Repräsentationen und Versionen in eCH-0160

1 Ausgangslage

Bei der Entwicklung und Spezifizierung von *arelda* im Bundesarchiv, später eCH-0160 SIP Ablieferungsschnittstelle, in den Nuller Jahren herrschte in der Archivwelt mehrheitlich die Meinung, dass in einer Digitalen Ablieferung nur abgeschlossene Dossiers und mithin abgeschlossene Dokumente vorkommen sollten. Versionen, Varianten, Doubletten und Kopien sollten vor der Ablieferung von der abliefernden Stelle aus den Dossiers entfernt werden, so wie das eigentlich auch bei der Aufbereitung einer Papierablieferung vorgesehen war. Damit war pro Dokument (mit den entsprechenden Metadaten) eine Primärdatei zu erwarten, in der das Dokument materialisiert ist. Dass mehrere Primärdateien ein Dokument repräsentieren, war nur für den Fall vorgesehen, dass aus technischen Gründen das Dokument nicht in einer Datei abgebildet werden konnte, z.B. *Single Page* TIFF für mehrseitige Dokumente. Darum ist es möglich, mit einem Dokument mehrere Dateien zu verbinden. Im Beispiel wäre dann die Reihenfolge der Seiten/TIFF-Dateien durch Dateinamen festgelegt worden (1.., 2.., 3..TIFF) – auch schon sehr unschön!

In diesem Paradigma waren Erhaltungsmassnahmen, d.h. Formatkonvertierungen, vollständig Teil des Ingest-Prozesses und weiterer Preservation-Prozesse im Archiv. Erst dabei würden unterschiedliche Repräsentationen des gleichen Dokuments entstehen (z.B. Konvertierung von Word-Primärdateien in PDF-Dateien). Diese Prozesse können dann auf Archivseite korrekt in PREMIS oder LMER[1] festgehalten werden.

Ein Paradigmenwechsel bei den Herstellern von Dokumentverwaltungssystemen und in den Archiven hat dann die Situation verändert. Die Hersteller von Dokumentverwaltungssystemen wollten den Dokumentabschluss mit der Konvertierung in ein unveränderbares Format finalisieren (damals dachte man, PDF sei das). Die Archive hingegen wollten, da nur eine Primärdatei pro Dokument vorgesehen war, dass diese Primärdatei in einem archivtauglichen Format abgeliefert würde, möglichst zeitnahe noch kontrolliert vom ursprüngliche Besitzer des Dokuments. Das war gewährleistet, wenn der Dokumentbesitzer beim Dokumentabschluss (zum Verschicken oder Drucken) das Dokument gezwungenermassen in PDF konvertieren müsste. Für die Archive war ein weiterer Hintergedanke, dass sich eine Konvertierung später im Archiv als schwierig herausstellen könnte und weil man ja nur eine Primärdatei archivieren wollte, wäre auch eine gute Kontrolle dieses entstehungsfernen Konvertierungsprozesses zwingend notwendig.

In der Folge wurde aufgrund technischer Innovationen die Konvertierung in PDF für die Hersteller von Dokumentverwaltungssystemen immer unwichtiger, die Compliance-Sicherheit kann heute auch mit anderen Mitteln auch über längere Zeit sichergestellt werden. Die Konvertierung in PDF wird heute von Dokumentverwaltungssystemen nicht mehr zeitnahe und im Prinzip nur noch für die Archive durchgeführt. Weil bei dieser PDF-Konvertierung keine manuelle Kontrolle mehr stattfindet, sondern ein mehr oder weniger kontrollierter Batchprozess die PDFs erzeugt, sind die Archive

[1] PREMIS hat sich weitgehend gegen LMER durchgesetzt, siehe auch: https://kost-ceco.ch/cms/dl/71d7e734490751c1e600021a5b8e9146/Vergleich_LMER-PREMIS.pdf

inzwischen dazu übergegangen, auch die ursprünglichen Office-Dokumente übernehmen zu wollen. Und wenn man schon die ursprünglichen Dokumente übernimmt, warum nicht gleich auch die Versionen zu diesen Dokumenten als Office-Dateien?

