KOST-Newsletter Quartal 2, 2014

Umgang mit Formaten

Die Publikation von Version 4.0 des Katalogs archivischer Dateiformate KaD war für die KOST Anlass, das Thema Formate Ende Juni im Rahmen eines Workshops ausführlicher zu betrachten.

KaD v4.0

Die neue Version 4.0 des KaD umfasst wie angekündigt nur wenige Änderungen. Neben Umstellungen im Layout, welche die wesentlichen Informationen leichter auffindbar machen sollen, sind dies in erster Linie das Einarbeiten der Preservation-Planning-Empfehlungen in den KaD. Die KOST-Geschäftsstelle arbeitet zusammen mit der Preservation-Planning-Expertengruppe seit einiger Zeit an Fragen der Bestandserhaltung. Erkenntnisse zum Bildformat TIFF, welche die KOST in Workshops und in einer ausführlichen Studie dargelegt hat, sind nun auch im KaD, der zentralen Ressource zu Dateiformaten, nachgetragen. Weitere Themen werden folgen.

Anwendung des KaD

Mehr denn je gilt: Der KaD ist keine Blaupause für die digitale Archivierung. Vielmehr muss er mit Sachverstand interpretiert werden und zusammen mit anderen Vorgaben in konkrete Richtlinien und Empfehlungen einfliessen. Um dies noch klarer zum Ausdruck zu bringen, verwendet die KaD-Bewertungsmatrix in der letzten Zeile einen neuen Farbcode:

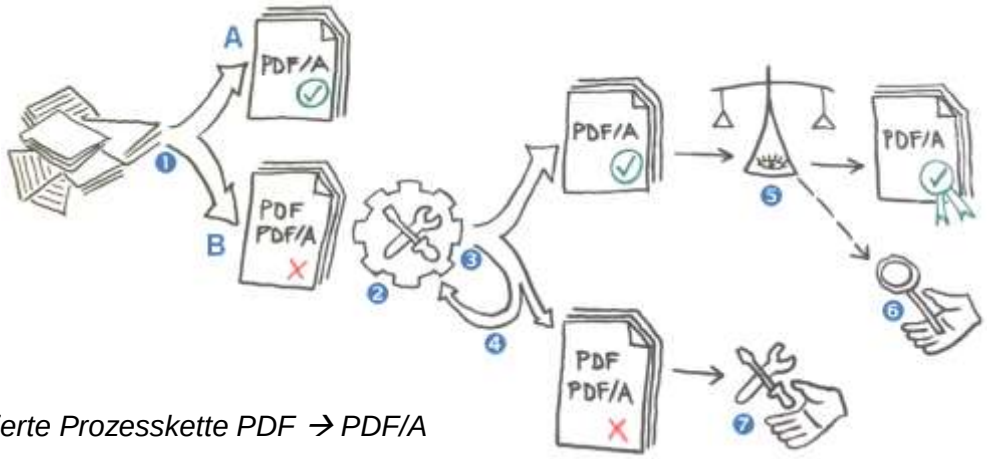
	TXT	PDF	PDF/A-1	PDF/A-2	PDF/A-3	ODF	OOXML	TIFF	JPEG	JPEG2000	PNG	DNG	PDF/A-2	WAV	MP3	Uncompressed Video	Digital Video	MPEG-2	MPEG-4
Gewichtete Summe	0.81	0.83	0.96	0.92	0.60	0.79	0.71	0.85	0.79	0.81	0.75	0.79	0.85	0.83	0.73	0.65	0.60	0.65	0.63
Logarithmisch gewichtete Summe	1.04	1.18	5.38	2.63	0.43	0.93	0.64	1.32	0.91	1.02	0.75	0.91	1.32	1.33	0.79	1.05	0.89	1.05	0.98

- Grün markiert sind die Formate, welche auf Grund der KaD-Analyse als besonders archivtauglich gelten können (gewichtete Bewertung 1.0 oder höher). Sie können den Kern von Archivempfehlungen bilden. Besonders nützlich sind sie dort, wo Archive auf Produktionsformate in der aktiven Phase des Lebenszyklus Einfluss nehmen können.
- Grau markiert sind Formate, die nicht die erste Wahl für die digitale Archivierung darstellen. Auf Grund ihrer Verbreitung werden die meisten dieser Formate jedoch auf kurze und mittlere Frist ohne Spezialaufwand lesbar bleiben. Ein Archiv, das existierende Unterlagen in solchen Formaten angeboten bekommt, kann in der Regel auf eine sofortige Konvertierung in eigentliche Archivformate verzichten, sofern es nicht eine strikte Formatpolitik verfolgt.
- Rot markiert sind Formate, deren gewichtete Bewertung unter 0.6 liegt. Diese Formate können explizit als für die Langzeitarchivierung nicht tauglich bezeichnet werden; von ihrer Übernahme ist abzuraten.

