Docmining : Une plate-forme de conception de systèmes d'analyse de documents

Eric Clavier, Sébastien Adam, Pierre Héroux, Maurizio Rigamonti

Docmining : Une plate-forme de conception de systèmes d'analyse de documents

Eric Clavier[1] — Sébastien Adam[2]— Pierre Héroux[2]— Maurizio Rigamonti[3]— Jean-Marc Ogier[4]

[1] France Telecom R&D
2 Avenue Pierre Marzin, 22307 Lannion Cedex France

[2] Laboratoire PSI – CNRS FRE 2645
Université de Rouen, Place Emile Blondel 76821 Mont Saint Aignan, France

[3] Laboratoire L3i
Université de La Rochelle

[4] DIUF, Université de Fribourg
Chemin du musée 3, 1700 Fribourg, Switzerland

Sebastien.Adam@univ-rouen.fr

Résumé *: Dans cet article, nous présentons une plate-forme logicielle, DocMining, qui permet la construction assistée de scénarios de traitement de document. La plate-forme présentée est issue d'un consortium regroupant quatre laboratoires de recherche et un partenaire industriel. Elle est dédiée à la conception et l'exécution de scénarios d'analyse de document. Elle permet un prototypage d'applications et l'intégration de bibliothèques de traitements hétérogènes. Nous montrons également comment exploiter les propriétés de cette plate-forme pour contribuer au problème de l'évaluation de performances en analyse de document.*

Mots-clés *: Scénarios d'analyse d'image de document, évaluation de performances, XML*

1 Introduction

La conception d'un système d'analyse d'image de document est une tâche fastidieuse qui requiert un assemblage rigoureux de composants ajustables très variés. Ces composants peuvent concerner des disciplines allant du traitement d'image à l'analyse sémantique, en passant par des outils de segmentation, d'extraction de caractéristiques, de classification, d'analyse structurelle ou syntaxique… De nombreuses équipes de recherche travaillent sur ces thématiques, avec des domaines d'application divers comme la reconnaissance de formulaires, de dessins techniques, de courriers manuscrits ou d'ouvrages anciens... Ces travaux ont donné lieu à une multitude de méthodes spécialisées, parfois concurrentes, et à plusieurs systèmes complets basés sur un enchaînement séquentiel de celles-ci [Dos 00][Boa 92]. Si ces systèmes s'avèrent généralement efficaces dans le contexte de l'application pour laquelle ils ont été conçus, force est de constater qu'ils ne le sont que dans ce cadre et sous réserve que les données analysées ne présentent pas de variabilités trop importantes. A notre connaissance, il n'existe actuellement pas de système totalement générique, capable de traiter aussi bien un schéma mécanique, un formulaire ou un courrier manuscrit sans modification importante du système.

Ce constat s'explique par deux difficultés majeures. La première est la nécessité d'expliciter et de formaliser une quantité considérable de connaissances pour que le système puisse s'adapter : les connaissances des spécialistes en analyse de document (sur le choix des traitements, leur paramétrage, leur enchaînement) et celles des spécialistes du domaine concerné par le document (concepts du domaine, relations, intentions…). La seconde traite de la nécessité d'intégrer un grand nombre de traitements différents dans le système, pour permettre une adaptation à des contextes variés.

Dans ce cadre, on se trouve confronté à deux options : la définition d'un système « sur-contraint » ou d'un système « sous-contraint ». Dans le premier cas, l'utilisateur se trouve face à un important ensemble de règles à respecter pour intégrer la connaissance, ce qui rend l'adaptation du système trop complexe. Dans le second cas, la solution peut se révéler trop « abstraite » et demander un travail trop important à l'utilisateur pour implémenter toute sa problématique sur la plate-forme.

Partant de ce constat, l'objectif du travail présenté dans cet article n'est pas de proposer un système générique d'analyse de document, mais plutôt un environnement logiciel, « DocMining », qui permet la construction assistée de scénarios de traitement de document. La plate-forme présentée est issue du consortium DocMining regroupant quatre laboratoires de recherche et un partenaire industriel. Elle est dédiée à la conception et l'exécution de scénarios d'analyse de

document. Elle permet ainsi le prototypage rapide d'applications et l'intégration de bibliothèques de traitements hétérogènes. Ce dernier point est en effet un problème récurrent lorsqu'il s'agit d'échanger des algorithmes entre groupes de recherches.

