Eléments de réflexion sur l'utilisation de corpus pour la construction d'ontologies

Xavier Aimé

Eléments de réflexion sur l'utilisation de corpus pour la construction d'ontologies

Xavier Aimé[1]

INSERM UMRS 1142, LIMICS, F-75006, Paris, France ;
Sorbonne Universités, UPMC Univ. Paris 06, UMR_S 1142, LIMICS, F-75006, Paris, France ;
Université Paris 13, Sorbonne Paris Cité, LIMICS, (UMR_S 1142), F-93430, Villetaneuse, France.
xavier.aime@inserm.fr

Résumé : L'Ingénierie des Connaissances (IC) est née d'une inter-disciplinarité alliant informatique, intelligence artificielle, psychologie cognitive, ergonomie et sciences de gestion. Aujourd'hui, il semblerait que l'IC soit plus orientée vers la dimension technologique, notamment de par les problèmes d'hétérogénéité et de quantité massive de données qu'elle doit gérer. Un petit retour aux sources, vers les Sciences Humaines et Sociales, ne serait pas préjudiciable. De nos jours, il est communément fait appel aux corpus de textes pour la construction d'ontologies de domaine, corpus essentiellement considérés comme des ensembles de termes plus ou moins structurés. Or un texte est avant tout un acte de communication, un outil de transmission d'un message de la part d'un émetteur doté d'une intention à destination d'un récepteur. Nous proposons dans cet article quelques éléments de réflexions à prendre en compte pour la constitution du corpus et l'extraction de termes en nous fondons notamment sur l'analyse de contenu utilisé en psychologie.

Mots-clés : Ontologies, Corpus, Psychologie sociale, Analyse de contenu.

1 Introduction

L'Ingénierie des Connaissances (IC), depuis plus de vingt-cinq ans, vise à fournir des modèles de connaissances pour des systèmes d'aide à la décision, pour faciliter l'interopérabilité entre systèmes d'informations ou encore pour augmenter les possibilités de recherche d'information. A la fin des années 1980, les chercheurs de ce domaine de l'Intelligence Artificielle souhaitaient bâtir des systèmes qui étaient censés reproduire la dimension *dynamique* des experts [1] : la résolution de problèmes et la prise de décisions. Pour cela, il fallait en premier lieu être en mesure de reproduire la dimension *statique* (et souvent implicite) de ces experts : leurs connaissances. Des approches pluri-disciplinaires ont été développées à cette fin, associant psychologues (cognitifs et sociaux), ergonomes et informaticiens. On peut citer parmi elles KOD (Vogel, 1988) ou encore KADS (Wielinga *et al.*, 1992).

Dans les années 1990, cette approche a cependant dû être abandonnée pour plusieurs raisons. La première d'entre elles est que l'expert ne peut être considéré comme l'unique « dépositaire d'un système conceptuel qu'il suffirait de mettre au jour » (Bourigault *et al.*, 2004). Les chercheurs se sont ainsi rendu compte qu'il était impossible de construire un duplicata numérique d'un expert. Non seulement, cette ambition est très couteuse en termes de temps et de finances,

1. Pour Aimé (2014), un individu est considéré comme expert s'il est en mesure de respecter les trois conditions suivantes : (1) la compétence, (2) la référence et (3) la légitimité. La compétence, ou expertise, est pour un individu la détention d'une information utile au groupe dont ses membres ne disposent pas. La référence est pour un individu le pouvoir d'incarner la norme (mais aussi les valeurs, les croyances...) du groupe. Enfin, la légitimité pour un individu est le pouvoir qui lui est donné (par accord social, consensus, règlement, etc.) pour diriger/orienter les conduites des autres membres du groupe.

mais en plus cela ne fonctionne pas. Une autre raison tient au fait qu'un expert, dans une situation donnée, prend rarement une décision seule. La dimension sociale et pragmatique de cette prise de décision, et de la détention de la connaissance d'un domaine, n'était en effet jusque-là pas véritablement prise en compte. Il ne s'agit désormais plus de modéliser la connaissance d'un seul individu ayant le statut d'expert, mais celle d'un groupe d'individus qui partagent un ensemble de connaissances, et ce au moyen de méthodes comme Comon-KADS (Schreiber *et al.*, 1994).

La fin des années 1990 et le début des années 2000 voient l'avènement des ontologies computationnelles dont l'objectif est de modéliser et de formaliser – pour un domaine donné – des connaissances consensuelles. Les méthodes de construction d'ontologies proposées sont fondées principalement sur trois points : (1) développement logiciel (par exemple, Methontology (Fernández-López *et al.*, 1997)), (2) diversification toujours plus grande des sources de connaissances et (3) réutilisabilité des ressources produites. « Les documents et les corpus sont désormais reconnus comme sources principales de connaissances, qu'il y a lieu d'organiser en systèmes utiles» (Prié, 2000). L'ère du *Personal Computer* est révolu. Désormais le monde est connecté, le web est participatif. Les sites internet ne sont plus figés et la seule propriété des informaticiens. Tout le monde est à la fois producteur et consommateur, juge et parti (Ganascia, 2009). Chacun est désormais en mesure de produire du contenu et de se prétendre expert. Les sources de connaissances explosent véritablement en quantité (pas forcément en qualité). De ce fait, la question – pour l'ingénieur des connaissances – n'est plus dans le *comment je prends* mais dans *qu'est-ce que je prends*, dans la sélection, dans le filtrage des ressources qui ne seront plus seulement celles de l'expert. Pour Smith (2012), « le travail des ontologues est une réponse au fait que les nombreuses organisations industrielles, gouvernementales et scientifiques, dont les activités reposent sur l'utilisation des ordinateurs, doivent faire de plus en plus face à des difficultés découlant de leur nécessité de combiner des données provenant de plusieurs sources hétérogènes. »

