Production de vérite terrain pour l'analyse et l'interprétation d'images de document

Pierre Héroux, Eugen Barbu, Sébastien Adam, Eric Trupin

Production de vérité terrain pour l'analyse et l'interprétation d'images de document

Pierre Héroux – Eugen Barbu – Sébastien Adam – Éric Trupin

LITIS - Université de Rouen
UFR des Sciences et Techniques
Avenue de l'Université
76800 Saint-Etienne du Rouvray

Pierre.Heroux@univ-rouen.fr

Résumé : *L'évaluation de performances en analyse et interprétation d'images de document est un problème récurrent. Si des bases d'images de document contenant des informations de vérité terrain existent, elles se révèlent souvent inadaptées pour l'évaluation d'un traitement ou d'un système complet dans un contexte particulier. Dans cet article, nous proposons une approche pour la génération automatique de vérité terrain par dérivation d'outils d'édition. Une implémentation de cette approche illustre la richesse des informations pouvant être produites.*

1 Introduction

La plupart des systèmes opérationnels en analyse et interprétation de documents sont utilisés dans un contexte donné et cette spécificité est utilisée comme connaissance *a priori*. Ces systèmes ne sont alors pas adaptés pour d'autres contextes. Ainsi, quelques questions doivent alors se poser lors de la conception d'un système d'analyse et d'interprétation d'images de documents dans un nouveau cas d'usage :

- Quelles sont les tâches devant être utilisées par la chaîne de traitement ?
- Comment les agencer au sein de cette chaîne de traitement ?
- Quel est l'algorithme le plus adapté pour accomplir une tâche donnée dans ce contexte ?
- Quels sont les paramètres les plus adaptés pour ces algorithmes ?

Pour répondre à ces questions, il est nécessaire de définir des procédures d'évaluation. Ces procédures d'évaluation peuvent intégrer une expertise humaine, mais des évaluations automatiques comparant les résultats des traitements automatiques avec des informations de vérité terrain sont souvent privilégiés car elles permettent d'agir sur un volume de données plus représentatif et reposant sur des métriques bien définies. Disposant de ces informations de vérité terrain sur des volumes de données importantes et de critères d'évaluation, la conception d'un système d'analyse peut être ramenées à un problème d'optimisation.

Dans cet article, nous proposons une approche pour la production automatique de vérité terrain utilisée pour l'évaluation de performances dans le contexte de la conception d'un système d'analyse et d'interprétation d'images de documents. La section 2 présente les motivations de ce travail et les argumentations des options prises dans notre proposition. La section 4 illustre par un exemple, la mise en œuvre de l'approche. Cette exemple illustre les possibilités offertes par la production automatique en terme de diversité et de richesse des informations produites.

2 Motivations

L'évaluation de performances en analyse d'images est un problème récurrent ayant, en particulier, motivé l'appel à projet Techno-Vision. La communauté scientifique de l'analyse d'images de documents est sensible à cette problématique, mais le nombre de bases mises à sa disposition pour l'évaluation est lui-même le signe qu'il n'y a pas de consensus ni sur la représentativité de ces bases, ni sur les informations de vérité terrain associée.

Parmi les bases faisant référence, citons celles proposées par l'Université de Washington [PHI 93a, PHI 93b], l'équipe MediaTeam [SAU 98], la Bibliothèque Nationale Américaine de Médecine [MED] et plus récemment PRImA [ANT 06].

Dans leur article [ANT 06], les auteurs plaident pour la constitution d'une base de documents visant deux objectifs majeurs dans l'optique d'une évaluation exhaustive des méthodes d'analyse. D'une part, les catégories de documents doivent être représentatives de celles rencontrées dans la vie courante. D'autre part, la représentation des documents entre catégories doit refléter un usage réaliste, tout conservant des documents suffisamment nombreux et variés au sein d'une même catégorie. Ces objectifs semblent difficiles à atteindre. Il est en effet difficile de borner la couverture de la base, visant une évaluation exhaustive. Par ailleurs, la définition d'une vérité terrain pour une telle base serait d'un coût prohibitif. Notons, par exemple, que la composition des trois CD-ROM délivrés par l'Université de Washington a coûté environ 2 millions de dollars, que des omissions y persistent et surtout qu'ils ne contiennent que des documents anglais et japonais.

