Ontology-Assisted Database Integration to Support Natural Language Processing and Biomedical Data-mining

Jean-Luc Verschelde[1], Mariana Casella Dos Santos[1], Tom Deray[1], Barry Smith[2,3] and Werner Ceusters[1]

[1]Language and Computing, Maaltecenter Blok A 3, Derbystraat 79, B-9051 Sint-Denijs-Westrem, Belgium, {Jean-Luc, Mariana, Tom, Werner}@landc.be, http://www.landcglobal.com/

[2]Institute for Formal Ontology and Medical Information Science, University of Leipzig

[3]Department of Philosophy, University at Buffalo, NY, phismith@buffalo.edu

Summary

Successful biomedical data mining and information extraction require a complete picture of biological phenomena such as genes, biological processes, and diseases; as these exist on different levels of granularity. To realize this goal, several freely available heterogeneous databases as well as proprietary structured datasets have to be integrated into a single global customizable scheme. We will present a tool to integrate different biological data sources by mapping them to a proprietary biomedical ontology that has been developed for the purposes of making computers understand medical natural language.

1 Introduction

Information systems available to pharmaceutical or biotech companies must employ a range of different data sources, each with their own data structure and mode of presentation. This diversity among the data sources hinders their integration, and thus hampers the complete apprehension of the information they contain [1,2].

Several academic and industrial research groups active in biology-related fields are growing ever more convinced that their research would advance at a greater pace and that they would gain better insight into the subjects of their research, if only different information sources could be integrated. Above all, research carried out during the target phase of the drug discovery process would profit from higher levels of integration. The latter would allow research scientists to expand out of the pure genome-driven target discovery to a situation where other types of relevant information are also made available. In this way, a wide-ranging query-strategy can be adopted, covering for instance, protein sequence information as well as clinical data.

In this paper, we present a solution to the problem of integrating different biological data sources, which involves mapping them semantically to a proprietary medical ontology, called LinKBase®, that has been developed to make computers understand medical natural language. We applied our integration strategy in two stages: First, we virtually expanded the LinKBase® medical ontology with domain knowledge in the field of molecular biology by taking over the concepts of the Gene Ontology™ [4]. Secondly, we mapped information from the protein database Swiss-Prot to the biomedical ontology. Until now, only those types of

protein information that are relevant for document ranking and information extraction purposes have been introduced into our knowledge system.

2 Materials and Methods

2.1 LinKBase®

LinKBase® comprehends various aspects of medicine, including anatomy, diseases, pharmaceuticals, and so on. These are represented via concepts that are interrelated using a rich set of possible relation types. By "concepts" we mean entries in the ontology that stand for real-world entities - not concepts in the minds of conscious beings that are abstractions of what these beings think the real-world entities are. Each concept is related to certain other concepts which provide the criteria which constitute its formal definition. A criterion is thus a reference from a source concept to a target concept using one relation from a range of different relation (or 'link') types. Our ontology contains 543 different link types, reflecting sometimes subtle semantic differences. They are divided into different groups, including spatial, temporal, and process-related link types. LinKBase® currently contains over 2,000,000 medical concepts with over 5,300,000 link type instantiations. Both concepts and links are language independent, but they are cross-referenced to about 3,000,000 terms in various[1] languages. Terms can be stored in different languages and can be linked to concepts, criteria, and link types via an intersection table which allows us to define both homonyms (single terms that have several different meanings or criteria/linked concepts/link types) and synonyms (multiple terms associated with one single criterion/concept/link type).

2.2 LinkFactory®, L&C's Ontology Management System (OMS)

LinkFactory® is an ontology management tool designed to build, manage, and maintain large, complex, language-independent ontologies such as LinKBase®. It is a bean-based multiple windows environment comprising more than 20 beans[2] that provide a wide range of functions. A selection of beans can be assembled in one or more frames and linked to one another. Different scenarios of bean configurations can be composed, depending on the user's focus. For example, if a ConceptTree bean (which provides the user with a hierarchical view of relations in the ontology) is linked to a FullDef bean, then selection of a source concept in the ConceptTree bean tells the FullDef bean to show the relations the selected concept has to other concepts (Figure 1). The FullDef bean shows also the full definitions, i.e. the necessary and sufficient conditions which must be satisfied if a real world entity, whose characteristics are (wholly or partially) known, is to be identified as an instance of the concept. A ReverseConcept bean linked to the ConceptTree bean shows the relations other concepts have to the selected concept.