Damit ist eigentlich der ganze Prozess der Erhaltung durch Formatkonvertierung vom Ingest und Archivsystem ins Dokumentverwaltungssystemen gerutscht. Bei einem Ingest aus einem solchen System via eCH-0160 SIP müsste diese Information, wie die verschiedenen Primärdateien zu einem Dokument miteinander in Zusammenhang stehen, ebenfalls mitgeliefert werden. Dafür bietet eCH-0160 nur ein Element "Eigenschaften" zur Datei an. "Eigenschaften" kann zwar als *Key-Value*-Paar-Element strukturiert sein, es gibt aber keine Vorgabe für diese Strukturierung. So bleibt sie gezwungenermassen applikationsabhängig und kann maschinell nur in Zusammenhang mit der Entstehungsapplikation ausgewertet werden.

2 Problemstellung

Ein Dokument kann in eCH-0160 mit mehreren Primärdateien verbunden sein:

Folgende Fälle können wir in nicht vollständiger Annäherung unterscheiden:

- Ein Dokument besteht aus technischen Gründen aus mehreren einzelnen Primärdateien.
  Bsp. Jede Seite wird in einem *Single Page* TIFF abgebildet, ein Video besteht aus Gründen einer maximalen Dateigrösse aus mehreren Einzeldateien (DVD mit VOB Dateien).
- Mehrere Primärdateien bilden einen jeweils unterschiedlichen Aspekt eines Dokuments ab.
  Bsp. Postkarte Vorderseite–Rückseite, Plan INTERLIS-Datei und GeoTIFF
- Repräsentationen einer Ursprungsdatei: Es gibt eine ursprüngliche Primärdatei und eine oder mehrere davon abgeleitete Repräsentationen, die wiederum abgeleitet sein können.
  Bsp. Das gleiche Dokument: DOCX-Datei konvertiert in PDF und anschliessend PDF konvertiert in PDF/A, parallel dazu DOCX-Datei konvertiert in PDF/A und weiter PDF konvertiert in PDF/A-2u.
- Repräsentationen eines Objekts/Dokuments in mehreren Primärdateien: Vom gleichen analogen Objekt werden auf unterschiedlichem Weg digitale Repräsentationen erstellt, es gibt keine ursprüngliche Primärdatei.
  Bsp. Eine Urkunde wird mit klassischen Mitteln digitalisiert und später in einem Forschungsprojekt mit einem 3-D Verfahren erneut digitalisiert.
- Die 1 : n Beziehung (Dokument : Datei) wird zur Darstellung eines Versionierungsprozesses benutzt: Die einzelnen Dateien sind auseinander hergeleitet, aber genau genommen noch nicht das fertige Dokument und untereinander auch inhaltlich nicht gleich.
  Bsp. Die gesamte Versionierung oder Teile davon werden in einzelnen Dateien abgebildet, eine Datei stellt dann das vollständige/finalisierte Dokument gesamt dar.

3 Die einzelnen Fälle im Detail

3.1 Technische Gründen für mehrere Primärdateien

Das ist der einfachste Fall und ursprünglich wohl der Grund für die 1 : n Modellierung der Beziehung Dokument : Primärdatei in eCH-0160.

In diesem Fall bilden mehrere Primärdateien zusammen ein Dokument. Meist sind die Primärdateien vom gleichen Dateityp (z.B. TIFF), der Zusammenhang ist additiv oder sequentiell. Für die maschinelle Interpretation wäre es wichtig, die Reihenfolge der Dateien zu kennen.

Beispiel: wir haben eine Reihe von Scans in Form von TIFF-Dateien; diese bilden zusammen ein Dokument. Für die Vermittlung und Erschliessung soll nun ein PDF mit OCR erzeugt werden. Das bedeutet, die TIFF-Dateien müssen in richtiger Reihenfolge zusammengefügt werden können.

Wir brauchen also eine Information zur Reihenfolge der verbundenen Dateien und eine weitere Information, dass es sich hier um den Typ Reihung *(Sequence)* handelt
Oft ist das Quellsystem selbst nicht in der Lage diese Reihenfolge korrekt zu dokumentieren, sie ergibt sich vielmehr implizit aus dem Dateinamen oder dem Erzeugungsdatum der Primärdateien.

Typ der Beziehung: **Reihung *(Sequence)***   *Ordnung der Dateien*: Nummerierung, ordnen nach Dateinamen oder ordnen nach Creation Date. Das Ordnungselement ist also eine Zahl, ein String oder ein Datum.

