Annotation, indexation et parcours de documents numériques

Consulter, explorer et maintenir les textes réglementaires

François Lévy — Adeline Nazarenko — Abdoulaye Guissé

LIPN UMR 7030 (Université Paris 13 et CNRS)
99, avenue J.-B. Clément, F-93430 Villetaneuse
prénom.nom@lipn.univ-paris13.fr

RÉSUMÉ. La mise sous forme numérique des documents renouvelle les rapports de l'annotation et de l'indexation, grâce aux traitements qu'elle rend possibles. Nous proposons un cadre unifié (appelé index) dans lequel ces deux opérations se complètent, permettant un jeu d'allers-retours entre des documents et une représentation sémantique. Ce cadre présente une bonne couverture théorique et rend compte d'activités diverses. Nous montrons son efficacité pratique à partir d'une expérience en cours sur l'indexation des textes réglementaires, dont nous expliquons la problématique et pour laquelle nous proposons une structure d'index originale. Cet exemple illustre en outre comment l'index peut devenir un espace de navigation entre le texte et le modèle sémantique, espace qui permet de nouveaux usages du document numérique.

ABSTRACT. The emergence of digital documents renews the relation between annotation and indexing, thanks to the processing it enables. We propose a unified framework (so called index) in which these two operations complement each other and support the interplay of bidirectional moves between documents and a semantic representation. This framework shows a good theoretical coverage and accounts for different activities. Its practical efficiency is illustrated on an ongoing experiment, which aims at indexing texts describing policies, which issue is explained and for which we present an original index structure. This example illustrates how that structure can be turned into a navigating space connecting a text and a semantic model, and how, on this ground, new functionalities can be built which enlarge the uses of digital documents.

MOTS-CLÉS : annotation sémantique, indexation, ingénierie des connaissances, navigation documentaire, règles métiers, réglementations.

KEYWORDS: semantic annotation, indexing, knowledge engineering, document navigation, business rules, policies.

DOI:10.3166/DN.13.3.121-152 © 2010 Lavoisier, Paris

1. Introduction

L'annotation et l'indexation sont des pratiques antérieures au développement des documents numériques, mais la multitude des usages tend à les fondre dans une notion de métadonnée relativement indifférenciée. Si le passage du support papier à l'ordinateur déplace les lignes, il nous semble utile de continuer à distinguer les deux problématiques pour montrer en quoi le support informatique enrichit chacune et leur permet de coopérer dans la construction d'un mode d'accès plus structuré au document, qui aurait été impensable autrefois.

L'argumentation proposée explique d'abord en quoi, selon nous, l'évolution technique fait bouger les conceptions traditionnelles de l'annotation et de l'indexation (section 2). Cette analyse permet de poser un cadre unifiant que nous appelons *index*. La section 3 propose un modèle de cette structure d'index et montre qu'il s'applique à diverses pratiques de l'indexation. La section 4 présente de façon beaucoup plus précise l'exemple d'index que nous avons proposé pour l'intégration de la documentation réglementaire dans les systèmes de gestion des règles métiers. Nous décrivons les besoins du domaine d'application et la structure d'index qui contribue à y répondre, en montrant concrètement comment un texte réglementaire peut être indexé. La section 5 développe, à propos de la gestion des règles métiers, les fonctionnalités de haut niveau auxquelles la navigation dans la structure d'index donne accès, tandis que la dernière section élargit la réflexion à d'autres exemples d'application.

2. De la note au parcours

2.1. Deux intentions différentes

À l'origine, une annotation est une remarque manuscrite notée en marge d'un texte¹. Le terme est plus restrictif que « note », la note n'étant pas en soi limitée par la place disponible dans les interstices du texte. Les notes de bas de page, notes de lecture, notes d'information, etc. sont en principe plus concises qu'un développement, un compte rendu, un rapport : l'éditeur érudit fait une note quand sa remarque mérite moins qu'un commentaire. L'annotation est plus lapidaire au contraire et laisse peu de place à l'argumentation. Elle reste souvent un texte libre dont on ne peut prévoir la forme, même s'il existe d'autres usages plus codés, comme les annotations typographiques. Un autre trait de l'annotation traditionnelle nous paraît significatif : insérée dans le texte, elle n'est accessible que par un parcours de celui-ci. L'inventaire des annotations suppose de tout lire mais, en général, cet inventaire n'est pas particulièrement intéressant : non structurées, les annotations ont peu de raison de se rassembler en une liste porteuse de sens ; on peut tout au plus en faire le décompte comme le professeur qui recense les barbarismes, solécismes et contresens pour évaluer une traduction en contexte scolaire. Dans le cas des annotations, l'intention essentielle va du texte aux annotations : c'est la lecture du premier qui gouverne l'accès aux secondes.

1. Voir l'article « annotation » du *Trésor de la Langue Française*.

Par étymologie, l'index est le doigt qui désigne, d'où la marque qui distingue et la liste qui pointe. L'index peut être un curseur placé dans un paquet de fiches pour repérer où s'est arrêté le parcours aussi bien que la liste des livres pointés par les autorités religieuses – interdits à la lecture. En informatique, c'est une structure intermédiaire qui mémorise l'adresse de certaines données et permet de les retrouver rapidement. En fin de livre, c'est une liste de termes associés aux numéros de pages où on peut les trouver. Dans une large mesure, la sémantique du lien d'index lui est externe. L'inscription dans la liste distingue un élément parmi la liste de tous les éléments et c'est la liste elle-même qui porte la sémantique de la distinction : l'interdiction pour les livres « mis à l'index », la pertinence des mots-clefs d'un index de fin de livre. Alors que chaque annotation se lit individuellement, un lien d'index ne s'interprète que collectivement, par des critères de choix identiques pour tous les liens : on peut indexer la première occurrence de chaque terme, ou ses occurrences les plus importantes, ou sa définition ; on peut aussi choisir d'indexer des concepts plutôt que des termes. Toujours est-il que l'intention va, cette fois, de l'index au texte et non l'inverse, c'est l'index qui guide la lecture, l'affranchit de la linéarité ou du hasard.

Ceci étant, si l'index s'utilise de l'extérieur vers le texte ou la collection indexée, il doit être construit et sa construction se situe dans un mouvement différent : il s'agit alors de parcourir l'ensemble visé pour en extraire les adresses qui doivent être mémorisées et l'entrée d'index qui y renverra. La sélection d'une adresse et le choix de la clef associée sont analogues au processus d'annotation décrit précédemment, à ceci près que la sémantique n'est pas libre, elle est commune à tout l'index.

Cette absence de solution de continuité entre les deux notions explique que le terme « annotation » soit parfois utilisé dans un sens large pour couvrir l'ensemble mais nous allons essayer de montrer que la différence d'intention reste pertinente dans le cadre du document numérique.

2.2. Ce que change la numérisation

En changeant de support, le texte change de contraintes. Nous ne rentrons pas ici dans une analyse générale de ces changements qui sort de notre sujet mais il faut noter, pour l'annotation, que la limitation spatiale posée par la marge disparaît. On peut annoter un document électronique *ad libitum*, à condition de pouvoir stocker ces données supplémentaires et de trouver un moyen de les présenter au lecteur sans détruire la lisibilité de l'original. L'émergence de standards d'annotation comme Annotea² et de nombreux outils d'annotation (Uren *et al.*, 2006) facilitent la construction de ces annotations tandis que les progrès techniques faits depuis vingt ans en matière d'interfaces et de présentation graphique rendent cet objectif atteignable (Hearst, 1999).

Pour les structures de données qui stockent informatiquement l'écrit, la marge n'est pas un objet. De fait, l'informatique stocke relativement rarement l'image de la page et sépare souvent dans ses structures de données le texte et des éléments de

2. www.w3.org/2001/Annotea/

présentation qui s'y appliquent. Or, sur papier, la marge offre une grande souplesse d'annotation : le texte ajouté peut avoir un ancrage et une portée précis (marqués par une astérisque ou des crochets, par exemple), une portée plus vague (un trait dans la marge) ou être simplement juxtaposée au texte. Le support électronique n'offre pas toujours la même liberté. L'annotation reste distincte du texte mais elle lui est rattachée avec une portée sans ambiguïté (le texte dans son entier ou un passage particulier).

Si le support électronique ne modifie guère la manière dont un index est utilisé, il en facilite beaucoup la construction. Le coût d'une recherche de chaîne de caractères étant incomparablement réduit quand on emploie l'ordinateur, on peut construire facilement des index dont les entrées sont caractérisées lexicalement : par exemple tous les paragraphes commençant par le mot-clef « Définition » ou toutes les premières occurrences des mots d'un dictionnaire spécialisé. Associée aux possibilités de présentation, cette facilité permet d'envisager l'utilisation d'index construits à la demande selon les besoins du lecteur. Nous considérons cependant qu'allier l'annotation et l'indexation permet de passer à une caractérisation sémantique à proprement parler.

L'annotation traditionnelle n'est accessible, avons-nous remarqué, que par le parcours du texte. Or le support électronique peut faciliter ce dernier : il devient possible de rassembler les annotations tout en notant leur position dans le texte. Une fois qu'elles sont ainsi rassemblées, c'est l'organisation conceptuelle de ces annotations – dans la mesure où celle-ci propose une structure d'accès significative pour le lecteur – qui constitue pour lui un moyen intéressant pour parcourir le document.

3. Structure d'un index

Nous décrivons ici comment les structures qui servent à l'indexation peuvent être spécifiées dans toute leur généralité. Tout d'abord les termes. « Indexation » est utilisé dans la suite pour désigner le processus qui consiste à poser des annotations sur un texte et à les rassembler en suivant une organisation sémantique. Nous parlons d'« index » quand il s'agit de la structure complexe qui en résulte : elle associe une structure sémantique et un texte *via* des annotations sémantiques particulières. Bien qu'on emploie facilement le terme d'index pour désigner la seule structure sémantique (c'est-à-dire l'ensemble des entrées d'index), celle-ci ne joue son rôle que parce que chaque entrée est associée à un ou plusieurs éléments textuels.