---

[1] At the moment of writing, 16 languages are supported, with primary focus on the following 7 languages: English (2,000,000 terms), Dutch (330,000 terms), Italian (141,000 terms), French (112,000 terms), Spanish (83,000 terms), Turkish (76,000 terms) and German (62,000 terms).

[2] A bean offers the user a GUI that exposes a small selection from the functionalities available on the LinKFactory(r) ontology server. Different beans serve different purposes - purposes that may or may not apply to all users. 16 different beans provide the user the necessary tools to browse and edit the entirety of the LinKBase(r) ontology. More advanced reasoning and search tools are available in 6 other beans. Last but not least, 6 additional beans provide flexible administration capabilities.

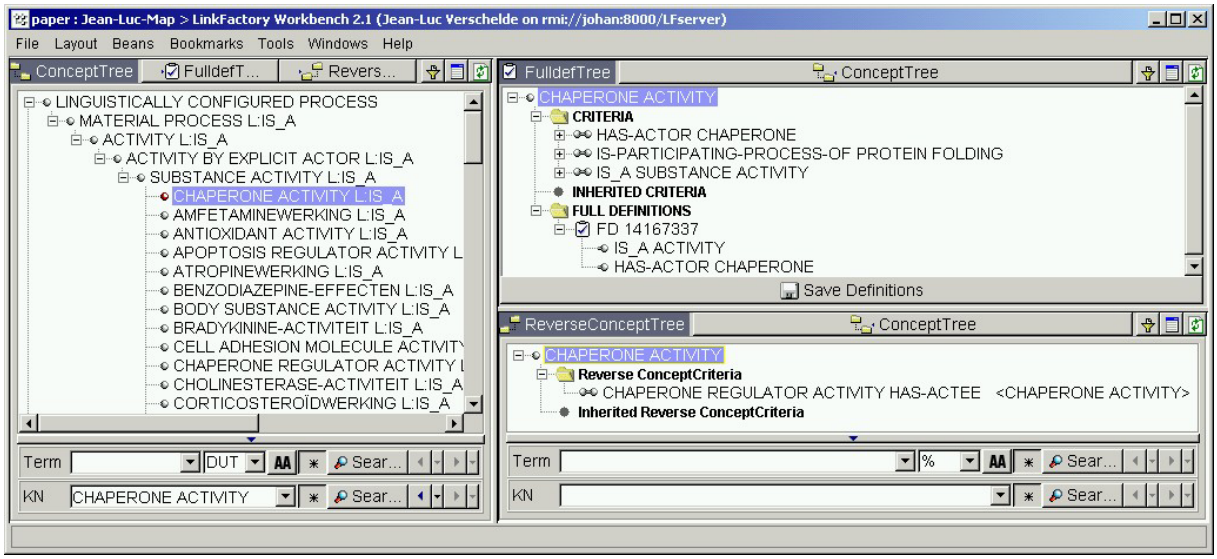

Figure 1: This frame shows a view of the semantic representation of <chaperone activity>[3], using the ConceptTree bean linked to a FullDef and ReverseConceptTreebean.

2.3 The Gene Ontology™ and Swiss-Prot

Integrating content is the holy grail of biomedical research and it is precisely here, we believe, that LinKBase® can bring interesting results. The applications of interest in this paper are situated in the field of molecular biology and bioinformatics. If we want to apply our NLP technology in these fields, we have to broaden our medical domain ontology with terminologies or ontologies that are drawn from this new domain, and it is for these purposes that we have incorporated the content of the Gene Ontology (GO).

