

Une méthodologie pour la rétroconversion des documents papier vers un format electronique

Pierre Héroux, Eric Trupin, Yves Lecourtier

▶ **To cite this version:**

Pierre Héroux, Eric Trupin, Yves Lecourtier. Une méthodologie pour la rétroconversion des documents papier vers un format electronique. Conférence International sur le Document Electronique, 1999, Damas, Syrie. pp.251-262. hal-01249442

HAL Id: hal-01249442
https://hal.science/hal-01249442

Submitted on 1 Jan 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Une méthodologie pour la rétroconversion des documents papier vers un format électronique

Pierre Héroux, Éric Trupin et Yves Lecourtier

Laboratoire Perception, Systèmes et Information
UFR des Sciences et Techniques
Université de Rouen
F-76821 Mont-Saint-Aignan Cedex

Pierre.Heroux@univ-rouen.fr

C et article présente une méthodologie pour la rétroconversion des documents papier vers un format électronique. Cette méthodologie se caractérise par une modélisation objet des documents couplée à une stratégie d'interprétation. La représentation proposée tente de répondre aux besoins d'un grand nombre d'applications dans le domaine du traitement des documents. En particulier, elle paraît adaptée à l'analyse et à l'interprétation des documents. Nous présentons également une stratégie d'interprétation des documents, associée à cette modélisation. Cette stratégie est basée sur l'extraction et l'identification des éléments du document, l'étude de la cohérence de la description globale du document et la remise en cause des traitements conduisant à des incohérences. Enfin, nous présentons des perspectives en terme d'implantation au sein d'un système d'interprétation de document.

T his paper presents a method to convert paper documents in an electronic format. This method consists in the definition of a convenient electronic format associated with a strategy for document understanding. The proposed representation of the document describes different abstraction levels. It is adapted to respond to many problems in the document processing domain. A strategy for document understanding is developed. It uses the proposed document representation and is based on extraction and identification of document entities, on the study of the consistency of the description and on the treatment of inconsistencies. Finally, future prospects are presented to implant this strategy in a document understanding system.

1. Introduction

Si le domaine de la reconnaissance et de l'interprétation des documents s'est considérablement développé au cours des dernières années, il n'en reste pas moins que beaucoup de progrès restent à faire en terme d'homogénéisation des formats de représentation des documents pour en faciliter l'échange et la réutilisation. En effet, beaucoup d'applications ayant été réalisées afin de répondre à des besoins ponctuels traitent des documents d'un type déterminé et utilisent les caractéristiques communes à tous les documents de ce type afin d'orienter, d'accélérer et de fiabiliser le traitement. Ces caractéristiques, utilisées comme connaissances *a priori*, ne sont alors pas exprimées explicitement dans la représentation des documents utilisée par ces applications. Ceci rend la représentation des documents traités inutilisable par d'autres applications. Des formats standards tels que PostScript PDF, RTF ou LaTeX, voire même des normes comme ODA, SGML ou encore XML, ont bien été définis en matière de représentation des documents, mais aucun de ces formats ne décrivant l'ensemble des vues du document, ne semble répondre à tous les besoins (réédition, conversion, diffusion, indexation, archivage). On ne trouve, en particulier, aucun format capable de décrire une large gamme de documents comprenant des élements textuels, des tableaux, des graphiques ou encore des images, et qui soit, de plus, adapté au problème de la rétroconversion des documents d'un support papier vers un format électronique.