3.1. *Modèle d'index*

Les entrées d'index sont contraintes, avons-nous dit, par la structure qui les rend efficaces en tant qu'accès au texte. Nous employons le terme de « modèle sémantique » pour désigner cette structure. Les fragments textuels annotés sont eux aussi contraints : ils doivent, pour la même raison d'efficacité, se prêter à une unité d'usage, en général par rapport à une forme de lecture projetée. Le texte dont il s'agit est donc structuré

par ces contraintes, ce que nous traduisons par la notion de « modèle de document ». Enfin, les liens qui attachent une entrée d'index à un élément textuel méritent qu'on puisse les saisir pour les inverser, les analyser, les qualifier. Nous parlons dans ce cas de « modèle de correspondance ». Plus formellement et de manière générale, nous définissons donc un index comme une structure composée de trois sous-structures.

Le *modèle de document* détermine quels fragments de texte constituent des unités documentaires pouvant supporter un lien d'indexation. En principe, n'importe quelle liste de caractères peut être annotée mais on se limite souvent, en pratique, à des intervalles continus³. On peut aussi typer les séquences de caractères en question pour distinguer des mots, des syntagmes, des phrases, des paragraphes, des sections de document, etc. La structure d'index dépend largement du modèle de document considéré. Dans le cas de la plate-forme KIM (Popov *et al.*, 2003), par exemple, seules les entités nommées peuvent servir de support à l'indexation. Dans un index traditionnel comme celui de la base MedLine⁴, l'unité documentaire généralement considérée est le document pris dans son ensemble. Dans d'autres cas, le modèle documentaire prend la forme d'un réseau de documents qui sont reliés les uns aux autres par des liens sémantiques⁵.

Le *modèle sémantique*, quant à lui, indique quelles unités sémantiques peuvent être associées aux unités documentaires et quelles relations ces unités sémantiques entretiennent entre elles. Ce modèle sémantique peut être un simple thésaurus (les documents de MedLine sont associés à des descripteurs extraits du thésaurus MESH⁶ (Névéol *et al.*, 2007)). Dans le cadre du web sémantique, il s'agit généralement d'un modèle ontologique mais il est souvent utilisé de manière partielle comme dans (Popov *et al.*, 2003; Amardeilh, 2007) où seules des instances peuvent être utilisées pour annoter⁷ alors que nous proposons de généraliser l'approche en donnant n'importe quel élément de l'ontologie (concept, rôle, instance de concept ou de rôle) comme cible du lien d'indexation.

Le *modèle de correspondance* associe des unités documentaires (ud), à des unités sémantiques (us). Dans le cas le plus simple, un lien d'indexation se représente comme un couple (ud_i, us_i) mais ces liens peuvent porter des propriétés : on peut les typer pour marquer le rôle que le fragment de texte joue par rapport à l'unité sémantique (définition *vs.* exemple), indiquer l'usage pour lequel ils sont proposés (spécialisé *vs.* grand public), leur associer un poids de pertinence, garder trace de leur l'origine, etc. Dans le cas général, un lien d'indexation se représente donc comme un triplet (l_i, ud_i, us_i) où l_i est la liste des propriétés du lien associant ud_i à us_i .

3. La contrainte a un avantage calculatoire certain mais elle a ses limites : elle ne permet pas de rendre compte des coordinations comme « La 1re et la 2e classe », où il faudrait pouvoir indexer « La 1re ... classe » aussi bien que « la 2e classe ».

4. www.ncbi.nlm.nih.gov/pubmed

5. Nous en verrons des exemples dans la section 6.

6. www.nlm.nih.gov/mesh

7. Elles sont créées à la volée lors de l'annotation pour peupler l'ontologie.

3.2. Discussion

Rien dans ce schéma n'impose de restreindre la structure sémantique à des concepts, des rôles et leurs instances. Les modèles ontologiques sont des outils commodes mais leur pouvoir d'expression est limité : il est plus restreint que celui de la logique du premier ordre, qui elle-même ne rend pas compte de tous les contenus langagiers. Les textes réels, fussent-ils dédiés à des applications techniques, ont des contenus qui débordent le cadre de ce qui peut se représenter sous la forme d'ontologies : la section 3.3 montre que cette structure sémantique peut être élargie pour mieux rendre compte des prédications, actions, événements que les textes réglementaires mentionnent ; pour d'autres cas d'usage, il faudrait modéliser les points de vue ou différents états chronologiques.

A noter également que les unités sémantiques ne s'identifient pas aux éléments mémorisés. L'indexation documentaire illustre très bien ce point. Chaque document possède une liste de mots-clefs et l'unité mémorisée est l'association (document, mot-clef) mais l'accès à l'ensemble documentaire se fait par des requêtes qui sont des combinaisons de mots-clefs et l'ensemble des documents associés est calculé dynamiquement⁸. Le modèle sémantique est en fait l'ensemble des requêtes acceptables pour le système mais ses unités restent *virtuelles* et seuls les mots-clefs sont réalisés. Ce processus repose sur l'existence d'opérateurs qui permettent de combiner les unités sémantiques pour en créer de nouvelles et qui se traduisent directement en termes d'indexation : par exemple, l'association d'un document à la clef composée k_1 ET k_2 doit se déduire de son indexation par k_1 et par k_2 .

Enfin, les liens d'indexation décrits dans le modèle de correspondance ne sont pas orientés. Le support informatique permet de les utiliser dans les deux sens. C'est ce jeu d'aller-retours entre deux modèles qui transforme véritablement l'index en espace de navigation.

3.3. Exemples d'index

La plupart des pratiques modernes d'indexation et d'annotation se rattachent au modèle d'index décrit ci-avant. Citons quelques exemples classiques.

L'exemple le plus simple et peut-être le plus emblématique est celui de la recherche d'information. Dans un moteur de recherche, le modèle de document se caractérise par sa taille – jusqu'à plusieurs milliards de pages web – et sa simplicité, un texte étant représenté par une séquence de mots. La structure sémantique de l'index est constituée par le vocabulaire des documents. Le modèle de correspondance associe aux mots-clefs les documents qui les contiennent. Dans ce cas, la richesse de l'index est moins portée par la structure de l'ensemble des clefs d'indexation (modèle sémantique), que

8. Pour des raisons d'efficacité, il arrive que l'ensemble des documents associés à un mot-clef soit précalculé mais le stockage intégral de toutes les combinaisons de mots-clefs et des ensembles documentaires associés n'est cependant pas envisageable.

par le nombre de liens d'indexation, les poids qui leur sont associés et le langage de requêtes qui permet dynamiquement de combiner différents mots-clefs en requêtes complexes (Baeza-Yates *et al.*, 1999). Dans le cas du moteur de recherche, le parcours va de l'index au texte plutôt que dans le sens inverse, même si on peut pondérer les mots-clefs par des informations issues des textes, à commencer par leur fréquence d'apparition.

Un exemple très traditionnel est celui des index de fins de livres (Mulvany, 1993; University of Chicago Press Staff, 2003) dont il a déjà été question précédemment et qui sont aujourd'hui revisités à la faveur de la mutation numérique pour faciliter et enrichir les index des ouvrages papier (Nazarenko *et al.*, 2007) mais aussi pour les adapter aux documents numériques et les intégrer à des moteurs de recherche (Anick, 2001) ou à des applications de navigation (Wacholder *et al.*, 2001). L'idée consiste en général à analyser le contenu du texte pour extraire les mots-clefs les plus pertinents et les organiser en un index qui donne accès aux différents segments de texte. L'originalité vient de ce que ces index peuvent être calculés dynamiquement à partir des documents retournés par une première requête, là où l'indexation classique est statique. Dans ce cas, la structure sémantique est donnée par l'ensemble des mots-clefs extraits et la structure qui lui est associée – liste plate, liste hiérarchisée, nuage de mots ou diverses métaphores visuelles (Hearst, 1999) –, la base documentaire est constituée de la liste des documents retournés par le moteur de recherche en réponse à une requête et les liens d'indexation sont calculés dynamiquement lors de l'extraction des mots-clefs. Il faut cependant souligner que ce type d'index est conçu pour naviguer de la structure sémantique vers le texte uniquement.

A l'opposé des moteurs de recherche se situe le modèle d'indexation du web sémantique qui vise à associer aux documents une sémantique formelle interprétable par des agents logiciels (Bontcheva *et al.*, 2003; Kalyanpur *et al.*, 2003). Dans ce cas, le modèle sémantique est généralement donné par une ontologie ou un ensemble d'ontologies et les liens d'indexation traduisent l'interprétation qui est faite du contenu documentaire. Les tailles de la collection documentaire, du modèle sémantique et de l'index lui-même sont beaucoup plus réduites que dans les exemples d'application précédents mais le modèle sémantique sert de support à des raisonnements dont nous montrons plus loin la richesse et l'intérêt. Dans ce type d'index, la navigation peut aller dans l'un ou l'autre sens, soit que l'ontologie éclaire le texte, soit que le texte serve à enrichir l'ontologie (Handschuh *et al.*, 2003).

Avec les folksonomies, le web 2.0 présente une nouvelle forme d'indexation, dite « sociale », où les liens d'indexation sont posés librement par les utilisateurs qui construisent de la sorte une structure sémantique représentée généralement sous la forme de nuages d'annotations ou de « tags ». Ces regroupements induits confèrent au modèle sémantique une structure plus riche qu'une simple liste de mots-clefs mais moins formalisée que dans le web sémantique. Le nombre des annotations compensant pour partie la pauvreté sémantique de l'index, cette structure sémantique lâche permet d'opérer des rapprochements entre documents ou entre annotateurs. Ce type d'annotation apparaît de plus en plus complémentaire de l'indexation classique des

documentalistes (Chan, 2009). Dans ce dernier cas de figure, on parcourt les liens d'indexation essentiellement pour accéder aux documents mais le parcours inverse existe aussi : il permet de regrouper des mots-clefs en nuages qui éclairent à leur tour la sémantique de la collection de documents.

Le tableau 1 montre comment ces différents types d'index se positionnent les uns par rapport aux autres.