Durant la décennie passée, et face à la croissance quasi exponentielle de la quantité de documents disponibles sur le Web, des réflexions ont été menées sur l'utilisation de documents textuels dans l'extraction de connaissances afin de construire des ontologies. L'idée se fonde sur « une vision unificatrice de la connaissance : le monde de la connaissance est découpé en domaines stables, dont chacun est équivalent à un réseau fixe de concepts, les termes étant les représentants linguistiques de ces concepts » (Bourigault *et al.*, 2004). Le risque est cependant de tomber dans le même piège des systèmes experts, à savoir considérer que toute la connaissance est exprimée dans le corpus de textes. Une place grandissante est laissée désormais aux méthodes issues du Traitement de la Langue Naturelle (TAL) et aux outils statistiques pour pouvoir gérer des corpus dont la taille est toujours de plus en plus grande. De plus, avec l'avènement du multimédia, les communications écrites comme orales doivent être prises en compte. Parallèlement, on peut constater, au fil des ans, que les recherches en IC sont de moins en moins pluridisciplinaires et deviennent de plus en plus informatiques. Elles s'intéressent de plus en plus à la gestion de la quantité de ces ressources au détriment de la gestion de leur qualité, pour des raisons évidentes de coût des projets, de nécessité de temps de traitement, etc. Plus que jamais, face à cette profusion de ressources et d'experts divers et variés, il est primordial de se poser la question de la qualité des documents et de leur utilisation pour construire des ontologies : est-ce que tout est dans le texte ? peut-on faire confiance aux textes ?

Cet article poursuit plusieurs objectifs. Le premier objectif est tout d'abord de poser quelques

définitions de la notion même de document et des ontologies, notamment sous l'angle de la psychologie sociale (cf. section 2). Le deuxième objectif est de présenter brièvement les grands points de méthodes de construction d'ontologies à partir de textes (cf. section 3). Le troisième objectif est de souligner la difficulté de construire un tel corpus à cet usage (cf. section 4). Enfin le dernier objectif est d'étudier et de proposer quelques pistes pour une meilleure prise en compte de certains critères lors de la construction d'ontologies à partir de textes (cf. section 5).

2 Définitions

L'une des vocations des ontologies est de permettre à un groupe d'individus de tenter d'avoir un consensus sur la signification des termes et des concepts qu'ils emploient. Les propos de cet article s'articulant autour de deux concepts, (1) le document et (2) l'ontologie, il semble pertinent d'essayer dans un premier temps d'en poser quelques bases de définitions.

2.1 Le document

De nombreux domaines s'intéressent au concept de document. De l'Ingénierie Documentaire à la Philosophie, en passant par l'IC et la Psychologie, les définitions sont nombreuses mais semblent néanmoins former un certain consensus. Ethymologiquement, le terme « document » vient du latin *docere* qui signifie enseigner. Selon l'Encyclopédie (1^{ere} édition, Volume 5, 1751), un document réfère à tous titres, pièces et autres preuves qui peuvent donner quelque connaissance d'une chose. Dans le même esprit, l'*International Institute for Intellectual Cooperation* définit un document comme étant toute base de connaissance, fixée matériellement, susceptible d'être utilisée pour consultation, étude ou preuve (par exemple un manuscrit, un imprimé, une représentation graphique ou figurée, ou encore un objet de collection). Plus synthétique, Buckland (1997) résume le document à toute expression de la pensée humaine ; ce que complète Ranganathan (1963), pour qui le document est synonyme d'une micro pensée incarnée apte à la manipulation physique, le transport à travers l'espace, et la conservation dans le temps. Pour Briet (1951), un document est « tout indice concret ou symbolique, conservé ou enregistré, aux fins de représenter, de reconstituer ou de prouver un phénomène physique ou intellectuel ». Selon Prié (2000), un document est « une trace de l'activité humaine, créée par un auteur et mise à disposition de lecteurs. [...] (il) a été créé dans un certain contexte de production, en vue d'un certain contexte de réception, il appartient également à un genre, et réfère de façon implicite ou non à d'autres documents, qui en tant qu'inter-texte en prescrivent également l'interprétation ». Bachimont & Crozat (2004b) abordent le document suivant trois axes : (1) la forme, (2) le signe et (3) le medium. « Le document comme forme renvoie au fait que le document est une forme physique perceptible dont la matérialité physique se prête à l'instrumentation technique. [...] Le document comme signe renvoie aux problèmes de manipulation, lecture et interprétation du contenu, en prenant en compte la double face matérielle et intelligible du signe. [...] Le document comme medium thématise le document comme objet social, objet de négociation et de transaction culturelle et économique. Il constitue un milieu d'échange entre des individus et des groupes qui s'articule et s'ancre dans la vie sociale ». (Bachimont & Crozat, 2004b)

En résumé, le document serait à la fois vue (1) selon une dimension cognitive comme une *mémoire persistante* avec une intention de preuve, et (2) selon une dimension sociale comme un *objet de communication*, que ce soit communication d'une connaissance ou transmission d'un

message doté d'une intentionnalité de la part d'un émetteur à destination d'un (ou plusieurs) récepteur(s) dans un contexte donné.

