

KOST-Newsletter Quartal 2, 2012

Formaterkennung und Formatvalidierung

Archivtaugliche Formate – und dann?

In der Praxis reicht es nicht aus, dass ein Archiv definiert hat, welche Formate es für die digitale Archivierung akzeptiert. Vielmehr muss bei jedem Ingest überprüft werden, ob diese Vorgaben tatsächlich eingehalten werden; sonst handelt sich das Archiv womöglich ein Risiko für die Bestandserhaltung ein. Diese Überprüfung vollzieht sich in zwei Schritten: zunächst in einer automatischen Formaterkennung, danach in der Validierung der gesamten Datei.



Was bedeutet das?

Die Formaterkennung identifiziert das Format einer Datei bis zu einer bestimmten, gewünschten Granularität. Sie stützt sich dabei auf das Vorhandensein besonders charakteristischer Eigenschaften; in der Regel sind dies bestimmte Byte-Sequenzen innerhalb der Datei. Die Formatvalidierung überprüft, ob eine Datei der Spezifikation ihres Formats entspricht. Dabei muss jede einzelne der in der Formatspezifikation verlangten Eigenschaften überprüft werden. Nur wenn alle Eigenschaften erfüllt sind, ist die Datei valide.

Für die Formaterkennung existieren mehrere, gut etablierte Datenbanken und Werkzeuge. Die Formatvalidierung ist technisch ungleich komplexer; entsprechend ist hier das Toolangebot beschränkt, insbesondere für Formate ausserhalb des Mainstreams.

Was tut die KOST?

Formaterkennung und -validierung sind in letzter Zeit verstärkt in den Fokus der Archive gerückt. Die KOST befasst sich seit zwei Jahren intensiv mit dem Thema.

- Grundlagen der Formaterkennung und -validierung sind in einer neuen Studie der KOST dargelegt, welche soeben auf der [KOST-Website](#) publiziert wurde. Die Studie untersucht auch verfügbare Datenbanken (z.B. PRONOM) und Werkzeuge im Bereich der Formaterkennung. Sie beruht auf dem Beitrag von Claire Röthlisberger zur diesjährigen Tagung des deutschen Arbeitskreises „Archivierung von Unterlagen aus digitalen Systemen“.
- Besonders reich ist die Tool-Landschaft im Bereich der Validierung von PDF/A-Dateien. Dieses Thema hat die KOST bereits vor einiger Zeit ausführlich untersucht und die Resultate in einer Studie präsentiert, die auf der [KOST-Website](#) erhältlich ist. Diese Untersuchungen sind auch die Basis für Beratungen durch die KOST-Geschäftsstelle.
- Im Bereich der Werkzeuge zur Formatvalidierung gibt es noch empfindliche Lücken, welche durch die Archivwelt geschlossen werden sollten. Nach ihren Erfahrungen mit dem Validator für SIPs (SIP-Val) Ende 2010 hat die KOST nun die Entwicklung eines Validators für SIARD-Dateien – *SIARD-Val* – in Angriff genommen (siehe Newsletter 4/2011). Wir freuen uns, *SIARD-Val* an einem Workshop am Donnerstag, 18.10.2012 den KOST-Trägern präsentieren zu können. Bitte reservieren Sie sich bereits das Datum, eine Einladung folgt rechtzeitig.

Weitere Aktivitäten der KOST

KOST.Forum no3: Digitale Archivierung im Staatsarchiv Genf

Die KOST freut sich, ihre Mitglieder nach über zwei Jahren Pause wieder zu einem KOST.Forum einladen zu können. Diese Veranstaltungsreihe hatten wir 2009 aus der Taufe gehoben, um den Trägerarchiven der KOST die Möglichkeit zu geben, im Kollegenkreis über Resultate und Erkenntnisse ihrer Arbeit zur digitalen Archivierung zu berichten.

Im Zentrum des dritten KOST.Forums steht das Staatsarchiv Genf. Dieses hat im Verlauf des letzten Jahres in einem intensiven Prozess strategische Grundlagen zur digitalen Archivierung erarbeitet und umgesetzt. Die Implementierung spielt sich in einer wegweisenden Zusammenarbeit mit dem Schweizerischen Bundesarchiv ab, das in einem Pilotprojekt seine Archivierungsplattform als Dienstleistung zur Verfügung stellt.