Ce choix peut paraître moins ambitieux que les approches visant la conception d'un système générique, mais nous pensons que notre approche pragmatique permet de faire un pas dans le sens de l'adaptabilité ou tout au moins du prototypage rapide de nouvelles applications. En outre, elle autorise, comme [Cou 01] ou [Pas 96], l'explicitation et donc l'archivage d'une partie des connaissances des experts, qui pourront être réutilisées pour d'autres cas d'usage, offrant ainsi un caractère de reproductibilité. Par ailleurs, DocMining ne se limite pas à la conception de systèmes d'interprétation mais permet d'aborder d'autres contextes applicatifs (benchmark, travaux collaboratifs, mesures de performances, ...).

L'article est organisé de la façon suivante. Nous présentons d'abord dans la section 2 les concepts qui ont conduit à la réalisation de cette plate-forme. Puis, nous décrivons son implémentation avant de proposer, dans la dernière section, deux cas d'usage dédiés à la problématique émergente de l'évaluation de performances.

2 Fondements de la plate-forme

Pour atteindre les objectifs énoncés dans l'introduction, la plate-forme propose à son utilisateur à la fois un ensemble de règles pour intégrer progressivement ses composants (modèles de documents, nouveaux types de données et nouveaux traitements) mais aussi des outils lui permettant de construire de façon interactive un enchaînement de ceux-ci en fonction de ses intentions et des tâches qu'il souhaite effectuer. DocMining (figure 1) s'articule autour de trois entités principales : le document, le traitement et le scénario :

- *le document* est l'élément central de l'architecture puisqu'il constitue un point d'accès commun à tous les traitements. Ainsi, toute opération réalisée par l'utilisateur peut faire l'objet d'un archivage dans le document, qu'il s'agisse d'un traitement, de son paramétrage, de ses entrées ou de ses sorties. La flexibilité et la modularité du document sont des garanties quant aux interactions rendues possibles entre les traitements disponibles et à la complexité des scénarios qu'il est possible de déclencher. Dans DocMining, le document ne contient pas uniquement les objets extraits par les traitements, il peut également contenir la trace des paramètres des traitements, leur durée d'exécution… Cette centralisation du point d'accès est une garantie pour faire face au traditionnel problème de l'éparpillement des données dans des structures distribuées.
- *le scénario* représente la tâche à appliquer au document ; son exécution coordonne l'accès des traitements constitutifs de la tâche au document par le biais d'étapes qui contiennent les critères de déclenchement des traitements ainsi que leurs paramètres. Le moteur de scénario chargé de son exécution gère l'enchaînement des traitements, la transmission des paramètres, ainsi que la mise à jour du document. Un scénario contient également d'autres composants, qui permettent de définir un contexte d'observation des traitements, d'adapter localement les données du document aux paramètres des traitements. Le scénario centralise donc la connaissance qu'a l'utilisateur sur une tâche de reconnaissance et évite l'éparpillement des données opérationnelles. Il est adapté à de nombreuses utilisations comme le partage des traitements entre groupes de recherches, l'intégration rapide de nouveaux traitements dans une chaîne ou la mise en place de campagnes d'évaluation de performances.
- *le traitement* représente une opération élémentaire paramétrable qui peut accéder à tout ou partie du document. En considérant le document comme un point d'accès commun, DocMining autorise l'interopérabilité entre traitements issus de bibliothèques hétérogènes. Si le traitement est capable de se « connecter » avec le document, il peut alors être enchaîné avec d'autres traitements. Pour interagir avec le document, le traitement doit respecter un contrat. Une partie du contrat est logicielle, l'autre est déclarative. La partie déclarative du contrat contient les types d'objets sur lesquels le traitement peut agir, le service rendu, les objets qu'il produit ainsi que ses paramètres d'application. En définissant de manière stricte comment un traitement peut interagir avec le document, l'approche permet d'ajouter facilement des traitements à la plate-forme. La modification de la partie déclarative du contrat (qui contient les éléments définissant l'accès au document) permet d'adapter le traitement à d'autres types de données que ceux prévus initialement, garantissant ainsi la réutilisabilité de traitements.

