Intégration de connaissances terminologiques dans les grandes bases d'objets Exemples en biologie moléculaire

Florence LEMAIRE et François RECHENMANN
INRIA Rhône-Alpes et IMAG/LIFIA
46, avenue Félix Viallet
38031 Grenoble cedex 1
{Florence.Lemaire,Francois.Rechenmann}@inria.fr

Les bases de connaissances et la résolution de problèmes

Les bases de connaissances sont le plus souvent vues en tant que composants des systèmes dits justement à base de connaissances. Ces systèmes reposent sur la recherche d'une séparation explicite entre l'expression des connaissances et les mécanismes d'exploitation de ces connaissances pour la résolution de problèmes. Lorsque ces connaissances sont spécifiques d'un individu ou d'un groupe d'individus, parce que acquises à travers leurs expériences passées, ces systèmes sont qualifiés d'experts. Le qualificatif "expert" fait ainsi référence à la nature des connaissances impliquées et aux performances du système, alors que le terme "à base de connaissances" fait référence à l'architecture informatique qui réalise la séparation connaissances / mécanismes d'exploitation. Cette séparation est réputée apporter des avantages décisifs quant aux possibilités, d'une part de faire évoluer les connaissances, d'autre part de fournir des justifications sur les solutions obtenues et des explications sur le processus de résolution. Ces avantages sont particulièrement appréciables lorsque les connaissances sont "expertes" : ces connaissances étant non consensuelles, d'une part il est difficile de les expliciter en une seule fois et le processus de construction de la base de connaissances est incrémental, d'autre part les solutions qu'elles permettent d'obtenir doivent être argumentées.

Les bases de connaissances comme modèles

Mais le développement de bases de connaissances peut se concevoir en dehors de toute préoccupation de résolution de problèmes. En effet, une base de connaissances peut être considérée comme un modèle d'un domaine, en particulier scientifique ou technique. Le bénéfice d'un processus de modélisation réside plus dans l'accroissement de la compréhension des entités et des phénomènes modélisés que dans l'obtention du modèle lui-même. Cet accroissement de compréhension résulte de l'explicitation et de la caractérisation des entités impliquées, ainsi que de leurs inter-relations. Cette situation est certes reconnue par les concepteurs de systèmes de résolution de problèmes à base de connaissances, mais la priorité est malgré tout donnée, fort justement compte tenu des objectifs poursuivis, à l'obtention d'un système efficace, et l'accroissement de compréhension est considérée comme un effet collatéral.

À l'inverse, la modélisation des connaissances dans un domaine tel que la biologie moléculaire vise d'abord, à travers l'effort d'explicitation, de structuration et d'organisation des connaissances, une meilleure compréhension des entités et de leurs interactions. Certes, la base résultante peut disposer de capacités prédictives et calculatoires pertinentes, mais sa seule existence justifie sa construction.

Objets et relations en biologie moléculaire

La macro-molécule d'ADN, principal constituant du génome, est le support matériel de l'hérédité et la mémoire du processus d'évolution. Sous la forme d'une séquence d'acides nucléiques symbolisés chacun par une des lettres A, C, G ou T, elle contient l'information nécessaire au développement et au maintien de tout organisme vivant. Certains fragments de cette séquence jouent un rôle particulièrement important, ce sont les gènes codant les protéines.

Mais de nombreux autres objets apparaissent au sein de réseaux d'interactions complexes. Le produit de l'expression d'un gène peut intervenir dans la régulation d'un ou de plusieurs autres gènes. Des gènes peuvent être organisés en des structures d'ordre supérieur. Ainsi, dans un opéron bactérien, un ensemble de gènes est sous le contrôle d'un seul signal de régulation.

Les bases de connaissances en biologie moléculaire cherchent à représenter ces différentes classes d'objets et les relations qu'elles entretiennent [Perrière&93].

Les modèles de connaissances[1] à objets se révèlent beaucoup mieux adaptés à la représentation des entités et de leurs interrelations que les modèles de règles. Ces derniers sont en revanche appréciés pour le développement de systèmes ayant pour vocation première la résolution de problèmes.

