Une entreprise commune des Archives suisses



Guide d'application

Table des matières Configurations système requises3 2 3 Introduction double validation PDF/A4 Installation de KOST-Val......5 Configuration de KOST-Val......5 Composants du fichier de configuration « kostval.conf.xml »......6 Exécuter la validation9 Exécuter en utilisant le GUI « KOST-Val fr.exe »......9 8.2 10 Description des étapes de validation......14 11.1 Licence « 3-Heights™ PDF/A Validator API »21 11.2 Licence « pdfaPilot CLI »......22

1 Préambule

KOST-Val est une application java pour la validation de la configuration et du contenu des fichiers TIFF (*Tagged Image File Format*), SIARD (*Software Independent Archiving of Relational Databases*), PDF/A, JP2 (JPEG 2000 ; *Joint Photographic Experts Group 2000*) JPEG ainsi que d'un paquet d'informations à verser (SIP ; *Submission Information Package*) pour le versement d'informations numériques. Cette application est sous licence GPL3+ et elle est mise à disposition du public par le CECO en Open Source. KOST-Val se base sur les composants inchangés d'autres fabricants qui sont intégrés directement dans le code source de KOST-Val. Les utilisateurs de KOST-Val sont tenus de suivre les dispositions de licence de tous ces composants. Des informations supplémentaires sont disponibles au chapitre 11.

KOST-Val remplit les exigences décrites ci-après.

<u>Validation TIFF</u>: KOST-Val lit un fichier TIFF et valide avec l'aide de Jhove la structure et le contenu ainsi que avec l'aide de ExifTool les caractéristiques clés telles que compression, espace colorimétrique et multipage. Les caractéristiques peuvent être configurées.

<u>Validation SIARD</u>: KOST-Val lit un fichier SIARD (eCH-0165¹ v1 et v2-2017) et valide la structure et le contenu.

Validation PDF/A: KOST-Val lit un fichier PDF ou PDF/A (ISO 19005-1 et 19005-2) et valide avec l'aide du « 3-Heights™ PDF/A Validator » de PDF-Tools ou avec l'aide du « pdfaPilot » de callas la structure et le contenu du fichier PDF. KOST-Val répartit les divers messages d'erreur dans les principaux groupes tels que polices, graphique et métadonnées. KOST-Val fournit seulement des versions limitée² de « 3-Heights™ PDF/A Validator » de PDF-Tools et « pdfaPilot » de callas.

Au module J les images JPEG et JP2 contenues dans le fichier PDF sont extraites et validées selon la configuration³. Il peut également être défini si la compression JBIG2⁴ est acceptée ou non.

<u>Validation JP2</u>: KOST-Val lit un fichier JP2 (ISO 15444) et valide avec l'aide de Jpylyzer la structure et le contenu.

<u>Validation JPEG</u>: KOST-Val lit un fichier JPEG (ISO 10918-1) et valide avec l'aide de BadPeggy la structure et le contenu⁵.

¹ La spécification peut être téléchargée à partir du site eCH : http://www.ech.ch/vech-web/page?p=dossier&documentNumber=eCH-0165.

² La restriction est principalement liée à la capacité maximale de 72'000 pages par an. Plus d'informations sur les licences sont dans chapitre 11. Cette restriction peut être enlevée en acquérant une licence du « 3-Heights™ PDF/A Validator API » de PDF-Tools et en l'activant avec le License manager respectivement par acquérant « pdfaPilot » de callas.

³ ISO19005 ne stipule pas que les images contenues doivent être valides. Cette validation des images est une condition posée par l'archivage, en particulier dans l'utilisation du PDF/A en tant que format d'image.

⁴ Au sujet du problème de la compression JBIG2 voir http://kost-ceco.ch/cms/index.php?jbig2-compression_fr. Le Groupe d'experts du preservation planning PPEG recommande pour le moment de ne pas utiliser la compression JBIG2 lors de la création de fichiers PDF.

⁵ Le message d'erreur « Not a JPEG file » est raffiné par KOST-Val et sortie différencié.

<u>Validation SIP</u>: KOST-Val lit un SIP (eCH-0160⁶ v1 et v1.1) et valide les points suivants de la spécification SIP. Les différents points sont validés dans des groupes tels que la structure des données, la validation du schéma et validation de la somme de contrôle. Auparavant une validation du format est effectuée.

Les résultats (y compris les messages d'incohérences ou d'erreurs) de chaque étape sont écrits dans un fichier de log.

Les étapes de validation / les vérifications sont exécutées les unes après les autres. Lorsque cela est possible, la validation continue en cas d'erreur afin de réduire le nombre de cycles de correction.

Des informations plus détaillées sur les formats individuels et des étapes de validation sont présentées dans l'annexe.

2 Configurations système requises

Microsoft Windows 98 ou plus récente Au moins 128 MB de RAM (512 MB ou plus recommandé) Au moins 20 GB d'espace sur le disque dur Java Runtime Environment (JRE), version 6 respectivement 1.6

3 Problèmes en suspens / Commentaires / Souhaits de modifications

Les points ouverts (erreurs, bogues, souhaits de modifications) sont visibles sur la plate-forme de développement GitHub (https://github.com/KOST-CECO/KOST-Val/issues) et peuvent être rapportés à kost-val@kost-ceco.ch.

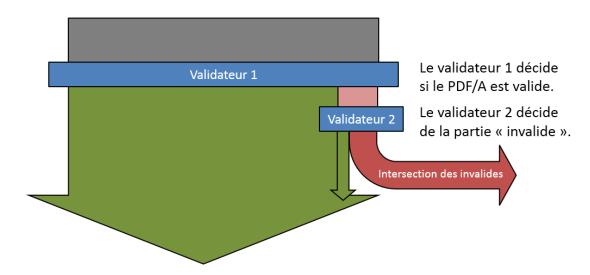
Cette liste peut être complétée par tous et est traitée par l'équipe de développement.

_

⁶ La spécification peut être téléchargée à partir du site eCH : http://www.ech.ch/vech-web/page?p=dossier&documentNumber=eCH-0160.

4 Introduction double validation PDF/A

Pour le PDF/A, le CECO propose la possibilité d'une double validation. Dans ce processus, un fichier PDF/A est d'abord contrôlé par une première validation. En cas de résultat invalide, un contrôle au moyen d'un deuxième validateur est effectué. Le PDF/A est considéré comme valide si au moins un des validateurs l'a identifié comme valide et comme invalide si les deux validateurs l'ont identifié comme invalide.⁷



La double validation PDF/A ne peut être utilisée que si les archives autorisent la prise en charge de fichiers PDF/A potentiellement invalides. Si ce n'est pas le cas, il ne faut pas faire la double validation.

