Treillis de concepts et ontologies pour interroger l'annuaire de sources de données biologiques BioRegistry

Nizar Messai, Marie-Dominique Devignes, Amedeo Napoli, Malika Smaïl-Tabbone

Treillis de concepts et ontologies pour interroger l'annuaire de sources de données biologiques BioRegistry

N. Messai, M-D. Devignes, A. Napoli, M. Smaïl-Tabbone

UMR 7503 LORIA, BP 239, 54506 Vandœuvre Lès Nancy, FRANCE.
{messai,devignes,napoli,malika}@loria.fr ; http://www.loria.fr/equipes/orpailleur

RÉSUMÉ. *Les sources de données biologiques disponibles sur le Web aujourd'hui sont multiples et hétérogènes. L'utilisation optimale de ces ressources nécessite de la part des utilisateurs des compétences à la fois en informatique et en biologie, à cause du manque de documentation et des difficultés d'interaction avec les sources de données. De fait, les contenus de ces sources de données restent souvent sous-exploités. Nous présentons ici une approche qui s'appuie sur l'analyse de concepts formels, pour organiser et rechercher des sources de données biologiques pertinentes pour satisfaire une requête donnée. Le travail consiste à construire un treillis de concepts à partir des méta-données associées aux sources. Le concept construit à partir d'une requête donnée est alors classifié dans le treillis. La réponse à la requête est ensuite fournie par l'extraction des sources de données appartenant aux extensions des concepts subsumant le concept requête dans le treillis. Les sources ainsi retournées peuvent être triées selon l'ordre de spécificité des concepts dans le treillis. Une procédure de raffinement de requête, s'appuyant sur des ontologies de domaines, permet d'améliorer le rappel par généralisation ou par spécialisation*

ABSTRACT. *Bioinformatic data sources available on the Web are multiple and heterogenous. The lack of documentation and the difficulty of interaction with these data sources require users competence in both informatics and biological fields for an optimal use of sources contents that remain rather under exploited. In this paper we present an approach based on formal concept analysis to classify and search relevant bioinformatic data sources for a given query. It consists in building the concept lattice from the binary relation between bioinformatic data sources and their associated metadata. The concept built from a given query is then merged into the concept lattice. The result is given by the extraction of the set of sources belonging to the extents of the query concept subsumers in the resulting concept lattice. The sources ranking is given by the concept specificity order in the concept lattice. An improvement of the approach consists in automatic query refinement thanks to domain ontologies. Two forms of refinement are possible by generalisation and by specialisation.*

MOTS-CLÉS *méta-données, bioinformatique, treillis de concepts, ontologies, sources de données.*

KEYWORDS: *metadata, bioinformatics, concept lattices, ontologies, data sources.*

1. Introduction

L'un des grands défis de la bioinformatique aujourd'hui est de permettre aux biologistes d'accéder efficacement aux données gisant dans les centaines de sources de données réparties à travers le monde. Le grand nombre de sources, leur hétérogénéité et la complexité des objets biologiques auxquels elles font référence rendent souvent difficile la mise en relation d'une requête avec la source appropriée. Souvent la requête elle-même doit être décomposée en requêtes élémentaires susceptibles d'être adressées à des sources distinctes. Différentes approches ont été adoptées pour unifier l'accès aux diverses sources de données face à une requête donnée. Des systèmes ont été produits, à partir d'entrepôts de données (tels que *GUS* (Davidson et al., 2000)), de fédération de bases de données (tels que *SEMEDA* (Köhler et al., 2003)), ou de médiateurs (tels que *TAMBIS* (Goble et al., 2001)). Certains de ces systèmes comme *TAMBIS* ou *SEMEDA* sont capables d'interpréter de façon sémantique la requête. Cependant la plupart des systèmes disponibles ne prend en compte qu'un petit nombre de sources de données et ne peut donc satisfaire un large éventail de requêtes. Le travail présenté ici vise à organiser les sources de données biologiques en fonction des connaissances disponibles sur ces sources afin de proposer aux utilisateurs les sources de données qui répondent au mieux à leur besoin. Le problème ici n'est pas l'interrogation des sources elles-mêmes mais plutôt l'identification et le choix parmi toutes les sources de données disponibles de celles qui sont les plus appropriées par rapport à un besoin donné.

Dans cet article, nous proposons de considérer ce problème comme un cas particulier de Recherche d'Information (*IR* pour *Information Retrieval*) dans lequel des sources de données plutôt que des documents font l'objet de la recherche et où l'indexation s'appuie sur les méta-données associées aux sources de données plutôt que les termes extraits des documents. Les sources de données sont annotées grâce à un annuaire appelé BioRegistry mis au point par des experts biologistes pour les besoins de recherche d'information. L'analyse de concepts formels (de l'anglais *Formal Concept Analysis* abrégé en *FCA*) est ensuite utilisée pour construire une classification flexible et dynamique des sources existantes. En effet la multiplicité et l'hétérogénéité des sources de données biologiques conduit à une méconnaissance sur les contenus et les propriétés de ces sources qui empêche de formuler correctement une requête sur l'annuaire des sources. La possibilité de naviguer dans une classification des sources et des propriétés partagées peut conduire à découvrir de nouvelles sources pertinentes et à améliorer la formulation de requêtes dans le cas de besoins précis. De plus les ontologies de domaines sont prises en compte dans l'indexation des sources, ce qui permet de traiter les requêtes de façon sémantique pour améliorer la recherche.

Nous commencerons par décrire et discuter brièvement en section 2 des travaux apparentés, impliquant *FCA* et *IR*, ainsi que l'usage d'ontologies de domaines, dans des problèmes similaires. La section 3 présentera le projet BioRegistry comme un annuaire structuré répertoriant les méta-données associées aux sources de données

biologiques. La formalisation de notre problème grâce à l'analyse de concepts formels sera détaillée en section 4 et les aspects relatifs à la recherche d'information seront développés en section 5, avec en particulier une technique originale de raffinement de requête. Une conclusion et quelques perspectives de ce travail seront présentées en section 6.

2. Travaux voisins

2.1. *Utilisation de treillis de concepts pour la recherche d'information*

Les treillis de concepts ont été appliqués dans la recherche d'information (IR) dès l'apparition de l'analyse de concepts formels (Wille, 1982). En effet une analogie existe entre les tableaux *objets* × *attributs* et *documents* × *termes*. La recherche d'information a par la suite été explicitement mentionnée comme étant l'une des applications possibles des treillis de concepts (Godin et al., 1995a). Les concepts formels sont vus comme des classes de documents pertinents pour un ensemble de contraintes données (représenté par une requête). La relation de subsomption (la relation d'ordre partiel entre les concepts du treillis) entre les concepts permet le passage d'un concept (ou d'une requête) à un autre concept plus général ou plus spécifique. Une approche pour la recherche d'information en utilisant les treillis de concepts, la recherche d'information par treillis, a été proposée dans (Carpineto et al., 2000, Carpineto et al., 2004). Dans les deux propositions (Godin et al., 1995a) et (Carpineto et al., 2000), la recherche d'information par treillis atteint des performances qui dépassent celles de la recherche booléenne classique. Une limite de la recherche d'information par treillis est la complexité du treillis (nombre de concepts) pour les contextes volumineux en nombre d'objets ou d'attributs. Cependant dans les applications réelles, la complexité théorique maximale n'est pas atteinte (Carpineto et al., 2000). De plus des solutions visant à contrôler la taille des treillis correspondant aux grands contextes ont été proposées dans (Pernelle et al., 2002) et (Stumme et al., 2002).

2.2. *Amélioration des performances de la recherche d'information en utilisant les ontologies de domaines*

Le raffinement de requête est un mécanisme visant à améliorer les performances de la recherche d'information en ajoutant à la requête de nouveaux termes liés à ceux initialement présents dans la requête (Carmel et al., 2002). La combinaison des ontologies de domaines et de l'analyse de concepts formels dans le but d'améliorer les performances de la recherche d'information a fait l'objet des propositions (Carpineto et al., 2000, Priss, 2000, Safar et al., 2004). Dans les deux premières propositions, un thésaurus est utilisé pour améliorer la qualité du processus de recherche d'information en enrichissant l'indexation dans le treillis par de nouveaux attributs aux concepts du treillis. Dans la troisième proposition, les ontologies de

domaines sont utilisées pour guider la construction du treillis selon les préférences des utilisateurs, ce qui permet d'éviter la construction du treillis complet. Les deux approches, (Carpineto et al., 2000, Priss, 2000) et (Safar et al., 2004), modifient directement le treillis soit en ajoutant des termes d'indexation soit en ne considérant qu'une partie du treillis.