Les modèles de connaissances à objets

L'idée fondamentale des modèles à objets est de rassembler dans une même unité syntaxique de description la connaissance relative à un objet ou une classe d'objets. Une telle unité est ainsi composée d'un nom et d'une liste d'attributs. Chaque attribut possède lui-même un nom et une liste de descripteurs, qui lui associent différentes informations : le type de ses valeurs, des moyens de calculer ces valeurs quand elles sont inconnues, par exemple des procédures de calcul (attachement procédural) ou des valeurs par défaut. Le type d'un attribut peut être une référence à un autre objet, introduisant la notion d'objet complexe : un objet qui est relié à d'autres à travers les valeurs de ses attributs. Une capacité intéressante des modèles à objets consiste à pouvoir introduire une description comme variante d'une autre. Dans ce cas, la nouvelle description hérite de la précédente ses attributs et descripteurs et ajoute ses propres éléments de description supplémentaires.

Les modèles à objets autorisent la description sous une forme hautement structurée de connaissances descriptives. Outre les mécanismes d'héritage d'attributs et d'obtention de valeurs par attachement procédural et par défaut, la classification permet de caractériser une entité en recherchant les classes auxquelles elle est susceptible d'appartenir compte tenu des valeurs de ses attributs, connues ou à déterminer.

Il est important de souligner ici la différence fondamentale qui distingue, malgré les similitudes apparentes, les modèles de connaissances à objets, qui sont destinés à décrire des objets du monde réel, des langages de programmation à objets, qui sont destinés à

[1] L'expression "modèle de connaissances" est employée par similitude avec l'expression "modèle de données". Elle est préférée à l'expression pléonastique "modèle de représentation de connaissances". Le terme "modèle" est ainsi employé dans deux acceptions différentes : pour désigner le formalisme dans lequel les connaissances sont exprimées, et pour désigner le résultat de ce processus de formalisation.

décrire des entités qui résultent déjà d'une étape préalable d'abstraction. Cette distinction apparaît très nettement à propos de la notion de comportement.

Dans un langage de programmation à objets, ce terme désigne l'ensemble des méthodes qui sont associées à une classe et qui portent sur les entités informatiques éléments de cette classe. Par exemple, une fenêtre d'une interface possède un comportement déterminé par les méthodes d'ouverture, de déplacement ou de changement de taille. Certes, ces méthodes constituent bien un modèle du comportement des fenêtres, mais cette classe d'objets résulte d'une étape préalable d'abstraction. En l'occurrence, la fenêtre est un artefact.

Dans un modèle de connaissances à objets, le terme "comportement" doit faire référence au comportement de l'objet réel, et non pas au comportement de l'entité informatique qui sert de support à la modélisation de l'objet réel. Ce dernier comportement est d'ailleurs entièrement déterminé par la sémantique du modèle de connaissances.

Utiliser un langage de programmation à objets pour modéliser directement des objets du monde réel conduit inéluctablement à mélanger ces deux niveaux de description. À titre d'exemple, modéliser les opérons directement par une classe dans un langage de programmation à objets conduirait à définir les méthodes décrivant le processus de régulation d'un opéron au même niveau que les méthodes qui impriment, ou sauvent sur disque, l'instance – entité informatique – qui "représente" cet opéron.

Le partage des connaissances

Le grand intérêt des systèmes de résolution de problèmes est de mettre à disposition d'un plus grand nombre des capacités de résolution détenues jusque là par quelques uns. Ceci constitue une première interprétation du partage de connaissances.

Dans une seconde interprétation, plus récente et illustrée par le "ARPA Knowledge Sharing Initiative", il s'agit de partager les connaissances dans l'état formalisé, sous lequel elles figurent dans une base. Par analogie avec le génie logiciel, l'espoir est de "réutiliser" ces connaissances pour le développement d'autres systèmes destinés à résoudre des classes de problèmes apparentées mais différentes [Neches&91]. Cette ambition rencontre une objection majeure liée au statut de modèle de toute base de connaissances : que ce soit en vue de la résolution de problèmes ou que ce soit pour mieux comprendre un domaine, elle est conçue avec des objectifs spécifiques. Réutiliser une telle base dans un contexte différent fait courir un double risque : celui d'obtenir une base finale inadaptée aux objectifs présents, et celui, plus subtil, de ne pas bénéficier de l'apport du processus même de construction, qui aura été ainsi court-circuité par la réutilisation.

