Reconnaissance et validation de format : théorie et pratique

Claire Röthlisberger-Jourdan, Centre de coordination pour l'archivage à long terme de documents électroniques (CECO), Berne

Table des matières

1 Bases

1.1 Le CECO

Le Centre de coordination pour l'archivage à long terme de documents électroniques est une entreprise commune des Archives fédérales suisses, des Archives nationales de la Principauté du Liechtenstein, de vingt-quatre Archives cantonales et de cinq Archives communales de Suisse. Il a pour mandat de soutenir ses membres dans l'archivage de documents électroniques. Il élabore notamment des standards et des directives, met à disposition des outils et des prestations en vue de résoudre des problèmes concrets et aborder les étapes de travail. Il fait également état des connaissances actuelles sur des thèmes particuliers au travers d'études et colloques. De plus, le CECO communique aux Archives participantes au cours de différentes manifestations les connaissances qu'il a réunies.

1.2 Le catalogue des formats de données d'archivage (Cfa)

1.2.1 Introduction

Le catalogue des formats de données d'archivage (Cfa) est un des premiers produits du CECO. Sa première version a été réalisée en 2007 en réponse à un souhait souvent formulé par de nombreuses Archives membres du CECO. Il poursuit deux objectifs. Premièrement, il désigne les formats théoriquement adaptés pour l'archivage dans l'état actuel des connaissances et qui peuvent servir de format de destination pour la migration ou la conversion. Deuxièmement, dans les contacts avec l'Administration, il sert de référence pour déterminer quels formats peuvent être utilisés (et donc recommandés) dans les cycles de vie actifs de l'archivage.

La version actuelle du Cfa, la deuxième, date de 2009.

La troisième version sera élaborée en 2012 et publiée au début 2013. En plus de bénéficier d'une mise à jour générale, le catalogue sera actualisé, en particulier dans les domaines des formats vidéo, textuels et graphiques.

1.2.2 Catégories de formats

Pour le catalogue des formats de données d'archivage, les catégories de formats ci-après sont pertinentes: données textuelles, données graphiques, données audio, données vidéo et données structurées (tableurs, bases de données). En revanche, les fichiers exécutables ne sont pas pertinents pour les archives étant donné que ces dernières ne conservent aucun logiciel.

1.2.3 Analyse et évaluation

Les formats contenus dans le catalogue ont fait l'objet d'une analyse selon différents points de vue (fig. 1):

1. Une évaluation sous l'angle des critères archivistiques révèle la mesure dans laquelle un format satisfait aux exigences de l'archivage et son aptitude à l'archivage ainsi que les risques potentiels de son utilisation. Dans ce but, un catalogue de six critères pondérés différents a été conçu.

2. Une analyse des bonnes pratiques établit l'évaluation de chaque format dans le monde archivistique. Cette vision se traduit par deux nouveaux critères: bonnes pratiques et perspectives.
3. Une classification des formats permet de comprendre les différentes évaluations selon les deux premiers points de vue et contribue au choix d'une recommandation. Des distinctions sont opérées entre les formats bien connus, largement utilisés et aptes à l'archivage du fait de leur stabilité, les formats potentiels qui sont certainement appelés à se répandre largement et les formats potentiels, pour la conception desquels l'aptitude à l'archivage a été déterminante, mais dont l'avenir est encore incertain.

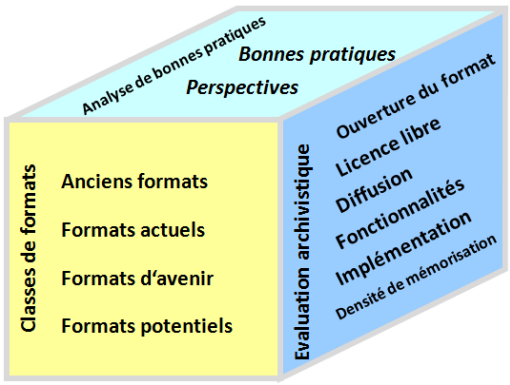

fig. 1: Critères d'analyse et d'évaluation (Cfa, version 1 et 2)

L'évaluation de chaque format est résumée dans une recommandation par catégorie de format établie par le CECO.

1.2.4 Mise en œuvre dans les Archives

Le Cfa n'est rien de plus qu'une ligne directrice. Il incombe à chaque institution d'archivage de décider quels formats elle accepte. Cette décision peut être différente d'une Archive à l'autre parce que les recommandations archivistiques contenues dans le Cfa sont soumises à des influences techniques, économiques ou politiques (fig. 2).

