bei „Elasticsearch“ die Community nicht so offengehalten wie bei „Solr“ [31][32].

3 Lösungsdesign

3.1 Projektplanung



3.2 Anforderungskatalog

Anforderung „Säuberung und Umstrukturierung“ (KKL_A01)

Nr. / ID	KKL_A01	**Titel**	Säuberung und Umstrukturierung

Ist Zustand

Der erste Teil der Dokumente (grösstenteils TIFF-Format) sind derzeit in einer Ordnerstruktur nach „Jahr“, „Monat“ und „Tag“ abgelegt. Diese Dokumente haben ausserdem einen zufälligen Dateinamen. In dieser Struktur existieren auch viele leere Ordner.

Der zweite Teil der Dokumente (grösstenteils WordPerfect-Format) sind derzeit in einer Ordnerstruktur nach „Jahr“ unterteilt. Teilweise wird hier noch weiter in Unterordner mit 8 Zeichen in beschreibender Form gegliedert. Die Dateinamen wurden ebenfalls mit 8 Zeichen kryptischen versehen und die Dateiendungen wurden als Autoren sowie Versionsbeschreibung verwendet (z.B. A03).

Soll Zustand

Alle Datei-Dubletten in der vorhandenen Dateistruktur sollen in der neuen Datenstruktur ignoriert werden. Als Dubletten werden Dateien bezeichnet, die den gleichen Namen sowie die gleiche Dateigrösse besitzt. Diese sollen ebenfalls bereinigt werden. Eine Dublette in der „Doks“-Struktur (TIFF-Format) hat gegenüber einer Dublette in der „Zwischenordner“-Struktur (WordPerfect-Format) Vorrang.

Bei allen WordPerfect-Dateien, bei denen mehrere Versionen existieren, soll nur die neuste Version berücksichtigt werden (wenn das Autorenkürzel sowie der Dateiname übereinstimmt).

Leere Ordner sollen in der neuen Datenstruktur ignoriert werden.

Grobschätzung des Aufwands

Da die Umstrukturierung grösstenteils automatisch ablaufen kann (z.B. mit Hilfetools wie „Bulk Rename Utility“) wurde der Aufwand auf 8 Stunden geschätzt.

Risiken

Risiko	**Beschreibung**	**Eintrittswahrscheinlichkeit**	**Massnahme**
Erstellungsdatum	Das Erstellungsdatum in den Metadaten stimmt nicht mit dem der Ordnerstruktur überein.	Mittel	Es wird das Erstellungsdatums der Ordnerstruktur übernommen.
Dateiendungen	Da die Dateiendungen vieler Dokumente angepasst wurden, kann nicht direkt festgestellt werden um welchen Dateityp es sich handelt.	Hoch	Mit Hilfe der Dateisignatur feststellen um welchen Dateityp es sich handelt (z.B. TrID).

Anforderung „WordPerfect Konvertierung ins PDF/A Format“ (KKL_A02)

Nr. / ID	KKL_A02	**Titel**	WordPerfect Konvertierung ins PDF/A Format

Ist Zustand

Zirka 42% aller Dokumente der Synopsis-Datenstruktur existiert im Corel WordPerfect-Format.

Soll Zustand

Alle WordPerfect Dokumente sollen in den PDF/A Standard konvertiert werden. Dabei ist es wichtig, dass die Formatierung bei den WordPerfect Dokumenten möglichst mit der des Originalformats übereinstimmt. Der Text des Dokuments soll möglichst identisch zum Originaltext als Textebene im PDF verfügbar sein.

Grobschätzung des Aufwands

Der Aufwand für die WordPerfect Dokumente mit Formatierung wurde auf 18 Stunden geschätzt.

Risiken

Risiko	**Beschreibung**	**Eintrittswahrscheinlichkeit**	**Massnahme**
WordPerfect Formatierung	Die Formatierung der WordPerfect Dokumente geht bei der Konvertierung ins PDF/A Format verloren.	Mittel	Betroffene Dokumente mit der Originalsoftware „Corel WordPerfect Suite 8“ öffnen und als PDF-Datei drucken.

Anforderung „TIFF Konvertierung ins PDF/A Format“ (KKL_A03)

Nr. / ID	KKL_A03	**Titel**	TIFF Konvertierung ins PDF/A Format

Ist Zustand

Zirka 38% aller Dokumente der Synopsis-Datenstruktur existiert im TIFF-Format. Dies sind eingescannte Dokumente, die teilweise eine schlechte Bildqualität besitzen.

Soll Zustand

Alle eingescannten Dokumente (TIFF-Format) sollen in den PDF/A Standard konvertiert werden. Der Text des Dokuments soll möglichst identisch zum Originaltext als Textebene im PDF verfügbar sein.

Grobschätzung des Aufwands

Da die TIFF-Formate keine Formatierung besitzen, jedoch eine schlechte Bildqualität, wurde der Aufwand auf 13 Stunden geschätzt.

Risiken

Risiko	**Beschreibung**	**Eintrittswahrscheinlichkeit**	**Massnahme**
TIFF-Bildqualität	Der Text der TIFF Dateien kann auf Grund der Bildqualität nicht oder nur schlecht erkannt werden.	Mittel	TIFF Dateien mit nicht erkennbarem Text werden ohne Texterkennung in die neue Datenstruktur aufgenommen.

Anforderung „Dokumentensuche“ (KKL_A04)

Nr. / ID	KKL_A04	Titel	Dokumentensuche

Ist Zustand

Neben der Dokumentenstruktur mit den Ordnern existiert eine Access 2016 / 365 Datenbank welche alle Dokumentennamen mit deren Pfade, eine Beschreibung, das Erstellungsdatum sowie Schlagwörter enthält. Über diese Datenbank wurde bisher mit den Schlagwörtern nach den Dokumenten gesucht.

Soll Zustand

Das Erstellungsdatum in den Eigenschaften der jeweiligen Datei soll in die neue Datenstruktur aufgenommen werden. Falls die Datei in einem Ordner der „Jahr, Monat, Tag“ Struktur liegt, so soll das Datum der Ordnerstruktur als Erstelldatum genommen werden. Falls die Datei in einem „Zwischenordner“ liegt, so soll jeder Ordner oberhalb der Datei als Schlagwort zum Dokument erfasst werden. Ausserdem soll der Original Dateiname, der Autor (über die Dateiendung), die Version (über die Dateiendung), die Beschreibung (über die vorhandene Access Datenbank) in die Datenbank übernommen werden.

Alle TIFF-Dateien sollen zudem als Bild gekennzeichnet sein.

Für alle Dokumente in der neuen Datenstruktur soll eine Möglichkeit bestehen zu wählen, ob das Dokument in der Dokumentensuche berücksichtigt werden soll oder nicht.

Ausserdem soll eine Volltextsuche für die in der Datenstruktur liegenden Dokumente (mittels Indexierung) ermöglicht werden.

Grobschätzung des Aufwands

Mit 32 Stunden wurde der Aufwand für diese Anforderung am höchsten geschätzt. Grund dafür ist, dass es sich um die Kernaufgabe dieses Wirtschaftsprojekts handelt.

Risiken

Risiko	Beschreibung	Eintrittswahrscheinlichkeit	Massnahme
TIFF-Bildqualität	Der Text der TIFF Dateien kann auf Grund der Bildqualität nicht oder nur schlecht erkannt werden.	Mittel	TIFF Dateien mit nicht erkennbarem Text werden ohne Texterkennung in die neue Datenstruktur aufgenommen.
TIFF-Handgeschriebenes	Handgeschriebene Texte können nicht gelesen werden.	Hoch	Handgeschriebene Texte werden ignoriert.

Anforderung „Web-Oberfläche“ (KKL_A05)

Nr. / ID	KKL_A05	**Titel**	Web-Oberfläche

Ist Zustand

Derzeit wird die Access 2016 / 365 Datenbank offline verwendet, um nach Dokumenten zu suchen.

Soll Zustand

Um in Zukunft die Dokumente auch übers Internet bereitstellen zu können, soll eine Web-Oberfläche bereitgestellt werden über welche angemeldet und berechtigte Benutzer nach den Dokumenten suchen und diese öffnen können. Es soll nach Dateinamen, Schlagwörtern, Erstellungsdatum und Volltext gesucht werden können. Es soll die Möglichkeit bestehen neue Benutzer anzulegen.

Grobschätzung des Aufwands

Der Aufwand der Anforderung „Web-Oberfläche“ wurde mit 8 Stunden relativ tief geschätzt. Grund dafür ist, dass es im Internet bereits eine Vielzahl von zur Verwendung verfügbaren Web-Oberflächen gibt.

Risiken

Risiko	**Beschreibung**	**Eintrittswahrscheinlichkeit**	**Massnahme**
Keine passende Web-Oberfläche verfügbar	Es wird keine passende Web-Oberfläche im Internet gefunden.	Niedrig	Es wird selbst eine Web-Oberfläche entwickelt.

3.3 Langzeitarchivierung

Da einer der wichtigsten Gründe für die Durchführung dieses Projekts die Langzeitarchivierung der vorhandenen Daten ist, musste in diesem Projekt ein besonderes Augenmerk auf diesen Aspekt gelegt werden.

Da digitale Daten mit der Zeit immer schwerer lesbar werden (werden diese nicht gepflegt), befasst sich die Langzeitarchivierung mit der Frage wie diese Daten für die Nachwelt möglichst gut erhalten bleiben. Neben den physikalischen Einwirkungen auf die Haltbarkeit gibt es auch einige weitere Faktoren:

- Dateien wurden mit Programmen erstellt, die heute nicht mehr verfügbar sind.
- Dateien wurden mit Programmen erstellt, die heute nur noch in Systemumgebungen von damals gelesen werden können.
- Dateien können zu keinem Programm mehr zugeordnet werden. Dies ist in diesem Projekt ein akutes Problem, da die Dateiendungen mit der man die Dateien meist den Programmen zuordnen kann, verwendet wurden um den Autor mit einem Kürzel sowie der Version der Datei zu beschreiben.
- Die Original Metadaten der Dateien können beim Öffnen und erneutem Speichern („Änderungsdatum“), durch Kopieren („Erstellungsdatum“) oder durch Umbenennung („Dateinamen“, „Autorenkürzel“, „Version“), Löschen, Verschiebung („Erstellungsdatum aus Ordnerstruktur“, „Schlüsselwörter“, „Originalpfad“) oder Überschreiben verloren gehen.
- Die Dateien selbst können mit der Zeit beschädigt werden.
- Verbindungen zwischen Dateien und zusätzlichen Daten (z.B. Access Datenbank mit „Beschreibung“) können verloren gehen.

3.3.1 Sicherung der Lesbarkeit

Um die Lesbarkeit der Dateien in der Synopsis Dateistruktur zu sichern entschied man sich für das PDF/A Format (siehe Punkt 3.6.1). Dies ist ein von der „International Origanization for Standardization“ (ISO) genormtes Dateiformat, welches festlegt welche Elemente der zugrundeliegenden PDF-Versionen im Hinblick auf die Langzeitarchivierung verwendet werden müssen. Die in diesem Projekt behandelten Dateien beschränken sich auf die Dateien im „WordPerfect“ Format, sowie die Dateien im „TIFF“ Format. Diese werden ins PDF/A Format konvertiert (siehe Punkt 3.6.2) womit nur noch ein PDF Lese Programm benötigt wird, um der Inhalt der neuen Datenstruktur zu lesen.

3.3.2 Sicherung der Dateien mit deren Metadaten

Um die Dateien mit deren Metadaten zu sichern entschied man sich für eine Speicherung in einer Datenbank. Dies hat folgende Vorteile:

- Das „Erstellungsdatum“ bleibt beim erneuten Speichern erhalten, da jeweils nur eine Kopie der Datei aus der Datenbank geladen wird und nicht die Originaldatei (diese Metadaten werden neben der Originaldatei separat in der Datenbank gespeichert).
- Der „Dateinamen“, der „Autor“ und die „Version“ bleiben bei einer Umbenennung der Datei erhalten, da jeweils nur eine Kopie der Datei aus der Datenbank geladen wird und nicht die Originaldatei (diese Metadaten werden neben der Originaldatei separat in der Datenbank gespeichert).
- Das „Erstellungsdatum aus Ordnerstruktur“, die „Schlüsselwörter“ sowie der „Originalpfad“ bleiben durch eine Verschiebung der Datei erhalten, da jeweils nur eine Kopie der Datei aus der Datenbank geladen wird und nicht die Originaldatei (diese Metadaten werden neben der Originaldatei separat in der Datenbank gespeichert).

Die Verwendung einer Datenbank birgt aber auch Nachteile:

- Alle Dateien an einem zentralen Ort, was bei einer Beschädigung der Datenbank zu einem Verlust aller Daten führen kann (Abhilfe hierfür schaffen regelmässige Datenbanksicherungen).
- Die Dateien können nicht mehr direkt (z.B. über den Windows Explorer) geöffnet werden, sondern müssen erst aus der Datenbank geladen werden.
- Wird die Datenbank nicht gepflegt, kann es dazu kommen, dass diese mit der Zeit veraltet ist und nur noch mit alten Datenbankprogrammen darauf zugegriffen werden kann.

3.3.3 Fazit

Das PDF/A Format ist ein bewährtes, ISO genormtes Format für die Langzeitarchivierung und es sind derzeit unzählige PDF Lese Programme vorhanden. Aus diesem Grund entschied man sich für das PDF/A Format.

Das grösste Risiko bei der Datenbank ist die Beschädigung bei einem zentrale Datenspeicher. Dieses Risiko kann jedoch durch regelmässige Datenbanksicherungen minimiert werden. Die Vorteile überwiegen hier klar die Nachteile, weshalb in diesem Projekt die Datenspeicherung mit deren Metadaten über eine Datenbank für die Langzeitarchivierung gewählt wurde.

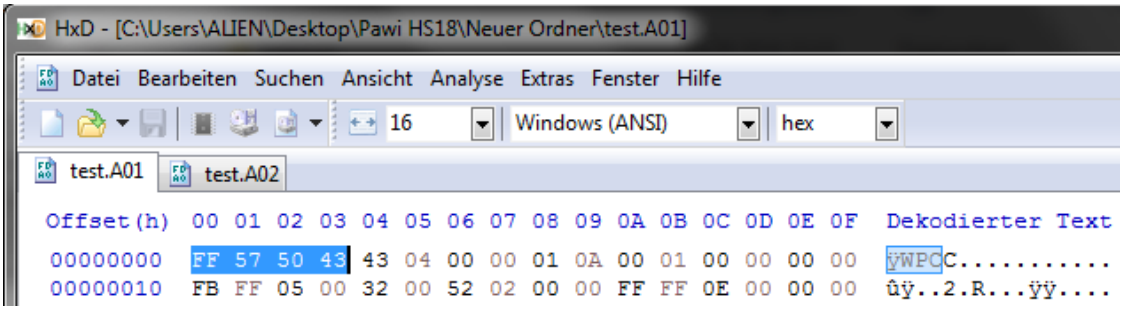
3.4 Dateitypbestimmung

Die Dateiendungen in der bisherigen Synopsis Dateistruktur wurden meist verwendet, um Autorenkürzel sowie Dokumentversion zu hinterlegen. Deshalb konnte nicht mehr direkt nachvollzogen werden, um welches Format es sich bei den jeweiligen Dateien handelte. Des Weiteren konnte nicht direkt bestimmt werden mit welchen Programmen diese erstellt wurden bzw. geöffnet werden können. Deshalb war eine Dateitypanalyse Voraussetzung für die weitere Arbeit am Projekt.