De façon pragmatique, nous nous plaçons, non pas dans un contexte d'évaluation exhaustive des méthodes d'analyse, mais plutôt dans l'optique de la conception d'un système d'analyse efficace pour un cas d'usage bien défini. Rien ne

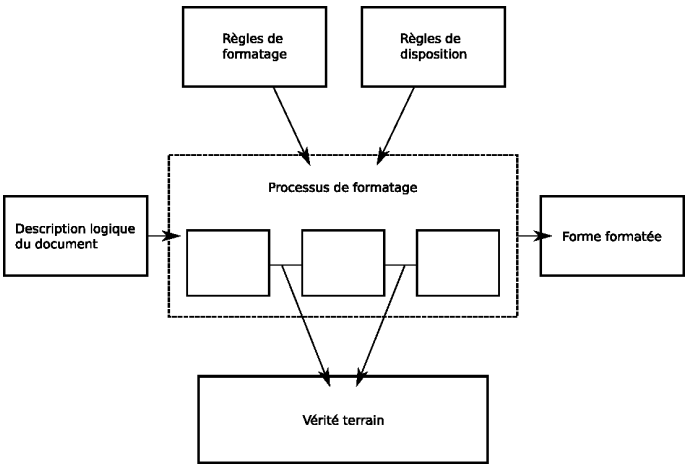


FIG. 1 – Vue générale de l'approche

garantit que les bases existantes contiennent des documents identiques à ceux du cas d'usage. Par ailleurs, les formats de vérités terrains associé à ces bases peuvent être incomplets ou inadaptés pour l'évaluation de certaines tâches.

Dans le contexte de la conception d'un système d'analyse, il est alors nécessaire non seulement de définir le format de vérité terrain apte à permettre l'évaluation, mais également de se constituer un corpus d'images de document représentatif du cas d'usage.

Une première approche à envisager consiste à définir la vérité terrain conforme au format approprié pour un corpus extrait du cas d'usage et pour un volume représentatif. Souhaitant, nous abstraire du coût d'une telle démarche, mais surtout de la subjectivité et des imperfections (omissions, erreurs…) toujours possibles lors de l'intervention d'un opérateur humain, nous proposons une approche de production automatique de vérité terrain non pas sur des images documents réelles mais sur des images créées de façon synthétique.

3 Description générale de l'approche

L'approche que nous proposons s'appuie sur une dérivation des logiciels de création de documents. Ces logiciels traitent une représentation interne des documents et produisent à partir de celle-ci une forme formatée pouvant être rendue à l'écran et/ou imprimée sur support papier. Ce processus consiste à appliquer un ensemble de règles de formatage (police, taille, alignement…) aux différentes entités logiques qui, confrontées aux contraintes de mise en page (dimension de la page, des marges, des colonnes…), permettent d'aboutir à la forme formatée. Le processus de formatage peut être vus comme un chaîne dont les traitements produisent des informations exploitées en interne par les traitements suivants.

Notre approche propose d'extraire parmi ces informations, celles pouvant être utiles pour constituer la vérité terrain associées à une image de document (cf. Fig. 1).

Dans la section suivante, nous proposons une mise en œuvre de cette approche en justifiant les choix technologies opérés.

4 Exemple de mise en œuvre

4.1 La chaîne d'édition des documents XML DocBook

Dans cet exemple de mise en œuvre, nous avons choisi de traiter des documents XML dont le contenu est balisé logiquement. Il est alors aisé d'accéder à la structure logique des documents.

Le processus de formatage est réalisé par application d'une suite de transformations encodées dans des feuilles de style XSLT. Un cas d'usage courant est la transformation des documents XML vers un équivalent (X)HTML. Dans notre cas, la forme formatée devra pouvoir être rendue sur un support papier. Les feuilles de style que nous utilisons sont donc spécialisées pour la transformation du document initial au format XSL-FO. Tout comme XML [BRA 98] et XSL-T [CLA 00], XSL-FO [ADL 01] est une recommandation du consortium W3C. De ce fait, il existe un nombre important d'outils disponibles permettant de traiter ces formats ouverts.

XSL-FO est un format qui vise précisément à exprimer la forme formatée d'un document sur un support papier. Les documents XSL-FO peuvent ensuite être convertis vers différents formats usuels (RTF, SVG, PostScript, PDF) donnant le même aspect visuel en utilisant des outils tels que FOP [APA], PassiveTeX [RAH] ou XEP [REN].