Dans notre travail, les ontologies de domaines sont prises en compte très tôt dans le processus de recherche d'information (dès la construction de BioRegistry). Ceci nous amène à proposer une méthode consistant à modifier la requête plutôt que le treillis pour améliorer le processus de recherche d'information (discuté en section 5.3).

2.3. *Algorithmes de construction incrémentale de treillis de concepts*

Le problème de la construction des concepts d'un treillis à partir d'un contexte formel a fait l'objet de plusieurs travaux de recherche. Une comparaison détaillée des performances des algorithmes proposés pour la génération de treillis et des diagrammes de Hasse correspondants est présentée dans (Kuznetsov et al., 2002). Parmi ces algorithmes proposés, quelques-uns effectuent une construction incrémentale des treillis de concepts à partir de contextes formels (Godin et al., 1995b, Carpineto et al., 1996 et Van der Merwe et al., 2004). Cet aspect est particulièrement intéressant pour l'application des treillis de concepts à la recherche d'information en général et à notre problème de recherche de sources de données biologiques en particulier. En effet, les requêtes utilisateurs peuvent être insérées dans le treillis représentant la collection de documents (ou de sources de données). Suite à cette insertion il est possible de déterminer les documents les plus pertinents répondants aux critères exprimés par l'utilisateur dans sa requête. La construction incrémentale des treillis de concepts permet l'ajout de nouveaux concepts et rend possible la prise en compte de nouvelles sources de données biologiques apparaissant sur le Web. Dans notre cas, cet ajout est essentiel pour tenir compte des indispensables mises à jour de l'annuaire BioRegistry (décrit en section 3).

3. Le projet BioRegistry

3.1. *Les sources de données biologiques*

Plus de sept cents sources de données biologiques sont connues de nos jours (Galperin, 2005). De nombreux efforts ont été consacrés jusqu'à présent aux problèmes posés par la standardisation de l'accès à ces sources, le traitement des requêtes en vue de leur distribution (répartition) sur les sources pertinentes, l'intégration des réponses, etc. Ces tâches nécessitent de concevoir des scénarios appropriés et de disposer d'une interopérabilité transparente entre les ressources. Des systèmes intégrés ont été développés à partir d'architectures d'entrepôts de

données ou de médiation (Davidson et al., 2000, Goble et al., 2001). Les solutions actuelles sont envisagées aussi dans le contexte du Web sémantique avec en particulier la composition de services Web (Oinn et al., 2004). Pour que toutes ces solutions deviennent réellement efficaces, il est indispensable que les connaissances disponibles sur les sources de données puissent être exploitées. Par exemple, une requête apparemment simple comme: « *Quels sont les gènes du chromosome X humain qui sont exprimés préférentiellement dans le cerveau ? »* met en jeu des données dites *de cartographie* et *d'expression*, qui peuvent à un instant donné être contenues dans une source unique ou dans des sources distinctes. Vraisemblablement, il sera possible de trouver plusieurs sources pertinentes pour chaque type de données. L'utilisateur pourra aussi préférer l'une de ces sources à cause de critères de qualité particuliers tels que la révision manuelle des données ou la fréquence de mise à jour, ou encore en raison de contraintes d'accès aux données.

Le catalogue de sources de données biologiques le plus documenté aujourd'hui est certainement *DBCAT* (Discala et al., 2000). Cet annuaire au format de fichier plat contient un ensemble de méta-données relativement restreint sur plus de 400 sources de données. Les possibilités d'interrogation sont cependant limitées par le fait que la plupart des champs sont de domaine ouvert (texte libre). D'autres annuaires sont développés aujourd'hui pour les services Web en bioinformatique comme par exemple dans les projets *MyGrid* et *BioMoby* (Lord et al., 2004, Wroe et al., 2003). La quantité d'information biologique accessible par les services Web est encore trop limitée pour répondre aux besoins des utilisateurs. Cependant cette situation pourrait changer et le besoin de modéliser et d'organiser les connaissances concernant les services Web en bioinformatique pourrait devenir aussi pressant qu'il l'est aujourd'hui pour les sources de données biologiques.

Afin de disposer d'un environnement adéquat pour tester nos propositions concernant la classification et la recherche de sources de données biologiques, nous avons décidé de construire notre propre annuaire, baptisé BioRegistry, dans lequel les méta-données associées aux sources de données biologiques sont organisées de façon dynamique, flexible et structurée. L'annuaire BioRegistry est introduit et détaillé dans les deux paragraphes suivants.

3.2. *Le modèle de méta-données de BioRegistry*

Les méta-données décrivent le contenu, la qualité et d'autres caractéristiques relatives à des données. Elle jouent un rôle essentiel dans les tâches d'indexation et d'interrogation d'ensembles de données. En 1995, un comité international d'experts a proposé un modèle standard de description de méta-données relatives aux ressources du Web : le *Dublin Core Metadata Initiative* ou *DCMI* (Dekkers et al., 2003). Ce standard est composé d'une quinzaine d'éléments : *identifier, title,*

creator, subject, description, publisher, contributor, date, type, format, language, source, rights, coverage, relation[1].

Bien que le modèle de méta-données du *DCMI* doit rester très simple, il inclut deux mécanismes permettant d'être plus précis dans la description : (i) des raffinements d'éléments (*element refinements*) tels que *created* et *modified* qui raffinent l'élément *date* pour exprimer une date de création ou de modification ; (ii) des schémas d'encodage tels que *vocabulary encoding schemes* permettant de spécifier qu'un terme est extrait d'un vocabulaire contrôlé ou *syntax encoding schemes* permettant de spécifier qu'une valeur est formatée selon certaines règles (par exemple une date dans le format *W3CDTF* : *YYYY-MM-DD*).

Le modèle des méta-données du *DCMI* étant très général, son utilisation pour décrire des ressources d'un domaine particulier nécessite souvent des raffinements voire des extensions. Une extension du *DCMI* est par exemple introduite par le *FGDC*[2], qui est le comité chargé de standardiser les méta-données sur les données géospatiales digitales. La richesse et la spécificité des sources de données biologiques nous ont également conduits à la proposition d'un modèle hiérarchique pour l'organisation des méta-données à attacher à ces sources. Le modèle de BioRegistry, schématisé sur la figure 1, comporte 3 sections (chaque élément souligné correspond à un élément du *DCMI*). La première section correspond aux méta-données associés à chaque source de données biologiques. La seconde section est consacrée à la documentation des ontologies de domaines utilisées (identifiant, version, localisation...). La troisième section contient les relations entre plusieurs sources de données.

Pour les besoins de cet article, nous nous focalisons sur la première section qui comporte quatre catégories de méta-données associées à une source de données :

– Identification de la source : nous retrouvons ici plusieurs champs du *DCMI* tels que *title*, *publisher*, *description*, *coverage/temporal*…

– Thèmes couverts par la source : nous retrouvons ici le champ *subject* du *DCMI* mais aussi un champ relatif aux organismes couverts par une source (par exemple, la source nommée *Mouse Genome DB* contient des données sur l'organisme *Mouse*).

– Qualité de la source : cette catégorie, absente du *DCMI*, est cruciale pour documenter la qualité d'une source biologique par rapport au mode de validation de ses entrées, la compatibilité par rapport aux standards, la couverture (nombre de *gènes*, nombre de *contigs*...) et l'existence de références croisées avec d'autres sources de données.

– Disponibilité de la source : cette catégorie regroupe les champs concernant les adresses des différents sites donnant accès à une source de donnée, ainsi que les contraintes d'accès pour les mondes académique et industriel (gratuité, authentification…).