The goal of the Gene Ontology Consortium is to create a controlled vocabulary that can be applied to all organisms for the description of the corresponding cellular components, molecular functions, and biological processes. The main purpose of GO is to provide conventions and a commonly accepted structured set of terms for annotating genes and gene products in a consistent way. In the GOA project [5], the GO vocabulary has been applied to a non-redundant set of proteins described in three databases - Swiss-Prot, TrEMBL and Ensembl - which when taken together provide complete proteome sets for Homo sapiens and other organisms. The Swiss-Prot database is composed of sequence entries that contain records comprising various types of data. Since our primary focus is natural language understanding applications (such as document ranking and information extraction), we confine ourselves here to the lexical information contained in the Swiss-Prot database tables providing protein names, gene names, and their synonyms.

2.4 Mapping of databases with MaDBoKS

The MaDBoKS (Mapping Databases onto Knowledge Systems) tool is an extension of the Linkfactory® ontology management system (OMS) that administers and generates mappings from external relational databases onto LinKBase®. Such a mapping defines the associations between the relational schemata (and population) of the database and LinKBase®. Through MaDBoKS, content from one or several databases can be retrieved and mapped onto LinKBase®. The MaDBoKS system is designed in such a way that all implicit and explicit

[3] Notation: 'GO concepts' , <LinKBase(r) concepts>, *GO relations* and LinKBase relations.

relationships between data from the different databases are mapped to the ontology[4]. Administration of the mapping mediates[5] the data contained in the different databases in such a way that it is associated with ontological information and the ontology is thereby virtually expanded with data and relations. The mapping tool can map column data as well as cell record data in such a way as to carry relationships over into the ontology. The MaDBoKS system meets the requirement that the ontology does not change upon coupling or decoupling of the databases. Also, the data in the databases is always used in its current state. This is because the flow of data to the ontology occurs in real time, so that every change in the database is automatically accessible to the ontology also. Applications that operate on the databases, for example gene annotation software, where the ontology concepts are used as references, can still be applied to the data as mapped.

This mapping of databases can be split up into two phases: an analysis phase and a physical mapping phase. In this first phase, the structure of the database (effectively: its model of reality) is analysed and mirrored within the ontology using the existing concepts and relations in LinKBase®. This structure is thereby mapped to the ontology and the resulting mapping information is stored in an XML-formatted mapping file. In the second phase, the database itself is mapped to the ontology. A generic mechanism then translates database data and relations to concepts and relations between the concepts of the ontology. As a matter of fact, high-level queries to the OMS are translated into database queries and queries to the LinKBase®. The results of these queries are processed and assembled in such a way that all results are presented to the user in a transparent manner, without his being aware that several sources are being questioned at once.

3 Discussion

3.1 Semantic heterogeneity

In the domain of molecular biology and related fields there exists not only a large number of databases containing specialized information, but also a variety of information systems designed to cope with data derived from different disciplines via processes of data integration. The demand for the latter from research groups working on multidisciplinary bioinformatics projects is becoming ever more intense, but such systems are difficult to build because of the heterogeneity of the databases involved. This heterogeneity is of two sorts: 1. Syntactic heterogeneity refers to differences in data models (different representation for the same semantic information) and data languages (different datatypes) and can be easily resolved. 2. Semantic heterogeneity refers to differences in the underlying meanings of the data represented. The aim of the integration process described here is that of developing a global scheme that is designed to integrate the local schemes of databases in a semantically correct way. In general, most of the existing global schemes are constructed on the basis of concepts and relations already present in the different databases. On this approach, each time additional databases have to be integrated, the global scheme has to be adapted or even reconceived from scratch. On the other hand, an ontology of the sort described here has from the start a much broader base for incorporating the content of heterogeneous databases and is much less dependent on the databases to be integrated.

---

[4] It is important to notice that the medical ontology LinKBase® is assumed as the ontology in use throughout this paper. However, LinKFactory® and MaDBoKS support any ontological content and remain fully functional in different, for example non-medical, contexts as well.

