L'annotation formelle de documents en (8) questions

Jérôme Euzenat

INRIA Rhône-Alpes
655 avenue de l'Europe, 38330 Montbonnot Saint-Martin, France
Jerome.Euzenat@inrialpes.fr

Résumé

Annoter un ensemble de documents informels à l'aide de représentations formelles appelle plusieurs questions qui doivent trouver une réponse si l'on veut développer un système cohérent. Ces questions sont liées à la forme et à l'objet des représentations retenues, à la nécessité d'utiliser de la connaissance indépendante du contenu des documents (ontologies, connaissance de contexte) et au statut du système résultant (grande base de connaissance ou éléments de connaissance distribués). Ces questions sont décrites et illustrées par l'annotation de résumés d'articles en génétique moléculaire.

Mots-clés : Web sémantique, recherche de documents par le contenu, annotation formelle, représentation du contenu, ontologie, connaissance de contexte.

1 MOTIVATIONS

L'un des buts les plus importants du web sémantique est l'annotation et l'indexation de pages web — décrites de manière informelle — par des descriptions — formelles — de leur contenu. Cependant, lancer un aspirateur de pages sur le web et attribuer à chaque page un contenu extrait automatiquement a peu de chances d'être utile. Dans la suite, nous attirons l'attention sur un ensemble de questions qui doivent trouver une réponse avant de commencer à développer un tel système. Ces questions sont liées au type et à l'usage des annotations et au statut épistémologique du contenu des sources considérées. Elles conduisent à définir la nature des descriptions de contenu, de la connaissance de contexte et du web sémantique considéré.

Les questions posées sont illustrées à l'aide de l'expérience de la construction d'une ontologie et de l'indexation et l'annotation des pages correspondantes afin de retrouver leur contenu de manière efficace. Bien que posées dans ce contexte, les questions qui nous intéressent sont applicables de manière générale. En particulier, elles sont indépendantes du médium (texte, image, multimédia…) ou du langage de représentation (le seul concept intervenant dans les questions est celui de

généricité). Elles ne sont donc pas liées aux techniques d'extraction de connaissance à partir de texte ou d'image.

1.1 Contexte

Escrire est une action INRIA réunissant trois équipes dont le but est la comparaison de trois langages de représentation de connaissance pour l'indexation et l'annotation par le contenu dans le web. Les langages comparés appartiennent aux familles des graphes conceptuels, des représentations de connaissance par objets et des logiques de descriptions. Le protocole de comparaison utilise un langage pivot défini pour le besoin de l'expérimentation et dans lequel ontologies, annotations et requêtes sont exprimées. L'expérience nécessite donc la traduction entre chaque langage et ce langage pivot. Il y a d'autres projets ayant pour but de comparer les langages de représentation de connaissance (Corcho & Gómez Pérez, 2000), mais la méthodologie d'Escrire consiste à les expérimenter en contexte.

Le domaine utilisé comme cas d'étude est celui des interactions géniques dans les premiers stades de développement de l'embryon de drosophile (*Drosophila melanogaster*) parce que le sujet nous était déjà connu (Euzenat *et al.*, 1997). Les textes utilisés, extraits de la base bibliographique Medline, ressemblent au suivant :

> Control of the initiation of homeotic gene expression by the gap genes giant and tailless in Drosophila.
> Reinitz J, Levine M
> Department of Biological Sciences, Fairchild Center, Columbia University, New York, New York 10027.
> The process of segmentation in Drosophila is controlled by both maternal and zygotic genes. Members of the gap class of segmentation genes play a key role in this process by interpreting maternal information and controlling the expression of pair-rule and homeotic genes. We have analyzed the pattern of expression of a variety of homeotic, pair-rule, and gap genes in tailless and giant gap mutants. tailless acts in two domains, one anterodorsal and one posterior. In its anterior domain tailless exerts a repressive effect on the expression of fushi tarazu, hunchback, and Deformed. In its posterior domain of action, tailless is responsible for the establishment of Abdominal-B expression and demarcating the posterior boundary of the initial domain of expression of Ultrabithorax. giant is an early zygotic regulator of the gap gene hunchback: in giant-embryos, alterations in the anterior domain of hunchback expression are visible by the beginning of cycle 14. giant also regulates the establishment of the expression patterns of Antennapedia and Abdominal-B. In particular, giant is the factor that controls the anterior limit of early Antennapedia expression.
> PMID: 1972684, UI: 90292349

Plusieurs autres fragments de résumés sont cités par la suite. Leur référence complète est accessible à travers Medline au moyen de leur identificateur Medline unique (UI) donné entre crochets (par exemple, [UI:90292349] pour le résumé ci-dessus).Les résumés de revues de biologie ont certains avantages du point de vue

de notre expérimentation : à l'instar des articles eux-mêmes, ils sont relativement précis et décrivent le résultat obtenu par l'article (un point qui n'est pas comparable aux papiers en informatique par exemple).