Diese Kategorisierung hat in erster Linie zum Ziel, unnötige bzw. kontraproduktive Konvertierungen zu verhindern.

Automatisierte Formatkonvertierung

Ein neues Forschungsgebiet der KOST sind automatisierte Formatkonvertierungen. Am Workshop „Dateiformate im Zentrum der digitalen Archivierung“ wurden dazu mögliche Workflows sowie ein *Proof of Concept* präsentiert. Solche automatisierten Konvertierungen verbinden Validatoren, Konvertierungstools und Werkzeuge für den automatischen visuellen Abgleich zu einer Prozesskette. Diese gestaltet die Konvertierung grosser Dokumentmengen effizienter und konzentriert menschliche Intervention auf genau diejenigen Dokumente, die Probleme bereiten.



Diskutierte Prozesskette PDF → PDF/A

In der Diskussion am Workshop von Ende Juni wurde die Verfügbarkeit von Tools für die automatisierte Formatkonvertierung allgemein begrüsst, jedoch Zweifel geäussert an ihrer Tauglichkeit für den Massenbetrieb. Ein Einsatz als Demonstrator oder für kleinere Aufträge wurde als wahrscheinlicher betrachtet. Die Geschäftsstelle wird ihren *Proof of Concept* KOST-Converter weiterentwickeln und den Trägerarchiven zur Verfügung stellen, um mögliche Anwendungsszenarien erproben zu können.

Validierungstools

Verschiedene Tools zur Formatvalidierung wurden am Workshop vorgestellt und getestet:

- Jpylyzer, ein Kommandozeilentool zur Validierung und Charakterisierung von JPEG2000-Dateien. Jpylyzer wird demnächst in KOST-Val integriert.
- MP3val, ein Validierungstool für MP3-Dateien.
- CSV Validator, ein Tool zur Validierung von CSV-Dateien gegen ein speziell dafür geschriebenes Schema. Die Anwendbarkeit ausserhalb des ursprünglichen Anwendungsfalls beim englischen Nationalarchiv scheint beschränkt.
- FITS, das File Information Tool Set, eine Kommandozeilenanwendung, die verschiedene Tools zur Formaterkennung und Dateicharakterisierung integriert.

Daneben wurden die aktuelle Version von KOST-Val, dem Formatvalidator der KOST, vorgestellt und weitere Entwicklungsziele skizziert. Neben der erwähnten Integration von jpylyzer steht für die KOST aktuell die Implementierung der dualen PDF/A-Validierung im Vordergrund. Diese ist ein Ansatz, um mit der Komplexität der PDF/A-Spezifikation besser und effizienter zurecht zu kommen. Nicht nur ist die PDF/A-Spezifikation ein umfangreiches System von Standards ohne klar umrissene Grenze; gewisse Bestimmungen sind so formuliert, dass sie von verschiedenen Tools verschieden interpretiert werden. Deshalb verwendet die duale Validierung zwei (qualitativ hochstehende gemäss KOST-Studie) PDF/A-Validatoren und betrachtet Dokumente als valid, die von mindestens einem der Validatoren als valid identifiziert werden. Der praktische Teil des Workshops demonstrierte die Nützlichkeit eines solchen Ansatzes.

Mögliche KOST-Projekte

Im letzten Newsletter haben wir die Projektbörse der KOST vorgestellt. Für die bereits eingereichten Projektvorschläge danken wir bestens. Zukünftig werden wir auf Seite 5 des Newsletters jeweils einen Überblick über die möglichen KOST-Projekte und die daran beteiligten Archive geben.

Newsletter CECO du 2e trimestre 2014

Gestion des formats

La publication de la version 4.0 du Catalogue des formats de fichiers d'archivage (Cfa) a été pour le CECO l'occasion de considérer plus en détail le thème des formats dans le cadre d'un atelier à la fin juin.

Cfa v4.0

Comme annoncé, la nouvelle version 4.0 du Cfa ne comprend que peu de modifications. Outre des transformations dans la mise en page destinées à faciliter la recherche des informations essentielles, elles concernent avant tout l'insertion dans le Cfa des recommandations sur la planification de la pérennisation. Le Bureau du CECO, en collaboration avec le groupe d'experts Planification de la pérennisation, travaille depuis quelque temps à des questions de conservation des collections. Des découvertes sur le format graphique TIFF que le CECO a présentées dans le cadre d'ateliers et d'une étude détaillée ont également été ajoutées dans le Cfa, la ressource essentielle sur les formats de fichiers. D'autres thèmes suivront.