Bitte reservieren Sie sich jetzt schon den Montag, 24. September 2012. Eine Einladung verschicken wir im August.

eCH: Meilensteine und neue Themen

Genau ein Jahr nach ihrer Gründung hat die [eCH-Fachgruppe Digitale Archivierung](#) den ersten Standardentwurf verabschiedet und wird ihn in den nächsten Tagen beim Expertenausschuss einreichen. eCH-0160 „Archivische Ablieferungsschnittstelle (SIP)“ beruht inhaltlich auf der SIP-Spezifikation des Schweizerischen Bundesarchivs, Version 4.0, in welche bereits Anregungen aus dem KOST-Projekt *bentō* eingeflossen sind. Der Standardentwurf wird als nächstes von den eCH-Referenten geprüft und danach in die öffentliche Vernehmlassung gegeben, an der sich alle Interessierten beteiligen können. Die Publikation von eCH-0160 wird im Spätherbst erwartet.

Die Fachgruppe hat zugleich ein neues Standardisierungsthema beschlossen und das Mandat einer bestehenden Themengruppe ausgeweitet. Neu in Angriff genommen wird die Standardisierung des SIARD-Formats für die Archivierung relationaler Datenbanken. Die Themengruppe, die sich um die Definition eines Austauschformats für Archivpakete kümmert, wird sich neu zudem mit der Metadaten-Schnittstelle aus dem Ingest-Prozess zum Archivinformationssystem befassen.

Katalog archivischer Dateiformate, Version 3 (KaD_v3)

Die neu gegründete *Expertengruppe Dateiformate* hat die Überarbeitung des KaD begonnen. Neben einer allgemeinen Revision der Inhalte und Bewertungen wird sie insbesondere in den Bereichen Text-, Bild- und Videodaten zusätzliche Grundlagen erarbeiten. Zudem wird sie einen Mechanismus zum *Preservation Planning* entwickeln.

Für dieses Projekt arbeitet die KOST mit anderen Akteuren und Arbeitsgruppen zusammen, insbesondere mit den VSA-Arbeitsgruppen *Records Management & Digitale Archivierung* und *Normen & Standards*, mit der eCH-Fachgruppe *Digitale Archivierung*, mit *Memo-riav* und mit dem Schweizerischen Tanzarchiv. Angestrebt wird eine Verabschiedung von KaD_v3 in der Frühjahrssitzung 2013 der KOST-Aufsichtskommission.

Deutsche Übersetzung des OAIS

Eine Arbeitsgruppe von *nestor*, dem deutschen Kompetenznetzwerk Langzeitarchivierung, hat eine [deutsche Übersetzung des Open Archival Information System OAIS](#) erarbeitet, des zentralen Standards für die digitale Archivierung. Die KOST beteiligte sich an diesem Projekt. Durch die genaue terminologische Arbeit wird die Verständlichkeit nicht nur der Hauptpunkte, sondern auch der weniger bekannten Konzepte des OAIS erleichtert.

Newsletter CECO du 2e trimestre 2012

Reconnaissance et validation de format

Des formats adaptés pour l'archivage - et après ?

Dans la pratique, les Archives ne peuvent se contenter de définir quels formats elles acceptent pour l'archivage numérique. Elles doivent en outre vérifier, lors de chaque entrée, si ces directives sont bien observées, faute de quoi elles pourraient prendre des risques pour la conservation. Cette vérification se fait en deux étapes: tout d'abord une reconnaissance automatique de format puis une validation du fichier dans son ensemble.



Qu'est-ce que cela signifie ?

La reconnaissance de format identifie le format d'un fichier jusqu'à la granularité désirée. Pour ce faire, elle s'appuie sur l'existence de caractéristiques particulières. Il s'agit généralement de certaines séquences d'octets à l'intérieur d'un fichier. La validation du format vérifie si un fichier correspond bien à la spécification de son format. Il faut vérifier chaque caractéristique exigée dans la spécification du format. Le fichier n'est valide que si toutes les spécifications sont satisfaites.

Il existe plusieurs bases de données et outils bien rodés pour la reconnaissance de format. La validation de format est, quant à elle, beaucoup plus complexe d'un point de vue technique. L'offre d'outils est donc plus limitée, en particulier pour les formats moins courants.

Que fait le CECO ?

Ces derniers temps, les Archives s'intéressent de plus en plus à la reconnaissance et à la validation de formats. Le CECO s'occupe intensément de ce sujet depuis deux ans.