3 Mise en œuvre

La recherche d'un format de document flexible et modulable nous a naturellement orientés vers l'utilisation de XML. La plate-forme peut donc prendre en charge tout document XML. Néanmoins, XML n'est qu'un formalisme, nous avons donc mis en place un schéma XML qui définit une structure de document élémentaire ainsi qu'un ensemble de composants graphiques basiques. La structure du scénario respecte également le formalisme XML, elle correspond à un schéma dont les balises sont interprétées par le moteur de scénario (ImTrac – figure 1). Les balises correspondant aux étapes contiennent les critères de déclenchement d'un traitement exprimé au moyen d'une expression *XPath*. Cette dernière correspond à l'objet devant être présent dans le document pour que le traitement puisse être appliqué. La puissance des expressions *XPath* permet de définir des critères de déclenchement très précis qui peuvent même tenir compte de la valeur d'un attribut.

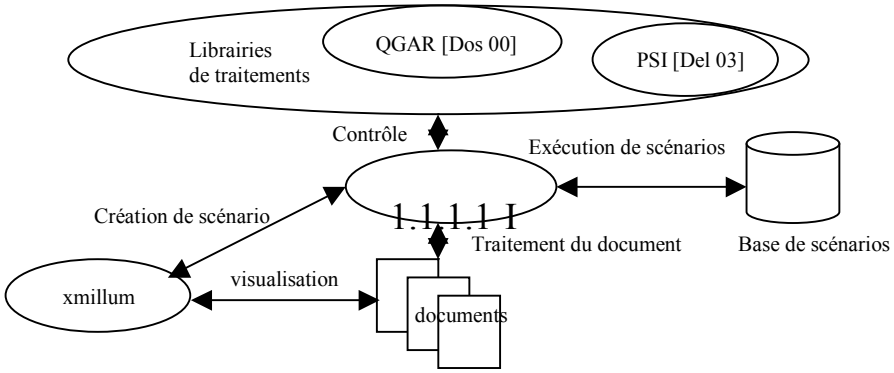

FIG. 1 - Architecture globale de la plate-forme

Outre les traitements, un scénario peut contenir d'autres composants (illustrés dans l'exemple de scénario proposé sur la figure 2) de granularité plus faible :

– *les instructions* (1), qui correspondent à du code interprété lors de l'exécution du scénario. Elles peuvent être placées avant et après un traitement. Elles permettent par exemple de définir des structures de contrôle (tests, boucles) sur les traitements en offrant la possibilité de définir des critères d'arrêt.

– *les adaptateurs de support* (2), qui permettent d'interpréter différemment la valeur d'un paramètre. Ce composant autorise différents modes d'extraction des paramètres et permet d'adapter un traitement à un jeu de données pour lequel il n'était pas prévu initialement et lui confèrent une réutilisabilité plus grande.

– *les gâchettes* (3), qui sont déclenchées automatiquement après l'exécution d'un traitement. Il s'agit de composants transverses au scénario qui autorisent la mise en place d'un contexte d'observation, la mise à jour d'un affichage…

– *les variables* (4), qui sont des composants atomiques pouvant être partagés par les autres composants. Elles définissent le contexte courant du scénario.

En termes d'outils d'analyse de document, les traitements disponibles sont issus des travaux des différentes équipes du consortium DocMining . Le lecteur intéressé en trouvera des présentations détaillées dans les références [Del 03][Dos 00][Hit 00]. Chacun des traitements possède un contrat logiciel correspondant à l'implémentation d'une interface java. La partie déclarative du contrat respecte un schéma XML et contient le nom de la classe, les paramètres du traitement, les objets pris en charge exprimés au moyen d'une expression XPath, ainsi que les objets générés exprimés par un brin XML. Ces deux derniers éléments sont utilisés lors de la construction automatique de scénario car ils définissent un comportement *a priori*.

