# Xed : un outil pour l'extraction et l'analyse de documents PDF

Maurizio Rigamonti, Karim Hadjar, Denis Lalanne, Rolf Ingold

# Xed : un outil pour l'extraction et l'analyse de documents PDF

Maurizio Rigamonti – Karim Hadjar – Denis Lalanne – Rolf Ingold

Groupe DIVA, Département d'Informatique de l'Université de Fribourg
Chemin du musée 3, 1700 Fribourg, Suisse

{prénom, nom}@unifr.ch

**Résumé** *: PDF est devenu le format de prédilection pour l'échange de documents. Cependant, son utilisation se limite à la visualisation et à l'impression. De nouveaux besoins d'extraction du contenu et de recherche sont nés du fait de l'utilisation grandissante du format. Pour cette raison, de nouveaux outils ont fait leur apparition sur le marché. Ces derniers se limitent malheureusement à l'extraction automatique du contenu, sans prendre en considération ni la structure physique ni la structure logique du document.*

*Nous proposons, dans cet article, une nouvelle approche palliant les insuffisances des outils d'extraction. Cette méthode combine a) des méthodes d'extraction appliquées aux fichiers PDF avec b) des méthodes d'analyse d'image de document visant à extraire la structure physique. Cet article décrit les différentes étapes nécessaires pour réaliser cette tâche.*

**Mots-clés** *: Extraction de documents PDF, analyse d'image de documents, structures physiques et logiques, reconstitution de l'ordre de lecture, XML.*

## 1 Introduction

De nos jours, de nombreux documents électroniques sont sauvegardés sous forme de fichier PDF [ADO]. Le format PDF (acronyme de **P**ortable **D**ocument **F**ormat) est un format universel d'échange de documents, qui permet de décrire la mise en page d'un document composé de texte, d'images et de graphiques complexes. PDF a été conçu pour succéder à PostScript quand l'Internet est devenu un phénomène de masse : le besoin d'échanger les documents nécessitait un format indépendant d'une plate-forme ou d'une application spécifique.

Cependant, la plupart des documents convertis dans ce format ne sont pas structurés, en dépit des fonctionnalités qui *a priori* sont offertes par le format PDF. Par exemple, un bloc de texte correspond rarement à un bloc physique ou logique de texte. Vraisemblablement, deux raisons sont à l'origine du problème : l'ordre de création du fichier et la génération faite par des outils automatiques. Pour simplifier, la position des objets PDF (texte, images, graphiques, styles, etc. ) dans un fichier correspond à leur ordre de génération. La génération des fichiers PDF faite par des convertisseurs automatiques pourrait être une autre justification à l'ordre chaotique des objets : le format offre une syntaxe riche et des fonctionnalités très complexes, donc difficiles à manipuler. Dans l'état actuel, un fichier PDF se limite à décrire l'**apparence** d'un document et toute notion structurelle a été abandonnée. Par conséquent, il ressort qu'une des seules limites du format PDF est la difficulté d'indexer le contenu du document. Un exemple trivial est donné par la recherche du titre «Orange mécanique » dans une base de données constituée de journaux au format PDF. Tous les journaux qui contiennent ces deux mots clés conjointement seront restitués. Mais l'absence d'informations topologiques et logiques implique que n'importe quel journal qui contient les deux mots fera partie du résultat. Ainsi, un journal contenant un article sur «l'orange, fruit de la saison» et un article sur «un nouveau bijou de la mécanique» pourrait être restitué, sans pourtant satisfaire la requête de l'utilisateur. Afin de répondre correctement à des requêtes constituées de plusieurs mots clés, comme c'est souvent le cas, il est donc nécessaire de reconstituer les structures aussi bien physiques que logiques d'un document PDF.

