Chapitre 7

Les enjeux de l'Ingénierie des connaissances*

Pour référencer ce chapitre :

AUSSENAC-GILLES N., CHARLET J., REYNAUD C. Chapitre 7 - Les enjeux de l'Ingénierie des Connaissances, in *Information-Interaction-Intelligence : le point sur le I3*, Eds(s): Sèdes F., Ogier J-M., Marquis P., Toulouse : Cépaduès Editions, p 244-266, 2012

*Ce chapitre est la version raccourcie d'un chapitre publié chez Hermès au sein d'une collection sur l'Intelligence Artificielle coordonnée par Henri Prade, Odile Papini et Pierre Marquis.

Les enjeux de l'Ingénierie des connaissances*

Nathalie Aussenac-Gilles[1], Jean Charlet[2], Chantal Reynaud[3]

[1] IRIT – CNRS, université de Toulouse ;
aussenac@irit.fr
[2] INSERM UMRS 872 Paris ; AP-HP Paris ;
Jean.charlet@upmc.fr
[3] LRI – Université Paris-Sud & INRIA Saclay Île-de-France ;
chantal.reynaud@lri.fr

Résumé. L'Ingénierie des Connaissances propose des concepts, méthodes et techniques permettant de modéliser, de formaliser et d'acquérir des connaissances dans les organisations dans un but d'opérationnalisation, de structuration ou de gestion au sens large. L'Ingénierie des Connaissances trouve ainsi ses champs d'application dans les domaines où l'objectif est de modéliser les connaissances et les mettre à disposition comme support à une activité ou à un raisonnement. Les innovations du domaine comprennent des méthodes, des logiciels et interfaces d'aide à la modélisation, ainsi que des représentations conceptuelles ou formelles. Nous dressons ici un rapide panorama des questions et des avancées majeures de l'Ingénierie des connaissances, dont les résultats récents sont marqués par l'essor du web sémantique, des ontologies et du web de données liées.
Mots clefs : Ingénierie des connaissances, modélisation conceptuelle, ontologies, web sémantique, Web de données.
Abstract : Knowledge Engineering offers concepts, methods and techniques for modeling, formalizing and acquiring knowledge in organizations. The goal is to structure, operationalize, and manage knowledge in a large acceptation. Knowledge Engineering scope is application domains where it is necessary to model knowledge and make it available as activity support or reasoning. Innovations of the field include methods, software and interfaces to support modeling, and conceptual or formal representations. We propose here a brief overview of issues and major advances in knowledge engineering, including the recent results marked by the rise of semantic web, ontologies, and linked open data.
Keywords : Knowledge engineering, Conceptual Modelization, Ontologies, Semantic Web, Web of data.

*Ce chapitre est la version raccourcie d'un chapitre publié chez Hermès au sein d'une collection sur l'Intelligence Artificielle coordonnée par Henri Prade, Odile Papini et Pierre Marquis.

1 Introduction

L'acquisition des connaissances est apparue comme une discipline avec un objet de recherche, les systèmes à bases de connaissances, à la fin des années 80. À la suite des développements des systèmes experts durant la décennie, la question de la modélisation et de l'acquisition des connaissances pour ces systèmes est devenue cruciale et justifia de nombreux travaux de thèses, que ce soit avec des problématiques très cognitives ou plus orientés vers la définition de représentations. Depuis le début des années 2000, la perspective s'est élargie et l'Ingénierie des Connaissances (IC) propose des concepts, méthodes et techniques permettant de modéliser, de formaliser et d'acquérir des connaissances dans les organisations dans un but d'opérationnalisation, de structuration ou de gestion au sens large.

L'IC trouve ainsi ses champs d'application dans les domaines où l'objectif est de modéliser les connaissances et les mettre à disposition comme support à une activité ou à un raisonnement. Les applications concernées sont celles liées à la gestion des connaissances, à la recherche d'information (sémantique), à l'aide à la navigation, à l'aide à la décision. Enfin, l'IC entretient des relations étroites avec le Web Sémantique, qui a un statut particulier en raison de forts recouvrements avec l'IC via le partage de nombreux outils et méthodes (ontologies, langages de représentation des connaissances, raisonnements, etc.).

Dans la suite de ce chapitre, nous allons, en section 2, présenter, selon un axe historique, les modélisations utilisées en IC, puis, en section 3, développer les principaux problèmes de recherche qui se posent dans le domaine. La section 4 permettra de synthétiser les enjeux méthodologiques du domaine avant de conclure.

2 Modélisations utilisées

2.1 La notion de modèle conceptuel

Autour des années 1990, il a été proposé de construire un système à base de connaissances (SBC) en commençant par la description des connaissances du système indépendamment de leur implémentation. Le support pour rendre compte de cette représentation a été appelé *modèle conceptuel*. Il s'agit de construire un modèle adapté à la nature des connaissances à décrire pour pouvoir ensuite le représenter dans des formalismes adéquats pour le SBC.

La façon dont sont décrites les connaissances conditionne la construction du SBC et surtout le fait que l'on puisse comprendre et s'approprier son fonctionnement. L'acquisition et l'IC ont repris à leur compte les travaux de Newell (1982) qui, le premier, a différencié les connaissances à représenter dans un système et l'implémentation de ce système. Newell a fait apparaître la nécessité d'un niveau de description des systèmes qui ne soit pas celui des symboles et langages informatiques, le *niveau des connaissances*. Ce niveau de description doit permettre de décrire le comportement du système observé indépendamment de son implémentation formelle. Ce système est de plus considéré comme un agent rationnel qui dispose de connaissances, doit atteindre des buts, sait effectuer des actions, et est rationnel, c'est-à-dire qu'il choisit (avec ses connaissances) l'enchaînement des actions qui va le mener le plus directement au but. Dans cette même proposition, Newell distingue

les *connaissances du domaine* des *connaissances de raisonnement* à savoir des actions et des buts, modélisés via des *méthodes* et des *tâches*.

2.2 Les modèles de raisonnement

Les modèles de raisonnement décrivent de façon abstraite le processus de résolution à mettre en œuvre dans un SBC en termes de tâches et de méthodes, les tâches étant réalisées par des méthodes. Cette distinction a été adoptée pour rendre compte de raisonnements (Klinker *et al.*, 1991 ; Schreiber *et al.*, 1994 ; Tu *et al.*, 1995) car elle présente l'avantage de décrire séparément le but visé de la façon de l'atteindre. Pour décrire une résolution de problèmes, on peut mettre en évidence des tâches de différents niveaux, les tâches de plus bas niveau poursuivant des sous-buts pour les tâches plus générales. Les méthodes décrivent comment un but peut être atteint à l'aide d'une série d'opérations réalisées dans un certain ordre.

2.3 Des modèles conceptuels aux ontologies

À la suite des travaux sur les raisonnements, il a semblé intéressant de construire le modèle conceptuel d'une application en combinant réutilisation de composants de modèles de raisonnement et abstraction de connaissances du domaine. Une analyse des connaissances du domaine devient alors nécessaire pour mettre en correspondance les connaissances du domaine et leurs rôles dans le raisonnement (Reynaud *et al.*, 1997). Les connaissances du domaine sont ainsi décrites en un noyau, l'*ontologie du domaine*, regroupant les classes d'entités du domaine, les concepts, et les relations entre ces entités. Ainsi, l'ontologie définit le vocabulaire logique qui permet d'exprimer des faits et des connaissances du domaine sur lesquelles raisonner. Certains concepts, dits primitifs, sont définis par leur place dans la hiérarchie des classes. D'autres, dits concepts définis, s'expriment sous forme de conditions nécessaires et suffisantes, à partir des concepts primitifs et de leurs relations.

Le concept d'ontologie a été introduit en intelligence artificielle par le projet ARPA *Knowledge Sharing Effort* (Neches *et al.*, 1991) et c'est Gruber (1993) qui en a proposé une première définition en IC. Une définition actuelle, proposée dans (Studer *et al.*, 1998), fait consensus dans le domaine :

> *An ontology is a formal, explicit specification of a shared conceptualisation. Conceptualisation refers to an abstract model of some phenomenon in the world by having identified the relevant concepts of that phenomenon. Explicit means that the type of concepts used, and the constraints on their use are explicitly defined. Formal refers to the fact that the ontology should be machine-readable. Shared reflects the notion that an ontology captures consensual knowledge, that is, it is not private of some individual, but accepted by a group.*

Ainsi, l'ontologie répond à des exigences complémentaires et symétriques : *a*) en tant que spécification, elle définit une *sémantique formelle* pour l'information permettant son exploitation par un ordinateur ; *b*) en tant que reflet d'un point de vue – partiel – sur un domaine, que l'on cherche le plus consensuel possible, elle fournit une sémantique qui doit permettre de lier la forme exploitable par la machine avec sa signification pour les humains – on parle alors de *sémantique interprétative*. Le fait que l'ontologie soit formelle est à la fois sa qualité par les raisonnements qu'elle permet et son handicap, par la difficulté qu'il y

a à la construire et la représenter formellement dans un langage de représentation des connaissances.