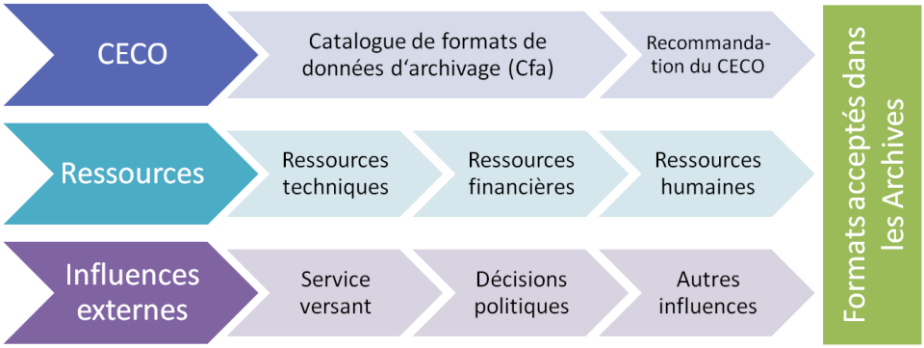

fig. 2: Influences à l'œuvre en matière d'acceptation de formats

2 Reconnaissance et validation de format: introduction

2.1 Vérification des formats acceptés

Dans la pratique, les Archives ne peuvent se contenter de définir quels formats elles acceptent pour l'archivage numérique. Le respect de cette décision doit être vérifié lors de chaque entrée. Cette vérification se fait en deux étapes: tout d'abord une reconnaissance automatique de format puis une validation du fichier dans son ensemble.

2.2 Différence entre reconnaissance et validation de format

2.2.1 Explication de base

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.

S'il n'existe pas de validateur particulier pour un format, il est alors uniquement possible de procéder à une validation approximative en ouvrant les fichiers au moyen d'un logiciel de visualisation (lecteur). Toutefois, étant donné que les logiciels de visualisation ont une certaine tolérance aux erreurs, le fait qu'un fichier s'affiche sans message d'erreur ne donne en soi aucune certitude quant à sa validité.

2.2.2 Exemple introductif

Un fichier au format PDF/A-1b contient des séquences d'octets caractéristiques particulières (fig. 3). Il commence par les signes %PDF-1 et contient à un endroit quelconque la notice de métadonnée RDF qui l'identifie en tant que PDF/A, version 1, niveau de conformité b. On voit bien ces séquences d'octets lorsqu'on ouvre un fichier PDF/A avec un éditeur de texte ordinaire:

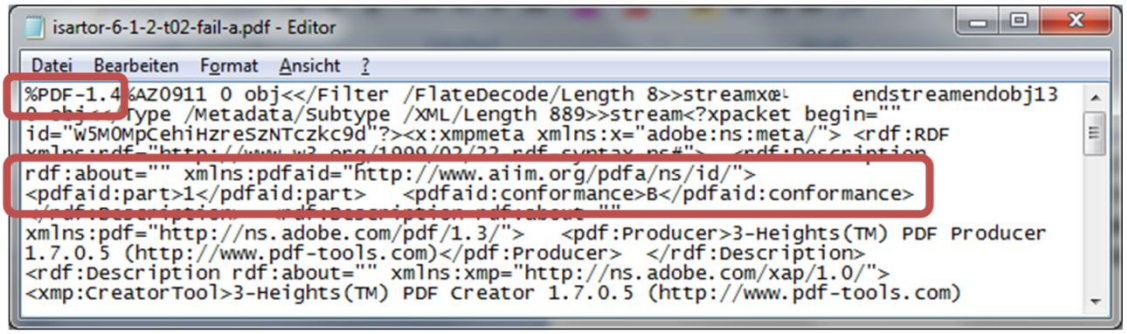

fig. 3: Séquences d'octets caractéristiques du PDF/A-1b

Si ces éléments manquent, le logiciel de reconnaissance de format ne reconnaît pas le fichier comme étant un PDF/A-1b, même si l'extension est .pdf. À l'inverse, il reconnaît à tort un fichier textuel comme étant un PDF/A-1 (fig. 4) si la séquence d'octets est présente. (On ne voit l'erreur que si l'on essaie d'ouvrir un tel fichier avec un lecteur PDF (fig. 5) ou lors de la validation PDF/A.)