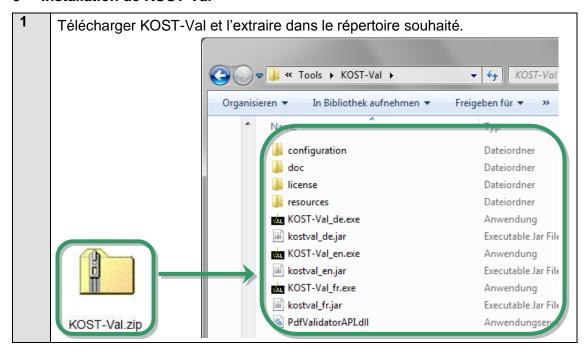
Pour la double validation, « 3-Heights™ PDF/A Validator » de PDF-Tools et « pdfaPilot » de callas sont utilisés. Si seulement un validateur est activé (voir 6.1.2), il effectuera automatiquement une validation simple.

Le constat que même des validateurs PDF/A de qualité peuvent donner des résultats différents est à la base de la conception de la double validation. Cela tient d'une part au fait que la norme PDF/A originelle implique un ensemble d'autres normes et standards qui ne sont pas implémentés jusque dans les derniers détails de manière obligatoire. D'autre part, certains objectifs de la norme sont formulés de telle sorte qu'ils peuvent légitimement être implantés de différentes manières. Pouvoir implémenter de manière homogène et exhaustive tous les outils importants de la spécification est de la musique d'avenir. C'est pourquoi la double validation de KOST-Val offre une solution provisoire.

.

⁷ C'est pourquoi la double validation ne peut être effectuée qu'avec des validateurs PDF/A de qualité. Les versions les plus récentes de « 3-Heights™ PDF/A Validator » de PDF Tools et de « pdfaPilot » de callas répondent à ces exigences élevées.

5 Installation de KOST-Val



6 Configuration de KOST-Val

Toutes les configurations de KOST-Val peuvent être effectuées dans le fichier de configuration « kostval.conf.xml » dans le sous-dossier « configuration ». Pour des modifications d'un fichier XML, un éditeur XML ou Notepad++8 doit être utilisé.



Le fichier « KaD_SignatureFile_V72.xml » et le fichier « jhove.conf » sont aussi enregistrés dans le dossier « configuration ». Ils sont décrits dans le chapitre 0.

-

⁸ L'application portable Notepad++ peut être téléchargée à partir du site http://portableapps.com/de/apps/development/notepadpp_portable.

6.1 Composants du fichier de configuration « kostval.conf.xml »

Le fichier de configuration « kostval.conf.xml » est divisé en différentes parties, qui sont décrites en détail dans le chapitre 12.1.

La configuration délivrée permet la validation immédiate de PDF/A, TIFF, SIARD, JP2, JPEG et SIP.

Ci-après une brève description des composants.

6.1.1 Partie générale

balise xml	Description : Valeur initiale
<pathtologfile></pathtologfile>	Désignation du chemin vers le répertoire logs de KOST-Val : logs
<pathtoworkdir></pathtoworkdir>	Chemin d'accès au répertoire de travail temporaire du KOST-Val : temp_KOST-Val
<pathtodroidsignature></pathtodroidsignature>	Chemin d'accès au fichier de signature DROID utilisé : configuration\KaD_SignatureFile_V72.xml
<showprogressonwork></showprogressonwork>	Afficher que KOST-Val travaille: yes

6.1.2 Partie PDF/A

balise xml	Description : Valeur initiale		
<pdftools></pdftools>	Indique si la validation PDF/A avec PDF Tools doit avoir lieu : yes Si les deux <callas> et <pdftools> ont la valeur yes, une double validation est effectuée (voir aussi chapitre 4) .</pdftools></callas>		
<callas></callas>	Indique si la validation PDF/A avec callas doit avoir lieu : yes Si les deux <callas> et <pdftools> ont la valeur yes, une double validation est effectuée (voir aussi chapitre 4) .</pdftools></callas>		
<nentry></nentry>	Indique si une erreur ou un avertissement est émis par callas lorsque l'entrée N ne correspond pas à: W		
<pdfa1></pdfa1>	Indique quelle conformité doit être atteinte au moins dans la version 1 : 1B		
<pdfa2></pdfa2>	Indique quelle conformité doit être atteinte au moins dans la version 2 : 2B		
<pdfaimage></pdfaimage>	Indique si la validation d'images (JPEG et JP2) doit être effectuée : no		
<jbig2allowed></jbig2allowed>	Indique si la compression JBIG29 est autorisée : yes		

_

⁹ Au sujet du problème de la compression JBIG2 voir http://kost-ceco.ch/cms/index.php?jbig2-compression_fr. Le Groupe d'experts du preservation planning PPEG recommande pour le moment de ne pas utiliser la compression JBIG2 lors de la création de fichiers PDF.

6.1.3 Partie SIARD

balise xml	Description : Valeur initiale	
<siardvalidation></siardvalidation>	Indique si la validation SIARD doit avoir lieu : yes	

6.1.4 Partie JP2

balise xml	Description : Valeur initiale	
<jp2validation></jp2validation>	Indique si la validation JP2 doit avoir lieu : yes	

6.1.5 Partie JPEG

balise xml	Description : Valeur initiale	
<jpegvalidation></jpegvalidation>	Indique si la validation JPEG doit avoir lieu : yes	

6.1.6 Partie TIFF¹⁰

balise xml	Description : Valeur initiale
<tiffvalidation></tiffvalidation>	Indique si la validation TIFF doit avoir lieu : yes
<allowedcompressionx></allowedcompressionx>	Indique si l'algorithme de compression X est autorisé ou non. Les algorithmes suivants sont admises dans l'état initial : Uncompressed, CCITT 1D, T4/Group 3 Fax, T6/Group 4 Fax, LZW, PackBits
<allowedphotointery></allowedphotointery>	Indique si l'espace colorimétrique Y est autorisé ou non. Les espaces suivants sont admis dans l'état initial : WhitelsZero, BlackIsZero, RGB, RGB Palette
<allowedbitspersamplez></allowedbitspersamplez>	Indique si Z bits per sample (par canal) est autorisé ou non. Les valeurs suivantes sont admises dans l'état initial : 1, 4, 8, 16
<allowedmultipage></allowedmultipage>	Indique si TIFF multipage est autorisé ou non : 1
<allowedtiles></allowedtiles>	Indique si la structure en tuiles est autorisée ou non :
<allowedsize></allowedsize>	Indique si la taille du fichier de 1000 MB (~ 1 GB) et plus est autorisée ou non : 0

KOST-Val_Guide_d_application_v1.8.2.docx Bg/Km/Rc, 16.10.2017

¹⁰ Les valeurs initiales sont fondées sur la recommandation Cfa - planification de la pérennisation (http://kost-ceco.ch/cms/index.php?preservation_fr), qui repose en partie sur le TIFF de base. Si nécessaire, les fichiers TIFF de base ont été précisés ou développés.