Dans cet article, nous proposons une stratégie pour l'interprétation et la rétroconversion des documents ainsi qu'une modélisation associée, basée sur le concept objet, capable de représenter une large gamme de documents. Dans la section 2., nous présentons les fonctionnalités auxquelles doit pouvoir répondre une modélisation pour être utilisée dans un grand nombre de cas. En particulier, cette modélisation doit pouvoir représenter le résultat (intermédiaire ou définitif) d'une interprétation de document. Nous montrons en quoi les formats de description fréquemment utilisés ne sont pas adaptés. Dans cette section, nous proposons une modélisation basée sur le concept objet qui répond aux besoins exprimés. La section 3. présente les spécifications des objets utilisés ainsi que les méthodes employées pour les identifier. Dans la section 4., nous décrivons la stratégie que nous suivons dans l'interprétation d'un document quelconque. Nous abordons plus précisément les méthodes d'identification des différents éléments du document, l'analyse de divers niveaux de cohérence dans la description globale du document. Enfin nous décrivons la remise en cause des traitements opérés afin de remédier aux incohérences de la description. La section 5. de cet article présente les fonctionnalités auxquelles doit pouvoir répondre un système de rétroconversion des documents. Nous annonçons également des perspectives qui permettent d'implanter ces fonctionnalités au sein d'un système de rétroconversion se basant sur la stratégie et la modélisation proposée.

2. Modélisation des documents

2.1. Vues du document

Le document est l'ensemble constitué par une information et le support sur lequel est inscrit cette information. La rétroconversion consiste à faire passer l'information d'un support papier vers un support électronique. La difficulté de cette tâche réside dans le fait que la représentation du document doit pouvoir exploiter les possibilités offertes par ce nouveau support (réédition, indexation, archivage, diffusion...). Il est d'usage de distinguer plusieurs vues du document. Nous nous intéresserons plus particulièrement à la description de l'image, de la vue physique et de la vue logique du document. Nous ne décrirons pas, dans cet article, les dimensions traitant de niveaux d'abstraction supérieurs tels que le niveau sémantique et le niveau pragmatique... La vue physique représente la façon dont les diverses entités graphiques ou matérielles sont organisées. La vue logique décrit le document du point de vue de l'articulation du propos. Cette structure est souvent représentée de façon hiérarchique en terme de sections, chapitres, parties, sous-parties, paragraphes, éventuellement alinéas...

L'acquisition numérique de l'image du document est la première étape d'un processus de rétroconversion. Cette acquisition permet d'obtenir l'image du document au format électronique. La reconstitution automatique de la structure physique est réalisée après une phase de prétraitement, par la localisation des éléments graphiques de l'image (segmentation). Ces éléments sont ensuite associés de façon à constituer des éléments composites et à reconstituer finalement la structure physique du document. Cette phase du traitement est appelée analyse de document. La reconstitution de la structure logique du document est une étape supplémentaire dans l'interprétation du document, qui vise à extraire l'information véhiculée. Cette étape est souvent réalisée en donnant aux éléments physiques du document un équivalent logique. Dans la plupart des systèmes existants, l'interprétation fait appel à des connaissances sur le type de document traité. Ces connaissances peuvent porter sur la structure logique générique attachée à une classe de documents, partie de la structure logique commune à tous les documents de la classe considérée. Ces connaissances peuvent également porter sur des feuilles de styles qui établissent les mises en forme appliquées aux divers éléments logiques et qui permettent d'effectuer une correspondance entre les éléments physiques et les éléments logiques.

2.2. Modélisations existantes

Les formats électroniques de description de document, sont de plus en plus utilisés en raison des fonctionnalités supplémentaires qu'offre ce support par rapport au support papier. En particulier, l'archivage, l'indexation, la diffusion et la réédition sont facilités grâce à un encombrement moindre que celui du papier et par la puissance de calcul des ordinateurs. Cependant, rares sont les formats qui, en offrant toutes ces possibilités, décrivent toutes les dimensions documents. Le format Tiff ne décrit que l'image. Au contraire, l'image