	Modèle de document	Modèle de correspondance	Modèle sémantique
Moteur de recherche	Séquence de mots	Simple liens pondérés	Vocabulaire
Index de fin de livres	Séquence de mots, pages, sections	Simple liens rarement pondérés	Terminologie spécialisée
Web sémantique	Séquence de mots	Simple liens	Ontologie
Folksonomie	Ensemble de documents	Liens pondérés par l'usage	Ensemble ouvert de tags

Tableau 1. *Divers types d'index*

4. Indexation des documents réglementaires

Cette section montre comment la structure d'index précédente peut être utilisée dans un contexte particulier, celui de la gestion des règles métiers qui vise à mettre en œuvre les politiques d'entreprises dans des systèmes d'information. Nous montrons l'intérêt de l'indexation des documents réglementaires qui expriment les politiques d'entreprises.

4.1. Contexte applicatif : la gestion des règles métiers

4.1.1. La gestion des règles métiers

« Un système de gestion de règles métiers (en anglais *business rules management system*, ou *BRMS*) est un logiciel qui gère et supporte les règles métiers d'une organisation ou d'une entreprise »⁹ et assiste ainsi la prise de décision de cette organisation ou de cette entreprise. Il peut s'agir du calcul de l'impôt d'un contribuable comme de l'attribution de points à des clients ou du contrôle qualité d'un processus industriel.

La mise au point d'un tel système repose sur l'acquisition des règles du métier considéré, leur représentation dans une forme interprétable par la machine ainsi qu'un

9. Article « Système de gestion de règles métiers » de Wikipedia, février 2010.

moteur d'inférence efficace capable d'orchestrer un ensemble éventuellement très important de règles et de contrôler leur application sur des cas particuliers.

Cette question de l'acquisition des règles métiers a été abordée jusqu'ici sous deux angles différents. Le premier consiste à définir un langage contrôlé et à demander aux personnes en charge d'écrire les règles de les écrire dans ce langage contrôlé. C'est l'approche adoptée dans Oracle Policy Automation (OPAdoc, n.d.) qui vise l'application de textes réglementaires à des cas particuliers ou dans le projet Sparcle d'IBM (Carolyn Brodie, 2006; Reeder *et al.*, 2007) qui porte sur les règles d'accès à l'information dans un gros système informatique. Ces deux projets utilisent en fait des langages contrôlés fortement contraints, accompagnés dans le second cas d'une assistance à l'écriture par listes de mots. Ces langages restent en principe assez familiers à l'utilisateur mais ils sont loin des textes qui formulent les politiques à l'intention de l'entreprise ou de ses clients. Il faut donc tout reformuler manuellement.

La seconde approche s'intéresse davantage à la langue naturelle et tente d'avancer dans l'analyse automatique des textes réglementaires dont les documents de spécification en langue naturelle peuvent être vus comme un cas particulier. A. Joshi *et al.* ont étudié la conformité de procédures à des textes réglementaires ((Dinesh *et al.*, 2006; Dinesh *et al.*, 2007; Dinesh *et al.*, 2008)). Ils décrivent ces derniers par des arbres syntaxiques abstraits, dont ils estiment qu'un ingénieur peut les traduire dans une logique CTL¹⁰. Les procédures sont par ailleurs modélisées comme des graphes de transition et la conformité du modèle à la logique est vérifiée avec des techniques de model-checking. Dans le domaine plus large des spécifications en langage naturel, (Ambriola *et al.*, 2006) proposent une architecture générale dont la première phase est aussi l'extraction d'un arbre syntaxique qui est ensuite traité par divers modules de modélisation. Malheureusement, le langage analysé reste fortement contraint, ce qui nous ramène à la première stratégie. L'architecture proposée par (Santiago *et al.*, 2009) traduit du langage naturel en formules logiques à partir desquelles des tests peuvent être générés mais le module d'analyse du langage naturel n'en est encore qu'au stade de prototype.

Le projet Ontorule¹¹, dans lequel s'inscrit ce travail, propose une troisième voie. L'objectif est de concevoir une plate-forme d'acquisition, de maintenance et d'exécution qui rende les règles du système accessibles aux gens du métier de manière à ce qu'ils puissent à la fois contrôler l'écriture des règles, assurer leur mise à jour et comprendre le pourquoi des décisions proposées par le système. Cette exigence d'accessibilité se traduit par une double contrainte :

- les règles doivent être exprimées à un niveau conceptuel, avec un vocabulaire qui soit explicite et, autant que faire se peut, maîtrisable par les gens des métier. Ce vocabulaire est exprimé dans une ontologie formelle du domaine et les règles doivent faire référence aux éléments de cette ontologie ;

10. CTL est une logique temporelle utilisée pour le model-checking.

11. Projet EU-IST Integrated Project 2009-231875 ONTORULE ("Ontologies meet business rules").

– l'ensemble des connaissances du système (ontologie et règles) doit être documenté, ce qui suppose de faire le lien avec les politiques et réglementations telles qu'elles sont rédigées, quand elles le sont.

Dans le contexte de la gestion des règles métiers, le document informatique a ainsi un double rôle : il est à la fois source de connaissances et document de référence. Même si les deux aspects sont liés, nous ne développons pas le premier qui renvoie à la question de l'acquisition des connaissances à partir de textes (Cimiano, 2006; Aussenac-Gilles *et al.*, 2008; Mondary *et al.*, 2008). Nous mettons l'accent sur le deuxième aspect qui relève de la problématique de l'indexation. Dans le contexte de la gestion des règles métiers, la structure d'index telle qu'elle a été définie précédemment montre tout son intérêt : elle permet d'articuler le texte et un modèle sémantique qui est ici composé non seulement d'une ontologie mais aussi d'une base de règles (voir section 4.2). L'index construit ainsi un espace de navigation dont nous montrons l'utilité pour la gestion des règles métiers dans la section 5.

4.1.2. Exemples de textes réglementaires

Il existe une grande diversité de textes réglementaires. Nous en présentons quelques exemples ici.

Le premier exemple est celui sur lequel s'appuie le reste de l'article. Il s'agit de la description publique du règlement du programme de fidélisation (AAdvantage) de la compagnie American Airline. Ce texte, d'une dizaine de pages, publié sur le site de la compagnie¹², décrit les droits et devoirs du client lorsqu'il utilise le programme et explique les règles d'attribution des bonus et points de fidélité aux passagers de la compagnie, à partir des kilomètres (*miles*) de vol qu'ils effectuent. Voici un extrait de ce règlement :

AAdvantage flight mileage credit is determined on the basis of nonstop distances between the airports where your flight originates and terminates. On connecting flights, you'll receive mileage credit for each segment of your trip. On single-plane flights, you'll receive the nonstop origin-destination mileage.

Mileage credit cannot be earned for the same flight in more than one of the following programs : the AAdvantage program or any other loyalty program in which American Airlines participates¹³.

12. Ce texte sous copyright est utilisé avec l'autorisation d'American Airline, que nous remercions (www.aa.com).

13. Dans AAdvantage, les kilomètres de vol crédités sont déterminés sur la base des distances en vol direct entre les aéroports où votre vol commence et se termine. Pour les vols avec correspondance, vous recevez un crédit kilométrique pour chaque segment de votre trajet. Pour les vols sans changement d'avion, vous recevez le kilométrage du vol direct de l'origine à la destination. On ne peut gagner des crédits kilométriques pour le même vol dans plusieurs des programmes suivants : le programme AAdvantage ou tout autre programme de fidélité auquel American Airline participe.

Un autre exemple provient du règlement financier des projets européens¹⁴ :

There is only one pre-financing payment (advance payment) during the life of the project. It will be received by the coordinator at the beginning of the project and in any case within 45 days of the entry into force of the grant agreement (unless a special clause stipulates otherwise). The coordinator will distribute it to the other beneficiaries :

- Once the minimum number of beneficiaries as required by the call for proposals have signed and returned Form A (accession form), and
- Only to those beneficiaries who have signed and returned Form A.

Like any other payment, the coordinator will distribute the pre-financing to the other beneficiaries in conformity with the ECGA and the decisions taken by the Consortium

Un troisième exemple est donné par l'accord signé par les présentateurs des "tutorials" organisés dans le cadre de la conférence AAAI-10¹⁵ :

Tutorial Speaker Reimbursement

Each tutorial speaker receives a grant of \$500.00 to assist with his or her travel to the conference. A maximum of two speakers per tutorial (or \$1000 per tutorial) will be awarded travel grants. More information regarding travel funds is included in the attached speaker agreement.

Comme on peut le constater, ces textes réglementaires décrivent un aspect métier de l'entreprise ou d'une organisation. Il s'agit de textes où l'entité en question définit des règles de gestion internes (pour ses agents) ou externes (pour ses clients ou son public) sous forme de contraintes, droits, devoirs, juridictions, autorités, etc.

Pour l'essentiel, la terminologie est spécialisée et employée de manière précise, même si on rencontre des cas d'ambiguïté ou des erreurs dans l'emploi de certains mots. Il est évidemment important, pour la compréhensibilité des règles, qu'elles soient exprimées dans des termes aussi proches que possible de cette terminologie, d'où l'importance d'une ontologie qui fixe le vocabulaire conceptuel à utiliser.

Les règles peuvent être exprimées sous des formes très diverses. Même quand elles sont exprimées de manière concise (en quelques phrases comme dans les exemples ci-dessus) elles font référence à des termes qui sont définis par ailleurs et peuvent être en dépendance avec d'autres règles. On constate qu'en dépit des efforts de précision dans la rédaction, il reste des implicites dans ces textes. La phrase qui fixe la somme globale allouée pour un tutorial (*\$1000 per tutorial*) semble contredire la précédente, une apparente incohérence qui ne pose pas de problème au lecteur humain, lequel comprend le sens général et rétablit la bonne interprétation sans même en avoir conscience, mais qui en poserait davantage à un moteur d'inférence si les deux règles étaient formalisées de manière indépendante.

14. *European Community Grant Agreement*, article 6 page 9.

15. *AAAI-10 Tutorial Speaker's Agreement*

4.2. Structure de l'index

Le modèle d'index présenté dans la section 3.1 peut être mis en œuvre pour la gestion des règles métiers. Il est appliqué ici au système de gestion de points de fidélité destinés aux voyageurs utilisant régulièrement les services de la compagnie American Airline. La figure 1 montre un fragment de ce texte différent de celui cité ci-avant où sont sélectionnées quelques règles métiers spécifiant la période de validité des points de fidélité.

subscribers receive their summary and information every month, regardless of activity level. Your summary includes flight and participant mileage earned, along with AAdvantage program information and special promotions.