6.1.7 Partie SIP

balise xml	Description : Valeur initiale
<allowedlengthofpaths></allowedlengthofpaths>	Nombre maximal de caractères autorisé dans la lon- gueur du chemin : 179
<allowedsipname></allowedsipname>	Spécifications pour la composition du nom SIP : SIP_[1-2][0-9]{3}[0-1][0-9][0-3][0-9]_\\w{3}
<allowedformats></allowedformats>	Récapitulatif des formats de fichiers autorisés, se composant de sous-éléments avec des informations détaillées. Les formats suivants sont admis dans l'état initial : TXT, PDFA1, PDFA2, TIFF, JP2, JPEG, WAVE, MP3, MP4, MJ2, CSV, SIARD, WARC

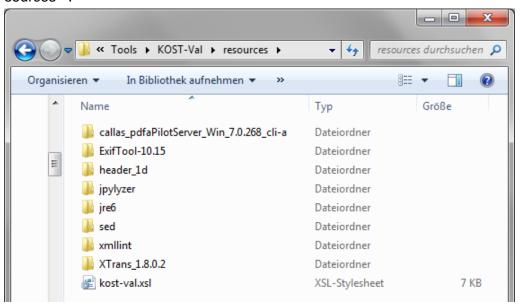
6.2 « KaD_SignatureFile_V72.xml » et « jhove.conf »

Le fichier « KaD_SignatureFile_V72.xml » et le fichier « jhove.conf » sont aussi enregistrés dans le dossier « configuration », qui n'a pas besoin d'être ajusté.

- « KaD_SignatureFile_V72.xml » est nécessaire pour la reconnaissance du format. Il comprend l'état du DROID au 28.08.2012 et a été en plus adapté par le CECO dans le domaine de la reconnaissance des formats de données d'archivage¹¹.
- « jhove.conf » est requis pour la validation de Jhove intégré dans KOST-Val.

7 Ressources de KOST-Val

Toutes les ressources de KOST-Val sont stockées dans le sous-dossier « ressources ».



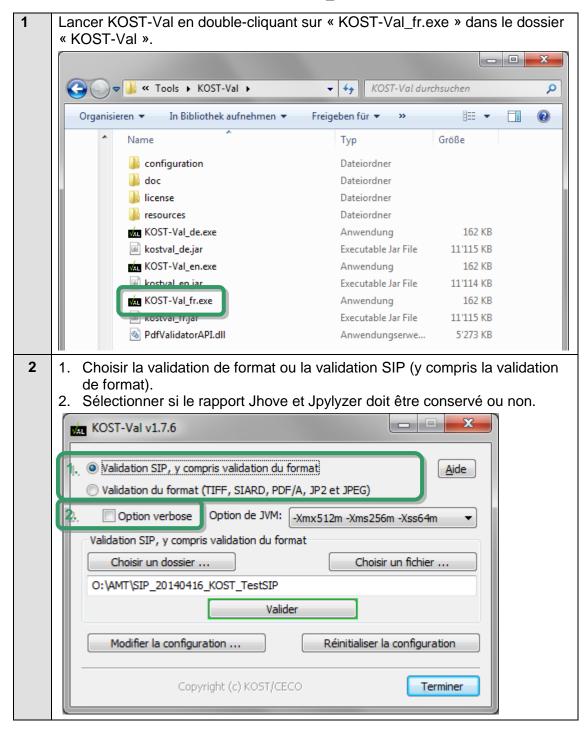
_

¹¹ A la place du fichier de signature de DROID, KOST-Val utilise un fichier de signature compatible développé par le CECO (voir aussi https://github.com/KOST-CECO/KaD_SignatureFile), KaD-SignatureFile. Ce dernier ne comprend que les formats analysés dans le Cfa et permet de les reconnaître avec la granularité recommandée par le CECO. Il aboutit ainsi à un gain significatif d'efficacité et à une facilité d'utilisation augmentée par rapport au fichier de signature de DROID.

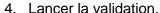
8 Exécuter la validation

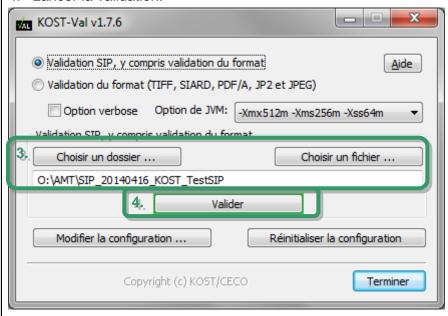
! KOST-Val n'est pas thread-safe! Cela signifie que KOST-Val ne peut pas être exécuté plusieurs fois simultanément sans interférences. L'exécution simultanée de plusieurs instances de KOST-Val peut entraîner des erreurs comme une copie de travail manquante.

8.1 Exécuter en utilisant le GUI « KOST-Val_fr.exe »



3. Entrer directement le lien ou sélectionner le fichier ou le dossier à valider.



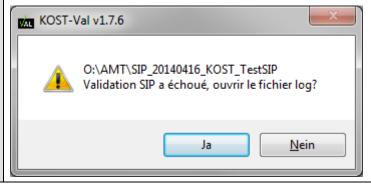


Note:

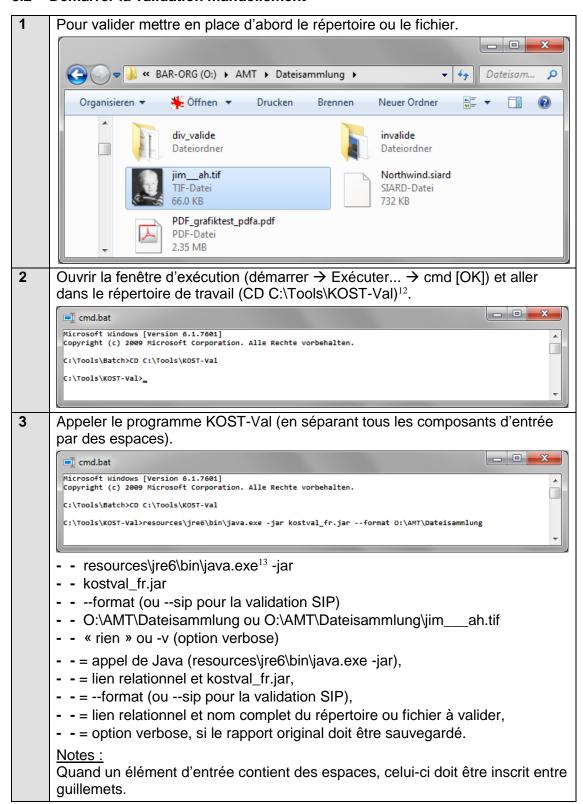
Si nécessaire, la valeur de la mémoire virtuelle de Java peut être rapidement adaptée. -Xmx doit être adaptée sur « Out of Memory » et -Xss à des erreurs « Stack Overflow ».