n'est pas décrite explicitement dans les formats tels que LaTeX ou les normes SGML (Standard Generalized Markup Language) ou XML (eXtended Markup Language) pour lesquels l'image du document est obtenue après une phase de formatage. Le format RTF (Rich Text Format), quant à lui, fait abstraction de la structure logique du document. D'autre part, LaTeX, SGML, XML et RTF ne decrivent pas les éléments graphiques. Ceux-ci ne sont inclus dans le corps du document que par une référence à un fichier graphique externe au document. Le format PDF (Portable Document Format) paraît le plus complet pour décrire les différentes vues du document. En effet, ce format permet l'accès aussi bien à l'image du document, qu'à son contenu textuel et à une ébauche de structure logique grâce à des liens hypertextes. Cependant, la modélisation de la structure physique (mise à part la notion de page) fait défaut. En effet, une représentation de la structure physique paraît indispensable pour l'extraction de l'information à partir de l'image du document. Finalement, il semble que la norme ODA [Int89] soit la plus complète en terme de représentation des structures physique et logique du document.

2.3. Modélisation proposée

Nous proposons, dans cette partie une modélisation des documents qui, non seulement semble adaptée à décrire une large gamme de documents, mais également répond aux besoins d'une majeure partie des exploitations qu'on peut en faire. Cette modélisation, par la richesse des renseignements qu'elle exprime, est en particulier adaptée à la rétroconversion des documents. Il s'agit d'une modélisation objet qui se veut être la plus complète possible. L'interrogation d'une base documentaire se ramène alors à l'expression d'une requête sur les champs modélisés par les objets.

La modélisation que nous proposons décrit l'image, la structure physique et la structure logique du document. Il est bien évident que cette représentation reste incomplète. Elle devra être enrichie par la description des niveaux d'abstraction supérieurs donnant ainsi accès au sens et à l'information proprement dite véhiculée par le document.

Le niveau d'interprétation le plus bas représente l'image du document. Elle résulte de l'acquisition numérique. Nous nous restreignons dans la pratique à une acquisition codée sur 256 niveaux de gris à une résolution de 300 dpi effectuée par un scanner de format A4. Le document pouvant contenir plusieurs pages, le niveau image du document peut en réalité contenir plusieurs acquisitions. Nous avons choisi de conserver l'image du document dans la modélisation complète, la forme formatée correspondant à une des vues du document. Ceci permet également de pouvoir remettre en cause les traitements effectués dès le début de l'interprétation sur une zone précise de l'image du document indépendamment du reste. De plus, cela permettra de décrire les objets physiques par une simple référence à l'image en précisant leur position et leur dimension.

Les niveaux d'interprétation suivants, dans le sens des abstractions crois-

santes, correspondent respectivement à la vue physique et à la vue logique du document. Nous avons choisi de modéliser ces deux vues en nous inspirant de la norme ODA. La représentation des vues physique (cf. figure 1) et logique (cf. figure 2) est indépendante. Cependant chacune d'elles est hiérarchique, le document étant un objet composite composés d'objets physiques et logiques composites et d'objets physiques et logiques de base.

Les objets, qu'ils soient composites ou de base, physiques ou logiques, peuvent être regroupés en classes. Une classe d'objets est décrite par un objet générique. Un objet générique décrit toutes les caractéristiques communes à tous les objets de la classe. De cette façon, la description d'un objet spécifique se fait par une référence à l'objet générique décrivant la classe à laquelle il appartient, et par la description des caractéristiques non exprimées au niveau de l'objet générique. Le processus d'identification de classification est la détermination de la classe d'un objet. Nous décrivons dans la section 3. les diverses méthodes de classification que nous employons.

La norme ODA modélise les vues physique et logique d'un document par des arborescences représentant les divers niveaux d'inclusion des objets, chaque objet contenant comme attribut la liste des objets le composant. Nous enrichissons cette description en décrivant l'ensemble des composants d'un objet composite par un graphe décrivant les relations qui existent entre les composants de l'objet. De cette façon, nous représentons, par exemple, des relations de voisinage entre deux objets composites. Cette représentation offre de plus des possibilités supplémentaires pour la classification des objets. Ces nouvelles possibilités sont décrites dans la section 3..