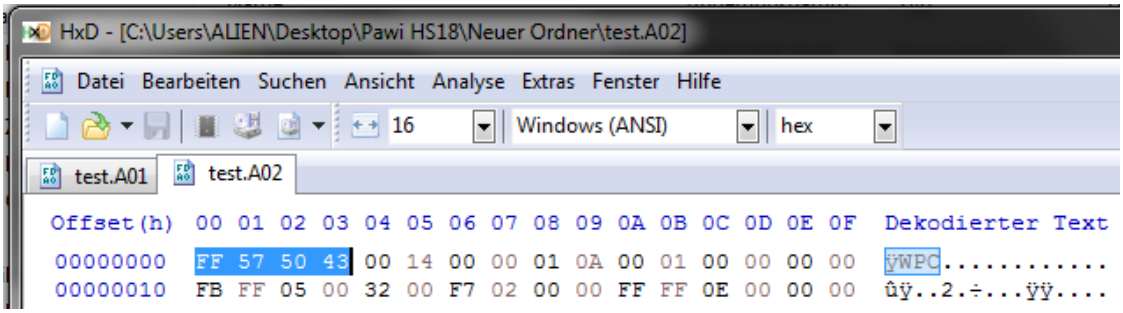
3.4.1 Dateikopf (Manuelle Analyse)

Da in der bisherigen Synopsis Dateistruktur vor allem WordPerfect und TIF/TIFF Dateien liegen, wurde versucht mittels Dateikopf die WordPerfect- sowie TIF/TIFF Dateien der Synopsis Dateistruktur zu bestimmen. Da sich der Dateikopf immer zu Beginn einer Datei (teilweise mit einem Offset von einigen Bytes) befindet, wurden zuerst mehrere WordPerfect-Dateien der Synopsis Dateistruktur mit Hilfe des Hex Editors „HxD v2.1“ [1] gelesen und die ersten Bytes auf Übereinstimmungen überprüft [2][3]. Dabei konnte folgendes Muster erkannt werden:

Die ersten 4 Bytes eines WordPerfect Dokuments in der bisherigen Synopsis Datenstruktur sehen in HEX-Format wie folgt aus:

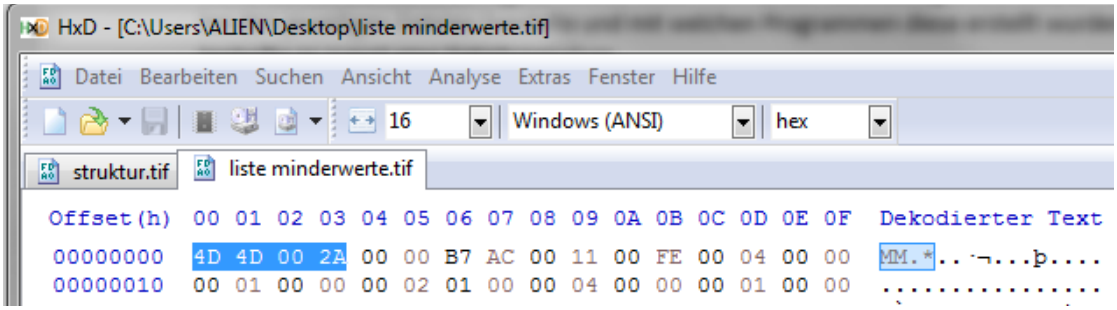


Ein Vergleich mit einer anderen WordPerfect Datei in der Synopsis Datenstruktur zeigt die Übereinstimmung der ersten 4 Bytes:

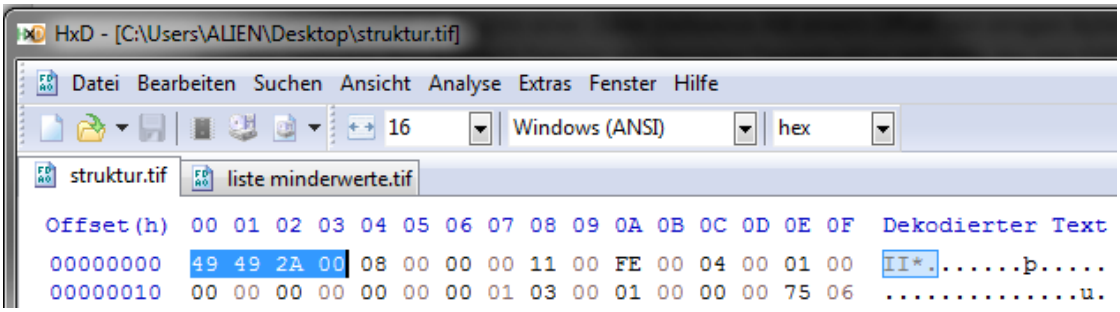


Dasselbe wurde für die TIF/TIFF-Dateien der bisherigen Synopsis Datenstruktur gemacht und die ersten Bytes auf Übereinstimmungen überprüft. Dabei konnte folgendes Muster erkannt werden:

Die ersten 4 Bytes eines TIF/TIFF Dokuments in der bisherigen Synopsis Datenstruktur sehen in HEX-Format entweder wie folgt aus:



oder so:

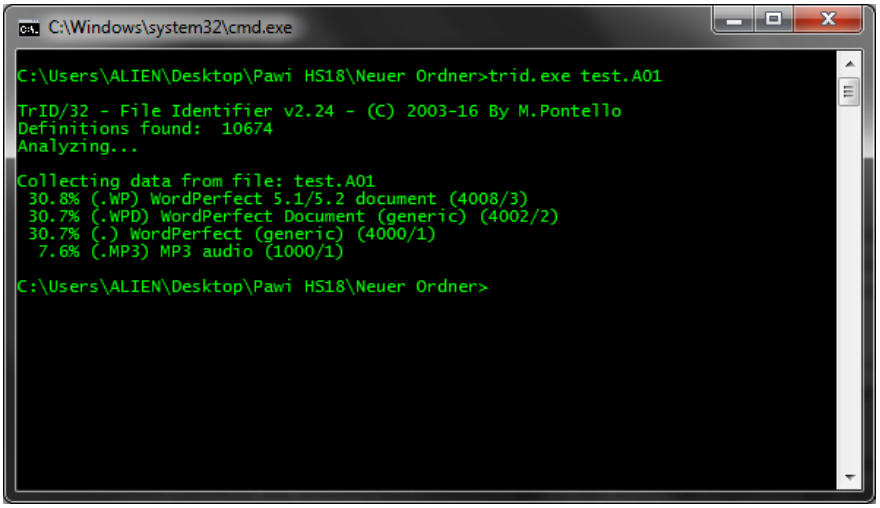


Mithilfe dieses Musters und einer eigens dazu entwickelten Anwendung in der Programmiersprache C# konnten alle Dateien der Synopsis Datenstruktur ausgegeben werden, welche den Dateityp WordPerfect oder TIF/TIFF hatten. Insgesamt konnten so 22‘435 WordPerfect- und 20‘359 TIFF-Dateien in der bisherigen Synopsis Datenstruktur identifiziert werden.

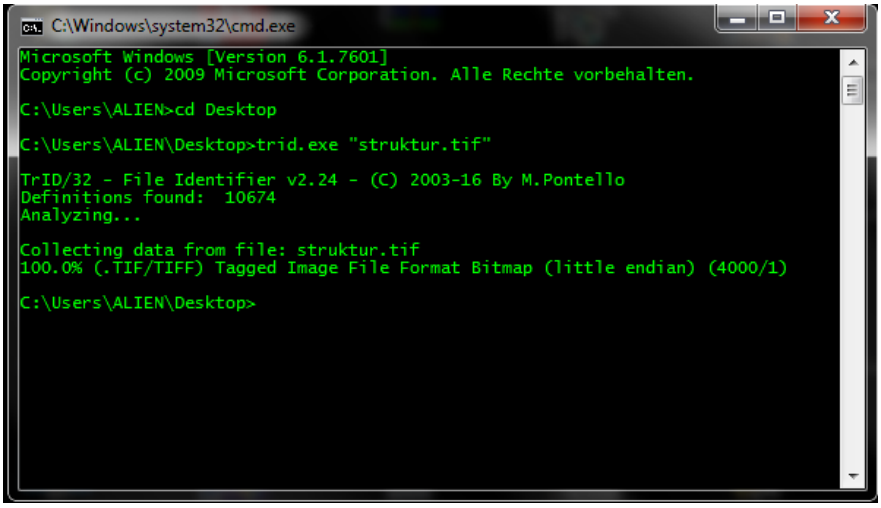
Mit Hilfe der Webseite „Files Signatures“ [3] kann man die Dateisignatur im HEX-Format über eine Datenbank den möglichen Programmen zuordnen.

3.4.2 Dateisignatur (Automatische Analyse)

Eine weitere Möglichkeit die WordPerfect- sowie TIFF-Dateien der bisherigen Synopsis Datenstruktur zu identifizieren bietet die Freeware „TrID v2.24“ [4] die einzelnen Dateiendungen auf mögliche Dateitypen und mögliche Programme geprüft. Ein Beispiel für die Dateisignaturprüfung eines WordPerfect Dokuments mittels „TrID“:



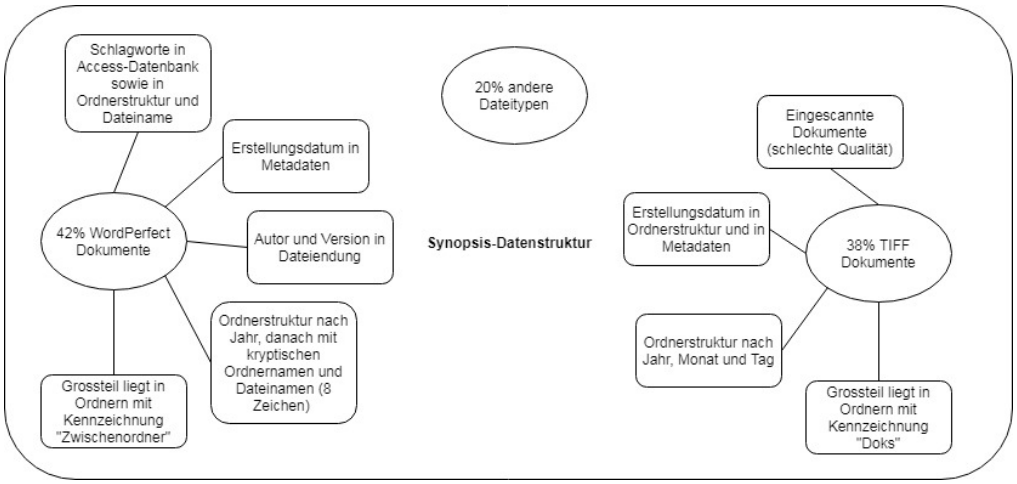
Ein Beispiel für die Dateisignaturprüfung einer TIF/TIFF Datei mittels „TrID“:



Leider kann hier nicht immer 100% davon ausgegangen werden, dass die von „TrID“ aufgelisteten Programme und Dateiendungen passen.

3.5 Datenbereinigung

Da die Metadaten der bisherige Synopsis Datenstruktur in verschiedenen Systemen verstreut liegen (Dateien und Schlüsselwörter in Ordnerstruktur, Beschreibungen und Erstellungsdatum in Access Datenbank, sowie ein weiteres Erstellungsdatum in den Dateieigenschaften), darin Dubletten existieren und für die neue Datenstruktur nur die neuste Version der verschiedenen WordPerfect Dateien relevant sind, musste zuerst eine Bereinigung durchgeführt werden (siehe Anforderungskatalog „KKL_A01“).



3.5.1 Auslesen der Metadaten

Die Dateien in der lokalen Ordnerstruktur enthalten als relevante Metadaten in den Eigenschaften das „Erstellungsdatum“, die „Dateigrösse“, den „Dateiname“ sowie den „Dateipfad“.

Bei den Dateien in den „Doks“-Ordnern (grösstenteils TIFF-Dateien) gibt es zudem ein weiteres „Erstellungsdatum“ als Teil der Ordnerstruktur (Dateien sind nach Ordnern „Jahr“, „Monat“ und „Tag“ abgelegt). In der vorhandenen Access Datenbank sind zudem die Metadaten „Beschreibung“ für die Dateien in den „Doks“-Ordnern vorhanden.

Die Dateien in den „Zwischenordner“-Ordnern (grösstenteils WordPerfect-Dateien) enthalten zudem den „Autorenkürzel“ und die „Version“ in der Dateiendung. Ebenfalls enthält hier die Ordnerstruktur die dazugehörigen „Schlagworte“ (jeder Ordnername sowie Dateiname als Schlagwort).

3.5.2 Auslassung von alten Versionen

Bei den Dateien in den „Zwischenordner“-Ordnern kann aus der Dateiendung neben dem Autor auch die Version ausgelesen werden. Falls die Dateiendung mit einem Buchstaben anfängt und darauf zwei Zahlen folgen, so kann davon ausgegangen werden, dass es sich bei den beiden Zahlen um die Version der Datei handelt (sowie der Buchstabe als „Autorenkürzel“). Hat eine Datei den gleichen Dateinamen wie eine andere und sie unterscheiden sich in der Version, so wird nur die neuere der beiden Dateien für die neue Datenstruktur berücksichtigt.

3.5.3 Entfernen von Dubletten

Falls eine Datei in der alten Datenstruktur exakt den gleichen Dateinamen sowie Dateigrösse besitzt, kann diese als Dublette gekennzeichnet werden. Die Dubletten werden entfernt, dabei ist jedoch zu beachten, dass die Metadaten beider Dateien in der bestehenbleibenden Datei zusammengefasst werden. Dateien in den „Doks“-Ordnern haben immer Vorrang vor den Dateien in den „Zwischenordner“-Ordnern. Kommen beide Dateien von den „Doks“-Ordnern oder von den „Zwischenordner“-Ordnern, so erhält die zuerst ausgelesene Datei den Vorrang.

3.5.4 Auslassung von Dateien ohne Inhalt

Dateien, welche keinen Inhalt haben, können ausgelassen werden.