Application du Cfa

Le Cfa n'est pas un schéma unique pour l'archivage numérique. Il s'agit plutôt l'interpréter avec toute la compétence requise et de l'intégrer conjointement avec d'autres prescriptions dans des lignes directrices et recommandations concrètes. La matrice d'évaluation Cfa utilise un nouveau code couleur dans la dernière ligne:

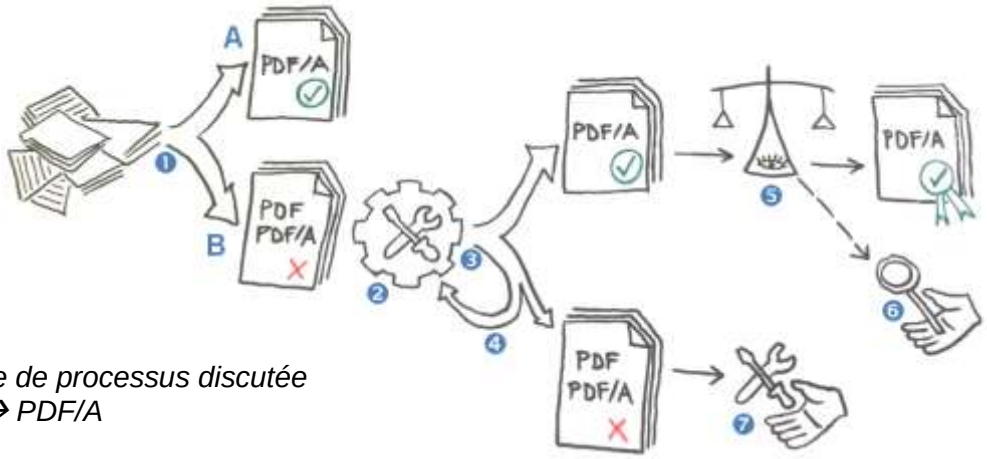
	TXT	PDF	PDF/A-1	PDF/A-2	PDF/A-3	ODF	OOXML	TIFF	JPEG	JPEG2000	PNG	DNG	PDF/A-2	WAV	MP3	Uncompressed Video	Digital Video	MPEG-2	MPEG-4	
Gewichtete Summe	0.81	0.83	0.96	0.92	0.60	0.79	0.71	0.85	0.79	0.81	0.75	0.79	0.85	0.83	0.73	0.65	0.60	0.65	0.63	0
Logarithmisch gewichtete Summe	1.04	1.18	5.38	2.63	0.43	0.93	0.64	1.32	0.91	1.02	0.75	0.91	1.32	1.33	0.79	1.05	0.89	1.05	0.98	

- en vert, les formats qui, sur la base de l'analyse Cfa, peuvent être considérés comme particulièrement adaptés pour l'archivage (évaluation pondérée de 1.0 ou plus). Ils constituent le coeur des recommandations archivistiques. Ils sont particulièrement utiles dans les endroits où les Archives peuvent exercer une influence sur les formats de production dans la phase active du cycle de vie.
- en gris les formats qui ne représentent pas le premier choix pour l'archivage numérique. En raison de leur diffusion, la plupart de ces formats restent toutefois lisibles à court et moyen termes sans effort particulier. Les Archives qui se voient proposer des documents existants dans de tels formats peuvent en règle générale décider de ne pas convertir immédiatement dans des formats d'archivage proprement dit dès lors qu'elles ne suivent pas une politique stricte de gestion des formats.
- en rouge les formats, dont l'évaluation pondérée se situe en dessous de 0.6. Ces formats peuvent être définis explicitement comme non adaptés pour l'archivage à long terme. Il est déconseillé de les prendre en charge.

Cette catégorisation vise surtout à éviter des conversions inutiles et contreproductives.

Conversion automatique de format

Les conversions automatiques de format représentent un nouveau champ de recherche du CECO. Des processus potentiels ainsi qu'une démonstration de faisabilité (*proof of concept*) ont été présentés à l'atelier. Ces conversions automatiques allient validateurs, outils de conversion et instruments pour la synchronisation visuelle automatique en une chaîne de processus. Celle-ci rend la conversion de grandes quantités de documents plus efficace et concentre l'intervention humaine sur les documents qui posent problème.



Chaîne de processus discutée
PDF → PDF/A

Au cours de la discussion durant l'atelier de la fin juin, les participants ont généralement salué le fait que des outils étaient disponibles pour la conversion automatique de format, ils ont cependant émis des doutes sur leur aptitude pour la gestion de masse. Ils ont envisagé qu'il était plus vraisemblable de les utiliser pour des démonstrations ou pour des mandats de plus petite envergure. Le Bureau développera et mettra à disposition des Archives membres son convertisseur CECO avec démonstration de faisabilité afin de pouvoir tester des scénarios d'applications envisageables.