Ce schéma de publication est applicable pour tous les documents XML pour peu que des feuilles de styles XSL-T les transformant au format XSL-FO soient définies.

Nous avons choisi pour illustrer notre approche une implémentation traitant des documents XML se conformant à la DTD DocBook en raison de son caractère généraliste. DocBook [WAL 99] est un standard approuvé par OASIS (Organization for the Advancement of Structured Information Standards). Ce format est de plus en plus utilisé dans le monde de l'édition. L'éditeur O'Reilly en a fait son format pivot. Et de plus en plus d'outils, tels que le traitement de texte de la suite OpenOffice prennent en charge ce format. DocBook est également le format sous lequel sont rédigées les documentations des logiciels au sein du LDP (Linux Documentation Project). Ces nombreux documents étant publiées sous FDL (Free Documentation Licence) peuvent exploités pour créer des bases de documents à moindre coût.

Un document XML DocBook peut représenter une section, un article, un chapitre, une partie, un livre ou un ensemble de livres. La DTD DocBook permet un balisage logique du contenu textuel grâce à plus de 300 balises différentes parmi lesquelles on retrouve des éléments exprimant la structure logique (paragraphe, hiérarchie de sections et sous-sections, article, chapitre, partie…), mais on trouve également des éléments permettant de déclarer des tableaux, des figures, des bibliographies, des liens internes ou externes… Enfin, les documents XML DocBook peuvent inclure des fragments se référant à d'autres DTD comme SVG pour les graphiques vectoriels ou encore MathML pour l'expression de formules mathématiques.

Parmi les nombreux outils s'interfaçant avec le format XML DocBook, on note en particulier l'existence de feuilles de styles XSLT développées par Bob Stayton [STA 05] dédiées à cette DTD. Ces feuilles de style permettent la trans-

formation des documents XML DocBook vers plusieurs formats dont, en particulier, XSL-FO. Ces feuilles de style sont conçues pour être aisément paramétrables de telle sorte qu'il est possible d'implanter aisément ses propres règles de formatage, via un dispositif de redéfinition des règles utilisées par défaut. Plusieurs documents XML DocBook peuvent alors être transformés en utilisant les mêmes règles de formatage. D'un autre côté, plusieurs paramétrages des feuilles de style peuvent être appliqués au même document pour obtenir autant de formes formatées du même document initial, tel que le montre la figure 2.

4.2 Génération de la vérité terrain

Pour notre application de génération automatique de vérité terrain nous dérivons le schéma de publication DocBook. Mais plutôt que de produire des documents format RTF, PostScript ou PDF, nous générons une image numérique pour chaque page du document, information plus proche de ce que les systèmes d'analyse et d'interprétation d'images de documents sont amenés à traiter. Notre implémentation actuelle autorise la génération d'images aux formats TIFF, PNG, JPEG, à diverses résolutions et à des taux de compression paramétrables.

Notre système est apte à exporter les informations internes utilisées à différents stades du processus de formatage, ces informations pouvant être utiles pour une tâche d'évaluation. Parmi la grande variété d'informations disponibles, on trouve :

La structure logique : elle est directement accessible puisque les documents XML DocBook initiaux sont balisés par une hiérarchie d'éléments logiques.

Les règles de formatage appliquées : Les feuilles de style utilisées pour transformer le document dans sa forme formatée incluent directement les règles de formatage appliquée aux éléments logiques. Il est alors possible de connaître quels sont les polices de caractères, attributs de polices (taille, style, couleur) et attributs de positionnement (gauche, droite, centré, justifié) utilisés pour chaque type d'élément logique.

structure physique : La structure physique est une hiérarchie d'objets physiques (page, colonne, bloc, ligne, zone de mot, zone de caractère). À chacun des éléments physiques est associé un ensemble d'information résultant du processus de formatage :

- position et dimension
- type de contenu : texte, graphique, tableau, équation, image…
- l'étiquette logique associée : titre, auteur, paragraphe, légende, entrée bibliographique…
- pour les éléments textuels, des attributs de police : nom de la police de caractères utilisée, taille, style (gras, italique, barré), couleur…
- attributs de disposition : alignement, indentation, espace vertical précédent et suivant…
- le contenu textuel
- une référence vers l'élément correspondant dans la structure logique.

