Vom Papier ins Netz - Digitalisierung von Findmitteln im STABS

Vielen Dank für Einladung

Ich möchte kurz meine Firma vorstellen, damit Sie einen Eindruck bekommen, was wir für Archive tun können. Die Firma besteht seit 1999 und beschäftigt inzwischen - neben meiner Person - 2 Archivare und mehrere Hilfskräfte sowie Praktikanten der Fachhochschule Potsdam, Fachrichtung Archiv. Der Schwerpunkt liegt im Moment auf der Digitalisierung von Findmitteln, aber auch die klassische Bestandserschließung gehört zu unserem Angebot. Seit neuestem digitalisieren wir auch Film- und Tondokumente in Zusammenarbeit mit der Tonabteilung des Studio Babelsberg.

Nun möchte ich einen kurzen Überblick über den Ablauf des Projektes zur Digitalisierung der Findmittel im Staatsarchiv Basel-Stadt geben. Es gab mehrere Arbeitsschritte zu tun:
1 Scannen der Vorlagen
2 Inhaltliche Besprechung, Festlegung der Felder = Erfassungsanweisungen
3 die eigentliche Erfassung bzw. automatische Verarbeitung
4 Kontrolle der erfassten Daten auf Fehlerfreiheit von Inhalt und Struktur
5 Transformation in ein durch ScopeArchiv lesbares Format
6 Abgabe von Testdateien, Korrektur, Endabgabe aller Dateien
7 Speichern und Archivieren der Daten und Bilder

Nun zu 1., dem Scannen: In diesem Projekt wurde das Scannen der Vorlagen vom Staatsarchiv Basel selbst organisiert, es gibt in der Stadt eine preiswerte Möglichkeit hierzu. In meinem späteren Vortrag werde ich über ein Projekt berichten, bei dem wir selbst die Vorlagen gescannt haben. In diesem Fall wurden uns die Bilddaten also vom Staatsarchiv auf DVDs zur Verfügung gestellt.

2 In einem ersten Gespräch vor Ort wurden die Findmittel vorgestellt und es gab für mich die Möglichkeit, Fragen zu den Wünschen bezüglich der Erfassung zu stellen. In diesem Zusammenhang war ich auch für die Erläuterung dankbar, wie diese Findmittel entstanden sind, und in welchem Beständekontext sie stehen, mit diesem Wissen lassen sich Einzelfragen besser verstehen. Gemeinsam mit Herrn Kansy habe ich die benötigten Datenfelder festgelegt und von ihm die DTD (document type definition) für die xml-Struktur von scopeArchiv erhalten. Die archivische Analyse umfasst die Feldbelegung, die hierarchische Struktur der Datensätze sowie bestandsspezifische Besonderheiten, z.B. ob Bemerkungen zu übertragen sind oder nicht. Auch die Vereinheitlichung von Laufzeitangaben wurde besprochen.

3 Die Digitalisierung sollte mit dem Tool für Retrokonversionen docWORKS vorgenommen werden. Es handelt sich hierbei um die Software, die in Zusammenarbeit mit den Nordrhein-Westfälischen Archiven entwickelt wurde. Dieses Programm vereinigt OCR, das heißt die Schrifterkennung von Abby-Finereader mit einer automatischen Strukturierung.

Wir haben mit den einfachsten Findmitteln angefangen, diese in docWORKS einzulesen. Einfach heißt, daß keine oder wenige handschriftliche Anmerkungen vorhanden sind, eine gute die Bildqualität vorliegt (die war aber immer in Ordnung),

aber ganz besonders, daß die Struktur homogen ist. Am besten funktioniert die Verarbeitung von Findmitteln in Tabellenform.

Allgemein ist zu den Findmitteln zu sagen:
Insgesamt wurden in diesem Projekt 9000 Findbuchseiten des Staatsarchivs Basel-Stadt zur digitalen Endstufe geführt. Von diesen 9000 Seiten konnten ca. 4500 Seiten mit docWORKS verarbeitet werden, davon allein 1200 Seiten Hausurkunden, die eine gleichförmige Struktur aufwiesen. Der Rest der Findlisten wurde nach speziellen Anweisungen manuell erfasst. Insgesamt wurde nach der Strategie verfahren, nach einer gewissen Zeit des Austestens der Möglichkeiten von docWORKS abzubrechen, wenn sich die Vorlagen als nicht umsetzbar erwiesen.