3.5.5 Auslassung von Dateien und Verzeichnissen

Folgende Dateien konnten in Absprache mit dem Auftragsgeber auf Grund von irrelevanten Inhalten ausgelassen werden:

1991 zwischenordner wp	ADRESSE, BERIALT, INPUTS		
1992 zwischenordner wp	ADRESSEN		
1993 zwischenordner wp	ADMINIST	ADRESSEN, AGENDA, INTERN, KONTO,SEECLUB, STUNDEN, sowie die restlichen 8 Files im Ordner ADMINIST	
1994 zwischenordner	ADMINIST	ABLAGE, ADRESSEN, AGENDA, INTERN, KONTO, RASTER, SEECLUB	
1995 zwischenordner wp	ADMINIST	ABLAGE, ADRESSEN, AGENDA, INTERN, KONTO, RASTER, SEECLUB	
1995 zwischenordner 2 disk	ADMINIST	ABLAGE, ADRESSEN, AGENDA, FAXSACHE, INTERN, KONTO, RASTER, SEECLUB, TERMINE	
1996 zwischenordner	ADMINIST	ABLAGE, ADRESSEN, FAXSACHE, INTERN, KONTO, MUSTER, RASTER, SEECLUB	
1997 zwischenordner wp	ADMINIST	ABLAGE, ADRESSEN, AN-OKB, FAXSACHE, INTERN, KONTO, LIEFERSC, MUSTER, PCBETREU, RASTER, SEECLUB	
1998 zwischenordner wp	ADMINIST		
1999 zwischenordner wp inkl bh	L99 plus	Administ	Ablage, Adressen, Intern, Liefersch, Quittungen, Raster, Rechnung, Seeclub, Termine

3.5.6 Spezialfälle

Um die Automatisierung der Bereinigung (siehe Punkt 3.5) möglichst einfach und reibungslos durchzuführen, mussten noch einige Anpassungen an der alten Synopsis Datenstruktur gemacht werden:

- Der Ordner „1991 zwischenord**nder** wp“ musste in „1991 zwischenord**ner** wp“ umbenannt werden.
- Der Ordner „Doks1997**11**_oktober“ musste in „Doks1997**10**_oktober“ umbenannt werden.

3.5.7 Automatisierung

Um die Metadaten automatisch aus der bisherigen Datenstruktur sowie der Access Datenbank auszulesen, die Texterkennung mittels OCR durchzuführen (siehe Punkt 3.7), die Dateien in durchsuchbare PDFs zu konvertieren (siehe Punkt 3.7.1) und diese in die neue Synopsis Datenstruktur (Datenbank) zu laden, wurde eine Anwendung in C# geschrieben. Der komplette Weg von einer WordPerfect-Datei in der bisherigen Synopsis Datenstruktur zu einem durchsuchbaren PDF inkl. Metadaten in der neuen Synopsis Datenbank dauert ca. 1 Sekunde. Bei den TIFF-Dateien sind dies ungefähr 10 bis 15 Sekunden. Dieser Zeitunterschied kommt daher, dass bei den WordPerfect-Dateien die Textebene im resultierenden PDF automatisch über die Druckfunktion im WordPerfect erstellt wird. Bei den TIFF-Dateien muss erst ein „Image-only PDF“ (siehe Punkt 2.1.1) erstellt werden, die Bilder davon extrahiert werden, die Bilder mittels Texterkennung (OCR) analysiert werden und zum Schluss die Bilder als durchsuchbares PDF zusammengefügt werden.

3.5.8 Erkennung von irrelevanten Dokumenten

Auf Wunsch des Auftraggebers wurden zusätzliche Möglichkeiten zur automatischen Erkennung von irrelevanten Dokumenten geprüft. Da nach der Konvertierung von WordPerfect- sowie TIFF-Dateien ins PDF-Format die inhaltlichen Texte des Dokuments für die Volltextsuche bekannt sind (werden mittels PDFBox.Net aus der Textebene ausgelesen), besteht hier eine Möglichkeit, um z.B. Dokumente mit wenig oder irrelevantem Inhalt automatisch auszusondern.

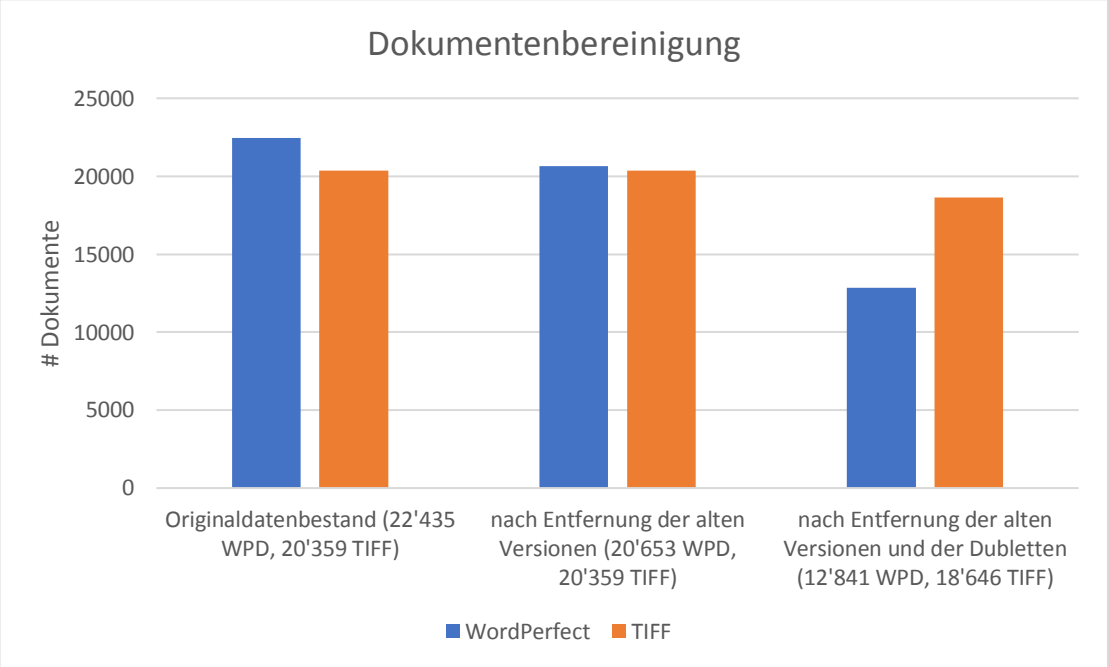
3.5.9 Problematische und spezielle Dateien

Bei der Konvertierung der Dateien in durchsuchbare PDFs (siehe Punkt 3.6) gab es folgende problematische Dateien:

- 35 Dateien konnten aufgrund von Fehlern nicht konvertiert werden.
- 119 Dateien konnten aufgrund eines Passwortschutzes nicht konvertiert werden.
- 63 Dateien wurden nicht konvertiert da sie keinen Inhalt hatten.
- 50 Dateien wurden als „reine“ (kein Text vorhanden) Bild-Dateien erkannt.
- 304 WordPerfect-Dateien enthalten im Inhalt ein Datumstextbaustein, der beim Öffnen das aktuelle Datum übernimmt. Um das daraus konvertierte durchsuchbare PDF nicht zu verfälschen werden diese manuell von Herrn Dr. Held korrigiert und später in der Synopsis Datenbank ersetzt.

3.5.10 Ergebnis

Insgesamt sind in der alten Synopsis Datenstruktur 52‘821 Dateien vorhanden. Durch die Reduzierung auf die WordPerfect- sowie TIFF-Dateien konnte dies auf 42‘794 Dateien (davon 22‘435 WordPerfect- und 20‘359 TIFF-Dateien) gesenkt werden. Nach dem Entfernen von alten Dateiversionen waren noch 41‘010 Dateien (20‘651 WordPerfect- und 20‘359 TIFF-Dateien) relevant. Das Entfernen von Dubletten minimierte die Anzahl auf 31‘486 Dateien (12‘840 WordPerfect- und 18‘646 TIFF-Dateien) welche zu durchsuchbaren PDFs konvertiert, deren Metadaten ausgelesen und in die Datenbank eingefügt werden konnten. Dabei sind die problematischen Dateien (Punkt 5.7.1) sowie die ignorierten Ordner und Dateien bereits inbegriffen.



3.6 Konvertierung

3.6.1 PDF Typen, Versionen und Standards

Da die PDF Version 2.0 erst seit 2017 verfügbar ist, wird dies von den meisten PDF Konvertern gar noch nicht unterstützt. Ausserdem basieren alle derzeitigen Standards auf der Version 1.7. Deshalb wurde mit dem Auftraggeber die PDF Version 1.7 als Ausgabeformat der WordPerfect- sowie TIFF-Dateien gewählt.

Da der PDF Standard „PDF/A-1“ keine „Vereinfachung der Durchsuchbarkeit von Texten in Unicode-Format“ enthält, wurde der Standard „PDF/A-2u“ als Ziel gesetzt. Der Standard PDF/A-3 wird nicht benötigt, da keine PDFs erstellt werden, welche nicht PDF/A konforme Dokumente einbetten (Container).

Aus diesem Grund wurde das Ziel der PDF-Konvertierung auf dem PDF Standard „PDF/A-2u“ mit der PDF Version 1.7 festgelegt.

3.6.2 PDF/A Konvertierung – WordPerfect und TIFF

Bewertung

Bei den Softwarelösungen wurde ein Augenmerk auf die Eigenschaften „Preis“ (Wie teuer ist die Lösung?), „Automatisierung“ (Wie gut lässt sich die Lösung automatisieren?), „Ergebnis“ (Gibt es grosse Unterschiede zum Originaldokument?), „Unterstützte PDF Versionen und Standards“ (Welche PDF-Versionen und Standards werden unterstützt?) und „Textebene vorhanden“ (Ist beim erstellten PDF des WordPerfect-Dokuments eine Textebene vorhanden?) gelegt.

Kriterium	0	1	2	3
Preis	Mehr als CHF 100.00	Von CHF 50.00 bis 100.00	Von CHF 0.00 bis 50.00	Gratis (Freeware)
Automatisierung	Kein Batch-Job möglich	Batch-Job über Anwendung	Batch-Job über Druckfunktion möglich	Batch-Job per Programmiersprache oder Command Line möglich
Ergebnis	Mehr als 3 Unterschiede zum Originaldokument	2-3 Unterschiede zum Originaldokument	1 Unterschied zum Originaldokument	Keine Unterschiede zum Originaldokument
Unterstützte PDF Versionen und Standards	Kein PDF/A Standard	Min. 1 PDF/A Standard und min. 1 Version	Min. 2 PDF/A Standards und min. 1 Version	Über 2 PDF/A-Standards und min. 2 Versionen
Textebene vorhanden	Keine Textebene vorhanden	-	-	Textebene vorhanden

OfficeToPDF v1.8 [10]

Preis	Automatisierung	Ergebnis	Unterstützte PDF Versionen und Standards	Textebene vorhanden
Open-Source	Hoch – Command Line	Abstände von Nummern zu Text nicht vorhanden (Seite 11)	PDF 1.7, PDF/A-3a	Ja

FoxPDF WordPerfect to PDF Converter v3.0 [11]

Preis	Automatisierung	Ergebnis	Unterstützte PDF Versionen und Standards	Textebene vorhanden
$29.90 (CHF 30.00)	Tief – In Anwendung	Andere Schriftart, Seiten stimmen nicht, Fusszeilen stimmen nicht	PDF 1.4, PDF/A-1a (invalid)	Ja

ABCpdf .NET Converter (Standard Single Licence) v11.2.2 [12]

Preis	Automatisierung	Ergebnis	Unterstützte PDF Versionen und Standards	Textebene vorhanden
$329.00 (CHF 323.00)	Hoch – Visual Studio (C#)	Andere Schriftart, Seiten stimmen nicht, Fehlende Nummern (Seite 14), Fusszeilen stimmen nicht	PDF 1.5, PDF/A-1a, PDF/A-1b, PDF/A-2a, PDF/A-2b, PDF/A-2u	Ja

Corel WordPerfect X9 – Publish to PDF (Home & Student Edition) [13]

Preis	Automatisierung	Ergebnis	Unterstützte PDF Versionen und Standards	Textebene vorhanden
CHF 200.95	Hoch – Visual Studio (C#)	Perfekt	PDF 1.2, PDF 1.3, PDF 1.4, PDF/A-1a (invalid), PDF/A-1b	Ja

PDF-XChange v7.0 build 326.1 (Standard) [14]

Preis	Automatisierung	Ergebnis	Unterstützte PDF Versionen und Standards	Textebene vorhanden
€37.00 (CHF 43.00)	Mittel – Über Druckfunktion	Perfekt	PDF 1.3, PDF 1.4, PDF 1.5, PDF 1.6, PDF 1.7, PDF 2.0, PDF/A-1a, PDF/A-1b, PDF/A-2a, PDF/A-2b, PDF/A-2u, PDF/A-3a, PDF/A-3b, PDF/A-3u	Ja

PDF24 Creator v8.6.1 (Private) [15]

Preis	Automatisierung	Ergebnis	Unterstützte PDF Versionen und Standards	Textebene vorhanden
Freeware	Mittel – Über Druckfunktion	Seiten stimmen nicht	PDF 1.2, PDF 1.3, PDF 1.4, PDF 1.5, PDF 1.6, PDF 1.7, PDF/X-3, PDF/A-1, PDF/A-2, PDF/A-3	Ja

Ergebnis

Software	Preis (0-3)	Automatisierung (0-3)	Ergebnis (0-3)	PDF Versionen + Standards (0-3)	Textebene (0, 3)	Gesamt (0-15)
OfficeToPDF	3	3	2	1	3	12
FoxPDF	2	1	1	1	3	8
ABCpdf	0	3	0	2	3	8
Corel WordPerfect	0	3	3	2	3	11
PDF-XChange	2	2	3	3	3	13
PDF24 Creator	3	2	2	3	3	13

Fazit

Im Vergleich wurden jeweils pro Kriterium 0-3 Punkte vergeben. Am Ende wurden alle Punkte zusammengezählt. Am besten schnitten die Lösungen „PDF-XChange“ sowie „PDF24 Creator“ mit jeweils 10 von 12 möglichen Punkten ab. Da bei der Lösung „PDF-XChange“ jedoch keine Unterschiede zum Originaldokument erkennen liessen und der Preis mit umgerechnet CHF 43.- günstig ist, wurde diese Lösung zur PDF Konvertierung gewählt, wobei die schwierigere Automatisierung über die Druckfunktion in Kauf genommen wurde.

3.7 Texterkennung (OCR)

Bewertung

Bei den Softwarelösungen wurde ein Augenmerk auf die Eigenschaften „Preis“ (Wie teuer ist die Lösung?), „Automatisierung“ (Wie gut lässt sich die Lösung automatisieren?) und „Ergebnis“ (Gibt es grosse Unterschiede zum Originaldokument?) gelegt.

Kriterium	0	1	2	3
Preis	Mehr als CHF 100.00	Von CHF 50.00 bis 100.00	Von CHF 0.00 bis 50.00	Gratis (Freeware)
Automatisierung	Kein Batch-Job möglich	Batch-Job über Anwendung	Batch-Job über Druckfunktion möglich	Batch-Job per Programmiersprache oder Command Line möglich
Ergebnis	Mehr als 3 Unterschiede zum Original-dokument	2-3 Unterschiede zum Originaldokument	1 Unterschied zum Originaldokument	Keine Unterschiede zum Originaldokument

Tesseract-OCR [16]

Preis	Automatisierung	Ergebnis
Open-Source	Hoch – Visual Studio (C#)	Probleme bei zu kleinen Texten Probleme bei schlechter Auflösung

IronOcr v4.4.0 [18]

Preis	Automatisierung	Ergebnis TIFF
Freeware	Hoch – Visual Studio (C#)	Probleme bei kleinen Textstellen Probleme bei schlechter Auflösung Probleme bei Umlauten

Ergebnis

Software	Preis (0-3)	Automatisierung (0-3)	Ergebnis (0-3)	Gesamt (0-9)
Tesseract-OCR	3	3	1	7
IronOcr	3	3	1	7

Fazit

„Tesseract-OCR“ sowie „IronOcr“ schneiden im Vergleich gleich gut. Jedoch hatte „IronOcr“ in den Tests zusätzlich Problem mit den Umlauten. Aus diesem Grund, sowie der grösseren Community bei „Tesseract-OCR“ entschied man sich für „Tesseract-OCR“.