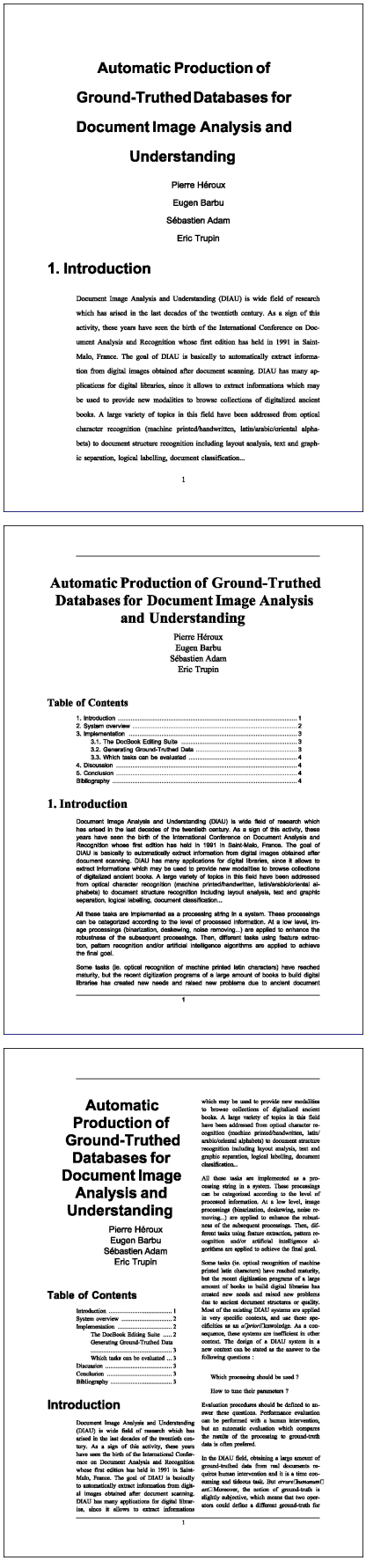
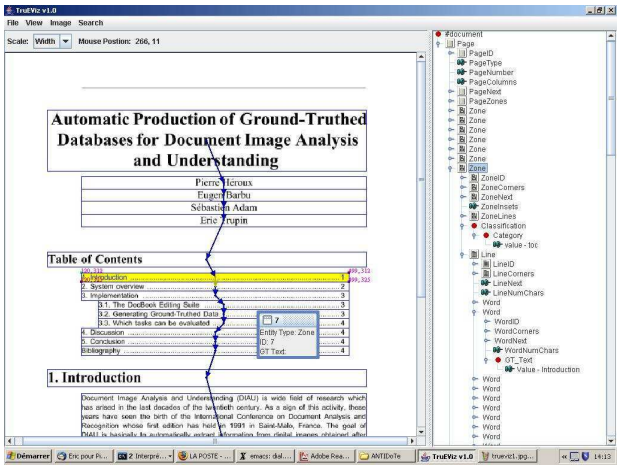


FIG. 2 – Une même page formatée de trois manières différentes

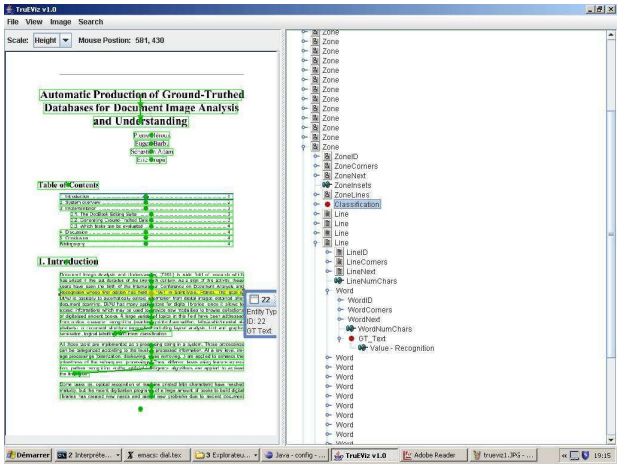
Selon le scénario défini pour l'évaluation, le format de la vérité terrain peut contenir l'ensemble ou une partie uniquement de ces informations. Ces informations peuvent être encodées selon différents formats. Nous aurions pu présenter des sorties au format RDIFF, DAFS [FRU 95], ces formats étant pris en charge par les outils Pink Panther [YAN 98] et PSET [MAO 02] d'évaluation de segmentation de pages, mais une sortie au format TrueViz nous semble plus apte à illustrer la variété des informations pouvant être produites. Trueviz est un format XML ayant initialement été conçu pour encoder des informations de vérité terrain pour l'évaluation d'outils de reconnaissance optique de caractères, mais cette DTD étant extensible, elle est également adaptée pour l'évaluation d'autres tâches d'analyse et d'interprétation d'images de documents. Par ailleurs, une interface graphique TRUEVIZ [KAN 01, MAO 01] permettant la visualisation et l'édition de vérité terrain est mise à disposition. Cette interface est multi-plateforme et elle est livrée comme un projet dont les codes sources sont libres. Elle peut être étendue pour s'adapter à des besoins plus spécifiques comme pour le projet d'évaluation de segmentation de pages d'article médicaux de la US National Library of Medecine.

