

Evolution et maintenance des ressources termino-ontologique : une question à approfondir

Nathalie Aussenac-Gilles*, Anne Condamines† et Florence
Sèdes*

* Institut de Recherches en Informatique de Toulouse
{aussenac, sedes}@irit.fr

† Equipe de Recherche en Syntaxe et Sémantique
anne.condamines@univ-tlse2.fr

Résumé

Ce numéro spécial s'intéresse à l'évolution et à la maintenance des ressources terminologiques ou ontologiques en lien avec l'évolution des textes à partir desquels ou pour lesquels elles sont construites. Cette problématique s'inscrit dans une réflexion de longue haleine sur les liens entre textes et connaissance dans un contexte de traitements automatisés, réflexion dont les principaux éléments sont rappelés dans la présente introduction.

Mots-clés : Modélisation de connaissances, terminologie, traitement du langage naturel, linguistique de corpus, maintenance.

Abstract

This special issue raises the question of the maintenance and the evolution over time of terminological or ontological resources, in relation with the evolution of the texts from which or for which they have been designed. This problem results from a long term reflection about the connexion between texts and knowledge in the context of their automatic processing. This introduction reports the core ideas of this reflection.

Key words: Knowledge modelling, Terminology, Natural Language Processing, Corpus Linguistics, Maintenance.

1. INTRODUCTION

De longue date, de nombreuses disciplines (sciences de l'information, terminologie et, plus récemment, traitement automatique des langues et ingénierie des connaissances) ont constitué et utilisé des ressources lexicales, renvoyant aux vocabulaires et même aux connaissances de domaines particuliers, pour répondre à des besoins précis. Ces ressources vont des thésaurus et langages documentaires utilisés en sciences de l'information aux ontologies manipulées par l'ingénierie des connaissances, en passant par les terminologies et autres bases de données lexicales, terminologiques ou sémantiques servant à l'étude des langues, au traitement informatique de l'écrit ou encore aidant à la traduction. La plupart du temps, ces ressources se présentent sous la forme de réseaux relationnels constitués de noeuds reliés par des arcs ; ce mode de représentation commun est un des éléments qui met en évidence la parenté entre ces différentes ressources. Pour accentuer l'existence de similarités entre ces ressources, on leur donne le nom générique de Ressources Terminologiques-Ontologiques (RTO).

Ces ressources ont été longtemps considérées comme très stables d'une utilisation à l'autre mais aussi d'une période à l'autre. Plusieurs explications peuvent justifier cette vision d'un fonctionnement très figé :

- D'un point de vue théorique, plusieurs points de vue qui sous-tendent les disciplines concernées ont considéré comme nécessaire de fixer les sens. Ainsi, le structuralisme s'est appuyé sur la seule synchronie pour développer la notion de système. La terminologie à la Wuster a préconisé la normalisation, qui passe par le figement, pour améliorer la communication. Dans les deux cas, les systèmes doivent nécessairement être maintenus éloignés de la réalité des usages discursifs qui introduisent la variabilité.
- D'un point de vue économique, il était plus efficace de viser une terminologie figée, c'est-à-dire dont l'élaboration pouvait être financée une fois pour toutes.
- D'un point de vue opérationnel, il était plus facile et, du moins le pensait-on, plus efficace, d'élaborer des ressources sans faire intervenir la notion de variation.

La prise en compte de l'usage réel des ressources dans les textes a représenté un tournant majeur dans la constitution de ces RTO il y a une quinzaine d'années. En effet, les liens étroits que ces ressources entretiennent avec les textes et les documents ont alors été interrogés grâce à la mise sur support numérique des documents, au développement

d'outils d'exploration textuelle et aux travaux sur la structuration informatique des ressources que cette numérisation a entraînés. Il était possible désormais d'évaluer rapidement la pertinence d'une ressource existante pour accéder à un texte et, inversement, de mettre en œuvre des méthodologies de construction de RTO à partir de textes. Les relations entre textes et RTO et le statut du corpus se sont alors trouvés au cœur de la problématique de la construction de ces ressources.