2.2 Les ontologies

Outre son aspect informatique et son origine philosophique, une ontologie computationnelle peut être considérée comme un objet de la *psychologie sociale*, comme une sorte de *représentation sociale* (dans son approche structurale) formalisée. Jodelet (1989) définit les représentations sociales comme « un ensemble de connaissances / croyances correspondant à un système d'interprétation du réel construit conjointement par un groupe afin de gérer la réalité. » Cette définition est très proche de celle de Gruber (1995) concernant les ontologies computationnelles. Selon cet auteur, ces connaissances / croyances partagées (en règle général de sens commun) facilitent la communication interindividuelle et limitent les conflits. Elles ont un impact sur le plan individuel de par le fait qu'elles définissent l'identité de l'individu comme membre du groupe, mais également son mode de pensée par la sémantique de chaque concept qu'elle définit consensuellement pour le groupe (norme). Ce côté social des ontologies est saillant tant dans la phase de leur construction (question du consensus) que dans celle de leur utilisation (question de l'appropriation), et plus globalement dans la question de la norme qu'elles constituent. Selon Stewart & Fraïssé (2009), les représentations sociales modélisent « ce que les gens pensent connaître et sont persuadés de savoir à propos d'objets, de situations, de groupes donnés. Ces connaissances de sens commun que sont les représentations sociales permettent alors de saisir la signification d'un objet ou d'une situation. Mais cette signification n'est pas inhérente à l'objet de représentation. C'est une réalité construite, appropriée par un individu ou un groupe et intégrée dans son système de valeurs dépendant de son histoire et du contexte social qui l'environne. »

Une ontologie reflète un aspect contextuel d'une catégorisation. Par exemple, modéliser le fait que l'eau *est un* liquide transparent est correct à nos conditions normales de température et de pression atmosphérique. Cette modélisation diffère si nous changeons l'un de ses paramètres, en dessous de zéro degré l'eau *est un* solide blanc (ou transparent). Une ontologie reflète également un aspect social d'une catégorisation. Par exemple, modéliser le fait qu'un chien *est un* animal de compagnie non comestible est correct pour la plupart d'entre nous. Mais cette modélisation diffère dans certains pays d'Asie où un chien *est un* animal comestible.

C'est à ce niveau que l'ontologie, pour un domaine, un endogroupe [2] et un contexte donnés, semble être le plus à même de modéliser une représentation sociale. Si on se réfère au modèle structuré de Abric (1987), d'un point de vue structurel (et intensionnel), le noyau central regrouperait les concepts de haut niveau contenus dans l'ontologie de haut niveau et dans l'ontologie de domaine. Le système périphérique contiendrait essentiellement les concepts de bas niveaux. D'un point de vue extensionnel (les instances), il s'agit principalement d'instances des concepts de bas niveau et donc rattachables au système périphérique. Enfin d'un point de vue expressionnel, il est envisageable de considérer les termes les plus typiques pour dénoter chaque concept comme membre du noyau central du fait de leur aspect consensuel et leur stabilité dans le temps au niveau du groupe.

2. Ce terme est issu de la théorie de l'identité en psychologie sociale. Il s'agit d'un groupe d'individus partageant un ensemble de valeurs ou d'intérêts.

Voyons maintenant quelques éléments de méthodologies pour construire une ontologie à partir d'un corpus.

3 Quelques éléments de méthodologie

Le processus de construction d'une ontologie à partir de corpus de textes est assez simple et se compose (très) schématiquement de quatre grandes étapes.

La première étape, primordiale pour la qualité de l'ontologie, est la *constitution du corpus*. « Le corpus se constitue de documents produits dans le contexte où le problème à résoudre se pose. Ce sont par exemple des documentations techniques, des ouvrages de références, des documents de travail, des manuels propres au domaine ou à l'industrie concernée, ou bien encore la transcription d'interviews menées avec des spécialistes » (Bachimont, 2000). Ce corpus doit être suffisamment large pour couvrir tout le domaine, et consensuel pour répondre à l'objectif d'une ontologie qui est – par définition – la formalisation d'une conceptualisation consensuelle.

La deuxième étape réside dans la *sélection des candidat termes* au moyen d'outils comme SYNTEX (Bourigault *et al.*, 2005) ou BIOTEX (Lossio-Ventura *et al.*, 2014). Ces syntagmes nominaux sont associés au moins à une notion de fréquence. Plusieurs stratégies peuvent dès lors être adoptées. Par exemple, on peut ne prendre que les candidats termes de fréquence élevée ou décider que certains candidats termes de fréquence faible ont aussi une importance dans le domaine. Selon Bourigault *et al.* (2004), « sachant qu'il ne pourra analyser tous les candidats termes extraits du corpus, l'analyste doit adopter une stratégie optimale qui, étant donné le temps qu'il consacre à l'analyse terminologique et le type de la ressource à construire, lui garantit que, parmi les candidats qui auront échappé à son analyse, la proportion de ceux qui auraient pu être pertinents est faible. » On voit ici que le principal critère de sélection réside sur la notion de fréquence (quand elle est élevée), ou sur une pertinence supposée par l'ontologue et validée par un expert. La distribution statistique ne pouvant tenir lieu à elle seule de sémantique, il est nécessaire à ce stade de passer par une étape de normalisation sémantique dans un paradigme donné comme par exemple l'un de ceux proposés par Rastier *et al.* (1997) : (1) le paradigme *référentiel* où chaque terme est lié à l'objet qu'il représente, (2) le paradigme *psychologique* où chaque terme est associé à une image mentale (une représentation psychologique construite et stockée en mémoire sémantique) ou (3) le paradigme *différentiel* où chaque terme est associé aux termes voisins suivant le principe de communauté (avec le père et les frères) et le principe de différence (avec le père et les frères) – paradigme choisi et développé en IC par Bachimont (2000).