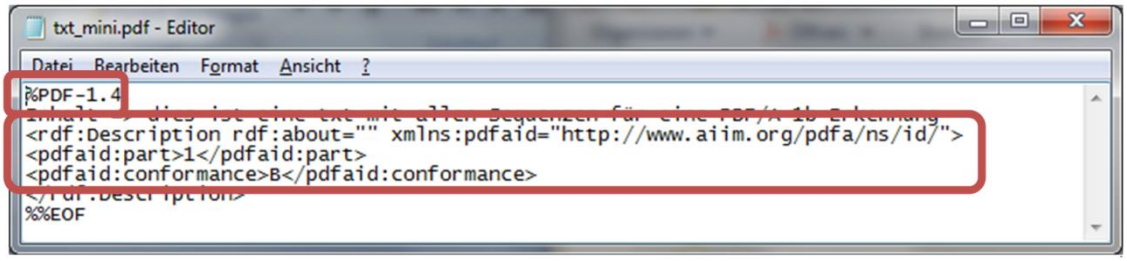

fig. 4: Fichier textuel contenant la séquence PDF/A

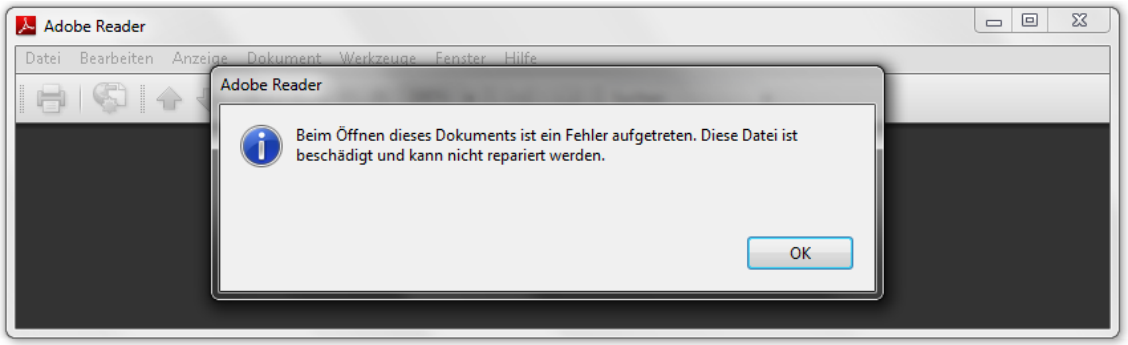

fig. 5: Message d'erreur du logiciel de visualisation en cas de fichier PDF endommagé

L'exemple suivant montre que l'ouverture d'un fichier PDF au moyen d'un lecteur PDF ne suffit pas. Il s'agit ici d'un fichier PDF qui contient la séquence d'octets mentionnée, mais qui est protégé par un mot de passe. Ceci enfreint la spécification du PDF/A. Lors de l'ouverture avec Adobe Reader, le fichier, détecté uniquement sur la base de la reconnaissance de format intégrée, sera identifié à tort comme PDF/A (fig. 6).

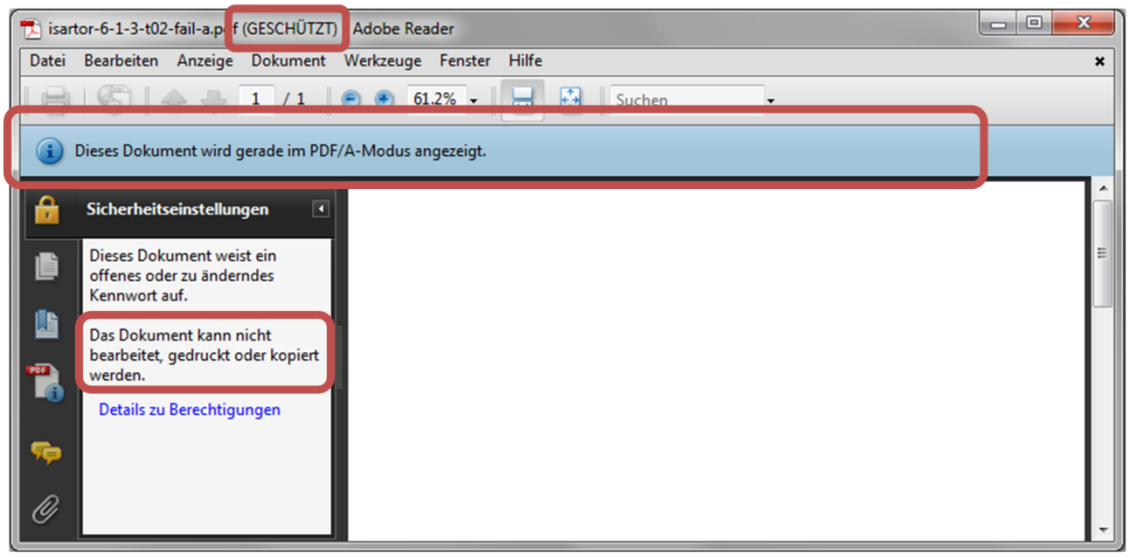

fig. 6: Limites du lecteur PDF

Dans ce cas, la simple reconnaissance de format se trompe. Seul un validateur PDF/A contrôle toutes les caractéristiques et garantit qu'il s'agit d'un PDF/A-1b valide, qui, à l'avenir également, pourra être transmis ou converti correctement (fig. 7).

PDF/A Validation Results (Before Conversion):

Run DateTime: 2012-Apr-19 15:14:49

Fail:

Formaterkennung und Formatvalidierung_v1.0.pdf

Error ID	*Message*	*Obj Refs*
e_PDFA14	Contains compressed object streams	9, 10, 12, 13, 15, 16, 40
e_PDFA15	Contains cross-reference streams	
e_PDFA341	The font is not embedded	82
e_PDFA361	Widths in embedded font are inconsistent with /Widths entry in the font dictionary	23, 24, 65, 73, 75, 78, 80, 82
e_PDFA723	The XMP Metadata stream is not valid	14
e_PDFA2331	Device-specific color space used, but no GTS_PDFA1 OutputIntent	1, 37

Pass:

Adobe_konvertiertes_Dokument_pdfa.pdf

PDF/A Revalidation Results (After Conversion):

Run DateTime: 2012-Apr-19 15:14:49

Fail:

Pass:

Generated using PDFTron PDF/A Manager V1.02.

fig. 7: Résultats de la validation de deux fichiers PDF

2.3 Utilisation et utilité de la reconnaissance de format

Vu les limites exposées et le flou de la reconnaissance de format, et parce qu'il faut toujours procéder à une validation en fin de compte, on peut se demander si une reconnaissance de format est vraiment nécessaire.

