KOST-Newsletter Quartal 4, 2018

Immer Ärger mit PDF/A!

Anfragen zu Problemen mit PDF/A-Dokumenten treffen mittlerweile im Wochentakt bei der KOST-Geschäftsstelle ein. KOST-Val, das Format- und Paketvalidierungstool der KOST, identifiziert viele PDF/A-Dokumente in den Archiven als invalid. Und die Preservation-Planning-Expertengruppe PPEG der KOST hat bereits mehrmals problematische Eigenschaften von PDF/A-Dokumenten identifiziert, ausführlich untersucht und zu korrigieren versucht. Was ist bloss los mit unserem Lieblingsarchivformat, das «Archiv» sogar im Namen trägt?

Es gibt zwei Gründe für diese Häufung von Problemen. Zunächst ist offensichtlich, dass PDF/A für staatliche Archive das wichtigste Dateiformat ist und auf absehbare Zeit bleiben wird. Das Format erfährt daher eine überproportionale Aufmerksamkeit, und es existiert auch bereits ein riesiges Corpus an PDF-Dokumenten. Dass wir von anderen Formaten weniger hören, liegt zweifellos daran, dass wir sie weniger beachten.

Der andere Grund ist, dass die Benutzer an PDF und die Archive an PDF/A immer umfangreichere Anforderungen stellen. PDF/A ist längst kein simples Format für Textdokumente mehr, sondern auch ein Archivformat für Bilder, Tabellen, CAD-, GIS- und andere Daten. Und PDF/A soll auch immer mehr leisten: Die Volltextsuche in grossen PDF-Sammlungen beispielsweise war zur Zeit der Spezifikation von PDF 1.0 erst eine Vision – heute ist sie ein unabdingbarer Bestandteil bei Verwaltung von und Zugang zu rasant wachsenden Archivbeständen. Entsprechend wird die fehlende digitale Repräsentation von Schriften in validen PDF/A-1b-Dateien von einem Detail zum substanziellen Problem.

Um diese Situation zu meistern, müssen wir unsere Aufmerksamkeit der Formatvalidierung zuwenden. Was ist Formatvalidierung, und weshalb validieren wir? Die erste Frage ist schnell beantwortet: Die Formatvalidierung überprüft, ob eine Datei gemäss der Spezifikation ihres Dateiformats korrekt aufgebaut, also valide ist. Die zweite Frage ist komplexer. Unrealistisch wäre es zu meinen, dass das Archiv nur valide Dateien übernehmen kann und soll: Die Erfahrung zeigt, dass zu viele Dateien abgeliefert werden, die mehr oder weniger von der PDF/A Spezifikation abweichen. Diese nicht zu übernehmen, entspräche einer Nachbewertung qua Technik. Andererseits ist es oft unmöglich, die Dateien erneut und korrekt generieren zu lassen, und eine Konvertierung im Archiv ist heikel. Gefragt sind also eine korrekte Interpretation des Validierungsergebnisses und ein daraus resultierendes differenziertes Vorgehen.

Die KOST-Geschäftsstelle empfiehlt ihren Trägern deshalb folgendes:

- Wo eine Möglichkeit besteht, auf die Erzeugung von PDF/A-Dateien in der Verwaltung Einfluss zu nehmen, sollte diese unbedingt genutzt werden: Auf die JBIG2-Komprimierung ist wenn immer möglich zu verzichten; die PDF/A-Version 2u ist den anderen existierenden Versionen vorzuziehen. Solche Präferenzen können oftmals zentral festgelegt werden.
- Die Ergebnisse der Formatvalidierung sind auf jeden Fall zu dokumentieren. Gewisse Fehler können mit einiger Zuversicht ignoriert werden, weil alle gängigen Viewer mit ihnen wohl auch in Zukunft umgehen können. Andere beeinträchtigen gewisse Aspekte der Nutzung, z.B. digitale Lesbarkeit. In diesen Fällen ist die Information der Benutzerinnen und Benutzer von grosser Bedeutung.

Vor allem empfehlen wir Ihnen, sich von Problemen mit PDF/A nicht ins Bockshorn jagen zu lassen. PDF/A bleibt ein zentrales Archivformat, und ein bisschen vertieftere Kenntnisse über seine Eigenschaften und Versionen erlauben es, Probleme richtig einzuschätzen und ihre Auswirkungen in Grenzen zu halten.

Die soeben publizierte Version 1.9.x von KOST-Val erlaubt neu die Überprüfung, ob Schriften korrekt digital repräsentiert sind. Zudem ermöglicht KOST-Val nun, gewisse Fehlermeldungen auszublenden beziehungsweise zu ignorieren und damit den Prozess der Validierung differenzierter zu gestalten. KOST-Val 1.9.x verwendet neu die Java-Version 1.8 und ist deswegen leider langsamer als seine Vorgänger. Die KOST-Geschäftsstelle empfiehlt dennoch dringend, auf die neue Version zu upgraden.

