XCDF : Un format canonique pour la représentation de documents

Jean-Luc Bloechle – Maurizio Rigamonti – Denis Lalanne – Rolf Ingold

Groupe DIVA, Département d'Informatique de l'Université de Fribourg
Boulevard de Pérolles 90, 1700 Fribourg, Suisse

{prénom.nom}@unifr.ch

Résumé *: Accéder au contenu structuré d'un document PDF est une tâche complexe dépendante de méthodes de pré-traitement et de rétro-ingénierie. Cet article décrit le format canonique XCDF utilisé pour la représentation des résultats d'extraction et d'analyse des structures physiques de documents PDF. Ce format est positionné par rapport aux autres recherches, puis détaillé d'un point de vue théorique. XED, l'outil réalisant la transformation de fichiers PDF vers le format XCDF est ensuite brièvement présenté. L'intérêt de XCDF est finalement illustré à l'aide de plusieurs exemples d'applications concrètes mettant en évidence son rôle central lors d'analyses de plus haut niveau.*

Mots-clés *:* PDF, XML, structures physiques et logiques, représentation de documents

1 Introduction

PDF (Portable Document Format [ADO]) est un format universel permettant la représentation d'un vaste éventail de documents électroniques. Au cours des années, ce format s'est considérablement enrichi: interaction, multimédia, annotations, etc. Cette généralité, unie à une représentation compacte de l'information ainsi qu'à une impression fidèle sur n'importe quel système, a permis au format PDF de s'imposer comme standard pour l'échange de documents à travers l'Internet. La richesse du format implique une quantité excessive d'opérateurs, parfois redondants, ayant provoqué le développement de générateurs PDF très divers et souvent complexes. Par défaut, ces générateurs de fichiers PDF se concentrent uniquement sur la préservation de la mise en page d'un document source; en conséquence 1) ils ne conservent pas les structures du document lorsque celles-ci sont présentes; 2) ils ajoutent de l'information supplémentaire, voire du bruit ainsi que de la sur-segmentation, au contenu original [RIG 05-1] ; 3) ils ne génèrent pas un fichier de manière unique (un générateur peut représenter un tableau sous forme de graphiques tandis qu'un autre sous forme d'images). Ce manque de structuration rend ardue la réédition, le copier-coller ainsi que l'indexation par des systèmes spécifiques. Ceux-ci sont alors obligés de pré-traiter les documents PDF afin d'en extraire le contenu et de le structurer [ANJ 01-2, LAW 99]. Pour faciliter la réutilisation de ce contenu il est donc nécessaire de 1) reconstituer les entités textuelles homogènes (mots, lignes, blocs) extraites des documents PDF et 2) définir un format canonique capable d'organiser le contenu original en fonction des structures extraites. En ce sens, différentes recherches ont été conduites; une synthèse de ces méthodologies est exposée dans la section 2. Notre proposition de représenter les documents électroniques d'une façon unique et structurée à l'aide du format XCDF est décrite dans la section 3. XED est notre outil de transformation de fichiers PDF vers le format XCDF (section 4) dont l'utilisation est illustrée par trois applications (section 5). Finalement, les extensions de XCDF et ses futures utilisations sont annoncées dans la conclusion.

2 Taxonomie de méthodes existantes pour l'analyse de PDF

Aujourd'hui, plusieurs travaux et recherches ont été accomplis [BLO 06], afin d'exploiter le contenu des documents PDF, d'en extraire les structures physiques et logiques, puis d'en dériver d'autres annotations comme l'ordre de lecture.

L'analyse de l'image du document profite des méthodes qui ont mûri pendant les dernières décénies et qui sont appliquées à des documents généralement idéaux, sans bruits et imprimés à haute résolution [HAD 03], afin de retrouver le contenu et les structures originales. L'analyse du contenu électronique du document [PAK 98] profite de techniques partiellement dérivées de celles de l'analyse d'image. Ces méthodes utilisent directement les primitives PDF [RIG 05-1]. Dans [HAD 04, RIG 04], nous avons proposé de mélanger les deux méthodologies afin de pouvoir analyser chaque type de document PDF.