---

1. http://dublincore.org/documents/dcmi-terms/
2. Federal Geographic Data Committee http://www.fgdc.gov/fgdc/fgdc.html

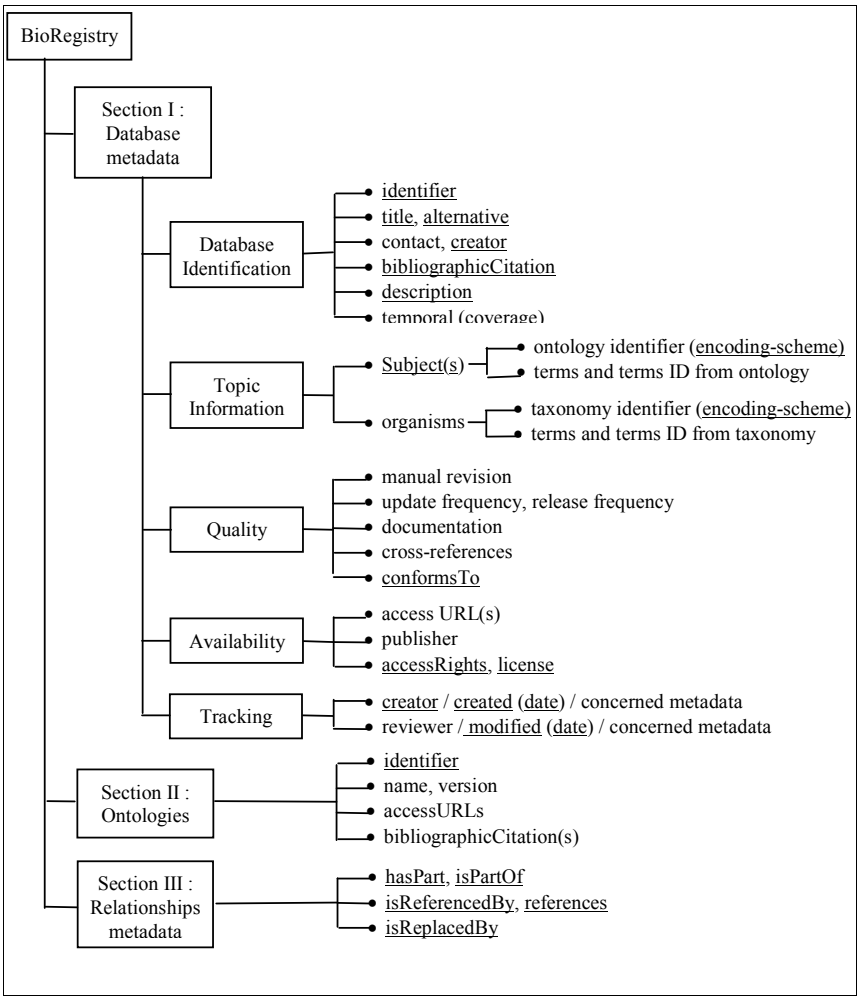

Figure 1. *Représentation schématique du modèle de BioRegistry*

Nous avons suivi les recommandations du *DCMI* en utilisant autant que possible des types de données standards pour les différents champs de notre modèle (*syntax encoding schema*) et surtout en nous appuyant sur des vocabulaires standards ou ontologies de domaines lorsqu'ils existent (*vocabulary encoding schema*). Ainsi, nous utilisons : (i) pour renseigner les sujets traités par une source de données, le

thésaurus biomédical *MeSH* maintenu par la *NLM*[3] ; (ii) pour renseigner les organismes couverts par une source, une ontologie extraite de la taxonomie des organismes vivants du *NCBI*[4] utilisée notamment pour annoter les séquences de *GenBank* et *EMBL*. Cette taxonomie étant très riche, nous en avons extrait une hiérarchie de termes présentée dans la figure 2 en partant des feuilles correspondant aux organismes modèles en biologie (tels que *la souris*, *la levure de bière*, *le riz* etc.) jusqu'à la racine (*tout organisme*) en ne conservant que les nœuds structurants. Dans la suite, cette hiérarchie de terme sera considérée comme une ontologie simplifiée qui nous servira lors de l'illustration de notre approche. En supposant que chaque nœud de l'ontologie soit défini par les propriétés communes aux groupes d'organismes correspondants aux nœuds fils, la relation entre les nœuds est une relation de spécialisation (relation d'ordre partiel « est – un »).

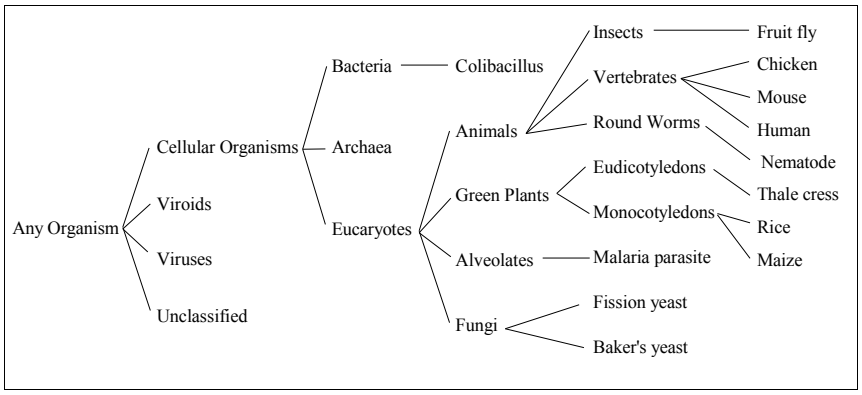

Figure 2. *Ontologie des organismes vivants (OntoBR)*

Notons enfin que le modèle de BioRegistry est ouvert et peut intégrer d'autres ontologies de domaines qui pourront coexister grâce au système d'association d'un terme d'indexation avec la référence de l'ontologie dont il est issu (cf. éléments *Subjects* et *Organisms* sur la figure 1).

Il faut encore souligner que le modèle hiérarchique de méta-données de BioRegistry a été implanté sous la forme d'un *schéma XML*. Des exemples de documents XML décrivant des sources de données biologiques dans notre modèle peuvent être consultés sur le site Web consacré à BioRegistry[5].

3. http://www.nlm.nih.gov/mesh/
4 The NCBI Taxonomy Homepage :
http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/
5. http://bioinfo.loria.fr/Members/devignes/Bioregistry/presentationBioregistry/view

3.3. *Interrogation de BioRegistry*

Une première façon d'exploiter directement BioRegistry est l'interrogation par formulaire : le biologiste doit alors entrer les valeurs de quelques champs de méta-données et reçoit en réponse une liste triée de sources de données de BioRegistry répondant à sa requête. Néanmoins cette approche oblige l'utilisateur à construire une requête, ce qui peut se révéler inefficace sans une vision globale des sources de données contenues dans BioRegistry. Une fonctionnalité de navigation à travers des groupes de sources de données partageant un nombre variable de méta-données communes serait en effet une aide précieuse pour l'utilisateur dépassé par le nombre et l'hétérogénéité de ces sources. Dans l'exemple de la requête exprimée en section 3.1 : « *Quels sont les gènes du chromosome X humain qui sont exprimés préférentiellement dans le cerveau ?* », une approche par navigation permettrait de repérer non seulement s'il existe des sources contenant à la fois des données de *cartographie* et d'*expression*, mais aussi à partir de sources voisines partageant la méta-donnée « cartographie », celles qui contiennent seulement des données sur les gènes humains ou plus particulièrement sur ceux du chromosome X humain. L'utilisateur découvrira également que parmi les sources de données dites « d'expression » certaines sont généralistes, d'autres ne recouvrent que certains types cellulaires ou certains protocoles expérimentaux. Connaissant alors l'existence et les propriétés de ces sources de données, il lui sera plus facile de formuler une requête précise pour laquelle il attendra une réponse exhaustive. Pour faire face à ces exigences, nous avons résolu d'appliquer l'analyse de concepts formels au contenu de BioRegistry afin de rendre possibles (i) une classification flexible des sources de données sur la base du partage de propriétés et (ii) une interrogation exhaustive de BioRegistry.

Par rapport à l'application de l'analyse de concepts formels à un problème de recherche d'information traditionnel (*documents* × *termes*), l'avantage ici est que BioRegistry contient beaucoup moins d'objets (de l'ordre d'un millier de sources) que la plupart des corpus documentaires, limitant ainsi l'espace de recherche et la complexité du traitement de la requête. Nous considérons ceci comme une condition suffisante pour un passage à l'échelle.

Les ontologies de domaines utilisées pour certains champs de BioRegistry (section 3.2) sont utilisables par le biologiste pour sélectionner les termes de sa requête. Elles sont également exploitées pour le raffinement de la requête en cas d'échec afin d'améliorer le rappel (section 5.3).