Outils de validation

Différents outils de validation de formats ont été présentés et testés durant l'atelier :

- Jpylyzer, un outil en ligne de commande pour la validation et la caractérisation de fichiers JPEG2000. Jpylyzer sera prochainement intégré dans KOST-Val ;
- MP3val, un outil de validation pour fichiers MP3 ;
- CSV Validator, un outil pour la validation de fichiers CSV par rapport à un schéma rédigé spécialement à cet effet. Les possibilités d'application hors scénario d'application d'origine auprès des Archives nationales anglaises semblent limitées ;
- FITS, File Information Tool Set, une application en ligne de commande qui intègre différents outils pour la reconnaissance de format et la caractérisation de fichiers.

Par ailleurs, la version actuelle de KOST-Val, le validateur de format du CECO, a été présentée et des objectifs pour son développement ultérieur ont été esquissés. Outre l'intégration déjà mentionnée de jpylyzer, le CECO a pour priorité d'implémenter la validation PDF/A duale. Cette dernière constitue une approche pour mieux maîtriser, et de manière plus efficace, la complexité de la spécification PDF/A. La spécification PDF/A est non seulement un vaste système de standards sans délimitation claire, mais certaines dispositions sont formulées de sorte qu'elles peuvent être interprétées différemment par différents outils. C'est pourquoi la validation duale utilise deux validateurs PDF/A (de qualité élevée selon l'étude du CECO) et considère comme valides les documents qui ont été identifiés comme tels par au moins un des validateurs. Le volet pratique de l'atelier a démontré l'utilité d'une telle approche.

Projets CECO potentiels

Nous avons présenté la bourse aux projets du CECO dans la dernière Newsletter. Nous vous remercions pour les propositions de projets déjà remises. Nous donnerons à l'avenir à la page 5 de la Newsletter un aperçu des projets CECO potentiels et des Archives qui y participent.

Mögliche KOST-Projekte / Projets CECO potentiels

Diese Rubrik listet die Projektvorschläge der Trägerarchive auf, welche aus Sicht der Geschäftsstelle am ehesten für ein KOST-Projekt in Frage kommen. Die ausführlichen Projektbeschreibungen sowie die vollständige Liste der Projektvorschläge finden Sie in der KOST-Projektbörse.

Cette rubrique répertorie les propositions de projets des Archives membres qui, du point de vue du Bureau, entrent très vraisemblablement en ligne de compte pour devenir un projet CECO. Vous trouvez la description détaillée du projet ainsi que la liste complète des propositions de projets dans la bourse aux projets du CECO.

ARTS - StAZH	Spezifikation zur Archivierung elektronischer Steuerdossiers *Spécifications pour l'archivage des dossiers fiscaux électroniques*
Einwohnerregister - StadtASG	Archivierung der Einwohnerregister *Archivage des registres des habitants*
GIS - StAZH	Spezifikation zur Archivierung von GIS Daten *Spécifications pour l'archivage des données GIS*
Kompass3 - StAZH	Archivierung der Daten der Berufsbildungsämter *Archivage des données des services de la formation professionnelle*
ViaCar/CARI - StAGR, StAZH	Archivierung der Daten der Strassenverkehrsämter *Archivage des données des services des automobiles*

Archive, die an der Mitarbeit an einem dieser Projekte interessiert sind, werden gebeten, dies der Geschäftsstelle mitzuteilen (info@kost-ceco.ch).

Les Archives intéressées à collaborer a l'un de ces projets sont priées de le communiquer au Bureau (info@kost-ceco.ch).

Veranstaltungshinweise / Calendrier des événements

Nachfolgend Hinweise auf Veranstaltungen von Archiven, Organisationen und Firmen, die für die KOST-Trägerarchive relevant sind und in der Schweiz stattfinden.

Ci-après, le calendrier des événements organisés en Suisse par des archives, organisations et entreprises sur des thèmes importants pour les Archives membres du CECO.

29.01.15 Vorstellung SIK-GIS Konzeptstudie «Nachhaltige Verfügbarkeit und Archivierung von Geodaten»
Présentation de l'étude conceptuelle CSI-SIG "Disponibilité assurée dans la durée et archivage des géodonnées»
Universität Bern
Université de Berne

Wenn Sie einen Veranstaltungshinweis im KOST-Newsletter publizieren wollen, kontaktieren Sie uns bitte unter info@kost-ceco.ch.

Si vous souhaitez publier un événement dans le calendrier de la Newsletter du CECO, veuillez s.v.pl. nous contacter à l'adresse info@kost-ceco.ch.