La troisième étape va consister en (1) des regroupements de candidats termes synonymes dénotant un même concept, et (2) en une *structuration hiérarchique de ces concepts*. Ces étapes peuvent se réaliser au moyen de l'outil TERMINAE (Aussenac-Gilles *et al.*, 2008), dont on soulignera au passage la possibilité de visualiser le contexte des candidats termes.

La quatrième étape réside dans la *validation* de la conceptualisation obtenue au moyen d'outils collaboratifs tels WebProtégé (Tudorache *et al.*, 2013) ou, d'un point de vue plus formel, par des outils tels OOPS ! (Poveda-Villalón *et al.*, 2014).

4 De la difficulté de choisir un corpus

Pas de bon corpus, pas de bonne ontologie, pas de bons résultats. « La constitution d'un corpus est très délicate de manière générale car le corpus conditionne largement le type et la nature des traitements que l'on peut effectuer sans que l'on ait forcément loisir de choisir le type de données le plus adéquat. Le choix d'un corpus introduit des biais sans qu'il soit toujours loisible de les apprécier » (Bachimont, 2000). S'il n'existe pas a priori de règles établies, Bourigault *et al.* (2004) énumèrent néanmoins quelques conditions à respecter. Il y a, en premier lieu, la question de la *disponibilité des ressources*, disponibilité que l'on peut décliner suivant trois critères : (1) critère de *connaissance* (on ne peut y répondre sans les experts du domaine), (2) critère *légal* (cela nécessite parfois de se plier à certaines contraintes administrative comme l'anonymisation préalable ou une déclaration CNIL), et (3) critère *social* (cela nécessite aussi de savoir qui donne l'accès à la connaissance, d'où une bonne connaissance de l'écosystème, et parfois le respect des voies hiérarchiques et des contraintes politiques). En deuxième lieu, il « convient de s'assurer auprès des spécialistes que les textes choisis ont un *statut suffisamment consensuel* pour éviter toute remise en cause ultérieure de la part de spécialistes ou d'utilisateurs ». Enfin, en troisième lieu, le corpus doit avoir une taille convenable pour couvrir assez largement le domaine étudié, mais aussi pour pouvoir être utilisé manuellement par l'ontologue pour vérification. La question sociale est ici tout aussi importante que la question du contenu du corpus. Le contexte de production est également à prendre en compte tant pour la sélection des documents que pour la sélection des termes (un auteur n'utilise pas toujours les mots désignant exactement ce dont il veut parler). On peut dès lors regretter que la dimension *acte de communication* de ces documents soit rarement abordée en IC. La question de l'objet d'utilisation de l'ontologie est également primordiale. Il doit en effet y avoir adéquation entre le contenu de l'ontologie construite et ce à quoi elle va servir. Dans le cadre de la réflexion menée dans cet article, ce contenu est obtenu principalement à partir d'un corpus de textes. Or, contrairement aux apparences, ce corpus n'est pas toujours porteur des bonnes, ou nécessaires, connaissances pour élaborer l'ontologie. A titre d'illustration, prenons trois domaines : (1) la Psychiatrie, (2) la Mode et l'Habillement et (3) le Droit.

4.1 Domaine de la Psychiatrie

Le domaine de la Psychiatrie comporte de multiples approches théoriques (psychanalytique, systémique, humaniste, cognitiviste, béhavioriste, biomédicale) plus ou moins (in)compatibles, ce qui engendre de nombreuses conséquences tant sur le plan de la conceptualisation (choix des documents pour parler d'une même pathologie) que celui de l'analyse sociale préalable (choix des auteurs, puis des experts pour valider l'ontologie construite). Entre l'approche psychanalytique (et ses nombreux courants) et l'approche biomédicale, les points de vue sont très différents, si ce n'est inconciliables dans une même ontologie. Une fois l'approche théorique choisie, il reste à déterminer les documents en tenant compte du contexte de production. Afin de construire un corpus de référence pour construire une ontologie de la psychiatrie pour un service donné, nous avons questionné au préalable les praticiens pour savoir si de tels documents étaient utilisables en l'état. Nous avons été confrontés à plusieurs cas de figures qui illustrent assez bien la complexité de la tâche. Le contexte de production est – dans ce cas précis – un praticien qui doit écrire un document décrivant un patient atteint d'une pathologie mentale. Ce

document, un acte de communication, est destiné à un autre praticien qui connait le domaine, à la famille du patient qui n'appartient pas au corps médical, et au patient qui souffre d'une pathologie pouvant altérer son jugement et dont l'état peut être aggravé à la lecture du dit document. En conséquence, certains praticiens nous ont révélé ne rien écrire, la transmission à leur confrère se faisant uniquement par oral. D'autres nous disent que le vocabulaire est très adapté à la situation, voir édulcoré. Par exemple, des praticiens ne parlent jamais de *relations sexuelles*, préférant l'usage du syntagme *relations intimes*. De même, ils préfèrent parler de *personnalité complexe* plutôt que de détailler certains éléments de la pathologie. Les termes utilisés sont alors plus à prendre comme des codes dont seuls les praticiens connaissent la véritable signification. Enfin, une grande partie révèle qu'il y a beaucoup d'implicite. Si nous nous tournons maintenant vers les grandes classifications utilisées en psychiatrie, se pose le problème du consensus car – ici peut-être plus que dans les autres domaines de la médecine – ces classifications sont porteuses d'idéologies. Les détracteurs du manuel diagnostique et statistique des troubles mentaux (DSM) soulignent le fait qu'aujourd'hui le problème n'est pas tant dans la classification (qui sera toujours critiquable) que dans les critères diagnostiques qui seraient largement influencés par l'industrie pharmaceutique et déplacerait suivant leur bon vouloir le seuil d'atteinte de telle ou telle pathologie (nécessitant de facto tel ou tel traitement). En résumé, il est assez difficile dans ce domaine de constituer un corpus le plus consensuel possible. Il n'est pas de bonne solution, tout juste la moins mauvaise.