3.2 Unterschiedliche Aspekte dargestellt durch mehrere Primärdateien

In dem Fall, wo mehrere Primärdateien unterschiedliche Aspekte eines Dokuments oder genauer vielleicht eines Archivobjekts abdecken, gibt es keine natürliche Ordnung, nur die domainspezifische Beschreibung erläutert die Zuordnung zum Objekt und eine mögliche Ordnung zwischen den Primärdateien.

Beispiel:
Eine Postkarte wird durch ein TIFF für die Vorderseite und eines für die Hinterseite repräsentiert.
Ein Plan wird durch eine Vektordatei in INTERLIS-Format und durch ein Geo-PDF dargestellt.

Es handelt sich hier nicht um Repräsentationen, die das Gleiche in unterschiedlicher technischer Form darstellen, sondern um jeweils ganz andere Informationen zum gleichen Objekt. So kann das Geo-PDF nicht aus der Vektordatei hergeleitet werden und umgekehrt.
Vorder- und Rückseite einer Postkarte könnte man alternativ auch als *Sequence*, ein Dokument bestehend aus zwei Dateien sehen.

Typ der Beziehung: **Aspekt *(Aspect)***
Ordnung der Dateien: Keine weiteren in eCH-0160 spezifizierten Metadaten, alle weiteren Metadaten sind domainspezifisch.

3.3 Primärdateien sind Repräsentationen eines Dokuments

Hier wollen wir davon ausgehen, dass die mit dem Dokument verbundene Primärdateien grundsätzlich das gleiche darstellen und sich nur formal unterscheiden (Dateiformat) und dass diese Repräsentationen mit technischen Mitteln auseinander abgeleitet werden können, wobei diese Ableitung nicht unbedingt umkehrbar sein muss.

Beispiel: Eine Word-Datei in Word2000-Format wurde beim Dossierabschluss in PDF 1.4 konvertiert und beide Dateien abgeliefert.

Repräsentationen können auch in komplexen, nicht linearen Beziehungen zueinander stehen. Solche Konvertierungen in Repräsentationen finden aber primär erst beim *Ingest* und oder später im Archiv statt. Dazu ein Beispiel: Ein Video in DigiBeta wird im *Ingest* in ein *preservation master* in FFV1 und eine *access copy* in MP4 konvertiert, später wird das *preservation master* File in eine sog. *Mezzanine*[2] Kopie in ProRes für die interne Weiterverarbeitung konvertiert.
Bedingt durch die heute weitgehend etablierte Langlebigkeit von verbreiteten Dokumentformaten, können CMS-Systeme getrost auf fortlaufende Formatkonvertierungen während der aktiven Zeit im Dokumentzyklus verzichten. So finden wir wohl selten mehr als zwei Repräsentationen vom gleichen Dokument in einer Ablieferung. Sind es nur zwei Repräsentationen, ist die Abhängigkeit *per se* linear, nämlich Original und Ableitung.
Komplexere Abhängigkeiten zwischen Repräsentationen können jedoch nur durch eine mitgelieferte PREMIS-Datei abgebildet werden.

Typ der Beziehung: **Repräsentationen *(Representation)***
Ordnung der Dateien: Nummerierung, ordnen nach Dateinamen oder ordnen nach Creation Date. Das Ordnungselement ist also eine Zahl, ein String oder ein Datum. Datei Nummer eins ist die Original- oder Ursprungsdatei, die folgenden Dateien sind Ableitungen davon.
Komplexere Abhängigkeiten zwischen Repräsentationen werden durch mitgelieferte PREMIS-Dateien abgebildet.

3.4 Primärdateien sind Versionen eines Dokuments

Versionen sind unterschiedliche Entwicklungsstufen eines Dokuments zu einem bestimmten Zeitpunkt. Jede Version hat also einen bestimmten Zeitstempel, der ihrer Entstehung entspricht (genaugenommen das Speicher- oder *Check In* Datum). Die letzte, nicht mehr veränderte Version wird zum finalen Dokument.
Es gibt offenbar zwei verschiedene Lösungsansätze in Dokumentverwaltungssystemen. Beim einen Ansatz wird jeweils die letzte eingecheckte Version zum finalen Dokument, bei anderen Ansatz wird das finale

[2] Ein zwischen oder *post production* Kopie in der Broadcast Industrie, siehe https://memoriav.ch/wp-content/uploads/2014/07/VARRFP.pdf

Dokument erst beim Schritt "Dokumentabschluss" erzeugt und wird dann auch im SIP zu einem neuen Dokument.