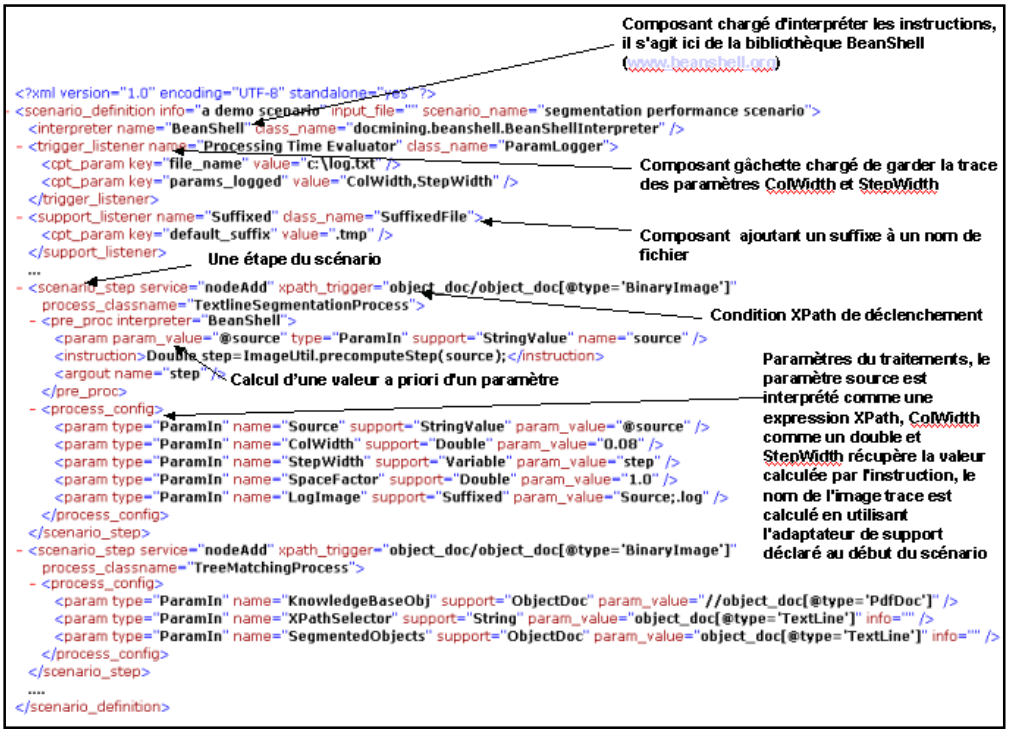

Fig. 2 - Un exemple de scénario illustrant les différents composants proposés

4 Deux cas d'usage pour l'évaluation de performances

Dans cette section, nous proposons deux cas d'usage de la plate-forme DocMining visant d'une part à démontrer l'intérêt de celle-ci dans le contexte de campagnes d'évaluation de performances.

4.1 La problématique de l'évaluation de performances

L'évaluation de performances est récemment devenue une problématique majeure dans le domaine de l'analyse d'image. La raison de l'émergence de cette problématique est relativement simple. Depuis maintenant près de 20 ans, toutes les équipes travaillant dans le domaine de l'analyse de document proposent régulièrement de nouvelles méthodes, à différents niveaux de la chaîne d'interprétation. Malgré cette richesse potentielle, force est de constater que nul n'est capable aujourd'hui de préciser quelle méthode est la plus efficace dans un contexte donné. A une granularité plus faible, cette impossibilité d'évaluer les performances d'outils sur des données représentatives rend complexe l'optimisation du paramétrage des méthodes. Pour pallier ces difficultés, trois aspects majeurs sont à prendre en compte : la constitution de bases de données accompagnées de la « vérité terrain », la proposition d'une métrique pour mesurer l'adéquation entre les résultats obtenus par les méthodes à évaluer et la vérité terrain et enfin la conception d'une plate-forme permettant une intégration rapide et flexible de nouvelles méthodes proposées par différentes équipes, avec les problèmes bien connus d'interopérabilité que pose une telle intégration. Les deux scénarios que nous proposons ci-après proposent des éléments de réponse dans ce contexte.