4.2 Domaine de la Mode et de l'Habillement

Depuis l'apparition, au Moyen-Âge, des premiers mots inventés pour désigner un vêtement dans les langues romane et germanique, à l'origine du français et de l'anglais modernes, le vocabulaire de l'habillement a évolué, voire muté, et n'a cessé de se développer en cohérence avec la propre évolution de la garde-robe occidentale. Le nombre de vêtements d'allure distincte en usage il y a plus de deux mille ans, de même que celui des termes spécifiques les désignant, ne dépasse pas la cinquantaine. En 2012, ce chiffre s'élève à plus de cinq cents.

Dans le cadre de nos travaux pour la création de VETIVOC, une ressource termino-ontologie du domaine du Textile, de la Mode et de l'Habillement (Aimé *et al.*, 2014), nous sommes amenés à étudier les vocabulaires utilisés dans les différents champs couverts par ce domaine. Plusieurs ressources à caractère consensuel sont disponibles. La grande majorité des publications dédiées à la mode (95 %) sont positionnées soit sur le créneau dit « beaux livres » (autrement dit, des ouvrages privilégiant l'aspect iconographique à l'aspect textuel), soit sur le créneau littéraire avec des textes historiques, sociologiques ou hagiographiques. Les 5 % des publications restantes sont des dictionnaires spécialisés qui peuvent être répartis en deux groupes : (1) les dictionnaires spécialisés proposant une approche généraliste de la mode et (2) les lexiques exclusivement consacrés aux vêtements ou aux accessoires, avec une approche sélective. Cependant, il est un autre type de ressources – consensuelles – à considérer car bien plus consultés que les ouvrages spécialisés évoqués plus haut : les magazines et les catalogues commerciaux. Véritables prescripteurs, on pourrait penser qu'une telle responsabilité mériterait une rigueur accrue en ce qui concerne les termes employés pour désigner les vêtements et les accessoires. Pourtant, face à l'ampleur, la complexité du vocabulaire de la mode, et le nombre restreint d'outils qui le formalisent, les erreurs et les imprécisions sont fréquentes dans la presse comme dans les catalogues commerciaux.

A titre d'illustrations, prenons l'exemple d'un article de la revue *FashionMag* [3] en date du 11 janvier 2015 sur la pré-collection de Kenzo. Les créations y sont entre autres décrits par les matières utilisées. On y note, d'une part, l'usage mélangé de termes en français et en anglais (alors qu'il existe leur correspondance en français). Mais on y repère, d'autre part, un mauvais usage de ces termes. Ainsi, il est notifié la présence de « shearling », i.e. de mouton retourné. Or la collection est en fausse fourrure ; il ne peut s'agir de vrai mouton retourné. L'emploi de ce terme (1) n'informe donc pas le lecteur avec clarté si son anglais est insuffisant et (2) il induit en plus erreur le lecteur anglophone en lui laissant supposer qu'il s'agit de vrai fourrure. Balteiro (2014) a analysé cette même tendance dans la presse spécialisée espagnole. L'auteur constate que le côté « tendance » de ces anglicismes y est très souvent favorisé en dépit de larges glissements sémantiques.

4.3 Domaine du Droit

Une autre illustration de la difficulté de choisir un corpus est fournie par le domaine juridique. La disponibilité des sources juridiques n'est aujourd'hui plus vraiment un problème, puisque l'accessibilité de la loi est un objectif de valeur constitutionnelle et internet en favorise aisément la mise à disposition physique des textes juridiques. N'apparaît pas non plus problématique la taille convenable de ce corpus, car il y a assez de textes juridiques pour construire une ontologie digne d'intérêt. Au contraire, il peut être nécessaire de limiter la taille du corpus pour assurer une certaine cohérence à l'ensemble.

En revanche, le statut suffisamment consensuel des textes juridiques choisis peut être un sujet de débat. En effet, on ne peut appréhender le domaine juridique de manière unitaire. D'importantes distinctions sont à opérer et elles font souvent l'objet de débats. On peut ainsi d'abord classer le domaine juridique selon les diverses grandes branches du droit : droit administratif, droit pénal, droit civil, droit international, droit européen, etc. On peut également subdiviser ce domaine en fonction de l'objet qu'il entend régir : on distingue ainsi le droit des libertés publiques, le droit des affaires, le droit de la fonction publique, le droit fiscal, le droit de l'urbanisme, le droit de l'environnement, le droit de la famille, le droit des collectivités territoriales, le droit du travail, etc. Ces deux grands critères de classification peuvent se croiser : on trouve ainsi le droit pénal des affaires ou le droit européen de l'environnement. Chaque sous-domaine juridique correspond à une communauté d'expertise distincte. D'un sous-domaine à un autre, les mêmes termes juridiques peuvent ainsi avoir des nuances et des implications différentes. Il est donc important de construire un corpus en tenant compte de ces spécificités et en remettant les termes dans le contexte juridique adéquat.