Sa fonction principale est d'opérer un tri. Les Archives doivent recourir à l'aide des machines dès que la quantité de données à contrôler est importante. La suite du processus peut être abordée automatiquement en se basant sur la reconnaissance de format. Dans le graphique ci-après, par exemple, la reconnaissance de format oriente pour traitement le fichier test.pdf soit vers la validation PDF/A, soit vers la conversion PDF en PDF/A soit, autre possibilité, vers la conversion DOC en PDF/A (fig. 8).

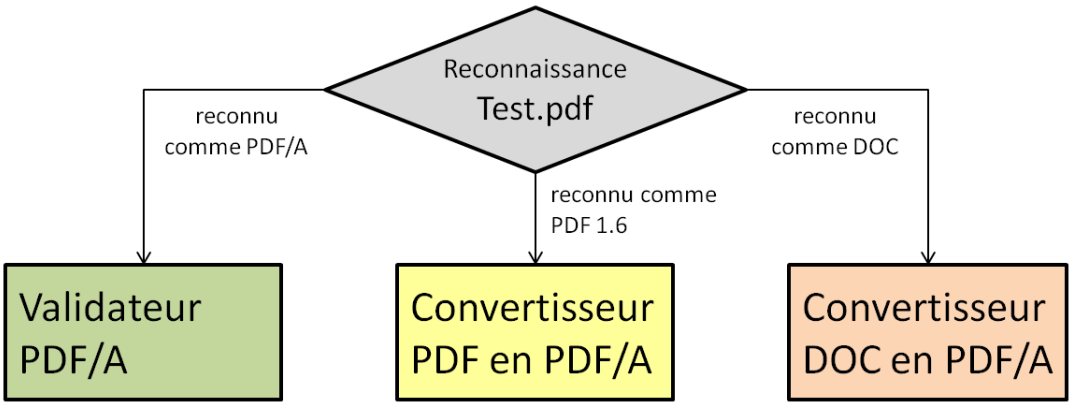

fig. 8: Reconnaissance de format: tri automatique

3 Reconnaissance de format

3.1 Exigences requises pour la reconnaissance de format

Pour être utilisé dans le processus standard des Archives, un outil de reconnaissance de format doit répondre aux exigences suivantes:

- Aptitude au mode batch
- Installation sans droits d'administrateur
- Reconnaissance rapide et de qualité
- Liste avec caractéristiques de chaque format, accessible au public et éditable.

3.1.1 Aptitude au mode batch

L'aptitude au mode batch est la condition impérative pour que la reconnaissance de format puisse être réellement intégrée au processus standard. En outre, les résultats doivent sortir sous une forme définie afin que la suite du processus puisse être lancée correctement.

3.1.2 Installation

Les logiciels devraient pouvoir être installés autant que possible sans droits d'administrateur ou alors aucune installation ne devrait être nécessaire. Cela facilite beaucoup le travail, surtout dans la phase de test.

3.1.3 Qualité et vitesse

En principe, la qualité l’emporte sur la vitesse. Les tests menés par le CECO se basent toujours sur l'installation visant la meilleure qualité possible avec l'influence négative que cela implique sur la vitesse.

En ce qui concerne la qualité, la granularité désirée est également importante. Du point de vue du CECO, la granularité doit être fine afin de pouvoir par exemple au moins différencier un fichier PDF (par exemple PDF-1.4 ou PDF-1.7) d'un fichier PDF/A (par exemple PDF/A-1b ou PDF/A2u). La granularité dépend également de la base de données de formats de fichiers utilisée.

3.1.4 Base de données de formats de fichiers

Les caractéristiques de chaque format doivent se baser sur une liste ou une base de données accessible au public et éditable. Comme le nombre de formats potentiels à identifier et de versions de formats est très élevé, et que l'identification est, de ce fait, relativement complexe, il est indispensable que cette base de données soit, si possible, utilisée et entretenue par un grand nombre d'acteurs. La base de données la plus répandue dans les milieux archivistiques et qui répond à ces critères est la base de données PRONOM des Archives nationales du Royaume-Uni (TNA).

3.2 Les quatre outils de reconnaissance de format les plus connus

Le tableau ci-après décrit brièvement les quatre logiciels de reconnaissance de format pour l'archivage numérique les plus connus et disponibles sur le marché:

- DROID DROID (Digital Record Object Identification) est un logiciel dévelop-

pé par les TNA afin de procéder à l'identification automatisée de formats de fichiers à l'aide de leur base de données PRONOM. DROID est une application Java indépendante de la plateforme et il peut être utilisé aussi bien depuis une interface graphique (Graphical User Interface, GUI) que depuis une console ou depuis une interface basée sur des lignes de commandes (Command Line Interface, CLI).