L'analyse du contenu électronique est à son tour composée de méthodes extensives et de restructuration. Les premières analysent le contenu du document afin de reconstituer les structures originales et y ajouter des annotations (tags PDF) sans réorganisation des primitives du document électronique. Ces techniques ont été appliquées avec des résultats intéressants dans plusieurs travaux [BAG 04, HAR 04, LOV 95]. L'objectif des techniques de restructuration est de représenter le document électronique en utilisant un format différent du PDF, par exemple XML, pour permettre d'accéder facilement l'information. Le cas le plus intéressant de restructuration est celui de la ré-ingénierie, qui vise à réorganiser le contenu du document en fonction des structures découvertes [ANJ 01-1, CHA

05, DEJ 06, FUT 03, RAH 03, JPE, XED]. La conversion est un cas particulier de restructuration dans lequel aucune structure n'est extraite, le fichier PDF étant simplement transformé dans un format plus facile à manier [BLO 06].

3 XCDF, un format canonique

Les méthodes d'extraction et de restructuration de documents PDF suggèrent la définition d'un nouveau format pour la représentation de l'information d'une manière unique et structurée. Un format canonique pour la représentation des documents électroniques est indispensable afin de permettre aux utilisateurs et aux chercheurs d'accéder facilement à cette information et ainsi de pouvoir lui appliquer des traitements de plus haut niveau.

Il est important de clairement différencier deux concepts majeurs: la structure physique et la structure logique. La structure physique exprime la mise en page d'un document, elle est universelle : un format unique pour la description de tout document statique. A l'inverse, la structure logique exprime les fonctions logiques et hiérarchiques des diverses parties d'un document. Ces fonctions dépendent de la classe d'un document ainsi que du type d'application visé. Le format canonique se positionne comme un format générique et compact pour l'expression de la structure physique des documents statiques. Ce format doit permettre une manipulation aisée de son contenu pour tout traitement ultérieur; une extraction de structure logique simple doit pouvoir être effectuée par une feuille de style XSLT par exemple.

La figure 1 présente le rôle central du format XCDF. Bien que ce format puisse théoriquement représenter tout document statique, il n'est actuellement extrait qu'à partir de documents PDF par XED, un outil spécialement développé pour l'occasion (cf. section 4). Des applications utilisant le format canonique ont également été développées (cf. section 5).

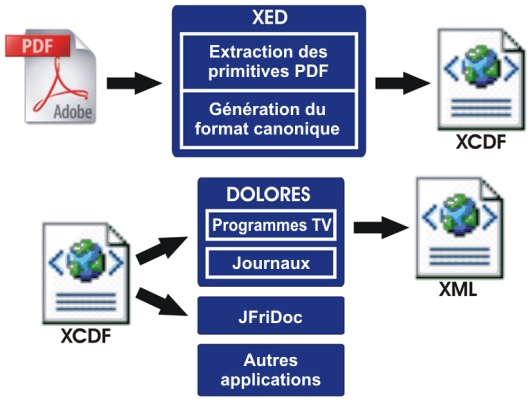

FIG. 1 - XCDF, un format pivot

La définition de ce format canonique doit être simple, universelle, complète et unique. **Simplicité :** Le format doit simplifier au maximum sa lecture et son utilisation. **Universalité :** Tout document statique doit pouvoir être représenté dans ce format. Sa mise en page doit être préservée à l'impression. **Complétude :** Aucune perte d'information visuelle ne doit survenir entre un document source et sa version canonique. Toutes les données typographiques et topologiques doivent être conservées. **Unicité :** Toutes les primitives (textes, graphiques et images) doivent être représentées d'une manière unique. Contrairement au format PDF, un affichage visuel donné ne doit avoir qu'une seule représentation possible.