4. Treillis de concepts pour la classification des sources de données de BioRegistry

4.1. *Terminologie et définitions de base de l'analyse de concepts formels*

Dans cette section nous rappelons brièvement les définitions et les résultats concernant l'analyse de concepts formels nécessaires pour la suite de cet article (voir aussi (Ganter et al., 1999)).

DEFINITION.1. — (Contexte formel) Un contexte formel est un triplet $\mathcal{K} = (G, M, I)$ où G est un ensemble d'objets, M est un ensemble d'attributs et I est une relation binaire entre G et M appelée relation d'incidence de $\mathcal{K}$ et vérifiant :

$I \subseteq G \times M$ et $(g, m) \in I$ (notée aussi gIm) signifie que l'objet $g \in G$ possède l'attribut $m \in M$.

DEFINITION.2. — Soit $\mathcal{K} = (G, M, I)$ un contexte formel. Pour tout $A \subseteq G$ et $B \subseteq M$, on définit :

- $A' = \{m \in M \mid \forall g \in A, gIm\}$
- $B' = \{g \in G \mid \forall m \in B, gIm\}$.

Intuitivement, A' est l'ensemble des attributs communs à tous les objets de A et B' est l'ensemble des objets possédant tous les attributs de B.

L'opérateur $'$ est appelé opérateur de dérivation et s'applique aussi bien aux sous ensembles de G qu'aux sous ensembles de M. Cet opérateur peut se composer avec lui même, pour partir d'un sous-ensemble d'objets A, produire A' et à partir de A' produire le sous-ensemble d'objets A'' (la notation '' est utilisée pour marquer la composition).

Les opérateurs $'$ et $''$ vérifient les propriétés suivantes pour A, $A1$, $A2$ des sous ensembles de G et B, $B1$, $B2$ des sous ensembles de M :

- $A1 \subseteq A2 \Rightarrow A2' \subseteq A1'$, $B1 \subseteq B2 \Rightarrow B2' \subseteq B1'$
- $A \subseteq A''$ et $A' = A'''$, $B \subseteq B''$ et $B' = B'''$ ($'''$ est la composition de $'$ et $''$)
- $A \subseteq B' \Leftrightarrow B \subseteq A'$

L'opérateur composé $''$ définit une fermeture sur l'ensemble des parties de G, noté $\wp(G)$, ou sur l'ensemble des parties de M, noté $\wp(M)$. Rappelons qu'une fermeture h sur un ensemble partiellement ordonnée $(E, \leq)$ est extensive ($\forall x \in E, h(x) \geq x$), monotone croissante ($\forall x, y \in E, x \geq y$ alors $h(x) \geq h(y)$) et idempotente ($\forall x \in E, h(h(x)) \geq h(x)$). Un élément $x \in E$ est dit fermé pour h si et seulement si $x = h(x)$.

DEFINITION.3. — (Concept formel) Soit $\mathcal{K} = (G, M, I)$ un contexte formel. Un concept formel est un couple (A, B) tel que $A \subseteq G$, $B \subseteq M$, $A' = B$ et $B' = A$. A et B sont respectivement appelées extension (*extent*) et intension (*intent*) du concept

formel *(A,B)*. On note $\mathcal{C}$ *(G,M,I)* l'ensemble des concepts formels associés au contexte formel $\mathcal{K}$= (*G, M, I*).

Un sous ensemble *B* de *M* est l'intension d'un concept formel dans $\mathcal{C}$ *(G,M,I)* si et seulement si *B'' = B* (*B* est fermé pour *''*) et, de façon duale, un sous ensemble *A* de *G* est l'extension d'un concept formel dans $\mathcal{C}$ *(G,M,I)* si et seulement si *A'' = A* (*A* est fermé pour *''*).

Les concepts de $\mathcal{C}$ *(G,M,I)* sont ordonnés par une relation de subsomption entre concepts (notée $\sqsubseteq$) qui se définit par :

(*A1*, *B1*) $\sqsubseteq$ (*A2*, *B2*) si et seulement si *A1* $\subseteq$ *A2* (ou de façon duale *B2* $\subseteq$ *B1*), (*A1*, *B1*) et (*A2*, *B2*) étant deux concepts formels de $\mathcal{C}$ *(G,M,I)*. (*A2*, *B2*) est dit *subsumant* de (*A1*, *B1*).

Cette relation de subsomption permet d'organiser les concepts formels en un treillis complet, *(*$\mathcal{C}$ *(G,M,I),* $\sqsubseteq$*)*, appelé treillis de concepts ou encore treillis de Galois (Barbut et al. 1970), qui est noté $\underline{\mathcal{B}}$(G,M,I).

Rappelons pour finir qu'un treillis est un ensemble partiellement ordonné *(E,*$\leq$*)* tel que tout couple d'éléments *(x,y)* dans *E* × *E* admet une borne inférieure (ou *infimum*) notée $x \wedge y$ et une borne supérieure (ou *supremum*) notée $x \vee y$. Le treillis est complet si toute partie *S* de *E* admet une borne inférieure notée ΛS et une borne supérieure notée V*S*. En particulier, un treillis complet admet un élément minimal (*bottom*) noté $\perp$ et un élément maximal (*top*) noté $\top$.

4.2. *Construction du treillis de concepts associé à BioRegistry*

Dans cette section nous appliquons les définitions de l'analyse de concepts formels pour formaliser le contenu de BioRegistry.

Dans la suite, la formalisation de BioRegistry s'appuie sur un contexte formel $\mathcal{K}_{bio}$ *= (G, M, I)* où *G* est un ensemble de sources de données biologiques (par exemple *Swissprot*, *RefSeq*,...), *M* est un ensemble de méta-données (par exemple *manual revision*, *human organism*,...) et *I* est une relation binaire entre *G* et *M* (relation d'incidence de $\mathcal{K}_{bio}$). La relation *(g, m)* $\in$ *I* (ou *gIm*) signifie que la méta-donnée *m* est associée à la source de données *g*. Un exemple de contexte formel $\mathcal{K}_{bio}$ est donné dans le tableau 1. Les noms complets des sources de données biologiques et ceux des méta-données sont donnés dans le tableau 2.

Considérons dans le contexte $\mathcal{K}_{bio}$ = (*G, M, I*) donné par le tableau 1 l'ensemble de sources de données *A = {S1,S2,S4}* $\subseteq$ *G*. L'ensemble de méta-données communes à toutes les sources dans *A* est *A' = {AO,MR,PS}* $\subseteq$ *M*. De façon duale, pour l'ensemble de méta-données *B = {MR,NS,PS}* $\subseteq$ *M*, l'ensemble des sources possédant toutes les méta-données dans *B* est *B' = {S2}*.

Le couple (*A, A'*) est un concept formel de $\mathcal{K}_{bio}$ alors que (*B', B*) ne l'est pas. En effet *A'' = {S1,S2,S4} = A* tandis que *B'' = {AO,MR,NS,PS}* $\neq$ *B = {MR,NS,PS}*.

Dans ce dernier cas, le concept formel associé à B = *{MR,NS,PS}* est le couple *({AO,MR,NS,PS},{S2})*.