4.2 Un scénario dédié à l'évaluation de performances en segmentation

Ce premier scénario vise à proposer des éléments de solution au problème de l'évaluation de performances d'outils d'extraction de la structure de documents. On peut décomposer celui-ci en trois étapes majeures :

- *La première étape concerne la génération de données accompagnées de la « vérité terrain »*. Dans notre approche, ce problème est résolu puisque c'est le scénario lui même qui permet de générer à la fois l'image de document à traiter et la vérité terrain. Pour ce faire, deux solutions sont possibles : La première consiste à créer un fichier XML instancié depuis une DTD classique (DocBook, TEI, MathML…) que nous transformons en un fichier au format PDF, en utilisant des outils disponibles dans la communauté. Les feuilles de styles utilisées dans ce cadre peuvent alors être modifiées en vue de créer des fichiers PDF avec des formatages différents. Ces fichiers sont ensuite transformés en fichiers image afin de pouvoir y appliquer les algorithmes à tester. Par ailleurs, nous conservons le fichier XML d'origine qui constitue pour le système d'évaluation la « vérité terrain ». L'autre solution consiste à prendre comme source initiale un ensemble de documents PDF déjà existants et de les « segmenter » au moyen de librairies pouvant lire ce format. Il est alors possible d'extraire du fichier les lignes et caractères qui le compose. La création de leur équivalent XML constitue alors notre vérité terrain. La première approche permet de générer des corpus propres à tester un algorithme alors que la deuxième permet de réaliser des tests sur des corpus représentatifs de la réalité des documents actuels.
- *La seconde partie du scénario consiste à appliquer les algorithmes à tester sur l'image générée précédemment*. C'est ici que la flexibilité de la plate-forme intervient puisqu'il est très simple, en utilisant les concepts décrits à la section 2 d'intégrer de nouveaux outils par simple définition de leur contrat et éventuellement l'utilisation d'adaptateurs de support pour mettre en forme les données. Dans le cadre de nos tests, deux approches différentes ont été intégrées : la première, issue de [Rob 02] exploite une approche purement ascendante, alors que la seconde, issue de [Par 96] est une approche hybride.
- *La dernière partie du scénario concerne la comparaison des résultats obtenus par les méthodes avec la vérité terrain*. Là encore la plate-forme montre un intérêt important puisque, grâce à l'utilisation intensive d'XML et son association avec des expressions XPath, il est aisé d'extraire à la fois les résultats obtenus par les outils et la vérité terrain. Cela permet alors de constituer des graphes que l'on peut ensuite comparer ensuite en utilisant l'algorithme proposé par Yanikoglu [Yan 98].

L'application de ce scénario nous permet de comparer quantitativement mais aussi qualitativement les résultats obtenus par différents algorithmes de segmentation. La comparaison qualitative est facilitée par l'utilisation de l'interface xmillum [Hit 00] qui permet de naviguer dans les résultats obtenus (figure 3).

4.3 Un scénario dédié à l'évaluation de performances en reconnaissance

Ce second scénario concerne également la problématique de l'évaluation de performances, mais dans un contexte de reconnaissance de formes. Le corpus considéré est le « chicken dataset », fourni par le TC 5 de l'IAPR, qui est composé d'images de découpes de poulet répartis en 5 classes : ailes, blancs, cuisses, dos et pilons. Associés aux images noir et blanc (figure 4), un ensemble de caractéristiques correspondant au codage du contour des formes est fourni. Comme pour le premier scénario, nous ne cherchons pas ici à montrer la supériorité d'une approche de reconnaissance sur une autre, mais plutôt l'intérêt de la plate-forme pour comparer les résultats obtenus par différentes approches. Dans ce contexte, des traitements issus de plusieurs librairies ont été intégrés pour valider l'interopérabilité de la plate-forme. Citons par exemple un extracteur de caractéristiques issu de la PSIlib [Del 03] basé sur les invariants de Fourier-Mellin [Ada 00], un extracteur de caractéristiques issus de la librairie openCV basé sur les moments de Hu [Hu 62] et une API de classification fournie avec la plate-forme.