Il est donc nécessaire de restructurer l'information brute extraite du PDF pour la convertir en XCDF. En effet, la représentation des primitives du format canonique est très structurée. Les primitives textuelles sont découpées en blocs homogènes, eux-mêmes découpés en lignes contenant des primitives lexicales (mots, signes de ponctuation, nombres, espaces et caractères spéciaux). Les primitives graphiques sont représentées sous forme de lignes et courbes de Bézier, les images par leurs boîtes englobantes ainsi qu'une référence sur leurs fichiers sources. La figure 2 expose la DTD partielle du format canonique. Les principales balises XML y sont décrites tandis que leurs attributs ont été omis par soucis de synthèse.

```
<!ELEMENT document (fonts?,page+)>
<!ELEMENT fonts (font+)>
<!ELEMENT page (image*,graphic*,
                textblock*)>
<!ELEMENT graphic (path+)>
<!ELEMENT textblock (textline+)>
<!ELEMENT textline (token+)>
```

FIG. 2 - DTD du format canonique

Chaque primitive visible contient des attributs sur sa position, sa couleur, son style ainsi que d'autres spécifiques à sa fonction. Par exemple, la balise "token" (unité lexicale) possède un attribut spécifique nommé "content" représentant une primitive textuelle encodée dans le standard UTF-8 (pour plus de détail, cf. [BLO 06]).

Bien que le format XCDF soit assez évolué, il fait encore l'objet d'étude pour certaines de ses caractéristiques comme la gestion des formes ou celle des symboles composés de plusieurs graphiques. La définition de ressources comme les polices et la couleur n'ont pas encore atteint un niveau totalement satisfaisant.

4 Extraction du format canonique

XED est l'outil développé permettant la transformation de documents PDF vers XCDF. Il est constitué de trois modules distincts effectuant respectivement 1) la lecture d'un fichier PDF ; 2) son analyse et restructuration vers le format canonique ; et 3) la lecture d'un fichier XCDF. L'algorithme de ré-ingénierie implémenté dans XED a été exécuté sur un large ensemble de données allant de documents à structures simples vers d'autres à structures complexes comme les journaux. Les résultats obtenus se sont révélés très satisfaisants ceci sans aucune calibration spécifique du système à un ensemble de données. Les sous-sections suivantes présentent successivement les trois modules développés ainsi qu'une évaluation de la génération du format canonique.

4.1 Module de lecture d'un fichier PDF

Le premier module convertit un fichier PDF en une arborescence d'objets Java équivalente. Pour ce faire, le fichier PDF source est d'abord lu. Puis, l'information ASCII et binaire contenue y est décodée. Les primitives textes, graphiques et images extraites peuvent alors être normalisées. La position de chaque primitive est calculée en coordonnées absolues, tandis que celle contenue dans un fichier PDF dépend de transformations géométriques spécifiées par des matrices relatives à l'état graphique courant. Finalement, l'arborescence d'objets Java équivalente au fichier PDF d'origine est générée. Une description plus détaillée de ce module a été présentée dans [HAD 04, RIG 04].

4.2 Module de restructuration vers le format canonique

Le deuxième module intervient lors de la reconstitution de la structure physique du document électronique [BLO 06]. L'arbre Java généré à la sortie du premier module est analysé et traité afin de le représenter dans le format canonique. La tâche principale effectuée à ce stade est la fusion des primitives textuelles en unités lexicales, lignes et blocs (cf. figure 3).

FIG. 3 - Fusion des primitives textuelles

La fusion des lignes en blocs s'appuie sur le paradigme des composantes connexes en traitement d'image, les lignes de texte jouant le rôle de pixels et le voisinage étant exprimé par des règles d'adjacence. Cette technique assure une fusion en blocs respectueuse des sous blocs (cf. figure 4).