Si on le souhaite, la configuration peut être adaptée provisoirement. La configuration provisoire est remise à son état initial en appuyant sur le bouton « Terminer » et « Réinitialiser la configuration »

A la fin de la validation, le résultat est affiché. Le fichier KOST-Val-Log peut alors être ouvert.



8.2 Démarrer la validation manuellement



KOST-Val_Guide_d_application_v1.8.2.docx Bg/Km/Rc, 16.10.2017

¹² Le disque est par exemple changé avec CD /D.

¹³ Utiliser la commande java -jar si le Java Runtime Environment (JRE) version 6 est la version standard. JRE7 peut causer des collisions. KOST-Val est massivement plus lent avec JRE8.

Il est également possible d'appeler KOST-Val depuis tout autre emplacement. Cependant, cela nécessite l'utilisation de liens absolus, non seulement dans l'appel du programme lui-même, mais aussi dans le fichier de configuration « kostval.conf.xml ».



- - C:\Tools\KOST-Val\ resources \jre6\bin\java.exe -jar
- - C:\Tools\KOST-Val\kostval_fr.jar
- --format (ou --sip pour la validation SIP)
- - O:\AMT\Dateisammlung ou O:\AMT\Dateisammlung\jim___ah.tif
- - « rien » ou -v (option verbose)
- La validation du fichier est terminée dès que « Valid » ou « Invalid » est affiché dans la fenêtre cmd. Le dossier a été validé une fois que le prompt (C:\Tools\KOST-Val>) apparaît.

```
C:\Tools\KOST-Ual\resources\jre6\bin\java.exe -jar kostval_fr.jar --format 0:\AMT\Dateisammlung KOST-Ual\resources\jre6\bin\java.exe -jar kostval_fr.jar --format 0:\AMT\Dateisammlung KOST-Ual\resources\jre6\bin\java.exe -jar kostval_fr.jar --format 0:\AMT\Dateisammlung KOST-Ual\resources\jre6\bin\java.exe -jar kostval_fr.jar --format 0:\AMT\Dateisammlung\refteq\formattering KOST-Ual\refteq\formattering KOST-Ual\refteq\
```

Les résultats détaillés se trouvent dans le fichier « kost-val.log.xml ».

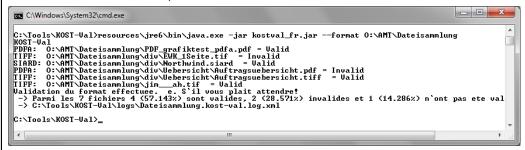
Le résultat de la validation globale (fichier correct/erroné) est aussi délivré et visible dans le statut *exit* du programme, afin que la validation puisse être liée dans la chaîne de traitement automatisée. Le statut *exit* peut prendre les valeurs suivantes :

- 0 tout est OK, valide
- 1 erreur dans l'appel du programme
- 2 échec de la validation, invalide

9 Évaluer la validation

- 1 Les étapes de validation sont décrites dans le chapitre 10.
- Après le démarrage de KOST-Val le type de validation et le fichier sont affichés. Si la validation a été réalisée sans erreurs, « Valid » est affiché.

Si la validation a échoué, Invalid est affiché.



Si l'option « -v » (verbose pour « bavard ») est entrée, le rapport original est écrit dans le fichier log. L'option « -v » est uniquement intéressante pour les spécialistes dans de rares cas, puisque tous les messages d'erreur sont transférés à 100 % vers KOST-Val.

3 La sortie du log spécifie des informations détaillées supplémentaires sur les étapes de validation invalides, en particulier les étapes invalides avec l'erreur.

KOST-Val

Format (PDF/A, TIFF, JP2, SIARD, JPEG):

Von den 7 Dateien sind 4 (57.143%) valid, 2 (28.571%) invalid und 1 (14.286%) wurden nicht validiert. Folgende Dateien wurden nicht validiert (.db): - O:\AMT\Dateisammlung\Thumbs.db Invalid: Validierung: TIFF -> O:\AMT\Dateisammlung\div\EWK_1Seite.tif Die Jhove-Valldierung wurde nicht bestanden (Status: Well-Formed, but not valid).
Jhove ErrorMessage: JPEGProc not defined for JPEG compression
Compression: JPEG (old-style) ist nicht zulaessig. A) Allgemeines PDF/A-Eintrag nicht vorhanden [callas] Geraeteabhaengige Farbe verwendet aber kein PDF/A Output Intent vorhanden [callas] C) Grafiken Geraeteabhaengige Farbe verwendet aber kein PDF/A Output Intent vorhanden [callas]
Das Dokument enthaelt geraetespezifische Farbraeume, [PDF Tools: iCategory_3]
Schrift nicht eingebettet (und Text-Darstellungsmodus nicht 3) [callas]
CIDSet in Schrift-Untergruppe ist unvollstaendig (Schrift enthaelt Glyphen, die nicht gelistet sind) [callas]
Das Dokument enthaelt Schriften ohne eingebettete Font-Porgramme oder Codierungsinformation (CMAPS), [PDF Tools: iCategory_8]
Die Dokumentmetadaten sind nicht vorhanden, inkonsistent oder beschaedigt. [PDF Tools: iCategory_16] C) Grafiken
C) Grafiken
D) Schriften
D) Schriften
D) Schriften H) Metadaten Validierung: PDFA-1B -> O:\AMT\Dateisammlung\PDF_grafiktest_pdfa.pdf Validierung: SIARD -> O:\AMT\Dateisammlung\div\Northwind.siard dataOwner enthaelt der Initialwert "(...)" und sollte korrekt ausgefuellt werden. dataOriginTimespan enthaelt der Initialwert "(...)" und sollte korrekt ausgefuell ert "(...)" und sollte korrekt ausgefuellt werden. w) Warnung
Oatedright inhepsite entire the control of the control

10 Description des étapes de validation

10.1 TIFF

Chaque étape de validation TIFF est décrite ci-après de façon détaillée.

A détection

Le fichier TIFF est-il reconnu comme tel ?

B Jhove

La validation Jhove est-elle réussie?

C compression

La compression utilisée est-elle autorisée selon le fichier de configuration ?

D espace colorimétrique

L'espace colorimétrique utilisé est-il autorisé selon le fichier de configuration ? N'existe il pas du « Planar Configuration »?

E BitsPerSample

Les BitsPerSample utilisés sont-ils autorisés selon le fichier de configuration ?

F Multipage

L'utilisation de la propriété « multipage » est-elle autorisée selon le fichier de configuration ?

G tuiles

L'utilisation de la propriété « tuiles » est-elle autorisée selon le fichier de configuration ?

H grandeur

La taille du fichier est-elle autorisée selon le fichier de configuration ?