2. PROBLEMATIQUE INTERDISCIPLINAIRE : CONSTRUIRE DES RESSOURCES A PARTIR DE TEXTES

Ces interrogations ont débouché sur une convergence qui a permis à la fois de développer l'interdisciplinarité et de renouveler chacune des disciplines concernées.

Dans un premier temps, cette nouvelle situation a conduit à concentrer les efforts sur deux facettes, parfois étudiées de manière indépendante : la nature et la construction de ces ressources d'une part, et l'optimisation de leur utilisation d'autre part. La réflexion s'est ainsi développée dans différentes directions¹ : modèles de stockage et langages de représentation de ces ressources, modes d'interprétation des données textuelles, constitution des corpus, définition d'outils et techniques adaptés pour l'analyse de textes (méthodes linguistiques, statistiques, techniques d'apprentissage, ...), prise en compte des applications dans la constitution des données, etc. Dans la continuité de la vision « classique », une hypothèse forte sous-tendait ces propositions : celle d'une relative stabilité des modèles, des besoins et des usages, justifiant la stabilité des ressources à construire et assurant leur réutilisabilité.

Le recul pris au niveau théorique et les expériences réelles menées pendant ces quelques quinze années ont conduit au constat que l'hypothèse de stabilité devait elle-même être réévaluée. En effet, différents paramètres peuvent influencer sur les analyses et le choix de constitution des modèles :

- *Le genre textuel.* La variation du fonctionnement linguistique des textes est connue de longue date. Une des façons de la prendre en compte est de l'étudier dans ses corrélations avec la variation extra-

¹ Cf les travaux de l'Action Spécifique « Corpus et Terminologies » du RTP-DOC du département STIC du CNRS. <http://www.irit.fr/ASSTICCOT>

linguistique en faisant intervenir la notion de genre textuel. Nous entendons par genre textuel une caractérisation des textes qui englobe la notion classique de genre en littérature. Nous en élargissons le sens pour intégrer, en plus des éléments relatifs à leur rédaction (classiquement pris en compte), des éléments relatifs à leurs interprétations possibles par les lecteurs ainsi qu'à leur analyse automatique. Le genre textuel pourrait ainsi permettre de prendre en compte la variation tout en constituant un palier de stabilisation de description des fonctionnements.

- *Le type d'application visé.* Si l'objectif d'utilisation des ressources est pris en compte depuis longtemps par les sciences de l'information, ce n'est que plus récemment que l'ingénierie des connaissances et la terminologie textuelle ont pris conscience de l'importance de l'application dans les choix d'interprétation des données textuelles et les choix de modélisation. Identifier des catégories d'applications pourrait permettre de systématiser les modes de constitution des RTO en fonction de ces catégories.
- *Les critères de validation retenus.* La validation concerne des vérifications des RTO à différents moments de leur construction. Les critères de vérification peuvent varier dans la mesure où ils concernent l'adéquation de ces ressources avec les connaissances, d'une part, et avec leur rôle attendu auprès des utilisateurs, d'autre part : deux éléments qui font intervenir une dimension « subjective ».
- *Les modes d'évaluation.* L'évaluation concerne la satisfaction du cahier des charges lorsque la ressource est construite. Les expériences rendant possibles cette évaluation ne sont pas uniques, la situation étant rendue plus complexe encore par le fait que la RTO n'est souvent qu'un des éléments du dispositif à valider (l'application cible).

Dans la continuité d'un effort d'évaluation, il est devenu crucial de s'intéresser à un autre paramètre, laissé de côté de manière plus ou moins délibérée jusqu'à maintenant : celui de l'évolution des ressources dans le temps, c'est-à-dire de leur maintenance et de leur pertinence dans de nouveaux contextes. Le paradoxe inhérent à la constitution d'une ressource termino-ontologique devient alors flagrant : une RTO doit à la fois « normaliser » des connaissances, c'est-à-dire leur donner un statut de référence à un moment donné et pouvoir être utilisée pour accéder à des connaissances qui évoluent, parfois très rapidement, dans des contextes dynamiques. Il devient alors nécessaire de s'interroger sur l'adéquation d'une RTO dans un contexte d'évolution dans le temps des pratiques, des textes, des connaissances et des vocabulaires.