Sources\Méta-données	*NS*	*PS*	*AO*	*An*	*Ve*	*Hu*	*Mo*	*MR*
S1		×	×					×
S2	×	×	×					×
S3	×					×		
S4		×	×					×
S5	×	×				×		
S6	×			×				
S7		×					×	
S8		×			×			

Tableau 1. *Un exemple de contexte formel BioRegistry ($\mathcal{K}_{bio}$)*

Nom de la source	Symbole
Swissprot	*S1*
RefSeq	*S2*
TIGR-HGI	*S3*
GPCRDB	*S4*
HUGE	*S5*
ENSEMBL	*S6*
Mouse Genome DB	*S7*
Vega Genome Browser	*S8*

Méta-données	Abréviation	Catégorie
Nucleic sequence	*NS*	Subject
Proteic sequence	*PS*	Subject
Any organism	*AO*	Organism
Animals	*An*	Organism
Vertebrate	*Ve*	Organism
Human	*Hu*	Organism
Mouse	*Mo*	Organism
Manual revision	*MR*	Quality

Tableau 2. *Noms complets des sources de données biologiques et de leurs méta-données*

On note par $\mathcal{C}_{bio}$ l'ensemble des concepts de $\mathcal{K}_{bio}$. La figure 3 (faite par le logiciel *ToscanaJ* (Becker et al., 2004)) montre le treillis de concepts $(\mathcal{C}_{bio}, \sqsubseteq)$ (noté dans la suite par $\mathcal{L}(\mathcal{C}_{bio})$) correspondant au contexte formel de BioRegistry $\mathcal{K}_{bio}$ donné par le tableau 1. Cette représentation des treillis de concepts est la plus utilisée dans la communauté de l'analyse de concepts formels. Elle s'appuie sur l'héritage à la fois des attributs (dans notre cas les méta-données) et des objets (dans notre cas les sources de données) entre les nœuds représentant les concepts du treillis. Les attributs sont placés au plus haut dans le treillis : à chaque fois qu'un nœud n est étiqueté par un attribut m, tous les descendants de n dans le treillis héritent l'attribut m. De façon duale, les sources de données sont placées au plus bas dans le treillis : à

chaque fois qu'un nœud *n* est étiqueté par un objet *g* tous ses ancêtres héritent à l'envers l'objet *g*. Ainsi l'extension d'un concept *C* est obtenue en considérant tous les objets qui apparaissent sur les descendants du nœud *n* dans le treillis et l'intension de *C* est obtenue en considérant tous les attributs qui apparaissent sur les ancêtres du nœud *n* dans le treillis (Vogt et al., 1995 et Becker et al., 2004). Considérons par exemple le nœud non étiqueté au milieu de la figure 3. Le concept correspondant à ce nœud est *({S2,S5},{NS,PS})*.

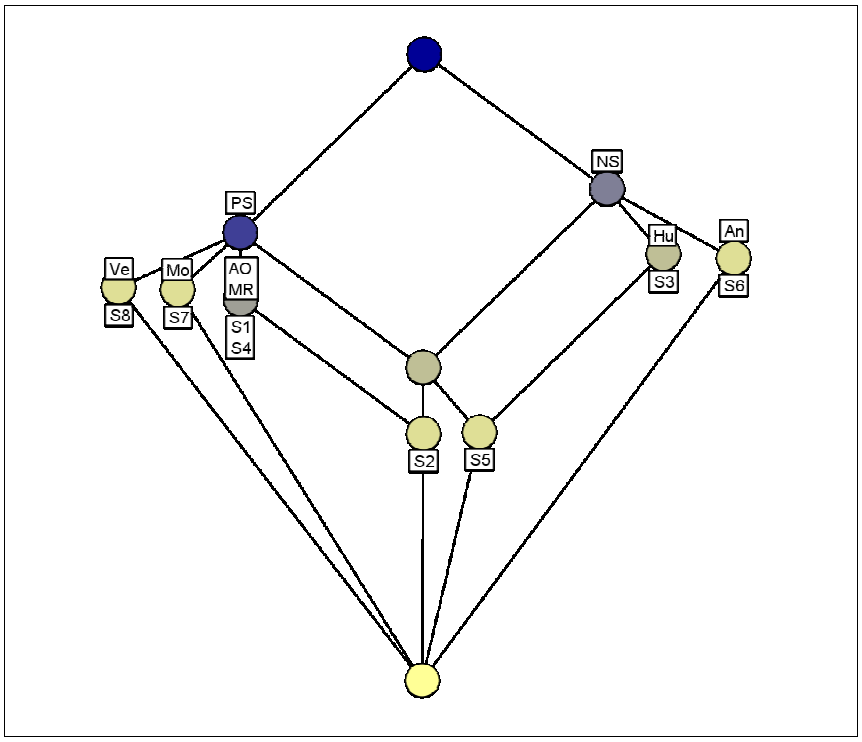

Figure 3. *Le treillis de concepts* $\mathcal{L}(C_{bio})$ *correspondant au contexte formel* $\mathcal{K}_{bio}$

Dans toute la suite, chaque nœud est désigné par le concept formel qui lui correspond dans le treillis $\mathcal{L}(C_{bio})$.

Une caractéristique importante du contexte formel $\mathcal{K}_{bio}$ est que l'ensemble *M* de méta-données est soigneusement choisi durant la construction de BioRegistry. Cette particularité nous motive dans le choix de l'algorithme de Godin (Godin et al., 1995b) pour construire le treillis de concepts correspondant puisque le contexte est

petit (en nombre de sources et de méta-données) et peu dense[6] (Kuznetsov et al., 2002). De plus, comme déjà mentionné dans la section 2.3, cet algorithme est incrémental et prend en compte l'ajout de nouveaux concepts dans un treillis existant. Cet aspect est essentiel pour la méthode d'interrogation détaillée dans la section 5.

4.3. *Classification flexible des sources de données de BioRegistry*

La formalisation du contenu de BioRegistry sous la forme d'un treillis de concepts permet soit de naviguer dans le treillis soit d'interroger les données du treillis et extraire des informations pertinentes pour un objectif donné. Par exemple un utilisateur intéressé par les méta-données de type *Subjects* (voir section 3.2) associées à l'ensemble des sources peut définir un contexte formel où les attributs ne sont constitués que des méta-données de la catégorie *Subjects*. L'ensemble d'objets correspondant comprend toutes les sources de données pour lesquelles au moins une méta-donnée de cette catégorie a été renseignée. Un treillis peut alors être construit, permettant de visualiser le partage des méta-données de la catégorie *Subjects* dans BioRegistry. Dans un autre cas, l'utilisateur peut être intéressé par la classification des sources de données relatives à l'organisme humain. Un nouveau contexte formel est construit automatiquement, dans lequel l'ensemble des objets comprend toutes les sources de données de BioRegistry dont la méta-donnée *Organism* a la valeur *Human*. L'ensemble des attributs est constitué de toutes les méta-données associées à cet ensemble de sources. Un nouveau treillis peut alors être construit pour répondre à ce nouveau besoin.

5. Interrogation du treillis de concepts de BioRegistry

5.1. *Recherche des sources de données biologiques pertinentes*

Une fois que le treillis de concepts $\mathcal{L}(\mathcal{C}_{bio})$ est construit, la recherche des sources de données pertinentes peut commencer. De la même façon que (Godin et al., 1995a) et (Carpineto et al., 2000), nous définissons un concept requête $Q = (Q_A, Q_B)$ où Q_A est un nom pour désigner une extension recherchée et Q_B est l'ensemble de méta-données décrivant les sources recherchée par la requête. Considérons par exemple la requête cherchant les sources ayant les méta-données Q_B = {*Nucleic Sequences*, *Human*, *Manual Revision*}. En utilisant les abréviations données dans le tableau 2, le concept réifiant la requête est $Q = (Q_A, \{NS,Hu,MR\})$.

---

[6] La densité d'un contexte formel, $\mathcal{K} = (G,M,I)$, est le rapport entre le cardinal de la relation I (le nombre de couples dans la relation) et le produit des cardinaux de G et de M, $|G| \times |M|$. Intuitivement c'est le rapport entre les cases occupées et les cases vides dans la table représentant le contexte formel.

Une fois défini, le concept Q est inséré dans le treillis $\mathcal{L}(C_{bio})$ en utilisant l'algorithme de construction incrémentale de Godin (Godin et al., 1995b). Le treillis obtenu est noté $\mathcal{L}(C_{bio} \oplus Q)$ où $C_{bio} \oplus Q$ désigne le nouvel ensemble de concepts résultant de l'insertion de la requête. Le treillis $\mathcal{L}(C_{bio} \oplus Q)$ correspondant à l'exemple mentionné précédemment est représenté à la figure 4. Les concepts entourés par des pointillés sont soit des nouveaux concepts soit des concepts modifiés suite à l'insertion de la requête Q dans le treillis. Ces concepts sont les seuls qui partagent des méta-données avec la requête et qui peuvent ainsi contenir des sources de données pertinentes.