Notre sujet est l'élaboration d'un cadre d'annotation simple pour de tels résumés : le but était de produire une « ontologie » du domaine et des annotations à partir du contenu des résumés afin d'aider les biologistes à retrouver des articles en rapport avec leurs travaux. Les annotations ont été produites à la main, en gardant à l'esprit que l'annotation du contenu — mais pas l'ontologie — puisse être plus tard engendrée automatiquement à partir des textes (Proux *et al.*, 2000). Elles ont donc été aussi systématiques que possible.

1.2 Terminologie

Définir d'abord la terminologie a deux fonctions : éviter les mésententes et formuler plus précisément certains problèmes qui seront soulevés par la suite.

On qualifiera de *générique* une assertion qui s'applique à plusieurs individus et d'individuelle une assertion qui s'applique à un individu particulier dans le domaine d'interprétation. Une classe, une relation ou une règle sont des entités génériques (par exemple, « gène ») ; un objet ou une assertion sur un objet sont des assertions individuelles (par exemple, "Antennapedia").

Un *schéma* est un ensemble d'assertions génériques. Il décrit des entités génériques utilisées pour exprimer le contenu. Une *description* est un ensemble d'assertions individuelles. Schéma et description sont des notions purement syntaxiques. Si ces notions sont communes en bases de données ou en logiques de description, elles le sont moins en logique ou dans les langages à objets (où une classe peut faire référence à des instances).

La *connaissance de contexte* est un ensemble d'assertions qui peuvent être schématiques ou descriptives et qui sont présupposés lorsque l'on cherche à accéder au contenu d'un document. Pour un document, il s'agit de la connaissance que l'auteur pense partagée par ses lecteurs (dans le texte cité ci-dessus, le lectorat est supposé savoir qu'"Antennapedia" est un gène homéotique : ceci n'est pas écrit dans le résumé mais nécessaire à sa bonne compréhension). Bien entendu, il est possible que la connaissance de contexte soit vide.

Une *ontologie* (*O*) est un ensemble d'assertions qui décrit les concepts impliqués dans le domaine. L'ontologie étant de la connaissance commune utilisée pour comprendre le contenu, elle fait évidemment partie de la connaissance de contexte. Cependant, on préfèrera distinguer l'ontologie, composée en partie de connaissance générique, et le reste de la connaissance de contexte qui contient exclusivement des descriptions individuelles. C'est cette dernière que nous identifierons par (*K*).

Par la suite, le mot *document* (*D*) fera référence aux résumés d'articles extraits de Medline. Plus précisément, il va dénoter la partie résumé des entrées Medline (sans le titre, les auteurs, ni les noms de journaux). Le *contenu* (γ) est le sens du document envisagé sous l'angle le plus général. Dans notre contexte, ce mot

dénotera le sens que nous voulons représenter. Lorsqu'il est exprimé dans un langage formel et attaché au document, on l'appellera *annotation* (*A*).

Lorsque l'on dispose de documents et de représentations de leur contenu, on peut considérer plusieurs opérations :

— L'*extraction* qui, à partir d'un document *D*, extrait la représentation de son contenu *A* ;
— La *génération* qui, à partir d'une représentation formelle *A*, engendre une représentation documentaire (texte, image…) *D* ;
— L'*indexation* qui, à partir d'un ensemble de documents et de représentations formelles, engendre une fonction des seconds vers les premiers permettant de retrouver les documents à partir des représentations ;
— L'*annotation* qui, à partir d'un ensemble de documents et de représentations formelles, engendre une fonction des premiers vers les seconds permettant de retrouver un contenu formel à partir des documents.

Dans le présent article, on ne s'attache pas à la différence entre indexation et annotation. Notre but est de considérer le rapport entre document et annotation et l'expression de celle-ci pour la tâche d'indexation. Le mot annotation ne sera pas employé pour l'opération qu'il dénote mais l'objet qu'il dénote (c'est-à-dire la représentation formelle du contenu).

Le but d'un tel schéma d'annotation est que la combinaison de l'ontologie, de la connaissance de contexte et de l'annotation permette de reconstruire le contenu. En d'autres termes, si |= est la conséquence logique, on veut avoir, $O \cup K \cup A \models \gamma$.

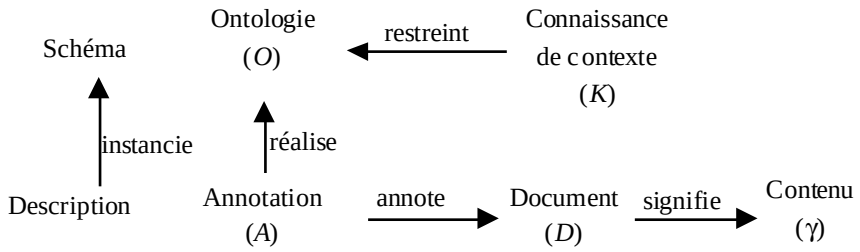

Fig. 1 – *Les différents termes utilisés et leurs relations. Il y a deux graphes déconnectés dont les relations pourront être complétés ou non suivant les réponses aux questions évoquées. Ici "réalise" signifie que les annotations peuvent être à la fois une instanciation et une extension de l'ontologie.*

La figure 1 résume les relations entre les différentes notions. Certaines des questions que nous souhaitons discuter sont liées à cette figure :

— Les annotations sont-elles uniquement une description ou peuvent-elles être une extension de l'ontologie (c'est-à-dire contenir des éléments de schéma) ?
— À l'inverse, les ontologies peuvent-elle définir certains éléments de description ou sont-elles restreintes aux schémas ?
— Y a-t-il une différence entre ontologie et connaissance de contexte (la dernière est-elle incluse dans la première) ?