En fonction de l'utilisation de l'ontologie, celle-ci va être plus ou moins riche en relations et en concepts définis : si par exemple, l'ontologie est utilisée dans une application de recherche d'information simple, l'ontologie servira d'abord à répertorier les concepts du domaine associés à des termes de ce même domaine. On parle alors *d'ontologie légère*. Si l'ontologie doit servir à décrire des domaines où l'on veut modéliser des raisonnements, elle sera en général plus volumineuse, comportera plus de relations contribuant à construire des concepts définis et, de façon générale, à décrire les concepts sur lesquels on veut raisonner. On parle ici d'*ontologie lourde*.

Par rapport à leur construction, les ontologies ont suscité de nombreuses attentes fondées sur une idée de généricité. Plusieurs recherches ont contribué à définir ce qui pouvait être générique dans une ontologie et quelles pouvaient être les méthodes à mettre en œuvre pour construire les parties non génériques. Ainsi, il est maintenant classiquement admis de distinguer dans une ontologie :

- la *top*-ontologie, niveau le plus élevé, structure les connaissances avec des catégories dont l'organisation dépend de réflexions philosophiques. La *top*-ontologie est partagée, générique, et la construction d'une ontologie peut s'inspirer de celles proposées par quelques équipes de recherche (*top-level categories* de Sowa[1], SUMO[2], DOLCE[3], …). Toutefois, aucune norme n'a encore été dégagée[4] ;
- l'ontologie noyau ou *core*-ontologie, fournit les concepts structurants d'un domaine et décrivant les relations entre ces concepts – en médecine p. ex., on y trouve des concepts de diagnostic, signe, structure anatomique et des relations comme celles liées à la localisation d'une pathologie sur une structure anatomique (cf. GFO-Bio[5]); dans le droit, l'ontologie LKIF-Core[6] propose les notions de norme, action légale et rôle légal ;
- l'ontologie du domaine, c'est-à-dire les concepts du domaine tels qu'ils sont manipulés par les professionnels. C'est la partie la plus spécifique de la modélisation. Il n'y a pas de frontière claire entre la *core*-ontologie et l'ontologie du domaine ; elles sont construites de façon concomitante. Les approches à partir de corpus, complétées par la réutilisation de thésaurus ou de terminologies du domaine modélisé, sont souvent utilisées pour construire cette partie des ontologies (*Cf.* 4.1).

3 Problèmes considérés et résultats

Un des faits marquants de l'évolution de l'IC est d'avoir diversifié les sources de connaissances utilisées dans les systèmes de traitement d'information « intelligents », et ce pour tirer profit à la fois de leur complémentarité et de l'évolution des techniques disponibles pour les analyser. Nous faisons tout d'abord un panorama des sources de

[1] http://www.jfsowa.com/ontology/toplevel.htm
[2] http://www.ontologyportal.org/
[3] http://www.loa-cnr.it/DOLCE.html
[4] http://ontolog.cim3.net/cgi-bin/wiki.pl?UpperOntologySummit/UosJointCommunique
[5] http://www.onto-med.de/ontologies/gfo-bio/index.jsp
[6] http://www.estrellaproject.org/lkif-core/

connaissances sur lesquelles les méthodes d'IC se sont successivement focalisées ; nous montrons ensuite comment des méthodes et techniques de modélisation proposées en IC permettent de dépasser ces problèmes ; enfin, nous présentons les résultats portant sur la réutilisation de modèles puis le lien avec les travaux en représentation des connaissances.

3.1 Les sources de connaissances

Historiquement, la connaissance a d'abord fait référence à une expertise humaine dont devait rendre compte la base de connaissances de *systèmes experts*. Ces connaissances étaient techniques et spécialisées. Elles correspondaient à des savoir-faire de haut niveau rarement verbalisés, qu'il s'agissait de pérenniser et de transmettre à un système informatique sous forme de règles de production[7]. Les systèmes experts ont évolué vers des *systèmes à base de connaissances* dont le rôle était d'aider les utilisateurs, en privilégiant l'efficacité et non la fidélité à un raisonnement humain. Les systèmes mettent alors en œuvre des modes de raisonnement qui leur sont propres. L'objectif est de réaliser des tâches, soit de manière autonome, soit de manière coopérative en s'adaptant aux différents contextes d'utilisation et aux profils des utilisateurs. Les connaissances mises à disposition, comme support à un raisonnement ou à une activité, correspondent alors à des savoirs techniques, consensuels et partagés, sous forme de règles ou de plan d'action, ou à des descriptions structurées et finalisées d'un domaine.

L'évolution historique des systèmes informatiques « à base de connaissances » permet d'identifier les différentes dimensions des connaissances à prendre en compte : les connaissances individuelles expertes ; les connaissances liées aux pratiques, aux activités et aux usages individuels ; les connaissances portant sur les organisations ; les connaissances consensuelles et partagées relatives à des domaines d'application ; les connaissances de sens commun ; les connaissances provenant du recoupement de données ou d'informations réparties sur le web.

C'est pour accéder à ces différents types de connaissances que de nouvelles sources ont été prises en compte. Ainsi, la place occupée par les documents a augmenté avec la disponibilité croissante des documents numériques. L'IC s'intéresse aux documents, en particulier les documents textuels, comme porteurs de sens et révélateurs de connaissances depuis les premières études sur l'acquisition des connaissances pour les systèmes experts. Les documents sont exploités pour leur contenu, en complément ou en alternatives aux experts et spécialistes du domaine. Des données peuvent également être la source de connaissances via des mécanismes d'extraction de connaissances à partir de données (ECD). Enfin, des composants de modèles de connaissances pré-existants peuvent être réutilisés lorsqu'ils portent sur des connaissances consensuelles et partagées. Il peut s'agir de méthodes de résolution de problèmes applicables à différents domaines (les bibliothèques de méthodes sont un des résultats majeurs des projets européens KADS et CommonKADS[8] (Schreiber *et al.*, 1999), de modèles du domaine ou ontologies correspondant à des représentations structurées définissant les concepts d'un domaine, ou de ressources telles que des bases de données lexicales (ex : WordNet[9] qui répertorie, classe et met en relation le contenu sémantique et lexical de la langue anglaise) ou des thesaurus

---

[7] Pour un historique sur les systèmes à base de connaissances, lire (Stefik 1995 ; Aussenac *et al.* 1996 ; Charlet *et al.* 2000).

[8] http://www.commonkads.uva.nl/

[9] http://wordnet.princeton.edu/wordnet/

correspondant à des vocabulaires normalisés sur un domaine, constitués d'ensembles structurés de termes.

3.2 Comment passer des sources de connaissances aux modèles : questions de recherche

Une des problématiques centrales et originales de l'IC est bien d'outiller – techniquement et méthodologiquement – le passage des sources mentionnées en 3.1 aux modèles identifiés en partie 2. Ces techniques font souvent appel à des logiciels mais aussi à des cadres d'analyse ou à des grilles d'observation venant d'autres disciplines. La recherche en IC est bien celle d'une ingénierie au sens où il s'agit d'innover autant dans la création d'outils, langages, méthodes que dans leur sélection et leur adaptation lorsqu'ils existent, mais surtout dans leur agencement pertinent au sein de guides méthodologiques et de plateformes de travail intégrées. L'innovation porte autant sur la nature et le développement de ces outils que dans la définition des conditions de leur utilisation et de la complémentarité de leurs interactions, pour traiter des types particuliers de connaissances, à chaque étape du développement d'une application.

Depuis près de 20 ans, les recherches méthodologiques en IC soulèvent des problématiques transverses, reformulées et renouvelées en fonction des sources de connaissances, de la nature des modèles construits ainsi que des utilisations et raisonnements prévus à partir de ces modèles.

3.2.1 Comment construire un modèle ?

Deux courants méthodologiques complémentaires ont défini d'abord des étapes et des techniques différentes (Aussenac *et al.*, 1992). Les démarches ascendantes privilégient les analyses de données, d'abord sur la base des besoins identifiés puis en fonction des parties de modèle à renseigner. Ces démarchent mettent l'accent sur les logiciels et techniques de recueil, d'extraction, de fouille et d'identification de connaissances, puis sur l'aide à la caractérisation abstraite des connaissances (classification, structuration, identification des méthodes et raisonnements). A l'inverse, les méthodes descendantes privilégient la *réutilisation* pour construire des modèles par adaptation et assemblage de composants existants, le recueil et l'extraction étant au service du choix des composants puis de la spécialisation du modèle. Pour unifier l'ensemble, on peut considérer la modélisation comme un cycle alternant phases ascendantes et descendantes, progressant d'étapes essentiellement consacrées au recueil ou à la réutilisation à des phases de représentation de plus en plus formelle. La plupart des méthodes et outils présentés en 3.3 combinent ces deux courants.