10.2 SIARD

Ce chapitre décrit plus en détail les étapes de validation. Ces critères sont un extrait des critères obligatoires de la spécification SIARD.

A lisibilité

Le fichier SIARD peut-il être lu ?

B validation de la structure des données

La structure primaire des données est-elle correcte ?

C validation du header

Le dossier header est-il valide?

D validation de la structure

La structure de metadata.xml (folder) correspond-elle à la structure des fichiers de content ?

E validation des colonnes

Les données du fichier metadata.xml ont-elles été transférées correctement dans le fichier tableZ.xsd?

F validation des lignes

Les données du fichier metadata.xml ont-elles été transférées correctement dans le fichier tableZ.xsd?

G validation des tables

Les noms de colonnes sont-ils uniques dans la table ?

H validation du content

Les fichiers XML dans le contenu sont-ils valides par rapport à leur définition de schéma (XSD) ?

I détection du format SIARD

Le fichier SIARD est-il détecté comme SIARD ?

J données primaires supplémentaires

Tous les fichiers ou dossiers contenu dans le dossier « content » sont-ils décrits dans metadata.xml? Furent xyz et abc rempli et non laissé sur (...)?

W avertissement

Furent <dataOwner> et <dataOriginTimespan> rempli et non laissé sur (...)?

10.3 PDF/A

Les messages du rapport initial de callas sont attribués aux modules basés sur des mots. Les messages d'erreur (« Message ») sont pris en charge 1:1. Lors de la validation avec PDF Tools, les catégories sont attribuées aux modules, et le texte est affiché.

A validation générale

La validation des conditions générales telles que fichier PDF/A non corrompu est-elle réussie ?

B validation de la structure

La validation de la structure (par exemple BOF et EOF) est-elle réussie?

C validation graphique

La validation graphique (par exemple, espace de couleur valide) est-elle réussie ?

Le nombre de composants dans l'entrée N du PDF/A OutputIntent correspond-il au profil ICC ?¹⁴

D validation des polices

Les polices sont-elles intégrées correctement dans le fichier PDF/A (par exemple, les caractères utilisés sont-ils tous intégrés) ?

E validation de la transparence

Les exigences en matière de transparence sont-elles remplies pour le fichier PDF/A ?

F validation des annotations

Le fichier PDF/A ne contient-il que des annotations licites (p.ex. pas d'annotation 3D) ?

G validation des actions & interactions

Le fichier PDF/A ne contient-il que des actions licites (p.ex. pas de JavaScript) ?

Le fichier PDF/A ne contient-il que des interactions licites ?

H validation des métadonnées

Les métadonnées sont-elles correctement incluses dans le fichier PDF/A?

I validation de la disponibilité

Les exigences à l'accessibilité (niveau de conformité A) sont-elles remplies ?

J validation d'images & JBIG2 (configurable)

Le fichier PDF/A ne contient-il que des images JPEG et JP2 valides ? Le fichier PDF/A contient-il la compression JBIG2 ?

KOST-Val_Guide_d_application_v1.8.2.docx Bg/Km/Rc, 16.10.2017

¹⁴ A cet effet, on effectue un test supplémentaire avec « pdfaPilot » de callas, qui émet un avertissement ou une erreur en fonction de la configuration (voir 6.1.2). Le « 3-Heights ™ PDF/A Validator » de PDF Tools, effectuée cette vérification en standard, mais elle est critiquée dans la iCategory_1, qui est émise en raison de la composition sous « A) validation générale ».

10.1 JP2

Les tests échoués de Jpylyzer sont attribués aux modules. Des messages d'erreur correspondant à ces derniers sont émis. Ce chapitre décrit plus en détail les étapes de validation de JP2.

A Détection et Jpylyzer

Le fichier JP2 est-il reconnu comme tel ? La validation avec Jpylyzer a-t-elle été réussie ?

B Validation des métadonnées

Les métadonnées ont-elles été saisies correctement ?

C Validation de l'image

L'image est-elle construite correctement ?

D Validation étendue

Les tests Jpylyzer additionnels ont-ils été réussis ?

10.2 JPEG

Les messages d'erreur de BadPeggy sont attribués aux modules et traduits (original en anglais). Ce chapitre décrit plus en détail les étapes de validation de JPEG.

A Détection et BadPeggy

Le fichier JPEG est-il reconnu comme tel ? La validation avec BadPeggy a-t-elle été réussie ?

B Données corrompues

BadPeggy génère-il un message d'erreur, ce qui indique des données corrompues°?

C Structure de fichier invalide

BadPeggy génère-il un message d'erreur, ce qui indique une structure de fichier non valide°?

D Autres problèmes

BadPeggy génère-il un message d'erreur, qui n'a pas encore été attribué et traduit°?

10.3 SIP

Ce chapitre décrit plus en détail les étapes de validation. Ces critères sont un extrait des critères obligatoires de la spécification SIP.

Module 1 : Consistance paquet et XML

1A lisibilité

Le SIP (ZIP / ZIP64) peut-il être ouvert sans erreur ?

1B validation de la structure des données

Existe-t-il une structure de répertoire primaire correcte ?

1C nommage de la structure et des fichiers

Les noms correspondent-ils aux restrictions de la spécification ?

1D validation du schéma metadata.xml

Les métadonnées correspondent-elles aux fichiers schéma dans header/xsd?

1E déterminer le type du SIP

Le type SIP est transmis et affiché : GEVER ou FILE.

1F existence des données primaires

Existe-t-il des fichiers primaires dans le répertoire content/ ?¹⁵

Module 2 : Consistance du fichier

2A données primaires manquantes

Tous les fichiers référencés existent-ils ?

2B données primaires supplémentaires

N'existe-t-il pas de fichiers primaires dans le SIP?

2C validation de la somme de contrôle

Les sommes de contrôle des fichiers sont-elles correctes ?

2D description des données primaires GEVER

Tous les fichiers référencés sont-ils aussi enregistrés dans ordnungsystem?

Module 3 : Consistance du format de fichier et de la date

3A détection des formats

Effectuer une reconnaissance et comparer si les formats sont autorisés.

3B formats supplémentaires

Lister tous les fichiers avec un format non spécifié et indiquer le format.

3C validation de formats

Les formats sont-ils valides ?

3D validation de période

Les indications temporelles dans (metadata.xml)/ablieferung sont-elles consistantes ?

KOST-Val_Guide_d_application_v1.8.2.docx Bg/Km/Rc, 16.10.2017

¹⁵ Les fichiers primaires peuvent manquer dans un GEVER-SIP qui est utilisé uniquement pour l'archivage d'une structure de système de classement, mais jamais dans un FILE-SIP.

11 Droits d'auteur

KOST-Val a été développé par le CECO. Le CECO en possède tous les droits. KOST-Val a été publié par le CECO en 2012 sous la licence GNU General Public License v3+.