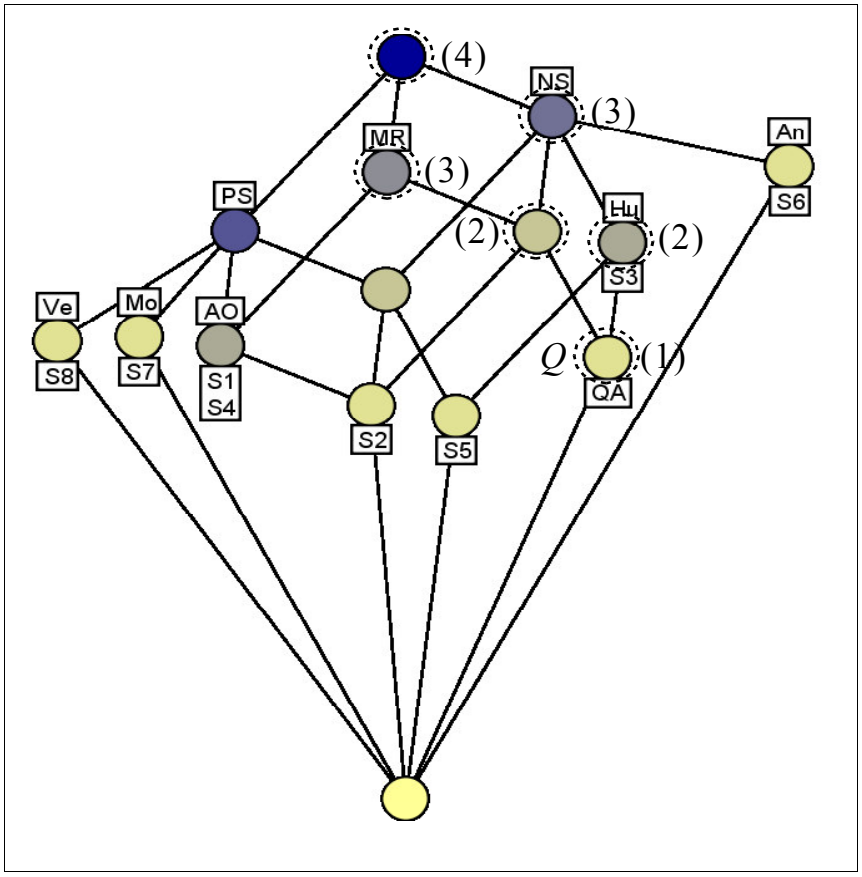

Figure 4. *Le treillis de concepts* $\mathcal{L}(C_{bio} \oplus Q)$

DEFINITION.4. — Une source de données S est *pertinente* pour une requête donnée $Q = (Q_A, Q_B)$ si et seulement si S est caractérisée par au moins une des méta-données de Q_B. Le *degré de pertinence* de S est donné par le nombre de méta-données que S partage avec Q_B.

Cette définition de la pertinence est à la base du processus de recherche détaillé dans la suite et illustré par un exemple. Elle est différente de la notion de *voisinage* utilisée dans (Carpineto et al., 2000), qui peut aboutir à l'obtention de documents ne partageant aucun terme avec la requête, ce qui ne correspond pas à nos besoins.

Etant donné une requête $Q = (Q_A, Q_B)$, toutes les sources de données pertinentes sont dans l'extension de Q et de ses subsumants dans le treillis de concepts (les concepts représentés en pointillé dans la figure 4) puisque l'intension de chacun de ces concepts est incluse dans Q_B (l'intension du concept requête). Dans ce qui suit nous notons par $\mathcal{P}_{\text{sources}}(Q, \mathcal{C}_{\text{bio}})$ l'ensemble des sources de données pertinentes pour la requête considérée dans l'ensemble de concepts formels $\mathcal{C}_{\text{bio}}$ du contexte formel $\mathcal{K}_{\text{bio}}$. Intuitivement, l'algorithme de recherche de sources de données pertinentes donné dans la figure 5, cherche d'abord à insérer le concept requête dans le treillis $\mathcal{L}(\mathcal{C}_{\text{bio}})$ ce qui produit le treillis $\mathcal{L}(\mathcal{C}_{\text{bio}} \oplus Q)$. Ensuite, l'ensemble des sources de données apparaissant dans l'extension de Q dans le treillis $\mathcal{L}(\mathcal{C}_{\text{bio}} \oplus Q)$ (si elles existent) est inséré dans la liste $\mathcal{P}_{\text{sources}}(Q, \mathcal{C}_{\text{bio}})$. Soit $SUBS(Q, \mathcal{C}_{\text{bio}}, 1)$ l'ensemble des subsumants directs (à distance 1) de Q dans $\mathcal{L}(\mathcal{C}_{\text{bio}} \oplus Q)$. L'ensemble des sources de données qui apparaissent dans les extensions des concepts de $SUBS(Q, \mathcal{C}_{\text{bio}}, 1)$ et qui ne sont pas déjà dans $\mathcal{P}_{\text{sources}}(Q, \mathcal{C}_{\text{bio}})$ sont ajoutées au résultat à cette étape. L'étape suivante consiste à déterminer $SUBS(Q, \mathcal{C}_{\text{bio}}, 2)$ l'ensemble des subsumants des concepts de $SUBS(Q, \mathcal{C}_{\text{bio}}, 1)$ (ou encore les subsumants de distance 2 de la requête). De la même façon que pour $SUBS(Q, \mathcal{C}_{\text{bio}}, 1)$, les nouvelles sources dans les extensions des concepts de $SUBS(Q, \mathcal{C}_{\text{bio}}, 1)$ sont ajoutées au résultat $\mathcal{P}_{\text{sources}}(Q, \mathcal{C}_{\text{bio}})$. La même opération est effectuée jusqu'à atteindre un ensemble $SUBS(Q, \mathcal{C}_{\text{bio}}, n)$ qui est vide (condition d'arrêt de l'algorithme). À chaque étape, les sources de données apparaissant dans un concept à intension vide sont ignorées. Le rang d'une source dans $\mathcal{P}_{\text{sources}}(Q, \mathcal{C}_{\text{bio}})$ peut être mémorisé selon la distance de la source (où du premier concept dans lequel apparaît la source) à la requête dans $\mathcal{L}(\mathcal{C}_{\text{bio}} \oplus Q)$.

Dans la figure 4, les numéros à côté des concepts représentent les itérations de l'algorithme expliqué précédemment. Dans la première itération le concept requête $Q = (Q_A, \{Hu, MR, NS\})$ est considéré. Dans cet exemple l'extension $Q_A = \varnothing$. Aucune source n'est ajoutée au résultat à cette étape. La deuxième itération permet d'ajouter au résultat $\mathcal{P}_{\text{sources}}(Q, \mathcal{C}_{\text{bio}})$ les sources de données *S3*, *S5* et *S2* car les concepts $(\{S3, S5\}, \{Hu, NS\})$ et $(\{S2\}, \{MR, NS\})$ (formant l'ensemble $SUBS(Q, \mathcal{C}_{\text{bio}}, 1)$) subsument le concept requête $Q = (Q_A, \{Hu, MR, NS\})$. À la troisième itération les sources *S1, S4* et *S6* sont ajoutées à $\mathcal{P}_{\text{sources}}(Q, \mathcal{C}_{\text{bio}})$ car $SUBS(Q, \mathcal{C}_{\text{bio}}, 2) = \{(\{S2, S3, S5, S6\}, \{NS\}), (\{S1, S2, S4\}, \{MR\})\}$. À la quatrième itération $SUBS(Q, \mathcal{C}_{\text{bio}}, 3) = \{(\{S1, S2, S3, S4, S5, S6, S7, S8\}, \{\})\}$. Le seul concept formant $SUBS(Q, \mathcal{C}_{\text{bio}}, 3)$ est un concept ayant une intension vide. Aucune source n'est donc ajoutée à

$\mathcal{P}_{sources}(Q, C_{bio})$ à cette étape et l'algorithme s'arrête puisque $SUBS(Q, C_{bio}, 4) = \varnothing$. Ainsi $\mathcal{P}_{sources}(Q, C_{bio})$ est constitué des sources ordonnées comme suit :

1) *({S3, S5}{Hu, NS})* et *({S2}{NS, MR})*

2) *({S6}{NS})*et *({S1, S4}{MR})*.