- Les bases de la reconnaissance et de la validation de formats figurent dans une nouvelle étude du CECO qui vient d'être publiée sur le [site du CECO](#). L'étude examine aussi les bases de données (par exemple PRONOM) et outils disponibles dans le domaine de la reconnaissance de format. Elle repose sur la contribution de Claire Röthlisberger à la journée thématique de cette année de la communauté de travail allemande "Archivierung von Unterlagen aus digitalen Systemen".
- La palette d'outils dans le domaine de la validation de fichiers PDF/A est particulièrement riche. Le CECO a déjà analysé ce thème en détail il y a quelque temps et présenté les résultats dans une étude disponible sur le [site du CECO](#). Ces analyses servent également de base aux conseils fournis par le bureau du CECO.
- Dans le domaine des outils de validation des formats, il existe encore des lacunes importantes que le monde des Archives devra combler. Les expériences des Archives avec le validateur pour SIP (*SIP-Val*) à la fin 2010 ont permis au CECO de s'attaquer au développement d'un validateur pour fichiers SIARD – *SIARD-Val* – (voir newsletter 4/2011). Nous nous réjouissons de pouvoir présenter aux membres du CECO SIARD-Val lors d'un atelier qui aura lieu le jeudi 18.10.2012. Nous vous prions de réserver d'ores et déjà cette date, une invitation suivra en temps utile.

Autres activités du CECO

CECO.Forum N°3 : l'archivage numérique aux Archives d'État de Genève

Après une pause de plus de deux ans, le CECO a le plaisir de pouvoir convier à nouveau ses membres à un CECO.Forum. Nous avons inauguré cette série en 2009 dans le but de donner aux archives membres du CECO la possibilité de présenter à leurs homologues les résultats et connaissances issus de leur travail sur l'archivage numérique.

Les Archives d'État de Genève se tiennent au coeur du troisième CECO.Forum. Dans le courant de l'année dernière, elles ont élaboré et mis sur pied, au cours d'un processus intense, des bases stratégiques pour l'archivage numérique. L'implémentation se déroule à la faveur d'une collaboration novatrice avec les Archives fédérales suisses, qui mettent à disposition, dans le cadre d'un projet pilote, leur plateforme d'archivage.

Veuillez réserver d'ores et déjà le lundi 24 septembre 2012. Une invitation sera envoyée en août.

eCH: quelques jalons et de nouveaux thèmes

Un an exactement après sa création, le groupe spécialisé *archivage numérique eCH* a adopté son premier projet de standard et va le soumettre ces prochains jours au comité d'experts. Le contenu d'eCH-0160 "Interface de versement aux Archives (SIP)" repose sur la spécification SIP des Archives fédérales suisses, version 4.0, qui s'inspire déjà de suggestions issues du projet CECO *bentō*. Le projet de standard sera ensuite examiné par les experts eCH puis envoyé en consultation publique. Toute personne intéressée peut participer à cette consultation. La publication de eCH-0160 est prévue pour la fin de l'automne.

Parallèlement, le groupe spécialisé a décidé d'un nouveau sujet de standardisation et étendu le mandat d'un groupe de travail existant. Le groupe s'attèle à la standardisation du format SIARD pour l'archivage des bases de données relationnelles. Le groupe de travail chargé de définir un format d'échange pour paquets d'échange d'informations s'occupera désormais en plus de l'interface de métadonnées entre le processus *Entrées* et le logiciel de gestion d'archives.

Catalogue des formats de données d'archivage, version 3 (Cfa_v3)

Le groupe d'experts *Formats de fichiers*, créé récemment, a entamé l'élaboration du Cfa. En plus de procéder à une révision générale des contenus et à une évaluation, il élaborera en particulier des bases dans les domaines des données textuelles, graphiques et vidéo. En outre, il développera un processus de planification de la pérennisation (*preservation planning*).

Pour ce projet, le CECO travaille en collaboration avec d'autres acteurs et groupes de travail, en particulier ceux de l'AAS, *Records Management & archivage électronique* et *Pour les normes et standards*, le groupe spécialisé *archivage numérique eCH*, *Memo-riav* ainsi qu'avec la Collection suisse de la danse. Le Cfa_v3 devrait être adopté lors de la séance de la commission de surveillance du CECO du début 2013.

L'OAIS traduit en allemand

Un groupe de travail de *nestor*, le réseau de compétences allemand pour l'archivage à long terme, a élaboré une [traduction allemande de l'Open Archival Information System OAIS](#), le modèle de référence standard pour l'archivage numérique. Le CECO a participé à ce projet. Le travail terminologique précis facilite la compréhension non seulement des points principaux de l'OAIS, mais également de ses concepts moins connus.

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.

24.09.12 KOST.Forum no3:
Digitale Archivierung im Staatsarchiv Genf
CECO.Forum no3
L'archivage numérique aux Archives de l'Etat de Genève
Schweizerisches Bundesarchiv, Bern

18.10.12 Workshop SIARD-Val
Atelier SIARD-Val
Schweizerisches Bundesarchiv, Bern

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.