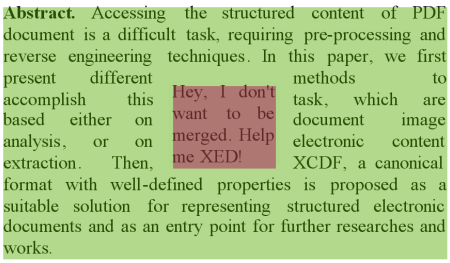

FIG. 4 - Fusion des lignes en utilisant le paradigme des composantes connexes

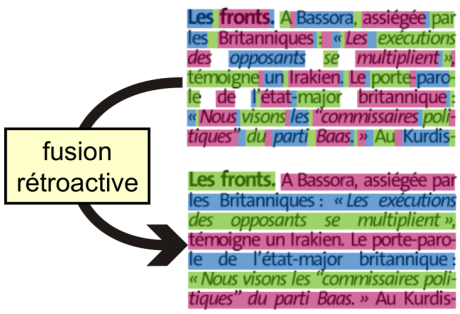

FIG. 5 - Résultat de la fusion rétroactive

Suite à la construction des blocs, une fusion en ligne dite "rétroactive" effectue un second passage sur les lignes d'un bloc, ceci permet de récupérer les sur-segmentations potentielles, généralement dues à une justification des paragraphes (cf. figure 5).

Ce module est actuellement adapté pour les langues latines, les langues arabes et orientales n'étant pour l'instant pas prises en compte. Un point essentiel de notre système réside dans la généralisation des méthodes de fusion. En effet, l'extraction du format canonique sur n'importe quel fichier PDF ne nécessite aucun paramétrage particulier ; les seuils n'étant pas statiques, mais générés dynamiquement par des caractéristiques textuelles. Relevons que ces seuils tendent toujours vers des minimums, cela pour éviter toute sous-segmentation. La figure 6 présente un extrait du fichier XCDF généré à partir de l'exemple de la figure 3.

```
<textblock x="81" y="374" w="145" h="137">
  <textline x="81" y="374" w="145" h="32">
    <token size="32" content="Bush"/>
    <token size="32" content=" "/>
    <token size="32" content="plans"/>
  </textline>
  <textline x="81" y="409" w="140" h="32">
    <token size="32" content="to"/>
    <token size="32" content=" "/>
    <token size="32" content="support"/>
  </textline>
  [...]
</textblock>
```

FIG. 6 - Le format canonique, XCDF

4.3 Module de lecture d'un fichier XCDF

Le module de lecture de fichiers XCDF permet de recréer l'arborescence d'objets Java correspondant à un fichier canonique extrait. Il devient ainsi possible de considérer le format canonique comme format pivot permettant ensuite d'effectuer des analyses de plus haut niveau sur les documents.

De plus, l'API offre une implémentation graphique de chaque primitive XCDF permettant la création aisée d'une interface graphique évoluée.

4.4 Evaluation de la génération du format canonique

Une évaluation de la génération du format canonique a été effectuée sur un ensemble représentatif de unes de journaux latins: La Liberté, Le Monde et The International Herald Tribune. Pour chacun de ces journaux, dix unes ont été extraites et représentées dans le format canonique. Les pourcentages d'unités lexicales, de lignes et de blocs de texte correctement détectés naviguent tous entre 97% et 99% [RIG 05-1].

5 Application du format XCDF

Cette section présente trois cas d'applications basées sur le format XCDF : un analyseur logique, un navigateur multimédia et un outil de création de PDF purs.

5.1 Logical Restructuring : Dolores

A partir du format canonique, le système Dolores (**Do**cument **Lo**gical **Res**tructuring) vise à récupérer la structure logique de documents. Contrairement à la structure physique extraite par XED, la structure logique n'est pas unique. Il est indispensable de définir une structure logique spécifique à l'application ainsi qu'à la classe de documents considérés. Dolores se focalise sur le développement de processus de restructuration permettant de reconstituer l'information logique à partir du format canonique. Cette sous-section présente un exemple concret d'application de Dolores pour l'extraction automatique de programmes TV, puis expose brièvement une application plus complexe : la restructuration logique de journaux.

Des programmes de télévision ont été téléchargés durant une semaine à partir du site Internet de la télévision suisse romande. Six chaînes différentes ont été traitées donnant un total de 42 fichiers PDF. Sur cette base, 42 fichiers XCDF correspondant ont été générés, après avoir 1) supprimé les informations superflues, 2) localisé les horaires, 3) récupéré le texte associé et 4) étiqueté les blocs de texte suivant la graisse de la fonte (cf. figure 7).