Des critères supplémentaires relatifs à des préférences données peuvent être considérés pour raffiner l'ordre des sources de données dans le résultat $\mathcal{P}_{sources}(Q, C_{bio})$. En effet dans le cas des sources au même rang dans $\mathcal{P}_{sources}(Q, C_{bio})$, il est possible de raisonner par rapport à la fraîcheur des données contenues. La fréquence de mise à jour des sources et/ou la date de la dernière mise à jour peuvent favoriser une source par rapport à une autre. La qualité des données peut aussi être un critère permettant de raffiner l'ordre. En effet on peut s'appuyer sur la conformité des données contenues dans une source à un standard pour favoriser cette source par rapport à celles qui ont le même rang dans le résultat $\mathcal{P}_{sources}(Q, C_{bio})$.

Entrée :

- une requête $Q = (Q_A, Q_B)$ où $Q_A = \varnothing$ et Q_B est un ensemble de méta-données
- le treillis de concepts $\mathcal{L}(C_{bio})$ correspondant au contexte formel $\mathcal{K}_{bio}$

Sortie :

- le treillis de concepts $\mathcal{L}(C_{bio} \oplus Q)$
- un ensemble de sources pertinentes pour la requête Q

Début :

1. Construire le concept $Q = (Q_A, Q_B)$
2. Insérer Q dans le treillis $\mathcal{L}(C_{bio})$ pour obtenir le treillis $\mathcal{L}(C_{bio} \oplus Q)$
3. Chercher dans $\mathcal{L}(C_{bio} \oplus Q)$ le nouveau concept $Q = (Q_A \cup Q_B', Q_B)$
4. *niveau* := 0
5. $SUBS(Q, C_{bio}, niveau) := \{Q\}$ (initialisation de la recherche des subsumants de la requête Q dans le treillis en fonction de leur niveau où distance à Q)
6. $\mathcal{P}_{sources}(Q, C_{bio}) := \varnothing$ (initialisation du résultat)
7. Tant que $SUBS(Q, C_{bio}, niveau) \neq \varnothing$ faire

   7.1. $\mathcal{P}_{sources}(Q, C_{bio}, niveau) := \varnothing$

   7.2. Pour tout concept C = (A, B) $\in SUBS(Q, C_{bio}, niveau)$ faire

      a. Si B $\neq \varnothing$ alors

a.1. $\mathcal{P}_{sources}(Q, C_{bio}, niveau) := \mathcal{P}_{sources}(Q, C_{bio}, niveau) \cup A$

b. Fin Si

7.3. Fin pour

7.4. $\mathcal{P}_{sources}(Q, C_{bio}) := \mathcal{P}_{sources}(Q, C_{bio}) \cup \mathcal{P}_{sources}(Q, C_{bio}, niveau)$

7.5. Construire $SUBS(Q, C_{bio}, niveau+1)$ l'ensemble des subsumants directs des concepts de $SUBS(Q, C_{bio}, niveau)$

7.6. $niveau := niveau + 1$

8. Fin tant que

9. Retourner $\mathcal{P}_{sources}(Q, C_{bio})$, l'ensemble des sources pertinentes pour Q

Fin

Figure 5. *Algorithme de recherche des sources de données pertinentes*

5.2. *Raffinement de requête à partir d'ontologies de domaines*

La réponse à une requête donnée peut être vide. Par exemple dans le contexte formel de BioRegistry donné dans le tableau 1, une requête cherchant une source contenant des données relatives à l'organisme *Chicken* aura un résultat vide. Cependant, des sources de données pertinentes relatives à cette requête peuvent exister mais les méta-données qui les décrivent ne correspondent pas à ce qui est exprimé dans la requête. Pour résoudre ce problème nous proposons un mécanisme de raffinement de requête à partir d'ontologies de domaines.

Pour ce faire, nous modifions la requête et non le treillis contrairement aux propositions (Carpineto et al., 2000, Priss, 2000, Safar et al., 2004). En effet, nous préservons la structure du treillis et nous modifions la requête en y ajoutant des méta-données qui sont en relation avec les méta-données de la requête initiale dans une ontologie donnée. Cette stratégie présente le double avantage d'être automatisable et d'éviter l'introduction de redondance dans le treillis.

Les méta-données ajoutées à la requête sont soit plus spécifiques soit plus générales que celles de la requête initiale. Ceci nous amène à définir deux types de raffinement : le raffinement par généralisation et le raffinement par spécialisation. Il est important de rappeler ici que nous ne sommes pas confrontés à des problèmes de synonymie entre les méta-données de la requête et les termes de l'ontologie puisque les méta-données utilisées lors de la construction de BioRegistry sont des termes extraits d'ontologies de domaines.

Le raffinement par généralisation pour une méta-donnée *m* de la requête considère des méta-données plus générales que *m* dans l'ontologie. Dans l'exemple cité précédemment (méta-donnée *Chicken : Q = (Q_A, {Ch})*) et en considérant l'ontologie *OntoBR* de la figure 2, les méta-données qui peuvent être considérées et ajoutées dans la requête lors d'un raffinement par généralisation sont *Vertebrates, Animals, Eucaryotes, Cellular Organisms*, et *Any Organism*, c'est à dire tous les ascendants de *Chicken* dans cette ontologie. Cependant, certaines de ces méta-données ne figurent pas dans le contexte formel $\mathcal{K}_{bio}$ donné par le tableau 1 ce qui est le cas de *Cellular Organisms* et *Eucaryotes*. Cela signifie que ces méta-données ne caractérisent aucune source de données du contexte $\mathcal{K}_{bio}$, ce qui rend inutile leur ajout à la requête. Seules les méta-données présentes dans $\mathcal{K}_{bio}$ sont considérées lors du processus de raffinement par généralisation. Ainsi, pour l'exemple de la requête *Q = (Q_A, {Ch})*, le raffinement par généralisation produit la nouvelle requête *Q = (Q_A, {Ve, An, AO})* où *Ve = Vertebrate, An = Animals,* et *AO = Any Organism*. L'application de l'algorithme de recherche en considérant cette nouvelle requête donne le résultat suivant :

1- *({S6 }{An}), ({S8}{Ve}) et ({S1, S2, S4 }{AO}).*

De façon duale, le raffinement par spécialisation pour une méta-donnée m de la requête considère les méta-données plus spécifiques que m dans l'ontologie. Dans l'exemple précédant, la méta-donnée *Chicken* n'a pas de descendant dans l'ontologie *OntoBR,* ce qui rend impossible le raffinement de la requête par spécialisation pour cette méta-donnée.

Un autre exemple plus parlant, est de considérer la requête composée de la méta-donnée *Eucaryotes* (la requête *Q = (Q_A, {Eu})*, qui ne retourne aucune réponse puisque *Eucaryotes* n'existe pas dans le contexte formel $\mathcal{K}_{bio}$. Le raffinement par spécialisation permet d'ajouter à la requête les descendants de la méta-donnée *Eucaryotes* dans *OntoBR*, qui apparaissent dans le contexte formel $\mathcal{K}_{bio}$, c'est à dire *Animals*, *Vertebrate*, *Human* et *Mouse*. Ainsi le raffinement par spécialisation de la requête Q = (Q_A, *{Eu}*) produit la nouvelle requête *Q = (Q_A, {An, Ve, Hu, Mo})* où *An = Animals, Ve = Vertebrate, Hu = Human* et *Mo = Mouse*. L'application de l'algorithme de recherche en considérant cette nouvelle requête donne le résultat suivant :

1- *({S6 }{An}), ({S8}{Ve}),({S3, S5}{Hu}) et ({S7}{Mo}).*

Il est possible de combiner les deux types de raffinement de requête en ajoutant à la requête à la fois les descendants et les ancêtres de la méta-donnée considérée dans l'ontologie *OntoBR*. Dans les deux types de raffinement, le nombre de méta-données ajoutées à la requête peut être contrôlé en considérant par exemple uniquement les ancêtres les plus proches de la méta-donnée considérée dans l'ontologie (dans le cas du raffinement par généralisation) et/ou ses descendants les plus proches (dans le cas du raffinement par spécialisation).