3a Der erste Schritt in docWORKS ist die OCR-Kontrolle, die studentische Hilfskräfte übernommen haben. Die räumliche Nähe zur Fachhochschule Potsdam macht es möglich, Studenten zu beschäftigen, die archivisches Vorwissen haben, und keine Grundsatzerläuterungen mehr abgegeben werden müssen. Die Einarbeitung in die Arbeitsschritte Zonenprüfung und OCR-Kontrolle sind in docWORKS noch relativ einfach.
Komplizierter wird es, wenn es um die Verarbeitung der Struktur mit den selbst erstellten Regeln geht. Hier übernehmen ausschließlich ausgebildete und in docWORKS geschulte Archivare die Analyse und Verarbeitung der Dokumente.

POWERPOINT:

A	**Vorlesungen, Seminare und Uebungen an der Universität Basel** (Vgl. auch PA 904)	
A 1	Uebung: Die mittelalterliche Stadt [1 Faszikel]	SS 1969
A 2	Seminar: Probleme der spätmittelalterlichen Stadtgeschichte [2 Faszikel]	SS 1972
A 3	Repetitorium der Schweizergeschichte [1 Faszikel]	WS 1964/65

Als Beispiel für eine Verarbeitungsregel dient folgende. Sie ist eine der am meisten verwendeten Regeln:

Bei dieser Regel muß die Vorlage vertikale Linien aufweisen. Daran erkennt das Programm die Stellen, an der der Text vertikal geschnitten werden muß. Zusätzlich wird definiert, wieviele Spalten vorliegen (BILD) sowie daß in der ersten Spalte die Signatur, in der zweiten der Titel und in der dritten Spalte die Jahreszahlen enthalten sind.

POWERPOINT:

```
<PROCESS name="VM_PA962">
<RULE name="ProcessRecords"/>
<parameters>
<param name="NumberOfCol" value="3"/>
<param name="HasHeading" value="0"/>
<param name="HasVLine" value="1"/>
<param name="HasPageNumber" value="0"/>
<param name="IsHeadlineUnderlined" value="0"/>
```

<param name="CutMultipleEntry" value="1"/>
</parameters>
<COLUMNS>
<COL name="0" rule="SetAndVerifyEntryForVLineBooks"/>
<COL name="1" rule="AnalyzeMainContentVLine"/>
<COL name="2" rule="SetAndVerifyYearForVLineBooks"/>
</COLUMNS>
<KEYWORDLIST name="YearType" words="WL_ErziehungJahreszahlen"/>
<KEYWORDLIST name="EntryTypeForCut" words="WL_PA962EntryTypeForCut"/>
<KEYWORDLIST name="EntryType" words="WL_PA962Signatur"/>
</PROCESS>

Wortlisten
Zusätzlich gibt es Wortlisten, die definieren, welche Zeichenfolgen in bestimmten Feldern verwendet werden dürfen. In den meisten Fällen sind sowohl Buchstaben als auch Ziffern erlaubt. Deswegen sind diese Wortlisten fast über die gesamte Bearbeitungszeit gleich geblieben. Es gibt Wortlisten für die Signatur und die Laufzeit. In manchen Fällen kann z.B. in der Signatur ausgeschlossen werden, daß Buchstaben vorkommen.

Nachbearbeitung
Wenn man sich für die Verarbeitung in docWORKS entschieden hat, und die Strukturierung auch erfolgreich abgeschlossen wurde, gibt es trotzdem Fehler bzw. Ungenauigkeiten, die nach dem Export nachgearbeitet werden müssen:

Nach dem Export übernehmen wieder unsere Archivare die erste Kontrolle der Struktur, und sie entscheiden, ob eventuell eine Neubearbeitung vorgenommen werden muß. Sind die Dateien überwiegend in Ordnung, erfolgt eine zweite, genauere Kontrolle des Exports übernehmen wieder Hilfskräfte, hier kommt es besonders auf die Richtigkeit der Signaturen an, aber auch auf die korrekte Feldzuteilung und auf das Löschen von überflüssigen Absatzmarken.