1.

AAdvantage members must have mileage earning or redeeming activity once every 18 months in order to retain their miles. Miles will first expire under this 18-month policy on December 15, 2007. If your account has no qualifying activity in any 18-month period, all miles in the account will expire except for those miles earned prior to July 1, 1989 in accounts established prior to January 1, 1989 whose mileage credit will not expire. **Each qualifying activity extends the expiration date of all unexpired mileage credit in your account for 18 months from the date of the qualifying activity.** Qualifying activity is defined as redeeming any AAdvantage award or accruing mileage credit on any eligible American, American Eagle®, AmericanConnection® or AAdvantage airline participant as well as accruing mileage credit with participating hotels, car rental companies, credit cards, telecommunication providers, and other service providers offering AAdvantage mileage credit.

Figure 1. Fragment de texte du corpus AAdvantage

4.2.1. Prise en compte des règles

L'indexation des règles soulève une double difficulté : une règle possède une sémantique complexe qui n'est pas le plus souvent représentable par une ontologie et elle concerne des fragments de texte plus larges que ceux reliés aux éléments de l'ontologie.

En suivant le standard SBVR (*Standard Business Vocabulary and Rules*¹⁶) destiné à la représentation des vocabulaires et règles métiers des organisations, nous distinguons trois types de règles.

– Les règles opératoires aboutissent à une prise de décision. Elles sont généralement déclenchées par un événement particulier et conduisent à l'exécution d'une action (respectivement décrits dans les parties condition et conclusion de la règle). Nous les représentons de manière informelle comme suit :

Quand condition vraie Necessaire action ...

16. www.omg.org/spec/SBVR/1.0/

	Modèle de document	Modèle de correspondance	Modèle sémantique
Gestion des textes réglementaires	Document structuré	Simple liens	Ontologie et base de règles

Tableau 2. Index pour la gestion des règles métiers

Exemple : « Un membre du programme AAdvantage qui voyage sur un segment de vol éligible doit voir son compte crédité du nombre de points correspondant à la base kilométrique de ce segment de vol. »

– Les règles structurelles décrivent des propriétés qui restent valides dans tous les états de l'application. Elles peuvent être représentées sous la forme :

Si condition vraie Alors conclusion vraie

Exemple : « Un segment de vol a une base kilométrique unique. »

– Les règles de dérivation indiquent des procédures de calcul ou d'inférence qui permettent d'ajouter de la connaissance à celle acquise.

Ces représentations et ces types de règles font partie intégrante du modèle sémantique que nous avons proposé, même si celui-ci est encore susceptible d'évolution, notamment en ce qui concerne les types de règles et les relations entre règles (subordination, exception, etc.).

4.2.2. Schéma d'ensemble

La structure d'index proposée pour les systèmes de gestion des règles métiers repose donc sur le modèle présenté en section 3.1 mais avec une structure sémantique enrichie par rapport au modèle ontologique (tableau 2).

Cette structure sémantique est composée d'une ontologie et d'une base de règles comme présenté dans la figure 2. Nous distinguons, dans la base de règles, deux sous-ensembles de règles qui correspondent respectivement aux règles structurelles et aux règles opératoires. Toutes ces règles sont considérées comme des unités sémantiques au même titre que les concepts, rôles ou instances de l'ontologie : elles peuvent être mises en correspondance avec une unité documentaire. Comme les règles sont elles-mêmes exprimées dans les termes de l'ontologie, on dispose en réalité de liens d'indexation supplémentaires, non pas entre le texte et la structure sémantique mais internes à cette dernière. Nous en verrons l'utilité dans la section 5.

4.2.3. Implémentation

Pour implémenter cet espace de navigation autour des textes, de l'ontologie et des règles, il est important que le format de représentation de la structure d'index soit uniforme et compatible avec celui des ressources dont nous disposons *a priori*. Les textes réglementaires sont en général sous forme brute (.doc, .pdf, etc.), d'où la

Figure 2. Structure d'index proposée pour la gestion des textes réglementaires. Le document est structuré et est indexé par rapport à un modèle sémantique complexe composé d'une ontologie et de bases de règles interreliées

nécessité de les avoir dans un format pouvant supporter le modèle documentaire prévu *i.e.* un texte vu comme l'arborescence de ses unités textuelles (sections, sous-sections, paragraphes... phrases). L'ontologie est en OWL (*Ontologie Web Language* (Hendler *et al.*, 2004)), ce qui facilite l'accès direct aux différents objets conceptuels et permet une grande portabilité sur les outils existants. Les règles métiers telles que nous les extrayons des textes doivent être traduites en règles formelles et opératoires. la plupart des langages de règles ont une représentation en RIF (Rule Interchange Format). OWL et RIF sont des standards W3C du web sémantique qui reposent sur des langages dérivés de XML. Nous utilisons cette même famille de standards pour représenter le texte, les règles métiers et l'ensemble des liens du modèle de correspondance.

Le texte est représenté dans un format XML calqué sur la structure du document initial, comme illustré ci-après. Les balises marquent l'ensemble des unités textuelles allant du document à la phrase. Celle-ci est le plus bas niveau représenté dans l'arborescence car la segmentation en mots reste implicite. Considérons le premier exemple de la section 4.1.2 tiré du document AAdvantage. La troisième phrase de l'exemple est la 25e du texte¹⁷.

```
<-- Fragment du modèle de document -->
<Document rdf:ID = "#AAdvantage">
  <Section rdf:ID = "#Sect1">
```

17. Mileage credit cannot be earned for the same flight in more than one of the following programs : the AAdvantage program or any other loyalty program in which American Airlines participates. (On ne peut gagner des crédits kilométriques pour le même vol dans plusieurs des programmes suivants : le programme AAdvantage ou tout autre programme de fidélité auquel American Airline participe.)

```

<Paragraphe rdf:ID="#Para6">
  ...
  <Sentence rdf:ID="#S25">
    <content>
      Mileage credit cannot be earned for the same flight...
    </content>
    <start_offset>3725</start_offset>
  </Sentence>
</Paragraphe>
</Section>
</Document>

```

Les deux exemples suivants présentent respectivement un fragment de l'ontologie (la définition en OWL du concept **Mileage_Credit**) et la représentation en XML de la règle *R3* décrite par la phrase étiquetée *S25* dans le document. La règle fait référence au concept **AAProgram** qui regroupe tous les programmes de fidélité auxquels American Airlines participe. Comme en RIF l'étiquette <Frame> signale un prédicat avec variables qui peut faire l'objet de traitements particuliers.

```

<-- Fragment de l'ontologie (modèle sémantique) -->
<owl:Class rdf:ID="Mileage_Credit">
  <rdfs:label xml:lang="en">Mileage Credit</rdfs:label>
</owl:Class>

<--Exemple de règle (modèle sémantique) -->
<Rule id="R3">
  <Definition>
    <typedef name="AAOtherProgram">
      <derivedFromType name="AAProgram"/>
      <excludeValue name="AAAdvantage"/>
    </typedef>
  </Definition>
  <Declare>
    <var name="F" type="Flight"/>
    <var name="M" type="AAMember"/>
    <var name="P" type="AAOtherProgram"/>
  </Declare>
  <When><Frame> granted_in(?F,?P,?M) </Frame></When>
  <do>
    <Not>
      <Frame> granted_in(?F,#AAAdvantage,?M) </Frame>
    </Not>
  </do>
</Rule>

```

Les liens entre le texte et les ressources sémantiques sont décrits sous forme de triplets RDF (*Resource Description Framework* (Ora et al., 1999)). L'extrait ci-après

présente deux exemples de liens. Le premier associe la chaîne de caractères située entre les positions 0 et 14 de la phrase *S25* et le concept *Mileage_Credit*. La seconde associe toute la phrase *S25* à la règle *R3*.

```
<!-- Exemple de liens (modèle de correspondance) -->
<!--Correspondance texte-ontologie-->
<TextLink rdf:ID="#link1">
  <contentInSentence rdf:resource="#S25"/>
  <start_offset>0</start_offset>
  <end_offset>14</end_offset>
  <isSourceConcept rdf:resource="#Mileage_Credit"/>
</TextLink>

<!-- Correspondance texte-règle -->
<TextLink rdf:ID="#link2">
  <contentInSentence rdf:resource="#S25"/>
  <start_offset>0</start_offset>
  <isSourceRule rdf:resource="#R3"/>
</TextLink>
```

Pour manipuler l'index au niveau sémantique et exploiter la richesse de sa structure, nous avons conçu une interface de navigation dans cette structure. Cet espace de navigation se décompose en trois zones principales comme le montre la figure 3. Le texte est affiché dans une fenêtre centrale (zone 2), phrase par phrase dans cette version préliminaire. Le modèle sémantique est présenté en deux parties : la zone 1 à gauche donne accès à l'ontologie et les règles sont visibles dans la zone 3 à droite. Ce modèle est unifié dans la mesure où les règles sont liées aux éléments de l'ontologie. Les occurrences du concept sélectionné dans l'ontologie sont automatiquement marquées en couleur (rouge à l'écran, plus foncé dans la figure) dans le texte de même que l'unité documentaire associée à une règle sélectionnée (en vert à l'écran, plus claire ici).

4.3. Exemple du crédit kilométrique

Revenons plus en détail sur le texte d'American Airlines présenté en section 4.1.2. Trois concepts décrivent le parcours effectué par le voyageur. Un vol (Flight) représente un trajet entre deux aéroports. Le concept se subdivise en vol avec correspondance (Connecting_Flight) où l'on utilise plusieurs avions, et vol sans correspondance (Single-Plane_Flight) qui peut comporter des escales, mais pas de changement d'avion. Dans le cas où le trajet total utilise plusieurs avions, il se décompose en Segments parcourus dans le même avion, depuis le point où l'on y monte jusqu'à celui où on le quitte. Tout vol, qu'il s'agisse d'un vol direct, sans changement d'avion, d'un segment de vol ou d'un trajet composé de plusieurs segments, est associé à deux rôles : son point de départ et sa destination qui sont tous les deux remplis par des aéroports. Le concept de distance en vol direct n'est pas un concept physique : il s'agit

Figure 3. Visualisation de l'espace de navigation

d'une distance conventionnelle entre deux aéroports, détachée du trajet réel effectué, renvoyant souvent à une réglementation interne. Il est désigné dans l'ontologie par Actual_Base_Miles (base kilométrique effective), terme plus souvent employé dans d'autres parties du texte. Chaque segment entre dans un rôle has_base_mileage qui n'est pas nommé dans le texte mais qui lie un segment à sa base kilométrique effective.