3.7.1 „Image-only PDF“ zu „Searchable PDF“

Bewertung

Bei den Softwarelösungen wurde ein Augenmerk auf die Eigenschaften „Preis“ (Wie teuer ist die Lösung?), „Automatisierung“ (Wie gut lässt sich die Lösung automatisieren?) und „Ergebnis“ (Gibt es grosse Unterschiede zum Originaldokument?) gelegt.

Kriterium	0	1	2	3
Preis	Mehr als CHF 100.00	Von CHF 50.00 bis 100.00	Von CHF 0.00 bis 50.00	Gratis (Freeware)
Automatisierung	Kein Batch-Job möglich	Batch-Job über Anwendung	Batch-Job über Druckfunktion möglich	Batch-Job per Programmiersprache oder Command Line möglich
Ergebnis	Mehr als 3 Unterschiede zum Original-dokument	2-3 Unterschiede zum Originaldokument	1 Unterschied zum Originaldokument	Keine Unterschiede zum Originaldokument

Acrobat Pro DC

Preis	Automatisierung	Ergebnis
$15.00 (15.00 CHF) / Monat	Tief – In Anwendung	Probleme bei schlechter Auflösung

PDF-XChange v7.0 build 326.1 + Ghostscript v9.24 + iTextSharp v2.0.50727 + Tesseract-OCR v3.05.02 + hOcr2Pdf.Net v1.0

Preis	Automatisierung	Ergebnis
PDF-XChange: €37.00 (CHF 43.00) Ghostscript: Freeware iTextSharp: Freeware Tesseract-OCR: Open-Source hOcr2Pdf.Net ist Open-Source	Hoch – Visual Studio (C#)	Probleme bei zu kleinen Texten Probleme bei schlechter Auflösung

Ergebnis

Software	Preis (0-3)	Automatisierung (0-3)	Ergebnis (0-3)	Gesamt (0-9)
Acrobat Pro DC	2	1	2	5
PDF-XChange + Ghostscript + iTextSharp + Tesseract-OCR + hOcr2Pdf.Net	2	3	1	6

Fazit

Die Software „Acrobat Pro DC“ bietet zwar das bessere Endergebnis, ist jedoch schlecht automatisierbar. Aus diesem Grund entschied man sich hier für die Lösung von „PDF-XChange + Ghostscript + iTextSharp + Tesseract-OCR + hOcr2Pdf.Net“.

3.7.2 Auslesen der Textebene eines „Searchable PDF“

Bewertung

Bei den Softwarelösungen wurde ein Augenmerk auf die Eigenschaften „Preis“ (Wie teuer ist die Lösung?), „Automatisierung“ (Wie gut lässt sich die Lösung automatisieren?) und „Ergebnis“ (Konnte aller Text der Textebene gelesen werden?) gelegt.

Kriterium	0	1	2	3
Preis	Mehr als CHF 100.00	Von CHF 50.00 bis 100.00	Von CHF 0.00 bis 50.00	Gratis (Freeware)
Automatisierung	Kein Batch-Job möglich	Batch-Job über Anwendung	Batch-Job über Druckfunktion möglich	Batch-Job per Programmiersprache oder Command Line möglich
Ergebnis	Mehr als 3 Unterschiede zum Original-dokument	2-3 Unterschiede zum Originaldokument	1 Unterschied zum Originaldokument	Keine Unterschiede zum Originaldokument

TikaOnDotnet v1.17.1

Preis	Automatisierung	Ergebnis
Freeware	Hoch – Visual Studio (C#)	Keine Probleme

PDFBox in .NET v1.8.9

Preis	Automatisierung	Ergebnis
Freeware	Hoch – Visual Studio (C#)	Keine Probleme

Ergebnis

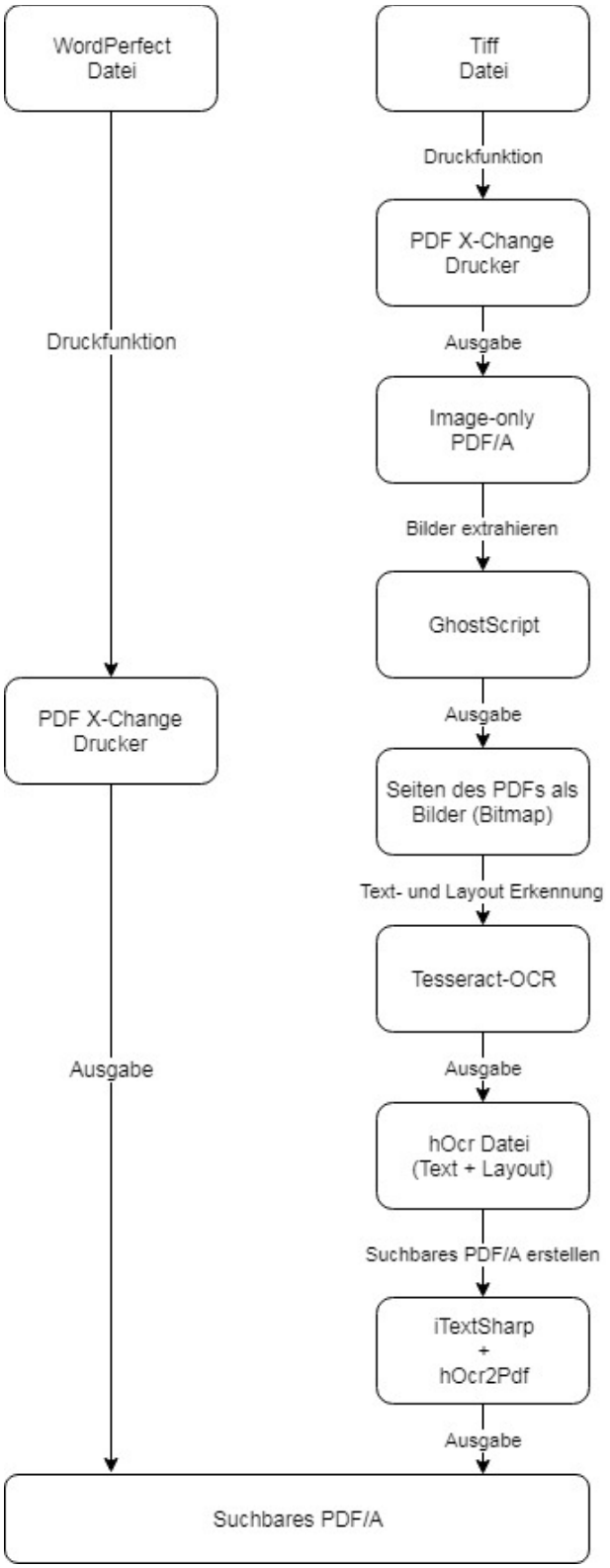
Software	Preis (0-3)	Automatisierung (0-3)	Ergebnis (0-3)	Gesamt (0-9)
TikaOnDotnet	3	3	3	9
PDFBox in .NET	3	3	3	9

Fazit

Da die beiden geprüften Softwarelösungen in allen Bereichen gleich gut sind, wurde zusätzlich die Verwendung in der .NET Umgebung angeschaut. In diesem Bereich hat „PDFBox in .NET“ das bessere Ergebnis erzielt, da sich die Schnittstelle der Biblothek einfacher verwenden lässt, weshalb man sich für „PDFBox in .NET“ entschied.

3.7.3 Konvertierung mit OCR

Die folgende Grafik soll den Ablauf der Konvertierung von einer WordPerfect- sowie TIFF-Datei zu einem „Searchable PDF“ nochmals verständlicher aufzuzeigen.



3.8 Volltextsuche und Indexierung

Bewertung

Bei den Volltextsuche-Lösungen wurde ein Augenmerk auf die Eigenschaften „Preis“ (Wie teuer ist die Lösung?), „Implementierung Indexierung“ (Wie einfach lässt sich die Indexierung implementieren?), „Implementierung Volltextsuche“ (Wie einfach lässt sich die Volltextsuche implementieren?), „Mitgelieferte Funktionalität“ (Welchen Funktionsumfang bietet die Lösung an?) und „Geschwindigkeit Indexierung“ (Wie schnell wird der Index erstellt?), „Geschwindigkeit Volltextsuche“ (Wie schnell werden komplexe Suchanfragen abgearbeitet?) gelegt.

Kriterium	0	1	2	3
Preis	Mehr als CHF 100.00	Von CHF 50.00 bis 100.00	Von CHF 0.00 bis 50.00	Gratis (Freeware)
Implementierung Indexierung	Sehr hoher Aufwand	Hoher Aufwand	Mittlerer Aufwand	Kleiner Aufwand
Implementierung Volltextsuche	Sehr hoher Aufwand	Hoher Aufwand	Mittlerer Aufwand	Kleiner Aufwand
Mitgelieferte Funktionalität (ausgenommen Indexierung und Volltextsuche)	Kein Zusatzfeature	1 Zusatzfeature	2-3 Zusatzfeatures	Mehr als 3 Zusatzfeatures
Geschwindigkeit Indexierung	Weniger als 1 MB/sec	Zwischen 1 MB/sec und 2 MB/sec	Zwischen 2 MB/sec und 3 MB/sec	Mehr als 3 MB/sec
Geschwindigkeit Volltextsuche (Komplexe Abfragen)	Über 20 sec	Zwischen 10 sec und 20 sec	Zwischen 4 sec und 10 sec	Unter 4 sec

Microsoft SQL-Volltextsuche (SQL Server 2014 Express)

Preis	Implementierung Indexierung	Implementierung Volltextsuche	Mitgelieferte Funktionalität	Geschwindigkeit Indexierung	Geschwindigkeit Volltextsuche
Freeware, wenn Microsoft SQL Express verwendet wird	Einfach mittels iFilter	Einfach mittels Abfragen	Keine bekannt	1 MB/sec	Über 20 sec

Lucene v3.0.3

Preis	Implementierung Indexierung	Implementierung Volltextsuche	Mitgelieferte Funktionalität	Geschwindigkeit Indexierung	Geschwindigkeit Volltextsuche
Open-Source	Einfach mittels Java oder .NET	Einfach mittels Java oder .NET	- Scoring - Faceting - Fielded Searching - etc.	3 MB/sec	Unter 4 sec

Solr v7.5

Preis	Implementierung Indexierung	Implementierung Volltextsuche	Mitgelieferte Funktionalität	Geschwindigkeit Indexierung	Geschwindigkeit Volltextsuche
Freeware	Nach Konfiguration durch Solr	Nach Konfiguration durch Solr	- Ranked Searching - Faceting - Fielded Searching - etc.	3 MB/sec	Unter 4 sec

Elasticsearch v6.5

Preis	Implementierung Indexierung	Implementierung Volltextsuche	Mitgelieferte Funktionalität	Geschwindigkeit Indexierung	Geschwindigkeit Volltextsuche
Freeware	Nach Konfiguration durch Elasticsearch	Nach Konfiguration durch Elasticsearch	- Ranked Searching - Faceting - Analytics - etc.	3 MB/sec	Unter 4 sec

Ergebnis

Software	Preis (0-3)	Implementierung (0-6)	Funktionalität (0-3)	Geschwindigkeit (0-6)	Gesamt (0-18)
Microsoft SQL Volltextsuche	3	4	0	1	8
Lucene	3	4	3	6	16
Solr	3	4	3	6	16
Elasticsearch	3	4	3	6	16

Fazit

Da „Solr“ sowie „Elasticsearch“ auf „Lucene“ aufbauen haben diese drei Produkte die gleiche Bewertung erhalten. Jedoch laufen „Solr“ und „Elasticsearch“ als separater Dienst und bietet eine Schnittstelle für die Suchergebnisse an, während „Lucene“ direkt in die Applikation implementiert werden kann.

Die „Microsoft SQL Volltextsuche“ bietet zwar eine einfache Implementation dieser zur Verfügung, bietet jedoch sonst fast keine Funktionalitäten, für komplexen Suchanfragen müssen erst eigene SQL Abfragen gebaut werden wodurch die Geschwindigkeit leidet. Aus diesem Grund schied hier die „Microsoft SQL Volltextsuche“ als Lösung aus [33][34].

Für eine kleinere Anwendung mit einer geringen Datenmenge (die konvertierten Dokumente der Synopsis Datenstruktur sind ca. 4GB gross) wäre ein separater Dienst ein „Overkill“, da in diesem Projekt nur die Funktionalität der Dokumentensuche benötigt wird. Aus diesem Grund entschied man sich für eine direkte Implementation von „Lucene“ in die Anwendung.

3.8.1 Lucene - Berechnung der Trefferquote

Die Berechnung der Trefferquote von den Lucene Suchergebnissen beruhen auf dem „Boolean model“ und der Formel „Practical scoring function“, welche auf dem Konzept „Term frequence/inverse document frequency“ sowie dem Model „Vector space model“ beruht [61].

Boolean Model

Das „Boolean Model“ verwendet die „AND“, „OR“ und „NOT“ Operationen, um zu bestimmen, ob ein Suchbegriff in einem Dokument vorkommt oder nicht.

Term Frequency/Inverse Document Frequency (TF/IDF)

Nachdem Lucene die Dokumente erhalten hat, indem der Suchbergriff vorkommt, müssen diese nach Trefferquote geordnet werden. Diese wird mit folgenden drei Werten berechnet:

- Term frequency: Wie oft kommt der Suchbegriff in einem Dokument vor. Je mehr er im Dokument vorkommt, desto mehr Gewicht erhält das Dokument bei diesem Wert.
- Inverse document frequency: Wie oft kommt der Suchbegriff in allen Dokumenten die Lucene gefunden hat vor. Je grösser die Anzahl an Dokumenten, in denen der Suchbegriff vorkommt, desto kleiner fällt die Gewichtung für dieses Dokument aus bei diesem Wert.
- Field-length norm: Wie gross ist das Feld, in dem der Suchbegriff im Dokument vorkommt. Wird der Suchbegriff z.B. im Dokumentnamen gefunden, so wird dies höher gewichtet als ein Treffer im Dokumenteninhalt.

Vector Space Model

Da meist nicht nur nach einem Suchbegriff gesucht wird, wird das sogenannte „Vector Space Model“ verwendet. Dabei werden alle drei Gewichtungen (siehe oben) für jeden Suchbegriff und jedes Dokument in einen Vector gelegt und miteinander verglichen. So können die Gewichtungen unter den Dokumenten miteinander verglichen werden und die Trefferquote daraus berechnet werden.