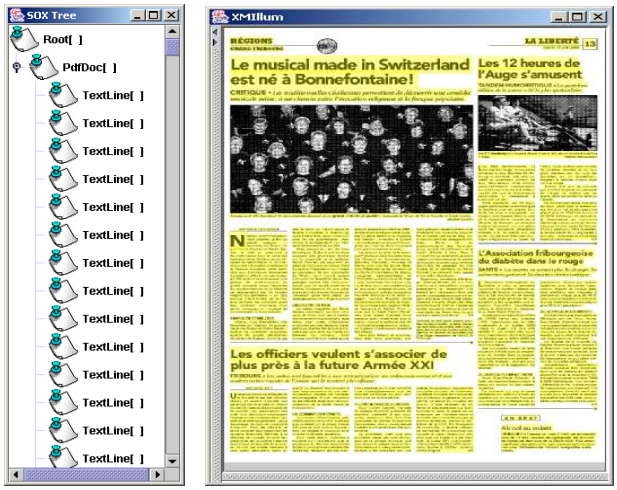

FIG. 3 - visualisation des résultats obtenus à l'aide de xmillum [Hit 00]

Le scénario construit peut être décomposé en trois étapes classiques :

- *Importation des données fournies par le TC5* : cette étape permet de construire la base de connaissance constituée par les références aux images de découpes. Chacune de ces références est complétée par une description de la découpe au moyen d'une chaîne représentant le contour de la forme. Cette étape est grandement facilitée par un composant permettant d'importer des images dans une structure et de lier chaque image à une étiquette. Les descripteurs fournis par le TC5 sont également importés lors de cette étape en utilisant des expressions XPath.
- *Extraction des caractéristiques* : cette étape permet d'ajouter des descripteurs aux formes de la base de connaissance. Par défaut, le moteur de scénario ne fait aucune hypothèse sur la structure du document ce qui permet d'ajouter à la structure existante n'importe quel fragment XML. Alors que la quantité et la complexité des données augmentent, leur manipulation n'est pas plus complexe puisque nous disposons d'un point d'accès unique, le document, et de méthodes de localisation dans cette structure.
- *Classification des découpes par chacun des jeux de descripteurs* : à l'issue de l'étape précédente, il est possible de mesurer le pouvoir discriminant de chacun des descripteurs. Le scénario consiste d'abord à générer aléatoirement des bases de test et d'apprentissage. Pour ce faire, aucune séparation physique de la base de connaissance n'est réalisée. On se contente de marquer les éléments qui la constituent. Un opérateur (K Plus Proches Voisins) permet alors de classifier les exemples de la base de test pour finalement mesurer le taux de reconnaissance. Même si les données ont des formats différents, aucune modification de code n'est nécessaire puisque ce sont les expressions XPath qui adaptent les données d'entrées aux formats d'entrée du traitement. Par ailleurs, l'approche choisie pour évaluer les performances peut facilement être modifiée pour appliquer par exemple un « Leave One Out » ou une « cross validation ». Comme dans le cas du premier scénario, xmillum permet de naviguer au sein des résultats obtenus.

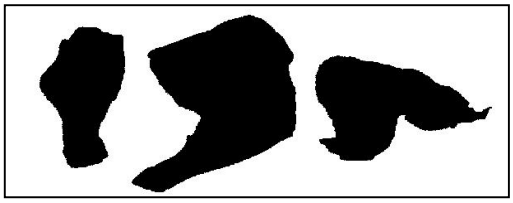

Fig. 4 - Exemples d'images du corpus « chicken dataset » de l'IAPR

5 Conclusion

Docmining est une plate-forme dédiée à la construction de nouveaux systèmes ayant trait à l'analyse de document. Cette plate-forme est opérationnelle et permet actuellement aux différents laboratoires membres du consortium de construire rapidement de nouveaux systèmes en exploitant des outils proposés par les différentes équipes. Deux cas d'usage ont été proposés dans cet article, ils avaient pour but de construire des systèmes d'évaluation de performances de scénarios d'analyse de document. Bien entendu, les possibilités offertes par Docmining dépassent largement ce cadre, comme nous l'avons déjà montré dans [CLA 03]. Des travaux sont donc encore en cours sur cette plate-forme, et ce à plusieurs niveaux. Parmi ceux-ci, on peut citer des travaux concernant l'optimisation de scénarios d'analyse en utilisant des algorithmes évolutionnaires, la poursuite des travaux concernant l'évaluation de performances et l'intégration d'une couche supplémentaire à la plate-forme pour permettre une programmation graphique, à la Khoros, des scénarios d'analyse.