Minimalanforderungen an ein digitales Langzeitarchiv

Seit ihrer Publikation 2009 haben die *Minimalanforderungen an die digitale Archivierung* der KOST als Richtschnur für ihre Arbeit gedient. Im November 2018 hat die Aufsichtskommission die Version 2.0 dieses Grundlagendokuments verabschiedet. Der Fokus verschiebt sich leicht von der *digitalen Archivierung* zum *digitalen Langzeitarchiv*, weswegen vor allem im Bereich der Infrastruktur und Sicherheit neue Anforderungen hinzukommen. Das Dokument wurde darüber hinaus inhaltlich gestrafft und neu nach dem Raster des nestor-Kriterienkatalogs organisiert. Die Minimalanforderungen an ein digitales Langzeitarchiv sind auf der KOST-Website publiziert.

xIsadg

Die Verzeichnungsschnittstelle xIsadg ist eine konzeptionelle XML-Implementation von ISAD(G). Im KOST-Projekt 17-034 xIsadg haben das Stadtarchiv Zürich, die Staatsarchive Appenzell Ausserrhoden, Basel-Stadt, Bern, Luzern, St.Gallen und Thurgau und die KOST-Geschäftsstelle bisherige Anpassungen und Erweiterungen an der letzten gemeinsamen Schemaversion 1.6 zusammengeführt und eine neue Version xIsadg 2.1 spezifiziert. In einem Anschlussprojekt überträgt eine reduzierte Projektgruppe (Staatsarchive ZH, BS, SG, LU, Stadtarchiv Zürich, KOST-Geschäftsstelle) das XML Data Dictionary ein eine Ontologie in der formalen Sprache OWL (Web Ontology Language). Abfragen auf Verzeichnungsinformationen in xIsadg sind dann mit den Mitteln von Linked Data im Sinne des Semantic Webs möglich.

SIARD-Format

2013 hat der Verein eCH das SIARD-Format zur Archivierung von Inhalten aus relationalen Datenbanken als eCH-0165 standardisiert. Inzwischen ist SIARD bei diversen Archiven in Europa im Einsatz. Um diese besser in die Weiterentwicklung des Formats einzubeziehen, hat das Schweizerische Bundesarchiv vorgeschlagen, die Verantwortung für SIARD dem neugegründeten Digital Information LifeCycle Interoperability Standards Board (DILCIS) Board zu übertragen. Das DILCIS Board ist eine Initiative der EU und des DLM Forums für die Bewirtschaftung von Standards zur digitalen Archivierung.

Der Expertenausschuss von eCH hat auf Antrag der Fachgruppe Digitale Archivierung die fehlerbehaftete Version 2.0 von eCH-0165 ausser Kraft gesetzt. Offizielle letzte eCH-Version ist wieder die Version 1.0. Die aktuelle Arbeitsversion 2.1 von SIARD ist bis zur Standardisierung durch das DILCIS Board auf der KOST-Website zugänglich.

Newsletter CECO du 4e trimestre 2018

Toujours des soucis avec le PDF/A !

Chaque semaine, des demandes concernant des problèmes liés au PDF/A parviennent au bureau du CECO. KOST-Val, l'outil de validation de paquets d'information et de formats du CECO identifie de nombreux documents PDF/A dans les archives comme non valides. Le groupe d'experts Preserveation Planning du CECO a décelé à de nombreuses reprises déjà des propriétés problématiques dans des documents PDF/A, les a examinées en détail et tenté de les corriger. Qu'arrive-t-il à notre format d'archivage favori dont l'intitulé contient même le mot « archive » ?

Il y a deux raisons à cette accumulation des problèmes. Il est tout d'abord évident que le PDF/A est, et demeurera dans un avenir proche, le format de fichiers le plus important dans les Archives d'État. Il est à ce titre l'objet d'une attention supérieure à la moyenne et il existe également déjà un corpus gigantesque de documents PDF. Si nous entendons moins parler des autres formats, c'est sans nul doute que nous les observons moins.

L'autre raison tient au fait que les utilisateurs ont d'immenses attentes envers le PDF, respectivement les archives envers le PDF/A. Ce dernier n'est depuis longtemps plus un simple format pour documents textuels, mais également un format d'archivage d'images, de tableaux, de données issues de la CAO, de géodonnées et autres données. Et on attend toujours plus de performances du PDF/A : la recherche en plein texte dans les grandes collections de PDF par exemple n'était qu'une vision d'avenir à l'époque de la spécification du PDF 1.0, elle est désormais un élément indispensable pour gérer et accéder aux fonds d'archives qui croissent à la vitesse grand V. De ce fait, l'absence de représentation numérique de polices d'écriture dans les fichiers PDF/A-1b valides passe du statut de problème mineur à celui de problème majeur.