Le problème n'est donc plus tant de partager les bases de connaissances que de partager leurs processus de construction. Deux solutions sont envisageables. La première consiste à permettre une construction concourante des bases à travers la spécification de protocoles d'interaction adaptés [Rechenmann93], [Euzenat95]. La seconde vise à compléter la base de connaissances formelles par des connaissances contextuelles. Il est important de souligner que ces connaissances contextuelles ne peuvent être exprimées dans le même modèle que les connaissances de la base, sous peine d'empiler les contextes sans limite. Il s'agit donc d'associer aux éléments de connaissances formelles, typiquement des classes ou des instances dans un modèle à objets, des énoncés en langage naturel qui commentent et justifient les choix de modélisation.

Connaissances formelles et textuelles

Une première solution immédiate consiste à associer aux noms de classes, d'attributs et d'instances un texte explicatif ou justificatif. La transformation de ces morceaux de texte en nœuds d'un réseau hypertexte peut être automatisée si les noms de la base sont

systématiquement utilisés pour les indexer [Grivaud92]. Il est alors possible de naviguer dans la base de connaissances (en empruntant les relations entre objets) et dans l'hypertexte, de passer à tout moment de la base à l'hypertexte et inversement.

Une grande partie des connaissances sur lesquelles repose la modélisation d'un domaine, tel que la biologie moléculaire, provient des articles scientifiques publiés dans les revues. Compléter une base formelle de ses sources revient donc à relier les noms de la base aux articles, ou parties d'articles, qui concernent les concepts nommés. La limitation majeure de la solution précédente est que la liaison entre la base et son contexte repose exclusivement sur les noms utilisés dans la base. Or, ces noms ne sont pas toujours significatifs. De plus, ils peuvent apparaître dans les textes sous des variantes lexicales, sémantiques ou multilingues. Il est donc indispensable d'établir une interface entre la base de connaissances formelles et les textes. Un lexique est susceptible de jouer ce rôle [van de Riet93].

En fait, dans ce contexte, un lexique est amené à jouer simultanément plusieurs rôles complémentaires [Lemaire95]. Le premier est effectivement de fournir une indirection dans la liaison entre les entités formelles de la base et les textes qui les commentent et les justifient. Mais le lexique constitue également une interface entre la base et l'utilisateur lui-même, donnant à ce dernier la possibilité, en navigant d'un terme à l'autre, de prendre connaissance de la base de connaissances. Cette navigation doit permettre à l'utilisateur de mieux appréhender le domaine traité dans la base et de situer celui-ci par rapport à des domaines connexes. C'est à travers cette prise de connaissance que des modifications ou des ajouts peuvent être apportés aux connaissances formelles. Enfin, un lexique peut conduire à un choix plus judicieux et systématique des noms de classes et d'attributs.

Un exemple en biologie moléculaire

Dans une base de connaissances sur le génome bactérien, la classe `opéron` est décrite par les attributs `structure`, `activateur` et `répresseur`. Ses sous-classes immédiates distinguent les opérons suivant leur mode de régulation : par induction ou par répression. À un niveau encore inférieur, les opérons identifiés dans une espèce particulière de bactérie sont précisément décrits. Par exemple, pour l'opéron tryptophane d'*E. coli*, l'attribut `répresseur` prend la valeur *trpR* et les gènes dans l'attribut `structure` sont *trpE*, *trpD*, *trpC*, *trpB* et *trpA*. Ces valeurs sont elles-mêmes des instances de la classe `gène` et de ses sous-classes.

Dans le lexique, le terme *répresseur* est décrit comme "une protéine multimérique se liant à l'ADN". La notion de *répresseur* peut être reliée à deux types de *contrôle négatif*: la *répression* et l'*induction.* La définition de la notion de répression, "Capacité des bactéries d'empêcher la synthèse de certaines enzymes quand leurs produits sont présents par l'inhibition de la traduction ou de la transcription", met en évidence d'autres relations avec des notions telles que la transcription ou la synthèse enzymatique.

À coté des connaissances formalisées, il est donc proposé d'énoncer des connaissances terminologiques destinées à les relier à leur contexte d'expression et d'acquisition. Deux problèmes se posent : le premier de nature conceptuel, le second de nature technique.

