

UN LEXIQUE AMÉLIORANT L'INTELLIGIBILITÉ D'UNE BASE DE CONNAISSANCES

Florence Lemaire

Inria Rhône-Alpes
655 route de l'Europe
38330 Montbonnot
Florence.Lemaire@inrialpes.fr

Résumé

Le partage de connaissances se heurte au problème de la compréhension de connaissances décrites par une personne autre que l'utilisateur. Pour améliorer la lisibilité des connaissances représentées dans une base d'objets, nous proposons de les relier à leurs sources et justifications textuelles au moyen d'un lexique. Le lexique est l'élément central de l'environnement de partage de connaissances, car il permet de relier ces objets à des documents. Il constitue une interface entre l'utilisateur et la base de connaissances et facilite donc l'accès au contenu de celle-ci. La définition de l'organisation d'un tel lexique demande de s'intéresser aussi bien aux problèmes liés à la constitution d'une terminologie qu'à ceux de la représentation de connaissances qui relève de la psychologie cognitive.

1. Introduction

La possession de la carte de l'île au trésor ne permet pas toujours l'accès au trésor. En effet, la compréhension de la carte nécessite l'accès à un certain nombre d'informations non représentées telles que: la localisation de l'île, la signification des symboles, l'échelle choisie, la date de la carte. Ce problème de compréhension de connaissances représentées par une personne autre que l'utilisateur est central pour la réalisation de grandes bases de connaissances.

L'élaboration de grandes bases de connaissances scientifiques (plusieurs dizaines de milliers d'entités) est un processus long et incrémental qui requiert la participation de nombreux spécialistes de domaines connexes qui doivent réaliser un partage de connaissances. Dans ce cadre, le partage de connaissances se définit comme la possibilité pour chaque spécialiste du domaine de comprendre toutes les connaissances de la base qui lui sont nécessaires pour la progression de sa recherche. Cette compréhension doit être suffisamment complète pour lui permettre de modifier ou enrichir la base de manière cohérente avec la représentation existante des connaissances.

Dans le cas d'une base de connaissances scientifiques utilisant une représentation de connaissances par objets, une connaissance est une classe, une instance ou un attribut, par la suite nous utiliserons le terme d'élément de connaissance formelle (ecf) pour désigner ces connaissances. Un ecf est dépendant d'informations qui ne sont pas représentées dans la base de connaissances, mais qui sont nécessaires pour en comprendre le contenu. Par conséquent, face à un ecf, l'utilisateur peut se poser des questions concernant la terminologie, la validité scientifique et les choix de modélisation effectués.

Dans une représentation d'objets, une connaissance est un objet décrit en utilisant des noms (noms de classes, d'attributs et d'instances, par exemple) qui sont aussi des termes du vocabulaire spécialisé du domaine [Bourigault, 1995]. Lors de la modélisation, ils peuvent être utilisés dans une acception très

spécifique et l'utilisateur peut donc se poser des questions de terminologie. La réponse à ces questions nécessite au minimum l'accès à la définition du terme, qui doit permettre d'en comprendre la signification, mais aussi l'usage particulier qui en est fait dans la base de connaissances.

- La validité d'une connaissance concerne les informations théoriques et expérimentales qui justifient la présence d'une connaissance dans une base. Ces informations se trouvent dans des publications exposant la connaissance en question.
- L'explication des choix de modélisation doit, par exemple, permettre de répondre à la question: pourquoi tel attribut est défini dans telle classe? La justification des choix de modélisation nécessite souvent d'interroger le concepteur, ce qui implique qu'il soit accessible et qu'il se souvienne de ses choix de modélisation.

Ces trois types de questions concernent d'une part le processus d'abstraction et d'autre part les objets modélisés, c'est-à-dire deux niveaux différents d'informations. La réponse à la plupart des questions se trouve, sous forme textuelle, dans des documents à l'origine de la représentation des connaissances. Toutes ces informations ne sont pas formalisables. De plus leur formalisation au sein de la base de connaissances aurait pour effet de déplacer le problème de la compréhension des connaissances formelles aux informations explicatives formalisées. Cependant, il est nécessaire d'intégrer ces informations de façon que qu'elles soient accessibles lors de la consultation ou la modification de la base.

Le problème consiste donc à relier les ecf à des informations textuelles de façon à permettre leur consultation par un spécialiste du domaine.

Dans une première partie, nous montrons que ce problème peut être envisagé comme relevant de la documentation, mais que les solutions existantes ne sont pas adaptées à des bases de connaissances scientifiques. Dans la seconde partie, nous décrivons notre solution, un environnement de partage de connaissances composé de trois structures: une base de connaissances, une base de documents et un

lexique. Le lexique est la structure pivot du fonctionnement de cet environnement et nous exposons dans la troisième partie certains des problèmes que soulève sa conception. Pour des raisons de lisibilité, nous nous focalisons sur le processus de consultation de la base et nous évoquons peu les problèmes liés à l'élaboration et à la modification de son contenu.

2. Un problème de documentation

Le partage des connaissances est dépendant de la façon dont les connaissances représentées sont explicitées car la représentation formelle est rarement autosuffisante pour être intelligible.

2.1. Insuffisances des solutions existantes

Le problème de la lisibilité d'une connaissance abstraite ou formalisée n'est pas spécifique aux bases de connaissances et des réponses ont été proposées dans le cadre du génie logiciel, des systèmes experts ou de l'acquisition de connaissances.