Plusieurs produits et études ont essayé d'extraire les primitives de base (texte, images et graphiques) ou d'enrichir et de structurer le format [LOV 95, BAG 03] d'un fichier PDF. Cependant, très peu de produits existants se préoccupent d'analyser le résultat obtenu. Aucun d'entre eux ne cherche à reconstituer la structure logique du document. Nous sommes convaincus que a) l'extraction des primitives d'un fichier PDF et conjointement b) l'application d'algorithmes classiques d'analyse de l'image d'un document permettront non seulement de reconstituer la structure physique du document, mais aussi sa structure logique. Ainsi, l'indexation, l'archivage, la ré-édition d'un document PDF, etc. s'en trouveront considérablement améliorés.

L'approche que nous proposons consiste à extraire tous les objets primitifs d'un fichier PDF, à les convertir dans un format plus maniable et plus structuré comme par exemple en XML. Ainsi, le document pourra être plus facilement manipulé et analysé. De plus, les résultats des analyses contribueront à structurer davantage l'information, structure qui pourra être ré-injectée dans le document PDF.

Cet article présente notre méthode et notre outil Xed (e**X**tracting **e**lectronic **d**ocuments). Dans une première phase, le texte, les images et les graphiques sont extraits et transformés soit en SVG (Scalable Vector Graphics) soit en XML (Extensible Markup Language). Ensuite, le résultat est mis en correspondance avec celui des

méthodes d'analyse de la structure physique du document. Cette collaboration permet non seulement de structurer les informations extraites dans des structures de document, mais aussi de bénéficier de la complémentarité des deux méthodes. Finalement, le texte à l'intérieur de chaque bloc physique est reconstitué en utilisant les informations topologiques contenues dans le PDF.

Cet article se compose de la façon suivante : la section 2 présente l'état de l'art ; la section 3 motive l'analyse conjointe du document électronique et de l'image du document ; la section 4 présente notre approche et l'outil Xed, contenant des composantes d'extraction du contenu et d'analyse de la structure physique ; la section 5 discute de l'évaluation de notre approche ; finalement, la section 6 conclut cet article et elle présente les perspectives futures.

# 2 État de l'art

Cette section est composée de deux sous-sections, qui présentent respectivement l'état de l'existant du point de vue des outils d'extraction et l'état de l'art en ce qui concerne l'analyse d'image de document.

## 2.1 Outils pour l'extraction et l'analyse de PDF

Plusieurs produits permettent d'extraire et, dans certains cas, d'analyser des fichiers PDF. Une analyse de ces applications et librairies est discutée dans cette section.

- **JPedal** [JPE] est une librairie Java qui permet l'extraction de primitives contenues dans un document PDF et l'analyse de la structure physique. Le site officiel de JPedal offre des services d'extraction pour la conversion en XML du document PDF et d'analyse du résultat.
- **SVGImprint** de **Mattercast** [MAT] est un produit commercial qui extrait toutes les primitives qui se trouvent dans un fichier PDF et les convertit en SVG.
- **Glance** [GLA] offre plusieurs applications pour l'extraction du contenu d'un fichier PDF, notamment pour le texte et pour les images. Il convertit le texte en ASCII ou en Unicode.
- **BCL** [BCL] met à disposition des outils pour le traitement de documents PDF. **BCL Jade** est un plug-in pour Adobe Acrobat qui permet d'extraire du texte, des données tabulaires et des graphiques.

Ces produits ont été testés sur une base de données hétérogène, composée de documents de différents types (journaux, rapports techniques, diapositives, etc.) générés à l'aide de différentes applications.