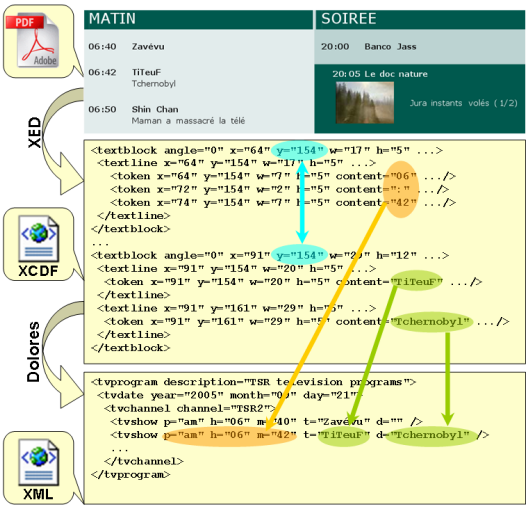

FIG. 7 - Extraction du format logique d'un horaire TV

Les résultats parfaits obtenus sur un problème simple ont prouvé l'utilité et la facilité d'utilisation du format canonique et ouvert la perspective vers une restructuration de documents à structure logique complexe comme les journaux. Cette classe de documents possède en effet des caractéristiques intéressantes : une mise en page riche contenant un grand nombre d'informations typographiques et topologiques ainsi que des hiérarchies logiques profondes. Aucune information topologique ou typographique n'est nécessaire dans la structure logique puisqu'elle fait directement référence au format canonique au travers d'identificateurs uniques. Un article sur deux pages n'aura ainsi pas d'incidence sur la représentation du format. Un processus utilisant des réseaux de neurones ainsi qu'un arbre d'automates déterministes nous a permis d'extraire cette structure logique à partir du format canonique (cf. [BLO-06]). Les premières expériences effectuées ont donné des résultats probants.

5.2 Le navigateur multimédia JFriDoc

JFriDoc, un navigateur de réunions multimédia centré document (cf. figure 8), utilise XED pour l'extraction des structures physiques de documents. Ces structures sont ensuite annotées manuellement avec des informations logiques à l'aide de l'outil Inquisitor [RIG 05-2]. Enfin, ces annotations sont comparées automatiquement avec la transcription des dialogues de la réunion afin de les aligner thématiquement et d'enrichir les documents d'indices temporels.

FIG. 8 - JFriDoc, navigateur multimédia

5.3 Génération de PDF épurés

Un projet récent concerne la création de fichiers PDF épurés, dans lesquels un utilisateur puisse copier-coller les données textuelles tout en respectant les informations structurelles du document, ces informations pouvant également servir à une indexation plus précise et aisée. L'application générant de tels fichiers interprète simplement le format canonique, puis le convertit en PDF, en utilisant uniquement un sous-groupe d'opérateurs PDF préalablement déterminés et non redondants. Les structures physiques sont ainsi préservées 1) en regroupant dans le même objet PDF un bloc XCDF sans segmenter les mots et 2) en ajoutant les informations de la boîte englobante sous forme de tags. Cette deuxième technique peut également être utilisée afin d'enrichir le document PDF avec des structures logiques ainsi que d'autres annotations. Un premier prototype d'application est en phase de développement et n'a donc pas encore pu être évalué.

