Generating Semantic Networks to the PubMed

Giovani Rubert Librelotto[1], Mirkos Ortiz Martins[1], Henrique Tamiosso Machado[1], Juliana Kaizer Vizzotto[1], José Carlos Ramalho[2], and Pedro Rangel Henriques[2]

[1] UNIFRA, Centro Universitário Franciscano, Santa Maria - RS, 97010-032, Brasil
{librelotto, mirkos, htmachado, juvizzotto}@gmail.com
[2] Universidade do Minho, Departamento de Informática
4710-057, Braga, Portugal
{jcr, prh}@di.uminho.pt

Abstract. This paper presents a topic map approach to PubMed in order to create a knowledge representation for this information system. PubMed is a free search engine that gives very full coverage of the related biomedical sciences. With more than 17 millions of citations since 1865, PubMed users have several problems to find the papers desired. So, it is necessary to organize these concepts in a semantic network. To achieve this objective, we use the Metamorphosis system, choosing the keywords from MeSH ontology. This way, we obtain an ontological index for PubMed, making easier to find specific papers.

1 Introduction

Daily, a lot of data is stored into PubMed system. There is a problem that organization requires an integrated view of their heterogeneous information systems. In this situation, there is a need for an approach that extracts the information from their data sources and fuses it in a semantically network. Usually this is achieved either by extracting data and loading it into a central repository that does the integration before analysis, or by merging the information extracted separately from each resource into a central knowledge base.

Topic maps are an ISO standard for the representation and interchange of knowledge, with an emphasis on the findability of information. A topic map can represent information using topics (representing any concept), associations (which represent the relationships between them), and occurrences (which represent relationships between topics and information resources relevant to them). They are thus similar to semantic networks and both concept and mind maps in many respects. According to Topic Map Data Model (TMDM) [GM05], Topic Maps are abstract structures that can encode knowledge and connect this encoded knowledge to relevant information resources. In order to cope with a broad range of scenarios, a topic is a very wide concept. This makes Topic Maps a convenient model for knowledge representation.

This paper described the integration of data from PubMed information system using the ontology paradigm, in order to generate an homogeneous view of those resources. PubMed is introduced in section 2. This proposal uses an environment, called Metamorphosis (section 3), for the automatic construction of Topic Maps with data extracted from the various data sources, and a semantic browser to navigate among the information resources – it is described in section 4. The section 5) presents the concluding remarks.

2 PubMed

PubMed [oM07] is a free search engine that provides very full coverage of the related biomedical sciences, such as biochemistry and cell biology. It also offers access to the MEDLINE database [oM06] with citations and abstracts of biomedical research articles.

The PubMed core subject is medicine and its related fields. It is offered by the United States National Library of Medicine as part of the Entrez information retrieval system. The inclusion of an article in PubMed does not endorse the article's contents, as other indexes. Nevertheless, many PubMed citations contain links to full text articles which are freely available, often in the PubMed Central digital library.

MEDLINE database covers over 4.900 journals published around the world primarily from 1966 to the present and is composed of more than 17 millions of citations. Information about the journals indexed in PubMed is found in its Journals Database, searchable by subject or journal title, Title Abbreviation, the NLM ID (NLM's unique journal identifier), the ISO abbreviation, and both the print and electronic International Standard Serial Numbers (pISSN and eISSN). The database includes all journals in all Entrez databases. A PubMed entry includes among other information the following details: *PubMed identifier*, *Authors' name*, *Title*, *Journal*, *Publication date*, *Language*, and *Mesh terms*.

The PubMed database consists of three tiers of software. At the bottom is a database management system (DBMS) that manages a collection of facts. At the top is the web browser that transmits requests for data to the database and renders the responses as web pages. In the middle is a software layer that mediates between the DBMS and the web browser to turn data requests into database queries, and to transform the query responses into hypertext mark-up language (HTML).

The PubMed data structure is composed of citations metadata. Each citation has the same structure. The main part of its schema can be formalized by the following context free grammar:

```
MedlineCitation     ==> PMID, DateCreated, DateCompleted, Article,
                        MedlineJournalInfo, ChemicalList,
                        CitationSubset, MeshHeadingList
Article             ==> Journal, ArticleTitle, Pagination,
                        Abstract, Affiliation, AuthorList,
                        Language, PublicationTypeList
Journal             ==> ISSN, JournalIssue, Title
JournalIssue        ==> Volume, Issue, PubDate
PubDate             ==> Year, Month, Day, Hour?, Minute?, Second?
MedlineJournalInfo  ==> Country, MedlineTA, NlmUniqueID
ChemicalList        ==> Chemical+
Chemical            ==> RegistryNumber, NameOfSubstance
MeshHeadingList     ==> MeshHeading+
MeshHeading         ==> DescriptorName, QualifierName?
AuthorList          ==> Author+
Author              ==> LastName, ForeName, Initials
PublicationTypeList ==> PublicationType+
```

PubMed files are intended for automatic processing and therefore available in XML format. Each set of 30.000 PubMed citations is stored as an XML instance defined by a DTD. Notice that the context free grammar above was obtained direct and systematically from the PubMed DTD.
For these reasons, it was defined an XML Schema to PubMed files. The view of this structure is shown in figure 1.