3.8.2 Lucene – Analyser, Tokenizer und Filter

Der Lucene „Analyser“ [62] generiert aus den indexierten Dateien sogenannte „Tokens“. Dies macht er mit verschiedenen „Tokenizer“ und „Filter“. Die Tokens sind Begriffe nach denen gesucht werden kann. Als Beispiel kann der Satz „Dies ist ein Beispiel.“ mittels Analyser in die Begriffe „Dies“, „ist“, „ein“ und „Beispiel“ unterteilt werden („LetterTokenizer“). Somit wird das Dokument, welches den Beispielsatz enthält mit jedem der Begriffe gefunden werden. Lucene bietet verschiede Arten von Analyser an. Hier die gängigen:

- **StandardAnalyser**: Dies ist der standardmässig von Lucene verwendete Analyser. Er unterteilt die Texte in Begriffe durch Stopwörter wie Leerzeichen, Punkte, Kommas („StandardFilter“) und entfernt diese („StopFilter“). Ausserdem konvertiert er die Begriffe in Kleinbuchstaben („LowerCaseFilter“) und erkennt URLs sowie E-Mail Adressen als ganze Begriffe.
- **StopAnalyser**: Dieser Analyser unterteilt Texte mit den Sonderzeichen („LetterTokenizer“). Ebenfalls enthält er einen „StopFilter“ und einen „LowerCaseFilter“, kann jedoch URLs und E-Mail Adressen nicht erkennen.
- **SimpleAnalyser**: Der „SimpleAnalyser“ enthält den „LetterTokenizer“ sowie den „LowerCaseFilter“, entfernt jedoch keine Sonderzeichen aus den Begriffen.
- **WhitespaceAnalyser**: Der „WhitespaceAnalyser“ ermöglicht das unterteilen der Texte mit Hilfe der Leerzeichen („WhitespraceTokenizer“).
- **KeywordAnalyser**: Dieser Analyser nimmt den gesamten Text als Begriff auf. Dies ist bei Feldern wie z.B. „ID“ oder „ZIP Codes“ nützlich.
- **Language Analyser**: Der „Language Analyser“ ermöglicht die Unterteilung der Texte in Begriffe durch hinterlegen der gewünschten Sprache (z.B. „EnglishAnalyser“). Diese enthalten den „StandardTokenizer“, den „StandardFilter“, den „EnglishPossessiveFilter“, den „LowerCaseFilter“, den „StopFilter“, sowie den „PorterStemFilter“.
- **Custom Analyser**: Dieser Analyser ermöglicht die Definition eines eigenen Analysers mit deren Filtern und Tokenizern.
- **PerFieldAnalyserWrapper**: Der „PerFieldAnalyserWrapper“ gehört nicht zu den Analysern, ermöglicht jedoch die Verwendung von verschiedenen Analysern für verschiedene Felder.

Fazit

Für dieses Projekt reicht die Verwendung des „StandardAnalyser“, da die Stopwörter nicht benötigt werden und er die Texte nach Wörtern in Begriffe unterteilt.

3.8.3 Lucene – Query Parser

Der Lucene „Query Parser“ [63] ermöglicht die Verwendung von bestimmter Syntax zur Suche von indexierten Dateien und deren Feldern. Als Beispiel kann mit einem „MultiFieldQueryParser“ per Suchsyntax angegeben werden, in welchen Feldern man genau suchen möchte, welche man ausschliessen möchte usw. Hier einige gängige „Query Parser“:

- **Default Query Parser**: Ermöglicht die Suche in einem oder allen Feldern.
- **MultiFieldQueryParser**: Ermöglicht die Suche in einem, in allen oder in Kombinationen von Feldern (Einschluss und Ausschluss von Feldern).
- **PrecedenceQueryParser**: Dieser Parser ermöglicht die gleiche Suche wie der „Default Query Parser“ jedoch mit besserer Unterstützung für die Suche mit Vorrang.
- **Surround**: Ermöglicht die gleiche Suche wie der „MultiFieldQueryParser“ jedoch ermöglicht er zusätzlich die Suche mittels „Span queries“ mit welchen man angeben kann, in welchem Bereich sich die Resultate befinden müssen.
- **Xml-Query-Parser:** Ermöglicht die Suche mittels XML Datei, in der angegeben wird, wie genau und nach was gesucht werden soll.

Fazit

Für dieses Projekt wurde der „MultiFieldQueryParser“ verwendet, da es für die Dokumentensuche von Vorteil ist, wenn man nach mehreren Feldern suchen kann.

3.9 Web-Anwendung

Für die Web-Anwendung wurde die Oberfläche „HTML5 Boilerplate“ [60] geprüft. Dieses wurde jedoch nicht verwendet, da der Funktionsumfang für dieses Projekt zu gross ist. Der Aufwand den Funktionsumfang für dieses Projekt zu verkleinern wäre grösser gewesen, als eine eigene Oberfläche zu entwickeln. Aus diesem Grund und der derzeitigen Programmierkenntnisse wurde die Webanwendung mit Hilfe des ASP.NET Framework entwickelt.

3.9.1 ASP.Net Framework

ASP.Net [36] ist ein „Web Application Framework“ von „Microsoft“, womit sich dynamische Webseiten, Webanwendungen und Webservices entwickeln lassen. In diesem Projekt wird es für die Erstellung der Webanwendung verwendet. Dabei wurde auch das von „Microsoft“ empfohlene „Code-Behind“ Modell verwendet, bei welchem die Logik getrennt in einer separaten Datei pro Seite abgelegt wird. Mit Hilfe von „Event Handler“ kann auf Interaktionen des Webanwendungsnutzer bzw. Ereignisse reagiert werden.

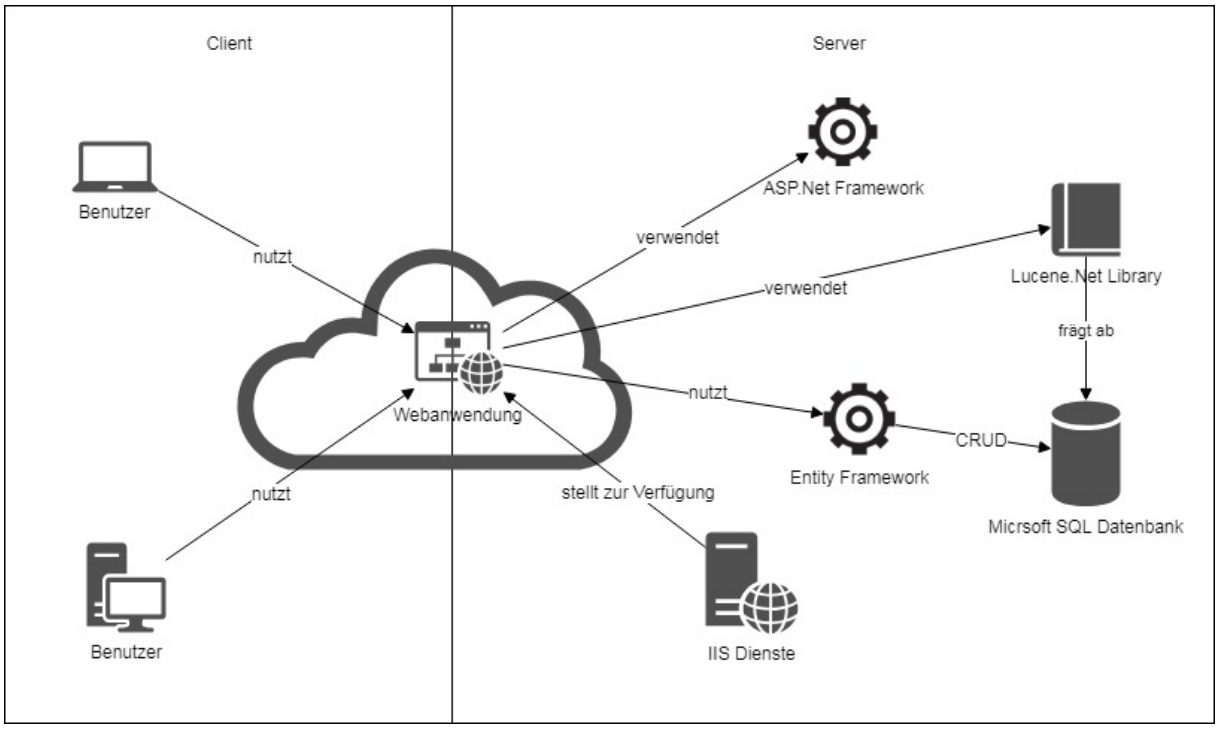
3.9.2 Entity Framework

Das Entity Framework [37] ist ein Framework von „Microsoft“ für objektrelationale Abbildung. Diese Technik wird im Englischen als „Object-relational mapping“ (ORM) bezeichnet. In diesem Projekt wird der ORM dafür genutzt, um auf die Daten in der Datenbank der Anwendung zuzugreifen, diese zu bearbeiten oder zu löschen (CRUD Operationen) ohne dabei direkt SQL Abfragen verwenden zu müssen.

3.9.3 IIS Dienste

Die Internet Information Services (IIS) [38][39] kommen ebenfalls aus dem Hause „Microsoft“. Es ist eine Dienste Plattform für PCs und Server, über welche sich z.B. Webseiten im Netzwerk bereitstellen lassen.

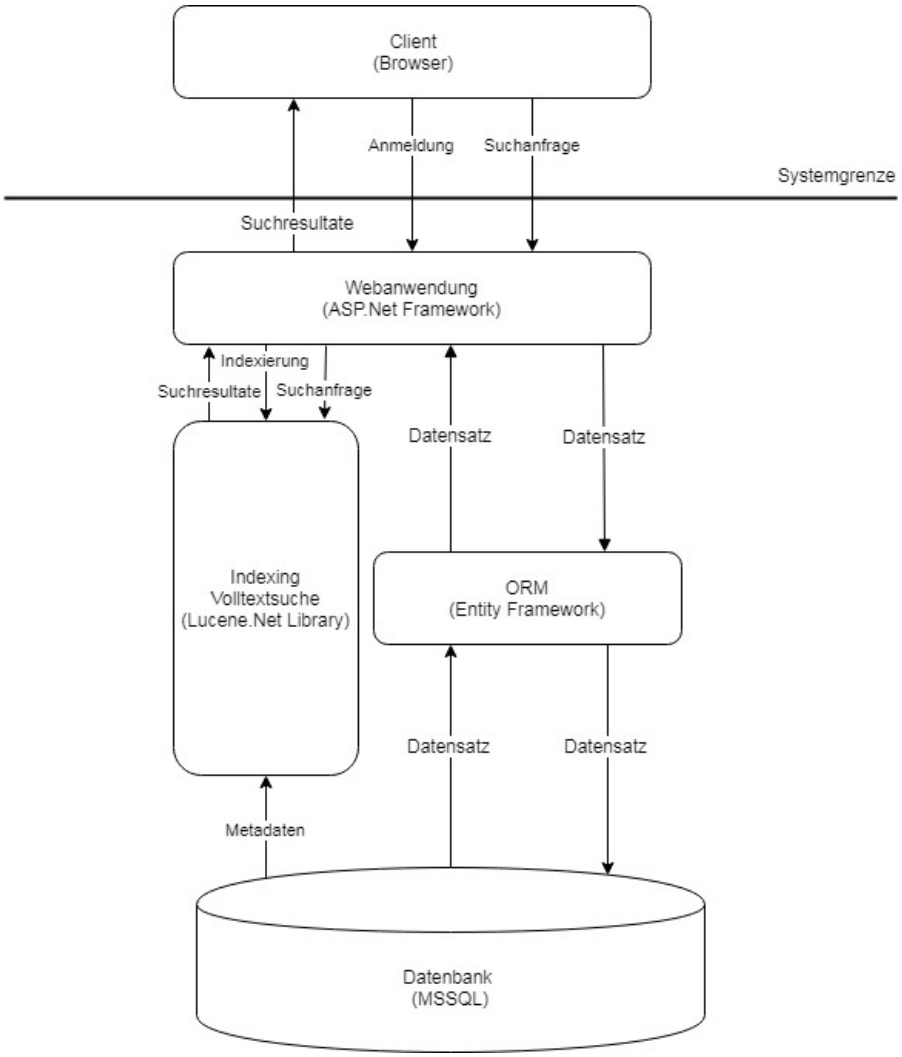
3.9.4 Abhängigkeiten



3.10 Systemmodell und Architektur

Die Anwendung wurde in einer Model-View-Control (MVC) Architektur aufgebaut. Die „View“ ist die Webanwendung, der „Control“ liegt durch das ASP.Net Framework ebenfalls in der Webanwendung (Code-behind). Das „Model“ wurde mittels MSSQL Datenbank und dem „Entity Framework“ umgesetzt.

Die folgende Grafik zeigt alle Beteiligten Akteure, die mit der Synopsis Webanwendung interagieren.



3.10.1 Datenbank

In der Datenbank („Microsoft SQL“ [40]) dient als zentraler Datenspeicher. Darin liegen alle Dokumente der neuen Synopsis Datenstruktur mit deren Metadaten, die Gruppen, die Benutzer, die Sichtbarkeiten sowie den Gruppenberechtigungen.

3.10.2 ORM

Der ORM welcher mittels „Entity Framework“ umgesetzt wurde, stellt die Schnittstelle zwischen Webanwendung sowie Datenbank her. Über ihn laufen alle Abfragen der Datensätzen (Dateien, Benutzer, Gruppen, Sichtbarkeiten, Gruppenberechtigungen), welche die Webanwendung aus der Datenbank benötigt oder in die Datenbank schreiben möchte (CRUD Operationen).

3.10.3 Volltextsuche und Indexierung

Dies ist neben dem ORM eine weitere Schnittstelle zwischen Datenbank und Webanwendung. Der Index für die Suchabfragen wird von „Lucene.Net“ [41] erstellt und verwaltet. Dazu werden alle notwenigen Metadaten aus der Datenbank abgefragt, indexiert und in lokale Dateien abgelegt. Die Suchanfragen der Webanwendung werden hier analysiert und mit Hilfe des zuvor erstellten Index verarbeitet, danach die Suchergebnisse an die Webanwendung zurückgeliefert.