Ces réponses sont de deux types (cf. tableau 1): documentation ou définition d'une structure complémentaire. Les solutions relevant de la documentation telles que les commentaires ou les hypertextes consistent en la connexion d'informations complémentaires (souvent de nature textuelle) à la connaissance à expliciter. L'autre solution consiste à définir une structure particulière relativement indépendante des connaissances formalisées et dont le rôle est de fournir des explications ou de justifier celles-ci. Les modules d'explication des systèmes experts sont un cas particulier d'une telle structure.

Pour résoudre le problème de la lisibilité d'une grande base de connaissances scientifiques, le choix d'une solution s'inspirant de la documentation semble approprié. En effet, l'utilisateur est un spécialiste qui cherche à enrichir sa connaissance du domaine. Les questions qu'il peut formuler sont multiples et concernent différents aspects de la connaissance. Par conséquent les réponses à fournir sont elles aussi multiples. De plus, la construction est un processus incrémental et la documentation doit accompagner ce processus.

Tableau 1: récapitulation des grands types de solutions. Les trois premières solutions relèvent de la documentation, les deux autres définissent une structure particulière destinée à l'explication.

solutions	exemples	remarques
annotation simple: une courte justification est associée à la connaissance au niveau de sa description	commentaire d'un programme	ne permet pas des explications fournies et complexes
annotation hypertexte: un nœud hypertexte est associé à chaque connaissance représentée	(Moia, 1990) et (Gaines, 1990)	l'existence d'un lien physique entre la connaissance et l'information limite la quantité d'informations et leur organisation
hypertexte indépendant: un hypertexte gère les informations et est connecté aux connaissances formelles	ColiGene (Grivaud, 1992) (Samuels, 1994)	problèmes liés aux hypertextes: production de l'hypertexte et guidage de l'utilisateur
générateur d'explications: une base de connaissances explicatives permet à un module de générer des justifications d'actions réalisées par un système expert	(Wick, 1992)	justification d'actions exécutées par un système
ontologie: hiérarchie de haut niveau d'abstraction explicitant la conceptualisation	(Gruber, 1993)	la formalisation est explicitée au moyen d'un autre formalisme pas de gestion des sources

2.2. Documenter des connaissances scientifiques

Une première expérience de documentation a posteriori d'une base de connaissances scientifiques existante - ColiGene - a été effectuée au moyen d'une structure hypertexte (Grivaud, 1992).

ColiGene est une base de connaissances dédiée à l'étude de l'expressivité des gènes chez *E.coli*. (Perrière, 1993; Schmeltzer, 1993). Dans cette base, les connaissances modélisées sont représentées sous

forme de classes. Chaque classe possède un nom et est constituée d'un ensemble d'attributs. Les classes sont structurées en une hiérarchie de classes et de sous-classes. Chaque nom de classe et d'attribut est documenté par des nœuds hypertextes. Certains mots des textes de ces nœuds sont reliés à d'autres nœuds. L'utilisateur a donc la possibilité de naviguer dans la base de connaissances ou à travers les nœuds hypertextes (Figure 1).

L'hypertexte élaboré a principalement un but didactique. Les nœuds contiennent de courtes descriptions assez générales des concepts représentés dans la base.

La réalisation de la documentation d'une base scientifique, destinée à des spécialistes, nécessite la conception d'une structure gérant l'accès à des informations beaucoup plus nombreuses et plus complexes que celles utilisées dans cette maquette. En effet, la documentation d'une seule entité complexe de la base ColiGene - l'instance opérationnelle tryptophane, par exemple - requiert déjà plus d'une centaine d'articles scientifiques. C'est-à-dire qu'une revue de questions sur l'opération tryptophane fait référence à au moins une centaine d'articles. Dans le cadre de l'étude de l'expressivité des genres, plusieurs aspects de l'opération

tryptophane peuvent être documentés : la structure de la séquence, les interactions avec la polymérase, la séquence du ribosome, l'action du tryptophane ou de ses analogues... La documentation d'un seul aspect - par exemple l'effet de la température - fait elle-même référence à une dizaine d'articles. ColiGene est un exemple de base de petite taille (6000 objets) et une documentation complète d'une telle base peut être estimée à un minimum de 20 000 articles.

L'utilisation d'une structure hypertexte semble inadéquate pour un accès rapide à un article pertinent dans une telle quantité de documents. Il paraît donc nécessaire de définir une structure intermédiaire organisant et structurant les documents tout en liant aux ecf.