1.3 Étendue et limites

Ce travail a été développé dans le contexte d'une expérimentation contrôlée qui peut éclairer son extension au web tout entier. Il est présenté comme un ensemble de questions concernant les caractéristiques des systèmes d'annotation formelle de documents. Ces questions sont posées avant d'être illustrées et discutées, puis la réponse retenue dans le cadre d'Escrire est présentée. Elles sont rassemblées dans trois sections allant du plus local au plus global : l'annotation d'un document, la connaissance de contexte concernant plusieurs documents puis le système dans son ensemble. Nous terminons en présentant les réponses à ces questions dans le cadre de différents systèmes.

Plusieurs de ces questions sont connues de disciplines comme l'extraction d'information à partir de texte, la représentation de connaissance ou la recherche d'information. Cependant, elles prennent un autre visage lorsqu'elles sont placées dans le contexte du web parce qu'elles interagissent entre elles. Elles ne sont pas liées spécifiquement au texte et peuvent s'appliquer aux images ou aux documents multimédia.

Bien que l'indexation de documents soit différente de la construction d'ontologies à partir de documents (Maedche & Staab, 2000), les questions posées devraient aussi se révéler utiles pour cette dernière tâche (Craven *et al.*, 2000).

Enfin, nous ne considérons pas le problème de relier les représentations formelles aux documents informels (Kahan *et al.*, 2001) et la discussion est indépendante du type de représentation de connaissance utilisée (ces problèmes sont en général égaux devant les questions posées).

2 QUE REPRÉSENTER ?

À l'origine on trouve les documents porteurs d'un certain contenu. Affecter des annotations à cette connaissance informelle requiert de prendre des décisions à propos du type d'information à représenter et de la forme de cette représentation. Ceci est étroitement lié aux types de requêtes auxquelles on voudrait être capable de répondre et donc à une application particulière. Cependant, le contenu lui-même peut exiger des traits de représentation qui ne peuvent être ignorés.

2.1 Quel aspect du contenu doit être représenté ?

Les annotations qui peuvent être attachées aux documents ont des statuts divers : lié au médium (date de création, longueur, encodage, format…), métadonnées (auteur, date de production…), identificateur (par exemple, un "Document object identifier" ou l'UI de Medline), descripteurs de contenu pris dans une table préétablie (par exemple, mots clés, catégories) et représentation du contenu (par exemple, un résumé). Medline fournit déjà beaucoup d'annotations sur les articles parmi lesquelles le langage, l'ISSN, le type de publication et les annotations par le "Medical Subject Headings" (qui est une organisation hiérarchique de termes). Ces annotations ne sont pas forcément une représentation

du contenu : un article de génétique moléculaire mentionne rarement le fait qu'il concerne la génétique moléculaire ou s'il est rédigé en anglais.

L'expérimentation d'Escrire est concernée par la représentation du contenu dans un langage formel. Elle ne concerne pas les métadonnées mais strictement le contenu des résumés (qui, eux-mêmes, peuvent être vus comme une représentation informelle du contenu des articles). Même en cas de représentation du contenu, plusieurs types d'information peuvent être représentés :

— La structure grammaticale du texte. Pour la phrase : "The process of segmentation in Drosophila is controlled by both maternal and zygotic genes" [UI:90292349], l'étiquetage grammatical peut-être (Hasida, 2001) :

```
<np opr="sbj">
   <ad sem="df.sg">The</ad>
   <n sem="sg">process</n>
   <adp><ad opr="arg">of</ad>
         <nph><n sem="sg" root="segment">segmentation</n></nph>
   </adp>
   <adp><ad
opr="loc">in</ad><np><name>Drosophila</name></np></adp>
</np>
<vph fun="gov">
   <v root="be" fun="aux" sem="sg">is</v>
   <v root="controll" fun="gov" sem="past">controlled</v>
</vph>
<adp><ad opr="mns">by</ad>
   <np fun="comp">
      <ajp syn="p"><io>both</io>
            <aj>maternal</aj>
            <io>and</io>
            <aj>zygotic</aj>
      </ajp>
      <n root="gene" sem="pl">genes</n>
   </np>
</adp>
```

— La structure rhétorique, c'est-à-dire la structure argumentative du texte : en général, les résumés considérés peuvent être représentés par : « état de l'art » et « expérience » conduit à « conclusion ». Par exemple, le résumé [UI:90292349] ci-dessus peut-être représenté comme « Assertion : phrases #1-2 ; Description d'expérience : phrase #3 ; Résultats : phrases #4-9 ». Une structure rhétorique plus profonde pourrait aussi être représentée (par exemple, "in particular" dans la dernière phrase du résumé relie les phrases #8 et #9 par une relation de généralité).