Les figures 3(a),3(b) et 3(c) montrent l'interface TRUEVIZ. Sur la partie gauche, on visualise l'image du document sur laquelle se superposent les régions d'intérêts constituant la vérité terrain pour la segmentation de pages à différents niveaux (zone, ligne, mot). La partie droite de l'interface montre la vérité terrain via une une structure arborescente dont les branches peuvent être déployées ou contractées. Ces figures montrent plusieurs niveaux de la vérité terrain ayant été automatiquement produite. On remarque sur la vue image que la notion d'ordre de lecture est inclue dans la vérité terrain de même que les coordonnées des différentes zones.

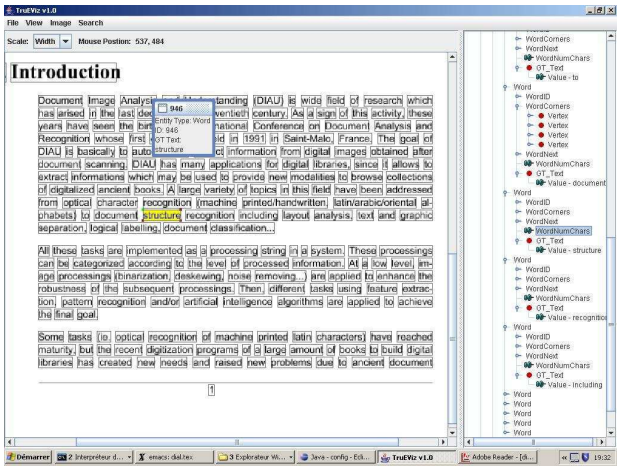
La représentation arborescente est plus détaillée de telle sorte qu'on y remarque la notion d'inclusion des mots dans les lignes, et des lignes dans les zones. On remarque également que la zone mise en avant correspond à un nœud de l'arbre étiqueté comme un élément de table des matières (toc) contenant une ligne dont le second mot est "introduction". Nous aurions de la même manière pu illustrer, qu'il est également possible d'accéder aux attributs de police de caractères des mots.



(a) Vue des zones



(b) Vue des lignes



(c) Vue des mots

FIG. 3 – Visualisation sous TRUEVIZ de la vérité terrain générée automatiquement

5 Discussion

Notre proposition énonce une procédure permettant de produire automatiquement des données synthétiques et les données de vérité terrain permettant la comparaison de traitements ou la détermination optimale de leurs paramètres dans un contexte particulier.

Cette proposition est étayée par un exemple de mise en œuvre présentée dans la section précédente. Cette illustration montre la variété des informations qu'il est possible de produire, permettant au moins l'évaluation des tâches suivantes :

- Reconnaissance optique de caractères
- Reconnaissance optique de fontes
- Segmentation de pages
- Discrimination de texte/graphique/tableau/image/équation
- L'analyse de la structure physique
- L'étiquetage physique

– L'interprétation de la structure logique
– La classification d'images de documents

Cependant, le format de vérité terrain peut être spécialisé pour des scénarii d'évaluation plus spécifiques tels que, par exemple, la localisation des zones correspondant à un élément logique particulier (figure, titre, auteur...) ou la propriété des fontes utilisée sa mise en forme.

Le choix d'opter pour la génération d'images de documents synthétiques engendre quelques limites. En premier lieu, il ne peut s'agir que de documents dont le contenu textuel est exclusivement imprimé. Cela exclus le traitement de documents incluant d'écriture manuscrite. Par ailleurs, une limite propre à XSL-FO restreint notre implémentation à ne produire que des blocs de textes rectangulaires isothétiques. Cependant, des propositions [SIL 05] sont actuellement faites pour étendre les fonctionnalités de cette norme.