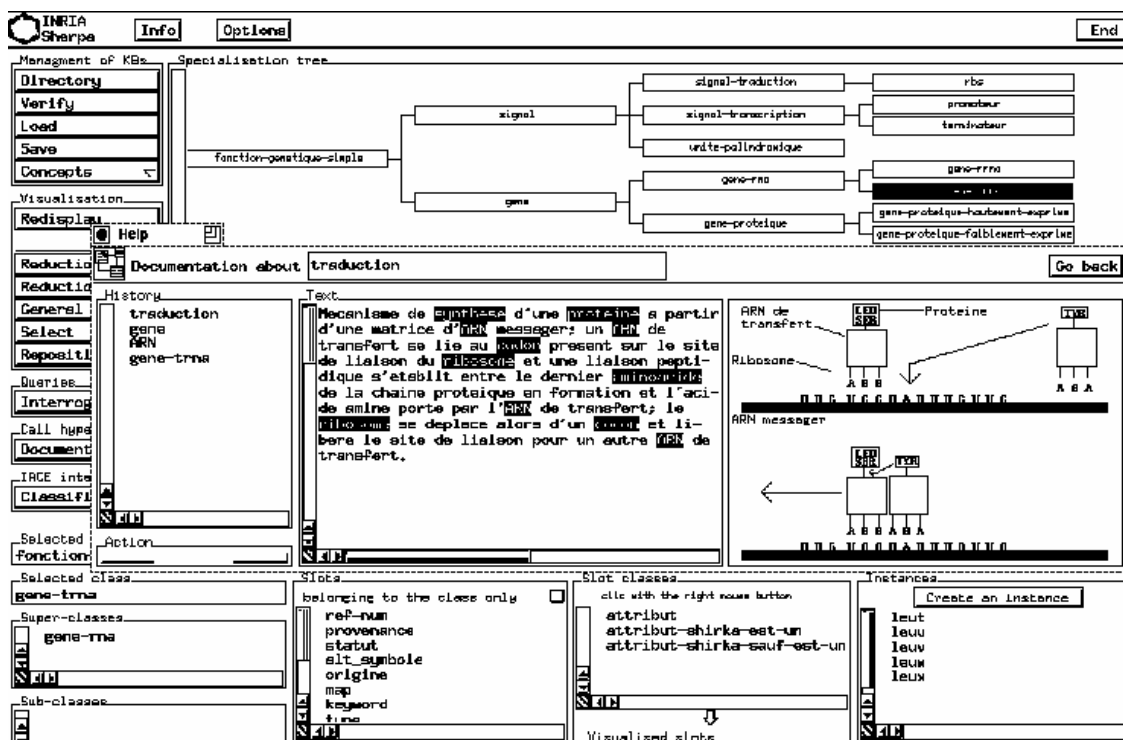


Figure 1: L'hypertexte associé à ColiGene. Chaque nom de classe est documenté par un nœud hypertexte. L'utilisateur peut consulter le graphe d'héritage ou la structure hypertexte et peut passer d'une structure à l'autre.

1.3. Un niveau de représentation intermédiaire

Le niveau intermédiaire est destiné à aider l'utilisateur à passer d'un ecf à un document l'explicitant. Cette opération nécessite deux étapes:

- la transformation de l'ecf qui intéresse l'utilisateur en une représentation qu'il manipule plus facilement. Cette étape requiert l'interprétation de l'objet formalisé et sa traduction en un schéma mental chez l'utilisateur.
- la traduction des interrogations générales dans la mémoire de travail de l'utilisateur par cette représentation de connaissances. Cette traduction doit conduire à une description de l'information recherchée assez précise pour n'accéder qu'aux quelques documents pertinents.

On considère que l'utilisateur est un spécialiste du domaine traité dans la base, on peut donc supposer que les entités qu'il manipule le plus facilement sont probablement les termes appartenant au vocabulaire du domaine. Dans certains cas, d'autres entités sont préférées, par exemple les cartes géométriques pour certains types de recherches et de raisonnements sur une base en biologie moléculaire.

Les entités du vocabulaire manipulées par l'utilisateur interviennent à deux niveaux lors d'une consultation. En effet, une partie des termes du vocabulaire du domaine est utilisée pour décrire les connaissances de la base (noms de classes, d'instances et d'attributs). De plus, ces termes permettent aussi de décrire le contenu des documents. Le vocabulaire du domaine constitue un intermédiaire pertinent entre la base et les documents.

2. Un environnement de partage de connaissances

Comme nous l'avons montré précédemment, l'accès à un document pertinent à partir d'un ecf, alors que celui-ci peut être documenté par de nombreux documents, nécessite de définir une structure intermédiaire qui guide l'utilisateur vers le bon document en utilisant le vocabulaire du domaine. Le lexique SiLex joue ce rôle dans notre environnement.

2.1. Un environnement composé de trois bases

Nous proposons un environnement de partage de connaissances composé de trois bases: une base de connaissances, une base de documents et un lexique SiLex (Figure 2).