3.10.4 Webanwendung

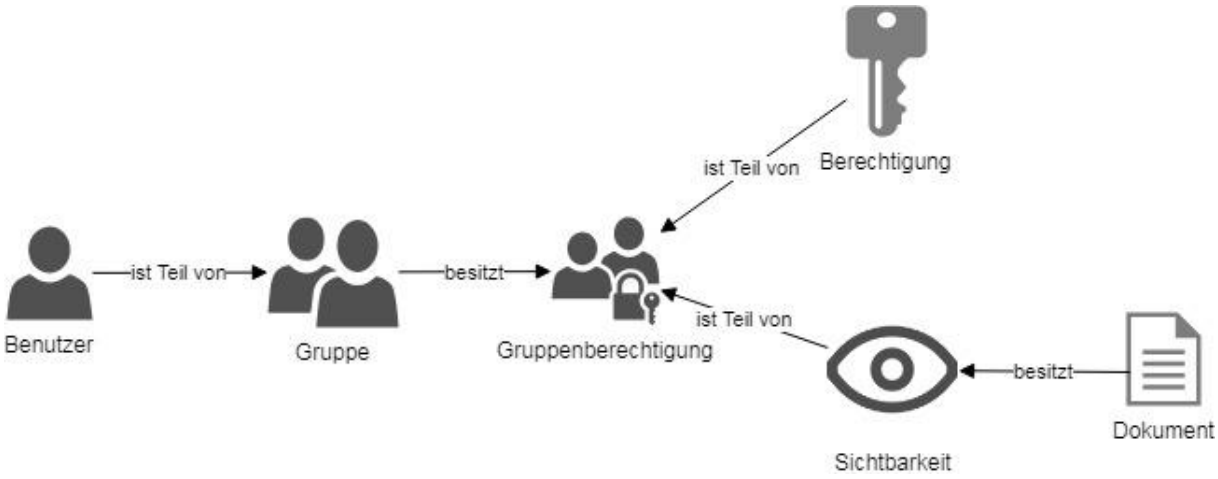
Die Webanwendung welche auf dem „ASP.Net Framework“ aufgebaut wurde, enthält zum einen die komplette Anwendungslogik sowie die Weboberfläche, über die der Besucher navigieren sowie interagieren kann. Alle Interaktionen, die der Besucher über den Browser mit der Webanwendung tätigt werden hier verarbeitet und darauf reagiert. Für die Oberfläche wurde zudem folgende JavaScript Frameworks verwendet: „jQuery“ [42], „jQuery UI“ [43] sowie „Notify“ [44].

Berechtigungen

Um die Webanwendung möglichst dynamisch zu gestalten, wurde diese wie folgt aufgebaut:

- Es existieren Benutzer mit Anmeldedaten
- Es existieren Gruppen, zu denen die Benutzer zugeordnet werden müssen
- Die Gruppen besitzen Gruppenberechtigungen (Zugriff auf die verschiedenen Webanwendungsbereiche, sowie Zugriff auf die Sichtbarkeiten)
- Es existieren Sichtbarkeiten, zu denen die Dokumente zugeordnet werden müssen

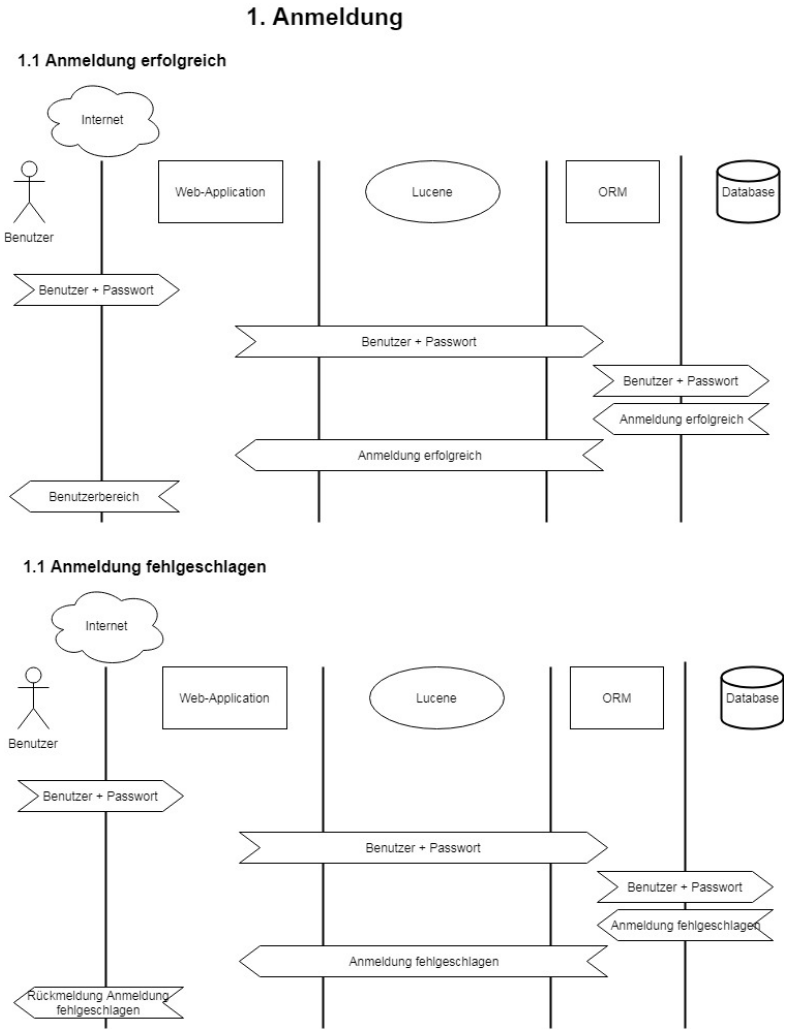
Durch diese Aufteilung können Benutzer zu Gruppen und Dokumente zu Sichtbarkeiten zugeordnet werden. Den Gruppen werden dann Berechtigungen (Zugriff auf Bereiche der Webanwendung) sowie Zugriff auf Sichtbarkeiten (bzw. Dokumentgruppen) erteilt.



Suchberechtigungen

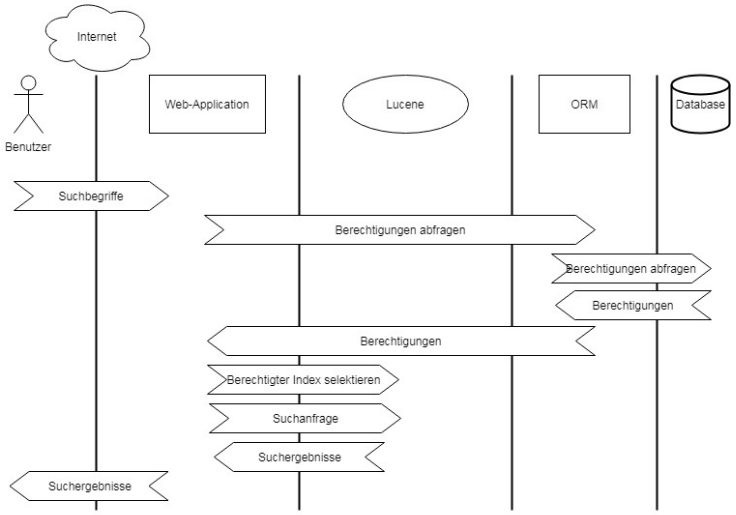
Da in der Anwendung den Dokumenten auch eine „Sichtbarkeit“ zugeordnet werden soll und je nach Gruppenberechtigung ein Benutzer nur bestimmte Dokumente mit einer „Sichtbarkeit“ suchen darf, traf man hier auf eine Schwierigkeit die sich mit der „Indexierung“ ergibt (da immer im gesamten Index gesucht wird, ausser man legt verschiedene Indizes an). Abhilfe schuf das sogenannte „Permission Filtering“ [35].

3.11 Sequenzdiagramme

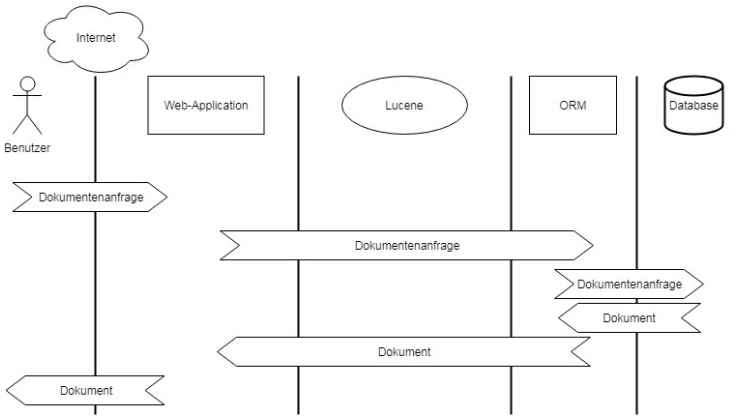


2. Suche

2.1 Suche (Voraussetzungen: 1.1, 3.1)

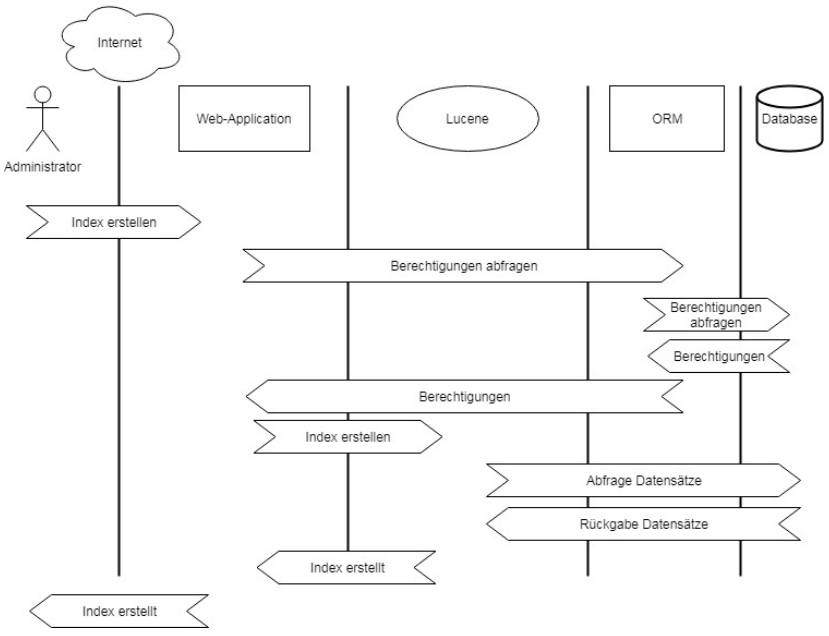


2.2 Dokumenten Download (Vorraussetzungen: 2.1)



3. Indexierung

3.1 Index erstellen (Voraussetzungen: 1.1)



4 Implementation

4.1 Systemvoraussetzungen

Für die Installation der Synopsis Web Anwendung sowie der Datenbank müssen folgende Kriterien erfüllt sein:

- Min. Microsoft Windows Server 2012 (min. Windows 7 SP1)
- Min. Microsoft SQL Server 2014 (inkl. Management Studio)
- Min. IIS 8.0 (ist beim Microsoft Windows Server 2012 bereits vorinstalliert)
- .NET Framework 4.6.1
- Festplattenspeicher: min. 30GB (SQL Server 2014: 6 GB, Datenbank: 10 GB, .NET Framework: 2.5 GB, ca. 10GB Reserve)
- Min. 1 GB RAM

4.2 Installation

4.2.1 Einrichten der Datenbank

Um die neue Synopsis Datenstruktur zu installieren, wird ein „Microsoft SQL Server Express 2014“, sowie das „SQL Server 2014 Management Studio“ benötigt.

1. Die Seite https://www.microsoft.com/de-ch/download/details.aspx?id=42299 aufrufen
2. Die Sersprache wählen und auf „Herunterladen“ klicken
3. Die für den Server passende Version wählen (32bit oder 64bit) – In diesem Fall war es „SQLEXPR_x64_ENU.exe“ sowie „SQLManagementStudio_x64_ENU.exe“
4. „SQLEXPR_x64_ENU.exe“ starten („New SQL Server stand-alone installation or add features to an existing installation“) und „Microsoft SQL Server Express 2014“ mit den Standardeinstellungen installieren (der „Instance Name“ sollte sich im Schritt „Installation Type“ für später notiert werden).
5. „SQLManagementStudio_x64_ENU.exe“ starten („New SQL Server stand-alone installation or add features to an existing installation“) und beim Schritt „Installation Type“ die Option „Add features to an existing instance of SQL Server 2014“ sowie die im letzten Schritt installierte SQL Instanz auswählen. Im nächsten Schritt „Feature Selection“ die Funktion „Management Tools – Basic“ mit den darunterliegenden Funktionen auswählen. Die Installation von „SQL Server 2014 Management Studio“ durch einen Klick auf „Next >“ bis am Ende abschliessen.
6. „SQL Server Management Studio“ starten und zur im Schritt 4 notierten SQL Instanz verbinden
7. Im „Object Explorer“ Rechtsklick auf „Databases“ und auf „Restore Database“. Im neuen Fenster unter „Source“ die Option „Device“ auswählen und mit einem Klick auf die Schaltfläche „…“ die Sicherung der Synopsis2018 Datenbank auswählen („Synopsis2018.bak“). Mit einem Klick auf die Schaltfläche „OK“ den Wiederherstellungsprozess starten.
8. Im „Object Explorer“ den Ordner „Security“ öffnen, den darunterliegenden Ordner „Logins“ öffnen, Rechtsklick auf das Login „NT AUTHORITY\SYSTEM“ und „Properties“ anwählen. Im neuen Fenster unter „User Mapping“ die „Synopsis2018“ Datenbank anwählen und im unteren Bereich die Optionen „public“ sowie „db_owner“ aktivieren. Mit einem Klick auf die Schaltfläche „OK“ die Einstellung speichern.

4.2.2 Einrichten des IIS

Um die Webanwendung auf dem Server zu hosten, werden die Internet Information Services (IIS), benötigt.

1. Den „Server Manager“ starten und im „Dashboard“ auf „Add roles and features“ klicken.
2. Bis zum Abschnitt „Server Roles“ durchklicken und unter „Web Server (IIS)“ die Funktion „Application Development“ (sowie alle darunterliegenden Funktionen) aktivieren.
3. Bis zum Abschnitt „Features“ durchklicken und die Funktion „.NET Framework 4.6 Features“ (sowie alle darunterliegenden Funktionen) aktivieren.
4. Die Installation abschliessen
5. Im Verzeichnis „C:\inetpub\wwwroot“ einen neuen Ordner mit dem Namen „Synopsis2018“ erstellen und die Webanwendung („Synopsis2018.zip“) in dieses Verzeichnis entpacken
6. Die IIS öffnen und auf der linken Seite unter „Application Pools“ über die Schaltfläche „Add Application Pool…“ einen neuen Pool mit dem Namen „Synopsis2018“ hinzufügen. Den neuen Pool „Synopsis2018“ anklicken und auf der rechten Seite auf „Advanced Settings…“ klicken. Im neuen Fenster unter dem Abschnitt „Process Model“ den Punkt „Identity“ auf „LocalSystem“ anpassen. Die Option darunter „Idle Time-out (minutes)“ kann hier auf die gewünschte Zeit angepasst werden, wie lange ein Benutzer ohne Reaktion eingeloggt bleibt (in Minuten).
7. Auf der linken Seite unter „Sites“, „Default Web Site“ sollte nun das Verzeichnis mit der Webanwendung erscheinen. Mit einem Rechtsklick auf dieses Verzeichnis und „Convert to Application“ öffnet sich ein neues Fenster. Bei „Application pool“ den im letzten Schritt hinzugefügt Pool („Synopsis2018“) anwählen und bei „Physical path“ den Pfad zur Webanwendung angeben („C:\intetpub\wwwroot\Synopsis2018“). Mit einem Klick auf „OK“ bestätigen.