-Generell sollten die Signaturen erst nach dem Export aus docWORKS kontrolliert und verändert werden. Korrigiert man die Signaturen in der Zonenprüfung, werden diese in jedem Falle einer erneuten OCR unterzogen, sodaß vorher durchgeführte Änderungen wieder verloren gehen. Dies bezieht sich unserer Erfahrung nach nur auf die Signaturen. Korrekturen an Titel und Laufzeiten bleiben in den weiteren Bearbeitungsstufen erhalten. Trotzdem waren nach dem Export auch Korrekturen im normalen Text nötig. Bestimmte Sonderzeichen, z.B. die kurzen Bindestriche "-" wurden als langer Bindestrich "–" umgesetzt. Dieser lange Bindestrich wird während des Exports als Fragezeichen "?" umgesetzt, welcher Umstand eine spätere manuelle Korrektur erfordert. Ein automatisches Suchen und Ersetzen ist hier nicht möglich, da Fragezeichen auch Teil von Zitaten sein können.

Nicht verarbeitete Vorlagen
Schnell fiel auf, daß einige Vorlagen nicht mit dem Retrokonversions-Tool umgesetzt werden konnten, obwohl sie zunächst so aussahen. Hier mußte leider viel Lehrgeld bezahlt werden, denn auch auf den ersten Blick "einfache" Findmittel sind begonnen worden, waren dann aber doch nicht zu verarbeiten.

Die Hälfte der Vorlagen (ca. 4500 Seiten) konnte aus verschiedenen Gründen nicht mit docWORKS bearbeitet werden. Hier sollen einige Beispiele und Gründe genannt werden.

1.) "Springende Spalten". Hier befinden sich die Laufzeitangaben nicht eindeutig in einer Spalte. Zusätzlich erschweren die Handschriften die Erkennung und erhöhen den Aufwand für die OCR-Korrektur. Dies kommt durch eingeklebte Titel, ein typisches Beispiel für nachbearbeitete gedruckte Findbücher.

POWERPOINT:

	6.	Basler Kompagnie . . . . . 1667—1668	
	5.	Luxemburg 1876- 1927	
M	1.	Mecklenburg . . . . . 1754- 1822—1881- 1962	
	2	Monaco	1886
N	1.	Nassau.	
	1.	Allgemeines und einzelnes . . . 1765—1866	

2.) Zu eng beschriebene Zeilen. Sie suggerieren zunächst zwar ein richtiges Schneiden zwischen den Verzeichnungseinheiten. Die vertikale Trennung ist in diesem Fall eindeutig und funktioniert auch richtig. Schwierig sind hier eher die horizontalen Abgrenzungen, also die Datensätze voneinander. Größtenteils werden hier Elemente über Kreuz vertauscht, das heißt, die Laufzeit des danachfolgenden Datensatzes erscheint schon bei der Verzeichnungseinheit davor und umgekehrt.

POWERPOINT:

P *	Gersbach, Fritz	1979
P	Gerster, Rita	
P	Gerster, Dr. Walter PD	1984
P *	Gessler, Lotte	1980
P	Gessler, Regina	2016
P	Geyl, Prof.Dr. P.	

Diese Findmittel und diejenigen, bei denen schon nach der ersten Sichtanalyse feststand, daß sie manuell erfasst werden müssen, wurden durch die Partnerfirma editura erfasst. Die allgemeinen Erfassungsanweisungen wurden so umgeschrieben, daß sie für betriebsfremde Arbeitskräfte schnell umgesetzt werden können. Die Texterfassung erfolgte in Rumänien zunächst auch in der DTD von docWORKS, erst später, als die Transformation fehlerfreie Daten produzierte, wurde auf die dann gültige Übernahmen.xsd umgestellt.

