KOST-Newsletter Quartal 3, 2017

KOST-Studie PDF/A-Validatoren

Die KOST-Träger wissen schon lange: Archivtaugliche Dateiformate einzufordern, ohne ihre Einhaltung zu überprüfen, ist nicht einmal die halbe Miete. Deshalb beschränkt sich die KOST nicht darauf, ihren Katalog archivischer Dateiformate KaD laufend zu aktualisieren, sondern beschäftigt sich seit Jahren intensiv mit der Formatvalidierung. Grundsätzliche Überlegungen, wie sie in der Studie „Formaterkennung und Formatvalidierung: Theorie und Praxis“ von 2012 angestellt wurden, bilden das theoretische Fundament, das im Multi-Format-Validator KOST-Val konkretisiert wird. Gegenwärtig validiert KOST-Val Dateien in den Formaten TIFF, SIARD, PDF/A, JPEG2000 und JPEG sowie SIP nach eCH-0160.

Von besonderer Bedeutung ist die Validierung bei dem für die digitale Archivierung zentralen Format PDF/A. Aus diesem Grund hat die KOST bereits 2010 sieben Validatoren für PDF/A getestet und analysiert und die Resultate als Studie veröffentlicht. Nach sieben Jahren schien eine komplette Überarbeitung dieser Untersuchung angebracht. Wir freuen uns, Ihnen nun die neue KOST-Studie zu PDF/A-Validatoren vorstellen zu können. Sie umfasst noch vier auf dem Markt wichtige Produkte, darunter neu den im Rahmen des PREFORMA-Projekts entwickelten Open-Source-Validator veraPDF.

Eine vollständige Analyse von PDF/A-Validatoren erfordert ein umfassendes, korrektes und möglichst unpubliziertes Testset von Dateien, die die separate Analyse sämtlicher interessierender Anforderungen erlauben. Ein solches Testset stand 2010 mit der „Bavaria-Testsuite“ zur Verfügung. Diese ist jedoch inzwischen nicht nur nicht mehr aktuell, sondern auch zu gut bekannt. Ein neues Testset zu erarbeiten, übersteigt die gegenwärtigen Möglichkeiten der KOST. Deshalb beruht die Neuauflage der PDF/A-Validatoren-Studie auf einem Validatoren-Benchmarking. Dafür wurde ein Testset von 2980 verschiedenen realen PDF-Dateien aus dem Umfeld der KOST und der nestor-Arbeitsgruppe Formaterkennung verwendet, die allen untersuchten Validatoren als Input dienten.

Die folgenden Aspekte wurden festgehalten (neben der Version, dem Tester und dem Testzeitpunkt):

- Kosten: Preis des Produkts inklusive Wartungsvertrag für ein Jahr.
- Geschwindigkeit: Dauer der kompletten Validierung in der Testumgebung.
- Robustheit: Anzahl der unkontrollierten Ausgaben im Lauf der kompletten Validierung.
- Einigkeit: In 82.58% der Fälle waren sich alle Validatoren einig, und bei lediglich 3.52% der Fälle gab es kein eindeutiges Resultat (2 gegen 2). Festgehalten wurde die Abweichung von der Mehrzahl der anderen Validatoren beim Testergebnis valid oder invalid. Angegeben wurde zusätzlich, wie sich die Abweichungen auf valide und invalide Dokumente verteilen.
- Genauigkeit: Da sich ein automatisiertes Mapping der Fehlermeldungen als nicht realisierbar herausgestellt hat, wurde eine manuelle Qualitätskontrolle über 30 Testdateien durchgeführt, welche alle die Validierung nicht bestanden hatten. Festgehalten wurde der Prozentsatz der übereinstimmenden Fehlermeldungen.

Die Analyseergebnisse sind in der folgenden Tabelle zusammengestellt. Bitte beachten Sie, dass von veraPDF zwei verschiedene Versionen getestet wurden.