3 Metamorphosis

The main idea behind Metamorphosis is close the gap between Topic Map technology and its users. Metamorphosis is being developed to become a Topic Map workbench easy to use and accessible to a common user (we are not there yet). Figure 2 shows the usage scenario proposed in this paper. It illustrates some of the interaction between the system components, information resources and users.

1. Metamorphosis Repository (MMRep) is the central component that takes care of Topic Map storage and management. All the other components interact with MMRep.
2. Topic Map Discovery (TMDiscovery) is a Topic Map driven browser that allows users to navigate inside the Topic Maps stored in MMRep.
3. Topic Map Extractor (Oveia) automates the task of Topic Map harvesting; it enables the user to specify the extraction task and generates a Topic Map in XTM syntax that can be uploaded into MMRep. Oveia implements some extraction mechanisms with which is possible to populate an ontology.
4. Information resources that we want to access.
5. Web interface driven by a topic map stored in MMRep that provides access to information resources.

Metamorphosis can be used to prototype web interfaces or to expose information systems on the web. To do this the user only needs to specify a topic map for each view he wants. Information integration is accomplished by concept integration in the topic map: to integrate two information systems we need to specify the two sets of concepts in the same topic map and specify the associations that will materialize that integration.
In the next sections we are going to discuss the main components of this workbench prototype: Metamorphosis Repository, Topic Map Discovery, Oveia and XTche.
This way, Metamorphosis let us achieve the semantic interoperability among heterogeneous information systems because the relevant data, according to the desired information specified through an ontology, is extracted and stored in a