3. QUESTIONS MISES DE COTE : EVOLUTION DANS LE TEMPS ET MAINTENANCE

A l'origine de ce numéro, il nous a semblé que la réflexion pluridisciplinaire (concernant des chercheurs en linguistique, terminologie, traitement automatique des langues, ingénierie des connaissances, sciences de l'information) était suffisamment mûre pour qu'il soit désormais envisageable de prendre en compte la déstabilisation que peut représenter la prise en compte de la dimension temporelle².

Quatre problématiques sont concernées par la prise en compte de l'évolution dans la constitution de RTO à partir de textes : évolution du contexte, repérage de l'évolution dans les textes, RTO et prise en compte de l'évolution, relation entre RTO et contexte évolutif.

3.1. Evolution du contexte

C'est d'abord l'évolution du contexte qui va entraîner une éventuelle évolution de la RTO. Il faut alors identifier quels éléments de ce contexte de production et/ou d'utilisation risquent de remettre en cause la pertinence de la ressource. Si ces éléments contextuels sont repérables pour une époque donnée, on peut se demander s'il est envisageable d'anticiper l'évolution des besoins concernant cette ressource afin de les prendre en compte lors même d'en constituer une version initiale.

3.2. Repérage de l'évolution dans les textes

Lorsqu'il est probable qu'une évolution du contexte va avoir des conséquences sur l'interprétation des textes autant que sur la terminologie et les connaissances utilisées dans les nouveaux textes. Les textes peuvent ainsi fournir des indices de termes ou concepts ayant évolué. Il devient nécessaire de mettre en œuvre des méthodes pour repérer cette évolution : construction d'observables à différentes époques, analyses des différences lexicales, des différences de distribution ... Si la linguistique diachronique (la lexicologie surtout) a mis en évidence des évolutions sur de longues périodes, elle a moins travaillé sur des périodes courtes (qui peuvent concerner certains projets et leurs documents associés) et surtout, elle a peu travaillé sur des méthodes de repérage, c'est-à-dire sur les indices qui peuvent être utilisés pour mettre au jour une évolution.

² Cf. également les travaux menés dans le cadre de l'Action Spécifique Fant-AS-STIC du RTP-DOC du département STIC du CNRS <http://liris.cnrs.fr/~taccary/FANT-AS-STIC/>

3.3. RTO et prise en compte de l'évolution

La prise en compte de l'évolution dans la constitution même de la RTO pose des questions de différentes natures : comment concevoir le modèle de données pour qu'il puisse être modifié facilement ? doit-il rendre compte de l'historique des connaissances ? comment concevoir les outils de traitement automatique ou d'apprentissage pour les intégrer dans un processus cyclique et non linéaire ? Les outils peuvent soit être adaptés soit redéfinis pour s'adapter au besoin de maintenance : diagnostiquer les besoins, faciliter l'intégration des évolutions, gérer l'archivage des versions ...

3.4. Relation entre RTO et contexte évolutif

Si la RTO n'est plus fixée une fois pour toutes, il est nécessaire de trouver des méthodes et des outils qui permettent de mesurer son adéquation avec un nouveau contexte : nouveau corpus, nouveau besoin, modifications du contexte. Finalement, on peut se demander si dans certains cas il n'est pas préférable de reconstruire une ressource nouvelle plutôt que de vouloir adapter une ressource existante. En effet, les méthodes et les outils permettent de construire une RTO relativement facilement et le problème serait alors plutôt de comparer des RTO, constituées à des périodes différentes.

4. LES ARTICLES DE CE NUMERO

Le nombre relativement important de propositions en réponse à l'appel de ce numéro spécial témoigne d'un intérêt évident pour la thématique et, qui plus est, d'un intérêt pluridisciplinaire. En effet, les propositions émanaient de toutes les communautés concernées (terminologie, sciences de l'information, traitement automatique des langues et ingénierie des connaissances). En revanche, il est évident que la problématique n'est pas encore bien balisée. Plusieurs articles de bonne qualité n'ont pu être retenus à cause d'une contribution trop minime à la thématique de l'évolution : l'importance du problème est perçue mais encore assez peu travaillée. Un autre élément est apparu dans différents articles proposés : l'accent était plutôt mis sur l'évaluation que réellement sur l'évolution des RTO. En d'autres termes, ce qui était repéré, c'est plutôt le symptôme d'une adéquation remise en question du fait de l'évolution dans le temps que les moyens pour remédier à cette difficulté. Ce constat est sans doute explicable par le fait que la problématique est émergente et la réflexion

débutante. Mais il n'est pas sans intérêt de dresser un panorama des travaux existants sur une problématique qui commence à se structurer, surtout lorsqu'elle concerne plusieurs disciplines.