3.2.2 Comment exploiter la complémentarité entre sources de connaissances ?

La diversité des connaissances utilisées constitue un des moyens de parvenir à des modèles plus précis ou d'automatiser une partie de leur construction. Une méthode d'IC doit identifier les sources qu'elle permet de traiter, les outils et techniques permettant de les exploiter au mieux puis guider *l'articulation* de ces outils au sein d'une démarche, pour assurer l'exploitation complémentaire de leurs résultats et construire un modèle. C'est ce qu'illustrent les résultats présentés en 3.3.

3.2.3 Comment l'ingénierie des modèles intègre l'objectif de leur utilisation ?

Plusieurs travaux ont montré que les modèles conceptuels étaient d'autant plus pertinents qu'ils étaient finalisés, produits pour des applications spécifiques. L'IC ne se limite donc pas à produire des modèles mais elle est concernée par leur utilisation dans la mesure où cette utilisation détermine en partie leur contenu, leur structure, et de ce fait la manière de les construire. La visée d'utilisation d'un modèle a donc un impact sur les choix méthodologiques et sur les formalismes de représentation retenus (Bourigault *et al.*, 2004).

3.2.4 Comment favoriser la réutilisation des modèles ?

La réutilisation d'éléments de connaissances déjà structurés est souvent privilégiée pour réduire le coût de la modélisation de connaissances. Ceci n'est possible que si l'on connait bien les principes selon lesquels ces modèles ont été construits, si on peut les comparer, le combiner, si le fait d'en reprendre des fragments et de les recomposer au sein de nouveaux modèles est techniquement possible et a du sens. Ces questions se retrouvent aujourd'hui dans les travaux sur l'alignement, la réutilisation et la composition de parties d'ontologies pour en construire de nouvelles.

3.2.5 Comment assurer l'évolution des modèles en lien avec leur contexte d'utilisation ?

Les modèles de connaissances utilisés dans des applications s'intègrent dans un cycle de vie qui intègre leur évolution, rendue nécessaire suite à l'évolution des sources de connaissances, des connaissances du domaine, des besoins des utilisateurs. Depuis 2008, l'évolution des ontologies, posée comme un enjeu fort de leur utilisation, fait l'objet de recherches définissant un cycle d'évolution, des moyens d'identifier les connaissances à modifier tout en assurant la cohérence logique du modèle (Stojanovic, 2004 ; Luong, 2007).

3.3 La construction de modèles : techniques, méthodes et outils

Pour faire des propositions pratiques en matière d'accès aux connaissances à travers les personnes ou les documents qui sont supposés en fournir des indications, l'IC a forgé des solutions qui lui sont propres. Elle s'est pour cela largement inspirée de disciplines proches en fonction de la source de connaissances considérée : ces disciplines ont couvert successivement la psychologie cognitive puis l'ergonomie puis la terminologie et la linguistique de corpus.

3.3.1 L'expertise humaine comme source de connaissances

La construction de modèles nécessite d'accéder aux connaissances fournies par les différentes sources. Les techniques d'accès diffèrent selon les types de sources, certaines allant jusqu'à faire émerger des connaissances nouvelles, non explicitées. *Technique* fait ici référence à des « modes opératoires » préconisant des modes de choix ou de création de situations de production ou d'utilisation de connaissances, puis la manière de repérer-recueillir-extraire ou analyser ces données et enfin des propositions pour interpréter, dépouiller, structurer les fruits de cette analyse.

Concernant l'expertise humaine, les approches ont évolué d'une vue *cognitiviste*, faisant l'hypothèse d'une passerelle possible entre des représentations chez les individus aux représentations informatiques, à des approches *constructivistes* puis de cognition située, prenant en compte la dimension contextuelle et parfois collective des savoirs. Dans le

premier cas, il s'agit de la localiser, puis de rendre explicite des savoir-faire et de les représenter. Cette vue, qui correspond historiquement à celles des systèmes experts, considère que les connaissances sont accessibles chez un ou plusieurs experts et qu'il suffit de les expliciter pour construire un système produisant les mêmes raisonnements.

Cette vue *cognitiviste* a été progressivement remise en question pour mieux répondre au caractère situé des connaissances. Les savoir-faire n'étant accessibles qu'à travers leur mise en œuvre en situation de résolution de problème ou de décision, l'IC a repris à l'ergonomie des techniques d'analyse de la tâche et de l'activité. Des panoramas de ces techniques sont présentés dans Aussenac (1989), Shadbolt *et al.* (1999) ou Dieng-Kuntz *et al.* (2005). On y distingue les méthodes dites directes qui consistent à interroger l'expert, à le faire s'exprimer oralement de manière plus ou moins guidée, des méthodes indirectes, comme l'analyse de « grilles répertoires », qui se fondent sur une interprétation des éléments recueillis alors que l'expert exécute des tâches faisant appel à son expertise mais incluses dans son activité.

Un résultat important de ces travaux a été de poser les bases de l'acquisition des connaissances comme champ disciplinaire s'intéressant aux connaissances *pour elles-mêmes* avant de considérer leur formalisation et leur exploitation dans un système donné. L'adoption d'un point de vue *constructiviste* et la prise en compte de méthodes établies en génie logiciel ont ensuite conduit à des propositions méthodologiques guidant l'ensemble du processus d'acquisition des connaissances. Plusieurs des méthodes ainsi définies dans des projets d'envergure, essentiellement européens, seront présentées en 3.3.3.

3.3.2 Les documents textuels comme sources de connaissances

Atouts de l'analyse de documents textuels

L'utilisation de documents textuels en tant que source de connaissances pose le problème de leur sélection et de leur analyse, qu'il s'agisse de documents techniques, de documents liés à une activité ou à tout un domaine d'application. L'analyse de documents peut porter sur le langage naturel formant le texte ou sur la structure de ces textes, explicitée sur papier ou écran par une mise en forme matérielle, et électroniquement par un étiquetage (Virbel, 2001). On parle dans ces derniers cas de documents structurés ou semi-structurés – *e.g.* documents XML.

Les textes sont des ressources très riches en connaissances. L'analyse de textes a ainsi toujours été présente en IC mais la manière de l'aborder a radicalement changé après 1990. On ne cherche plus à reconstituer automatiquement les modes de compréhension d'un texte par un individu (Aussenac-Gilles *et al.*, 1995). L'analyse de textes doit son essor aux avancées du traitement automatique du langage (TAL) naturel écrit, qui ont débouché sur des logiciels d'analyse robuste et spécialisés. Cette maturité du TAL a été concomitante au déploiement des ontologies. L'analyse de textes écrit a alors trouvé un champ d'application privilégié dans la construction d'ontologies et leur utilisation pour l'annotation sémantique de documents. L'hypothèse forte derrière l'exploitation automatique de textes est qu'ils fournissent les éléments stables, consensuels et partagés d'un domaine (Bourigault et Slodzian, 1999 ; Condamines, 2003). Or ce n'est pas toujours le cas, et deux éléments clés conditionnent l'obtention d'un modèle de qualité : tout d'abord, la constitution d'un corpus pertinent en amont du processus ; ensuite, une contribution régulière de spécialistes du domaine ou de modélisation pour assurer l'interprétation des résultats. L'analyse des textes est aussi utilisée pour la construction de ressources proches des ontologies telles que les thesaurus, index, lexiques ou bases de connaissances terminologiques.

Techniques et outils pour l'analyse de textes

Les approches dites *linguistiques* s'appuient sur les formulations présentes dans les textes pour repérer des contextes riches en connaissances. Il peut s'agir de groupes nominaux ou verbaux ayant une cohérence forte, et dont l'usage laisse penser qu'ils sont des termes désignant des concepts du domaine, ou exprimant des relations entre concepts. Il peut s'agir aussi d'indices plus minces mettant en relation des éléments plus diffus et pour lesquels l'analyste humain doit reconstruire les liens de référence avant d'aboutir à des éléments de connaissance, des axiomes ou des règles. Les approches *statistiques* traitent le texte dans sa globalité et tirent profit des redondances, des régularités d'usage, des co-occurrences pour identifier des expressions figées et des termes, mais aussi des mots ou groupes de mots (clusters) ayant des comportements ou contextes linguistiques analogues.

Dans les deux cas, des analyses préalables du texte, comme le découpage des textes en phrases puis en mots (*token*) ou l'analyse grammaticale de ces mots, sont nécessaires. Plus ces traitements sont sophistiqués (comme l'analyse syntaxique complète de phrases), plus on peut définir des règles précises d'interprétation automatique. Malheureusement, la plupart des logiciels réalisant des analyses sophistiquées sont peu robustes et souvent disponibles pour quelques langues seulement, l'anglais étant très privilégié. De plus, ces analyses nécessitent parfois des ressources (lexiques, dictionnaires sémantiques, etc.) qui sont rarement disponibles dans toutes les langues.