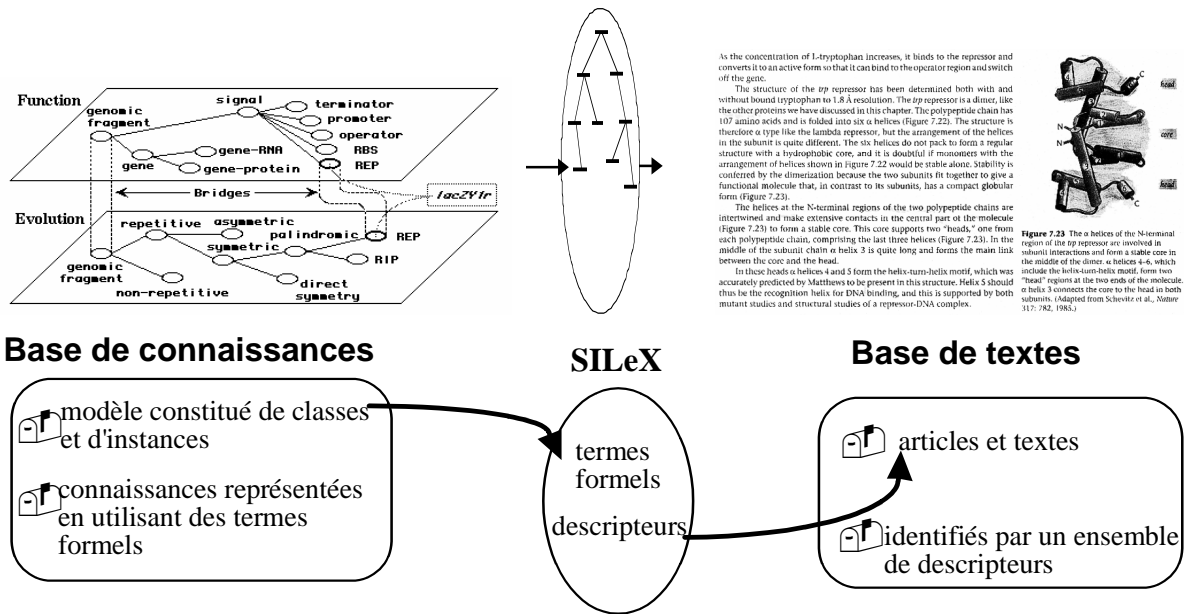


Figure 2: Un environnement composé de trois structures. La base de connaissances contient une modélisation d'un aspect du domaine. Le lexique SiLex contient le vocabulaire du domaine, les liens entre les termes sont représentés de façon explicite. La base de textes contient les documents expliquant et justifiant le contenu de la base de connaissances. Le lexique fait la liaison entre les connaissances formalisées et les documents.

- la base de connaissances est un modèle des connaissances impliquées dans un domaine particulier. Elle utilise une représentation par objets et propose des mécanismes d'inférence tels que l'attachement procédural ou la classification. La description des connaissances utilise des noms qui appartiennent à la terminologie du domaine et que nous nommerons les termes formels.
- la base de documents contient des publications, ainsi que des textes, schémas ou images produits spécialement pour documenter la base. Le contenu de chaque document peut être décrit par un ensemble de termes choisis par les concepteurs au moment de l'indexation du document; ces termes sont des descripteurs.
- le lexique regroupe tout le vocabulaire utilisé dans l'environnement, c'est-à-dire les termes formels, les descripteurs, mais aussi des mots plus généraux appartenant aussi au vocabulaire du domaine permettant de définir ce vocabulaire. Nous utilisons le terme 'lexique' en référence aux travaux de Van de Riet [1993] qui sont à l'origine de notre interface pour une structure de type lexique envisagée comme interface. Cependant, cette structure (le lexique) étant constituée

de termes spécifiques d'un domaine et d'une application, le terme 'terminologie' semble plus approprié. Mais la notion de terminologie sous-entend un travail de terminologues quant à l'analyse des termes utilisés, la façon dont ils sont employés, travail qui est hors du propos de cette étude. Afin de distinguer la structure proposée des lexiques et terminologies existantes, nous précisons, à chaque fois, son nom: SiLex, qui signifie 'un simple lexique'. Le lexique SiLex définit les termes utilisés, présente des relations entre eux et structure le vocabulaire en plusieurs sous-domaines.

La description des interactions entre l'utilisateur et la base lors d'une session de consultation (Figure 3) va nous permettre d'illustrer le fonctionnement de cet environnement et les avantages de l'intégration de SiLex.

Lors de la consultation de la base, l'utilisateur dispose d'une interrogation plus ou moins précise, mais il ne connaît pas nécessairement les termes formels pour l'exprimer. SiLex lui permet de choisir les termes pertinents pour accéder au contenu de la base formelle et lui permet ensuite d'accéder aux documents répondant à ses interrogations.

As the concentration of L-tryptophan increases, it binds to the repressor and converts it to an active form so that it can bind to the operator region and switch off the gene.

The structure of the *trp* repressor has been determined both with and without bound tryptophan to 1.8 Å resolution. The *trp* repressor is a dimer, like the other proteins we have discussed in this chapter. The polypeptide chain has 107 amino acids and is folded into six α helices (Figure 7.23). The structure is therefore a type like the lambda repressor, but the arrangement of the helices in the subunit is quite different. The six helices do not pack to form a regular structure with a hydrophobic core, and it is doubtful if monomers with the arrangement of helices shown in Figure 7.22 would be stable alone. Stability is conferred by the dimerization because the two subunits fit together to give a functional molecule that, in contrast to its subunits, has a compact globular form (Figure 7.23).

The helices at the N-terminal regions of the two polypeptide chains are intertwined and make extensive contacts in the central part of the molecule (Figure 7.23) to form a stable core. This core supports two "heads," one from each polypeptide chain, comprising the last three helices (Figure 7.23). In the middle of the subunit chain α helix 3 is quite long and forms the main link between the core and the head.

In these heads α helices 4 and 5 form the helix-turn-helix motif, which was accurately predicted by Matthews to be present in this structure. Helix 5 should thus be the recognition helix for DNA binding, and this is supported by both mutant studies and structural studies of a repressor-DNA complex.

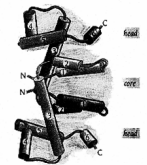


Figure 7.23 The α helices of the N-terminal region of the *trp* repressor are involved in subunit interactions and form a stable core in the middle of the dimer. α helices 4-6, which include the helix-turn-helix motif, form two "heads" regions at the two ends of the molecule. α helix 3 connects the core to the head in both subunits. (Adapted from Scheraga et al., *Nature* 317: 782, 1983.)

