KOST-Newsletter Quartal 4, 2013

KOST-Val

Formaterkennung und Formatvalidierung

Um Gewissheit zu erlangen über die Formate der Dateien, die es übernimmt, muss ein Archiv diese erkennen und validieren. Die Formaterkennung identifiziert das Format einer Datei bis zu einer bestimmten, gewünschten Granularität. Die Formatvalidierung ist viel aufwändiger: Sie überprüft, ob eine Datei der Spezifikation ihres Formats entspricht, validiert also jede einzelne der in der Formatspezifikation verlangten Eigenschaften.

Grundsätzlich existieren zwei Arten von Formatvalidatoren: Generische Validatoren (wie beispielsweise JHOVE) nehmen eine Formaterkennung vor und validieren Dateien gegenüber dem dabei erkannten Format. Spezifische Validatoren (wie beispielsweise PDFTron) prüfen, ob eine Datei einem vom Benutzer vorgegebenen Format entspricht. In der Regel sind die spezifischen Validatoren genauer in der Formatanalyse.

Die zentrale Bedeutung der Formaterkennung und -validierung für die digitale Archivierung ist in einer KOST-Studie im Detail beschrieben und wurde bereits im Newsletter 2012/2 vorgestellt.

Formatvalidierung im Archiv

Die praktische Anwendung der Formatvalidierung im Archiv sieht sich mit zwei wesentlichen Problemen konfrontiert:

- Den Arbeitsschritt der Formatvalidierung im Ingest-Prozess mit einer Vielzahl von Validatoren bestreiten zu müssen, ist langfristig keine akzeptable Lösung. Mehrere Validatoren bedeuten mehrfachen Tool-Unterhalt (Installationen, Updates), unterschiedliche Bedienungen, mehrfachen Aufwand für die Validierung und verschiedenartigen Output.
- Als archivtauglich kann in vielen Fällen nicht einfach ein generisches Format gelten, sondern nur eine bestimmte Version oder ein bestimmtes Profil eines Formats. Ein Archiv kann beispielsweise beschliessen, nicht einfach PDF zu verlangen, sondern PDF/A-1b oder -2b, oder nicht einfach TIFF, sondern "Baseline TIFF V6". Viele existierenden Formatvalidatoren können diesen Detaillierungsgrad nicht liefern, da sie nur generische Formate validieren.

KOST-Val

Der Formatvalidator KOST-Val ist die Lösung, welche die KOST für diese Probleme anbietet. Als integriertes, modulares und konfigurierbares Tool erlaubt er die Validierung verschiedener Formate und Profile in einem Arbeitsgang. Er bindet Validatoren aus anderen Quellen ein und verfeinert bei Bedarf ihren Output. Für Formate, für die kein oder kein genügender Validator existiert, bietet KOST-Val eigene Validierungsmodule an. Aktuell können die Formate PDF/A, TIFF und SIARD validiert werden; weitere Module sind in Planung. KOST-Val ist eine java-basierte Konsolenanwendung und liefert einen ausführlichen Validierungslog. Das Ergebnis der Gesamtvalidierung wird ebenfalls ausgegeben und im *exit*-Status des Programms sichtbar, so dass die Validierung in eine automatisierte Verarbeitungskette eingebunden werden kann.

Die bisherigen Standalone-Validatoren SIARD-Val und TIFF-Val gehen in KOST-Val auf und werden nicht mehr weiterentwickelt.

Und jetzt?

KOST-Val steht unter der Open-Source-Lizenz GPL 3.0 und kann von der KOST-Website unter http://kost-ceco.ch/cms/index.php?kost_val_de heruntergeladen werden. Entwicklungsplattform ist das Open-Source-Repository GitHub (https://github.com/KOST-CECO/KOST-Val). Weitere Dateiformate werden laufend hinzugefügt; entsprechende Wünsche können bei GitHub registriert (https://github.com/KOST-CECO/KOST-Val/issues) oder direkt an claire.roethlisberger@kost.admin.ch gemeldet werden.

COPTR

Als neue, zentrale Plattform für Formatvalidatoren und alle anderen Arten von Tools für die digitale Archivierung hat das britische SPRUCE-Projekt kürzlich das *Community Owned digital Preservation Tool Registry COPTR* vorgestellt. Die Validatoren der KOST sind auf COPTR ebenfalls dokumentiert.