En matière de construction d'ontologies, l'analyse de texte répond à deux types de besoins (Maedche, 2002) : l'identification des concepts et de leurs propriétés ou relations, tâche appelée « ontology learning » (Buitelaar *et al*., 2005) ; la recherche d'instances de concepts et de relations dans les textes, ou « ontology population ». La modélisation de vocabulaires a motivé la réalisation de logiciels spécialisés débouchant sur des résultats de plus haut niveau qui apportent des briques à intégrer dans un modèle : extracteurs de termes – Terminoweb (Barrière et Akakpo, 2006), Syntex-Upery (Bourigault, 1992), TermExtractor (Drouin, 2003), extracteurs de relations – Caméléon (Aussenac-Gilles *et al.*, 2008), Nooj[10], RelExt (Schutz et Buitelaar, 2005), langages de patrons, extracteurs d'entités nommées (Poibeau et Kosseim, 2001) et de termes recherche d'instances ou de relations entre instances (comme avec la plate-forme KIM[11]). La construction de modèles à partir de textes a également bénéficié de plateformes de TAL (GATE[12], Linguastream[13], UIMA[14]) qui permettent de développer des chaînes de traitement spécialisées pour la construction d'ontologies.

3.3.3 Plateformes de modélisation

Les plateformes de modélisation intègrent l'accès à des sources de connaissances, ou à leurs traces, des techniques et logiciels des types que nous venons de présenter ainsi que des techniques et des langages de modélisation. Elles implémentent une méthodologie en définissant une chaîne de traitements qui guide le processus de modélisation. Dans la suite, nous présenterons en premier lieu les travaux les plus significatifs en matière de modélisation du raisonnement, résultats marquants des années 90, puis nous nous

[10] url nooj
[11] http://www.ontotext.com/kim/
[12] http://gate.ac.uk/
[13] http://linguastream.org/
[14] http://domino.research.ibm.com/comm/research_projects.nsf/pages/uima.index.html

focaliserons sur les méthodes et plateformes de construction d'ontologies qui sont d'actualité depuis une dizaine d'années.

Méthodes de modélisation du raisonnement

Des guides méthodologiques ont été établis pour mieux maîtriser le développement de gros projets de SBC. Ces méthodes ressemblent dans leur principe à celles qui existent en génie logiciel car elles accordent beaucoup d'importance à la modélisation. Dans les deux cas, il s'agit de gérer les cycles de développement et de construire un ou des modèles du système à concevoir.

Ces travaux ont fait évoluer la notion de modèle conceptuel, en accordant une place importante au modèle de raisonnement, et ont renouvelé les langages associés, en articulant les notions d'inférence et de tâche. Du point de vue méthodologique, ces recherches ont montré en quoi les primitives de modélisation fournissent une grille pour le recueil et l'interprétation de connaissances, et en cela guident leur modélisation. De ces travaux, en particulier les résultats sur les Tâches Génériques de Chandrasekaran (1983), a émergé l'intérêt de disposer d'éléments de modèles génériques, et réutilisables par instanciation à une application particulière, puis avec la méthode CommonKADS que ces éléments soient adaptables et modulaires.

Faisant suite aux travaux sur les Tâches Génériques et sur les méthodes à limitation de rôles (Marcus et McDermott, 1989), et à partir des propositions de L. Steels dans l'approche componentielle COMMET (et dans l'atelier KREST) (Steels, 1990), plusieurs travaux ont distingué explicitement les concepts de tâche et de méthode. Cette distinction présente l'avantage de décrire séparément le but visé de la façon de l'atteindre et rend possible la définition explicite de différentes façons d'atteindre un même but par l'association de plusieurs méthodes à une même tâche.

Ces différents travaux ont débouché sur la méthode et la plateforme la plus aboutie, CommonKADS (Schreiber *et al.,* 1999). CommonKADS permet la construction de plusieurs modèles liés entre eux, nécessaires à la spécification d'un système à base de connaissances, dont le modèle organisationnel permet de prendre en compte l'utilisation des connaissances dans leur environnement. Le modèle d'expertise du système est désormais reconnu comme nettement différent d'un modèle cognitif de l'expert humain. Ce modèle d'expertise est décrit selon trois points de vue : les tâches, les modèles du domaine, les méthodes (*Cf* 2.1 et 2.2).

Méthodes et plateformes de construction d'ontologies

La construction d'ontologies soulève des problèmes variés :

- définir ce que doit être le contenu de l'ontologie et s'assurer de sa qualité ;
- exploiter efficacement les sources de connaissances disponibles via, par exemple, des processus d'analyse de textes ou de réutilisation d'ontologies existantes ;
- faciliter / assister par des outils la tâche du modélisateur dans le processus de construction ;
- définir un cadre méthodologique et la démarche à mettre en œuvre dans l'enchaînement des différentes tâches.
- des plateformes de gestion d'ontologies offrent des environnements uniformes et cohérents de développement d'ontologies, qui assurent une aide à ces différentes tâches, intègrent pour cela divers outils, et s'appuient sur une méthodologie qui assure l'enchaînement de ces tâches au cours du processus de développement.

Parmi les nombreuses méthodes de construction d'ontologies[15], nous limiterons dans le cadre de cet article à la présentation des méthodes OntoClean, ARCHONTE et OntoSpec, qui toutes trois se focalisent sur la qualité du contenu de l'ontologie.

La méthode OntoClean a été conçue par Guarino et Welty (2004). Les premières idées sont décrites dans une série d'articles publiés en 2000, le terme OntoClean est apparu en 2002. Inspirée de la notion d'ontologie formelle et de principes de philosophie analytique, cette méthode constitue un apport important en tant que première méthode formelle de l'ingénierie ontologique. Elle propose d'analyser les ontologies et de justifier les choix ontologiques en exploitant des méta-propriétés de classes formelles indépendantes de tout domaine d'application, initialement au nombre de quatre (l'identité, l'unité, la rigidité et la dépendance).

La méthode ARCHONTE (ARCHitecture for ONTological Elaborating) élaborée par Bachimont *et al.* (2002) est une méthode ascendante de construction d'ontologies à partir des textes du domaine en trois étapes. Une première étape consiste à choisir les termes pertinents du domaine, et à les normaliser sémantiquement en précisant les relations de similarité et de différences que chacun des concepts entretient avec ses frères et son père (*principe de la sémantique différentielle*). La seconde étape consiste à formaliser les connaissances (*engagement ontologique)*. Il s'agit de construire une *ontologie différentielle* en ajoutant des propriétés ou annotations, et en définissant les domaines et co-domaines des relations. Enfin, une troisième étape consiste à opérationnaliser l'ontologie dans un langage de représentation des connaissances. On obtient alors une *ontologie computationnelle*.

Enfin, OntoSpec (Kassel, 2002) est une méthode de spécification semi-informelle d'ontologies. Sa conception a été motivée par le fait que les définitions en langue naturelle associées aux entités conceptuelles permettent à des utilisateurs de l'ontologie de participer à son élaboration avec un ingénieur de la connaissance. Par ailleurs, cette méthodologie propose un cadre constitué d'une typologie de propriétés pouvant intervenir dans les définitions de concepts et de relations et de règles, pour paraphraser en langue naturelle ces types de propriétés. Ce cadre est un guide pour modéliser et facilite ensuite le passage à une ontologie formelle.

Les plateformes de construction d'ontologie sont en général construites autour d'un éditeur d'ontologies. C'est le cas de Protégé[16], créé à l'université de Stanford, très utilisé pour créer ou modifier des ontologies RDFS ou OWL. A côté de cette fonction d'édition, beaucoup d'autres fonctionnalités peuvent être intégrées dans des plateformes, parmi lesquelles on trouve : des fonctions de traduction de Schema XML, d'aide à l'accès à des modules d'ontologies ou d'aide au partitionnement d'ontologies, des modules de traduction de vocabulaires, des accès à des moteurs de recherche d'ontologies, à des étiqueteurs morpho-syntaxiques (par ex. *Tree-Tagger*[17]), des modules d'aide à la personnalisation d'ontologies, de génération de documentation, d'aide à la gestion de leur évolution, de leur évaluation, des fonctionnalités d'alignement ontologies, des services de raisonnement et d'inférence, des services d'aide à la navigation, de visualisation ou encore des aides à la réutilisation d'ontologies. L'ensemble de ces fonctionnalités correspondent, à titre d'exemples, à des plugins de la plateforme Neon[18].

---

[15] Pour une synthèse des principales méthodologies de construction d'ontologies, se reporter à (Fernandez-Lopez et Gomez-Perez, 2002).