Notice : This product includes software developed by the Apache Software Foundation (http://www.apache.org/).

KOST-Val se base sur les composants inchangés suivants d'autres fabricants qui sont intégrés directement dans le code source de KOST-Val :

Programme tiers/Composants tiers	Version	Licence
3-Heights™ PDF/A Validator API		
http://www.pdf-tools.com	4.9.25	Voir chapitre 11.1
Apache Commons http://commons.apache.org/		Apache License 2.0
- commons-collections-3.2.1.jar	3.2.1	
- commons-configuration-1.6.jar	1.6	
- commons-digester-1.8.jar	1.8	
- commons-lang-2.4.jar	2.4	
- commons-logging-1.1.1.jar	1.1.1	
Apache log4j http://logging.apache.org/log4j/	1.2.12	Apache License 2.0
Apache log4j http://logging.apache.org/log4j/	1.2.12	Apache License 2.0
Apache Xalan-Java http://xml.apache.org/xalan-j/	2.7.0	Apache License 2.0
Apache Xerces http://xerces.apache.org/	2.7.1	Apache License 2.0
BadPeggy http://coderslagoon.com/	2.0	GPL°v3 License
DROID http://digital-preservation.github.io/droid/	5.0.3	3c BSD License
iText http://itextpdf.com/	5.5.5	AGPL v3 License
Jdom 2.0.0 http://www.jdom.org/	2.0.0	jdom License
Jhove http://hul.harvard.edu/jhove/	1.5	LGPL v2.1 License
Junit°4.4 http://www.junit.org/	4.4	CPL v1.0
Spring Framework API		
http://static.springsource.org/spring/docs/3.0.x/api/	3.0.0	Apache License 2.0
zip64 <u>http://sourceforge.net/projects/zip64file/</u>	1.02	GPL v2+ License

KOST-Val utilise les composants inchangés suivants d'autres fabricants qui sont livrés avec KOST-Val :

Programme	tiers/Composants tiers	Version	Licence
ExifTool http://www.sr	no.phy.queensu.ca/~phil/exiftool/	10.15	PERL respective GPL v3.0 License
Jpylyzer	http://jpylyzer.openpreservation.org/	1.17.0	LGPL°v3.0 License
NSIS°v2.46	http://nsis.sourceforge.net/Main_Page	2.46	zlib/libpng License
pdfaPilot CL	https://www.callassoftware.com	7.0.268	Voir chapitre 11.2
GNU sed	https://www.gnu.org/software/sed	4.4	GPL v3+ License
XML.nsh	http://nsis.sourceforge.net/XML_plug-in	2.0	zlib/libpng License
Xmllint	http://xmlsoft.org/xmllint.html/	20630	MIT License
XTrans	http://sourceforge.net/projects/xtrans/	1.8.0.2	GPL v2 License

Les utilisateurs de KOST-Val sont tenus de suivre les dispositions de licence de tous ces composants qui se trouvent dans le répertoire KOST-Val\license.

11.1 Licence « 3-Heights™ PDF/A Validator API »

Pour l'utilisation de la version limitée de « 3-Heights™ PDF/A Validator » de PDF Tools, le CECO a convenu des conditions d'accord individuel suivantes aux Conditions Générales de licence avec PDF Tools :

2. Individuelle Vereinbarung

Dieses Vertragsverhältnis regelt die Client-Lizenz zwischen der PDF TOOLS als Lizenzgeber und der KOST als Lizenznehmer gemäss nachfolgenden Spezialbestimmungen:

- PDF Tools AG erteilt für KOST eine kostenfreie OEM-Lizenz für das 3-Heights™ PDF/A Validator API als Zusatzfunktion ihrer eigenen Validator-Software (KOST-Val).
- Die Lizenz schliesst den Gebrauch der Software (KOST-Val) durch Gedächtnisinstitutionen, bestehend aus Archiven oder Bibliotheken, deren Zulieferer und der KOST selbst, ein.
- Der OEM-Lizenzschlüssel, welcher fest in KOST-Val eingebunden ist, darf nicht ausserhalb der Applikation (KOST-Val) verwendet werden.
- Die Lizenz ist zeitlich unbegrenzt, jedoch bezüglich Durchsatz pro Installation begrenzt (72'000 Seiten pro Jahr).
- Für die Verteilung der Software (KOST-Val) an den Anwender ist die KOST zuständig.
- Der First Level Support der Anwender erfolgt durch KOST, Second Level Support Fälle leitet KOST an PDF Tools AG weiter.
- Wenn der Anwender weitergehende Bedürfnisse hat, z.B. höherer Durchsatz, Integration in andere Applikationen etc. kauft er die Software (3-Heights™ PDF/A Validator API) direkt bei PDF Tools AG.
- Die KOST darf weiterhin den Quellcode von KOST-Val Open Source publizieren und KOST-Val gratis und ohne Registrierung abgeben.

Les points suivants sont déterminants pour l'utilisateur :

- La licence comprend l'utilisation du logiciel (KOST-Val) par des institutions patrimoniales composées des archives ou des bibliothèques, leurs fournisseurs et le CECO lui-même.
- La clé de licence OEM qui est fermement liée à KOST-Val ne peut pas être utilisée hors de l'application (KOST-Val).
- La licence est illimitée dans le temps, mais le débit par installation est limité (72'000 pages par an).
- Le support de niveau 1 aux usagers est effectué par le CECO. Le CECO transmet à PDF Tools AG les cas relevant du support de niveau 2.
- Si l'usager a d'autres besoins, par exemple un débit plus élevé, intégration dans d'autres applications, etc., il doit acheter le logiciel « 3-Heights™ PDF/A Validator API » directement auprès de PDF Tools AG.

Les utilisateurs de KOST-Val sont tenus de respecter les présentes dispositions de la licence.

11.2 Licence « pdfaPilot CLI »

Pour l'utilisation de la version limitée de « pdfaPilot CLI » de callas, le CECO a convenu des conditions d'accord individuel suivantes aux Conditions Générales de licence avec callas :

2. Individuelle Vereinbarung

Dieses Vertragsverhältnis regelt die Lizenz zwischen der callas software als Lizenzgeber und der KOST als Lizenznehmer gemäss nachfolgenden Spezialbestimmungen:

- callas software erteilt für die KOST eine kostenfreie Lizenz für callas pdfaPilot CLI für Windows zur innerbetrieblichen Nutzung und zur Integration in ihren eigenen Validator "KOST-Val".
- Die Lizenz schliesst die Distribution von KOST-Val an "Anwender" (Gedächtnisinstitutionen, Archive oder Bibliotheken und deren Zulieferer) ein.
- Für die Distribution von KOST-Val an diese Anwender ist die KOST zuständig und darf KOST-Val auch gratis und ohne Registrierung an diese abgeben.
- Die Lizenz ist zeitlich unbegrenzt, jedoch bezüglich Durchsatz pro Installation begrenzt auf 72'000 Seiten pro Jahr.
- Die KOST darf den eigenen Quellcode von KOST-Val Open Source publizieren. callas pdfaPilot CLI ist hiervon ausgenommen.
- First Level Support der Anwender erfolgt durch die KOST. Second Level Support leistet callas software gegenüber der KOST.