FIG. 1 - Un détail de la une de « Le Soir » du 10 septembre 2003

La figure 1 et la table 1 représentent respectivement un extrait de la première page du journal « Le Soir » du 10 septembre 2003 et le résultat de son extraction avec l'outil **pdw** de **Glance** (l'angle, la police et la taille des caractères ont étés omis).

| X | Y | Width | Text |
|---|---|---|---|
| 83.3 | 1421.8 | 20.3 | Co |
| 104.0 | 1421.8 | 15.3 | ct |
| 118.9 | 1421.8 | 29.5 | eau |
| 83.3 | 1403.8 | 14.4 | le |
| 101.4 | 1403.8 | 34.4 | trop |
| 139.4 | 1403.8 | 54.2 | connu, |
| 83.3 | 1385.8 | 14.4 | le |
| 101.4 | 1385.8 | 29.9 | mal |

TAB. 1 – le résultat de son extraction avec **pdw** de **Glance**

La table 1 indique clairement que l'ordre de lecture des lignes de texte n'est pas reconstitué dans sa totalité. Par exemple, le mot « Cocteau » n'a pas été recomposé. La table ci-dessous résume les fonctionnalités des produits existants sur la base de six critères : les types de primitives extraites du PDF (texte, graphiques et images), le format de sortie des données (SVG ou XML) et le respect de l'ordre de lecture.

| | Extraction | | | SVG | XML | Ordre de lecture |
|---|---|---|---|---|---|---|
| | Texte | Images | Graphiques | | | |
| **JPedal** | oui | oui | - | - | oui | oui |
| **BCL** | oui | oui | - | - | - | - |
| **Glance** | oui | oui | - | - | - | - |
| **SVG Imprint** | oui | oui | oui | oui | - | - |

TAB. 2 – Comparaison des produits existants, par rapport aux primitives extraites du PDF, au format d'extraction et au respect de l'ordre de lecture

La table 2 montre que seulement JPedal respecte l'ordre de lecture, qui est reconstitué à l'aide d'algorithmes d'analyse. Vraisemblablement, le problème de l'ordre de lecture non respecté est intrinsèque à la génération des documents PDF. La visualisation étant le but principal, la structure interne du document n'est pas prise en compte. Pourtant, nous supposons que la problématique principale à laquelle nous sommes confrontés est que la structure d'un fichier PDF ne respecte ni la structure physique, ni la structure logique du document. Nous proposons donc d'appliquer une analyse classique de l'image du document, au lieu d'essayer d'introduire des règles complexes et des heuristiques dans les outils d'extraction de la version électronique du document.

## 2.2 Analyse de la structure d'un document

L'analyse des structures d'un document est un problème très important aussi bien dans le domaine de l'analyse d'image de documents, que dans les systèmes de reconnaissance des documents. Un des aspects

principaux de l'analyse de la structure d'un document est l'extraction des propriétés physiques et logiques des régions d'un document. Les propriétés physiques mettent en évidence la topologie du document, tandis que les propriétés logiques représentent la fonction des régions (e.g. titre, article, section, figure, etc.).
Bien que le domaine de l'analyse de document soit en permanente évolution, tous les buts n'ont pas été atteints. Par exemple, l'analyse de documents à structures complexes comme les journaux ne donne des résultats satisfaisants [ANT 03, HAD 01, HAD 03] qu'après l'intervention d'un utilisateur humain. Cependant, souvent l'analyse de documents se focalise sur des documents papier qui ont été scannés. Ces images présentent plusieurs défauts : basse qualité et résolution, bruit et déformations, etc. Dans notre cas, nous utilisons des images idéales de très grande qualité, créés à partir de documents électroniques.
La littérature propose de nombreux algorithmes et techniques pour l'analyse géométrique et structurelle des documents, comme par exemple des approches basées sur la morphologie, les profils de projection, l'analyse des textures, l'analyse de la structure du fond, etc. [NAG 00]. Les méthodes d'analyse se sont concentrées dans un premier temps sur des structures physiques simples [HAR 94] et ont récemment abordé des structures plus complexes, comme celles des quotidiens. Actuellement, la philosophie CIDRE[1] suggère d'utiliser des modèles de documents et des systèmes incrémentaux d'apprentissage assisté par l'utilisateur [HAD 02]. Les modèles de document [HU 01] sont définis par un opérateur humain ou pendant une phase d'apprentissage qui nécessite de nombreuses données fonds de vérité.