[16] http://protege.stanford.edu/

[17] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

[18] http://www.neon-toolkit.org/wiki/Neon_Plugins#Neon_Plugins

Certaines des plateformes sont spécialisées pour traiter un type de données. Text2Onto (Cimiano et Völker, 2005) rassemble des logiciels de fouille de textes et de génération d'informations structurées à partir de documents faiblement structurés. Pour construire une ontologie, Text2Onto est associé à la plate-forme KAON (Karlsruhe Ontology Management Infrastructure) (Oberle *et al.,* 2004). DaFOE4App (Differential and Formal Ontology Editor for Applications) (Charlet *et al.,* 2010), dont la conception reprend les principes de TERMINAE (Aussenac-Gilles *et al.,* 2000) et d'ARCHONTE (Bachimont *et al.,* 2002), est une plateforme pour construire des ontologies à partir de textes et de thesaurus. Cette plateforme met en avant la dimension linguistique et couvre toutes les étapes allant de l'analyse d'un corpus (annoté au sein d'une plateforme de TAL) à la définition d'une ontologie formelle du domaine. Elle garantit la persistance, la traçabilité et le dimensionnement des modèles (plusieurs milliers de concepts).

3.4 Réutilisation de modèles

Tout comme le génie logiciel vise la réutilisabilité de composants logiciels, l'acquisition des connaissances cherche à favoriser la réutilisation de composants de connaissances. Cette réutilisation peut se concevoir de différentes manières.

Initialement proposée dans le cadre du projet KADS-I, la réutilisation de modèles de raisonnement consiste à reprendre et adapter à la réalisation de tâches spécifiques des modèles de tâches exprimés dans une terminologie indépendante de tout domaine d'application. Cette approche est séduisante mais l'adaptation d'un modèle de raisonnement à un domaine pose deux types de problèmes. D'une part, une application met souvent en œuvre plusieurs raisonnements auxquels correspondent plusieurs modèles qu'il faut pouvoir distinguer et combiner. D'autre part, la réutilisation et l'adaptation de modèles génériques prédéfinis à une application donnée s'avère un travail difficile et long à réaliser. Une des raisons en est que la tâche à réaliser et la base de connaissances du système doivent tous deux être exprimés dans les termes d'un domaine d'application donné, alors que les méthodes réutilisables, issues de bibliothèques, sont exprimées à l'aide d'un vocabulaire générique. Ainsi, adapter des éléments de résolution de problèmes à une application est d'abord un problème de mise en correspondance de termes. Ce constat a alors ouvert la voie à des approches plus flexibles, au sein desquelles les éléments réutilisés et adaptés sont de granularité plus fine. Il ne s'agit plus de réutiliser des modèles génériques complets de raisonnement mais des éléments de raisonnement.

Les besoins en réutilisation de modèles de connaissances du domaine ont conduit à définir la notion d'ontologie. Ces représentations structurées qui définissent les concepts d'un domaine se sont multipliées. Leur réutilisation d'une application à une autre est une des motivations de leur existence. La réutilisation d'ontologie vise à réduire les difficultés de leur développement ex-nihilo, apparues comme un élément de blocage pour certaines applications. Cette réutilisation pose différents types de problèmes : la sélection des ontologies à réutiliser, l'aide à la réutilisation d'ontologies volumineuses et difficilement compréhensibles, et enfin l'intégration. Nous abordons le second, le plus difficile à résoudre.

Pour réduire les difficultés d'adaptation d'ontologies réutilisées ou de construction d'ontologies volumineuses, la notion de patterns de connaissances, directement issue des *design patterns* utilisés en génie logiciel, a été introduite dans le domaine de l'ingénierie ontologique par (Clark *et al.*, 2000) puis dans les travaux sur le Web sémantique par (Gangemi *et al.*, 2004), (Rector *et al.*, 2004) et (Svatek, 2004). Les patterns de

connaissances sont des représentations récurrentes et partagées de connaissances explicitement représentées comme des modèles généraux, ré-utilisables après transformation (par renommage symbolique) pour créer des représentations spécifiques. Ils peuvent être utiles pour construire des ontologies plus rapidement en aboutissant à un résultat de meilleure qualité, en résolvant, par exemple, des problèmes de conception indépendamment de toute conceptualisation comme l'a proposé le groupe de travail du W3C « Semantic Web Best Practices and Deployment »[19]. Les patterns peuvent aussi faciliter l'application de bonnes pratiques (Pan *et al.*, 2007) ou encore guider la résolution de problèmes de contenu (Gangemi, 2005). Une bibliothèque de patterns de connaissances a été proposée dans le cadre du projet européen NeOn[20] qui distingue les patterns structurels, des patterns de correspondance, de contenu, de raisonnement, de présentation et lexico-syntaxiques (Presutti *et al.*, 2008) ainsi qu'une méthodologie de conception d'ontologies à base de patterns, eXtreme Design (XD) (Daga *et al.*, 2010). Enfin, dans les dernières années, beaucoup de travaux de recherche ont porté sur la conception d'outils d'alignement d'ontologies (Euzenat *et al.*, 2007).

3.5 Représentation des connaissances dans les modèles

L'IC n'a pas pour principal objectif de construire des langages de représentation des connaissances mais puisque les chercheurs spécifient des connaissances ou des modèles, ils ont souvent participé et participent encore à l'élaboration ou l'évolution de langages au sein de groupes de normalisation, comme le W3C. Comme pour la modélisation, les langages de représentation des connaissances ont d'abord été liés aux modèles de raisonnement puis se sont intéressés aux ontologies (*Cf.* 2, 2.1, 2.2) pour revenir maintenant aux raisonnements.

Si on s'intéresse aux langages de représentation des ontologies, on peut noter, dans les années 1980, le succès de la logique d'une part, et du langage des graphes conceptuels d'autre part. Les graphes conceptuels proposent à la fois une formalisation logique et une symbolique graphique (intéressante à l'époque où on n'avait pas d'IHM puissantes pour afficher des réseaux sémantiques ou des arbres à déployer et refermer à la demande). OWL s'est ensuite imposé : langage résultant de la fusion des travaux des projets DAML et OIL, défini comme une sur-couche de XML, il s'est stabilisé en 3 langages, OWL Lite, OWL-DL, OWL-full dans le cadre des travaux du W3C. La différence entre chacun de ces langages découle de choix de représentativité *versus* calculabilité. Au-dessus, on a les langages liés à l'expression de règles d'inférences. La question du choix, en termes de standards, des langages pour exprimer ces règle n'est pas encore tranchée et il existe un certain nombre de candidats parmi les langages de règles permettant de prendre en compte les ontologies comme SWRL[21], Description Logic Programs (DLP)[22] (Hitzler *et al.*, 2005), le *Rule Interchange Format* (RIF)[23] ou le langage de requêtes SPARQL[24].

Il existe d'autres langages que ceux repérés ici. Récemment, le langage SKOS (pour Simple Knowledge Organisation System) a été défini pour représenter des modèles peu

[19] http : /www.w3.org/2001/sw/BestPractices/
[20] http://www.neon-project.org
[21] http://www.w3.org/Submission/SWRL/
[22] http://logic.aifb.uni-karlsruhe.de/wiki/DLP.
[23] http://www.w3.org/2005/rules/wiki/RIFWorkingGroup
[24] http://www.w3.org/TR/2008/REC-rdf-sparql-query-20080115/

formels pour lesquels on ne veut justement pas inférer de conséquences logiques précises comme pour les ontologies. SKOS permet de représenter en particulier la relation de généralisation versus spécialisation (broader-than et narrower-than – BT/NT), très utilisée dans les thésaurus sans imposer les inférences logiques que OWL impose sur la relation de subsomption. C'est d'autant plus nécessaire que les inférences logiques sur la relation de subsomption ne sont valides que sur les ontologies, construites pour respecter ces contraintes, et pas sur les thésaurus.

Par ailleurs, il est intéressant de noter que les applications utilisant des thésaurus et les ontologies sont arrivés à un état de maturité tels que ces différentes ressources – *i.e.* thésaurus et ontologies – sont elles-mêmes impliquées dans des processus de développement qui amènent à utiliser les différents langages de représentation des connaissances à des étapes différentes du processus et pas toujours de la façon prévue par les concepteurs de ces langages. Ainsi un thésaurus et une ontologie utilisés conjointement dans une application vont être modélisés en OWL pour cette application, mais ils seront développés, l'un en SKOS, l'autre en OWL.

4 Enjeux méthodologiques et applicatifs actuels

Les enjeux actuels de l'IC sont à la fois d'ordre méthodologique et applicatif. Pour y répondre, un certain nombre de principes, qui ont fondé la discipline, fournissent un cadre :

- la nécessité d'une démarche pluridisciplinaire prenant en compte les analyses d'autres disciplines telles que la psychologie cognitive, l'ergonomie, les sciences des organisations, la linguistique, le traitement automatique des langues, la recherche d'information, la gestion documentaire ;
- le besoin de penser l'ergonomie des systèmes en amont de la construction du projet, en particulier, prendre en compte les usages des systèmes construits et leur intégration dans le système de traitement de l'information dans lequel ils s'insèrent ;
- l'intérêt d'une démarche de modélisation approfondie, faisant cohabiter des modèles différents à chaque temps du processus d'élaboration du système opérationnel.