DROID est gratuit et sous licence BSD pour DROID v4.0. Il peut être téléchargé sur http://sourceforge.net/projects/droid/.

- File — File est un programme UNIX de reconnaissance de format de fichier. Les *nombres magiques* utilisés pour la reconnaissance sont indiqués dans un fichier textuel ASCII du nom de *Magic*.
  File est gratuit et peut être téléchargé notamment sur ftp://ftp.astron.com/pub/file/ .
- Fido — Fido (Format Identification for Digital Objects) est un outil CLI pour Windows et Linux destiné à l'identification de formats de fichiers. Il a été développé par Open Planets Foundation (OPF) pour faciliter l'intégration dans des procédures automatisées. Fido utilise les signatures PRONOM. Toutefois, il les traduit en expressions rationnelles pour les utiliser directement.
  Fido est disponible sous licence Apache 2.0 et peut être téléchargé sur http://github.com/openplanets/fido/downloads .
- Tika — La boîte à outils Apache Tika™ permet de détecter et d'extraire des métadonnées et des contenus textuels structurés à partir de différents documents en utilisant une des bibliothèques d'analyse syntaxique existantes. C'est la partie "Mime Magic Detection" qui est utilisée pour la reconnaissance de format.
  Tika est un projet de la Apache Software Foundation sous licence Apache, version 2.0. La dernière version Tika peut être téléchargée gratuitement sur http://tika.apache.org/download.html .

3.2.1 Évaluation sommaire des logiciels de reconnaissance de format

Asger Blekinge a testé pour le projet Scape trois des outils mentionnés. Les résultats des tests sont publiés sur le blog de Open Planets Foundation[1]. Vous en trouvez ci-après une version résumée et complétée (tableau 1). Notons que la granularité désirée ainsi que les paramètres utilisés influencent les résultats de manière importante.

[1] Asger Blekinge, Identification tools, an evaluation
http://www.openplanetsfoundation.org/blogs/2012-02-23-identification-tools-evaluation

Nom	Forces	Neutre	Faiblesses
DROID	Installation Base de données PRONOM Qualité		Aptitude mode batch Vitesse
File	Aptitude mode batch Installation Vitesse		Qualité Autre base de données
Fido	Aptitude mode batch Base de données PRONOM Qualité	Installation	Vitesse
Tika	Aptitude mode batch Installation Vitesse		Qualité Autre base de données

Tableau 1: Évaluation sommaire des logiciels de reconnaissance de format

3.2.2 Analyse des bases de données de formats utilisées

Les quatre outils de reconnaissance de format utilisent trois sources différentes pour reconnaître chaque fichier. Afin que chacun puisse mieux se représenter les choses, chaque source sera expliquée à l'aide de la notice pour PDF 1.4. Pour les cas où cette granularité n'est pas donnée, la notice PDF a été utilisée.

La base de données PRONOM

La base de données PRONOM a été élaborée par les TNA. Elle est continuellement actualisée et complétée par la communauté mondiale de ses utilisateurs et par des concepteurs de logiciels. La base de données PRONOM est utilisée aussi bien par DROID que par Fido. Elle contient les notices d'environ 860 formats.

L'illustration ci-après représente l'extrait d'une notice PDF 1.4:

Name	Acrobat PDF 1.4 - Portable Document Format		
Version	1.4		
Other names	PDF (1.4)		
Identifiers	MIME: application/pdf Apple Uniform Type Identifier: com.adobe.pdf PUID: fmt/18		
External signatures	File extension: pdf		
Internal signatures	**Name**	PDF 1.4	
	Description	Header and footer	
	Byte sequences	**Position type**	Absolute from BOF
		Offset	0
		Byte order	
		Value	255044462D312E34
		Position type	Absolute from EOF
		Offset	0
		Byte order	
		Value	2525454F(46|460A|460D|460D0A|460D00)

fig. 9: Notice PDF tirée de la base de données PRONOM

En plus du type MIME répandu (application/pdf), PRONOM publie également son propre identifiant, le PRONOM Persistent Unique Identifier (PUID, dans l'exemple fmt/18) qui montre la granularité correspondante.

Sous "Byte sequences" il est indiqué que le PDF 1.4 doit commencer par "%PDF-1.4" (écriture hexadécimale 255044462D312E34) et se terminer par "%%EOF" ou "%%EOF." ou "%%EOF.." (écriture hexadécimale 2525454F46, 2525454F460A, 2525454F460D, 2525454F460D0A ou 2525454F460D00).

La collecte de données Magic

Dans l'application File se trouve également la collecte de données Magic. Cette dernière comprend environ 230 fichiers, qui, eux-mêmes, contiennent dans certains cas plusieurs identifications de format, mais une seule notice générale concernant le PDF.