5.3. *Choix entre raffinement par généralisation et par spécialisation*

Dans le cas où la méta-donnée considérée est une feuille ou la racine de l'ontologie, le problème du choix ne se pose pas puisque dans les deux cas un seul type de raffinement est possible (raffinement par généralisation pour une feuille et par spécialisation pour la racine). Cependant, lorsque la méta-donnée considérée n'est ni une feuille ni la racine de l'ontologie, les deux types de raffinement sont possibles et le choix peut être fait selon les préférences de l'utilisateur. En effet, si l'utilisateur accepte des sources de données dont une partie seulement correspond à sa requête, alors le raffinement par généralisation peut être choisi. Si par contre il accepte des sources de données correspondant à une partie seulement de sa requête, alors le raffinement par spécialisation peut être choisi. La composition de sources de données correspondant partiellement à une requête peut alors permettre d'obtenir une réponse complète. Dans les deux cas, il est utile de procéder à un ordonnancement à posteriori des sources de données ajoutées au résultat suite au raffinement de la requête. Cet ordonnancement doit être basé sur la similarité entre les méta-données indexant ces sources, d'une part et les méta-données constituant la requête, d'autre part (Ganesan et al., 2003).

6. Conclusion et perspectives

Dans cet article, nous avons présenté une approche combinant l'analyse de concepts formels et les ontologies de domaines pour résoudre un problème de recherche d'information en bioinformatique. BioRegistry, en tant qu'annuaire structuré de méta-données relatives aux sources de données biologiques, constitue un support d'application approprié pour la théorie de l'analyse de concepts formels et ses potentialités en termes de passage à l'échelle et de flexibilité. L'approche décrite vise à organiser pour les interroger des sources de données biologiques. La construction de treillis de concepts apparaît comme un moyen privilégié de fournir des vues personnalisées sur les sources de données biologiques et d'organiser les connaissances sur ces sources. De plus un mécanisme de raffinement de requête dépendant d'ontologies de domaines est proposé pour améliorer le processus de recherche d'information.

La mise à disposition prochaine d'une version étendue de l'annuaire BioRegistry (plusieurs centaines de sources de données) permettra de mettre en œuvre une évaluation de cette approche en grandeur réelle. Cette expérimentation pourra à son tour conduire à de nouvelles perspectives telles que la définition de critères de pertinence adaptés aux préférences de l'utilisateur pour classer les résultats et la prise en compte des relations entre les méta-données au cours du processus de classification par analyse de concepts formels ainsi qu'à l'extension de l'analyse de concepts formels pour le traitement de tableaux multivalués.

Remerciements

Nous voulons remercier Shazia Osman pour sa contribution à la conception et à la construction de BioRegistry. Ce travail est soutenu par la région Lorraine dans le cadre du PRST « Intelligence Logicielle ».

7. Bibliographie

Barbut M., Monjardet B., « *Ordre et Classification : Algèbre et Combinatoire* », Tome II, Hachette, 1970.

Becker P., Correia J. H., « The ToscanaJ suite for implementing Conceptual Information Systems », Formal Concept Analysis: Foundations and Applications *Formal Concept Analysis: Foundations and Applications*, LNAI 3626, p. 324-348, B. Ganter, G. Stumme, R. Wille, editor, Springer-Verlag, 2005.

Carpineto C., Romano G., « A Lattice Conceptual Clustering System and Its Application to Browsing Retrieval. », *Machine Learning*, vol. 24, n° 2, 1996, p. 95-122.

Carpineto C., Romano G. « *Concept Data Analysis: Theory and Applications* », John Wiley & Sons, 2004.

Carpineto C., Romano G., « Order-theoretical ranking », *Journal of the American Society for Information Science*, vol. 51, n° 7, 2000, p. 587-601.

Carmel D., Farchi E., Petruschka Y., Soffer A., « Automatic query refinement using lexical affinities with maximal information gain », *SIGIR'02 : Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, Tampere, Finland, 2002, ACM Press, p. 283-290.

Davidson S., Crabtree J., Brunk B., Brian P., Schug J., Tannen V., Overton G. C., Stoeckert C. J., « K2/Kleisli and GUS : experiments in integrated access to genomic data sources », *IBM systems journal*, vol. 40, 2000, p. 512-531.

Dekkers M., Weibel S., « State of the Dublin Core Metadata Initiative », *D-Lib Magazine*, vol. 9, n° 4, 2003.

Discala C., Benigni X., Barillot E., Vaysseix G., « DBCAT : a catalog of 500 biological databases », *Nucleic Acids Research*, vol. 28, n° 1, 2000, p. 8-9.

Galperin M. Y., « The Molecular Biology Database Collection : 2005 update », *Nucleic Acids Research*, vol. 33, 2005, National Center for Biotechnology Information and National Library of Medicine and National Institutes of Health.

Ganter B., Wille R., « *Formal Concept Analysis* », Springer, mathematical foundations édition, 1999.

Ganesan P., Garcia-Malia H., Widom J., « Exploiting hierarchical domain structure to compute similarity », *ACM Transactions on Information Systems*, vol. 21, n° 1, 2003, p. 64-93.

Goble C. A., Stevens R., NG G., Bechhofer S., Paton N. W., Baker P. G., Peim M., Brass A., « Transparent Access to Multiple Bioinformatics Information Sources », *IBM Systems Journal Special issue on deep computing for the life sciences*, vol. 40, n° 2, 2001, p. 532-552.

Godin R., Mineau G., Missaoui R., « Méthodes de classification conceptuelle basées sur les treillis de Galois et applications », *Revue d'intelligence artificielle*, vol. 9, n° 2, 1995, p. 105-137.

Godin R., Missaoui R., Alaoui H., « Incremental Concept Formation Algorithms Based on Galois (Concept) Lattices. », *Computational Intelligence*, vol. 11, 1995, p. 246-267.

Köhler J., Philippi S., Lange M., « SEMEDA : ontology based semantic integration of biological databases », *Bioinformatics*, vol. 19, 2003, p. 2420-2427.

Kuznetsov S., Obiedkov S., « Comparing Performance of Algorithms for Generating Concept Lattices », *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 14, 2002, p. 189-216, Taylor & Francis.

Lord P., Bechhofer S., Wilkinson M. D., Schiltz G., Gessler D., Hull D., Goble C., Stein L., « Applying semantic Web services to Bioinformatics : Experiences gained, lessons lernt », *ISWC*, 2004, LNCS 3298, p. 350-364, S.A. McIlraith et al., editor, Springer-Verlag, 2004.

Oinn T., Addis M., Ferris J., Marvin D., Senger M., Greenwood M., Carver T., Glover K., Pocock M. R., Wipat A., Li P., « Taverna : a tool for the composition and enactment of bioinformatics workflows », *Bioinformatics*, vol. 20, 2004, p. 3045-3054.

Pernelle N., Rousset M.-C., Soldano H., Ventos V., « ZooM : a nested Galois lattices-based system for conceptual clustering », *Journal of Experimental and Theoritical Artifial Intelligence (JETAI)*, vol. 14, n° 2, 2002, p. 157-187.

Priss U., « Lattice-based Information Retrieval », *Knowledge Organization*, vol. 27, n° 3, 2000, p. 132-142.

Safar B., Kefi H., Reynaud C., « OntoRefiner : a user query refinement interface usable for Semantic Web Portals », *Proceedings of the ECAI 2004, Workshop on Application of Semantic Web Technologies to Web Communities*, August 2004.

Stumme G., Taouil R., Bastide Y., Pasquier N., Lakhal L., « Computing Iceberg Concept Lattices with Titanic », *Data Knowledge Engineering,* vol. 42, n° 2, 2002, p. 189-222.

Van der Merwe D., Obiedkov S. A., Kourie D. G., « AddIntent : A New Incremental Algorithm for Constructing Concept Lattices », *Proceedings of ICFCA Concept Lattices, Second International Conference on Formal Concept Analysis, ICFCA 2004*, Sydney, Australia, February 23-26, 2004, LNAI, vol. 2961, p. 372–385, P. W. Eklund, editor, Springer, 2004.

Vogt F., Wille R., « TOSCANA : A graphical tool for analyzing and exploring data », *Graph Drawing*, LNCS, vol. 894, p. 226-233, R. Tamassia and I.G. Tollis, editor, Springer-Verlag, Berlin-Heidelberg, 1995.

Wille R., « Restructuring lattice theory : an approach based on hierarchies of concepts », *Orderd sets*, vol. 23, 1982, p. 445-470, Rival editor.

Wroe C., Stevens R., Goble C., Roberts A., Greenwood M., « A suite of DAML+OIL Ontologies to Describe Bioinformatics Web Services and Data », *International Journal of Cooperative Information Systems special issue on Bioinformatics*, vol. 12, n° 2, 2003, p. 197-224.