6 Conclusion

Cet article présente XCDF, un format canonique pour la représentation de documents électroniques statiques de façon unique et structurée, facilitant les recherches et analyses de plus haut niveau. Une taxonomie des systèmes d'extraction et d'analyse de documents PDF a été brièvement présentée, en organisant les systèmes suivant leur méthode d'analyse. Cet article a ensuite détaillé le format XCDF ainsi que XED, le système développé pour la restructuration de documents PDF. Ce dernier est composé de trois modules : le premier nécessaire à l'extraction des données d'un document PDF, le deuxième analysant les données extraites et les

restructurant au format canonique et le troisième offrant une API de manipulation de fichiers XCDF. La dernière section de cet article expose trois applications de XCDF : Dolores, un système pour la restructuration logique de documents électroniques, le navigateur multimédia JfriDoc, puis un prototype pour la génération de PDF conformes à XCDF. Dans le futur, nos travaux se concentreront sur 1) l'extension de la spécification XCDF, 2) le développement et l'évaluation de Dolores sur la classe des journaux et 3) l'évolution de l'application générant des PDF épurés.

Références

[ADO] Adobe PDF reference, http://partners.adobe .com/asn/tech/pdf/specifications.jsp

[ANJ 01-1] Anjewierden A., AIDAS: Incremental logical structure discovery in PDF document, *ICDAR'01*, 2001, pp. 374-377

[ANJ 01-2] Anjewierden A., Kabel S., Automatic indexing of documents with ontologies, *BNAIC'01*, 2001, pp. 23-30

[BAG 04] Bagley S.R., Brailsford D.F., Hardy, M.R.B, Creating reusable well-structured PDF as a sequence of component object graphic (COG) elements, *DocEng'03*, 2003, pp. 58-67

[BLO 06] Bloechle J.-L., Rigamonti M., Hadjar K., Lalanne D., Ingold, R., XCDF: A Canonical and Structured Document Format, *DAS'06*, 2006, pp. 129-140

[CHA 05] Chao, H., Fan, J., Capturing the Layout of electronic Documents for Reuse in Variable Data, *ICDAR'05*, 2005, pp. 940-944

[DEJ 06] Dejan H., Meunier J.L., A System for Converting PDF Documents into Structured XML Format, *DAS'06*, 2006, pp. 129-140

[FUT 03] Futrelle R.P., Shap M., Cieslick C., Grimes, A.E., Extraction, layout analysis and classification of diagrams in PDF documents, *ICDAR'03*, 2003, pp. 1007-1012

[HAD 03] Hadjar K., Ingold R., Arabic Newspaper Page Segmentation, *ICDAR'03*, 2003, pp. 895-899

[HAD 04] Hadjar K., Rigamonti M., Lalanne D., Ingold R., Xed: a new tool for eXtracting hidden structures from Electronic Documents, *DIAL'04*, 2004, pp. 212-221

[HAR 04] Hardy M.R., Brailford D., Thomas P.L., Creating Structured PDF Files Using XML Templates, *DocEng'04*, 2004, pp. 99-108

[JPE] JPEDAL, http://www.jpedal.org

[LAW 99] Lawrence S., Bollacker K., Lee Giles C., Indexing and Retrieval of Scientific Literature, *CIKM'99*, 1999, pp. 139-146

[LOV 95] Lovegrove W.S., Brailsford D.F., Document analysis of PDF files: methods, results and implications, *Electronic Publishing*, 1995, pp. 207-220

[PAK 98] Paknad M.D, Ayers R.M., Method and apparatus for identifying words described in a portable electronic document, *U.S. Patent 5,832,530*, 1998

[RAH 03] Rahman F., Alam H., Conversion of PDF documents into HTML: a case study of document image analysis, *Asilomar CSS'03*, 2003, pp. 87-91

[RIG 04] Rigamonti M., Hadjar K., Lalanne D., Ingold R., Xed: un outil pour l'extraction et l'analyse de documents PDF, *CIFED'04*, 2004, pp. 85-90

[RIG 05-1] Rigamonti M., Bloechle J.-L., Hadjar K., Lalanne D., Ingold R., Towards a Canonical and Structured Representation of PDF Documents through Reverse Engineering, *ICDAR'05*, 2005, pp. 1050-1054

[RIG 05-2] Rigamonti M., Lalanne D., Evéquoz F., Ingold R., Browsing multimedia archives through implicit and explicit cross-modal links, *MLMI'05*, 2005, pp. 14-25

[XED] XED online, http://diuf.unifr.ch/diva/xed