Le premier problème est lié à la définition des caractéristiques qui permettent de décrire un terme d'un point de vue lexical par contraste notamment aux caractéristiques concernant l'objet dénoté par ce terme. La bonne définition de ces caractéristiques est importante pour deux raisons. D'une part, elle doit permettre d'éviter des confusions au sujet de la signification des relations entre les différentes entités. Ainsi la relation de subsumption entre deux lexèmes n'a pas le même sens que celle entre deux concepts. En effet, la subsumtion entre deux lexèmes concerne par exemple des constructions grammaticales alors que celle entre deux objets concerne des propriétés de ceux-ci observés dans le monde réel. D'autre part, elle doit permettre de limiter la redondance inévitable entre les connaissances exprimées dans le modèle formel et celles exprimées

dans le lexique. Ce problème se poserait de façon particulièrement aiguë si la constitution de la base terminologique précédait la constitution de la base formelle, démarche souvent préconisée en acquisition de connaissances, mais qui n'est pas en accord avec la perception, qui a été exposée précédemment, de l'intérêt de construire des bases de connaissances.

Le second problème, plus simple, concerne la représentation des termes au sein du lexique. Il serait tentant, pour des raisons d'économie et de minimalité, d'utiliser le modèle à objets, qui sert à représenter les connaissances de la base, pour représenter les entités du lexique. Il est clair qu'il s'agirait là d'un bel exemple de confusion entre modèle de connaissances et langage de programmation. En l'occurrence, la confusion serait faite entre le terme, qui peut être représenté par une entité informatique dans un langage de programmation à objets, et le concept du monde réel dénoté par le terme et déjà décrit dans la base formelle.

Une telle distinction entre connaissances terminologiques et connaissances sur les objets doit donc être préservée tant au niveau des informations utilisées pour leur description qu'au niveau du type de représentation choisi. Cette distinction doit permettre une évolution indépendante des deux structures, même si à l'origine l'élaboration du lexique n'est pas indépendante de l'objectif assigné à la base de connaissances. L'existence de cet objectif initial permet de s'interroger sur le degré de réutilisabilité d'une base terminologique pour annoter d'autres bases.

Conclusion

Les connaissances terminologiques apparaissent donc comme un complément naturel des connaissances formelles. Elles permettent de placer la base de connaissances dans son contexte d'élaboration et donc de faire participer ses utilisateurs au processus de modélisation dont elle est l'aboutissement. C'est à travers ce partage différé du processus de modélisation que réside l'espoir de partager les connaissances de la base avec des personnes qui n'ont pas participé à sa construction.

Références bibliographiques

[Euzenat 95]
Jérôme Euzenat
Building consensual knowledge bases: context and architecture
KB&KS'95 (Second International Conference on Building and Sharing of Very Large-Scale Knowledge Bases), Enschede (NL), 10-13 avril 1995

[Grivaud&92]
Sylvain Grivaud, François Rechenmann
Navigation dans les bases de connaissances associant objets et hypertexte
Actes Représentations Par Objets (RPO), La Grande Motte (FR), p. 262-280, (22-23 juin) 1992

[Lemaire95]
Florence Lemaire
Knowledge Bases, Texts and Lexicon
KB&KS'95 (Second International Conference on Building and Sharing of Very Large-Scale Knowledge Bases), Enschede (NL), 10-13 avril 1995

[Neches&91]
R. Neches *et al.*
Enabling Technology for Knowledge Sharing
AI Magazine, Vol. 12, n° 3, p. 36-56, 1991

[Perrière&93]
Guy Perrière, Frank Dorkeld, François Rechenmann, Christian Gautier
Object-oriented knowledge bases for the analysis of prokaryotic and eukaryotic genomes
Actes 1st international conference on Intelligent Systems for Molecular Biology (ISMB'93), Bethesda (US), p. 319-327, 6-9 juillet 1993

[Rechenmann93]
François Rechenmann
Building and sharing large knowledge bases in molecular genetics
dans K. Fuchi et T. Yokoi (Eds), Knowledge Building and Knowledge Sharing, Ohrosha (Tokyo) and IOS Press (Amsterdam), 1994

[van de Riet93]
R.P. van de Riet
Linguistic instruments in knowledge engineering, a research proposal and some experiments
dans K. Fuchi et T. Yokoi (Eds), Knowledge Building and Knowledge Sharing, p. 200-207, Ohrosha (Tokyo) and IOS Press (Amsterdam), 1994