4.2.3 Einrichten der Webanwendung

Damit die Webanwendung auch eine Verbindung zur Datenbank herstellen kann, müssen hier noch einige Anpassungen gemacht werden.

1. Die Datei „Web.config“ im Verzeichnis „C:\inetpub\wwwroot\Synopsis2018“ mit einem Texteditor öffnen.
2. „SQL Server Management Studio“ starten und zur SQL Instanz verbinden. Rechtsklick auf die Instanz (oberster Punkt im „Object Explorer“) und den Wert bei „Name“ kopieren. Unter dem Abschnitt „connectionStrings“ in der „Web.config“ den Wert „ALIEN-PC\SQLEXPRESS“ durch den kopierten Wert ersetzen (kommt zweimal vor).
3. Die „value“ beim Wert „LuceneIndexPath“ durch den gewünschten Pfad wo Lucene den Index abspeichern soll ersetzen.
4. Die Datei speichern.
5. Falls gewünscht kann auch eine sichere Verbindung via SSL eingerichtet werden (https://www.cool-it.at/blog/Februar-2018/Let-s-Encrypt-gratis-SSL-TLS-mit-IIS-und-Windows-S, https://farmcode.org/articles/lets-encrypt-webconfig-iis-redirect-for-http-to-https-allowing-http-access-to-the-well-known-folder/).
6. Mit einem Klick auf „Browse“ auf der rechten Seite in den IIS unter „Sites“, „Default Web Site“, „Synopsis2018“ kann die Webanwendung im Browser geöffnet werden. Beim ersten Aufruf der Webanwendung muss der Suchindex unter „Dokumentenverwaltung“ neu erstellt werden.

5 Validierung

5.1 Vergleich neuer Ist-Zustand mit Soll-Zustand

Anforderung „Säuberung und Umstrukturierung“ (KKL_A01)

Soll-Zustand	Ist-Zustand	Davon erreicht	Bemerkung
Dubletten sollen ignoriert werden.	Alle Dubletten wurden entfernt.	100%	Als Dubletten wurden Dokumente mit gleichem Dateinamen und Dateigrösse definiert. Es ist möglich, dass Dateien mit gleichem Inhalt existieren.
Nur die neuste Version der Dokumente sollen berücksichtigt werden.	Nur die neuste Version der Dokumente wurde übernommen.	100%	Als Dokument mit einer Versionsangabe wurde definiert, dass das erste Zeichen der Dateiendung ein Buchstabe und die beiden letzten je eine Zahl sein müssen. Hier wäre es möglich, dass noch Dateien existieren die früher in der Dateiendung falsch bezeichnet wurden.
Leere Ordner sollen ignoriert werden.	Leere Ordner wurden ignoriert.	100%	Da die neue Datenstruktur in einer Datenbank liegt, existieren hier auch keine Ordner mehr.

Anforderung „WordPerfect Konvertierung ins PDF/A Format“ (KKL_A02)

Soll-Zustand	Ist-Zustand	Davon erreicht	Bemerkung
Alle WordPerfect-Dateien sollen ins PDF/A Format konvertiert werden	98% der Dateien konnten erfolgreich ins PDF/A Format konvertiert werden.	98%	Die restlichen 2% konnten aufgrund von Fehlern, Passwortschutz oder leerem Inhalt nicht konvertiert werden.
Alle ins PDF/A Format konvertierten WordPerfect Dateien sollen möglichst mit der Formatierung des Originaldokuments übereinstimmen	Mit der Verwendung des „PDF XChange“ Druckertreibers der direkt übers Programm „WordPerfect“ angesprochen wird, konnte eine sehr hohe Übereinstimmung bezüglich der Formatierung erreicht werden.	99%	
Der Text des Dokuments soll möglichst identisch zum Originaltext als Textebene im PDF verfügbar sein.	Mit der Verwendung des „PDF XChange“ Druckertreibers der direkt übers Programm „WordPerfect“ angesprochen wird, konnte eine sehr hohe Übereinstimmung bezüglich der Textebene erreicht werden.	99%	

Anforderung „TIFF Konvertierung ins PDF/A Format“ (KKL_A03)

Soll-Zustand	Ist-Zustand	Davon erreicht	Bemerkung
Alle TIFF-Dateien sollen ins PDF/A Format konvertiert werden	99% der Dateien konnten erfolgreich ins PDF/A Format konvertiert werden.	99%	Die restlichen 1% konnten aufgrund von Fehlern, Passwortschutz oder leerem Inhalt nicht konvertiert werden.
Der Text des Dokuments soll möglichst identisch zum Originaltext als Textebene im PDF verfügbar sein.	90% der TIFF-Dateien konnte zufriedenstellend erkannt werden. Handgeschriebenes konnte ignoriert werden.	90%	Da die Texterkennung von der Bildqualität der TIFF-Dateien abhängig ist und diese in der alten Synopsis Datenstruktur eher schlecht war, wurden auch hier einige Texte unterschiedlich erkannt.

Anforderung „Dokumentensuche“ (KKL_A04)

Soll-Zustand	Ist-Zustand	Davon erreicht	Bemerkung
Suche nach „Erstellungsdatum“ soll möglich sein	„Erstellungsdatum“ ist in Datenbank abgelegt	100%	Umgesetzt mit der Lucene Suchsyntax
Suche nach „Dateiname“ soll möglich sein	„Dateiname“ ist in Datenbank abgelegt	100%	Umgesetzt mit der Lucene Suchsyntax
Suche nach „Schlagworten“ soll möglich sein	„Schlagwörter“ sind in Datenbank abgelegt	100%	Umgesetzt mit der Lucene Suchsyntax
Suche nach „Autor“ soll möglich sein	„Autor“ ist in Datenbank abgelegt	100%	Umgesetzt mit der Lucene Suchsyntax
Suche nach „Version“ soll möglich sein	„Version“ ist in Datenbank abgelegt	100%	Umgesetzt mit der Lucene Suchsyntax
Suche nach „Beschreibung“ soll möglich sein	„Beschreibung“ ist in Datenbank abgelegt	100%	Umgesetzt mit der Lucene Suchsyntax
Suche nach Dokumenteninhalt (Volltextsuche) soll möglich sein	Volltextsuche ist möglich (Volltext ist in Datenbank abgelegt)	100%	Umgesetzt mit der Lucene Suchsyntax
Möglichkeit, um Dokumente von der Suche auszuschliessen	Dokumente können von der Suche ausgeschlossen werden	100%	

Anforderung „Web-Oberfläche“ (KKL_A05)

Soll-Zustand	Ist-Zustand	Davon erreicht	Bemerkung
Weboberfläche für Dokumentensuche soll bereitgestellt werden	Weboberfläche für Dokumentensuche des bereinigten Synopsis Datenbestands ist über Browser erreichbar	100%	Umgesetzt mit ASP.NET sowie Lucene.Net
Möglichkeit zur Benutzerverwaltung	Über die Weboberfläche können Dokumenten, Gruppen, Benutzer, Sichtbarkeiten und Berechtigungen verwaltet werden	100%	

5.2 Zufriedenheit des Kunden

Dank der vorliegenden Semesterarbeit an der Hochschule Luzern (Studierender: Fabrizio Rohrbach; Leitung Prof. Michael Kaufmann) konnte eine Pendenz erledigt werden, das seit der Erstellung des KKL (1991-200) der Lösung harrt. In den 1990er Jahren setzte die Bauherren-Organisation und die Planer (aber noch die ausführenden Firmen) lokale PC-Netzwerke für Berichte, Korrespondenz, Tabellenkalkulation etc. ein, die Ablage dieser Dokumente erfolgte aber weitgehend auf Papier. Die Kommunikation erfolgte weitestgehend via Post und v.a. Fax. Trotzdem wurde ab 1997 im Hinblick auf bevorstehende Auseinandersetzungen zwischen der Trägerstiftung als Bauherr und dem Totalunternehmer eine Datenbank in einer frühen Version von MS Access angelegt. Dazu wurde ein Teil der Papierdokumente (insbesondere Korrespondenz von Dritten) eingescannt und verschlagwortet, andere Dokumente wurden im Format der damals weit verbreiteten Textverarbeitung WordPerfect abgespeichert. Mit der Datenbank konnten nicht-sprechende Filenamen aufgefunden werden, diese Files mussten dann in einer Ordnerstruktur von Jahreszahl – Monat – Tag manuell aufgesucht und mit der Originalsoftware oder als Bild in schlechter Auflösung gelesen werden. Die Datenbank funktionierte, wurde aber nie wirklich benützt, da die oben erwähnten rechtlichen Streitigkeiten mit einem pauschalen Vergleich erledigt wurden.

Im Hinblick auf das 20-Jahr-Jubiläum des KKL (1998 für den Konzertsaal, 2000 für das restliche Gebäude) tauchte die Idee auf, die damalige erhebliche Investition der Trägerstiftung nutzbar zu machen. Die detaillierte Dokumentation der Entstehungsgeschichte dieses Bauwerks sollte so einem grösseren Kreis von Interessierten und Fachleuten, insbesondere Historikern, Planern, Politikwissenschaftlern und Baujuristen zugänglich gemacht werden. Als ehemaliger Verantwortlicher der Bauherrschaft und als Initiant und Mit-Designer dieser halb vergessenen Datenbank wandte ich mich via Prof. Dr. René Hüsler, Direktor Departement Informatik, an Prof. Michael Kaufmann, der das Dokumenten-Aufbereitungs- und Such-Projekt unter dem Titel «Datenarchäologie» ausschrieb.

Die darauffolgende Zusammenarbeit mit Fabrizio Rohrbach verlief sehr positiv. Nach einer Sichtung und typologischen Auszählung der Dokumente wurde entschieden, zusätzlich zu den bereits in der alten Datenbank aufgenommenen Dokumenten (eingescannte Dokumente und verschlagwortete WordPerfect-Dokumente) die WordPerfect-Files aus den jährlich abgelegten Files der Bauherren-Organisation in die neu zu erstellende Datenbank zu übernehmen. Um die Datenmenge zu reduzieren bzw. nicht relevante Dokumente auszusortieren, entwickelte Fabrizio Rohrbach eine ganze Reihe von Test und entsprechende Programme. So wurden Doubletten, Vorversionen und Dokumente in spezifischen Pfaden ausgeschieden. Wiederum aufgrund von Tests entwickelte Fabrizio Rohrbach schliesslich ein Programm zur Konversion sowohl der Bild-Files als auch der WordPerfect Files in das PDF-Format wobei die Textebene für spätere Volltextsuche beibehalten bzw. erstellt wurde.

Die so kreierten PDF-Files sind nun via eine Internet-basiere Datenbank nach formalen Kriterien (Pfad, Daten, Autorenkürzel, etc.) und nach Volltext-Wörtern suchbar. Die Ergebnisse werden in übersichtlichen, von den Suchmaschinen her bekannten Listen sehr benutzerfreundlich präsentiert und können per Klick heruntergeladen werden. Zudem wurde die Datenbank mit einer reichen Nutzer-Verwaltung ergänzt. Damit können später entdeckte irrelevanten Dokumente unsichtbar gemacht, aber auch zusätzliche Dokumente hinzugefügt und zu verschlagwortet werden. So z.B. Dokumente im Format der Tabellenkalkulation Lotus 1-2-3, die aus zeitlichen Gründen im Rahmen der vorliegenden Arbeit nicht in die Datenbank aufgenommen werden konnten. Ebenso erlaubt es der Administrationsteil, Benutzer mit verschiedenen Zugriffsmöglichkeiten zu definieren bzw. Benutzergruppen zu schaffen. Auch dies ist für den Nutzwert der Aufbereitung zentral, weil der z.T. vertrauliche Charakter der Dokumente eine Selektion bzw. Legitimation der Nutzer erfordert.

Insgesamt hat Fabrizio Rohrbach in kurzer Zeit eine funktional sehr reiche und trotzdem benutzerfreundliche Suchmaschine für die KKL-Dokumente geschaffen. Mit einem geringen Rechercheaufwand können rasch und gezielt Dokumente zu bestimmten Vorgängen, Ereignissen oder Themen aufgefunden werden. Das Resultat der Arbeit übertrifft die anfänglichen Erwartungen deutlich. Die Herausforderungen der «Datenarchäologie» wurden gemeistert, obwohl kaum eines der damals verwendeten Programm bzw. Datenformate heute noch vorausgesetzt werden kann. Die Seminararbeit von Fabrizio Rohrbach ermöglicht es nun, die Dokumente über die Entstehungsgeschichte des KKL rechtzeitig zu dessen 20-jährigen Bestehen einem interessierten Benutzerkreis zur Verfügung zu stellen.

Thomas Held, 21.12.18

6 Schlussfolgerung

6.1 Erkenntnisse

Durch dieses Projekt konnte ich mir bisher nicht bekannte Technologien kennenlernen. Vor dem Start des Projekts war mir die Möglichkeit der „Indexierung“ sowie dessen Vorteile nicht bekannt. Auch die Vorzüge der Open-Source Software „Lucene“ für die Suchfunktion lernte ich in diesem Projekt zu schätzen. Die Verwendung eines OR-Mappers in der .Net Entwicklung war für mich ebenfalls Neuland, bisher hatte ich die Vorzüge des OR-Mappers nur im Modul „Applikationsentwicklung“ mit Bezug auf die Java Entwicklung kennengelernt.

Auch wurde mir bewusst, wie wichtig ein gutes „Requirements Engineering“ sowie „Change-Management“ ist. Die Anforderungen mussten wegen Unklarheiten einige Male überabeitet bzw. mit dem Auftraggeber in den Sitzungen neu ausgehandelt werden. Ebenfalls kamen neue Anforderungen hinzu (z.B. „Dokumentenbeurteilung durch Drittpersonen“ - siehe Protokoll vom 17.10.18), welche mit zusätzlichem Aufwand verbunden waren. Ein weiteres Beispiel wie wichtig das „Requirements Engineering“ ist, ergab sich gegen Ende des Projekts. Ich ging davon aus, dass eine neuere Dokumentversion immer den gleichen „Autorenkürzel“ besitzen muss, wie die alte Dokumentversion. Für den Auftraggeber hingegen war es klar, dass Dokumente mit dem gleichen Namen, jedoch nicht dem gleichen „Autorenkürzel“ dennoch in Bezug auf die Version zusammengehören. Aus diesem Grund musste gegen Ende des Projekts nochmals eine automatische Säuberung der Dokumente durchgeführt werden.

Auch zeigt sich, dass das kontinuierliche Testen während der Entwicklung enorm wichtig ist. Als Beispiel prüfte gegen Ende des Projekts der Auftraggeber eine Teilmenge der konvertierten Dokumente auf deren Relevanz. Dabei fiel auf, dass auch nach der Säuberung noch Dokumente mit gleichem Dateinamen sowie Dateigrösse vorhanden sind (Dubletten). Nach näherer Überprüfung konnte festgestellt werden, dass bei der automatischen Entfernung der Dubletten nicht alle Dubletten entfernt wurden. Deshalb musste am Ende des Projekts nochmals ein Bereinigungsdurchgang durchgeführt werden. Wären die Dokumente bereits früher geprüft worden, hätte man auf den Fehler in der automatischen Säuberung schneller reagieren können.