Betreuung

Während der ganzen Phase der Erfassung wurden Nachfragen der Mitarbeiter - sei es in Rumänien oder Fragen der Hilfskräfte in Deutschland, durch Archivare beantwortet. Wenn sie das Problem nicht selbst lösen konnten oder wenn es Darstellungswünsche des auftraggebenden Archives betrifft, dann erst wurde im Archiv nachgefragt. Dies geschah per mail mit Anhang des entsprechenden Bildes. Die Antworten aus Basel kamen erfreulicherweise meistens sehr schnell.

4 Anschließend wurden sowohl die Daten aus Rumänien als auch die Exporte aus docWorks, noch einmal auf Fehler hin geprüft. Typische Fehler von docWorks sind wie oben erwähnt die doppelte OCR der Signaturen und die fehlerhafte Umsetzung von Sonderzeichen. Typische Unsicherheiten der Erfassungskräfte in Rumänien sind das Entziffern von Handschriften, hier ist besonders die Grammatik schwierig. Die Struktur der Daten war in den meisten Fällen richtig.

doxTransform

Eine weitere wichtige und komplizierte Frage war die Transformation der Daten aus docWORKS in eine durch ScopeArchiv lesbare Form. In enger Zusammenarbeit mit dem Staatsarchiv wurden die Anforderungen an die xml-Dateien festgelegt. Unser Problem dabei war, daß wir das Programm ScopeArchiv nicht selbst anwenden und es uns nicht zur Verfügung stand, sodaß wir auf den Import und die Fehlermeldung des Staatsarchivs angewiesen waren, um die transformierten Dateien zu testen. Ich hoffe, daß wir Herrn Kansy und Herrn Müller nicht zu sehr beansprucht haben.

Das Tool doxTransform wurde zur Transformation, speziell der Laufzeiten und der Hierarchie von xml-Dokumenten von einer Struktur (dtd) in eine andere entwickelt, hier von docWorks zu ScopeArchiv. Möglich sind auch Transformationen in andere Datenbanken bzw. archivische Erschließungsprogramme. Wir haben nach einer automatischen Lösung gesucht, um diese wiederkehrenden und fehleranfälligen Arbeiten so effizient und zuverlässig wie möglich zu gestalten.

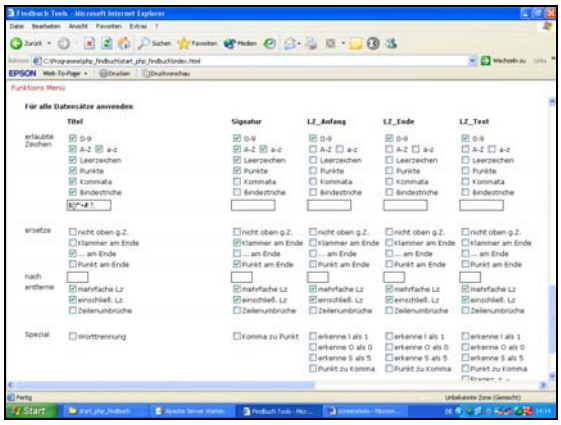
Die Software ermöglicht die Korrektur, die Nachbearbeitung sowie die Konvertierung von xml-Dateien verschiedenster Findmittel. Hierbei ist es unerheblich ob die Daten manuell in einer xml-Struktur erfasst wurden oder es sich um eine aus docWorks exportierte Datei handelt. Bei allen Programmteilen entsteht auf dem Bildschirm eine Tabellenansicht. In der linken Spalte ist der Originaleintrag enthalten, in der rechten Spalte sind dieselben Angaben in einem offenen Feld sichtbar, die manuell verändert werden können.

Das Programm basiert auf der Skriptsprache PHP, die als open source zur Verfügung steht. Für die Transformation wird der Apache http Server genutzt, der weitere Skriptsprachen zur Verfügung stellt, aus denen Websites, Datenbanken und xml-Strukturen erstellt bzw. bearbeitet werden können.

Die folgende Darstellung bezieht sich auf die Konvertierung von xml-Dateien aus docWORKS in das Zielformat ScopeArchiv.