Un autre critère de choix du corpus peut être la portée juridique des textes. On peut décider de construire un corpus contenant seulement les textes juridiquement contraignants, tels que les règlements européens, les lois et les règlements (décrets, arrêtés) en droit national. Ce sont eux qui créent des droits et des devoirs pour les individus, les entreprises et toute autre personne morale. On écartera ainsi les circulaires, les instructions, les décisions jurisprudentielles, ainsi que la doctrine juridique (articles de revues spécialisées, manuels et ouvrages). L'orientation de ces choix dépendra essentiellement de l'objectif poursuivi par l'ontologie. En d'autres termes, il faudra se poser la question de savoir à qui s'adressent le corpus et l'ontologie. S'ils visent

3. http ://www.fashionmag.com

un public spécialisé de juristes, juges et avocats, le fait de ne pas prendre en compte les textes d'orientation, telles que les circulaires et les instructions, les jurisprudences et les articles et ouvrages de doctrine, engendrera un corpus amputé et sera loin de remplir sa fonction d'aide à la décision. S'ils visent un public de non juristes qui cherche à trouver des éléments pour l'aider à prendre une décision en vue de résoudre un problème opérationnel, il peut être intéressant, pour ne pas noyer les utilisateurs, de limiter le corpus aux textes juridiquement contraignants en écartant l'insertion directe des autres composantes d'un bon corpus juridique (circulaires, instructions, jurisprudences, doctrine). Toutefois, ces dernières ne devraient pas être purement et simplement supprimées mais remplacées par des textes de vulgarisation permettant d'éclairer les textes juridiques contraignants dont la seule lecture n'est pas suffisante pour aider à la décision.

5 Discussion

La démarche de construction d'une ontologie à partir d'un corpus de textes s'apparente en de nombreux points au processus d'analyse de contenu. Ce processus a été développé il y a plusieurs décennies en Psychologie (elle s'apparente également à l'herméneutique ancienne, dans une tradition aristotélicienne). Il s'agit d'un « ensemble de techniques d'analyse des communications visant, par des procédures systématiques et objectives de description du contenu des énoncés, à obtenir des indicateurs (quantitatifs ou non) permettant l'inférence de connaissances relatives aux conditions de production/réception (variables inférées) de ces énoncés » (Bardin, 1977). Cette procédure, de type structuraliste, comporte deux étapes successives : (1) l'*inventaire des termes*, et (2) leur *classification*. Ainsi, l'analyse de contenu vise à décrire objectivement, systématiquement et quantitativement le contenu du corpus étudié. Cette méthode – telle que présentée par l'auteur – a pour objectif de construire une catégorisation de manière exhaustive à partir des termes utilisés dans le corpus et ainsi fournir à l'analyste (ou l'ontologue) un réseau sémantique reflétant le contenu du corpus étudié. Cette méthode se fonde sur l'hypothèse de Whorf quant à l'interdépendance de la langue et de la pensée (Whorf *et al.*, 2012). Pour l'auteur, la catégorisation en analyse de contenu « a pour objectif premier de fournir par condensation une représentation simplifiée des données brutes. » Sur cette représentation, l'analyste va opérer des inférences, avec comme hypothèse forte qu'il n'introduit pas de biais. Deux possibilités sont envisagées : (1) une *catégorisation a priori* où l'analyste élabore sa catégorisation à partir des termes recueillis, ou (2) une *catégorisation a posteriori* où l'analyste dispose d'une classification qu'il va peupler avec de nouvelles catégories ou avec des termes extraits. Bardin (1977) énumère cinq critères à respecter pour obtenir une bonne catégorisation : (1) l'*exclusion mutuelle* où un terme ne peut appartenir qu'à une seule catégorie (éviter au grand maximum les ambiguïtés) ; (2) l'*homogénéité* où « un même principe doit gouverner l'organisation des catégories » ; (3) la *pertinence* où les catégories construites doivent répondre à la problématique retenue ; (4) l'*objectivité* et la *fidélité* où si on a plusieurs analystes, on doit avoir la même grille de lecture, et les variables traitées doivent être clairement définies dès le départ ; et (5) la *productivité* où les catégories créées doivent permettre une inférence riche, des hypothèses nouvelles et représenter des données fiables.

Dans ce cadre, de nombreuses réflexions ont été menées sur le choix des termes et le sens à y associer. En effet, le sens donné à un terme, le type de message sous-tendu, et donc la validité du terme utilisé dépendent du contexte, lequel peut être exprimé en termes de sphères tels que pro-

posés par Searle et Austin dans la classification des actes de discours. Par exemple, Chabrol & Bromberg (1999) propose une grille de lecture permettant l'identification d'une cinquantaine de types d'actes de parole, regroupés en cinq sphères dont une sphère informationnelle regroupant tout ce qui est à propos des objets du monde (elle sert à construire un environnement manifeste ; c'est assez typiquement la sphère des connaissances).