Sur les figures 1 et 2, nous schématisons la modélisation physique et la structure logique d'un document synthétique. La vue physique est décrite par quatre niveaux hiérarchiques. Au plus haut niveau, on ne trouve qu'un seul objet, racine de la structure, qui correspond à la vue physique du document entier. Le niveau suivant décrit un premier niveau de blocs (bloc correspondant au titre, 1^{re} colonne, 2^{e} colonne, bloc correspondant au pied de page) et les relations de voisinage qui existent entre ces blocs. Le troisième niveau décrit un niveau de blocs plus fin et ainsi de suite. Il est à noter que tous les objets de base ne sont pas au niveau le plus bas. En effet, sur notre illustration, des objets présents sur le troisième niveau n'ont pas de composants.

Nous avons représenté sur la figure 2 ce que pourrait être la vue logique du document synthétique de la figure 1. Nous avons isolé trois niveaux hiérarchiques. À la racine, est décrite la vue logique du document dans sa globalité. Les niveaux suivants décrivent la structure logique selon un découpage hiérarchique de plus en plus fin.

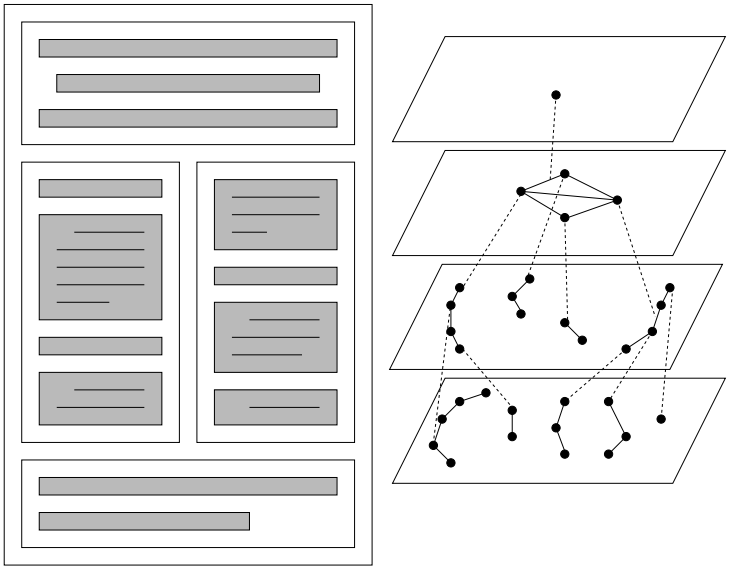


FIG. 1 – *Modélisation physique d'une page de document synthétique*

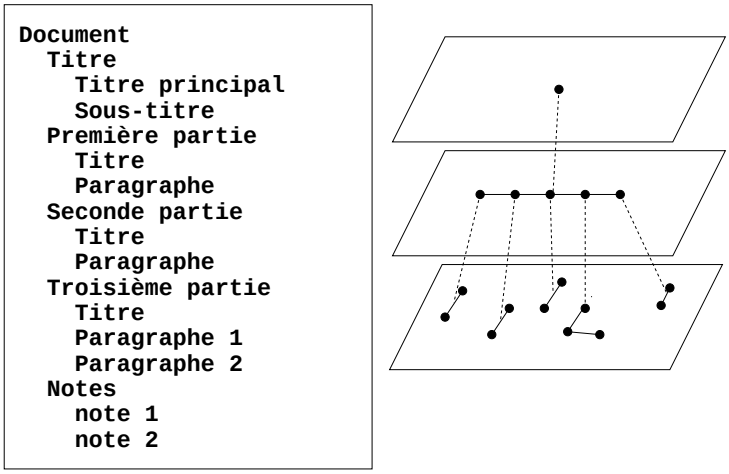


FIG. 2 – *Modélisation logique d'un document synthétique*

3. Classification des objets

3.1. Spécifications

Les structures physique et logique, même si elles représentent des vues différentes du document, utilisent le même formalisme, si bien que les objets physiques et logiques suivent les mêmes spécifications. En revanche, il existe une différence entre les objets spécifiques et les objets génériques décrivant une classe. Les principaux attributs des objets spécifiques sont les suivants :

- un vecteur de caractéristiques intrinsèques à l'objet. Ces informations sont extraites à partir des traitements bas niveaux ;
- une chaîne symbolique exprimant la liste des composants de l'objet ;
- le graphe modélisant la structure selon laquelle sont organisés les composants de l'objet considéré.