— La structure logique qui a trait à la représentation conceptuelle et relationnelle du contenu. Cette structure est celle représentée dans Escrire.

2.2 Quel est l'objet et la forme du contenu représenté ?

Actuellement, représenter formellement le contenu *complet* de même un texte simple ne semble pas raisonnable (une représentation immédiate — c'est-à-dire dirigée par la représentation syntaxique — en calcul des prédicats du résumé présenté ci-dessus est longue de quatre pages). Nous nous restreignons donc à des représentations d'assertions simples (similaire à ce que l'on peut exprimer avec les langages de SHOE (Heflin *et al.*, 1999) ou Ontobroker (Fensel *et al.*, 1998)). Il reste cependant plusieurs niveaux de représentation possibles :

— Références aux gènes de drosophile présents dans le document (c'est la base à laquelle s'arrêtent certains travaux (Tateisi *et al.*, 2000)) :

```
<objref name="Antennapedia" type="gene"/>
```

— Références aux classes de gènes de drosophile présents dans le document :

```
<objref name="homeotic"/>
```

— Référence aux gènes de drosophiles présents dans le document et assertion de leur classe :

```
<objref name="Antennapedia" type="homeotic"/>
```

— Assertion d'interactions entre gènes présentes dans le document :

```
<relation type="interaction">
   <role name="promoter">
      <objref type="gap" id="gt"/>
   </role>
   <role name="target">
      <objref type="homeotic" id="Antp"/>
   </role>
</relation>
```

— Assertion d'interactions entre gènes présentes dans le document et de leurs circonstances (localisation, effet…) :

```
<relation type="interaction">
   <role name="promoter">
      <objref type="gap" id="gt"/>
   </role>
   <role name="target">
      <objref type="homeotic" id="Antp"/>
   </role>
   <attribute name="effect">inhibition</attribute>
   <attribute name="location">anterior</attribute>
</relation>
```

Nous n'irons pas plus loin, principalement parce qu'il s'agit des informations que les biologistes recherchent et qui ne sont nulle part présentes dans un langage formel. Mais on pourrait ajouter d'autre information comme le contexte expérimental ou les conséquences sur le phénotype (c'est-à-dire les conséquences physiques de l'interaction sur une drosophile adulte).

Dans le contexte d'Escrire, les représentations étaient restreintes aux représentations des mentions de gènes, de classes de gènes et d'interactions entre gènes. Cela réduit la représentation des résumés au point que certains (même s'ils décrivent des articles liés aux interactions géniques ont une représentation formelle vide. À l'opposé, le résumé présenté ci-dessus contient des références à 7 gènes, 3 classes de gènes et 9 interactions (toutes distinctes). L'annotation actuelle du résumé [UI:90292349] est longue de deux pages de XML.

2.3 Les annotations sont-elles réduites à des descriptions ?

Jusqu'ici l'annotation reste simple : elle consiste à identifier des éléments particuliers dans le texte et à inclure leur description comme annotation. Une brève analyse du texte pourrait révéler qu'"Antennapedia" est un nom propre et doit donc dénoter un individu. Cependant, ce qui est clairement présenté comme un individu n'en est pas un. En effet, les articles scientifiques sont rarement au sujet d'individus : leur validité provient de leur universalité. "Antennapedia" dénote en fait le concept du gène "Antennapedia" présent dans toutes les cellules de toutes les drosophiles. Le représenter comme un individu ne pose pas de problème tant que le biologiste ne parle pas d'une instance particulière de ce gène (en fait, les résumés issus de Medline sont remarquablement homogènes sur ce point dans leur syntaxe et leur contenu).

À l'opposé, "homeotic genes" est un groupe nominal dénotant un ensemble d'individus et "gap" un nom propre dénotant une classe (il ne s'agit pas ici du mot "fossé" en anglais, mais d'une classe de gènes qui porte ce nom car la suppression d'un de ses membres introduit un fossé par perte de segments dans la segmentation du corps de la drosophile). Ceci est très pratique pour spécifier le type des objets décrits :

```
<objref name="Antennapedia" type="homeotic"/>
```

Cependant, certaines classes ont bien pour but de dénoter un ensemble d'objets, comme dans "…], Polycomp (Pc), acts as a repressor of the ANT-C and BX-C" [UI:89127495] où les complexes "ANT-C" et "BX-C" (qui sont des ensembles de gènes) sont utilisés comme des individus. On peut tirer deux leçons de cela : en ce qui concerne l'extraction à partir de texte, un individu (grammatical) peut dénoter une classe. La seconde leçon concerne la relation entre le texte et l'annotation. Il n'est pas vrai qu'une ontologie fournisse le type du contenu d'un document et les annotations doivent simplement se référer à des individus. Le contenu peut concerner des classes d'objets. Il peut affirmer des propriétés très fortes sur les classes (par exemple, le fait qu'une classe ne contienne qu'un seul élément ou qu'elle soit sous-classe d'une autre). On pourrait alors vouloir introduire des références explicites aux classes.

```
<classref name="homeotic"/>
```

2.4 Faut-il réifier certaines classes ?

De plus, le résumé présenté montre que les classes de gènes peuvent être réifiées (c'est-à-dire représentées comme des individus). Ceci soulève le problème suivant : les classes doivent-elles être exprimables dans les annotations (qui ne sont alors plus de simples descriptions) ou doivent-elles être réifiées ?