# 3 Une analyse conjointe du document électronique et de l'image du document

Notre approche propose de fusionner des techniques d'analyse du document électronique PDF avec des méthodes d'analyse de l'image du document. Les raisons pour lesquelles les deux méthodes doivent coexister et être fusionnées sont multiples et concernent principalement soit l'extraction du contenu soit l'extraction des structures du document. Les sous-sections 3.1 et 3.2 présentent respectivement les motivations liées à ces deux aspects.

## 3.1 Extraction du contenu du document

La méthode d'extraction du contenu d'un document PDF que nous proposons profite de la combinaison de l'analyse d'image et de l'extraction du fichier électronique pour palier aux limites intrinsèques de chacune des deux techniques prises indépendamment.
Les documents PDF contenant des images de document ne peuvent pas uniquement être analysés en utilisant l'extraction électronique. C'est le cas des anciens documents qui ont été digitalisés et inclus en tant qu'image dans le document PDF. Dans ce cas, il est nécessaire d'utiliser des techniques d'analyse d'image pour extraire la structure physique et d'OCR pour reconstituer le texte.
Les performances des OCRs et des OFRs ne peuvent pas garantir 100% de reconnaissance, particulièrement pour des langues non latines. L'information textuelle étant présente dans les documents PDF « idéaux », i.e. ne contenant pas d'image de texte, il est naturel de l'extraire directement en analysant le fichier électronique. L'extraction de fontes permet d'accéder à leur forme vectorielle, ceci nous permet d'éviter de recourir à la reconnaissance de fontes.
La grande majorité des documents modernes sont plutôt « idéaux », et très rarement constitués d'image de documents textuels. Cependant, afin de traiter tous les types de document, il est nécessaire de combiner des techniques d'extraction du contenu électronique dans le fichier PDF avec des techniques d'analyse d'image du document. De plus, les deux techniques peuvent s'enrichir mutuellement et garantir une segmentation plus robuste.

## 3.2 Extraction des structures du document

Concernant l'extraction de la structure physique, à première vue, il serait plus naturel de l'extraire directement à partir du fichier PDF, en se servant de sa structuration interne. Notre expérience nous a toutefois montré que cela pose de grandes difficultés parce que les informations structurelles ne sont pas toujours fiables. Par exemple, dans des documents multi-colonnes, l'ordre d'apparition des blocs de texte ne reflète en général pas l'ordre de lecture. Pire, il arrive que des portions de phrase ou des mots isolés n'apparaissent pas dans leur contexte mais de manière isolée à la fin d'un fichier. Nous avons de bonnes raisons de penser que ce type d'imperfection dépend de l'historique du document et des logiciels qui ont servi à le produire.
De plus, la séparation en zones de texte et en zones graphiques, tâche relativement aisée en appliquant des méthodes d'analyse d'image, peut s'avérer très compliquée en utilisant uniquement l'information contenue dans le PDF, surtout lorsqu'une zone graphique est constituée d'éléments textuels.
D'autre part, l'analyse des structures contenues dans le fichier PDF se révèle indispensable afin de segmenter une composition complexe en sous-zones textuelles, d'image et graphiques (ex. figure 2). Cette tâche serait particulièrement ardue en utilisant que l'analyse d'image avec des modèles de documents ou des règles.

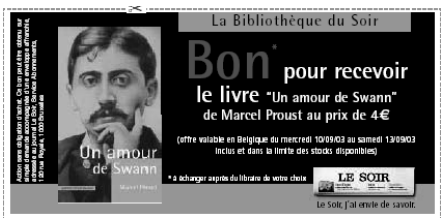


FIG. 2 – Un extrait de la une du journal « Le Soir » du 10 septembre 2003

[1] CIDRE acronyme pour Cooperative and Interactive Document Reverse Engineering, financé par le fond national Suisse pour la recherche scientifique, code 2000-059356.99-1

La figure 2 présente un exemple extrait de la une du journal « Le Soir », qui contient une mise en page complexe, une composition de textes, images et graphiques. La figure 3 décrit les primitives, contenues dans le fichier PDF, qui composent cette région du document : les rectangles blancs mettent en évidence les graphiques et les gris les images (le texte a été omis).