Les principaux attributs des objets génériques sont les suivants :

- un vecteur de caractéristiques représentant le centre de la classe dans l'espace de représentation numérique ;
- une matrice de variance-covariance décrivant la distribution des objets de la classe dans l'espace de représentation ;
- une grammaire contenant l'ensemble des règles devant être observées dans la chaîne symbolique des objets spécifiques de cette classe ;
- une chaîne modèle se rapprochant au mieux de toutes les chaînes symboliques des objets spécifiques de cette classe ;
- un graphe représentant la structure commune à tous les objets spécifiques de cette classe.

3.2. Comparaison et classification d'objets

Les attributs définis pour les objets nous autorisent à appliquer différentes méthodes afin de comparer des objets. La comparaison entre objets peut être réalisée par un calcul de distance qui caractérise la proximité des objets. Diverses distances peuvent être calculées entre les vecteurs de caractéristiques de deux objets spécifiques. On peut citer la distance de Hamming, la distance du maximum ou encore la distance euclidienne.

Pour comparer deux objets on peut également comparer les chaînes symboliques représentant la liste des composants de l'objet. Cette comparaison peut être effectuée par le calcul d'une distance d'édition telle que la définissent Wagner et Fisher [WF74].

Enfin, la comparaison de deux objets peut être réalisée par l'examen des graphes [Mic84] représentant les composants de l'objet et les liens qu'ils entretiennent. Pour effectuer cette distance entre deux graphes, nous recherchons le plus grand sous-graphe commun aux deux graphes comparés. Plus ce sous-graphe comporte de nœuds, plus grande est leur partie commune. Cependant, cela ne signifie pas qu'ils sont semblables car ils peuvent avoir une partie dissemblable encore plus grande. Pour tenir compte de ce phénomène, nous établissons

deux taux de recouvrement en ramenant la taille du plus grand sous-graphe commun (nombre de nœuds et nombre d'arcs) à la taille des graphes comparés. L'examen de ces deux taux de recouvrement nous renseigne davantage sur la similitude qu'il existe entre ces deux graphes. Si l'un d'entre eux est égal à 1, cela signifie qu'un des graphes est inclus dans l'autre. Si dans ce cas de figure, l'autre taux est faible, alors le graphe incluant est très grand par rapport au graphe inclus. Lorsque les deux graphes sont égaux, alors les deux taux de recouvrement sont égaux à 1. On peut alors définir une mesure de similarité entre les deux graphes de la façon suivante en considérant t_1 et t_2 les taux de recouvrement du plus grand sous-graphe commun sur chacun des deux graphes comparés :

$$d = \frac{1}{t_1 \cdot t_2} - 1.$$

Les distances entre objets apportent une indication sur le fait que deux objets appartiennent ou non à la même classe. Pour identifier la classe d'un objet, on détermine la classe la plus proche. Nous utilisons alors la distance entre un objet spécifique dont la classe est inconnue et les objets génériques représentant les classes connues. Outre les distances euclidienne, de Hamming et du maximum, qui sont utiles pour déterminer la distance entre l'objet et le centre de la classe, on peut également utiliser la distance de Mahalanobis qui tient compte, avec un *a priori* gaussien, de la dispersion des objets à l'intérieur de la classe. La chaîne symbolique de l'objet spécifique peut être comparée à la grammaire des objets génériques par analyse syntaxique. Cette analyse nous renseigne sur le fait que la chaîne symbolique est conforme aux règles exprimées dans la classe. Cette méthode étant très stricte, nous songeons à implanter une méthode de comparaison plus souple telle que la comparaison entre une chaîne symbolique de l'objet spécifique et une chaîne symbolique modélisant la classe. La comparaison de graphes telle qu'elle a été définie dans le paragraphe précédent est utilisée pour évaluer la distance entre le graphe des composants de l'objet spécifique et le graphe modèle de l'objet générique.