Exprimer les classes dans les annotations a l'avantage d'être naturel, mais pose le problème de la manipulation d'entités élaborées (classes, variables…). Le langage de requête doit alors être plus puissant et le schéma risque d'être dynamiquement modifié ce qui peut le mettre en péril. La réification est moins naturelle, mais évite d'introduire ces structures génériques. Par contre, si ces structures génériques existent aussi dans le système (par exemple, dans l'ontologie), elles se retrouvent dupliquées et il est nécessaire de maintenir la cohérence entre les deux représentations.

Si l'on choisit de réifier les classes, alors les assertions à représenter doivent être interprétées de manière concordante (le contrôle affecte-t-il toutes les instances de la classe "maternal" ou seulement l'une d'entre-elles?) et la construction correspondante (c'est-à-dire la quantification adaptée) doit être utilisée.

Certains langages comme RDF-Schema (Brickley& Guha, 2000), permettent de prendre en compte des ensembles d'ensembles à l'aide d'un mécanisme de réification. D'autres séparent délibérément les objets des classes (à l'instar de la séparation Abox/Tbox des logiques de description). Ces deux questions ont été débattues en représentation de connaissance pendant des années, mais savoir si les langages de description d'ontologies doivent disposer de cette capacité ou non reste une question ouverte. Ce qui est clair, dans le cadre d'un système d'annotations formelles, est qu'il est nécessaire de choisir entre une option ou l'autre et de s'y tenir (Woods, 1991).

La modélisation devient très difficile car pour le seul couple "Antennapedia-homeotic", on peut décider de les modéliser par une instance et une classe, une classe et une sous-classe ou une classe et une classe de classe. Dans Escrire, il a fallu réifier la notion de classe de gènes qui apparaît en tant que classe et en tant qu'ensemble de gènes nommés qui peuvent être manipulés dans les annotations et les requêtes.

3 LA NÉCESSITÉ DE LA CONNAISSANCE DE CONTEXTE

L'interprétation de documents requiert de la connaissance externe. Cette connaissance permet de disposer de la connaissance laissée implicite dans le corps des documents. L'approche actuelle consiste à considérer que la connaissance de contexte est contenue dans les ontologies. Comme évoqué dans la partie terminologique, il est temps de clarifier le statut des ontologies dans l'indexation et l'annotation de documents.

3.1 La connaissance de contexte est-elle nécessaire ?

Dans un article concernant la drosophile, aucun biologiste ne va prendre en compte des informations sur les procaryotes (formes de vies unicellulaires sans noyau) et sur les mammifères. En plus, ces biologistes vont se focaliser sur les gènes les plus souvent étudiés dans l'espèce. Il est donc délicat de déterminer la connaissance nécessaire à la lecture du résumé ci-dessus. Le professeur de biologie, l'étudiant en biologie et l'informaticien n'apprennent pas la même quantité d'information du résumé [UI:90292349]. Le professeur en biologie apprendra que « giant contrôle l'expression d'Antennapedia dans la partie antérieure de l'œuf », l'étudiant apprendra que « certains gènes zygotiques ont une influence sur la segmentation de la drosophile » et l'informaticien apprendra qu'« il y a des interactions entre gènes chez la drosophile ». L'information retirée dépend fortement de la connaissance initiale. En plus d'interroger ce qu'est le contenu d'un résumé, ceci pose donc le problème de délimiter cette connaissance requise.

Par exemple, dans le document ci-dessus, les auteurs tirent des conséquences sur l'influence de gènes gap ou homéotiques à partir d'expériences montrant l'interaction entre des gènes particuliers sans mentionner que certains sont des gènes gap et d'autres des gènes homéotiques. Comprendre ce résumé demande donc la connaissance des instances mentionnées dans celui-ci (c'est-à-dire, savoir qu'"Antennapedia" est un gène homéotique).

Ceci devrait être moins nécessaire dans le cas de romans de fiction où rien du monde qui ne soit générique doit être connu au préalable. Ce n'est pas le cas des travaux scientifiques (même s'ils sont destinés à émettre des jugements universaux, ils doivent parler « du monde ». Ainsi, Vénus se doit d'être la même planète dans tous les articles). L'absence de connaissance de contexte sur des individus interdit les requêtes impliquant des noms propres comme "Antennapedia". Il interdit aussi les requêtes inter-documents du type « X est-il inhibé par Z » (lorsqu'un article évoque « X active Y » et un autre « Y inhibe Z », si le « Y » ne peut être supposé commun, rien ne peut être conclu).

Déterminer et caractériser la connaissance de contexte est donc très important pour les applications d'annotations formelles.