Les points suivants sont déterminants pour l'utilisateur :

- La licence comprend la distribution de KOST-Val aux « utilisateurs » (institutions patrimoniales, archives, bibliothèques, et leurs fournisseurs).
- La licence est illimitée dans le temps, mais le débit par installation est limité à 72'000 pages par an.
- Le CECO peut publier leur propre code de KOST-Val open source mais pdfaPilot CLI est exclue.
- Le support de niveau 1 aux usagers est effectué par le CECO. Le CECO transmet à callas software les cas relevant du support de niveau 2.

Les utilisateurs de KOST-Val sont tenus de respecter les présentes dispositions de la licence.

12 Annexe

12.1 Propriétés du fichier de configuration « kostval.conf.xml »

Le fichier de configuration « kostval.conf.xml » est divisé en différentes parties, qui sont décrites en détail dans ce chapitre.

12.1.1 Partie générale

<pathtologfile>

Chemin d'accès au répertoire de log du KOST-Val. La valeur initiale est **logs**. Il est possible de saisir le chemin absolu ou le chemin relationnel comme dans la valeur initiale (du point de vue de kostval_fr.jar). Ce dossier ne doit pas être créé.

<pathtoworkdir>

Chemin d'accès au répertoire de travail temporaire du KOST-Val. La valeur initiale est temp_KOST-Val. Il est possible de saisir le chemin absolu ou le chemin relationnel comme dans la valeur initiale (du point de vue de kostval_fr.jar). Ce dossier ne doit pas être créé.

<pathtodroidsignature>

Chemin d'accès au fichier de signature conforme à DROID utilisé dans le répertoire. La valeur initiale est configuration\KaD_SignatureFile_V72.xml.

<showprogressonwork>

Afficher (yes) ou non (no) que KOST-Val travaille. La valeur initiale est yes.

12.1.2 Partie PDF/A

<pd><pdftools>

Indique si la validation PDF/A avec PDF Tools doit avoir lieu (yes) ou non (no). La valeur initiale est yes.

Si les deux <callas> et <pdftools> ont la valeur yes, une double validation est effectuée (voir aussi chapitre 4).

<callas>

Indique si la validation PDF/A avec callas doit avoir lieu (yes) ou non (no). La valeur initiale est yes.

Si les deux <callas> et <pdftools> ont la valeur yes, une double validation est effectuée (voir aussi chapitre 4).

<nentry>

Indique si une erreur (E) ou un avertissement (W) est émis par callas lorsque l'entrée N ne correspond pas. La valeur initiale est W.

<pdd><pdfa1>

Indique quelle conformité doit être atteinte au moins dans la version 1. Il peut être choisi entre *Accessible* (1A), *Basic* (1B) ou non permis (no). La valeur initiale est 1B.

<pdd><pdfa2>

Indique quelle conformité doit être atteinte au moins dans la version 2. Il peut être choisi entre *Accessible* (2A), *Basic* (2B), *Unicode* (2U) ou non permis (no). La valeur initiale est 2B.

<pd><pdfaimage>

Indique si une validation de l'image est effectuée (yes) ou non (no). Les images JPEG et JP2 contenus sont extraites et validées. La valeur initiale est no.

<jbig2allowed>

Indique si la compression JBIG2 est autorisée (yes) ou non (no). La valeur initiale est yes.

12.1.3 Partie SIARD

<siardvalidation>

Définition si la validation SIARD doit avoir lieu (yes) ou non (no). La valeur initiale est yes.

12.1.4 Partie JP2

<jp2validation>

Définition si la validation JP2 doit avoir lieu (yes) ou non (no). La valeur initiale est yes.

12.1.5 Partie JPEG

<jpegvalidation>

Définition si la validation JPEG doit avoir lieu (yes) ou non (no). La valeur initiale est yes.

12.1.6 Partie TIFF

Les valeurs initiales sont fondées sur la recommandation Cfa - planification de la pérennisation (http://kost-ceco.ch/cms/index.php?preservation_fr), qui repose en partie sur le TIFF de base. Si nécessaire, les fichiers TIFF de base ont été précisés ou développés.

<tiffvalidation>

Définition si la validation TIFF doit avoir lieu (yes) ou non (no). La valeur initiale est yes.

<allowedcompressionX>

Indique si l'algorithme de compression est autorisé (description) ou non (0). Les valeurs initiales sont Uncompressed, CCITT 1D, T4/Group 3 Fax, T6/Group 4 Fax, LZW, PackBits.

Х	Description	Valeur initiale	Origine
1	Uncompressed	Uncompressed	Baseline-TIFF
2	CCITT 1D	CCITT 1D	Baseline-TIFF
3	T4/Group 3 Fax	T4/Group 3 Fax	recommandation Cfa-PP
4	T6/Group 4 Fax	T6/Group 4 Fax	recommandation Cfa-PP
5	LZW	LZW	recommandation Cfa-PP
7	JPEG	0	TIFF v6 TN2
8	Deflate	0	TIFF v6
32773	PackBits	PackBits	Baseline-TIFF

<allowedphotointerY>

Indique si l'espace colorimétrique est autorisé (description) ou non (0). Les valeurs initiales sont WhitelsZero, BlackIsZero, RGB, RGB Palette.

Υ	Description	Valeur initiale	Origine
0	WhiteIsZero	WhiteIsZero	Baseline-TIFF
1	BlackIsZero	BlackIsZero	Baseline-TIFF
2	RGB	RGB	Baseline-TIFF
3	RGB Palette	RGB Palette	Baseline-TIFF
4	transparency mask	0	TIFF v6
5	CMYK	0	TIFF ∨6
6	YCbCr	0	TIFF ∨6
7	CIE L*a*b*	0	TIFF v6

<allowedbitspersampleZ>

Indique si le *bits per sample* (par canal) est autorisé (Z) ou non (0). Les valeurs initiales sont 1, 4, 8, 16.

Z	Description	Valeur initiale	Origine
1	1 bits per sample (par canal)	1	recommandation Cfa-PP
2	2 bits per sample (par canal)	0	TIFF v6
4	4 bits per sample (par canal)	4	Baseline-TIFF
8	8 bits per sample (par canal)	8	Baseline-TIFF
16	16 bits per sample (par canal)	16	recommandation Cfa-PP
32	32 bits per sample (par canal)	0	TIFF v6

<allowedmultipage>

Indique si TIFF multipage est autorisé (1) ou non (0). La valeur initiale est 1 et correspond à la recommandation de la planification de la pérennisation Cfa TIFF.