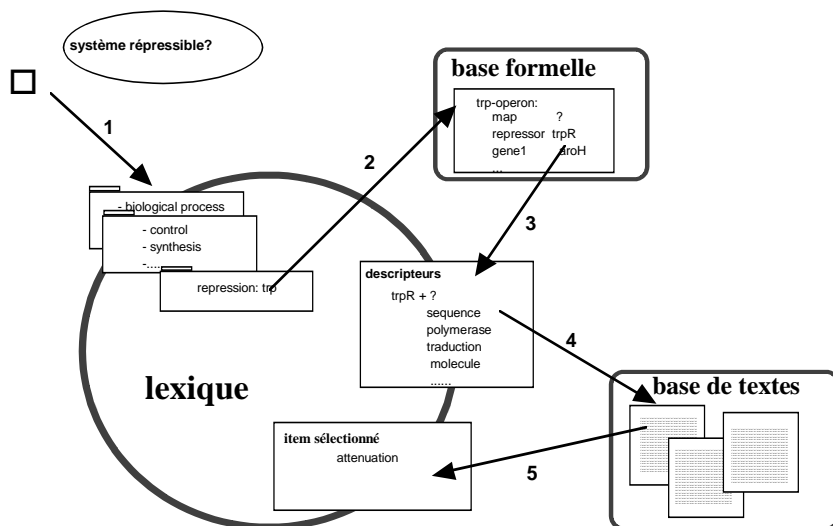


Figure 3: Scénario de consultation. La navigation dans le lexique SiLex permet à l'utilisateur de préciser sa requête (1) et de sélectionner le terme formel pertinent pour accéder au contenu de la base formelle (2). Dans celle-ci, il peut sélectionner un nom de classe ou d'instance. Il retourne ainsi dans le lexique SiLex (3) où il peut sélectionner un sous-ensemble des descripteurs associés à ce nom lors de l'indexation des documents. Il affine ainsi sa requête et accède à un nombre réduit de documents pertinents (4). Il peut sélectionner un des descripteurs associés au document qu'il consulte (5) et retourner ainsi dans le lexique pour poursuivre sa recherche.

2.2. SiLex: une interface

SiLex est la clé de voûte de l'architecture de l'environnement de partage de connaissances, car il constitue le lien entre la base formelle et la base de documents. C'est de plus l'élément central du fonctionnement de cet environnement, car il est au cœur de toutes les interactions avec l'utilisateur.

Un intermédiaire entre la base de connaissances et les documents

SiLex contient les termes formels utilisés dans la base de connaissances, les descripteurs du contenu des documents et des liens entre ces différents mots. Par conséquent, il permet à l'utilisateur de passer d'un ecf à un document. En effet, lorsqu'un utilisateur désire des informations complémentaires sur une connaissance de la base, SiLex l'aide à affiner sa requête en lui proposant une liste des descripteurs associés au nom de cette connaissance lors de l'indexation des documents. L'utilisateur sélectionne alors les descripteurs pertinents et accède au document correspondant sans avoir à parcourir un ensemble de documents comme dans un hypertexte.

Une interface entre l'utilisateur et la base de connaissances

En proposant la définition des termes formels, SiLex aide l'utilisateur à accéder au contenu de la base de connaissances. En effet, Van de Riet (1993) suggère d'utiliser un lexique comme une structure aidant l'utilisateur à choisir les termes adéquats quand il travaille avec un outil informatique ou une base de données. Le lexique est envisagé comme une structure améliorant le rôle des dictionnaires de données.

Dans notre environnement, l'utilisateur est supposé avoir une connaissance générale du vocabulaire du domaine. SiLex présente les termes suivant une hiérarchie conceptuelle et l'utilisateur peut aller des termes généraux à des termes plus spécifiques qui correspondent aux termes utilisés dans la base formelle (figure 4).

Un dépôt de vocabulaire utilisé dans l'environnement

Comme SiLex présente tout le vocabulaire utilisé dans l'environnement de façon structurée, à l'aide de hiérarchies, il offre une vue du domaine traitée dans la base et permet de le situer dans un contexte plus large. Il permet en effet d'exprimer des informations qui ne sont pas décrites dans la base. De plus, SiLex rend explicite les liens entre les descripteurs, il fournit donc un schéma des concepts qui décrivent le contenu informatif de la collection de documents (Agosti, 1992).

Une aide à la modélisation et à la documentation

L'existence de SiLex doit permettre aux concepteurs d'exprimer leurs connaissances au moyen d'un vocabulaire consensuel. En effet, lorsqu'un concepteur désire ajouter une nouvelle connaissance à la base formelle, la consultation de SiLex lui permet de choisir les termes pertinents pour décrire la connaissance; il peut ainsi vérifier que les termes qu'il utilise ont l'acception attendue et qu'il n'a pas oublié de dépendances entre connaissances. De plus, SiLex peut aider les concepteurs à indexer les documents qui justifient les connaissances ajoutées en proposant une organisation de l'ensemble du vocabulaire du domaine.

SiLex joue ainsi différents rôles qui sont liés à sa structure, à son contenu et à ses connexions avec les deux autres bases de l'environnement.