Beispiel: Es existieren von einem Dokument drei Versionen als Worddokumente. Die drei Versionen haben den gleichen Dateinamen, aber einen unterschiedlichen Entstehungszeitpunkt, dazu eine abschliessende finale Version des Dokuments in PDF. Das Dokument selber hat den gleichen Namen wie die Primärdateien. Es ist zu vermuten, dass die letzte Wordversion und die PDF-Version eigentlich Repräsentationen sind, wir wissen das aber ohne detaillierte Kenntnisse des Dokumentverwaltungssystems nicht.

Genaugenommen ist es unerheblich, ob die letzte Wordversion und die PDF-Version inhaltlich identisch, und damit Repräsentationen sind; wir haben auf dem Zeitstrahl, der die Versionierung abbildet, vier Objekte, die die einzelnen Schritte bei der Entstehung des Dokuments abbilden: Word X^1 -> Word X^2 -> Word X^3 -> PDF X^4

Typ der Beziehung: **Version *(Version)***   *Ordnung der Dateien:* nach Creation Date. Das Ordnungselement Datum.   Die letzte Datei nach Datum sortiert ist die finale Version

4 Lösungsvorschlag

Der Vorschlag versucht mit einem Minimum an Ergänzungen zu eCH-0160 v1.1 auszukommen und möglichst Rückwärtskompatibel zu bleiben.

Dazu wird das xml-Element *dateiRef* mit vier optionalen Attribute "reihung", "aspekt", "repraesentation" und "version" ergänzt, der Attributwert ist jeweils die Sortierreihenfolge.

Die Sortierung erfolgt nach sog. natürlicher Sortierfolge, d.h. sind alle Werte numerisch, wird numerisch sortiert, sonst alphanumerisch. Als Attributwert kann das *Creation Date*, der Dateiname oder ein numerischer Wert aus dem GEVER System verwendet werden.

Grundsätzlich gilt, die letzte Primärdatei repräsentiert das Dokument in der Form, wie es bei einer Überlieferung von Versionen, Repräsentationen etc. abgeliefert worden wäre.

5 Beispiele

Das Beispiel zeigt ein fiktives eingescanntes vierseitiges "Bibliographie von Planta" Ursprünglich wurden vier Seiten im format TIFF eingescannt (Multipage TIFFs waren offenbar nicht vorgesehen)

Die vom Scanndienst erstellten Dateinamen definieren in diesem Fall die Reihenfolge der Seiten.

Das *Creation Date* der TIFF Dateien ist offenbar nicht das wirkliche *Creation Date* beim Scanndienstleister, sondern das Datum der Akquisition der TIFF Dateien in der Amtstelle.

Zu einem Späteren Zeitpunkt wurde im Amt aus diesen vier TIFF Dateien zur besseren Verwendbarkeit eine PDF erstellt (2020-02-09).

Die Ordnung der Repräsentation ist also durch die zwei Datumsangaben "2019-04-15" und "2020-02-09" definiert. Die Datei mit dem Datum "2020-02-09" ist die aktuellste Repräsentation.

Die TIFF Dateien sind durch die alphanumerischen Werte "d00001.tif", "d00002.tif", "d00003.tif" und "d00004.tif" eindeutig geordnet.

```
   <titel>Bibliographie von Planta</titel>
   <autor/>
   <erscheinungsform>digital</erscheinungsform>
   <entstehungszeitraum>
      <von>
         <datum>2015-05-26</datum>
      </von>
      <bis>
         <datum>2019-04-15</datum>
      </bis>
   </entstehungszeitraum>
   <klassifizierungskategorie>Nicht Klassifiziert</klassifizierungskategorie>
   <zusatzDaten>
      <merkmal name="Ordnerpfad">\Ordner B docx</merkmal>
      <merkmal name="Dokumentdatum">26.05.2015</merkmal>
   </zusatzDaten>
   <dateiRef repraesentation="2019-04-15T12:01:00" reihung="1">d00001.tif</dateiRef>
   <dateiRef repraesentation="2019-04-15T12:01:00" reihung="2">d00002.tif</dateiRef>
   <dateiRef repraesentation="2019-04-15T12:01:00" reihung="3">d00003.tif</dateiRef>
   <dateiRef repraesentation="2019-04-15T12:01:00" reihung="4">d00004.tif</dateiRef>
   <dateiRef repraesentation="2020-02-09T15:01:00">d000024.pdf</dateiRef>
</dokument>
```