[5] Compare the description of mediators given by Wiederhold [3].

3.2 Formal ontologies and ontological clarity

Integrating data from different data sources is a delicate subject within the medical and biomedical domain. The ever-present ambiguity of terms, both within and between different databases, terminologies and ontologies, as well as the frequent misapplication of synonymy, makes this task highly error prone. LinKBase® is an ontology supporting the integration of data from different external data sources in a transparent way, capturing the exact intended semantics of the database terms, and filtering out erroneous synonyms. To achieve this goal, LinKBase® makes use of the rigor and formalism of philosophical formal ontology. LinKBase® is structured according to the Basic Formal Ontology (BFO), a philosophically inspired top-level ontology [6-11], which focuses on the entities in reality at different levels of granularity rather then on human conceptualizations thereof. BFO provides LinKBase® with a rigorous classification of all the entities in LinKBase®, a formal description of these entities, as well as a first-order logical description of all possible formal relations existing between these entities in reality. The theory is then used to constrain our modelling space in order to reduce the likelihood of error, to detect and correct previously introduced errors, and also to support the automatic generation of new relations between entities. A formal ontology based on rigorous philosophical ontological theory is a crucial tool to avoid mistranslation between the different sources when integrating databases. It can also be employed for advance detection of possible areas of ambiguity and problems arising in the modelling and mapping of external databases, as well as for semi-automatic detection of errors in the post-mapping phase.

3.3 Mapping of GO and Swiss-Prot onto LinKBase®

LinKBase® has served as a vital component of our NLP-based tools for information retrieval and extraction in the field of clinical medicine. However, our ontology was less complete with regard to biomolecular information. Since this information is crucial for proper indexing of the biomedical literature in which biological processes are described at the molecular level of granularity, the medical ontology has now been supplemented with terminology from GO and Swiss-Prot.

GO has a number of problems from the formal-ontological point of view, pertaining above all to its treatment of the notions of function and activity [9]. Yet, it is nonetheless an important source of terms for describing biological processes, molecular functions/activities and cellular components. Swiss-Prot, for its part, provides us with a corresponding facility regarding the names of genes and gene products acting in different biological processes.

The design of MaDBoKS allows mapping of databases on both the column and row level. The former means that a full table is mapped onto a concept from the LinKBase® ontology in such a way that all the data in that table then represents entities standing in a child-parent relation to the mapped LinKBase® concept. The latter provides mapping of those implicit relationships in a database (described below in our account of Swiss-Prot information mapping) in which parts of the population of a database are mapped onto ontology concepts. This flexibility in preparing mapping schemes allows us to fit parts of GO into the structure of LinKBase®.

As already mentioned, the mapping of the external databases starts with an analysis phase. During this phase, we carefully investigated the top-layer concepts of the three GO sub-domains that will act as the mapping layer between the LinKBase® concepts and GO terms. The concepts of this layer are compared to the existing concepts/terms in LinKBase® using a semi-automatic procedure. If no exact match is found, a new LinKBase® concept is created and necessary new criteria are associated therewith in order to capture in detail the semantics

of the new concept. Once the mapping layer is fully mirrored in LinKBase®, we map onto this layer the corresponding concepts from GO.

Swiss-Prot information is also mapped on the row level, but in addition, the relation between the two columns "Entry name" and "Protein name" is mapped as well. The column "Entry name" denotes all Swiss-Prot proteins containing a unique identifier. The column "Protein name" denotes a class of proteins. From an ontological point of view, there is a parent-child relationship between the data under "Protein name" and "Entry name". Part of the data in the "Protein Name" column was already represented in our LinKBase® as concepts under <protein> and are therefore mapped onto the children of <protein> with equivalent meanings. The result of this mapping is depicted in Figure 2, using the same frame as in Figure 1. Black concepts represent concepts originating from LinKBase®. The data taken from the database columns are shown in blue and relations between a LinKBase® concept and a database concept are in purple.