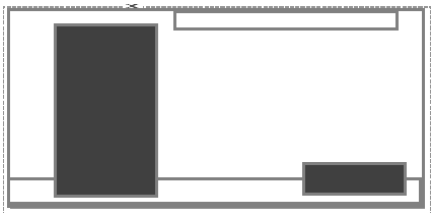

FIG. 3 – Les primitives extraites

Pour conclure, nous pensons que l'analyse à partir de l'image générée à partir du PDF présente pour nous l'avantage de considérer une représentation normalisée et cartésienne, quasi universelle.

La section qui suit présente notre approche, l'architecture de l'outil que nous avons développé, Xed, le composant d'extraction du contenu électronique, le composant d'extraction de la structure physique basé sur l'analyse d'image, et enfin la fusion de ces deux composantes. Finalement, une application utilisant notre outil est présentée permettant de restituer l'ordre de lecture d'un document.

# 4 Xed

Cette section est divisée en quatre sous-sections : la première sous-section présente l'architecture de Xed, qui permet d'extraire et d'analyser un document PDF ; la deuxième décrit l'algorithme d'analyse de la structure physique ; la troisième présente la mise en correspondance des objets extraits et des résultats de l'analyse de l'image ; enfin, la dernière sous-section finalise la reconstitution de l'ordre de lecture d'un document.

## 4.1 L'architecture de Xed

L'architecture de Xed est divisée en trois phases : la lecture, l'extraction et, enfin, l'analyse (voir figure 4).

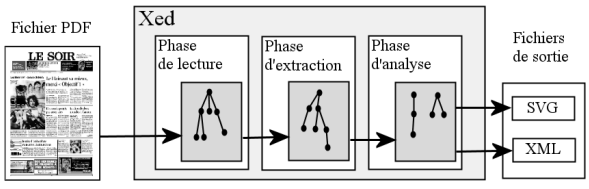


FIG. 4 - Les trois phases de Xed.

La lecture est la phase de plus bas niveau. Elle interprète le fichier PDF afin de le préparer pour l'extraction. Le fichier est d'abord transformé en arbre, dans lequel chaque objet PDF est stocké. Le contenu restant le même, la seule différence entre un objet PDF et son double dans l'arborescence est sa représentation.

L'objectif principal de la lecture est l'homogénéisation des données. Cette phase se justifie par le fait que dans certains cas le format PDF accepte plusieurs représentations pour un même objet. Par exemple, une chaîne de caractères peut être exprimée à l'aide d'une représentation soit hexadécimale soit littérale. Dans l'arbre qui représente le document dans Xed, la chaîne est toujours représentée littéralement.

Dans la phase suivante, Xed extrait les primitives de base du document PDF. L'extraction revient à parcourir l'arbre construit dans la phase précédente et à interpréter les informations contenues dans celui-ci pour reproduire le document original. Ce processus est analogue à celui qui est accompli par chaque outil d'impression de documents PDF. L'unique différence est que le document n'est pas imprimé, mais reconstruit dans une représentation interne (dite *document abstrait* dans la suite de cette section). Le document abstrait est créé page après page. La première étape consiste à charger toutes les ressources utilisées par la page, c'est à dire les polices de caractères, les espaces de couleurs, les styles et les états graphiques étendus. Une fois cette tâche terminée, Xed crée tous les objets primitifs contenus dans la page, c'est à dire le texte, les images et les graphiques (par exemple, des courbes). Durant la création des primitives, l'état graphique est mis à jour. Cet état permet de connaître la position de chaque objet et de définir son style (par exemple, la couleur, l'épaisseur de la ligne, les trames de remplissage, etc.). Une des fonctions du document abstrait est de conserver cet état graphique ; chaque primitive est strictement corrélée à la précédente et elle seule ne contient pas toute l'information qui permet de la situer dans le contexte du document. Par exemple, une figure dépend à la fois de la forme et d'un style. À son tour, le style contient des informations concernant la couleur et le trait, qui se trouvent dans l'état graphique à un moment donné. Une fois que le document abstrait a été créé, chaque objet a été enrichi avec des informations graphiques, qui sont implicites dans le document PDF.