Für alle Programmteile sind gewisse Analysen der Vorlagen nötig. So können Plausibilitätskontrollen für Signatur- und Laufzeitangaben voreingestellt und somit automatisch vom Programm überprüft werden.

<u>POWERPOINT:</u>



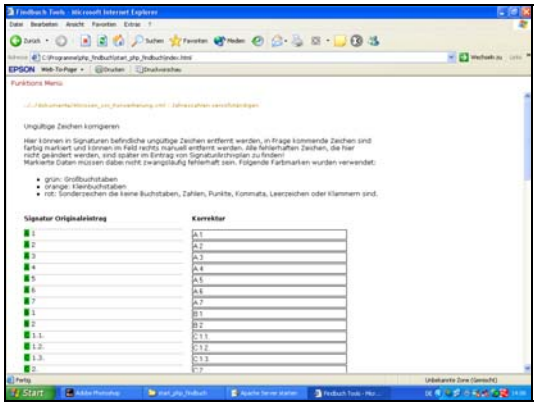
Funktion "Signaturprüpfung"
Im Programmteil Signaturprüfung werden die eingetragenen Signaturen auf ihre logische Richtigkeit überprüft, Buchstaben und Ziffern müssen also in aufsteigener Reihenfolge vorhanden sein. Es kann durch das Setzen von Häkchen an bestimmten Stellen zum Beispiel ausgeschlossen werden, daß Bindestriche vorkommen. Falls diese Zeichen doch in der xml-Datei vorkommen sollten, wird die betreffende Stelle farbig angezeigt und die entsprechende Stelle kann korrigiert werden.

Zudem erfolgt hier die Erweiterung unvollständiger Signaturen. In den Findmitteln waren meist nur die Endziffern und -buchstaben notiert, nicht aber die zusammengesetzte vollständige Signatur. Zum Beispiel hieß die erste Signatur "A", die zweite nur "1" statt "A 1". (SIEHE BILD SPRINGENDE SPALTEN; 2 bilder zurück)

Im vorliegenden Tool wird eine Datei erstellt, die das Feld SignaturArchivplan mit den Originaleinträgen sowie ein zusätzliches Feld für die automatisch generierte erweiterte Signatur enthält. Die Originalinformation geht also keinesfalls verloren.

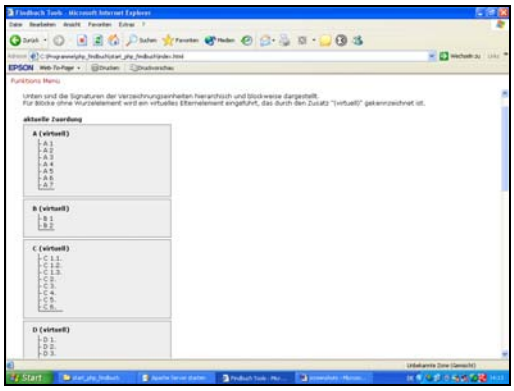
Es erfolgt zunächst die Überprüfung der eingetragenen Signaturen, falsche Einträge, die durch die OCR oder die manuelle Erfassung entstanden sind, können in der rechten Spalte manuell korrigiert werden.

POWERPOINT (hier sieht man, daß Buchstaben grün markiert sind, Sonderzeichen wären rot markiert etc.):



Nach Abschluss der Korrekturen werden die *Änderungen* in das Feld Archivplan eingelesen.

POWERPOINT:



Die Original-Signaturen werden automatisch in Blöcke eingeteilt. Zunächst sind die einzelnen Blöcke aufgelistet, die in dem entsprechenden Feld verändert werden können.

Eine falsche Zuordnung hat sich jedoch meist nicht ergeben, da die Signaturen mit besonderer Sorgfältigkeit sowohl bei der OCR als auch nach dem Export geprüft worden sind.

Falls doch ein Fehler auftaucht, können durch Setzen bzw. Entfernen von Häkchen Datensätze zu einem Block zusammengefasst bzw. in mehrere Blöcke getrennt werden. Es wird wieder eine Tabellensansicht generiert, in der man übersichtlich Änderungen vornehmen kann. Danach wird der Export gestartet, die exportierte Datei erhält hierbei eine Namenserweiterung zur Unterscheidung zur Originaldatei. Keine Datei wird also überschrieben, es entstehen immer neue Dateien.