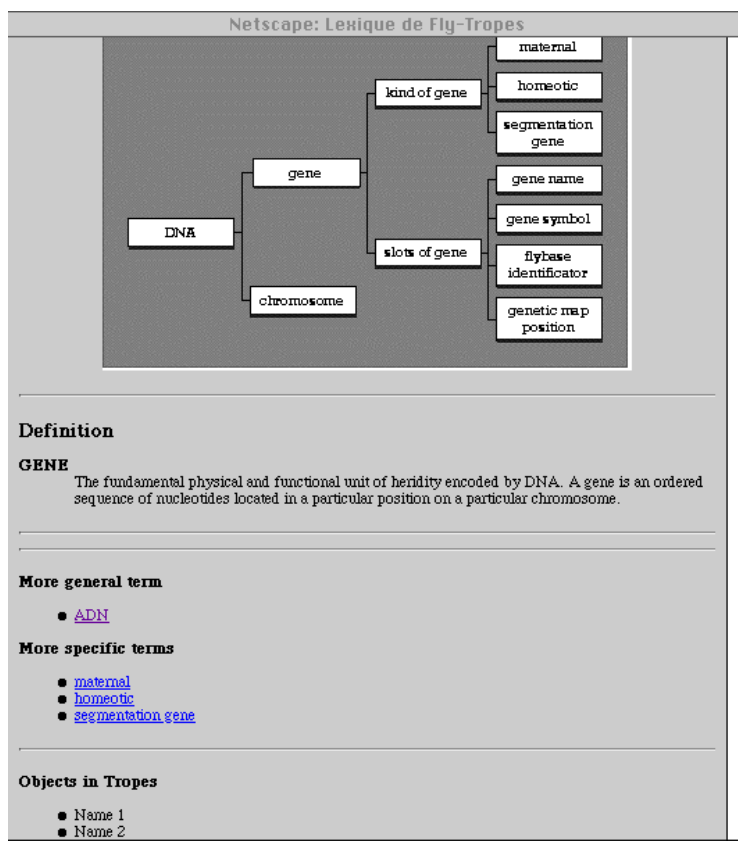


Figure 4: Exemple d'un Œcran de SiLex dŒfinissant un terme et le situant dans sa hiŒrarchie. Chaque terme est caractŒrisŒ par sa dŒfinition en langue naturelle, les attributs Œmore general termŒ et Œmore specific termsŒ qui permettent la navigation dans la hiŒrarchie de termes et les noms dŒrivŒs de ce terme qui sont utilisŒs dans les descriptions des ecf.

3. flaboration de SiLex

SiLex est la structure pivot de l'environnement de partage de connaissances. Sa rŒalisation nŒcessite une rŒflexion Œ trois niveaux diffŒrents: linguistique, psychologique et informatique afin de rŒsoudre trois types de problŒmes: qu'est ce qu'une information terminologique? Comment doit Œtre organisŒ le contenu du lexique ? Quelles doivent Œtre ses interactions avec les deux autres structures?

3.1. Une information terminologique

La rŒalisation de SiLex nŒcessite de rŒsoudre le problŒme de la caractŒrisation d'un terme, c'est-Œ-dire dŒterminer les informations nŒcessaires et pertinentes pour dŒfinir un terme. Ce problŒme existe dans plusieurs domaines tels que la linguistique, l'intelligence artificielle et la terminologie. Plusieurs faŒons d'expliquer un terme ont ŒtŒ proposŒes suivant le cadre dans lequel le terme doit Œtre utilisŒ.

Ainsi, dans le domaine du traitement de la langue naturelle, l'item doit Œtre caractŒrisŒ de faŒon Œ ce que sa signification soit manipulable par des modules informatiques. Ce qui signifie que la description des items doit permettre, lors de l'analyse d'une phrase, d'accŒder Œ son sens. La description des items comprend donc aussi bien leurs caractŒristiques

grammaticales et les constructions syntaxiques autorisŒes que leur signification. Le but de la description est de relier de faŒon non ambiguŒ un item Œ un concept. Plusieurs moyens sont utilisŒs pour rŒaliser cette liaison. Ainsi, dans Jacobs (1993), une entrŒe grammaticale est reliŒe Œ un ensemble de sens, la description du sens contenant aussi bien des informations syntaxiques que la localisation du concept dans une hiŒrarchie. Par contre dans Mikrokosmos (Onyskevych, 1992), les informations lexicales sont traitŒes dans un lexique alors que les concepts sont dŒcrits dans une ontologie. Le lien entre les deux types d'informations se fait par l'intermŒdiaire d'attributs dŒcrits dans le lexique et calculant le concept associŒ Œ l'entrŒe.

Le problŒme de la caractŒrisation d'un item est diffŒrent dans le cas de l'acquisition de connaissances. En effet, les structures terminologiques sont alors dŒfinies dans un objectif d'aide Œ la formalisation des connaissances. Le problŒme est alors de caractŒriser un terme de faŒon Œ ce qu'il n'y ait pas d'ambiguŒtŒ lors de son utilisation par des concepteurs. Dans les bases de connaissances terminologiques (Meyer, 1992; Bourigault, 1995), un terme est dŒcrit par des Œquivalents dans d'autres langues, des catŒgories morphologiques, des particularitŒs grammaticales et une dŒfinition en langue naturelle, ainsi que l'expression des relations conceptuelles. L'objectif d'une base de connaissances terminologiques est de servir de source d'informations pour la construction de

plusieurs bases de connaissances (Condamines, 1994).