Finalement, la dernière phase consiste à analyser les objets extraits du PDF. Le document abstrait est directement manipulé et sa morphologie est adaptée lorsque nécessaire. Les sous-sections suivantes décrivent en détails un exemple d'analyse qui peut être introduite dans cette phase : l'algorithme de restitution de l'ordre de lecture.

Une fois que les trois phases précédentes sont achevées, Xed convertit toutes les informations contenues dans le document abstrait sous forme d'un document XML. Xed a été implémenté en Java et utilise deux librairies, une pour la phase de lecture et une pour l'extraction. La phase d'analyse est encapsulée dans l'application principale. Enfin, les formats de sortie sont actuellement soit un format XML spécifique [HAD 04], soit du SVG. Cependant, l'architecture modulaire de Xed permet de générer en sortie toutes formes de format.

## 4.2 L'algorithme d'analyse de la structure physique du document

Dans la vision par ordinateur, le but de la segmentation d'images est de décomposer une image en régions homogènes, qui possèdent des propriétés similaires et

significatives. L'algorithme, utilisé pour segmenter une image de document, agit en plusieurs phases : a) d'abord, il extrait les filets et les cadres ; b) ensuite, il sépare le texte des images ; c) il reconnaît ensuite les lignes de texte et, enfin, d) il fusionne les lignes de texte en blocs. Chaque phase peut être appliquée indépendamment des autres. L'extraction de filets et des cadres utilise une approche ascendante. L'extraction des lignes de texte suit l'algorithme RLSA, suivie par la création de composantes connexes. Enfin, la fusion de lignes de texte en blocs homogènes est basée sur des règles.

Notre algorithme utilise des images TIFF non bruitées, générées à partir d'un fichier PDF. Le résultat de l'analyse est stocké dans un fichier XML, qui contient les informations concernant les filets, les images, les cadres, les lignes de texte extraites et les blocs formés par celle-ci.

### 4.3 Mise en correspondance des objets extraits avec Xed et par l'analyse de la structure du document

La mise en correspondance des résultats de l'analyse de la structure du document et des objets extraits du fichier PDF est montrée dans la figure 5. La sous-section précédente a montré quelle information est restituée par l'analyse logique du document. L'extraction des objets PDF permet de définir aisément des propriétés géométriques pour chaque primitive, c'est à dire la boîte englobante pour le texte, le rectangle défini par la hauteur et la largeur pour une image, et enfin des courbes et des polygones pour un graphique.

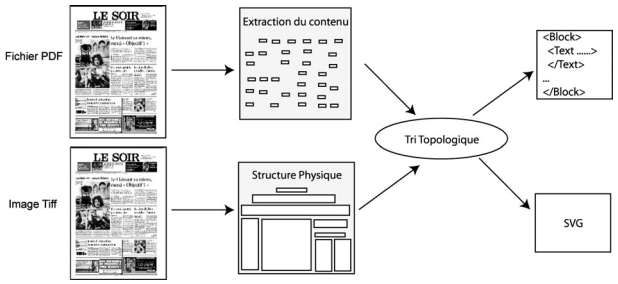


FIG. 5 - Tri topologique.