Si nous voulons maîtriser cette situation, nous devons porter notre attention sur la validation de formats. Qu'est-ce qu'elle représente et pourquoi validons-nous des formats ? La première réponse est toute trouvée : la validation de formats permet de vérifier si un fichier est correctement conçu par rapport à la spécification de son format de fichier, autrement dit s'il est valide. La seconde réponse est plus complexe. Il ne serait pas réaliste de penser que les archives ne peuvent ou ne doivent prendre en charge que des fichiers valides. L'expérience montre en effet que trop de fichiers ne respectant pas ou pas complètement la spécification PDF/A sont versés. Ne pas les prendre en charge signifierait effectuer une évaluation qui serait dictée par la technique. D'autre part, il est souvent impossible de générer une nouvelle fois et correctement les fichiers, et procéder à une conversion dans les archives est délicat. Il faudrait donc disposer d'une interprétation correcte du résultat de la validation et établir à partir de là un processus différencié.

Le CECO recommande donc à ses membres le processus suivant :

- Lorsqu'il existe une possibilité d'exercer dans l'administration une influence sur la génération de fichiers PDF/A, il faut absolument en profiter. Il faut si possible éviter la compression JBIG2 et préférer la version PDF/A 2u aux autres versions existantes. Ces préférences peuvent souvent être définies de manière centrale.
- Il faut dans tous les cas documenter le résultat de la validation de format. On peut ignorer certaines erreurs avec confiance parce tous les lecteurs actuels et même à venir sont capables de les gérer. D'autres nuisent à certains aspects de l'utilisation, par exemple la lisibilité numérique. Dans ces cas, il est très important d'informer les utilisateurs et utilisatrices.

Nous vous recommandons surtout de ne pas vous laisser décontenancer par les problèmes du PDF/A. Il demeure un format d'archivage capital et des connaissances un peu plus approfondies sur ses propriétés et versions permettent de cerner correctement les problèmes et de limiter leurs répercussions.

La version 1.9.x de KOST-Val fraîchement publiée permet depuis peu de vérifier si la représentation numérique des polices d'écritures est correcte. En outre, KOST-Val permet dorénavant de masquer ou d'ignorer certains messages d'erreur et de mener à bien le processus de validation de manière différenciée. KOST-Val 1.9.x utilise depuis peu la version java 1.8 et, de ce fait, il est malheureusement plus lent que la version précédente. Le bureau du CECO recommande toutefois vivement de passer à la nouvelle version.

Exigences de base pour les archives numériques à long terme

Depuis leur publication en 2009, les *Exigences de base pour l'archivage numérique* du CECO ont servi de lignes directrices pour son travail. La commission de surveillance a approuvé la version 2.0 de ce document fondamental en novembre 2018. L'accent s'est légèrement déplacé de *l'archivage numérique* vers *les archives numériques à long terme*, raison pour laquelle de nouvelles exigences surtout en matière d'infrastructure et de sécurité ont été ajoutées. Le contenu du document a en outre été allégé et il est désormais organisé en suivant la structure du catalogue de critères nestor. Les Exigences de base pour les archives numériques à long terme sont publiées sur le site du CECO.

xIsadg

L'interface de description xIsadg constitue une implémentation conceptuelle de ISAD(G). Dans le cadre du projet du CECO 17-034 xIsadg, les Archives de la ville de Zurich, les Archives de l'État d'Appenzell Rhodes-Extérieures, Bâle-Ville, Berne, Lucerne, Saint-Gall et Thurgovie ainsi que le bureau du CECO ont jusqu'ici adapté et enrichi la dernière version commune du schéma (1.6) et spécifié une nouvelle version d'xIsadg (2.1). Un groupe réduit d'institutions (Archives de l'État de ZH, BS, SG, LU, Archives de la ville de Zurich, bureau du CECO) transfère dans le cadre d'un projet complémentaire le dictionnaire de données XML dans une ontologie utilisant le langage formel OWL (*Web Ontology Language*). Les requêtes d'informations descriptives dans xIsadg sont alors possibles avec les moyens de données liées au sens du web sémantique.

Format SIARD

En 2013, l'association eCH a édicté la norme eCH-0165 pour le format SIARD servant à l'archivage de contenus provenant de bases de données relationnelles. Entre-temps, SIARD est utilisé dans différentes archives d'Europe. Afin de mieux pouvoir impliquer ces dernières dans les futurs développements du format, les Archives fédérales suisses ont proposé de transférer la responsabilité de SIARD au groupe d'experts international Digital Information LifeCycle Interoperability Standards Board (DILCIS Board) créé récemment. Ce groupe est le fruit d'une initiative de l'UE et du DLM-Forum pour la gestion des normes d'archivage numérique.

A la demande du groupe spécialisé Archivage numérique, le comité d'experts d'eCH a annulé la version erronée 2.0 de la norme eCH-0165. La dernière version officielle d'eCH est la version 1.0. La version de travail actuelle 2.1 de SIARD (en allemand) est accessible jusqu'à la normalisation via le groupe DILCIS Board sur le site internet du CECO.