Ci-après (fig. 10) la notice relative au PDF:

```
#------------------------------------------------------------------------------
# $File: pdf,v 1.6 2009/09/19 16:28:11 christos Exp $
# pdf:  file(1) magic for Portable Document Format
#

0   string      %PDF-       PDF document
!:mime  application/pdf
>5  byte        x       \b, version %c
>7  byte        x       \b.%c
```

fig. 10: Notice PDF tirée de la collecte de données Magic de File

La ligne d'identification (0 string %PDF- PDF document) indique qu'un PDF doit commencer par le texte "%PDF-" et qu'une reconnaissance réussie indique "PDF document". File indique également le type MIME (application /pdf).

Tika-mimetypes.xml

Tika-mimetypes.xml est disponible dans Tika sous org.apache.tika.detect.MagicDetector. Tika-mimetypes.xml contient environ 170 notices, dont une seule définissant le PDF.

Ci-après (fig. 11), la notice pour le PDF:

```
<mime-type type="application/pdf">
  <acronym>PDF</acronym>
  <comment>Portable Document Format</comment>
  <magic priority="50">
    <match value="%PDF-" type="string" offset="0" />
  </magic>
  <glob pattern="*.pdf" />
  <alias type="application/x-pdf" />
</mime-type>
```

fig. 11: Notice PDF tirée de Tika-mimetypes.xml

Les notices Tika se basent sur les notices du type MIME (par exemple application/pdf).

La ligne d'identification (<match value="%PDF-" type="string" offset="0" />) indique qu'un PDF doit commencer par le texte "%PDF-".

Conclusion

La qualité de la reconnaissance de format atteinte au moyen de File et Tika est insuffisante parce que la granularité n'est pas assez fine. Par conséquent, ces deux programmes n'ont pas été soumis à une évaluation détaillée. En revanche, DROID et Fido ont subi une évaluation détaillée compte tenu de leur qualité et de leur vitesse (critère secondaire). Pour le test, nous avons utilisé un échantillon de 664 fichiers ainsi que DROID_SignatureFile_V55 à chaque fois (pour que les deux outils soient mieux comparables).

3.2.3 Évaluation détaillée de Fido 1.0.0

Qualité et vitesse

Fido a eu besoin de moins de 6 minutes pour une reconnaissance avec une capacité mémoire théorique de 10 Go. Fido utilise correctement la base de données PRONOM.

Si on utilise la capacité mémoire standard de 128 Ko, Fido a certes besoin de moins de 20 secondes, mais il n'a pas reconnu 44 fichiers correctement parce qu'il n'a pas chargé tous les octets des grands fichiers et que, dans chaque cas, au moins une séquence d'octets importante ne se serait trouvée qu'à la fin du fichier. Dans notre test, des fichiers PDF/A et SIARD en particulier sont concernés. Par conséquent, compte tenu de la perte de qualité, une telle réduction de la capacité mémoire pour gagner beaucoup de temps n'est pas recommandée.

Cependant, contrairement à DROID, Fido offre la possibilité de restreindre les PUIDs utilisés pour la reconnaissance de format, soit en excluant certains PUIDs, soit en définissant les PUIDs qu'il faut utiliser. Dans un deuxième test, seuls 10 PUIDs ont été choisis pour la reconnaissance, à nouveau avec une capacité mémoire théorique de 10 Go. Cette fois, Fido a eu besoin de moins de 40 secondes pour effectuer la reconnaissance. Parmi les formats reconnus, Fido a identifié à tort, en se basant sur l'extension, 43 fichiers PDF comme étant des PDF/A. À part ce défaut de jeunesse de Fido, tous les PUIDs désirés ont été reconnus correctement. Dès que cette erreur sera réparée, cette forme de restriction est très recommandée si on travaille en principe avec un nombre restreint de formats adaptés pour l'archivage. Le petit solde éventuel de fi-

chiers qui n'ont pas été reconnus peut ensuite être identifié par Fido lors d'un deuxième passage sans restriction des PUIDs.

Aptitude au mode batch

Fido est un programme console, apte au mode batch à 100 %. Tous les paramètres jusqu'au format d'édition sont transmis via la fonction de reconnaissance de Fido.

Installation

En principe, il n'y a pas besoin de droits d'administrateur pour installer Fido. Cependant, comme Fido est un programme Python, ce dernier doit être installé sur l'ordinateur. Il est possible d'utiliser Python 2.7 en version application portable.

3.2.4 Évaluation détaillée de DROID 6.0.1

Qualité et vitesse

DROID a eu besoin d'environ une minute pour effectuer la reconnaissance avec une capacité mémoire théorique non limitée. Le résultat de la reconnaissance correspond aux résultats attendus à l'aide de la base de données PRONOM et est identique à celui de Fido.

Si l'on utilise une capacité mémoire de 128 ko, DROID n'a plus besoin que de 5 secondes, mais ne reconnaît pas correctement les mêmes 44 fichiers que Fido. Par conséquent, compte tenu de la perte de qualité, une telle réduction de la capacité mémoire pour gagner beaucoup de temps n'est pas recommandée.