3.2 La connaissance de contexte fait-elle partie de l'ontologie ?

Un problème annoncé est la distinction entre ontologie et connaissance de contexte. Parce que les ontologies sont de la connaissance commune (ou partagée), on devrait supposer que si l'ontologie n'est pas l'ensemble des types d'individus trouvés dans le contenu (dans le cas où des éléments génériques sont autorisés dans les annotations), elle est au moins la connaissance commune, c'est-à-dire la connaissance nécessaire à la compréhension du contexte.

Dans un tel cas, des individus (comme "Antennapedia") doivent faire partie de l'ontologie. Cependant, nous plaiderons plutôt pour que ce ne soit pas le cas. D'un côté, il y a des descriptions consensuelles des entités des résumés qui

peuvent être considérées comme « ontologiques », mais de l'autre côté, différentes applications demandent différentes connaissances de contexte. Par exemple, un système destiné à trouver l'expression des gènes pour des experts biologistes ne doit pas avoir la même connaissance de contexte qu'un système pour enseigner au lycée. Pouvoir changer celles-ci sans changer l'ontologie semble appréciable.

3.3 La connaissance de contexte peut-elle évoluer ?

Il y a des exemples de résumés dans lesquels les auteurs introduisent de nouveaux concepts ou, au moins, de nouvelles classes d'objets : «…] a new set of genes described here, which we call tube expansion genes" [UI:20347021]. Si celles-ci deviennent consensuelles, elles doivent être ajoutées à la connaissance de contexte pour interpréter d'autres documents. Ceci requiert l'évolution de la connaissance de contexte.

Ce problème peut conduire à un relativisme complet indiquant qu'il n'y a pas d'ontologies mais plutôt des cercles concentriques de connaissance de contexte de moins en moins acceptée. Dans une telle vision, la connaissance introduite dans un résumé pourrait faire partie d'un nouveau cercle de connaissance. Cette organisation est relativement proche de celle de Cyc (Lenat & Guha, 1989).

Plus radicalement, ceci devrait conduire à la conséquence que le web sémantique n'a pas besoin de connaissance de contexte (ou qu'il est sa propre connaissance de contexte) : il n'a qu'à récupérer de la connaissance sur le web. De plus, si des classes peuvent être définies dans les annotations, alors, aucune ontologie n'est nécessaire, le web sémantique n'a qu'à la faire naître !

Dans Escrire, l'ontologie et la connaissance de contexte font toutes deux partie de l'ontologie. Cette ontologie contient la description des objets partagés et elle ne peut évoluer. Il nous semble cependant préférable de séparer l'ontologie de la connaissance de contexte.

4 LE WEB SÉMANTIQUE : BASE DE CONNAISSANCE OU ÉLÉMENTS DE CONNAISSANCE DISTRIBUÉS ?

Le niveau le plus général est la base de documents annotés ou le web (sémantique) lui-même. Il est composé de ressources distribuées (documents, annotations, ontologies et connaissance de contexte). Une application doit appréhender ces ressources avec une idée claire de leur statut.

4.1 La base de documents annotés est-elle connaissance commune ou connaissance distribuée ?

Un problème qui apparaît lors de la modélisation est celui de la politique d'évaluation des requêtes. On peut chercher à répondre à une requête (q) avec uniquement la connaissance de contexte (K) et le contexte de chaque document (plus précisément, il va retourner les références des documents annotés par A, tels que : $K \cup O \cup A \models q$). Mais ce faisant, le système ne tire pas parti du reste de la base. Plus précisément, si quelqu'un demande les interactions entre les gènes "giant" et "spalt major", aucune réponse ne sera retournée car aucun résumé ne mentionne les deux gènes. Une deuxième manière de faire consiste à retourner l'ensemble minimal de documents dont les annotations ($A_1,\ldots A_n$), jointes à la connaissance de contexte, permettent de répondre à la requête ($K \cup O \cup A_1 \cup \ldots A_n \models q$). Typiquement, un biologiste peut apprécier de savoir qu'il existe un document évoquant l'interaction entre "giant" et "Antennapedia" [UI:90292349] et un autre évoquant l'interaction entre "Antennapedia" et "spalt major" [UI:92090726].

Cette discussion peut être transférée à l'interprétation des requêtes. En effet, dans le premier cas, la requête sera « quels documents mentionnent la régulation de "spalt major" par "giant" ? » alors que, dans le second cas, elle sera « Est-ce que (le contenu de la base de documents indique que) "giant" contrôle "spalt major" ? ».

Bien entendu, ceci soulève le problème de consistance : dans un langage suffisamment puissant, il y aura des ensembles de documents permettant de déduire n'importe quelle assertion parce que leur contenu mis ensemble est inconsistant. Pour peu que ces ensembles soient petits, ils seront candidats pour répondre à toutes les requêtes. La première approche a donc le mérite de confiner les inconsistances.

Au-delà de ce problème technique, il y a un problème épistémologique : ce n'est pas la même chose d'indexer un ensemble de documents à des fins de recherche d'information et d'agglomérer l'information pour construire une immense base de connaissance.