Les applications liées à l'IC constituent alors un vaste champ de recherche, d'expérimentation et de valorisation dans lequel doivent se développer des méthodologies innovantes. C'est cette articulation entre méthodologie et applications qui sera le fil conducteur des enjeux décrits ci-dessous.

4.1 Articuler la langue, les connaissances et leur support

La langue naturelle est le vecteur privilégié de l'expression des connaissances. Il est donc logique que l'IC se préoccupe de connaissances exprimées en langue naturelle (expression orale ou écrite) et qu'elle entretienne, de ce point de vue, des liens étroits avec les domaines de la linguistique, du Traitement automatique de la langue naturelle (TALN) d'un côté et de la Recherche d'Information (R.I.) de l'autre (*Cf.* 3.3.2).

Construction de modèles de raisonnement et d'ontologies

L'IC s'intéresse aux documents comme porteurs de sens et révélateurs de connaissances depuis les premières études sur l'acquisition des connaissances pour les systèmes experts (années 90). Par la suite, sous l'impulsion du groupe TIA dans les années 90, on a pu mettre

en avant l'utilité de corpus textuels générés durant une activité pour aider à construire l'ontologie de cette même activité, de ce domaine. Dans cette perspective, les documents d'un corpus sont considérés comme une source de connaissances complémentaire ou alternative aux experts et spécialistes du domaine. Le traitement de ces sources de connaissances passe d'abord par des outils de TALN mais aussi par des plateformes aptes à utiliser le résultat de ces outils pour construire les ontologies ou plus largement des terminologies.

Textes et documents en ingénierie des connaissances

Dans cette perspective, le document en tant que tel est central, il est un élément support de connaissance à part entière. La gestion des documents produits et utilisés au sein de l'activité individuelle et collective étudiée, mais aussi, en tant que telle, la gestion de fonds documentaires (images, sons, vidéos) intéresse alors l'IC. Ces applications font appel aux technologies relevant de la gestion documentaire et permettant le partage, la diffusion, l'archivage, l'indexation, la structuration ou la classification de documents ou de flux de documents. Parce que de plus en plus de projets d'IC intègrent la gestion de documents sous des formes très variées, les chercheurs du domaine ne peuvent s'affranchir d'une réflexion approfondie sur la notion de document, et particulièrement de document numérique. Ainsi, plusieurs chercheurs contribuent aux travaux du réseau thématique pluridisciplinaire sur le document (RTP-DOC) et à ses productions (Pédauque, 2003 ; Pédauque, 2005).

Recherche d'information avec des ontologies

Sous l'impulsion du Web sémantique qui, dans ses premiers attendus, utilisait les ontologies comme source de métadonnées pour indexer des documents, les ontologies sont arrivées au cœur d'applications de R.I. Cela amène des réflexions sur des ontologies utilisées spécifiquement pour la R.I. et ayant une composante linguistique forte, à minima des termes associés aux concepts. On parle alors de ressource termino-ontologique ou RTO (Reymonet *et al.*, 2007). Une dernière problématique de ce champ est la mise en œuvre d'environnements applicatifs où l'on fait cohabiter les ontologies avec les thésaurus (Vandenbussche et Charlet, 2009).

4.2 Faire face à l'explosion des données

Aujourd'hui nous assistons à une explosion des données disponibles. Les applications traitent des données de plus en plus nombreuses et diverses suscitant des besoins nouveaux pour les décrire et les intégrer. La description de ces données très nombreuses passe par l'élaboration de modèles au sein desquels la quantité d'information à prendre en compte peut être très importante. L'enjeu aujourd'hui est alors d'être capable de construire des modèles de très grande taille, en diminuant par exemple la quantité d'informations à prendre en compte simultanément. C'est ainsi que des travaux sur la modularité ont vu le jour, qui, appliqués à l'ingénierie des ontologies, visent la construction d'ontologies de très grande taille nécessaires aux applications. Citons, à titre d'exemple, les travaux sur la construction d'ontologies modulaires (Stuckenschmidt *et al.*, 2009) réalisés dans le cadre du projet Knowledge Web (2004-2007).

Gérer une grande masse de données dans un contexte distribué peut aussi nécessiter de prendre appui sur un ensemble d'ontologies existantes nécessitant d'être remaniées, alignées, transformées sous forme de modules ou intégrées avec des ressources non

ontologiques telles que des bases de données, des folksonomies ou des thesauri. Le projet NeOn (2006-2010) a eu pour objectif de proposer une méthodologie de construction de telles ontologies dites en réseau (Gomez-Perez *et al.,* 2009), incluant un support au développement collaboratif et la prise en compte de l'aspect dynamique et évolutif de ces ontologies. Il s'agit d'un enjeu important pour le développement de grosses applications à base d'ontologies.

4.3 Gérer l'intégration des connaissances par les ontologies

Concernant le problème de l'intégration de données diverses, une approche de plus en plus privilégiée, autant dans le domaine des bases de données que dans celui de la recherche d'information, est d'appuyer l'intégration des données sur une ontologie du domaine concerné. Les ontologies jouent en effet un rôle clé en intégration de sources multiples et hétérogènes. Elles peuvent aider à comprendre et interpréter des descriptions hétérogènes de contenus relatifs à un même domaine pour pouvoir ensuite plus facilement les mettre en relation (Asselé *et al.*, 2010).

Par ailleurs, des ontologies existent mais sont de très grande taille et donc difficiles à exploiter. C'est le cas, par exemple, des domaines de l'agronomie ou de la médecine pour lesquels existent des ontologies de plusieurs milliers de concepts. Dans ce cas, l'enjeu consiste à permettre de comprendre le contenu de ces ontologies pour aider à en extraire le sous-ensemble pertinent pour une application. La prolifération d'ontologies est la conséquence d'expériences d'usages et de réutilisation d'ontologies qui ont montré qu'elles ne représentent correctement et de façon consensuelle que des domaines réduits. En suivant ces travaux (Rosenbloom *et al.*, 2006), on parle d'*ontologies de référence* avec des visées de représentations larges et d'*ontologies d'interface* développées pour des applications spécifiques. Entre les 2 types d'ontologies, on a besoin de services d'alignements et, comme précédemment, de la possibilité d'extraire le sous-ensemble de l'ontologie de référence pertinent pour une application. C'est ce que permet un standard comme CTS2[25].

4.4 Tirer parti des nouvelles sources de connaissances

Nous discuterons deux sources qui constituent aujourd'hui des enjeux majeurs : le Web 2.0 et le Web des données liées.

Le Web 2.0 ou Web social (O'Reilly, 2005) accorde une place plus importante aux utilisateurs que le Web dans sa version initiale en leur permettant non seulement d'accéder en lecture à des pages Web mais également en écriture. L'internaute devient actif. Auteur et acteur, il peut utiliser le Web pour élaborer ses propres contenus et les partager. Ceci est rendu possible par des outils tels que les blogs, les réseaux sociaux, les outils collaboratifs, les plateformes de mise en relation, les services de partage en ligne. Ces outils et ces services sont de plus en plus utilisés dans les organisations (Lewkowicz *et al.*, 2001). Le contenu créé par un utilisateur ou dont il est propriétaire peut être exploité. Mais les applications qui gèrent ces contenus ont leur propre format de données et celles-ci sont de plus en plus distribuées et hétérogènes, ce qui pose des problèmes importants d'intégration d'information. De la même façon, le *taggage*[26], une pratique souvent utilisée pour

[25] http://www.3mtcs.com/resources/hl7cts

[26] On décrit ici du fait d'indexer un contenu avec des méta-données choisies par l'utilisateur. On parle alors de *foksonomies*.

regrouper des contenus jugés similaires et en facilité la recherche, présente des limites du fait de l'ambigüité et de l'hétérogénéité des *tags*. Ainsi, les applications Enterprise 2.0 (McAfee, 2006), qui tendent à se développer de plus en plus, constituent un terrain d'expérimentation et de valorisation des techniques de l'IC, tout en permettant au domaine de se renouveler en faisant des propositions nouvelles pour faciliter la navigation, l'interrogation ou l'extraction. Il s'agit, pour l'IC, d'un enjeu applicatif d'une importance majeure.