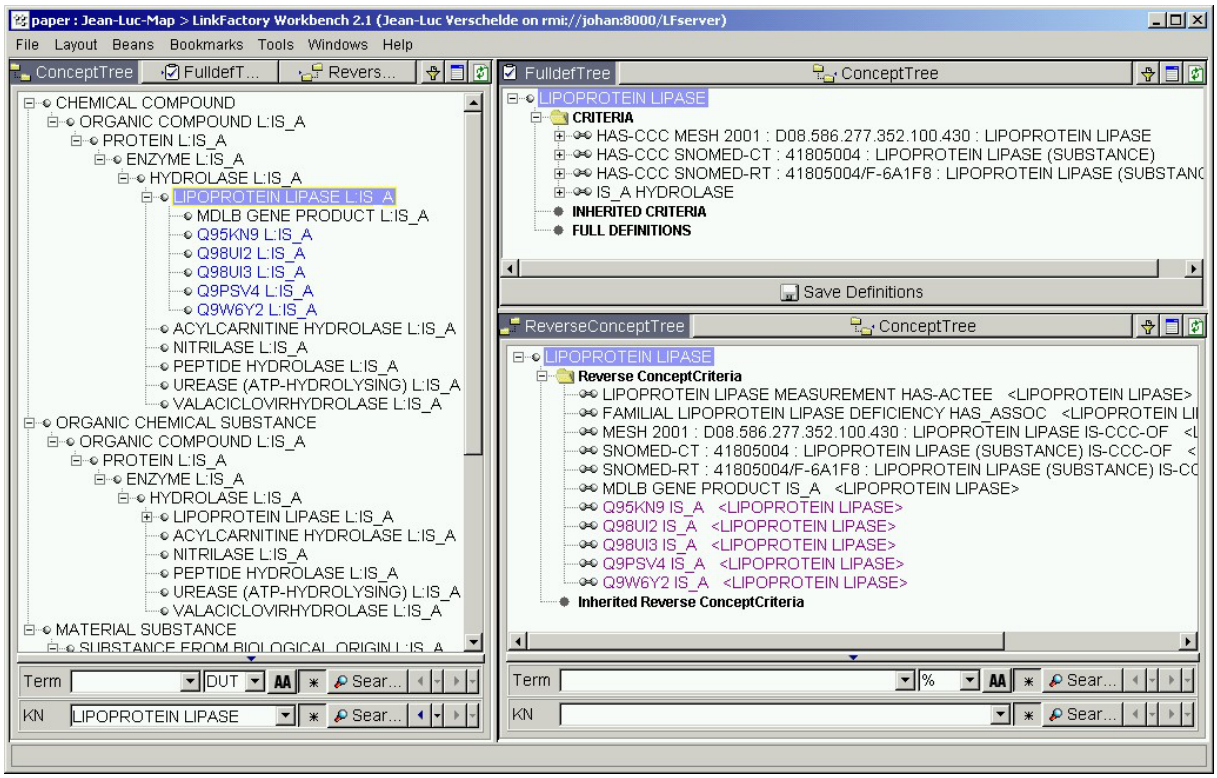

Figure 2 An example of mapping the database "Protein name" entry 'lipoprotein lipase' on its LinKBase® equivalent on row level.

The gene products or proteins of Swiss-Prot are annotated by associating them to different concepts in GO. Hence, for each protein one or more appropriate biological processes, molecular functions, or cellular components are assigned. We adopted these relations between proteins and GO concepts, but clarified the relations by using link types with coherent in-depth semantics. The relation between a protein and a function is elucidated in LinKBase® by a has-function link type. This link type reflects the relation between a substance and its function or goal. Between a protein and its location within the cell, a has-spatial-point-reference link type is used that covers all possible spatial relations (Figure 3).