Weitere Aktivitäten der KOST

arcun

Die Staatsarchive von Schaffhausen und Freiburg haben den produktiven Betrieb von *arcun* aufgenommen. Sie stossen damit zum Staatsarchiv Uri und zum Stadtarchiv Luzern, die bereits seit einigen Jahren digitales Archivgut auf *arcun* speichern.

Die KOST-Software *arcunTAR*, mit welcher die Archive ihr *arcun*-Konto verwalten, liegt neu in der Version 4.2 vor. Diese Version kann Verzeichnisse übertragen, ohne dass für jede bereits übertragene Datei ein Hashwert berechnet werden muss. Dadurch wird die Suche nach neuen Dateien in einem grösseren Verzeichnisbaum etwa um den Faktor 10 schneller als bis anhin. Dieser Zeitgewinn kann für Archive, die umfangreiche Datenbestände auf *arcun* verwalten, essentiell sein.

Preservation Planning

Die im letzten Newsletter angekündigte Preservation-Planning-Expertengruppe hat sich konstituiert und ein erstes Mal getagt. Erstes Thema waren die Empfehlungen zum Format TIFF, basierend auf der entsprechenden KOST-Studie. Die Diskussionen und Entscheide der Gruppe werden insbesondere zu präziseren Analysen und konkreteren Hilfestellungen im Katalog archivischer Dateiformate führen.

Die KOST dankt den Kolleginnen und Kollegen von den Staatsarchiven BS, BL, GE, NE, SG, UR und ZH, für ihr Engagement in dieser Expertengruppe.

Beratung von Trägerarchiven

In den letzten Monaten hat die KOST-Geschäftsstelle mit den folgenden Trägerarchiven beratend oder begleitend zusammengearbeitet:

- Staatsarchiv Neuenburg: Präsentation der und Gedankenaustausch über die Projekte in den Bereichen Records Management und digitale Archivierung (AENeas).
- Staatsarchiv Zürich: Evaluation einer Infrastruktur zur digitalen Archivierung.
- Staatsarchiv Basel-Landschaft: Revision des digitalen Langzeitarchivs SILO1 anhand der Minimalanforderungen an die digitale Archivierung.
- Stadtarchiv Zürich: Beratende Mitarbeit im Projekt „Richtlinien Records Management für die Stadtverwaltung Zürich“

Newsletter CECO du 4e trimestre 2013

KOST-Val

Reconnaissance et validation de format

Pour en avoir le cœur net sur les formats des fichiers qu'elles prennent en charge, les Archives doivent les reconnaître et les valider. La reconnaissance de format identifie le format d'un fichier jusqu'à la granularité souhaitée. La validation de format est beaucoup plus sophistiquée. En effet, elle vérifie si un fichier correspond à la spécification de son format et donc valide séparément chacune des caractéristiques exigées dans la spécification du format.

En principe, il existe deux sortes de validateurs de format. Les validateurs génériques (par exemple JHOVE) effectuent une reconnaissance de format et valident des fichiers par rapport au format ainsi reconnu. Les validateurs spécifiques (par exemple PDFTron) vérifient si un fichier correspond à un format donné par l'utilisateur. En général, les validateurs spécifiques fournissent une analyse de format plus précise.

L'importance capitale de la reconnaissance et de la validation de format pour l'archivage numérique est décrite en détail dans une étude du CECO et a déjà été présentée dans la Newsletter 2/2012.

Validation de format dans les Archives

L'application pratique de la validation de format dans les Archives se voit confrontée à deux problèmes essentiels:

- Devoir effectuer l'étape de la validation de format dans le processus d'entrées avec une multitude de validateurs n'est pas une solution acceptable à long terme. Plusieurs validateurs signifient l'entretien de multiples outils (installations, mises à jour), diverses manipulations, beaucoup d'efforts pour la validation et des résultats hétéroclites.
- Dans de nombreux cas, un format générique ne convient pas à l'archivage, mais seulement une version déterminée ou un certain profil d'un format. Par exemple, des Archives peuvent décider d'exiger non seulement le PDF, mais le PDF/A-1b ou -2b, ou non seulement le simple format TIFF, mais le «Baseline TIFF V6». Beaucoup de validateurs de format ne peuvent pas fournir ce degré de détail puisqu'ils ne valident que des formats génériques.