La mise en correspondance se réduit donc à calculer l'union entre les informations géométriques calculées pour les objets PDF et les informations topologiques obtenues par l'analyse de l'image. L'algorithme se compose de deux phases:

1. Pour chaque bloc reconnu par l'analyse de la structure, une liste vide d'objets est crée. De même, une liste des objets non assignés et une pour les objets non résolus sont initialisées. Chaque objet extrait du PDF est ensuite comparé à chaque bloc.
2. À la fin de la comparaison, trois cas peuvent se présenter :
   (a) Un objet PDF appartient à un **seul** bloc. Il est donc rajouté à la liste du bloc.
   (b) La boîte englobante de l'objet appartient à **plusieurs** blocs où elle présente **une** ou **plusieurs intersections** avec d'autres blocs. Dans ce cas, l'objet est ajouté à la liste des objets non résolus.
   (c) Si l'objet n'appartient à aucun bloc, il est introduit dans la liste des objets non assignés.

L'avantage de conserver une liste pour les objets non assignés et une pour les objets non résolus est de permettre au système d'analyse de la structure de corriger de faux résultats. Le cas d'un objet non résolu ou d'un objet non assigné se présente quand un bloc a été sur-segmenté ou quand sa surface est trop petite. De plus, un objet non assigné correspond au cas où un bloc n'a pas été créé, ce qui justifie le besoin de distinguer les deux listes.

### 4.4 Reconstitution de l'ordre de lecture dans un bloc

La reconstitution de l'ordre de lecture dans un bloc revient entre autre à regrouper le texte en lignes, qui sont à leur tour décomposées en mots. L'extraction de texte a démontré non seulement que la plupart des fichiers PDF ne sont pas structurés, mais aussi que les mots ne peuvent pas être considérés comme des objets primitifs. En effet, dans certains cas, un mot est décomposé en plusieurs sous-chaînes. De la même façon, le cas contraire peut se présenter ; plusieurs mots font alors partie d'un même objet de texte.

La reconstitution de l'ordre de lecture se fait à partir de la boîte englobant le texte. Chaque sous-chaîne est définie par ses coordonnées `x` et `y` (le repère `O(0,0)` est le coin en haut à gauche du document), la longueur, la hauteur et les informations concernant l'espace entre les caractères et les mots. Ces informations suffisent donc à recréer les lignes et les mots. Le procédé se déroule en trois phases distinctes :

1. Les objets sont triés par rapport à leur coordonnée `y`, de la plus petite à la plus grande. Les lignes sont ainsi partiellement reconstituées. Pour chaque région, les objets sont ensuite triés par rapport à leur cordonnée `x`.
2. Les objets, contenant du texte et se situant à une même hauteur, sont ensuite regroupés. Supposons que deux chaînes $t_1$ et $t_2$, avec $x_1$, $x_2$ et $w_1$, $w_2$ étant respectivement leur coordonnée horizontale à gauche et la longueur de leur boîte englobant, alors si $x_1 + w_1 >= x_2$, les objets sont fusionnés.
3. Etant donné que le PDF prend en compte le caractère espace ainsi que d'autres séparateurs, chaque chaîne de texte est parcourue. Si un symbole de séparation est trouvé, l'objet qui le contient est divisé en deux objets distincts.

## 5 Evaluation

L'approche que nous proposons dans cet article n'a pas encore été évaluée quantitativement. La base de donnée considérée jusqu'à présent est composée de document PDF idéaux, à structures complexes. Les résultats de

l'extraction que nous avons obtenus sur une centaine de unes de journaux francophones, anglophones, italophones et arabes sont satisfaisants à l'œil nu en ce qui concerne la préservation de la mise en page. Afin de s'assurer que l'extraction du contenu du document est correcte et qu'elle préserve effectivement la mise en page, il serait possible de prévoir une évaluation automatique en superposant l'image du document PDF avec celle obtenue par l'extraction et de calculer le taux de ressemblance à travers un calcul de distance. Enfin, afin d'évaluer les performances d'extraction des structures physiques et logiques, la constitution d'un fond de vérité devra être considéré, à l'aide d'outils semi-automatiques supervisés par un opérateur humain [RIG 03] et dotés d'apprentissage [HAD 02]. L'évaluation des résultats de l'analyse conjointe est strictement dépendante de celles de deux techniques.