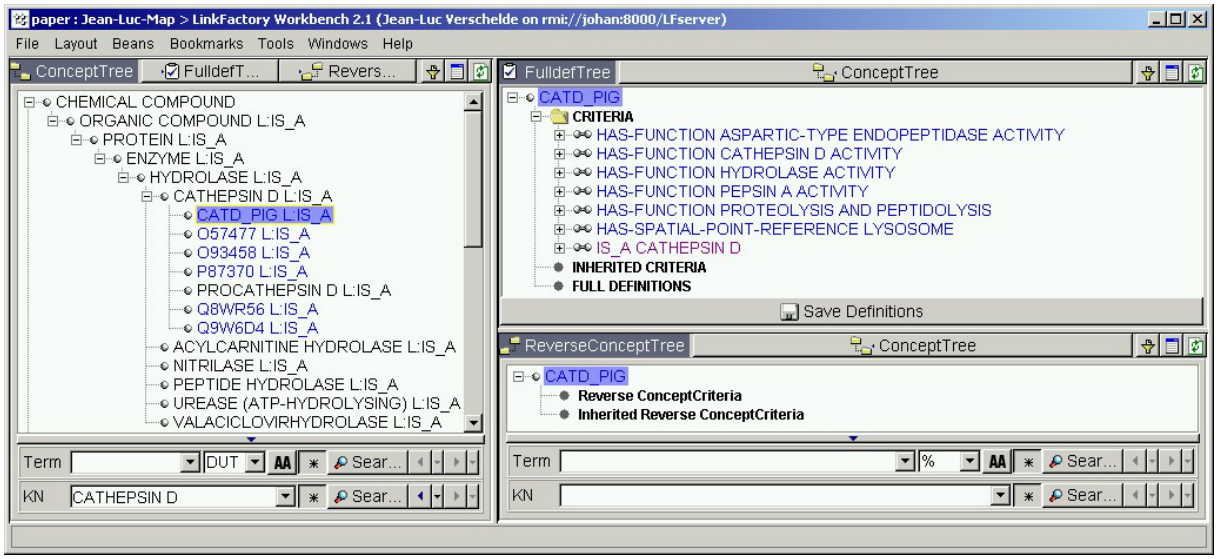

Figure 3 Cathepsin D from pig (<CATD_PIG>) is associated with different functions and is spatially related to the lysosome. <CATD_PIG> is an aspartic protease (has-function 'aspartic, type endopeptides activity') and is actively present in the lysosome (has-spatial-point-reference 'lysosome')

We identified the more general concepts of GO in LinKBase® and created new concepts in LinKBase® if they weren't recognized. We also enriched LinKBase® with concepts and ontological relations needed to capture the semantics of GO's top-level concepts. It was our objective, not to model fully the GO hierarchy in our ontology, but rather to map the GO hierarchy onto those GO top-layer concepts modelled in LinKBase®. This approach saved us much time and effort. Applying the same approach to other heterogeneous databases will allow LinKBase® to serve as a central reference framework for integration.

3.4 Improvement of GO's expressiveness

The structure and attributes of GO were examined in order to reveal problems that could endanger the semantic integrity of our extended LinKBase® and its applicability in the processing of biological data.

We examined GO's (problematic) part-whole relationship (*part-of*), identified different types of *part-of* relations in the three sub-domains, and proposed an improved representation of these variants:

- 'flagellum' *part-of* 'cell'. The flagellum is not part of every possible cell but only of some cells.
- 'membrane' *part-of* 'cell'. The membrane is part of any cell.
- 'flagellar membrane' *part-of* 'flagellum'. The flagellar membrane surrounds the flagellum.

'Flagellum' and 'membrane' are represented as children of <flagellum structure> and <membrane structure>, respectively. Both of the latter LinKBase® concepts are children of <subcellular structure> and both are dissective, which means that both are such that all parts of any instance of e.g. <flagellum structure> are also instances of <flagellum structure> (Figure 4). The same structuring principle [12] has been followed for the parts of <flagellum structure (sensu bacteria)> and <flagellum structure (sensu eukarya)>. All *part-of* relations in