Pour déterminer la classe d'un objet, nous comparons l'objet spécifique à identifier avec tous les objets génériques. Nous avons présenté précédemment trois méthodes de classification : une méthode statistique (distance entre les vecteurs caractéristiques), une méthode syntaxique (comparaison à partir de la chaîne symbolique) et une méthode structurelle (comparaison de graphe). Chacune de ces méthodes établit une liste valuée par l'inverse de la distance de l'objet à la classe, caractérisant une probabilité d'appartenance à la classe. Pour déterminer la classe de l'objet spécifique, il convient d'effectuer une fusion de ces données équivalente à une coopération de classifieurs. Il existe plusieurs façons d'effectuer la fusion de ces données :

- ordre d'apparition dans chacune des listes ;
- cumul pondérés des degrés d'appartenance ;
- multiplication des probabilités d'appartenance.

Le processus de classification des objets exposé ici permet de déterminer la classe d'un objet indépendamment du reste de la structure du document à ceci près que chaque objet fait référence à ses descendants immédiats via la chaîne symbolique et le graphe représentant la liste des descendants et les relations qu'ils entretiennent. Nous détaillons dans la suite comment la modélisation de l'ensemble du document est opérée.

4. Stratégie d'interprétation

Dans cette section nous présentons la stratégie d'interprétation observée lors de la rétroconversion des documents. L'interprétation des documents revient à identifier la classe de chaque objet. Notre stratégie suit un cycle, inspiré du cycle de « perception » utilisé dans [OMLL95], au cours duquel, après l'émission d'une hypothèse sur la classe de chaque objet, nous examinons la cohérence de la description globale du document. Les résultats des traitements ayant mené à des incohérences sont alors remis en cause. Ces remises en cause permettent de fournir de nouvelles hypothèses sur la classe des objets. Le cycle d'interprétation se termine lorsque la description globale du document est cohérente ou lorsque le nombre de cycles effectués devient trop important. Dans la suite de cette section, nous présentons l'initialisation du cycle d'interprétation, les divers niveaux d'incohérence et les diverses remises en cause auxquelles nous avons recours.

4.1. Initialisation du cycle

Après une phase d'acquisition, l'image du document est traitée. Dans un premier temps des traitements de bas niveaux permettent de binariser et de redresser l'image. Une phase de segmentation permet de localiser les objets physiques de base. Ces objets sont ensuite associés au sein d'objets composites en observant des règles portant sur la proximité et le contenu des objets. On se référera à [DTJ+97] pour plus de détails. On obtient de cette façon une première ébauche de la structure physique du document. La modélisation complète du document repose sur l'identification de l'ensemble des objets physiques et logiques. Dans un premier temps, nous réalisons une première classification des objets de la structure physique obtenue. Cette première classification des objets est réalisée indépendamment du contexte dans lequel ils se trouvent. Elle se fait donc uniquement grâce à la comparaison du vecteur de caractéristiques de l'objet à identifier avec ceux des objets génériques, ce vecteur de caractéristiques étant le seul des attributs de l'objet à ne pas faire référence aux composants de l'objet qui n'ont pas encore été reconnus. Cette première phase initialise la reconnaissance de la structure physique du document en proposant une hypothèse de reconnaissance de chacun des objets. De cette première structure physique, nous émettons une hypothèse sur la classe du document traité. À chaque classe de documents correspond une structure logique générique. Une

première structure logique est alors initialisée en exploitant l'hypothèse sur la classe du document. En réalité, pour chacune des classes d'objets, il est émis une probabilité sur le fait que l'objet à identifier appartienne à cette classe.