Aptitude au mode batch

L'utilisation de l'interface graphique (GUI) de DROID est particulièrement convaincante. En revanche, la version console est très insatisfaisante. Dans le cadre du test, il a été en particulier impossible de générer un résultat utilisable. Comme DROID est un logiciel libre Java, il serait théoriquement possible d'intégrer DROID directement dans une application Java. Malheureusement, le CECO n'a pas pu faire marcher la version 6 de DROID comme interface de programmation Java (Application Programming Interface). Il a certes été possible d'intégrer la version 5, mais dans ce cas, des fichiers très volumineux n'ont partiellement pas été reconnus correctement, et ce malgré une capacité mémoire non limitée.

Installation

Il n'y a pas besoin de droits d'administrateur pour installer DROID.

3.3 Conclusion sur les logiciels de reconnaissance de format

Son ampleur et sa granularité parlent en faveur de l'utilisation de la base de données PRONOM pour la reconnaissance de format. Pour les Archives, il est plus recommandé de participer au développement et à l'amélioration de PRONOM que d'élaborer sa propre base de données et ses propres outils de reconnaissance de format[2]. Par

[2] Ceci contredit donc le postulat de Steffen Bachmann et Katharina Ernst pour une reconnaissance de format pondérée sous la forme de modules („Formaterkennung – Ziele, Herausforderungen, Lösungsansätze“, in: Matthias Manke (Hg.), *Auf dem Weg zum digitalen Archiv. Stand und Perspektiven von Projekten zur Archivierung digitaler Unterlagen*. 15e Journée de la Communauté allemande de travail „Archivierung von Unterlagen aus digitalen Systemen“ des 2 et 3 mars 2011 à Schwerin. Publication des Landeshauptarchivs Schwerin, 2012). Les problèmes qu'ils décrivent lors de la reconnaissance de format reposent sur des particularités ou des er-

conséquent, pour les logiciels, ce sont les outils DROID ou Fido qui entrent en ligne de compte. Contrairement à Tika et File, tous les deux sont plus lents, ce qui est en majeure partie attribuable à la granularité de la reconnaissance ou à la taille de PRONOM.

Les forces et faiblesses de DROID et Fido ressortent clairement de l'évaluation détaillée. Pour qui n'utilise la reconnaissance de format que de temps en temps, l'interface graphique de DROID sera certainement très satisfaisante. Pour une mise en œuvre dans le cadre d'un processus automatique, l'utilisation de Fido est recommandée, car il n'y a aucune réserve à formuler sur son aptitude au mode batch. La vitesse joue un rôle secondaire lors de l'évaluation. Un autre avantage de Fido est que le programme est actuellement développé par au moins trois personnes venant d'Archives différentes et que son code source est beaucoup plus transparent, alors que le développement de DROID est perturbé par différents changements au niveau du personnel au sein des Archives nationales du Royaume-Uni.

4 Validation de format

4.1 Exigences requises pour la validation de format

Pour être utilisé dans les processus standards des Archives, un outil de validation de format doit répondre aux exigences suivantes:

- Aptitude au mode batch
- Installation sans droits d'administrateur
- Validation rapide et de qualité
- Traçabilité / Vérification de la validation

Les exigences requises pour la validation de format sont presque identiques à celles pour la reconnaissance de format. En ce qui concerne la qualité, il n'est pas seulement déterminant de savoir si le fichier est valide ou invalide, mais également que le message d'erreur correspond à la bonne erreur.

La traçabilité de la validation est une condition. Si ce n'est pas possible, par exemple s'il ne s'agit pas d'un logiciel libre, la validation doit être vérifiée pour chaque exigence à l'aide de fichiers de test invalides.

4.2 Logiciels de validation

Vu les exigences requises pour la validation, il est généralement nécessaire d'avoir un outil spécialisé par format.

4.2.1 PDF/A

La validation du PDF/A est très complexe et n'est pas simple à réaliser. De plus, comme la demande est très importante, il existe un marché pour les validateurs PDF/A et on ne peut pas se procurer de logiciels libres. La vérification de la qualité des validateurs a été effectuée par le CECO. Elle est basée sur les fondements des travaux de

---

reurs de PRONOM, par exemple le fait de renoncer à une reconnaissance de chaque version TIFF, remplacé par une notice générique pour le format TIFF (PUID fmt/353). En ce qui concerne les erreurs dans PRONOM, il s'agit si possible d'élaborer et de soumettre des propositions de corrections dans le but d'augmenter la qualité de la base de données. Bien entendu, il faut faire attention de toujours utiliser la version actuelle de PRONOM ou le dernier *SignatureFile* de DROID.

l'entreprise PDFlib. Les résultats sont réunis dans une étude sur les convertisseurs PDF/A et publiés sur le site du CECO[3].