Les documents pourraient être modélisés comme les croyances de personnes particulières (leurs auteurs). Cela permettrait d'introduire de la flexibilité dans la réponse aux requêtes, mais la représentation de ces croyances nécessiterait de définir comment elles se propagent d'une personne à une autre. Cette perspective est cependant au delà des efforts actuels.

L'équilibre est difficile à trouver parce que les documents doivent partager certains objets (noms de gènes par exemple) et pas d'autres (les interactions). Dans le cas d'Escrire, le but étant la recherche de documents, les requêtes sont interprétées comme des demandes de documents décrivant le phénomène. Ce choix est nécessairement dépendant de l'application.

5 COMMENT CES QUESTIONS S'APPLIQUENT-ELLES ?

Afin de donner une idée des réponses à ces questions, nous présentons dans la table 1 les réponses concernant quelques systèmes. Ceux-ci ne sont pas tous semblables et sont présentés ici afin de se faire une idée des techniques actuellement utilisées.

Le premier est Medline lui-même. Dans Escrire, il servira de témoin, c'est-à-dire de système suffisamment éloigné des autres pour relativiser les écarts existant entre eux. En effet, celui-ci annote les mêmes articles à l'aide de métadonnées et dispose d'un ensemble de termes hiérarchisés qui annotent les articles recensés.

Le second système est le schéma d'évaluation des MUC ("message understanding conferences" (Grishman & Sundheim, 1995)) et surtout celui qui a prévalu dans les 5 premières évaluations. Son but est de tester la capacité de systèmes de traitement du texte à trouver des schémas qui correspondent finalement à nos objets (ou à des évènements comme des attentats). Ils concernent des corpus relativement réguliers comme des dépêches d'agences. S'il y a une ontologie, elle est dans chacun des systèmes comparés et souvent sous une forme non explicite. Les systèmes utilisent en général de la connaissance de contexte comme les noms de villes et de pays.

	Medline	MUC-6	Escrire	Escrire NG	SHOE	KA2
Aspect du contenu	organismes sujets journaux auteurs	produits personnes organi-sations	gènes classes de gènes interactions		chercheurs laboratoires	chercheurs laboratoires publications
Objet et forme du contenu	objets	objets, relations, actions	objets, relations		objets, relations	objets, relations
Annotation = descriptions?	O	O	O	N	O	O
Réification de classes ?	N	N	O	O	N	N
Connaissance de contexte ?	O (journaux)	O (caché)	O	O	O	N
$K \subseteq O$?	N	—	O	N	O	N
Évolution du contexte ?	N	N	N	N	N	N/O
Connaissance commune ou distribuée ?	Distribuée	Distribuée	Distribuée		Commune	Commune

Tab. 1 – *Questions concernant l'annotation formelle de documents et les réponses produites par différents systèmes.*

Les réponses sont ensuite données dans le cadre d'Escrire et dans celui d'un Escrire « nouvelle génération » qui pourrait tirer parti de la présente discussion.

SHOE (Heflin *et al.*, 1999 ; Heflin & Hendler, 2000) est un langage d'annotation destiné à guider les agents sur le web. Il permet d'annoter des pages web avec un langage de représentation par objets. Il est présenté ici dans son utilisation-pilote consistant à indexer des pages de chercheurs et de laboratoires. Ontologies et annotations sont séparées, mais les ontologies peuvent contenir des éléments individuels et il est possible de faire référence à des morceaux d'ontologies plus spécifiques dans les annotations.

KA2 (Fensel *et al.*, 1998 ; Staab *et al,*. 2000) est l'application du système Ontobroker à la construction d'une base de connaissance distribuée (concernant l'acquisition de connaissance) à partir de pages web annotées. Comme SHOE, son but est de rapatrier la connaissance dans une base centralisée et de répondre à des requêtes. Mais l'ontologie est partie intégrante de l'application et ne permet pas d'intégrer de la connaissance de contexte : la seule source de connaissance provient des annotations.

Les systèmes que nous avons examinés sont relativement semblables (en particulier en ce qui concerne la forme de la connaissance représentée, y compris une absence de réification). Le seul grand clivage concerne l'interprétation des requêtes. Ceci est, semble-t-il, le socle minimal que peuvent offrir les systèmes d'annotation. Ils devront développer d'autres paramètres pour répondre à des demandes de représentations plus poussées.

L'un des problèmes apparaissant dans le tableau 1 est l'absence de prise en compte de l'évolution des ontologies et de la connaissance de contexte. Cette lacune risque de poser des problèmes.

6 CONCLUSION

Bien que le but d'« exprimer le contenu de documents dans un langage formel pour une recherche plus focalisée » soit attirant, l'implémenter sans précautions risque de conduire à des déconvenues. Différentes questions doivent être tranchées afin d'obtenir un résultat cohérent. Certaines d'entre-elles sont bien connues en traitement de la langue, en représentation de connaissance ou en recherche d'information, mais elles prennent une nouvelle dimension dans le présent contexte parce qu'elles interagissent.