6 Remerciements

Comme il est évoqué en introduction, le travail présenté dans cet article est issu d'un consortium regroupant quatre laboratoires de recherche et un partenaire industriel: Les laboratoires concernés sont le laboratoire

PSI de l'université de Rouen, le laboratoire L3I de l'université de la Rochelle, l'équipe QGAR du LORIA de Nancy et le laboratoire DIUF de l'université de Fribourg. Le partenaire industriel est un l'opérateur téléphonique France Telecom. Les auteurs tiennent donc à remercier les membres de ces équipes qui ont contribué à ce travail et à la relecture de cet article.

7 Bibliographie

[ADA 00] ADAM S., OGIER J.M., CARIOU C., MULLOT R., GARDES J., LECOURTIER Y., « "Utilisation de la Transformée de Fourier-Mellin pour la reconnaissance de formes multi-orientées et multi-échelles : application à l'analyse automatique de documents techniques », *Revue Traitement du Signal*, vol. 18, n°1, 2001, p. 17-33.

[BOA 92] Boatto L., and al. « An interpretation system for land register maps », IEEE Computer Magazine, vol. 25, n°7, 1992, p 25-33.

[CLA 03] CLAVIER E., MASINI G., DELALANDRE M., RIGAMONTI M., TOMBRE K., GARDES J., « Docmining : a cooperative platform for heterogeneous document interpretation according t user-defined scenarios », in the proceedings of Graphic Recognition Worskshop, Barcelone, Espagne, 2003, p. 21-32.

[COU 01] COÜASNON B., « DMOS : a generic document recognition method, application to an automatic generator of musical scrores, mathematical formulae and table structure recognition systems, *Proceedings of the 6th International Conference on Document Analysis and Recognition*, Seattle, USA, 2001, p. 215-220.

[DEL 03] DELALANDRE M., NICOLAS S., TRUPIN E. ET OGIER J.M., « Symbols Recognition by Global-Local Structural Approaches, Based on the Scenarios Use, and with a XML Representation of Data », *Proceedings of the 7th International Conference on Document Analysis and Recognition,* Edimbourg, Ecosse, 2003, p. 1002-1006.

[DOS 99] DOSCH P., AH-SOON C., MASINI G., SÁNCHEZ G., TOMBRE K. « Design of an Integrated Environment for the Automated Analysis of Architectural Drawings ». *In S.-W. Lee and Y. Nakano, editors,* Document Analysis Systems: Theory and Practice. Selected papers from Third IAPR Workshop, DAS'98, Nagano, Japan, November 4-6, 1998, *in revised version*, volume 1655 of *Lecture Notes in Computer Science*, 1999, p. 295-309.

[HIT 00], HITZ O., ROBADEY L., INGOLD R., « An architecture for editing document recognition results using XML », *In proceedings of the 4th IAPR workshop on Document Analysis System,* Rio de Janeiro, Bresil, 2000, p. 385-396.

[HU 62] HU M.K., « Visual pattern recognition by moment invariants », *IRE transaction on Information Theory*, vol. 8, 1962, p. 179-187.

[ROB 01] ROBADEY L., 2(CREM) : une méthode de reconnaissance structurelle de documents complexes basée sur des pattern bidimensionnels, PhD thesis, Departement d'informatique de l'université de Fribourg, 2001.

[PAR 96] PARODI P. et PICCIOLI G., « An efficient pre-processing of mixed-content document images for OCR systems », *Proceedings of the 13th International Conference on Pattern Recognition*, Vienne, Autriche, 1996, p. 778 -782.

[PAS 96] PASTERNAK B., Adaptierbares Kernsystem zur Interpretation von Zeichnungen, Dissertation zur Erlangung des akademischen Grades eines Doktors des Naturwissenschaften, Universität Hamburg, 1996.

[YAN 98] YANIKOGLU B. A. et VINCENT L., « Pink panther: a complete environment for ground-truthing and benchmarking document page segmentation », *Pattern Recognition* vol. 31, September 1998, p. 1191-1204.