Inventé par Tim Berners-Lee, le Web des données liées[27] réfère à un style de publication et d'interconnexion des données structurées sur le Web basé sur le modèle RDF. Les données liées ont l'avantage de fournir un mécanisme simple d'accès unique et normalisé au lieu de s'appuyer sur différents formats d'interface et de résultat. Les sources de données peuvent ainsi être plus facilement explorées par les moteurs de recherche, elles peuvent être accessibles à l'aide de navigateurs génériques de données, elles peuvent avoir des liens avec des sources de données différentes. Le nombre de données publiées selon les principes des données liées croît rapidement (on parle de milliards de triplets RDF disponibles sur Internet). Du fait de l'utilisation d'un vocabulaire non ambigu et relié, cette masse de données représente une source prometteuse que l'IC doit absolument considérer.

4.5 Évaluer la qualité des modèles

Enfin, une question fondamentale pour l'IC concerne l'évaluation de la qualité des modèles utilisés et des résultats produits. L'exploitation de connaissances de qualité médiocre conduit à des erreurs, des doublons, des incohérences qu'il faut éviter. Le thème de la qualité est devenu ces dernières années un des sujets d'intérêt à la fois émergeant dans le domaine de la recherche et critique dans les entreprises.

La qualité des modèles/ontologies peut être assurée de manière méthodologique, lorsque l'ontologie a été construite selon une méthode rigoureuse s'appuyant sur les fondements théoriques et philosophiques de ce qu'est une ontologie. Nous avons cité en partie 3.3 les principes de structuration ontologiques basés sur des critères différentiels de la méthode ARCHONTE, ou encore les méta-propriétés formelles de la méthode Ontoclean ; OntoSpec permettant d'appliquer ces méta-propriétés au fur et à mesure de la spécification de concepts.

D'autres travaux méthodologiques visent à passer de démarches « artisanales », essentiellement manuelles, dont le coût et la durée sont difficiles à estimer, à des approches plus systématiques, outillées et mieux maîtrisées. Des propositions méthodologiques mettent ainsi l'accent sur la réutilisation comme Methontology (Gomez-Perez *et al.*, 2004), sur des guides pratiques (Fridman-Noy et Hafner, 1997) ou sur l'analyse systématique de textes à l'aide de logiciels de TAL comme TERMINAE et les méthodes répertoriées dans (Maedche, 2002).

5 Conclusion

L'IC est un champ de recherche qui s'est beaucoup transformé ces dernières années. Pendant longtemps, ce domaine s'est intéressé à la production de modèles de connaissances

[27] Tim Berners-Lee parle de *Web of Data*. Il promeut un projet qui va dans ce sens, qui est le *Linked Open Data* (*LOD*).

selon un processus bien structuré sous le contrôle d'ingénieurs de la Connaissance. Les modèles élaborés, généralement complexes, étaient utilisés dans des applications spécifiques. Aujourd'hui, nous assistons à une transformation de l'IC. Les applications dans lesquelles des connaissances sont utilisées comme support à un raisonnement ou une activité se sont beaucoup diversifiées. Depuis 2000, elles concernent la gestion des connaissances au sens large, la recherche d'information sémantique, l'aide à la navigation, l'aide à la décision, beaucoup d'applications relevant également du domaine du web Sémantique. Cet élargissement se poursuit et de nouveaux champs d'application émergent encore, posant les problèmes de l'IC en des termes nouveaux.

La dernière décennie a connu une transformation majeure dans la façon dont les individus interagissent et échangent. L'information est dorénavant coproduite, partagée, classée et évaluée sur le Web par des milliers de personnes. Ces usages et les technologies sous-jacentes sont connus sous le nom de Web 2.0. L'ingénierie et la gestion des connaissances dans le contexte de communautés d'intérêt ou de pratiques dont l'émergence spontanée et l'activité sont permises par ces évolutions du Web (Web 2.0, Web social) sont des enjeux majeurs de la future décennie.

Enfin, à l'ère du Web Sémantique, un nouveau paradigme prometteur apparaît, celui du Web des données. Le Web des données, qui fait suite au Web des documents, entend faire face au déluge informationnel en connectant les données. Pour favoriser l'émergence de ce Web des données, le W3C a promu deux initiatives facilitant l'exploitation de données structurées : le Web Sémantique et les données liées (Linked Open Data). Les technologies du Web Sémantique fournissent un environnement de description des données, d'interrogation, de raisonnement. L'initiative des données liées a pour objectif de construire un réseau global des données en liant des données provenant d'horizons divers, permettant ainsi, à partir d'une donnée, d'obtenir l'ensemble des données qui en découlent. L'interconnexion des données qui permet de leur donner du sens sera une source importante d'innovation puisqu'elle augmentera la pertinence des contenus et aboutira à la constitution d'une base de données à l'échelle du Web. De nombreuses sources de données existent déjà, parmi lesquelles trône DBpedia qui structure le contenu de Wikipedia en triplets RDF de façon à rendre les informations de l'encyclopédie réutilisables. DBpedia est une source très puissante car elle est interconnectée avec d'autres sources de données, Geonames, MusicBrainz, etc.. L'IC se doit d'alimenter tous ces développements nouveaux, nécessitant le recours à des ontologies, leur alignement, la définition de techniques et de langages, tels RDFa, d'enrichissement des données par des métadonnées exploitées en cas de réutilisation, etc.

L'IC a vécu des changements d'orientation successifs. Les nouveaux champs d'application cités ci-dessus viennent renforcer ce caractère mouvant. C'est un domaine en constante évolution de l'intérieur – nouvelles analyses, nouvelles perspectives, manière originale de poser les problèmes, nouveaux concepts théoriques – et de l'extérieur – les types d'applications ciblés ayant changé au fil des années, les contributions d'autres disciplines viennent apporter des méthodes et des concepts nouveaux.

Références

Asselé Kama A., M. Giovanni, R. Choquet, J. Charlet et M.-C. Jaulent (2010). Une approche ontologique pour l'exploitation de données cliniques, *Journées Francophones d'Ingénierie des Connaissances (IC)*, Nîmes.

Aussenac, N. (1989). Conception d'une méthodologie et d'un outil d'acquisition de connaissances expertes. *Thèse de doctorat*, IRIT, Université Paul Sabatier, Toulouse.

Aussenac-Gilles, N., J.P. Krivine et J. Sallantin (1992). Acquisition des connaissances. *Revue d'Intelligence Artificielle (RIA)*, Hermès, Paris, Vol. 6, n° 2, 7-18.

Aussenac-Gilles N., D. Bourigault, A. Condamines et C. Gros (1995). How can knowledge acquisition benefit from terminology? *Proceedings of the 9th Knowledge Acquisition Workshop*. Banff, Univ. of Calgary (CA).

Aussenac-Gilles N., P. Laublet et C. Reynaud (1996). Acquisition et Ingénierie des Connaissances – tendances actuelles, *Cepaduès editions*, Toulouse.

Aussenac-Gilles N., B. Biebow et S. Sulzman (2000). Revisiting Ontology Design : a methodology based on corpus analysis. In R. Dieng et O. Corby (Eds.), *EKAW*, LNAI 1937, Springer-Verlag, 172-188.

Bachimont B., A. Isaac et R. Troncy (2002). Semantic Commitment for Designing Ontologies : a proposal. *EKAW*, Springer-Verlag, 114-121.

Bourigault D. et M. Slodzian (1999). Pour une terminologie textuelle. *Terminologies Nouvelles*, 19 : 29-32.

Bourigault D., N. Aussenac-Gilles et J. Charlet (2004). Construction de ressources terminologiques ou ontologiques à partir de textes : un cadre unificateur pour trois études de cas. *Revue d'Intelligence Artificielle (RIA)*, Numéro spécial sur les techniques informatiques de structuration de terminologies, M. Slodzian (Ed.), Hermès, Paris, Vol. 18, n°1, 87-110.

Chandrasekaran B. (1983). Towards a taxinomy of problem solving types. *The AI Magazine*, 4(1), 9-17.

Charlet J., M. Zacklad, G. Kassel et D. Bourigault (2000). *Ingénierie des connaissances : Evolutions récentes et nouveaux défis*, Eyrolles, Paris.

Charlet, J., S. Szulman, N. Aussenac-Gilles, A. Nazarenko, N. Hernandez, N. Nadah, E. Sardet, J. Delahousse, V. Teguiak et A. Baneyx (2010). DaFOE : une plateforme pour construire des ontologies à partir de textes et de thésaurus. *EGC*, 631-632

Cimiano P. et J. Völker (2005). Text2Onto, *NLDB*, 227-238.

Clark, P., J. Thompson et B. Porter (2000). Knowledge Patterns. In Anthony G. Cohn, F. Giunchiglia, and Bart Selman, eds. *KR2000: principles and Knowledge Representation and reasoning*, San Francisco, Morgan Kaufmann, 591-600.

Condamines A. (2003). Sémantique et corpus spécialisés : Constitution de bases de connaissances terminologiques. Mémoire d'habilitation à diriger des recherches en Linguistique de l'université de Toulouse 2. *Carnets de grammaire de l'ERSS* N°13 - Octobre.