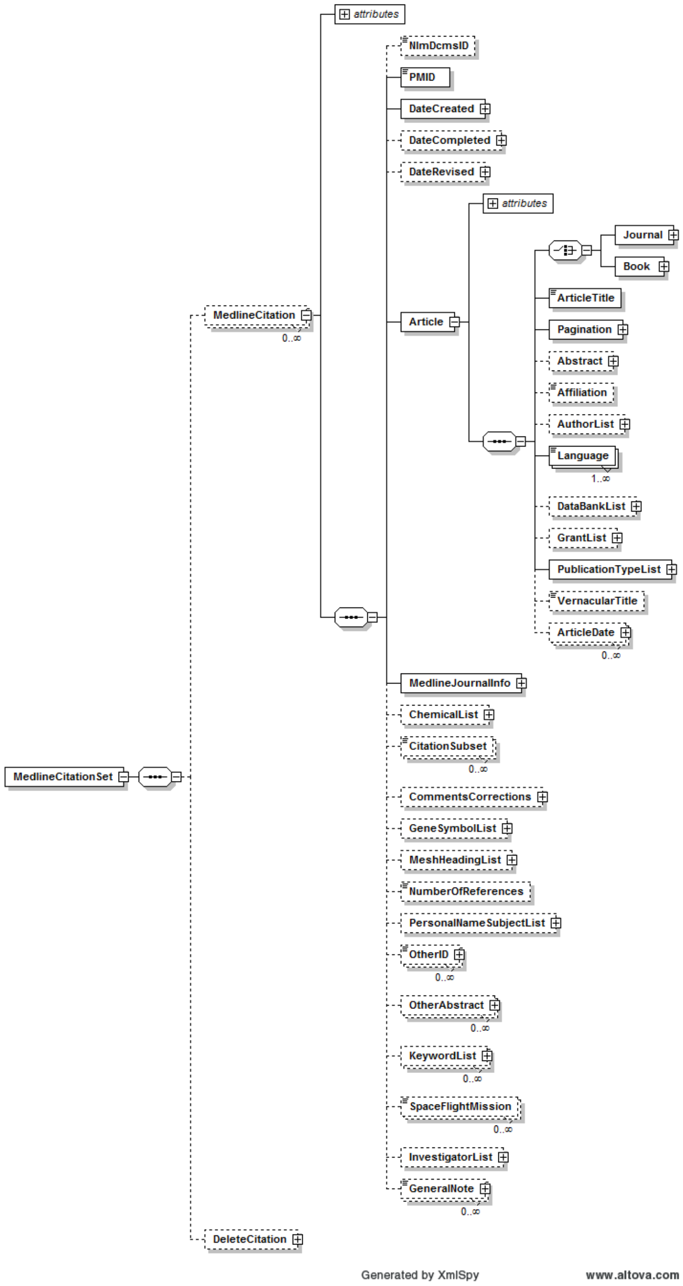


Fig. 1. PubMed's XML Schema

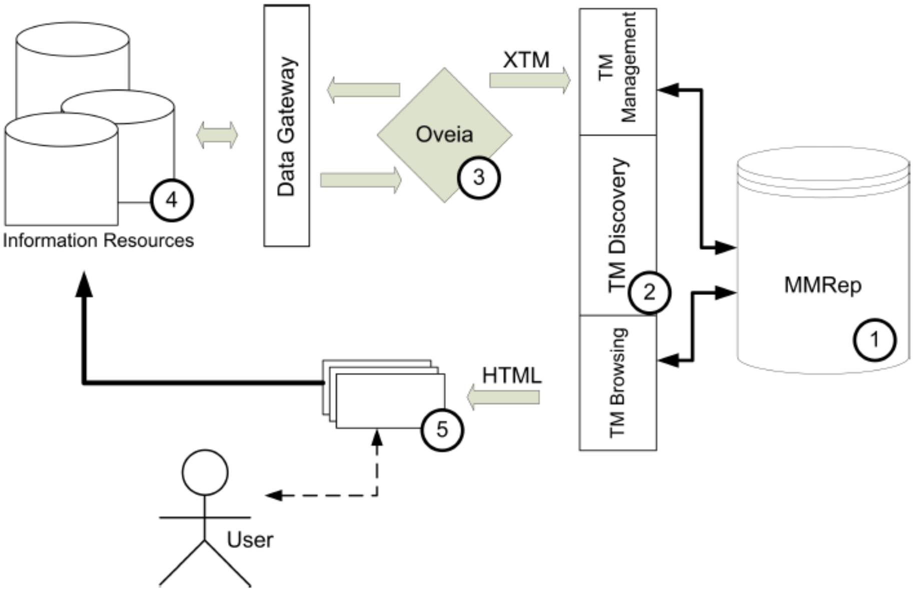


Fig. 2. Metamorphosis Functional Diagram

topic map. The environment validates this generated topic map against a set of rules defined in a constraint language. That topic map provides information fragments (the data itself) linked by specific relations to concepts at different levels of abstraction. Note that not all data items need to be extracted from the sources to the Topic Map. We only extract the necessary metadata to build the intended ontology. This ontology will have links to enable a browser to access all data items.

Thus the navigation over the topic map is led by a semantic network and provides an homogeneous view over the resources – this justifies our decision of call it semantic interoperability.

4 Topic Maps applied to PubMed

In order to obtain a semantic network from PubMed data, we divided this task in a few parts, as shown Figure 3.

In the first one, we created a relational database to store all contents of XML data obtained from PubMed data source. This database is generated according to the PubMed DTD using the Exult tool. An SQL script processes the result database to remove the redundant data and to erase several tables unnecessary. The final PubMed local database has XX tables.

To extract data from this database we use Metamorphosis [LRH06]. Metamorphosis has mechanisms to query the PubMed local database (Oveia) according to an ontology specification (XS4TM). Besides, there is a Web interface to make