<allowedtiles>

Indique si la structure en tuiles est autorisée (1) ou non (0). La valeur initiale est 0 et répond aux exigences des Baseline TIFFs.

<allowedsize>

Indique si une taille du fichier de 1000 MB (~ 1 GB) et plus est autorisée (1) ou non (0). La valeur initiale est 0 et correspond à la recommandation de la planification de la pérennisation Cfa TIFF.

12.1.7 Partie SIP

<allowedlengthofpaths>

Nombre maximal de caractères autorisé dans la longueur du chemin. La valeur initiale est **179** et correspond à la restriction de la spécification du SIP : moins de 180 caractères.

<allowedsipname>

Spécifications pour la composition du nom du SIP. La valeur initiale (expression Regex) est SIP_[1-2][0-9]{3}[0-1][0-9][0-3][0-9]_\w{3} et correspond à la restriction de la spécification du SIP: SIP_[dateversement]_[Nom du service versant](_[Référence]), pour lequel la date doit être saisie au format AAAAMMJJ et la référence peut être utilisée de manière optionnelle.

<allowedformats>

Récapitulatif des formats de fichiers autorisés, se composant de sous-balises avec des informations détaillées. Les valeurs initiales sont visibles dans le tableau suivant. La liste détaillée est conservée intentionnellement, de sorte que les formats non autorisés puissent être facilement supprimés et que le moins de formats possible aient à être ajoutés. Les indications concernant l'extension et le PUID sont visibles dans KaD_Si-gnatureFile V72.xml.

<puid> </puid>	<extension> </extension>	<comment> </comment>
TXT	txt	Plain Text file
PDFA1	pdf	PDF/A-1
PDFA2	pdf	PDF/A-2
TIFF	tif, tiff	Tagged Image File Format
JP2	jp2, jpg2, j2c	JPEG2000
JPEG	jpeg, jpg, jpe	JPEG
SVG	svg	Scalable Vector Graphics
WAVE	wav, wave	Waveform Audio
MP3	mp3	MPEG 1/2 Audio Layer 3
MP4	mp4, m4v, m4a, f4v, f4a	MPEG-4 Media File
MJ2	mj2, mjp2	MJ2 (Motion JPEG 2000)
MPEG2	mpeg, mpg, m2v	MPEG-2 Video Format
PRORES	mov	Apple ProRes
CSV	CSV	Comma Separated Values
SIARD	siard	SIARD
XML	xml	Ext. Markup Language
XSD	xsd	XML Schema Definition
XLS	xls, xlw	Microsoft Excel (xls)
WARC	warc	WARC
ARC	arc	Internet Archive
HTML	htm, html	Hypertext Markup Language

12.2 Structure du programme

KOST-Val a été développé selon les exigences suivantes :

Exigences fonctionnelles:

Validation TIFF: KOST-Val lit un fichier TIFF et valide avec l'aide de Jhove les points suivants:

Éta	pe de validation	Nom de l'étape
Α	(abandon en cas d'erreur)	détection
В		Jhove
С		compression
D		espace colorimétrique
Е		BitsPerSample
F		multipage
G		tuiles
Н		grandeur

Validation SIARD: KOST-Val lit un fichier SIARD et valide les points suivants:

Éta	pe de validation	Nom de l'étape
Α	(abandon en cas d'erreur)	lisibilité
В		validation de la structure des données
С		validation du header
D		validation de la structure
Е		validation des colonnes
F		validation des lignes
G		validation des tableaux
Н		validation du content
I		détection du format SIARD
J		données primaires supplémentaires
W		avertissement

Validation PDF/A : KOST-Val lit un fichier PDF et valide avec l'aide du 3-Heights™ PDF/A Validator de PDF-Tools ou avec l'aide du pdfaPilot de callas les points suivants :

Éta	pe de validation	Nom de l'étape
Α	(abandon en cas d'erreur)	validation générale
В		validation de la structure
С		validation graphique
D		validation des polices
Е		validation de la transparence
F		validation des annotations & interactions
G		validation des actions
Н		validation des métadonnées
I		validation de la disponibilité
J	(configurable)	validation d'image & JBIG2

Validation JP2 : KOST-Val lit un fichier JP2 et valide avec l'aide de Jpylyzer les points suivants :

Éta	pe de validation	Nom de l'étape
Α	(abandon possible en cas d'erreur)	détection et Jpylyzer
В		validation des métadonnées
С		validation de l'image
D		validation étendue

Validation JPEG: KOST-Val lit un fichier JPEG et valide avec l'aide de BadPeggy les points suivants:

Étape de validation		Nom de l'étape
Α	(abandon possible en cas d'erreur)	détection et BadPeggy
В		données corrempues
С		structure de fichier invalide
D		autres problèmes

Validation SIP: KOST-Val lit un SIP et valide les points suivants:

Étapo	e de validation	Nom de l'étape
1a	(abandon en cas d'erreur)	lisibilité
1b	(abandon en cas d'erreur)	validation de la structure des données
1c	(abandon en cas d'erreur)	nommage de la structure et des fichiers
1d	(abandon en cas d'erreur)	validation du schéma metadata.xml
1e		déterminer le type du SIP
1f		existence des données primaires
2a		données primaires manquantes
2b		données primaires supplémentaires
2c		validation de la somme de contrôle
2d		description des données primaires GEVER
3a		détection des formats
3b		formats supplémentaires
3c		validation de format
3d		validation de période

Les résultats (y compris les messages d'incohérences ou d'erreurs) de chaque étape sont écrits dans un fichier de log.

Le résultat de la validation globale (fichier correct/erroné) est aussi délivré et visible dans le statut *exit* du programme, afin que la validation puisse être liée dans la chaîne de traitement automatisée. Le statut *exit* peut prendre les valeurs suivantes :

- 0 tout est OK, valide
- 1 erreur dans l'appel du programme
- 2 échec de la validation, invalide

Les étapes de validation / les vérifications sont exécutées les unes après les autres. Lorsque cela est possible, la validation continue en cas d'erreur afin de réduire le nombre de cycles de correction.

Exigences non fonctionnelles:

Pour des tâches particulières, des programmes externes ou des frameworks Java appropriés sont utilisés.

L'application est constituée de manière modulaire, de sorte que la mise en place d'un ou plusieurs module(s) de validation supplémentaire(s) nécessite moins d'effort.

La sortie du programme et le fichier log permettent une lecture aisée du résultat de chaque validation et par conséquent l'utilisation de l'outil dans une chaîne de processus.

La sortie de la console se limite à la description du module de validation, les résultats valides ou invalides ainsi que le chemin vers le fichier. Toutes les informations additionnelles sont spécifiées dans le fichier log.

12.3 Principe de la validation de format