PDF/A Validatoren 2017	Callas: pdfaPilot	PDF Tools: 3Heights PDF Validator	PDFTron: PDF/A Manager	veraPDF	veraPDF
Kosten	Teuer EUR 5'399.-	Gering CHF422.-	Mässig USD 699.-	Gratis CHF 0.00	Gratis CHF 0.00
Geschwindigkeit	Gut 1:58:50	Sehr gut 0:18:00	Sehr gut 0:13:38	Mangelhaft 8:30:20	Mangelhaft 9:22:45
Robustheit	Sehr gut 0	Sehr gut 1	Sehr gut 3	Mangelhaft 49	Mangelhaft 55
Einigkeit **Total Abweichung**	Gut 2.38%	Sehr gut 0.87%	Ausreichend 5.74%	Gut 4.90%	Mangelhaft 19.53%
Rest Valid **Rest Invalid**	0.00% 2.38%	0.81% 0.07%	5.00% 0.74%	2.28% 2.62%	10.23% 09.30%
Genauigkeit	Gut 80.89%	Sehr gut 92.67%	Gut 82.11%	Gut 86.44%	Mangelhaft 26.56%
Getestete Version	CLI v7.0.267	Shell v4.9.20.0	v6.7152209	v1.6.1	v1.8.4
Tester	KOST	KOST	KOST	KOST	KOST
Testzeitpunkt	Juni 2017	Juni 2017	Juni 2017	Juni 2017	August 2017
Bemerkungen	pdfaPilot ist ein Konverter und kostet entsprechend viel. Einigkeit und Genauigkeit wären deutlich besser mit der Option „N-Eintrag im OutputIntent prüfen“.		PDF/A Manager ist auch ein PDF zu PDF/A Konverter. Die Kosten stammen aus dem Jahr 2010.		Einigkeit und Genauigkeit haben sich in der Version 1.8.4 massiv verschlechtert. Von den 30 invaliden PDF-Dateien wurden 18 als valide ausgegeben.

Die Ergebnisse des Benchmarking haben die KOST dazu veranlasst, in KOST-Val PDFTron durch pdfaPilot zu ersetzen (neben dem zweiten Validator 3Heights PDF Validator). Diese Ersetzung wird gegenwärtig durchgeführt. Der neue Open-Source-Validator veraPDF vermag die in ihn gesetzten Erwartungen aus Sicht der KOST momentan leider noch nicht zu erfüllen.

Change Requests für xIsadg

Im letzten Newsletter haben wir Sie über die Publikation von xIsadg 2.0, der Schnittstelle zwischen Ingest und Data Management, informiert. Um die Weiterentwicklung von xIsadg planen und steuern zu können, sammeln und publizieren wir entsprechende Change Requests auf der KOST-Website unter https://kost-ceco.ch/cms/index.php?cr_xisadg_de.

Newsletter CECO du 3e trimestre 2017

Étude du CECO : Validation PDF/A

Les membres du CECO le savent depuis longtemps : réclamer des formats adaptés pour l'archivage sans en vérifier la conformité ne représente que la moitié du travail. C'est la raison pour laquelle le CECO ne se contente pas de mettre à jour continuellement son Catalogue des formats de données d'archivage Cfa, mais qu'il s'active depuis des années à la validation de formats. Les réflexions de base comme celles développées dans l'étude « Reconnaissance et validation de format - Théorie et pratique » de 2012 constituent le fondement théorique qui trouve sa concrétisation dans le validateur multiformat KOST-Val. Ce dernier valide actuellement des fichiers dans les formats TIFF, SIARD, PDF/A, JPEG2000 et JPEG ainsi que des SIP conformément à eCH-0160.

La validation du PDF/A, format essentiel pour l'archivage numérique, revêt une importance particulière. C'est pour cette raison que le CECO a testé en 2010 déjà sept validateurs et analysé les résultats pour les publier dans une étude. Après sept ans, il semble que le moment est venu de remanier entièrement cette analyse. Nous avons le plaisir de vous présenter la nouvelle étude du CECO sur les validateurs de PDF/A. Cette étude comprend encore quatre produits importants du marché parmi lesquels le nouveau validateur open source veraPDF développé dans le cadre du projet PREFORMA.

Une analyse complète de validateurs de PDF/A nécessite un corpus de test global, correct et si possible non publié de fichiers qui permettent une analyse séparée de toutes les exigences qui nous intéressent. L'ensemble de test Bavaria a mis un tel corpus de test à disposition en 2010. Cet outil est entre-temps devenu non seulement obsolète, mais il est en plus également trop bien connu. Développer un nouveau corpus de test dépasse les possibilités actuelles du CECO, raison pour laquelle la nouvelle version de l'étude sur les validateurs de PDF/A se base sur un banc d'essai de validateurs (*benchmarking*). Nous avons utilisé un corpus de 2980 fichiers PDF réels différents provenant de l'environnement du CECO et du groupe de travail nestor sur la reconnaissance des formats qui ont servi à tester tous les validateurs analysés.