Für die „Säuberung und Umstrukturierung“ wurden zu Beginn des Projekts lediglich 8 Stunden geschätzt. Durch die zusätzlichen Arbeiten mussten hier insgesamt ca. 16 Stunden aufgewendet werden. Auch beim „Web-Interface“ wurden 8 Stunden zu wenig eingerechnet. Durch die Verwendung von „Lucene“ konnte jedoch Arbeitspacket „OCR, Volltextsuche und Indexing“ die verlorene Zeit wieder gut gemacht werden.

6.2 Ausblick

In diesem Projekt wurden die WordPerfect- sowie TIFF-Dateien der Synopsis Datenstruktur bereinigt und ins PDF/A Format konvertiert. Zudem wurden die konvertierten Dokumente mit deren Inhalt sowie Metadaten in eine Datenbank aufgenommen. Die dazu erstelle Weboberfläche bietet mittels Lucene die Dokumentensuche (auch innerhalb der Dokumente mittels Volltextsuche).

Der Auftraggeber wird diese in naher Zukunft nutzen, um mit der „Dokumentenbeurteilung durch Drittpersonen“ die Datenstruktur weiter zu säubern. Dazu werden den Beurteilern je einen Benutzer auf der Webanwendung, sowie die PDF Dokumente in einem lokalen Ordner zur Verfügung gestellt.

In einem weiteren Schritt wird der Auftraggeber die restlichen Dokumenttypen in der alten Synopsis Datenstruktur ebenfalls ins PDF/A Format konvertiert und in die neue Synopsis Datenstruktur aufnehmen.

Danach kann interessierten Personen einen Zugriff auf die Dokumentensuche in der neuen Synopsis Datenstruktur über die Webanwendung gewährt werden.

7 Quellenverzeichnis

[1] HxD - Freeware Hex Editor und Disk Editor | mh-nexus
Verfügbar unter: https://mh-nexus.de/de/hxd/ (Letzter Zugriff am 5. Dezember 2018)

[2] List of file signatures - Wikipedia 2018
Verfügbar unter: https://en.wikipedia.org/w/index.php?oldid=871804782 (Letzter Zugriff am 5. Dezember 2018)

[3] File Signature Database
Verfügbar unter: https://www.filesignatures.net/ (Letzter Zugriff am 5. Dezember 2018)

[4] Marco Pontello's Home - Software - TrID 2018
Verfügbar unter: http://mark0.net/soft-trid-e.html (Letzter Zugriff am 5. Dezember 2018)

[5] Types of PDFs: searchable PDF, true PDF, image-only PDF
Verfügbar unter: https://www.abbyy.com/en-au/finereader/pdf-types/ (Letzter Zugriff am 5. Dezember 2018)

[6] Portable Document Format - Wikipedia 2018
Verfügbar unter: https://de.wikipedia.org/w/index.php?oldid=183311191 (Letzter Zugriff am 5. Dezember 2018)

[7] PDF-Formate – Compart
Verfügbar unter: https://www.compart.com/de/pdf-formate-a-1-2-3-e-ua-vt-x (Letzter Zugriff am 5. Dezember 2018)

[8] PDF/A - Wikipedia 2018
Verfügbar unter: https://de.wikipedia.org/w/index.php?oldid=169465208 (Letzter Zugriff am 5. Dezember 2018)

[9] PDF/A-3 Übersicht 2016
Verfügbar unter: https://www.pdf-tools.com/pdf20/de/know-how/pdf-iso-standards/pdfa-3-uebersicht/ (Letzter Zugriff am 5. Dezember 2018)

[10] cognidox/OfficeToPDF
Verfügbar unter: https://github.com/cognidox/OfficeToPDF (Letzter Zugriff am 5. Dezember 2018)

[11] FoxPDF WordPerfect to PDF Converter, WordPerfect to PDF Converter, Convert WordPerfect to PDF, Convert WP5 to PDF, Convert WP6 to PDF, Convert WPF to PDF, Convert WPG to PDF, Convert WPG2 to PDF, WordPerfect to PDF, WP5 to PDF, WP6 to PDF, WPF to PDF, WPG to PDF, WPG2 to PDF etc. 2013
Verfügbar unter: http://www.foxpdf.com/WordPerfect-to-PDF-Converter/WordPerfect-to-PDF-Converter.html (Letzter Zugriff am 5. Dezember 2018)

[12] ABCpdf - C# PDF Library Component for .NET
Verfügbar unter: https://www.websupergoo.com/abcpdf-1.aspx (Letzter Zugriff am 5. Dezember 2018)

[13] WordPerfect Office | Free Trial
Verfügbar unter: https://www.wordperfect.com/en/product/office-suite/ (Letzter Zugriff am 5. Dezember 2018)

[14] PDF-XChange Standard
Verfügbar unter: https://www.pdf-xchange.de/pdf-xchange-standard/index.php (Letzter Zugriff am 5. Dezember 2018)

[15] PDF24 Creator
Verfügbar unter: https://de.pdf24.org/creator.html (Letzter Zugriff am 5. Dezember 2018)

[16] tesseract-ocr/tesseract
Verfügbar unter: https://github.com/tesseract-ocr/tesseract/wiki (Letzter Zugriff am 5. Dezember 2018)

[17] Tikaondotnet 2016
Verfügbar unter: https://kevm.github.io/tikaondotnet/ (Letzter Zugriff am 5. Dezember 2018)

[18] The C# OCR Library | Iron Ocr
Verfügbar unter: https://ironsoftware.com/csharp/ocr/ (Letzter Zugriff am 5. Dezember 2018)

[19] Scanned PDF to OCR (Textsearchable PDF) using C#
Verfügbar unter: https://tech.io/playgrounds/10058/scanned-pdf-to-ocr-textsearchable-pdf-using-c (Letzter Zugriff am 5. Dezember 2018)

[20] Ghostscript 2018
Verfügbar unter: https://www.ghostscript.com/ (Letzter Zugriff am 5. Dezember 2018)

[21] itext/itextsharp
Verfügbar unter: https://github.com/itext/itextsharp (Letzter Zugriff am 5. Dezember 2018)

[22] FairfieldTekLLC/Hocr
Verfügbar unter: https://github.com/FairfieldTekLLC/Hocr (Letzter Zugriff am 5. Dezember 2018)

[23] CodePlex Archive 2018
Verfügbar unter: https://archive.codeplex.com/?p=hocrtopdf (Letzter Zugriff am 5. Dezember 2018)

[24] Apache PDFBox | A Java PDF Library 2018
Verfügbar unter: https://pdfbox.apache.org/ (Letzter Zugriff am 5. Dezember 2018)

[25] How Full Text Search and IFilters Works in SQL SERVER | Ricoh Data Center
Verfügbar unter: https://www.ricohidc.com/kb/how-full-text-search-and-ifilters-works-in-sql-server/ (Letzter Zugriff am 5. Dezember 2018)

[26] PDF iFilter 64 11.0.01 2013
Verfügbar unter: https://supportdownloads.adobe.com/thankyou.jsp?ftpID=5542&fileID=5550 (Letzter Zugriff am 5. Dezember 2018)

[27] Windows TIFF IFilter Overview 2018
Verfügbar unter: https://docs.microsoft.com/en-us/previous-versions/windows/it-pro/windows-server-2008-R2-and-2008/dd834685(v=ws.11) (Letzter Zugriff am 5. Dezember 2018)

[28] Apache Lucene - Welcome to Apache Lucene 2018
Verfügbar unter: http://lucene.apache.org/ (Letzter Zugriff am 5. Dezember 2018)
[29] Apache Solr - 2018
Verfügbar unter: http://lucene.apache.org/solr/ (Letzter Zugriff am 5. Dezember 2018)
[30] Open Source Search & Analytics · Elasticsearch
Verfügbar unter: https://www.elastic.co/ (Letzter Zugriff am 5. Dezember 2018)
[31] Apache Solr vs Elasticsearch - the Feature Smackdown!
Verfügbar unter: http://solr-vs-elasticsearch.com/ (Letzter Zugriff am 5. Dezember 2018)
[32] Elasticsearch vs. Solr – Auswahl einer Open-Source-Suchmaschine
Verfügbar unter: https://www.searchtechnologies.com/de/blog/solr-oder-elasticsearch-als-top-open-source-suche (Letzer Zugriff am 5. Dezember 2018)
[33] Comparing Microsoft SQL Server Full-Text Search and Apache Lucene | DB Best Chronicles 2013
Verfügbar unter: https://www.dbbest.com/blog/lucene-vs-sql-server-fts/ (Letzter Zugriff am 5. Dezember 2018)
[34] Full Text Search Engines vs. DBMS | Lucidworks 2009
Verfügbar unter: https://lucidworks.com/2009/09/02/full-text-search-engines-vs-dbms/ (Letzter Zugriff am 5. Dezember 2018)
[35] Permission Filtering - Lucene Tutorial.com
Verfügbar unter: http://www.lucenetutorial.com/techniques/permission-filtering.html (Letzter Zugriff am 5. Dezember 2018)
[36] The Official Microsoft ASP.NET Site
Verfügbar unter: https://www.asp.net/ (Letzter Zugriff am 5. Dezember 2018)
[37] Übersicht – EF6 2018
Verfügbar unter: https://docs.microsoft.com/de-de/ef/ef6/ (Letzter Zugriff am 5. Dezember 2018)
[38] What is Entity Framework?
Verfügbar unter: http://www.entityframeworktutorial.net/what-is-entityframework.aspx (Letzter Zugriff am 5. Dezember 2018)
[39] Home The Official Microsoft IIS Site
Verfügbar unter: https://www.iis.net/ (Letzter Zugriff am 5. Dezember 2018)
[40] Microsoft® SQL Server® 2014 Express
Verfügbar unter: https://www.microsoft.com/de-ch/download/details.aspx?id=42299 (Letzter Zugriff am 10. Dezember 2018)
[41] Apache Lucene.Net 2014
Verfügbar unter: https://lucenenet.apache.org/ (Letzter Zugriff am 5. Dezember 2018)
[42] jQuery
Verfügbar unter: https://jquery.com/ (Letzter Zugriff am 5. Dezember 2018)
[43] jQuery UI
Verfügbar unter: https://jqueryui.com/ (Letzter Zugriff am 5. Dezember 2018)
[44] Notify.js 2018
Verfügbar unter: https://notifyjs.jpillora.com/ (Letzter Zugriff am 5. Dezember 2018)
[45] Lucene Query Syntax - Lucene Tutorial.com
Verfügbar unter: http://www.lucenetutorial.com/lucene-query-syntax.html (Letzter Zugriff am 5. Dezember 2018)
[46] Apache Lucene - Query Parser Syntax 2013
Verfügbar unter: http://lucene.apache.org/core/3_5_0/queryparsersyntax.html (Letzter Zugriff am 5. Dezember 2018)
[47] PDF Tools Online - Validate PDF
Verfügbar unter: https://www.pdf-online.com/osa/validate.aspx (Letzter Zugriff am 5. Dezember 2018)
[48] Working with Lucene.Net - ASP Free 2007
Verfügbar unter: https://www.aspfree.com/c/a/braindump/working-with-lucene-net/ (Letzter Zugriff am 5. Dezember 2018)
[49] Using Lucene.NET for Searching PDFs - Don't Panic Labs 2012
Verfügbar unter: https://dontpaniclabs.com/blog/post/2012/04/17/using-lucene-net-for-searching-pdfs/ (Letzter Zugriff am 5. Dezember 2018)
[50] Implementing Lucene.Net in Asp.Net web application 2018
Verfügbar unter: http://sonyblogpost.blogspot.com/ (Letzter Zugriff am 5. Dezember 2018)
[51] Simple search engine in C# with Lucene.NET 2012.
Verfügbar unter: https://csharpdreams.wordpress.com/2012/10/01/simple-search-engine-in-c-with-lucene-net/ (Letzter Zugriff am 5. Dezember 2018)
[52] Lucene.Net ultra fast search for MVC or WebForms site => made easy! – CodeProject
Verfügbar unter: https://www.codeproject.com/Articles/320219/Lucene-Net-ultra-fast-search-for-MVC-or-WebForms (Letzter Zugriff am 5. Dezember 2018)
[53] c# - How to hash a password - Stack Overflow
Verfügbar unter: https://stackoverflow.com/questions/4181198/how-to-hash-a-password/10402129#10402129 (Letzter Zugriff am 5. Dezember 2018)
[54] How To Install IIS In Windows Server 2016 2016.
Verfügbar unter: https://www.rootusers.com/how-to-install-iis-in-windows-server-2016/ (Letzter Zugriff am 5. Dezember 2018)

[55] interop - Read WPD (WordPerfect) files in .NET - Stack Overflow
Verfügbar unter: https://stackoverflow.com/questions/6662828/read-wpd-wordperfect-files-in-net (Letzter Zugriff am 5. Dezember 2018)
[56] It Could Be Done!: Dynamically Created Controls in ASP.NET 2018
Verfügbar unter: http://couldbedone.blogspot.com/2007/06/dynamically-created-controls-in-aspnet.html (Letzter Zugriff am 5. Dezember 2018)
[57] Store or save files/documents in SQL Server database using C#
Verfügbar unter: http://www.shabdar.org/sql-server/121-store-or-save-files-in-sql-server-database-using-c.html (Letzter Zugriff am 5. Dezember 2018)
[58] An Introduction to Entity Framework for Absolute Beginners – CodeProject
Verfügbar unter: https://www.codeproject.com/Articles/363040/An-Introduction-to-Entity-Framework-for-Absolute-B (Letzter Zugriff am 5. Dezember 2018)
[59] Convert PDF to instantly editable text, optical character recognition (OCR) software | Adobe Acrobat DC
Verfügbar unter: https://acrobat.adobe.com/us/en/acrobat/how-to/ocr-software-convert-pdf-to-text.html (Letzter Zugriff am 5. Dezember 2018)
[60] HTML5 Boilerplate: The web's most popular front-end template
Verfügbar unter: https://html5boilerplate.com/ (Letzter Zugriff am 5. Dezember 2018)
[61] Theory Behind Relevance Scoring
Verfügbar unter: https://www.elastic.co/guide/en/elasticsearch/guide/current/scoring-theory.html (Letzter Zugriff am 5. Dezember 2018)
[62] Guide to Lucene Analyzers | Baeldung
Verfügbar unter: https://www.baeldung.com/lucene-analyzers (Letzter Zugriff am 5. Dezember 2018)
[63] Exploring Query Parsers
Verfügbar unter: https://lucidworks.com/2009/02/22/exploring-query-parsers/ (Letzter Zugriff am 5. Dezember 2018)

8 Anhang

8.1 Sitzungsprotokolle

- Protokoll vom 18.09.18
- Protokoll vom 03.10.18
- Protokoll vom 17.10.18
- Protokoll vom 31.10.18
- Protokoll vom 28.11.18
- Protokoll vom 05.12.18
- Protokoll vom 12.12.18

8.2 Zusätzliches

- Bedienungsanleitung
- Projektplan
- Datenbank-Diagramm