Cordier M.O. et C. Reynaud (1991). Knowledge Acquisition Techniques and Second-Generation Expert Systems, A*pplied Artificial Intelligence (AAI)*, 5(3), 209-226.

Daga, E., E. Blomqvist, A. Gangemi, E. Montiel, N. Nikitina, V. Presutti et B. Villazon-Terrazas (2010). *NeOn D2.5.2. Pattern-based ontology design: methodology and software support, NeOn project*.

Dieng-Kuntz R., O. Corby, F. Gandon, A. Giboin, J. Golebiowska , N. Matta et M. Ribière, (2005). *Knowledge management : Méthodes et outils pour la gestion des connaissances*. Dunod 3e édition. 450 p.

Fernandez-Lopez M. et A. Gomez-Perez (2002). Overview and Analysis of Methodologies for building Ontologies, *Knowledge Engineering Review (KER)*, 17(2), 129-156.

Fridman-Noy, N. et C. Hafner (1997). The state of the art in ontology design: a survey and comparative review. *Artificial Intelligence Magazine*, 53-74.

Gangemi, A. (2005) Ontology design patterns for Semantic Web Content. Musen, M.A. et al. (eds.), *ISWC 2005*. LNCS, vol. 3729, Springer. Heidelberg, 262-276.

Gangemi, A., C. Catenacci et M. Battaglia (2004). Inflammation ontology design pattern: an exercise in building a core biomedical ontology with descriptions and situations. In Domenico Maria Pisanelli, (eds), *Ontologies in Medecine*. IOS Press. Amsterdam.

Gomez-Perez, A., M. Fernando Lopez et O. Corcho (2004). *Ontological Engineering: with examples from the area of Knowledge Management, e-commerce and the Semantic Web*, Springer-Verlag, London.

Gomez-Perez, A. et M.C. Suarez-Figueroa (2009). Scenarios for building ontology networks within the NeOn methodology, Poster, *K-CAP*, 183-184.

Gruber T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5, 199–220.

Guarino N. et C. Welty (2004). An overview of OntoClean. In S. Staab et R. Studer (Eds), *The Handbook of Ontologies*, Berlin : Springer-Verlag, 151-172.

Hitzler P., R. Studer et Y. Sure (2005). Description Logic Progams: A practical choice for modelling Ontologies. *1st WS on Formal Ontologies Meet Industry (FOMI)*.

Kassel G. (2002). OntoSpec : une méthode de spécification semi-informelle d'ontologies, *Journées Francophones d'Ingénierie des Connaissances*, Rouen, 75-87.

Klinker G., C. Bhola, G. Dallemagne, D. Marquès et Mc. Dermott (1991). Usable and reusable programming constructs. *Knowledge Acquisition*, 3, 117–136.

Lewkowicz M. et M. Zacklad. (2001). Une nouvelle forme de gestion des connaissances basée sur la structuration des interactions collectives. *Ingénierie et capitalisation des connaissances*, M. Grundstein et M. Zacklad (Eds.) (2001). Série Informatique et systèmes d'information, Hermes Science Europe LTD, Stanmore, 49-64.

Luong P. H. (2007). Gestion de l'évolution d'un web sémantique d'entreprise. *Thèse de doctorat*. Ecoles des Mines de Paris.

Maedche A. (2002). *Ontology learning for the Semantic Web*. Kluwer Academic Publisher.

Marcus S. et J. McDermott (1989). SALT: a knowledge acquisition language for propose and revise systems. *Artificial Intelligenc*e, 39(1), 1-38.

McAfee, A. (2006) Enterprise 2.0: The dawn of Emergent Collaboration, *MIT Sloan Management Review*, Vol. 47, n°3, 21-28.

Neches R., R. Fikes, T. Fini, T. Gruber, R. Patil, T. Senator et W.R. Swartout (1991). Enabling Technology for Knowledge Sharing, *AI Magazine*, Winter 91, 36-56.

Newell A. (1982). The knowledge level. *Artificial Intelligence*, 18, 87–127.

Oberle D., R. Volz, B. Motik et S. Staab (2004). An Extensible Ontology Software Development, *International Handbook on Information Systems*, Staab S. et R. Studer (Eds), Springer, 311-333.

O'Reilly, T. (2005). What is the Web 2.0: Design Patterns and Business Models for the next Generation of Software. http://www.oreillynet.com/lpt/a/6228.

Pan, J.Z., L. Lancieri,, D. Maynard, F. Gandon, R. Cuel, et A. Leger (2007). *Knowledge Web Deleverable* D.1.4.2.v2. Success stories and Best Practices.

Pédauque R. T. (2003). Le document : forme, signe et medium les re-formulations du numérique. STIC-CNRS, http://archivesic.ccsd.cnrs.fr/sic_00000511.html

Pédauque R. T. (2005). Le texte en jeu, permanence et transformations du document, STIC-SHS-CNRS, http://archivesic.ccsd.cnrs.fr/documents/archives0/00/00/14/01/index_fr.html.

Poibeau T. et L. Kosseim (2001). Proper Name Extraction from Non-Journalistic Texts. *Computational Linguistics in the Netherlands* 2000: *Selected Papers from the Eleventh CLIN Meeting*, W. Daelemans, K. Sima'an , J. Veenstra and J. Zavrel (Eds), 144-157..

Presutti, V., A. Gangemi, S. David, G. A. De Cea, M. C. Surez-Figueroa, E. Montiel-Ponsoda, et M. Poveda (2008). *NeOn D2.5.1*. A Library of Ontology Design Patterns: reusable solutions for collaborative design of networked ontologies. NeOn project.

Rector, A. et J. Rogers (2004). Patterns, properties and minimizing commitment: Reconstruction of the galen upper ontology in OWL. In A. Gangemi and S. Borgo, (eds.), *EKAW Workshop Core Ontologies in Ontology Engineering*.

Reymonet A., J. Thomas et N. Aussenac-Gilles (2009). Modélisation de ressources termino-ontologiques en OWL. *Journées Francophones d'Ingénierie des Connaissances (IC 2007), Grenoble (F)*, F. Trichet (Eds.), Cépaduès Editions, 169-180,

Reynaud C., N. Aussenac-Gilles, P. Tchounikine et F. Trichet (1997). The notion of role in conceptual modelling. E*KAW - European Knowledge Acquisition Workshop* - R. Benjamins & E. Plaza Eds, Springer Verlag, Lecture Notes in Artificial Intelligence, 221-236.

Rosenbloom S. T., R. A. Miller et K. B. Johnson (2006). Interface terminologies: facilitating direct entry of clinical data into electronic health record systems, *Journal of the American Medical Informatics*. Association 13(3):277–288.

Schreiber G., B. Wielinga H. Akkermans, W. Van de Velde et A. Anjewierden (1994). Cml: The commonkads conceptual modelling language. In L. Steels , G. Schreiber et W. V. De Velde, Coordinateurs, *EKAW'94*, LNCS, n° 867, 1–25.

Schreiber G., H. Akkermans, A. Anjewierden, R. DeHoog, N. Shadbolt, W. Van de Velde et B. Wielinga (1999). Knowledge Engineering and management: The CommonKADS Methodology, Cambridge, MA: MIT Press.

Steels, L. (1990). Components of expertise. *The Artificial Intelligence Magazine*. 11(2), 28-49.

Stefik, M. (1995). *Introduction to Knowledge Systems*, Morgan Kaufman.

Stojanovic, L. (2004). Methods and tools for ontology evolution. *PhD thesis*, Université de Karlsruhe, Institut AIFB, Karlsruhe.

Stuckenschmidt, H., C. Parent et S. Spaccapietra (2009). *Modular ontologies: Concepts, Theories and Techniques for Knowledge Modularization*, Springer.

Studer R., V. R. Benjamins et D. Fensel (1998). Knowledge Engineering: Principles and Methods, *Data and Knowledge Engineering*, vol. 25, 161-197.

Svatek, V. (2004). Design patterns for semantic web ontologies: Motivation and discussion. *7th conference on Business Information systems*, Poznan.

Tu S. W., H. Erikson, J. H. Gennari, Y. Shahar et M. A. Musen (1995). Ontology-based configuration of problem-solving methods and generation of knowledge-acquisition tools: application of PROTÉGÉ-II to protocol-based decision support. *Artificial Intelligence in Medicine*, 7, 257–289.

Vandenbussche P.-Y. et Charlet J. (2009). Méta-modèle général de description de ressources terminologiques et ontologiques. In Gandon F., coordinateur, *Actes des 20*es *Journées Ingénierie des Connaissances*, pages 193–204, Hammamet, Tunisie, 25-29 mai.

Virbel, J. et C. Luc (2001). Le modèle d'architecture textuelle: fondements et expérimentation. *Verbum*, Vol. XXIII, n°1, 103-123.

Sommaire