La première phrase du texte n'est pas une règle¹⁸ : elle donne une information, mais n'est pas assez précise pour déterminer exactement des actions ou des contraintes. La dernière est une règle de non cumul qui a un caractère opératoire¹⁹ : elle impose de vérifier les programmes de fidélité auxquels le bénéficiaire participe. La seconde phrase²⁰ comme la troisième²¹ indiquent des règles structurelles. Le mode de calcul est clair et suffisamment déterminé mais il apparaît dans une autre partie du texte que les kilomètres ne sont portés au crédit de l'utilisateur qu'à sa demande et à condition qu'il fournisse des documents (billet ou autres). En fait, les règles doivent

18. *AAdvantage flight mileage credit is determined on the basis of nonstop distances between the airports where your flight originates and terminates.* (Dans AAdvantage, les kilomètres de vol crédités sont déterminés sur la base des distances en vol direct entre les aéroports où votre vol commence et se termine.)

19. *Mileage credit cannot be earned for the same flight in more than one of the following programs : the AAdvantage program or any other loyalty program in which American Airlines participates* (On ne peut gagner des crédits kilométriques pour le même vol dans plusieurs des programmes suivants : le programme AAdvantage ou tout autre programme de fidélité auquel American Airlines participe.)

20. *On connecting flights, you'll receive mileage credit for each segment of your trip.* (Pour les vols avec correspondance, vous recevez un crédit kilométrique pour chaque segment de votre trajet.)

21. *On single-plane flights, you'll receive the nonstop origin-destination mileage.* (Pour les vols sans changement d'avion, vous recevez le kilométrage du vol direct de l'origine à la destination.)

être enchaînées : si la première tire les conséquences de la décomposition d'un vol en segments, seule la seconde précise le crédit attaché à chaque segment, le même que si l'on prenait un vol direct, donc sa base kilométrique effective.

La qualification des règles dans l'annotation reflète un point de vue, ici celui de l'entreprise : pour le client, le non cumul est simplement une règle structurelle. On peut noter que la dernière phrase de l'exemple renvoie à une autre règle structurelle implicite dans le reste de la citation : seul un membre du programme de fidélité peut se voir attribuer des crédits kilométriques. Les conditions de cette attribution sont précisées à plusieurs endroits. On pourrait facilement commettre une autre erreur : il faut rapprocher ce passage d'autres précisant le mode de calcul du crédit kilométrique en fonction de la classe de service et du statut du membre du programme considéré (or, élite, etc.). Nous utilisons ici une version simplifiée qui se contente d'attribuer à un membre un coefficient multiplicateur et qui représente le droit d'un client à bénéficier d'un programme pour un voyage par un prédicat *granted-in()*.

Les règles se traduisent par :

- R1** Si *connecting-flight(F)* et *AAmember(M)* et *has-segment(F, S)* et *granted-in(F, AAdvantage, M)* Alors *granted-in(S, AAdvantage, M)*
- R2** Si *single-plane-flight(F)* et *AAmember(M)* et *granted-in(F, AAdvantage, M)* et *has-miles(F, Abm)* et *has-coef(M, r)* Alors *earns(M, Abm × r)*
- R3** Quand *flight(F)* et *AAmember(M)* et *AAprogram(P)* et $P \neq \text{AAdvantage}$ et *granted-in(F, P, M)* Necessaire $\neg \text{granted-in}(F, \text{AAdvantage}, M)$

L'indexation de ce fragment par l'ontologie et les règles est représentée dans la figure 4 qui montre la richesse de l'index.

4.4. Vue d'ensemble de l'index

Après l'analyse détaillée qui précède, cette section vise à donner une vue d'ensemble de l'index construit. Sans entrer dans le détail – car ce n'est pas l'objet de cet article –, nous expliquons rapidement le processus de construction de l'index puis nous décrivons le résultat obtenu.

Dans le cas particulier d'American Airline, la construction de l'ontologie a été guidée par TERMINAE (Aussenac-Gilles *et al.*, 2008) qui permet à un ingénieur de la connaissance de construire une ontologie de domaine à partir d'un corpus d'acquisition. Le module d'annotation exploite les connaissances lexicales obtenues lors de la construction de l'ontologie pour annoter de nouveaux textes au regard de cette ontologie. Même si son développement n'est pas achevé, il permet déjà d'avoir une bonne couverture d'annotation pour tous les éléments de type concepts et instances de concepts (elle est plus faible pour les rôles). L'éditeur de règles permet de formaliser les règles à partir des fragments de textes réglementaires. Il est en cours de test. Cet

Figure 4. Extrait de l'index pour la phrase *On connecting flights, you'll receive mileage credit for each segment of your trip* (Pour les vols avec correspondance, vous recevez un crédit kilométrique pour chaque segment de votre trajet)

effort de modélisation et de développement a permis d'obtenir, dans ce cas d'usage, une ontologie et une base de règles réalistes ainsi qu'une bonne annotation du corpus réglementaire pour ce qui touche aux concepts, instances de concepts et règles. Un premier prototype de moteur de recherche permet de tester les premières fonctionnalités de navigation sur l'index ainsi créé. Ce sont ces expériences dont nous faisons état ici.

Le texte lui-même représente 11 pages et nous adoptons une structuration simple en phrases.

A partir des 973 candidats termes extraits du texte, une ontologie a été construite : elle comporte 137 concepts du domaine, auxquels s'ajoutent 50 concepts plus généraux extraits de LKIF²². 10 classes principales regroupent les différentes catégories de concepts en jeu. Cette ontologie comporte également une cinquantaine d'instances de concepts qui ont été créés à partir des termes ou des entités nommées extraites du texte, même si le texte est principalement une description générique des règles d'attribution de bonus. La construction de l'ontologie à partir de textes conduit également à

22. www.estrellaproject.org/lkif-core/

Règles structurelles	67
Règles opératoires	28
Règles de dérivation	12
Règles de politique	36
Politiques métier	56

Tableau 3. Répartition des règles

associer des termes à des concepts de l'ontologie, ce qui permet ensuite une annotation normalisée des textes. Par exemple, les termes *miles*, *mileage*, *mileage credit*, *actual miles*, *base miles* renvoient en l'espèce au même concept Mileage, de même que *ticket*, *flight* et *fare*, *codeshare flight* et *codeshare service*²³ sont utilisés dans les règles de manière interchangeable pour désigner la place réservée et payée par le client sur un vol donné.

A ce stade, l'ontologie est moins riche en rôles, l'accent ayant d'abord été mis sur les concepts du domaine. L'analyse en cours montre cependant l'importance des rôles dans le domaine (ex. [Passenger] earns [Benefit], un passager bénéficie d'un avantage, ou [Passenger] books [Flight], un passager réserve une place sur un vol). Le repérage et la formalisation de ces rôles sont importants pour guider l'interprétation du texte : l'argument de type Passenger n'est explicité dans le texte que pour la moitié des occurrences du premier rôle ci-dessus et jamais pour le second. Dans certains cas, l'ontologie permet de mettre en évidence des phénomènes de métonymie et d'éviter les erreurs d'interprétation, fréquentes pour qui ne connaît pas très bien le domaine. Dans « rail services earn miles » (litt. *les services ferroviaires obtiennent un crédit kilométrique*), il faut ainsi comprendre qu'une personne obtient un crédit kilométrique quand elle utilise des services ferroviaires.

La création des règles se fait à l'aide d'un éditeur de règles qui permet de reformuler progressivement les fragments du texte initial exprimant des règles en règles formelles. Ce travail de reformulation s'accompagne souvent d'étapes de décomposition ou de regroupement de règles. Nous avons identifié 199 règles dans le corpus qui se répartissent dans les 5 catégories du standard SBVR (voir tableau 3) : des règles structurelles, opératoires et de dérivation, déjà mentionnées à la section 4.2.1, auxquelles s'ajoutent des politiques, trop imprécises pour être opératoires²⁴ et des politiques métier très générales²⁵. La couverture de l'annotation du corpus est bonne pour les parties riches en règles. Si l'on considère les fragments de textes d'où les 199 règles identi-

23. *Codeshare* désigne un mode d'organisation où deux compagnies vendent des places sur le même vol.

24. *The member must adhere to the rules and limitations of the shareAAmiles program* (Les membres doivent respecter les règles et contraintes du programme shareAAmiles.)

25. *AAdvantage mileage accrual eligibility on airline participant routes is subject to change without notice* (Les règles d'éligibilité des crédits kilométriques obtenus sur les vols des compagnies partenaires sont susceptibles d'évoluer sans préavis.)

fiées sont dérivées, on note qu'ils comportent en moyenne 4 occurrences de concepts de l'ontologie. Le minimum est de 2 occurrences, ce qui signifie que toutes les règles sont reliées dans l'espace de navigation.

5. L'index comme espace de navigation

L'index se présentant comme une structure d'accès aux ressources indexées, la structure d'index proposée constitue un espace de navigation entre ressources documentaires et sémantiques à travers le modèle de correspondance. Selon la définition de ce dernier, le parcours entre les ressources n'est pas orienté et peut alors s'effectuer dans les deux sens. Nous présentons dans cette section les fonctionnalités de navigation qu'un tel index supporte puis nous montrons comment elles peuvent être utilisées pour la gestion des règles métiers en reprenant l'exemple précédent du texte d'American Airlines. La section 6 montrera ensuite comment cette approche peut être généralisée à d'autres types d'applications.

5.1. Fonctionnalités générales de navigation

L'index que nous avons présenté en section 3 ouvre en réalité trois espaces de navigation inter-reliés :

1) la navigation intra-documentaire permet de passer d'une unité documentaire à une autre suivant des parcours hiérarchiques ou linéaires, par des recherches de motifs, voire, quand ils existent, en suivant des liens hypertextuels ;

2) la navigation sémantique est très dépendante de la structure sémantique : les éditeurs d'ontologies²⁶ permettent par exemple de parcourir les ontologies et de répondre à des requêtes sur des structures ontologiques. La particularité du modèle sémantique présenté dans la section précédente vient de ce qu'il articule une ontologie et une base de règles, permettant ainsi, à partir d'une règle d'accéder aux concepts auxquels elle fait référence et inversement de trouver les règles qui font intervenir un concept donné ;

3) la navigation sémantico-documentaire repose sur l'index à proprement parler, permettant de passer du document à la structure sémantique et inversement.