the 'cellular component' domain of GO are modelled with the has-spatial-point-reference link type. In LinKBase®, incoming links are not inherited. This means that for instance the children of <cell> do not inherit an incoming has-spatial-point-reference link from the source concept <flagellum> and thus <flagellum> is only related to the universal concept <cell>. In LinKBase®, <cell> subsumes all different types of cells. Specific cell types that are known to have the cellular structure flagellum can be further detailed by linking them to <flagellum> with a corresponding special relationship. Many of the cellular components that were modelled as orphan concepts in GO are now properly structured. This method should also settle the problem annotators have in annotating genes or proteins where it is not clear from the literature whether the gene or protein should be pinpointed to a specific part of a cellular component or to the cellular component taken as a whole. For instance, in those cases where it is not clear if a protein is present in the nucleus and/or the nucleolus, we propose to annotate it to <nucleic structures>, which subsume the whole as well as the parts.

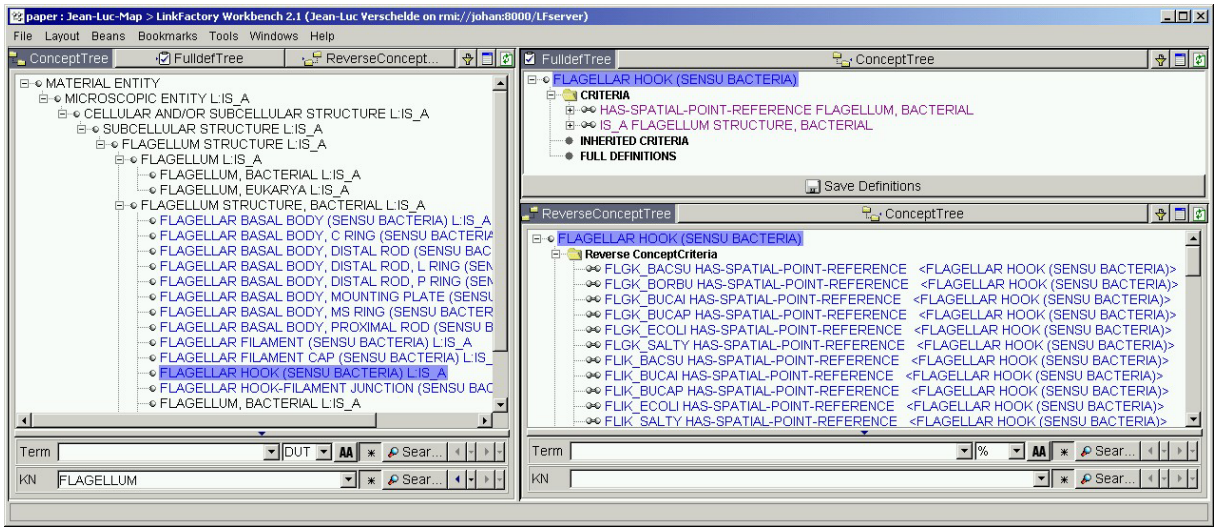

Figure 4 Representation of <flagellum> and its parts.

- 'regulation of signal transduction' *part-of* 'signal transduction'. The regulation is *part-of* the process called signal transduction.

Solution: < regulation of signal transduction> is-participating-process-of <signal transduction>. The target process <signal transduction> has the regulation process as one of its parts.

The different intended meanings of the *part-of* relation in the different sub-domains of GO are crystallized as link types in LinKBase® in such a way that they automatically acquire better semantics.

The use of the *is-a* relation in GO is intended to assign to a parent concept a child concept that inherits from the parent all associated criteria. This means that the child concept contains all the specifications of the parent concept but has at least one more. GO's *is-a* relation represents a kind of subsumption that has to be interpreted as follows - all instances of a child concept must always be an instance of the parent concept. However, GO also contains some inconsistencies in this respect, as is demonstrated by the following case. GO defines 'development' as "Biological processes aimed at the progression of an organism over time from an initial condition (e. g. a zygote, or a young adult) to a later condition (e. g. a multicellular animal or an aged adult)". 'Cellularization' is such a process and hence, is listed under 'development' and linked with an *is-a* relation (Figure 5). However, the same reasoning has been used for 'cellularization (sensu animalia)' that is subsumed by 'embryonic development (sensu animalia)', a process defined as a development process in its entirety. It is

clear that 'cellularization (sensu animalia)' is only part of the embryonic development process and erroneously the *is-a* relation is used here with the meaning *part-of*.