4.2. Cycle d'interprétation

Les phases suivantes du cycle d'interprétation sont réalisées en effectuant un examen de la cohérence de la description du document et en remettant en cause les traitements ayant conduit à des incohérences.

Nous avons isolé plusieurs niveaux d'incohérence. La remise en cause qui sera effectuée par la suite afin de palier l'incohérence dans la description dépend du niveau d'incohérence observé.

Le premier niveau de cohérence que nous définissons est appelé cohérence intrinsèque. Un objet est dit intrinsèquement cohérent si les caractéristiques qui lui sont propres sont plausibles au regard des connaissances dont on dispose, s'il est susceptible d'appartenir à une des classes d'objets connues.

Le second niveau de cohérence que nous abordons est intitulé cohérence contextuelle de voisinage. Ce niveau concerne le fait qu'un objet d'une certaine classe peut être rencontré dans l'environnement des objets de même niveau hiérarchique qui l'entourent. Un objet est dit cohérent au niveau contextuel de voisinage s'il existe au moins une classe d'objets pouvant l'inclure ainsi que ses voisins et dans la configuration observée.

Nous définissons également un niveau de cohérence hiérarchique. Un objet est dit hiérarchiquement cohérent si la classe à laquelle appartiennent ses composants est compatible avec la classe à laquelle lui-même appartient.

Enfin, nous définissons le niveau de cohérence d'interprétation qui traite de la possibilité que deux objets puissent être mis en correspondance à deux niveaux d'interprétation différents (cohérence entre la classe attribuée à un objet physique et celle attribuée à l'objet logique décrivant la même entité). Les structures physique et logique du document étant indépendantes dans la modélisation que nous utilisons, il n'est pas toujours possible d'établir cette cohérence d'interprétation à tous les niveaux. Cependant, les objets physiques et logiques de base doivent pouvoir correspondre. Il en est de même pour les racines des structures physiques et logiques. En effet, la classe de documents attachées à la structure physique doit être identique à celle attachée à la structure logique.

La remise en cause est le traitement appliqué afin de résoudre les incohérences rencontrées dans une modélisation. Les traitements effectués lors d'interprétations ayant conduit à des incohérences doivent être révisés. Nous avons isolé plusieurs niveaux de remise en cause.

En premier lieu, nous pouvons remettre en cause les classifications des objets qui ont été réalisées. Des éléments nouveaux peuvent alors être pris en compte afin de proposer une nouvelle classe d'appartenance pour un objet. En effet, la connaissance des classes affectés aux objets voisins ou hiérarchiquement inférieurs ou supérieurs permet de rendre plus robuste la classification. Cette

prise en compte du voisinage de l'objet s'effectue en intégrant la classification structurelle de l'objet. D'un autre coté, la classification peut être orientée. La détermination de la classe d'un objet nous donne une connaissance *a priori* sur la classe des objets qui le composent. Cette connaissance peut être introduite lors de la coopération des classifieurs au même titre que le résultat d'une classification statistique, syntaxique ou structurelle.

Nous pouvons également remettre en cause la phase de structuration. L'objet considéré existe-t-il ? N'est-il pas qu'une partie d'objet ? Ou au contraire, est-il le regroupement de plusieurs objet ? Cette remise en cause se traduit par la fusion ou la scission d'objet, par la destruction et l'apparition de nouveaux liens hiérarchiques. La remise en cause de la structuration est surtout guidée par les connaissances portant sur la structure des objets dont la classe est bien déterminée.

On peut également être amené à remettre en cause les données extraites. En effet, une nouvelle extraction des données de base localisée sur la zone dont l'interprétation est incohérente peut remédier au problème rencontré. De même, la reconnaissance de certains objets peut orienter vers l'emploi de traitements plus appropriés aux données que ceux ayant conduit à des interprétations incohérentes.