Nous mettons ici l'accent sur cette troisième forme de navigation qui repose directement sur l'index, même s'il est clair que ses potentialités sont décuplées du fait qu'elle est reliée aux autres et que l'on peut ainsi enchaîner des parcours dans les trois espaces de navigation.

26. Protégé (protege.stanford.edu/) ou NeonToolkit (neon-toolkit.org/wiki/Main_Page), par exemple.

La navigation dans l'index repose sur plusieurs opérations :

- le parcours de lien permet de retrouver les éléments de l'ontologie ou les règles qui annotent un fragment textuel donné ou, inversement, les passages textuels indexés par un élément de l'ontologie ou une règle ;
- au-delà du parcours à proprement parler, l'interrogation de l'index permet de retrouver toutes les unités sémantiques liées à une unité documentaire ou toutes les unités documentaires rattachées à une unité sémantique. Cela ouvre la voie à des calculs de pertinence, à des analyses d'ambiguïté, des mesures de couverture, etc. ;
- à partir de là, on peut mesurer la similarité des unités documentaires (resp. des éléments sémantiques) en comparant les liens qui les relient aux éléments sémantiques (resp. aux unités documentaires). De même qu'on mesure classiquement la similarité de deux mots en comparant les vecteurs formés par les unités sémantiques qui sont associées à chacun, on peut comparer deux règles sur la base des vecteurs d'unités documentaires auxquelles elles sont attachées. À ce stade, pour une règle *R* donnée, nous nous contentons d'ordonner toutes les règles voisines en fonction du nombre d'unités sémantiques qu'elles partagent avec *R* mais cette approche est réductrice. Il faudrait tenir compte d'autres facteurs comme la diversité des annotations, par exemple. D'autres mesures de similarité sont à l'étude ;
- ces opérations peuvent ensuite être généralisées à l'index dans son ensemble, ce qui permet de mesurer le degré de couverture de l'ontologie et de la base de règles dans les textes, d'identifier les zones documentaires qui sont mal indexées et les unités sémantiques qui n'ont pas de contrepartie documentaire, ou au contraire de repérer les unités documentaires les mieux indexées et les unités sémantiques les mieux ancrées dans le texte.

5.2. *Exploitation de l'index pour la gestion des règles métiers*

L'association de fonctionnalités élémentaires, *i.e.* le parcours successif de plusieurs liens, montre comment le texte et les annotations sémantiques s'éclairent l'un l'autre.

5.2.1. *Acquisition des règles métiers*

Écrire les règles métiers qui modélisent la réglementation en vigueur dans un domaine d'activité donné (par ex. les règles d'attributions des points de fidélités aux voyageurs d'une compagnie aérienne) est une tâche longue et difficile dont il est délicat de garantir la qualité. Le problème est d'autant plus critique que les réglementations sont en réalité amenées à changer assez régulièrement.

L'index tel que nous l'avons conçu doit permettre de guider ce travail de modélisation. Si le texte est déjà indexé au regard d'une ontologie, le fait de sélectionner dans le texte un fragment qui exprime une règle donne directement accès aux éléments sémantiques qui sont mentionnés dans ce fragment et qui ont des chances d'apparaître aussi dans la/les règles qui seront modélisées à partir de ce fragment.

Supposons en effet que la personne chargée de l'écriture des règles sélectionne la phrase *On connecting flights, you'll receive mileage credit for each segment of your trip*²⁷ et cherche à modéliser la règle qu'elle exprime pour obtenir la règle *R1* mentionnée ci-dessus. Si la phrase est déjà annotée comme dans la figure 5, tous les concepts sont déjà explicités (y compris quand les mots ne correspondent pas directement au label du concept comme dans *trip* qui est annoté *Flight*) et la mention d'un rôle devrait elle-même être identifiée, ce qui facilite grandement l'écriture de la règle : le travail peut se concentrer sur la sémantique de la règle elle-même plutôt que sur les éléments de vocabulaire à utiliser. L'environnement d'analyse des règles que nous construisons tire profit de la connexion avec l'ontologie pour proposer une étape de normalisation du vocabulaire : chaque terme annoté est, sous le contrôle de l'expert, remplacé par le terme vedette associé à ce concept.

Figure 5. Exemple de fragment annoté au regard de l'ontologie

5.2.2. Traçabilité du modèle sémantique

Un système de gestion de règles métiers prend des décisions par application de sa base de règles et il est souvent important de rendre compte des résultats, des inférences qui sont faites ou des décisions qui sont prises.

La première étape est généralement facile. Elle consiste à indiquer les règles qui ont été utilisées dans le processus inférentiel ayant conduit à la décision. Ce premier niveau d'explication peut être précieux pour les utilisateurs si les règles sont écrites à un niveau suffisamment conceptuel pour être compréhensibles. L'étape suivante consiste à expliquer les règles elles-mêmes et c'est là que la traçabilité assurée par l'index est précieuse : on peut remonter des règles aux fragments de textes dont elles sont issues même si ces fragments de texte sont incomplets ou ambigus.

Cette traçabilité est en fait déjà importante dans la mise au point des règles, notamment quand des problèmes d'incohérence apparaissent dans la base de règles. C'est

²⁷. Pour les vols avec correspondance, vous recevez un crédit kilométrique pour chaque segment de votre trajet.

le cas par exemple si plusieurs règles aboutissent directement ou indirectement à des conclusions opposées à partir des mêmes conditions initiales. Il faut alors corriger les règles pour maintenir le système en état de fonctionnement et le fait de pouvoir revenir au texte permet de confronter les fragments réglementaires dont les règles sont issues. On peut ainsi déterminer si l'incohérence a été introduite lors de la modélisation, suite à une interprétation erronée ou partielle d'un fragment textuel, ou si elle provient du texte qui est lui-même incohérent. L'index ne permet pas de lever l'incohérence mais de la circonscrire, ce qui est un préalable à sa résolution. Un simple inventaire des règles dispersées dans le texte et gouvernant le calcul des avantages en crédit kilométrique le montre bien²⁸ :

- *On single-plane flights, you'll receive the nonstop origin-destination mileage.*
- *If you are an Elite member, you will earn a minimum of 500 miles on applicable routes.*
- *Certain airline tickets are not eligible for mileage credit.*
- *You'll receive AAdvantage mileage credit only for the class of service on which your fare is based when you are ticketed.*
- *As an elite-status member you earn an elite status mileage bonus on the base or guaranteed minimum miles for each eligible flight.*

Comme souvent dans les langues naturelles, le texte pose une règle apparemment universelle – mais qui correspond en fait seulement au cas général – puis apporte des précisions sous forme d'exceptions (seconde et troisième phrases de la liste ci-avant) et de raffinements (les dernières phrases décrivent deux étapes de calcul, selon la classe et le statut). Il suffit qu'une de ces règles échappe au regroupement et soit traitée isolément pour aboutir à une contradiction. Un système de traçage remontant jusqu'au texte permet de reconstituer le groupe, puis l'interprétation correcte.

5.2.3. Analyse d'obsolescence

Une autre difficulté classique, à laquelle sont confrontés les systèmes de gestion des règles métiers et plus généralement les systèmes à base de connaissances, est liée

28. Le texte est un peu simplifié pour permettre au lecteur d'accéder directement à l'essentiel des contradictions latentes.

- *Pour les vols sans changement d'avion, vous recevrez le kilométrage du vol direct de l'origine à la destination.*
- *Si vous êtes un membre d'élite, vous gagnerez un crédit kilométrique de 500 miles minimum sur les vols éligibles.*
- *Certains tickets d'avion ne permettent pas d'obtenir de crédit kilométrique.*
- *Vous recevrez un crédit kilométrique de AAdvantage correspondant la classe pour laquelle le ticket a été émis.*
- *En tant que membre élite, vous obtenez un bonus kilométrique calculé en fonction du nombre de kilomètre de base ou garanti de chacun des vols éligibles que vous faites.*

à l'évolution de la réglementation qui fait que certains éléments de connaissances perdent de leur utilité quand la réglementation et la base de règles évoluent. C'est le cas du fragment de texte qui suit :

*If your account has no qualifying activity in any 18-month period, all miles in the account will expire except for those miles earned prior to July 1, 1989 in accounts established prior to January 1, 1989 whose mileage credit will not expire.*²⁹

qui se traduit par deux règles : l'une traite des comptes créés avant le 1er janvier 1989 et comportant du crédit acquis avant le 1er juillet de la même année, la seconde des autres comptes. Comme le crédit acquis ne se renouvelle pas mais tend au contraire à être consommé, la première règle va devenir obsolète à un moment donné et on pourra alors expurger le texte d'une exception devenue sans objet.

Là encore l'index a un rôle à jouer. S'il est facile de tenir un journal de l'utilisation des règles, et donc d'identifier celles qui cessent d'être utiles quand la base de règles ou ses conditions d'emploi évoluent, il est moins facile de repérer les parties de l'ontologie et les zones de textes qui deviennent obsolètes, à moins d'avoir un index permettant de remonter des règles obsolètes aux unités sémantiques et documentaires auxquelles elles sont liées et de vérifier en retour que ces unités ne sont pas reliées à d'autres règles.