Un texte est un acte de communication dont l'objet est à la fois (1) de transmettre une information à autrui, et (2) d'agir sur autrui. Pour Bachimont & Crozat (2004b), un document est un « objet matériel qui se déploie dans la temporalité d'une lecture et qui donne lieu à une interaction avec le lecteur ». Considérer une ontologie construite uniquement à partir de textes revient à (re-)définir la notion d'ontologie comme étant « le reflet d'une des façons dont la connaissance peut être perçue et dite » (Aussenac-Gilles & Sörgel, 2005). Selon Grice (1957), « the meaning (in general) of a sign needs to be explained in terms of what users of the sign do (or should) mean. » Cela a deux conséquences : (1) si, par définition, une ontologie est un objet formalisant une conceptualisation consensuelle, alors il doit également y avoir consensus au préalable sur ce corpus et donc sur le discours des auteurs, et (2) lorsqu'il construit une ontologie à partir d'un corpus de textes, l'ontologue se trouve influencé non seulement par sa propre subjectivité, par l'objectif de son analyse, ses options théoriques, les enjeux du projet, les besoins d'exhaustivité, mais aussi par les auteurs du corpus sur lequel il travaille.

Un autre point est également à prendre en compte : celui de la confusion entre la dimension syntaxique (et statistique) et la dimension sémantique. Selon Roche (2007), « les structures conceptuelles construites à partir de textes ne sont pas des ontologies au sens d'une conceptualisation d'un domaine au-delà de tout discours. Elles relèvent de la sémantique lexicale : il n'y a pas de concepts dans un texte, mais uniquement des usages linguistiques de ces concepts. La construction d'ontologies à partir de textes repose sur un ensemble d'hypothèses fortes. [...] La première est de dire que les experts peuvent traduire leurs connaissances ontologiques du domaine dans des textes et que ces derniers constituent un monde plus ou moins clos. La deuxième considère le processus de rétro-ingénierie comme possible, basée sur le fait que certaines catégories de mots et que certaines relations linguistiques traduisent un usage en langue de concepts et de relations conceptuelles. La troisième hypothèse postule que les structures lexicale et conceptuelle sont relativement isomorphes. Enfin, que la validation par les experts suffit à ériger la structure conceptuelle comme une ontologie du domaine. » L'ensemble de ces hypothèses font que l'ontologie ainsi construite ne repose finalement que sur le décret du statut conceptuel de constats syntaxiques fondés sur des fréquences. Mais construire une ontologie en dehors de ce cadre statistique est difficilement automatisable et donc fortement chronophage. On comprend dès lors pourquoi cette solution n'est pas aujourd'hui majoritairement prisée.

6 Conclusion

L'IC est née d'une inter-disciplinarité alliant informatique, intelligence artificielle, psychologie cognitive, ergonomie et sciences de gestion. Aujourd'hui, il semblerait que l'IC soit plus orientée vers la dimension technologique, notamment de par les problèmes d'hétérogénéité et de quantité massive de données qu'elle doit gérer.

Dans ses premières heures, l'IC – avec ses systèmes experts – était dans un idéal de biomimétisme de l'expert où la machine allait reproduire le fonctionnement de son cerveau. Nous assistons aujourd'hui à un véritable changement de paradigme en IC. Il ne s'agit plus de repro-

duire virtuellement et fidèlement, sous la forme d'une boite noire autonome, le fonctionnement d'un expert avec l'aide d'autres disciplines des Sciences Humaines. Il s'agit de fournir un outil informatique donnant de manières appropriée et intelligente des informations à un utilisateur qui est confronté à un problème, et qui vont l'aider à résoudre ce problème en lui apportant des connaissances supplémentaires adéquates. Les systèmes ne sont plus alors décisionnels, mais qualifiés d'*aide à la décision*.

Il nous semble donc qu'un petit retour aux sources ne serait pas préjudiciable. Nous n'inventons rien, nous ne faisons que rappeler certaines évidences que les ontologues d'aujourd'hui ne voient peut-être plus. Dans la lignée de Vygostky, il nous faut ré-aborder les connaissances comme étant enracinées dans le social et considérer les documents au-delà de leur aspect mémoriel. Une ontologie crée à partir d'un corpus de textes bien choisi et dans lequel les termes auront été extraits d'un point de vue pragmatique, puis utilisé en adéquation avec son milieu, offre d'énormes avantages que ce soit en termes d'appropriation (dans le sens de « faire sien le contenu et de l'intégrer comme une part de soi » (Bachimont & Crozat, 2004a)) ou en termes de représentation sociale. Pour connaître sa réutilisabilité, il est également nécessaire de lui adjoindre en méta-données un certain nombre d'informations sociales telles que les auteurs du corpus utilisé (et évaluer leur statut d'autorité ainsi que la consensualité du corpus), le contexte de production du corpus, quel groupe d'individus est destinataire du projet, dans quelle sphère ont été choisi les termes...

Remerciements

Je remercie Rossella Pintus et Sophie George pour leur éclairage respectifs sur les domaines du Droit et de la Mode/Habillement.

Références

ABRIC J. (1987). *Coopération, Compétition et Représentation Sociale*. Fribourg, Suisse : Delval.

AIMÉ X. (2014). Pour une approche écologique des ontologies computationnelles. *Intellectica – Dossier spécial "Philosophie du Web et Ingénierie des Connaissances"*, **1**(61), 189–210.

AIMÉ X., GEORGE S. & HORNUNG J. (2014). VETIVOC, une Ressource termino-ontologique modulaire du domaine du textile, de la mode et de l'habillement. *Revue d'Intelligence Artificielle*, **6**, 689–728.

AUSSENAC-GILLES N., DESPRES S. & SZULMAN S. (2008). The terminae method and platform for ontology engineering from texts. In *Proceedings of the 2008 Conference on Ontology Learning and Population : Bridging the Gap Between Text and Knowledge*, p. 199–223, Amsterdam, The Netherlands, The Netherlands : IOS Press.