En acquisition des connaissances et en traitement du langage naturel, la source du vocabulaire est un ensemble de textes (Rey, 1992; Rastier, 1996). Dans le cadre de l'environnement de partage de connaissances, le vocabulaire provient en partie d'un processus de modélisation. La définition du vocabulaire se fait en parallèle à la réalisation de la base de connaissances. Le problème n'est donc plus de relier un item textuel à un concept, mais de distinguer la définition d'une notion (ou concept) de sa description en tant que connaissance représentée dans un objectif particulier. Cette nouvelle façon d'appréhender le problème de la définition d'un terme par une focalisation sur la distinction entre définition et description peut apporter de nouvelles solutions aux problèmes de conception de lexiques. En effet, la définition des termes effectuée en parallèle à la description des objets dans une structure de représentation des connaissances doit aider à distinguer les aspects pertinents du point de vue de la définition.

3.2. L'organisation de SiLex

La description d'une information terminologique n'est pas le seul problème auquel est confrontée la réalisation de SiLex, l'organisation de son contenu interne est aussi un point crucial. En effet, SiLex est destiné à aider l'utilisateur à accéder à la base de connaissances et aux documents au moyen d'une navigation entre les termes utilisés dans ces deux structures. Son organisation doit donc offrir à l'utilisateur un moyen naturel de passer de termes assez généraux à des termes spécifiques correspondant à la question qu'il se pose. Le problème est donc d'organiser les items, non pas d'une façon compatible avec une modélisation du monde mais d'une façon efficace pour une navigation par un utilisateur humain. Ainsi, alors que dans une base de connaissances telle que ColiG ne il peut y avoir un lien entre Ogné et Ot-RNAO, dans SiLex ces deux items font partie du sous-domaine des objets biologiques, mais appartiennent à des classes différentes de termes et ne sont donc pas reliés de façon directe.

Il n'existe pas, dans la littérature, de critères stricts permettant d'évaluer les distances entre items de façon à les organiser. Cependant les notions psychologiques de réseaux sémantiques (Quillian, 1968) et de prototypes (Rosch, 1976) influent sur l'organisation d'un lexique. En effet, la notion de prototype, c'est-à-dire d'instance la plus représentative d'un concept et, par extension, de terme plus facilement utilisé, peut contraindre l'organisation du lexique en général. Mais cette notion est-elle directement applicable au vocabulaire de la biologie moléculaire par exemple? De même, des études ont été menées sur l'organisation du lexique mental mais les résultats actuels sont trop prospectifs pour fournir des indications de conception.

Pour pallier ce manque de critères et obtenir une structure efficace, deux solutions techniques sont intéressantes à envisager. Une première solution consiste à proposer une structure dynamique pour SiLex. C'est-à-dire que la structure informatique de SiLex doit permettre une réorganisation facile de son

contenu au cours de la conception. Cette réorganisation ne concerne pas uniquement l'ajout ou le retrait de termes, mais aussi la possibilité de définir de nouveaux domaines, de restructurer les hiérarchies et les liens entre celles-ci.

La seconde solution provient du fait que les facteurs intervenant lors d'une session de consultation sont peu connus. Il est donc intéressant d'utiliser les traces de navigation des consultations effectuées par les utilisateurs afin d'adapter en conséquence l'organisation du lexique. Ces traces sont facilement conservées avec les logiciels de navigation actuels, mais elles sont peu utilisées pour valider la qualité des organisations proposées: chemins effectivement parcourus, noeuds ouverts, etc.

SiLex est donc une structure dont le contenu et l'organisation sont difficiles à définir et qui peuvent évoluer au cours de la construction incrémentale de la base de connaissances. La gestion de ces évolutions concerne notamment le maintien de la cohérence au sein de l'environnement.

3.3. Intégration de SiLex dans l'environnement de partage

Le problème de la gestion de la cohérence au sein d'une structure est assez classique, mais celui de la gestion de structures relativement indépendantes et qui évoluent à des rythmes différents est plus nouveau. Or, SiLex ne peut pas être considéré indépendamment des deux autres structures auxquelles il est relié. En effet, il est fortement connecté à la base de connaissances et à la base de documents étant donné qu'il constitue l'interface d'accès de ces deux structures. L'évolution des trois structures ne se fait donc pas de façon totalement indépendante. L'évolution normale est l'ajout d'une nouvelle connaissance, accompagné de l'ajout de nouveaux documents et, si besoin, de nouveaux termes formels ou descripteurs. Un autre type d'évolution est le retrait de certaines connaissances qui ne sont plus valides, ce qui peut conduire à retirer les documents qui documentent ces seules connaissances et peut-être des termes associés.

D'un point de vue informatique, maintenir la cohérence entre trois structures consiste, par exemple au niveau le plus simple, à garantir que chaque terme utilisé dans l'environnement est défini dans SiLex. À un niveau plus complexe, on peut imaginer que si dans le lexique un terme est lié à un autre par un lien d'implication, la vérification de cohérence oblige l'utilisateur, quand il modifie l'objet formel défini par le premier terme, à vérifier la cohérence de la relation d'implication avec le second objet de la relation.

La gestion de la cohérence, pour être appropriée, doit aussi tenir compte de la dynamique des connaissances scientifiques. Ainsi, lors de la conception de SiLex, il est intéressant de tenir compte de l'observation selon laquelle dans la terminologie d'un domaine l'évolution porte généralement sur les définitions et selon laquelle il est rare qu'une notion soit renommée, même si son contenu conceptuel change.

La conception de l'environnement, et plus particulièrement de SiLex, ne peut donc se faire sans tenir compte des études existantes en terminologie et en linguistique. Cette conception doit de plus s'appuyer

sur les connaissances actuelles en psychologie concernant la représentation des connaissances et le traitement du langage.