Dans l'exemple des règles d'attribution de bonus, l'analyse d'obsolescence est un processus complexe de propagation récursive qui remonte des règles à l'ontologie puis au texte :

- étant donné un ensemble de règles $\mathcal{R}_o$ marquées comme obsolètes et l'ensemble des unités sémantiques de type ontologique $\{uo_1, uo_2 \dots uo_n\}$ liées à ces règles, on considère qu'une unité uo_i est obsolète si elle n'est liée qu'à des règles de $\mathcal{R}_o$. L'obsolescence se propage ; autrement dit, si uo_i entre dans la définition d'une autre unité ontologique ux_j , celle-ci risque de devenir elle-même obsolète ;

- étant donné un ensemble de règles $\mathcal{R}_o$ et un ensemble d'unités ontologiques $\mathcal{UO}_o$ obsolètes ainsi que l'ensemble des unités documentaires $\{ud_1, ud_2 \dots ud_n\}$ liées à ces règles et à ces unités ontologiques, on considère qu'une unité ud_i est obsolète si elle n'est liée qu'à des unités sémantiques de $\mathcal{R}_o \cup \mathcal{UO}_o$. Une unité documentaire qui contient une unité obsolète peut être elle-même obsolète ou nécessiter une mise à jour.

5.2.4. Mise à jour de la réglementation

Les textes de réglementation évoluant, il faut mettre à jour l'index. Les deux opérations élémentaires sont l'ajout et le retrait d'un texte, une modification étant un retrait suivi d'un ajout. Le texte ajouté, retiré ou modifié peut être un document entier ou un

29. Si votre compte ne montre aucun signe d'activité sur une période de 18 mois, tout le crédit kilométrique expire, à l'exception du crédit gagné avant le 1er juillet 1989 sur des comptes ouverts avant le 1er janvier 1989, lequel n'expire pas.

fragment de document, puisque, en pratique, l'évolution du document se détaille au niveau du paragraphe ou même de la phrase.

Le retrait d'un texte nécessite une analyse d'impact qui est similaire à l'analyse d'obsolescence mais avec le texte comme point de départ. On retire, avec le texte, toutes les unités sémantiques (unités ontologiques ou règles) qui ne sont liées qu'à ce texte-là.

L'ajout d'un texte nécessite une nouvelle phase locale d'annotation. Dans les cas simples, cette annotation n'a pas d'impact sur la structure sémantique, mais dans les cas plus complexes, l'ajout d'un texte amène à mettre à jour la structure sémantique elle-même. Il peut en effet être nécessaire, pour rendre compte du contenu du nouveau texte, d'introduire de nouveaux éléments dans l'ontologie ou de la réorganiser partiellement, de modifier des règles existantes ou d'en ajouter de nouvelles. Sans entrer dans le détail de ces processus complexes d'annotation et de mise à jour incrémentale de la structure sémantique, on peut souligner que l'index a un rôle particulier à jouer quand le texte ajouté vient en remplacer un autre assez proche dans la formulation : l'annotation du nouveau texte peut s'appuyer sur la trace des annotations de l'ancien.

6. Vers d'autres applications

Les sections qui précèdent ont montré la richesse de navigation que permet la structure d'index dans le cas de l'analyse des textes réglementaires mais la structure proposée est générale et d'autres applications peuvent être envisagées.

Il va de soi que la richesse de l'index et des fonctionnalités de navigation repose sur la qualité de l'annotation des documents ainsi que sur celle des modèles documentaire et sémantique considérés. L'approche envisagée pour l'analyse des textes réglementaire est coûteuse : elle suppose un travail de conceptualisation (construction d'une ontologie de domaine, édition des règles) qui ne peut être entrepris que pour les applications où cet effort est rentable et où la navigation sémantique apporte une réelle plus-value à une certaine communauté d'utilisateurs. C'est le cas de la gestion des règles métiers, domaine dont les enjeux économiques sont réels et où la fiabilité des informations est cruciale³⁰ mais d'autres champs d'applications émergent en parallèle.

6.1. *Edition savante et critique*

L'un des champs d'applications les plus connus est celui de l'édition savante et critique où l'on cherche à donner accès à l'œuvre d'un auteur ou d'un groupe d'au-

30. C'est la raison d'être du projet Ontorule coordonné par IBM/Ilog.

teurs³¹. On peut citer ici l'exemple d'HyperNietzsche³², projet d'édition de l'oeuvre de Nietzsche (d'Iorio, 2006), ou celui de l'édition électronique des manuscrits de Stendhal (Lebarbé *et al.*, 2009). Dans ces deux cas, l'index offre aux lecteurs la possibilité de parcours qui réorganisent le contenu documentaire selon des « critères génétiques, chronologiques ou thématiques » (d'Iorio, 2006), selon un axe critique où différents niveaux de granularité correspondent à différents niveaux de lecture et types d'utilisateurs (Lebarbé *et al.*, 2009). Il s'agit de s'affranchir de l'ordre arbitraire – en cela qu'il organise un parcours unique, préétabli – de l'édition « papier ».

La structure d'index que nous proposons s'applique à ce type d'application :

- le modèle documentaire est généralement plus complexe puisqu'il doit prendre en compte la dimension intertextuelle selon laquelle différents documents sont liés les uns aux autres par des liens chronologiques, génétiques ou de citation. Il doit aussi tenir compte de la structure propre des documents analysés qui peut être plus riche que celle des textes réglementaires considérés plus haut. Dans ce cas, le modèle documentaire repose généralement sur le standard de la TEI (*Text Encoding Initiative*³³). Un tel modèle documentaire multiplie les possibilités de navigations documentaires comme le montre (Lebarbé *et al.*, 2009) ;

- le modèle sémantique se réduit souvent à un ensemble de métadonnées documentaires mais il gagnerait à être explicité et formalisé en tant que tel pour ouvrir la voie à divers types de navigations sémantico-documentaires, notamment des lectures thématiques. Le degré de complexité et la richesse du modèle sémantique doivent être adaptés à l'application visée : simple liste de termes, thesaurus hiérarchisé, ontologie, « ontologie dynamique » (d'Iorio, 2006) reflétant l'évolution des objets d'étude dans le temps, dans l'espace ou selon les points de vue, voire des modèles plus riches encore ;

- la fonction de correspondance lie des unités documentaires de taille variable à des unités sémantiques diverses et peut leur associer différentes propriétés. Elle aussi doit être définie en fonction de l'application et des parcours visés.

L'index ainsi construit permet une « contextualisation dynamique » (d'Iorio, 2006) des unités documentaires non seulement dans un espace documentaire – ce que fait l'hypertexte – mais aussi dans un espace sémantique. Il faut cependant souligner que, si l'index permet des parcours « rhizomatiques » (Lebarbé *et al.*, 2009), les modes d'interprétation sous-jacents sont prédéfinis dans le modèle documentaire (lorsque les documents sont liés les uns aux autres par des liens sémantiques) et dans le modèle sémantique qui est associé aux documents : une lecture chronologique n'est possible que si des métadonnées temporelles sont associées aux documents et qu'un calcul peut être fait dessus ; les points de vue et les modalités ne peuvent être analysés dans les

31. Au sens où nous employons ces termes, l'édition savante fournit à la communauté de chercheurs les instruments de son travail, l'édition critique vise les variantes du texte et les conditions de son écriture, et les deux points de vue ont trop en commun pour qu'ils puissent être aisément distingués.

32. www.nietzschesource.org

33. www.tei-c.org/index.xml

textes réglementaires si le modèle ontologique et à base de règles ne permet pas de les représenter. Dans tous les cas, un index est conçu pour outiller des formes de lecture définies *a priori* et il ne saurait rendre compte de toute la diversité des interprétations qui peuvent être faite d'un texte donné.

6.2. Redocumentarisation du patrimoine numérisé

Des problématiques analogues à l'édition peuvent se trouver dans la sphère scientifique et technique. Le projet EnCORe (Colaix-Castillo *et al.*, 2005) de développement d'une encyclopédie numérique de la chimie organique peut être également relu dans ce cadre. Ce projet visait à procurer un accès unifié et intégré à un vaste ensemble de ressources spécialisées de chimie organique (les articles de recherche et revues, les bases de données de réactions chimiques, des banques d'images, des biographies de chimistes, etc.) pour faciliter la formation et la découverte scientifique dans le domaine de la chimie organique. On comprend que les unités documentaires sont plus hétérogènes encore que dans le cas précédent et elles sont liées les unes aux autres par de multiples liens de citation, d'auteur commun et d'illustration. Dans le cas de cette encyclopédie cependant, cette richesse dont le modèle documentaire doit rendre compte est telle que les risques de « désorientation » et de « surcharge cognitive » pour l'utilisateur sont majeurs (Conklin, 1987). La navigation purement documentaire, comme dans le cas d'internet, ne peut répondre aux besoins des utilisateurs : il faut un moteur de recherche sur une structure d'index et la qualité de l'encyclopédie dépend de la qualité du modèle sémantique sous-jacent. En l'occurrence, c'est un thesaurus et une ontologie qui étaient envisagés (Dragos *et al.*, 2009), même si le projet n'a pu voir le jour dans son ambition première.

La gestion et la publication de la documentation juridique soulève des problèmes similaires d'indexation de contenu et d'accès à l'information. Le projet Legilocal³⁴ vise à simplifier l'accès des citoyens aux données juridiques publiques locales. Il propose aux collectivités un bouquet de services qui leur permet de publier facilement leurs données sur leur site web, ce qui assure au citoyen un accès homogène aux données des différentes collectivités. Là encore, le modèle documentaire doit permettre de représenter en fait un réseau de relations sémantiques entre documents, les uns étant liés les uns aux autres par des liens de citation, de jurisprudence, de chronologie, etc. Dans ce cas de figure, le modèle sémantique envisagé présente une double originalité par rapport à ceux qui sont présentés ci-dessus : il articule d'abord une structure de thesaurus hiérarchisé préétabli par les experts du domaine et des annotations librement posées par les citoyens dans différents groupes de discussion ; il fait le lien entre différents niveaux de langue (langue juridique des agents de collectivités territoriales vs. langue commune).

34. Projet FUI coordonné par LexisNexis (2010-2012) et auquel nous participons.

À travers ces applications, c'est en réalité un enjeu de redocumentarisation³⁵ du patrimoine numérisé qui apparaît³⁶. Chaque application a ses besoins et ses contraintes propres mais le modèle général d'index que nous avons proposé permet de confronter et de rapprocher ces différents projets et/ou pratiques. Le développement de ces applications nécessite un effort important, similaire à celui que nous avons entrepris pour la gestion des règles métiers, car il faut dresser une typologie des utilisateurs et de leurs besoins, choisir des modèles documentaire et sémantique adaptés à chaque situation, les construire et développer des outils de navigation. Mais, c'est ainsi que des fonctionnalités de navigation adaptées à des usages spécialisés peuvent voir le jour.