La génération synthétique peut également se voir opposer des arguments sur la représentativité des images produites par rapport au cas d'usage réel, d'une part, du point de vue du type de document (type de contenu, mise en forme) et, d'autre part, du point de vue de la qualité des images. Ce second problème a fait l'objet de nombreux travaux relatifs à la modélisation des déformations subies par les images de documents par dégradation du support papier ou par les conditions de numérisation dont les plus significatifs sont [BAI 00, KAN 00]. D'autres approches donnant des résultats plus proches de documents réels peuvent être appliqués. Zi et Doermann [ZI 04] proposent en particulier de superposer à l'image synthétique le résultat de la numérisation sous diverses conditions de pages blanches. Il est toujours possible de procéder à l'impression des images et à la numériser.

Pour répondre au problème de la représentativité des documents, nous proposons que la définition des feuilles de style permettant le formatage des documents électroniques soit validée par une confrontation des images produites avec les images réelles, en utilisant éventuellement la transcription de quelques documents issus du corpus réel selon le schéma de la figure

6 Conclusion

Nous avons présenté dans cet article, une approche de génération automatique de vérité terrain pour l'évaluation de performances en analyse et interprétation d'images de document. Cette approche propose une alternative à l'évaluation de performances utilisant des bases dans lesquelles les documents ne sont pas représentatifs du cas d'usage ou dont le format de vérité terrain n'est pas adapté.

Une mise en œuvre de cette approche exploitant des documents XML DocBook illustre la souplesse en terme de format de vérité terrain. En effet, dans l'implémentation actuelle la vérité terrain peut être exprimée selon DAFS, RDIFF et TrueViz, mais d'autres formats éventuellement spécifiques peuvent aisément être implantés. La richesse des informations produites permet l'évaluation de performances pour un large panel de traitements d'analyse ou d'interprétation d'images de document. En effet, si beaucoup de bases existantes donnent des vérités terrains permettant l'évaluation d'OCR ou de segmentation de pages, nous offrons des possibilités supérieures telles que, par exemple, des informations pour la reconnaissance de fontes, la localisation de table des matières, ... jusqu'à la reconstruction de la structure logique du document.

Les informations produites de façon automatique présente également l'avantage de s'exempter des erreurs de saisie et de la subjectivité induites par l'intervention d'opérateurs humains.

L'utilisation de technologies XML rend aisée l'extension de la plateforme à la génération d'autres types de documents. En effet, avec un faible investissement, nous avons pu générer des images de documents et la vérité terrain associée à partir de documents XML TEI, produisant des bases de documents plus proches de ceux traités actuellement dans les projets de numérisation massive de fonds de bibliothécaires.

Nous envisageons d'adjoindre à la plateforme proposée des algorithmes publiés dans la littérature de dégradation synthétique des données produites, ainsi que des méthodes d'évaluation permettant la comparaison des résultats produits par des tâches d'analyse avec la vérité terrain, et ce, à différents niveaux. À terme, nous envisageons donner un libre accès à cette plateforme à la communauté scientifique, souhaitant qu'elle formule des suggestions quant à son évolution et à la définition de formats de vérité terrain enfin partagés, au moins pour certaines tâches.

Références

[ADL 01] ADLER S., BERGLUND A., CARUSO J., DEACH S., GRAHAM T., GROSSO P., GUTENTAG E., MILOWSKI A., PARNELL S., RICHMA J., , ZILLES S., *Extensible Stylesheet Language (XSL)*, W3C, `http ://www.w3.org/TR/xsl/`, 2001.

[ANT 06] ANTONACOPOULOS A., KARATZAS D., , BRIDSON D., Ground-Truth for Layout Analysis Performance Evaluation, BUNKE H., SPITZ A. L., Eds., *Document Analysis Systems VII, seventh International Workshop DAS 2006*, 2006, pp. 303-311.

[APA] APACHE, FOP, `http ://xmlgraphics.apache.org/fop/`.

[BAI 00] BAIRD H. S., State of the Art of Document Image Degradation Modeling, *Proceedings of the IAPR International Workshop on Document Analysis Systems*, 2000, pp. 1-16.

[BRA 98] BRAY T., PAOLI J., SPERBERG-MCQUEEN C. M., MALER E., , YERGEAU F., Extensible Markup Language (XML) 1.0, 1998, `http ://www.w3.org/TR/REC-xml/`.

[CLA 00] CLARK J., XSL Transformations (XSLT), 2000, `http ://www.w3.org/TR/xslt`.

