Darstellung von Repräsentationen und Versionen in eCH-0160

# Ausgangslage

Bei der Entwicklung und Spezifizierung von *arelda* im Bundesarchiv, später eCH-0160 SIP Ablieferungsschnittstelle, in den Nuller Jahren war man einhellig der 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]](#footnote-1) festgehalten werden.

Ein Paradigmawechsel 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 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.

# 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.

1. PREMIS hat sich weitgehend gegen LMER durchgesetzt, siehe auch: <https://kost-ceco.ch/cms/dl/71d7e734490751c1e600021a5b8e9146/Vergleich_LMER-PREMIS.pdf> [↑](#footnote-ref-1)