Enfin, il se peut que malgré les remises en cause effectuées, nous ne parvenions pas à une interprétation cohérente. Ceci est, en particulier, dû au fait que les objets observés n'appartiennent pas à des classes d'objets connus. En effet, le fait de ne pas pouvoir rattacher un objet à une classe connue ne signifie pas que l'objet n'existe pas dans sa forme observée. Il se peut que nous n'ayons pas connaissance de la classe à laquelle il appartient réellement. Auquel cas il peut être nécessaire d'apprendre cette nouvelle classe en la situant par rapport aux classes déjà connues par l'exploitation des similitudes et des différences qu'elle entretient avec ces dernières. Il s'agit, en fait, de la remise en cause de nos connaissances.

5. Conclusion et perspectives

Dans cet article, nous avons mis en évidence la nécessité de définir un format de description des documents adaptés aux plus grand nombre d'applications afin de faciliter l'échange de documents. Nous avons mis l'accent sur le fait qu'il existait des besoins au niveau de la rétroconversion des documents papier vers un format électronique. En particulier, ce processus nécessite la description de toutes les dimensions du document, des divers niveaux d'abstraction avec lequel il peut être considéré. Nous avons proposé une modélisation objet des documents qui décrit les structures physique et logique du document. Nous avons détaillé les spécifications des objets utilisés et les méthodes qui permettent de les identifier.

Enfin, nous avons proposé une stratégie d'interprétation des documents utilisant la modélisation objet décrite précédemment. Cette stratégie s'appuie

sur la classification des objets, une analyse de la cohérence de la description et la remise en cause des représentations incohérentes.

Cette stratégie doit être implantée dans un système de traitement des documents dont la fonction est d'interpréter une large gamme de documents. Cette interprétation est réalisée en utilisant la méthodologie (modélisation et stratégie) décrite dans cet article. Afin de fiabiliser les interprétations, nous souhaitons implanter dans ce système une interface avec un opérateur qui validera, corrigera la modélisation. L'opérateur pourra également proposer une modélisation pour les parties de document que le système n'aura pas réussi à établir automatiquement. Ceci pourra être l'occasion pour le système d'enrichir ses connaissances, d'apprendre ou de préciser la définition de nouvelles classes d'objets. L'apprentissage incrémental devra également pouvoir être effectué automatiquement en exploitant les similitudes et les différences avec les classes d'objets connues les plus proches de celles observées. Ces fonctionnalités donnent au système un caractère évolutif. Le caractère adaptatif correspond au fait que les documents du même type sont de plus en plus efficacement traités. Cette fonctionnalité peut être implantée en évaluant, en fonction du contexte, les traitements déclenchés lors de l'interprétation. Cette évaluation peut être assez simple et consister à savoir si dans un certain contexte, le traitement a, par la suite été remis en cause ou pas. De cette façon, en capitalisant « l'expérience », lors de l'interprétation de nouveau document on pourra, en fonction du contexte, s'orienter vers les traitements ayant le plus de chances d'aboutir.

Références

[DTJ+97] Sébastien Diana, Éric Trupin, Frédéric Jouzel, Jacques Labiche, and Yves Lecoutier. From acquisition to modelisation of a form base to retrieve information. In *4th International Conference on Document Analysis and Recognition*. IAPR, 1997.

[Int89] International Standard Organization. *ISO 8613 : Information Processing. Text and Office System, Office Document Architecture (ODA) and Interchange Format*, 1989.

[Mic84] Laurent Miclet. *Méthodes structurelles pour la reconnaissances de formes*. Eyrolles, 1984.

[OMLL95] Jean-Marc Ogier, Rémy Mullot, Jacques Labiche, and Yves Lecourtier. Interprétation de document par cycles «perceptifs» de construction d'objets cohérents. application aux données cadastrales. *Traitement du Signal*, 12(6):627–637, 1995.

[WF74] R. A. Wagner and M. J. Fisher. The string to string correction problem. *ACM*, 20(10), May 1974.