# 6 Conclusion

L'extraction de structures de haut niveau d'un document PDF s'avère être une tâche difficile et complexe, qui, en contrepartie, présente de nombreuses applications pratiques. Pour aboutir dans l'extraction de structures de haut niveau, il faut absolument disposer d'outils de bas niveau qui permettent d'extraire toute primitive d'un document PDF (texte, graphiques et images) mais aussi des structures cachées (structures physiques, logiques, ordre de lecture, etc.). Xed est l'approche que nous proposons afin d'accomplir cette tâche. L'originalité de notre méthode vient de la fusion de deux techniques : a) l'extraction de primitives PDF avec b) des techniques classiques d'analyse de l'image d'un document. Par exemple, l'utilisation d'algorithmes d'analyse de la structure physique d'un document nous a permis d'enrichir les informations brutes contenues dans un document PDF et de reconstituer l'ordre de lecture à l'intérieur d'un bloc de texte.

# Références


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

[ANT 03] Antanacopoulos A., Gatos B., Karatzas D., « ICDAR 2003 Page Segmentation Competition », ICDAR2003, août 2003, Edinburgh (Scotland), pp. 688-692.

[BAG 03] Bagley S. R., Brailsford D. F., Hardy M. R. B., « Creating reusable well-structured PDF as a sequence of component object graphic (COG) elements », Proceedings of the 2003 ACM symposium on Document engineering, novembre 2003, Grenoble (F), pp 58-67.

[BCL] BCL, http://www.bcltechnologies.com/document/index.asp

[GLA] Glance, http://www.pdf-tools.com/en/home.asp

[HAD 01] Hadjar K., Hitz O., Ingold R., « Newspaper Page Decomposition using a Split and Merge Approach », ICDAR'01, septembre 2001, Seattle (USA), pp. 1186-1189.

[HAD 02] Hadjar K., Hitz O., Robadey L., Ingold R., « Configuration REecognition Model for Complex Reverse Engineering Methods: 2(CREM) », DAS'02, Août 2002, Princeton (USA), pp. 469-479.

[HAD 03] Hadjar K., Ingold R., « Arabic Newspaper Page Segmentation », ICDAR'03, Août 2003, Edinburgh (Scotland), pp. 895-899.

[HAD 04] Hadjar K., Rigamonti M., Lalanne D., Ingold R., « Xed : a new tool for eXtracting Electronic Documents », DIAL'04, 23-24 janvier 2004, Palo Alto (USA), pp. 212-224.

[HAR 94] Haralick R. M., « Document image understanding: Geometric and logical layout », Proc. Internet. Conf. On Computer Vision and Pattern Recognition, 1994, pp. 385-390.

[HU 01] Hu J., Kashi R., Lopresti D., Nagy G., Wilfong G., « Why table ground truthing is hard », ICDAR'01, Septembre 2001, Seattle (USA), pp. 129-133.

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

[LOV 95] Lovegrove W. S., Brailsford D. F., « Document analysis of PDF files: methods, results and implications », Electronic publishing, juin et septembre 1995, pp 207-220.

[MAT] MatterCast, http://www.mattercast.com/default.aspx

[NAG 00] Nagy G., « Twenty Years of Document Image Analysis in PAMI », IEEE Transactions on Pattern Analysis and Machine Intelligence, Janvier 2000, Vol. 22, No 1, pp. 38-62.

[RIG 03] Rigamonti M., Hitz O., Ingold R., « A Framework for Cooperative and Interactive Analisys of Technical Documents », GREC 03: Fifth IAPR International Workshop on Graphics Recognition, Juillet 2003, Barcelona (Spain), pp. 407-414.