Conclusion

Le problème du partage des connaissances a été envisagé ici comme un problème de lisibilité des connaissances à partager. Ce problème de lisibilité est décrit comme la difficulté pour un utilisateur à traduire la représentation formelle réalisée par un autre en une représentation mentale qu'il puisse assimiler et manipuler. Afin d'améliorer la lisibilité des bases de connaissances, nous avons proposé un environnement qui facilite la documentation du contenu des bases et qui de plus prend en compte le fonctionnement et les connaissances de l'utilisateur en lui permettant d'utiliser sa propre connaissance du vocabulaire du domaine, grâce à SiLex, pour interagir avec la base de connaissances. Cet objectif contraint fortement la conception de SiLex tant au niveau de son contenu que de son organisation.

La réalisation d'un prototype d'environnement de partage de connaissances et son utilisation par des biologistes pour construire une base de connaissances doit nous permettre de vérifier la justesse de nos hypothèses. Ce prototype doit aussi nous permettre de voir si un tel environnement aide à une meilleure conception des bases de connaissances, notamment en obligeant les concepteurs à définir clairement le vocabulaire utilisé.

Remerciements

Je tiens à remercier Gabriel Otman dont les remarques m'ont aidé à améliorer cet article.

Bibliographie

[Agosti, 1992] Agosti M., Gradenigo G and Marchetti P.G., A Hypertext Environment for Interacting with Large Textual Databases. *Information Processing & Management*, 28(3):371-387, 1992.

[Bourigault, 1995] Bourigault D., Réflexions sur le concept de base de connaissances terminologiques. *Actes des 5es Journées Nationales du PRC-GDR Intelligence Artificielle*, Nancy, février 1995

[Condamines, 1994] Condamines A., Terminologie et Représentation des connaissances. *La Banque des Mots*, 6 : 29-44, 1994

[Gaines, 1990] Gaines B.R. et Linster M., Integrating a Knowledge Acquisition Tool, an Expert System Shell and a Hypermedia System. *International Journal of Expert Systems*, vol 3 (2): 105-129, 1990.

[Grivaud, 1992] Grivaud S. et Rechenmann F., Navigation dans les bases de connaissances associant objets et hypertexte. *Représentation par objets (RPO)*, La Grande Motte, 22-23 juin 1992

[Gruber, 1993] Gruber T.R., *Toward Principles for the Design of Ontologies Used for Knowledge Sharing*. Technical Report KSL 93-04, Knowledge Systems Laboratory, Stanford University, 1993.

[Jacobs, 1993] Jacobs P.S et Rau L., Innovations in text interpretation. *Artificial Intelligence* 63 : 143-191, 1993.

[Moia, 1990] Moia M., Expert systems and Hypertext: a promising integration for training. *Computational*

Intelligence II, F. Gardin and G. Mauri (Eds), Elsevier Science Publishers B.V (North Holland), pp 37-48, 1990.

[Meyer, 1992] Meyer I., Skuce D., Bowker L., Eck K. Towards a new generation of terminological resources: an experiment in building a terminological knowledge base, *Proceedings of COLING'92*, Nantes, 23-28 août, 1992

[Onyshkevych, 1992] Onyshkevych B. et Nirenburg S., Lexicon, Ontology, and Text Meaning. in *Lexical Semantics and Knowledge Representation*, J. Pustejovsky, ed. Heidelberg: Springer Verlag, 1992

[Perriere, 1993] Perrière G., Dorkled F., Rechenmann F. et Gautier C., Object-Oriented Knowledge Bases for the Analysis of Prokaryotic and Eukaryotic Genomes. *Proceedings of the First International Conference on Intelligent Systems for Molecular Biology*. Bethesda, MD, jul 6-9, 1993.

[Quillian, 1968] Quillian M.R., Semantic memory. in *Semantic information processing*, M. Minsky (ed), Cambridge, Mass., MIT Press, 1968

[Rastier, 1996] Rastier F., Le terme : entre ontologie et linguistique, *Banque des mots*, 7: 35-64, 1995

[Rey, 1992] Rey A., *La terminologie: noms et notions*. Que Sais-je?, PUF, 1992.

[Rosch, 1976] Rosch E., Mervis C.B., Gray W., Johnson D. et Boyes-Braem P., Basic objects in natural categories. *Cognitive Psychology* 8: 382-439, 1976

[Samuels, 1994] Samuels P., Hypertext for Computational Mathematics., In *Artificial Intelligence in Mathematics*. J.H. Johnson, S. McKee & A.Velle (Eds), Oxford University Press, 1994.

[Schmeltzer, 1993] Schmeltzer O., Mždigue C. Uvietta P. Rechenmann F., Dorkled F., Perrière G. et Gautier C., Building Large Knowledge Bases in Molecular Biology. *Proceedings of the First International Conference on Intelligent Systems for Molecular Biology*. jul 6-9, Bethesda, MD, 1993.

[Van de Riet, 1994] Van de Riet R.P., Linguistic Instruments in Knowledge Engineering. A research proposal and some experiments. in K.Fuchi, T.Yokoi (eds), *Knowledge building and knowledge sharing*, IOS Press, Amsterdam, 1994

[Wick, 1992] Wick M.R. and Thompson W.B. Reconstructive expert system explanation. *Artificial Intelligence* 54 : 33-70, 1992