Ci-après, le récapitulatif des produits des années 2009-2010 réunit les quatre validateurs les plus connus[4] (tableau 2):

PDF/A Validatoren		**Adobe: Adobe Acrobat 9.1**	**Intarsys: PDF/A Live**	**PDF Tools: 3Heights PDF Validator Shell**	**PDFTron: PDF/A Manager**
Geschwindigkeit & Robustheit: Sehr gut = <1 Minute & ohne Absturz Gut = 1 - 5 Minuten & ohne Absturz Ausreichend = >5 Minuten & ohne Absturz Mangelhaft = Absturz		Ausreichend	**Gut**	**Sehr gut**	**Sehr gut**
Genauigkeit: [5] Sehr gut = Mittelwert >=95% Gut = Mittelwert 90% - 94% Ausreichend = Mittelwert 75 - 89% Mangelhaft = Mittelwert <75%		**Gut**	**Sehr gut**	**Gut**	**Sehr gut**
Isartor testsuite (non-conforming)					
6.1 File structure	31x	97%	90%	100%	90%
6.2 Graphics	47x	100%	100%	100%	100%
6.3 Fonts	28x	100%	96%	100%	100%
6.4 Transparency	6x	100%	100%	100%	100%
6.5 Annotations	25x	96%	100%	100%	100%
6.6 Actions	37x	100%	100%	100%	100%
6.7 Metadata	27x	100%	100%	100%	96%
6.9 Interactive Forms	3x	100%	100%	100%	100%
Other non-conforming					
ISO 19005 violations	9x	89%	89%	89%	100%
XMP 2004 violations	5x	20%	60%	20%	80%
PDF 1.4 violations	8x	38%	100%	63%	75%
Conforming					
Real world	34x	88%	97%	85%	100%
PDFlib samples	8x	100%	100%	88%	88%
Advanced XMP	16x	100%	100%	100%	100%
Mittelwert (conforming / non-conforming)		**91%**	**97%**	**90%**	**95%**
Getestete Version:		9.1.0	5.0.4	1.8.32.1	1.00 (CLI)
Tester / Testjahr:		**PDF**lib / 2009	KOST / 2010	**PDF**lib / 2009	KOST / 2010
Bemerkungen:		Die Version 9.0 ist bei der Geschwindigkeit und Robustheit mangelhaft.	Test mit einer neueren Version nochmals komplett durchgeführt.		Report ist sehr übersichtlich (Kompakt mit guter Fehlermeldung).

Tableau 2: Extrait du récapitulatif des produits tiré de l'étude du CECO

Pour la validation du PDF/A il faut noter que le PDF/A a certes sa norme ISO, mais celle-ci ne fait qu'énumérer quelles fonctions de la version PDF servant de base sont obligatoires, recommandées, restreintes ou interdites. Dans les détails, ces fonctions sont interprétées individuellement de manière différenciée. En outre, tous les documents qui définissent conjointement la norme PDF/A sont très volumineux et très techniques. Dans certains cas, cela aboutit à des résultats de validation variables.

[3] Étude du CECO en vue d'établir un récapitulatif et d'évaluer divers programmes de validation de fichiers PDF/A, http://kost-ceco.ch/cms/index.php?pdfa_validatoren_fr.

[4] L'étude du CECO a également examiné le pdfaPilot de Callas. Vu que ce produit est contenu dans Adobe Acrobat sous le nom de Preflight Validator, et comme les résultats de ces deux validateurs ne présentent pas de différence significative, pdfaPilot de Callas ne figure pas dans le récapitulatif dans le but d'en améliorer la lisibilité. Le récapitulatif complet des produits est disponible dans l'étude.

[5] Pour la précision, ce n'est pas le résultat (réussi/échoué) uniquement qui est important; le message d'erreur doit en outre décrire au moins une erreur réelle.

4.2.2 TIFF

Pour la validation de fichiers TIFF, le CECO ne connaît que le validateur JHOVE. JHOVE (JSTOR/Harvard Object Validation Environment) est libre et gratuit. Le CECO n'a pas connaissance d'une vérification indépendante de la qualité de validation effectuée à l'aide du code source.

4.2.3 JPEG2000

Pour la validation de fichiers JPEG2000, en plus du validateur JHOVE le CECO connaît également le jpylyzer de OPF. Les deux validateurs sont libres et gratuits. Le CECO n'a pas connaissance d'une vérification indépendante de la qualité de validation effectuée à l'aide du code source.

4.2.4 WAV

Pour la validation de fichiers WAV, le CECO ne connaît que le validateur JHOVE. Ce dernier est libre et gratuit. Le CECO n'a pas connaissance d'une vérification indépendante de la qualité de validation effectuée à l'aide du code source.

4.2.5 SIARD

Il n'existait jusqu'en 2012 aucun validateur pour la validation de fichiers SIARD. C'est pourquoi le CECO développe actuellement l'application SIARD-Val (validateur pour fichiers SIARD), qui sera disponible gratuitement sous licence GPL3. La version bêta est prévue pour l'automne 2012. Ensuite de quoi une autre institution vérifiera la qualité de la validation à l'aide du code source.

4.3 Conclusion sur les logiciels de validation de format

Ce n'est pas uniquement la validation de format qui représente un travail de grande ampleur, mais également l'assurance de la qualité des outils de validation. C'est pourquoi il n'existe pour beaucoup de validateurs encore aucune vérification indépendante de la qualité de leurs résultats. Le monde des Archives devra combler ces lacunes. Le CECO prévoit donc de poursuivre son travail et ses recherches sur les outils de validation. Sur quels formats l'attention se portera-t-elle par la suite, la question reste en suspens.