La diversité des articles retenus illustre le large panorama des champs d'investigation ouverts : ces articles abordent, sous des angles et points de vue complémentaires, plusieurs des quatre problématiques présentées ci-dessus.

B. Rothenburger expose sa réflexion sur le mode de prise en compte ontologique et terminologique de l'évolution des connaissances dans les domaines techniques. Il s'agit ici de conserver ou reconstituer des connaissances sur la manière dont on est parvenu aux connaissances courantes d'une communauté, en développant l'approche dans le cadre de la gestion des connaissances dans les domaines techniques.

Les deux articles suivants s'intéressent à des éléments méthodologiques et à la nature des modèles favorisant l'évolution de modèles en lien avec des textes.

T. Roy et P. Beust proposent un modèle de représentation des concepts et de la terminologie d'un domaine. Une représentation construite à partir de textes selon ce modèle (Lucia) reflète le point de vue d'un individu, auteur des textes. Sa projection sur une collection de documents rend compte de la présence (ou non) des termes et concepts. Les auteurs mènent une réflexion sur la maintenance de ces modèles et montrent en quoi ils facilitent l'étude de corpus au cours du temps.

L'article de C. Chrisment, N. Hernandez, G. Hubert et J. Mothe s'intéresse à la mise à jour d'une ontologie à partir de l'analyse d'un corpus et de la gestion de types abstraits (concepts de haut niveau d'abstraction). Le processus de mise à jour ainsi défini est utilisé pour améliorer l'indexation de documents. Il a été testé avec succès dans le domaine de l'astronomie. L'article souligne ainsi un apport possible de l'ingénierie des connaissances à la recherche d'information.

Plus proche de l'analyse linguistique, A. Tartier propose une méthode de repérage de la variation terminologique basée sur un traitement automatique. Elle définit une mesure de distance entre formes terminologiques. Mise en œuvre sur un corpus diachronique, cette méthode vise à distinguer les changements éphémères des changements durables.

Toujours au niveau terminologique, dans le cadre du projet Gene Ontology (GO), dont l'objectif est d'établir un vocabulaire unique pour la description des gènes de différentes espèces, se pose le problème de

l'augmentation incessante du nombre de gènes à annoter et du volume de données textuelles à traiter. L'article de N. Grabar, C. Bousquet et M.-C. Jaulent aborde les limites de cette ressources, dues entre autres à la formulation linguistique et à la structure, et propose des solutions pour enrichir GO et en optimiser l'utilisation dans les applications automatiques.

Pour finir, l'article de A. Baneyx et J. Charlet offre une illustration des enjeux applicatifs de la maintenance. Les auteurs s'appuient sur leur expérience de construction d'ontologies dans le domaine médical pour illustrer les questions d'évaluation et de maintenance, jusqu'ici peu abordée par les recherches du domaine. En effet, ils montrent que les principes de bonne construction d'ontologies (comme la différenciation explicite et argumentée des concepts) ainsi que les techniques d'enrichissement d'ontologies à partir de corpus (comme la recherche de relations par patrons) offrent un double intérêt pour la maintenance.

Souhaitons que ce recueil permette à toutes les communautés scientifiques réunies autour de la Revue I3, de percevoir l'enjeu d'une étude pluri-disciplinaire des questions de repérage, de restitution et de gestion de l'évolution dans le temps des terminologies, des connaissances et de leurs modélisations. En donnant quelques pistes de réflexion, nous espérons que les articles rassemblés dans ce numéro soient le point de départ de réflexions croisées et de recherches nouvelles sur ces questions.