[FRU 95] FRUCHTERMAN T., DAFS : A Standard for Document Image Understanding, *Proceedings of the Symposium on Document Image Understanding Technology*, 1995, pp. 94-100.

[KAN 00] KANUNGO T., HARALICK R. M., BAIRD H. S., STUEZLE W., , MADIGAN D., Statistical, Nonparametric Methodology for Document Degradation Model Validation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, n° 11, 2000, pp. 1209-1223.

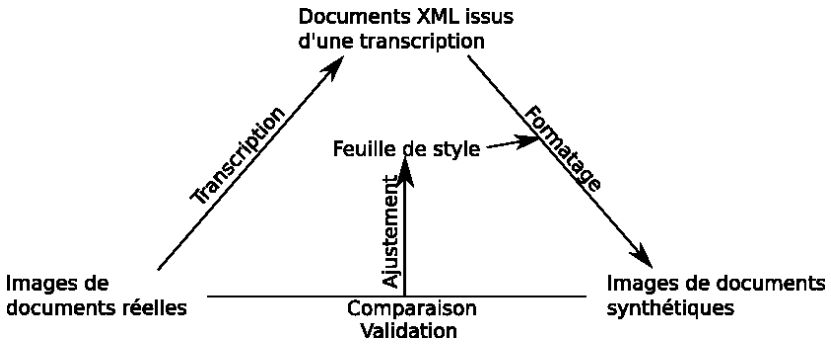


FIG. 4 – Cycle de définition et de validation des feuilles de styles

[KAN 01] KANUNGO T., LEE C. H., CZORAPINSKI J., , BELLA I., TRUEVIZ : a groundtruth/metadata editing and vizualization toolkit for OCR, *Proceedings of the SPIE Conference on Document Recognition and Retrieval*, 2001, pp. 1-12.

[MAO 01] MAO S., KANUNGO T., Empirical Performance Evaluation and its Application to Page Segmentation Algorithms, *IEEE Transactions on Pattern Analysis and Machine Interlligence*, vol. 23, n° 3, 2001, pp. 242-256.

[MAO 02] MAO S., KANUNGO T., Software Architecture of PSET : a page segmentation avaluation toolkit, *International Journal on Document Analysis and Recognition*, vol. 4, n° 3, 2002, pp. 205-217.

[MED] OF MEDECINE U. S. N. L., Medical Archive Records Groundtruth, `http ://marg.nlm.nih.gov/index.swf`.

[PHI 93a] PHILIPS I. T., CHEN S., HA J., , HARALICK R. M., English Document Database Design and Implementation Methology, *Proceedings of second annual Symposium on Document Analysis and Retrieval*, 1993, pp. 65-104.

[PHI 93b] PHILIPS I. T., HARALICK R. M., CD-ROM Document Database Standard, *Proceedings of the second International Conference on Document Analysis and Recognition*, 1993, pp. 478-483.

[RAH] RAHTZ S., PassiveTeX, `www.tei-c.org.uk/Software/passivetex/`.

[REN] RENDERX, XEP Engine, `http ://www.renderx.com/tools/`.

[SAU 98] SAUVOLA J., KAUNISKANGAS H., *MediaTeam Oulu Document Database*, MediaTeam, Oulu University, Finland, `http ://www.mediateam.oulu.fi/MTDB/`, 1998.

[SIL 05] DA SILVA A. C. B., DE OLIVEIRA J. B. S., MANO F. T. M., SILVA T. B., MEIRELLES L. L., MENEGUZZI F. R., , GIANETTI F., Support for arbitrary regions in XSL-FO, *Proceedings of the 2005 ACM Symposium on Document Engineering*, 2005, pp. 64-73.

[STA 05] STAYTON B., *DocBook XSL : The Complete Guide*, Sagehill Enterprises, 2005.

[WAL 99] WALSH N., MUELLNER L., *DocBook : The Definitive Guide*, O'Reilly, 1999.

[YAN 98] YANIKOGLU B. A., VINCENT L., Pink Panther : A complete environment for ground-truthing and benchmarking document page segmentation., *Pattern Recognition*, vol. 31, 1998, pp. 1191-1204.

[ZI 04] ZI G., DOERMANN D., Document Image Ground Truth Generation from Electronic Text, *Proceedings of the 17th International Conference on Pattern Recognition*, vol. 2, 2004, pp. 663-666.