KOST-Val

Le validateur de format KOST-Val est la solution proposée par le CECO pour ces problèmes. Cet outil intégré, modulaire et configurable permet de valider différents formats et profils en une seule étape. Il intègre des validateurs provenant d'autres sources et affine au besoin leurs résultats. Pour les formats pour lesquels il n'existe pas de validateur ou pas de validateur suffisamment performant, KOST-Val offre ses propres modules de validation. Il peut actuellement valider les formats PDF/A, TIFF et SIARD. D'autres modules sont en préparation. KOST-Val est une application console basée sur java et il fournit un journal de validation détaillé. Il fournit également le résultat de la validation globale et l'affiche dans le statut *exit* du logiciel, si bien que la validation peut être intégrée dans une chaîne de traitement automatisée.

Les anciens validateurs autonomes SIARD-Val et TIFF-Val sont intégrés dans KOST-Val et ne seront plus développés.

Et ensuite?

KOST-Val est sous licence *open source* GPL 3.0 et peut être téléchargé sur le site du CECO: http://kost-ceco.ch/cms/index.php?kost_val_fr. La plateforme de développement est le dépôt *open source* GitHub (https://github.com/KOST-CECO/KOST-Val). D'autres formats de fichiers sont ajoutés régulièrement. Des souhaits à ce propos peuvent être enregistrés auprès de GitHub (https://github.com/KOST-CECO/KOST-Val/issues) ou communiqués directement à claire.roethlisberger@kost.admin.ch

COPTR

Le projet britannique SPRUCE a récemment présenté le *Community Owned digital Preservation Tool Registry COPTR* , une nouvelle plateforme centrale pour validateurs de format et toutes sortes d'outils pour l'archivage numérique. Les validateurs du CECO sont également documentés sur COPTR

Autres activités du CECO

arcun

Les Archives d'État de Schaffhouse et Fribourg ont débuté l'exploitation d'*arcun*. Elles rejoignent ainsi les Archives d'État d'Uri et les Archives de la Ville de Lucerne qui stockent des archives numériques sur *arcun* depuis quelques années déjà.

Le logiciel du CECO *arcunTAR*, avec lequel les Archives gèrent leur compte *arcun*, est disponible en version 4.2. Cette version transfère des répertoires sans qu'il soit nécessaire de calculer une valeur de hachage pour chaque fichier déjà transféré. Par rapport à ce qu'on avait jusqu'ici, cela permet d'augmenter d'environ dix fois la rapidité de recherche de nouveaux fichiers dans une arborescence plus grande. Ce gain de temps est essentiel pour des Archives gérant des fonds de données de grande ampleur.

Preservation Planning

Le groupe d'experts *Preservation Planning* annoncé dans la dernière Newsletter s'est constitué et a siégé pour la première fois. Le premier thème abordé a été les recommandations sur le format TIFF, basées sur l'étude du CECO consacrée à ce sujet. Les discussions et décisions du groupe déboucheront notamment sur des analyses plus précises et des aides dans le Catalogue de formats de données.

Le CECO remercie les collègues des Archives d'État de BS, BL, GE, NE, SG, UR et ZH pour leur engagement dans ce groupe d'experts.

Activités de conseil auprès des Archives membres

Ces derniers mois, le Bureau du CECO a conseillé les Archives suivantes dans le cadre de leurs projets:

- Archives de l'État de Neuchâtel: Présentation des projets dans le domaine de la gestion des documents (*records management*) et de l'archivage numérique (AENeas) et échange d'idées.
- Archives d'État de Zurich: Évaluation d'une infrastructure pour l'archivage numérique.
- Archives d'Etat de Bâle-Campagne: Révision des archives numériques à long terme SILO1 en se fondant sur les exigences de base pour l'archivage numérique.
- Archives de la Ville de Zurich: Collaboration au travers de conseils dans le projet «Richtlinien Records Management für die Stadtverwaltung Zürich»