En effet, construire une base de connaissance géante requiert de partager les noms, ce qui conduit à les considérer dans la connaissance de contexte. Mais si le système extrait la connaissance des documents, il devrait aussi laisser cette connaissance de contexte évoluer avec le temps (et donc amorcer le web sémantique sans connaissance de contexte). Si, de plus, la connaissance générique est permise dans les annotations, alors il devient théoriquement possible d'amorcer un tel système sans ontologie préconçue.

Nous avons cherché à formuler et clarifier ces questions en les illustrant dans le contexte d'Escrire : une expérimentation simple par rapport à ce qui attend le web sémantique, mais utile pour les drosophilistes. Cependant, nous prétendons qu'elles sont pertinentes pour tout projet d'annotation formelle de documents.

Il n'existe certainement pas une réponse ou une combinaison de réponses meilleure que les autres. Il y a plusieurs combinaisons répandues (voir §5) et quelques-unes seront plus adaptées à certaines applications. À notre avis, laisser ces questions sans réponses dans un projet d'annotation (d'une partie) du web hypothèquera son développement.

REMERCIEMENTS

Ce travail a été partiellement financé par l'Action de recherche concertée Escrire de l'INRIA soutenue par France Telecom. L'auteur tient à remercier l'ensemble des équipes impliquées dans l'action Escrire (Rim Al-Hulou, Olivier Corby, Rose Dieng, Carolina Medina, Emmanuel Nauer, Amedeo Napoli, Yannick Toussaint, Raphaël Troncy) ainsi que Jeff Heflin et Gwendal Auffret pour leurs discussions et commentaires fructueux.

URLS

L'information sur Escrire peut être trouvée à `http://escrire.inrialpes.fr` et l'accès public à Medline (Pubmed) se trouve à `http://www.ncbi.nlm.nih.gov/PubMed`.

RÉFÉRENCES

BRICKLEY D., GUHA R. (2000) Resource description framework (RDF) schema specification 1.0. W3C candidate recommandation http://www.w3.org/TR/2000/CR-rdf-schema-20000327

CORCHO O., GÓMEZ PÉREZ A. (2000) A roadmap to ontology specification languages. *Lecture notes in computer science* 1937, p.80-96

CRAVEN M., DIPASQUO D., FREITAG D., MCCALLUM A., MITCHELL T., NIGAM K., SLATTERY S. (2000) Learning to construct knowledge bases from the world wide web. *Artificial intelligence* 118, 1-2, p. 69-113

EUZENAT J., CHEMLA C., JACQ B. (1997) A knowledge base for *D. melanogaster* gene interactions involved in pattern formation. Actes 5th international conference on intelligent systems for molecular biology, Halkidiki (GR), p. 108-119

FENSEL D., DECKER S., ERDMANN M., STUDER R. (1998) Ontobroker: the very high idea. Actes 11th FLAIRS, Sanibal Island (FL US), p. 131-135

GRISHMAN R., SUNDHEIM B. (1995) Design of the MUC-6 evaluation. Actes 6th Message understanding conference, Columbia (ML US)

HASIDA K. (2001) The GDA tag set, http://www.i-content.org/GDA/tagset.html

HEFLIN J., HENDLER J., LUKE S. (2000) SHOE: a knowledge representation language for internet applications. Rapport technique CS-TR-4078, University of Maryland, College Park (ML US)

HEFLIN J., HENDLER J. (2000) Semantic interoperability on the web. Actes 1st Extreme Markup Languages conference, Montréal (CA) , p. 111-120 http://www.cs.umd.edu/projects/plus/SHOE/pubs/extreme2000.pdf

KAHAN J., KOIVONEN M.-R., PRUD'HOMMEAUX É., SWICK R. (2001) Annotea : an open RDF infrastructure for shared web annotations. Actes 10th WWW conference, Hong-Kong (HK), p. 623-632

LENAT D., GUHA R. (1989) Building large knowledge-based systems: representation and inference in the Cyc project. Readings (MA US) : Addison-Wesley

MAEDCHE A., STAAB S. (2000) Mining ontologies from texts. *Lecture notes in computer science* 1937, p. 189-202

PROUX D., RECHENMANN F., JULLIARD L. (2000) A pragmatic information extraction strategy for gathering data on genetic interactions. Actes 6th International Conference on Intelligent systems for molecular biology, La Jolla (CA US), p. 279-285

STAAB S., ANGELE J., DECKER S., ERDMANN M., HOTHO A., MAEDCHE A., SCHNURR H.-P., STUDER R., SURE Y. (2000) Semantic community web portals. Actes 9th WWW Conference, Amsterdam (NL), p. 473-492

TATEISI Y., OHTA T., COLLIER N., NOBATA C., TSUJII J.-I. (2000) Building an annotated corpus in the molecular biology domain. Actes Coling workshop on « Semantic annotation and intelligent content », Luxembourg (LU), p. 28-34

WOODS W. (1991) Understanding subsumption and taxonomy: a framework for progress. In John Sowa (éd.), *Principles of semantic networks: exploration in the representation of knowledge*, p. 45-94. San-Mateo (CA US) : Morgan-Kauffman