AUSSENAC-GILLES N. & SÖRGEL D. (2005). Text Analysis for Ontology and Terminology Engineering. *Applied Ontology, IOS Press*, **1**(1), 35–46.

BACHIMONT B. (2000). Engagement sémantique et engagement ontologique : conception et réalisation d'ontologies en ingénierie des connaissances. In J. CHARLET, M. ZACKLAD, G. KASSEL & D. BOURIGAULT, Eds., *Ingénierie des connaissances : évolutions récentes et nouveaux défis*, p. 305–323. Paris, France : Eyrolles.

BACHIMONT B. & CROZAT S. (2004a). Instrumentation numérique des documents : pour une séparation fonds/forme. *Information-Interaction-Intelligence*, **4**(1), 95–104.

BACHIMONT B. & CROZAT S. (2004b). Réinterroger les structures documentaires : de la numérisation à l'informatisation. *Information–Interaction–Intelligence*, **4**(1), 59–74.

BALTEIRO I. (2014). The influence of English on Spanish Fashion Terminology : -ING forms. *Journal of English for Specific Purposes at tertiary level*, **2**(2), 156–173.

BARDIN L. (1977). *L'analyse de contenu*. Paris, France : PUF.

BOURIGAULT D., AUSSENAC-GILLES N. & CHARLET J. (2004). Construction de ressources terminologiques ou ontologiques à partir de textes : un cadre unificateur pour trois études de cas. *Revue d'Intelligence Artificielle*, **18**(4), 97–110.

BOURIGAULT D., FABRE C., FRÉROT C., JACQUES M. & OZDOWSKA S. (2005). Syntex, analyseur syntaxique de corpus. In *TALN 2005, Dourdan, 6-10 juin*.

BRIET S. (1951). *Qu'est-ce que Ia documentation*. Editions Documentaires Industrielles et Techniques.

BUCKLAND M.-K. (1997). What is a "document" ? *JASIS*, **48**(9), 804–809.

CHABROL C. & BROMBERG M. (1999). Préalables à une classification des actes de parole. *Psychologie française*, **44**(4), 291–306.

FERNÁNDEZ-LÓPEZ M., GÓMEZ-PÉREZ A. & JURISTO N. (1997). METHONTOLOGY : From ontological art towards ontological engineering. In *Proceedings of the The Fourteenth National Conference on Artificial Intelligence (AAAI'97), Workshop on ontological engineering*, p. 33–40, Stanford, USA.

GANASCIA J. (2009). *Voir et pouvoir : qui nous surveille ? Un essai sur la sousveillance et la surveillance à l'ère de l'infosphère*. Paris, France : Editions du Pommier.

GRICE H.-P. (1957). Meaning. *Philosophical Review*, (66), 377–388.

GRUBER T. (1995). Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human Computer Studies*, **43**(5/6), 907–928.

JODELET D. (1989). *Les représentations sociales*. Paris, France : PUF.

LOSSIO-VENTURA J., JONQUET C., ROCHE M. & TEISSEIRE M. (2014). BIOTEX : A system for Biomedical Terminology Extraction, Ranking, and Validation. In *Proceedings of the 13th International Semantic Web Conference (ISWC'14). Trento, Italy*.

POVEDA-VILLALÓN M., GÓMEZ-PÉREZ A. & SUÁREZ-FIGUEROA M. (2014). OOPS !(OntOlogy Pitfall Scanner !) : An On-line Tool for Ontology Evaluation. *International Journal on Semantic Web and Information Systems (IJSWIS)*, **10**(2), 7–34.

PRIÉ Y. (2000). Sur la piste de l'indexation conceptuelle de documents, une approche par l'annotation. *Document Numérique*, **4**(162), 11–35.

RANGANATHAN S. (1963). *Documentation and its facets*. London, UK : Asia Publishing House.

RASTIER F., CAVAZZA M. & ABEILLÉ A. (1997). *Sémantique pour l'analyse*. Paris, France : Masson.

ROCHE C. (2007). Dire n'est pas concevoir. In 18es *Journées Francophones d'Ingénierie des Connaissances (IC'2007)*, p. 157–168, Toulouse, France : Cépaduès. ISBN 978-2-85428-773-8.

SCHREIBER G., WIELINGA B., AKKERMANS H., VAN DE VELDE W. & ANJEWIERDEN A. (1994). CML : The CommonKADS conceptual modelling language. In *A future for Knowledge Acquisition*, p. 1–25. Springer.

SMITH B. (2012). How to do Things with Documents. *Rivista di Estetica*.

STEWART I. & FRAŠSÉ C. (2009). *La pensée sociale*, chapter Les schèmes cognitifs de base : un modèle pour étudier les représentations sociales, p. 99–119. Eres : Paris, France.

TUDORACHE T., NYULAS C., NOY N. & MUSEN. M. (2013). WebProtégé : A collaborative ontology editor and knowledge acquisition tool for the web. *Semantic web*, **4**(1), 89–99.

VOGEL C. (1988). *Génie cognitif*. Paris, France : Masson (Sciences Cognitives).

WHORF B.-L., J.-B.CARROLL, LEVINSON S.-C. & LEE P. (2012). *Language, thought, and reality : Selected writings of Benjamin Lee Whorf*. Mit Press.

WIELINGA B., SCHREIBER A. & BREUKER J. (1992). KADS : A modelling approach to knowledge engineering. *Knowledge Acquisition*, **4**(1), 5–53.