Figure 5 GO hierarchy of concepts related to 'development'

This ambiguity in the modelling of 'cellularization' could be resolved using our modelling approach based on formal rules. However, it is not our intention to change or remodel GO. This is because we have thus far been able to achieve all that we need merely by adding structuring information (for example in the <flagellum> case) and mapping GO relationships to link types in order to gain the virtues of a better semantics. Proper modelling of 'cellularization' would require changing the database information itself, which is not the task of database integration.

4 Conclusion

Our LinKBase® ontology is a representation of the medical domain. By mapping more specialized information sources like GO and protein databases, we were able to quickly expand the reach of our ontology and hence achieve a database warehousing system, within which all mapped databases are correctly related to each other in such a way that a global view of the dispersed information is possible. The MaDBokS system can be used to graft databases onto the ontology and thereby make the latter useable for a variety of applications. The flexibility of the MaDBoKS system and the speed with which databases can be integrated allows the prototyping of different integration protocols in relation to different sets of databases, and hence enables a fine-tuning of the integration process for specific applications such as data-mining and information extraction.

5 References

[1] T. Critchlow, M. Ganesh, R. Musick. Automatic Generation of Warehouse-Mediators Using an Ontology Engine. In Proceedings of the 5 th Interna-tionalWorkshop on Knowledge Representation meets Databases (KRDB'98). May 1998.

http://citeseer.nj.nec.com/critchlow98automatic.html

[2] A. Silvescu, J. Reinoso-Castillo, C. Andorf, V. Honavar and D. Dobbs. Ontology-Driven Information Extraction and Knowledge Acquisition from Heterogeneous, Distributed Biological Data Sources. Proceedings of the IJCAI-2001 Workshop on Knowledge Discovery from Heterogeneous, Distributed, Autonomous, Dynamic Data and Knowledge Sources.

[3] G. Wiederhold. Mediators in the Architecture of Future Information Systems. IEEE Computer, 38-49, March 1992.

http://www-db.stanford.edu/pub/gio/1991/afis.ps

[4] Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature Genet. 25(1): 25-29, 2000.

[5] E. Camon, M. Magrane , D. Barrell , D. Binns , W. Fleischmann , P. Kersey , N. Mulder , T. Oinn , J. Maslen , A. Cox , R. Apweiler. The Gene Ontology Annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro. Genome Res., 13(4):662-72, 2003.

[6] A. Flett, M. Casella dos Santos and W. Ceusters. Some Ontology Engineering Processes and their Supporting Technologies. Siguença, Spain, October 2002. EKAW2002.

[7] T. Bittner and B. Smith. A Theory of Granular Partitions. in: Foundations of Geographic Information Science, M. Duckham, M. F. Goodchild and M. F. Worboys, eds., London: Taylor & Francis Books, 117-151, 2003.

[8] F. Montyne, J. Flanagan. Formal ontology: The Foundation for Natural Language Processing. January 2003.

http://www.landcglobal.com/

[9] B. Smith, J. Williams and S. Schulze-Kremer. The Ontology of the Gene Ontology. Proceedings of AMIA 2003.

http://ontology.buffalo.edu/smith/

[10] Anand Kumar, Barry Smith. The Universal Medical Language System and the Gene Ontology: Some Critical Reflections. Proceedings of KI2003, Hamburg, September 2003

http://ontology.buffalo.edu/smith/

[11] B. Smith. Basic formal ontology.

http://ontology.buffalo.edu/bfo/

[12] U. Hahn, S. Schulz and M. Romacker. An ontological engineering methodology for part-whole reasoning in medicine. 1998.

http://citeseer.nj.nec.com/hahn98ontological.html