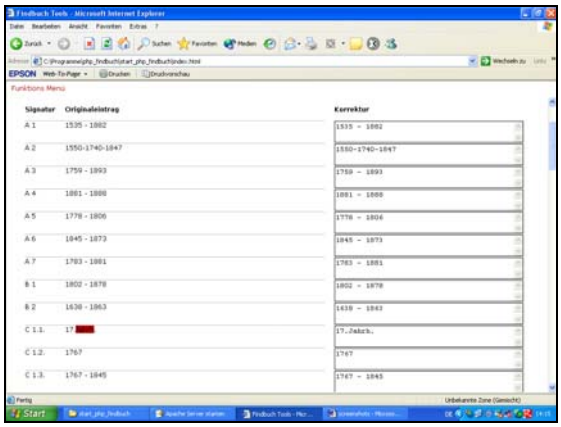
Die erweiterte nun vollständige Signatur findet in das Feld *Signatur* Eingang. Man unterscheidet also: die originale unvollständige Signatur wird in das Feld Archivplan, die erweiterte Signatur in das Feld Signatur eingelesen.

Funktion "Struktur & SignaturArchivplan"
Im Programmteil Struktur & SignaturArchivplan wird die soeben generierte Hierarchie der Signaturen in weitere Felder transformiert. Es erfolgt die Erstellung einer Datei, die zusätzlich die Felder VerzeichnungsId, VaterId und Zweigposition enthält.

Man wählt unter *Bearbeitete Dokumente* eine Datei aus, bei der bereits die Signaturprüfung durchgeführt wurde. Bei Dateien mit kompletten Signaturen in der Vorlage kann dieser Schritt übersprungen werden. Es ist für die Erstellung der Hierarchie unbedingt notwendig, daß die Signaturen vollständig ausgefüllt sind, da sonst die Zuordnung nicht automatisch erfolgen kann.

Die korrekte Hierarchie geht in die Datenfelder *VaterId* und *Zweigposition* ein. Es entstehen also Serien und Dossiers. Korrekturen der Hierarchie können selbstverständlich vorgenommen werden. Die im Export erstellte Datei erhält wieder eine Namenserweiterung. In diesem Stadium sind bereits die Felder SignaturArchivplan, VerzeichnungseinheitId, VaterId und Zweigposition erstellt worden.

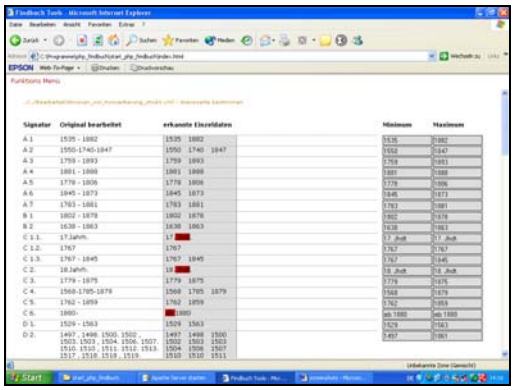
Funktion "Transformation"
POWERPOINT:



Im Programmteil Transformation erfolgt die Korrektur der Laufzeiten, die Ermittlung von Laufzeit-Anfang und Laufzeit-Ende, die Syntaxerstellung der Laufzeitfelder sowie die Konvertierung der gesamten Datei in das Übernahmeschema von ScopeArchiv.

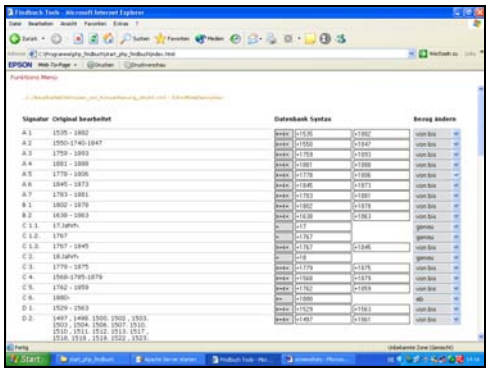
Es können nur Dateien mit gewissen Namenserweiterungen konvertiert werden, da das Vorhandensein bestimmter Felder, z. B. VerzeichnungseinheitId Voraussetzung für die Transformation ist. Es werden wie bei den Signaturen zunächst die Laufzeiten aus der Vorlage dargestellt. Fehlerhafte Einträge können hier manuell korrigiert werden.