7. Conclusion

Nous avons montré par une analyse détaillée des opérations d'annotation et d'indexation que, dans un contexte numérique, ces deux opérations deviennent complémentaires. Nous avons proposé un modèle d'index qui traduit cette complémentarité. Ce modèle est souple et englobe des activités documentaires aussi diverses que les index de fin de livres ou les folksonomies.

L'application de ce modèle à l'indexation des documents réglementaires nous permet de proposer et de mettre en œuvre une architecture adaptée à cette tâche spécifique et de valider, ce faisant, l'approche d'indexation proposée sur un cas d'usage concret. Le modèle sémantique est particulièrement riche puisqu'il comporte à la fois une ontologie et une base de règles. Une première implémentation de ce modèle a été réalisée. Elle rend possible le développement de fonctions de navigation élaborées qui contribuent à assurer la cohérence de l'ensemble des documents et des représentations sémantiques mais aussi à maintenir cette cohérence face à l'évolution naturelle de l'un et de l'autre.

Ces fonctions de navigation montrent comment le texte et le modèle sémantique s'éclairent mutuellement. Nous avons présenté ici les fonctions qui nous ont paru pertinentes pour le cas d'usage étudié mais l'index peut servir de support à d'autres formes de raisonnement. Nous pensons notamment à des processus de saturation de l'annotation, à l'évaluation du modèle sémantique en termes de couverture, à l'explicitation des éléments implicites dans le texte, à la mise à jour du modèle sémantique

35. « Pour définir la redocumentarisation, il faut commencer par s'entendre sur le terme 'documentarisation'. Documentariser, c'est ni plus ni moins 'traiter un document' comme le font, ou le faisaient, traditionnellement les professionnels de la documentation (bibliothécaires, archivistes, documentalistes) : le cataloguer, l'indexer, le résumer, le découper, éventuellement le renforcer, etc. [...]. L'objectif de la documentarisation est d'optimiser l'usage du document en permettant un meilleur accès à son contenu et une meilleure mise en contexte. »

« Le numérique est une opportunité formidable de redocumentarisation. [...] Il s'agit de traiter à nouveau des documents traditionnels qui ont été transposés sur un support numérique en utilisant les fonctionnalités de ce dernier. » (Salaün, 2010).

36. « patrimoine » est à entendre ici dans un sens large incluant la littérature scientifique et technique et débordant des seuls écrits.

au regard d'un flux de documents, à la détection des incohérences dans le texte, etc. Ces nouvelles formes de raisonnement restent à étudier, à formaliser et à tester sur des cas d'usage concrets pour en évaluer la pertinence.

Remerciements

Ce travail a été partiellement réalisé dans le cadre du projet ONTORULE (FP7 231875, <http://ontorule-project.eu/>). Nous remercions la compagnie American Airline de nous avoir autorisés à utiliser son corpus à des fins de recherche et de test dans le cadre de ce projet.

8. Bibliographie

- Amardeilh F., Web Sémantique et Informatique Linguistique : propositions méthodologiques et réalisation d'une plateforme logicielle, PhD thesis, Univ. Paris X, 2007.
- Ambriola V., Gervasi V., « On the Systematic Analysis of Natural Language Requirements with CIRCE », *Automated Software Engineering*, vol. 13, p. 107-167, 2006.
- Anick P., « The automatic construction of faceted terminological feedback for interactive document retrieval », in D. Bourigault, C. Jacquemin, M.-C. L'Homme (eds), *Recent Advances in Computational Terminology*, John Benjamins Publishing Company, Amsterdam/Philadelphie, p. 29-52, 2001.
- Aussenac-Gilles N., Despres S., Szulman S., « The Terminae Method and Platform for Ontology Engineering from Texts », in P. Buitelaar, P. Cimiano (eds), *Bridging the Gap between Text and Knowledge : Selected Contributions to Ontology learning from Text*, IOS Press, 2008.
- Baeza-Yates R., Ribeiro-Neto B. (eds), *Modern Information Retrieval*, Addison-Wesley Longman Publishing Company, 1999.
- Bontcheva K., Cunningham H., « The Semantic Web : A New Opportunity and Challenge for Human Language Technology », *Proceedings of Workshop on Human Language Technology for the Semantic Web and Web Services, 2nd International Semantic Web Conference, Sanibel Island*, p. 89-96, 2003.
- Carolyn Brodie Clare-Marie Karat J. K., « An Empirical Study of Natural Language Parsing of Privacy Policy Rules Using the SPARCLE Policy Workbench », *SOUPS 06*, p. 8-19, 2006.
- Chan L. M., « Social Bookmarking and Subject Indexing », *Satellite Pre-Conference, IFLA Classification and Indexing Section*, p. 1-16, August, 2009.
- Cimiano P., *Ontology Learning and Population from Text : Algorithms, Evaluation and Applications*, Springer, November, 2006.
- Colaax-Castillo C., Krief A., « EnCORe (Encyclopédie de chimie Organique Electronique) : an Original Way to Represent and Transfer Knowledge from Freshmen to Researchers in Organic Chemistry », *Proceeding of the 2005 conference on Towards the Learning Grid : Advances in Human Learning Services*, IOS Press, Amsterdam, The Netherlands, p. 231-237, 2005.
- Conklin J., « Hypertext : an introduction and survey », *Computer*, vol. 20, n° 9, p. 17-41, 1987.

- Dinesh N., Joshi A., Lee I., Sokolsky O., « Logic-based regulatory conformance checking », *14th Monterey Workshop*, Springer-Verlag, Berlin, Heidelberg, p. 147-160, 2007.
- Dinesh N., Joshi A., Lee I., Sokolsky O., « Reasoning about Conditions and Exceptions to Laws in Regulatory Conformance Checking », *DEON*, Springer-Verlag, Berlin, Heidelberg, p. 110-124, 2008.
- Dinesh N., Joshi A., Lee I., Webber B., « Extracting Formal Specifications from Natural Language Regulatory Documents », *ICoS-5*, Buxton, England, p. 1-10, 2006.
- d'Iorio P., « Nietzsche on New Paths : The HyperNietzsche Project and Open Scholarship on the Web », in M. C. Fornari (ed.), *Friedrich Nietzsche*, Edizioni e interpretazioni, p. 475-496, 2006.
- Dragos V., Nazarenko A., « Towards a semantic model to enhance knowledge sharing and discovery in organic chemistry », *Proceedings of the IADIS International Conference Information Systems (IS'09)*, Barcelona, Spain, February, 2009.
- Handschuh S., Staab S., « Deep Annotation for Information Integration », in S. Handschuh, S. Staab (eds), *Annotation for the Semantic Web*, IOS Press, Amsterdam, p. 25-45, 2003.
- Hearst M. A., *Modern Information Retrieval*, Addison-Wesley, Wokingham, UK, 1999.
- Hendler J., Horrocks I., al., « OWL web ontology language reference », *W3C Recommendation*, 2004.
- Kalyanpur A., Hendler J., Parsia B., Golbeck J., « SMORE - Semantic Markup, Ontology, and RDF Editor », <http://www.mindswap.org/papers/SMORE.pdf>, p. 1-5, 2003.
- Lebarbé T., Meynard C., « Nouvelles pratiques éditoriales, nouvelles lectures : les enjeux de l'édition électronique de manuscrits littéraires », *Mémoires du livre / Studies in Book Culture*, 2009.
- Mondary T., Despres S., Nazarenko A., Szulman S., « Construction d'ontologies à partir de textes : la phase de conceptualisation », *Actes des 18èmes journées francophones d'Ingénierie des Connaissances*, Nancy - France, p. 87-98, 2008.
- Mulvany N. C., *Indexing Books*, The University of Chicago Press, 1993.
- Nazarenko A., Aït El Mekki T., « Building back-of-the-book indexes », in F. Ibekwe-SanJuan, A. Condamines, M. T. C. Castellvi (eds), *Application-Driven Terminology Engineering*, John Benjamins Publishing Company, Amsterdam, p. 179-202, 2007.
- Névéal A., Shooshan S. E., Mork J. G., Aronson A. R., « Fine-Grained Indexing of the Biomedical Literature : MeSH Subheading Attachment for a MEDLINE Indexing Tool », *Proceedings of the AMIA Symposium*, p. 553-557, 2007.
- OPAdoc, « Documentation Oracle OPA », <http://www.oracle.com/technology/products/applications/policy-automation/index.html>, n.d.
- Ora L., Swick R., Resource Description Framework (RDF) Model and Syntax Specification, Technical report, W3C consortium, 1999. Recommendation.
- Popov B., Kiryakov A., Manov D., Kirilov A., Ognyanoff D., Goranov M., « Towards Semantic Web Information Extraction », *Proceedings of the Human Language Technologies Workshop (ISWC'03)*, Sanibel, Floride, p. 1-22, 2003.
- Reeder R. W., Karat C.-M., Karat J., Brodie C., « Usability Challenges in Security and Privacy Policy-Authoring Interfaces », in C. B. et al. (ed.), *INTERACT 07*, vol. 4663 of LNCS, IFIP, Springer, p. 141-155, 2007.

- Salaün J.-M., « "Industries du copyright" et fair use », Blog "Industries du copyright" et fair use, 2010. Mardi 8 juin.
- Santiago V., Vijaykumar N. L., da Silva J. D. S., « Natural Language Requirements : Automating Model-Based Testing and Analysis of Defects », *IX Workshop do Curso de Computação Aplicada*, LIT/INPE, p. 1-10, oct, 2009.
- University of Chicago Press Staff (ed.), *The Chicago Manual of Style*, University of Chicago press, Chicago, USA, 2003.
- Uren V., Cimiano P., Iria J., Handschuh S., Vargas-Vera M., Motta E., Ciravegna F., « Semantic Annotation for Knowledge Management : Requirements and a Survey of the State of the Art », *Journal of Web Semantics*, vol. 4, p. 14-28, 2006.
- Wacholder N., Nevill-Manning C., « Workshop report : The Technology of Browsing Applications, Workshop held in conjunction with JCDL 2001 », *SIGIR Forum*, vol. 35, n° 1, p. 16-19, 2001. <http://www.acm.org/sigir/forum/S2001-TOC.html>.