Les aspects suivants ont été consignés (en plus de la version, du testeur et de la date du test) :

- Prix : prix du produit y compris contrat de maintenance pour un an.
- Vitesse : durée de la validation complète dans l'environnement de test.
- Robustesse : nombre de sorties incontrôlées au cours d'une validation complète.
- Consensus : dans 82.58% des cas, tous les validateurs étaient d'accord et dans seulement 3.52% des cas il n'y avait pas de résultat net (2 contre 2). L'écart par rapport à la majorité des autres validateurs a été consigné dans le résultat du test qu'il soit valide ou invalide. Le résultat indique en plus la répartition entre documents valides et invalides.
- Précision : étant donné qu'il s'est avéré impossible de réaliser un mappage automatisé des messages d'erreur, nous avons effectué un contrôle manuel de la qualité sur plus de 30 fichiers du test qui n'avaient pas réussi la validation. Nous avons consigné le pourcentage des messages d'erreur concordants.

Les résultats de l'analyse sont rassemblés dans le tableau ci-après. Veuillez noter que nous avons testé deux versions différentes de veraPDF.

Validateurs PDF/A 2017	Callas: pdfaPilot	PDF Tools: 3Heights PDF Validator	PDFTron: PDF/A Manager	veraPDF	veraPDF
Prix	Élevé EUR 5'399.-	Bas CHF422.-	Modéré USD 699.-	Gratuit CHF 0.00	Gratuit CHF 0.00
Vitesse	Bien 1:58:50	Très bien 0:18:00	Très bien 0:13:38	Insuffisant 8:30:20	Insuffisant 9:22:45
Robustesse	Très bien 0	Très bien 1	Très bien 3	Insuffisant 49	Insuffisant 55
Consensus **Écart total**	Bien 2.38%	Très bien 0.87%	Suffisant 5.74%	Bien 4.90%	Insuffisant 19.53%
Reste valide **Reste invalide**	0.00% 2.38%	0.81% 0.07%	5.00% 0.74%	2.28% 2.62%	10.23% 09.30%
Précision	Bien 80.89%	Très bien 92.67%	Bien 82.11%	Bien 86.44%	Insuffisant 26.56%
Version testée	CLI v7.0.267	Shell v4.9.20.0	v6.7152209	v1.6.1	v1.8.4
Testeur	CECO	CECO	CECO	CECO	CECO
Date du test	Juin 2017	Juin 2017	Juin 2017	Juin 2017	Août 2017
Remarques	pdfaPilot est un convertisseur ce qui explique son prix élevé. Le consensus et la précision seraient nettement meilleurs avec l'option « contrôler information N dans OutputIntent) ».		PDF/A Manager est aussi un convertisseur de PDF en PDF/A Les prix datent de 2010		Le consensus et la précision se sont massivement détériorés dans la version 1.8.4. 18 des 30 fichiers PDF invalides sont passés pour valides.

Les résultats du banc d'essai ont entraîné le remplacement de PDFTron par pdfaPilot dans KOST-Val (en plus du deuxième validateur 3Heights PDF Validator). Ce remplacement est effectué actuellement. Du point de vue du CECO, le nouveau validateur open source veraPDF ne peut pour l'instant pas encore répondre aux attentes placées en lui.

Change requests pour xlsadg

Nous vous avons informé dans la dernière newsletter sur la publication de xlsadg 2.0, l'interface entre les entités *entrées* et *gestion des données*. Afin de planifier et piloter la suite du développement de xlsadg, nous recueillons et publions sur le site internet du CECO des demandes de modification à son sujet (change requests) : https://kost-ceco.ch/cms/index.php?cr_xisadg_de.

Potentielle KOST-Projekte / Projets potentiels du CECO

Die Liste der potentiellen KOST-Projekte listet Projektvorschläge auf, die für eine Betreuung als KOST-Projekt in Frage kommen, und für die weitere Partner gesucht werden.
La liste des projets potentiels du CECO répertorie les propositions de projets des archives membres qui entrent en ligne de compte pour devenir un projet CECO et pour lesquels des partenaires additionnels sont recherchés.

ArFin	Archivierung der staatlichen Finanzdaten
AEG	*Archivage des données financières de l'Etat*

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

20.03.18 Kolloquium „AIS-Modell"
Bern
Colloque « Modèle pour logiciels de gestion d'archives »
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.