POWERPOINT:



Nach dem Beenden der Korrektur werden Laufzeit-Anfang und Laufzeit-Ende (Minimum und Maximum) automatisch angezeigt. In vielen Findmitteln gibt es mehr als zwei Jahreszahlen, manchmal Tages- und Monatsangaben, manchmal auch drei Jahreszahlen, die mit Komma abgetrennt sind. Aus den unterschiedlichsten Vorlagen soll hier eine Vereinheitlichung vorgenommen werden.

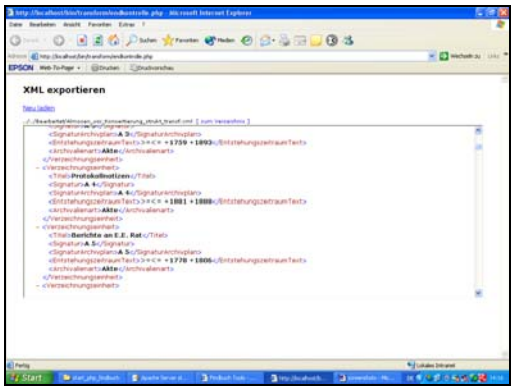
POWERPOINT:



Im nächsten Tool "Syntax erstellen" wird die Laufzeit in der Enddatei angezeigt. Änderungen können vorgenommen werden, indem in der rechten Spalte eine andere Auswahl getroffen wird, z.B. „sine dato“ statt ein angegebenes „ni“.

Ist die Syntax korrekt, erfolgt über den Button Export die Erstellung der endgültigen Konvertierungsdatei. Diese erhält wiederum eine Namenserweiterung "_transform". Nach Beendigung dieses Programmteils ist die Datei entsprechend den Scope-Anforderungen konvertiert und kann in die Scope-Datenbank eingelesen werden.

POWERPOINT:



Funktion "Fixieren/Validieren"
Das Zusatz-Tool *Fixieren/Validieren* ermöglicht das nachträgliche Einfügen von Leerzeichen in der Signatur, die Neuordnung der Tags, die Umbenennung doppelter Verzeichnungseinheiten sowie die erneute Durchzählung der Zweigpositionen.

Funktion "Dateiweit suchen und ersetzen"
Das Zusatz–Tool "Dateiweit suchen und ersetzen" ist eine praktische Dreingabe, die die Vorlage verbessern hilft. Es wird u.a. in solchen Fällen benötigt, wenn dateiweit einheitliche Veränderungen vorgenommen werden soll, z.B. wenn Abkürzung "u." zu "und" aufgelöst werden soll.

DoxTransform ist somit ein praktikables internes Hilfsmittel, um Kundenwünsche zur Konvertierung von Daten umzusetzen.

Abgabe an den Kunden

Nach der Bearbeitung werden die Dateien durch das Staatsarchiv Basel-Stadt importiert. Im Einzelfall erscheinen Fehlermeldungen, die wir selbstverständlich solange korrigieren, bis der Import reibunslos funktioniert.

Nach dem gesamten Projekt wurden die DVDs mit den Bilddaten zurückgegeben. Die abgelieferten Dateien speichern und archivieren wir im gesamten Gewährleistungszeitraum.

FAZIT: Das gesamte Projekt war auf 10 Monate veranschlagt. Inwzischen würden wir ein solches Projekt schneller bewältigen können, da mit jeder Digitalisierung die Erfahrung und die Routine wächst. Generell gilt aber, daß man zu Beginn die Vorlagen immer einfacher eingeschätzt, als sie wirklich sind. Dies ist die Herausforderung, die wir gerne annehmen.
In diesem Sinne... Herzlichen Dank fürs Zuhören.