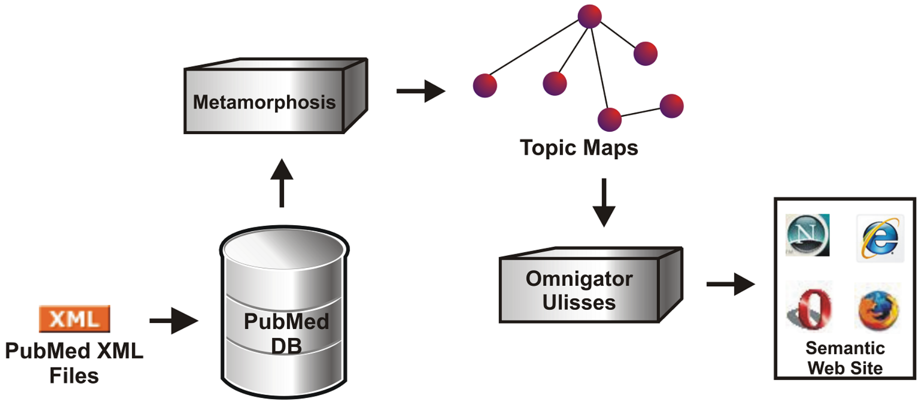


Fig. 3. The system's architecture

a query over the database. This interface has a text field to the user puts his query. After the query submission, Metamorphosis processes this string finding MeSH terms that describes the desired publications. These terms are structured in a RDF file [vAMMS07].
Using these MeSH terms as keywords, Metamorphosis searches articles that match with the user's query. This search processes includes several fields, like article's title, abstract, keywords, chemical substances, and MeSH terms. When an article satisfies the query, it will be mapped to a topic, as well its main fields, creating associations between them.
When the system receives a request, the required data will be collected from the selected databases at runtime. Then it will be further processed and converted into semantically relevant data by Metamorphosis. The resulting data has the standard XTM format. So, one of the advantages of this approach is that no new database will be created and no redundant data will be produced.
After end of the process, Metamorphosis has all topics and associations stored in its repository. The generated XTM documents can be then processed and displayed to the user by the presentation tier. This way, any topic maps navigator tool is able to browse the semantic network composed by these concepts. For instance, Ulisses [LRH04] allows the topic maps navigation over Metamorphosis' repository and XTM files (in last case, it is also possible to use Ontopia Omnigator [Ont02]). Information is interconnected within a huge knowledge network navigable in any direction.

4.1 Defining the Topic Maps concepts to PubMed citations

In order to define the topic map extraction from PubMed instances, the first task is to specify the main concepts (topic types). This way, the topic types in this domain are:

Article : each article is stored in a tag called $< MedlineCitation >$;

Author : the article authors are declared in $< Author >$;
Keyword : the keywords are MeSH terms. They are defined in $< MeshHeading >$;
Publication year : this metadata is in $//PubDate/Year$ path;
Journal : all journals are found in $< Journal >$ tag;
Language : the paper's language is define in $< Language >$;
Chemical substances : all chemical items cited in each paper are referenced in $< Chemical >$;

After the topics choice, the next step is the topic characteristics definition. Below we have the main ones:

Article : PMID (PubMed identifier), title, pagination, abstract, DOI, ...;
Author : initials, last name, middle name, and first name;
Keyword : descriptor and qualifier terms;
Journal : ISSN, title, abbreviation, volume, issue, and publication date;
Chemical substances : register number and substance name;

At this moment, all topics and its characteristics are defined. The final topic map definition step is the specification of association type. The main association types and some roles are described below:

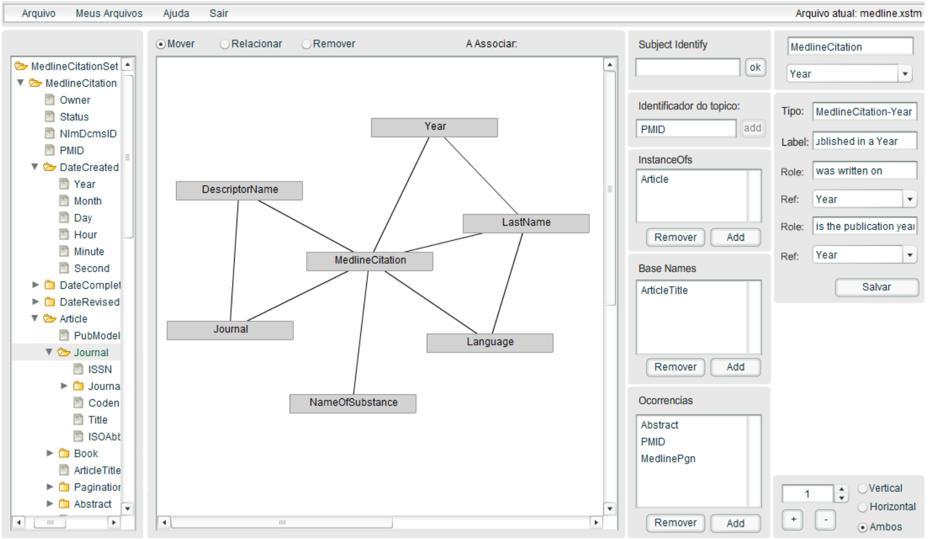
- Author writes article;
- Keyword describes article;
- Article was published in an year;
- Article is published in a journal;
- Article is written in a language;
- Article refers to chemical substances;
- Author publishes in an year;
- Author writes paper in a language;
- Journal refers to the keywords;

Looking at a TM we can think of it as having two distinct parts: an ontology and an object catalog. The ontology is defined by what we have been designating as topic type, association type, and association role. The catalog is composed by a set of information objects that are present in information resources (one object can have multiples occurrences in the information resource) and that are linked to the ontology.

The PubMed's topic map ontology defined above (topic types, roles, and association types) and the topic characteristics are mapped to an XS4TM specification as can be seen in next subsection.

The XS4TM specification describing the PubMed scenario was defined in a XS4TM Web editor. Figure 4 shows a view of this specification, which defines seven topic types, nine association types, and eighteen role types.

On the left side, XS4TM presents the XML tree extracted from PubMed's XML Schema. The topic types from this case study are shown in the center window. To create a new topic type, the user just needs to make a simple drag and drop



Arquivo atual: medline.xstm

Identificador do topico: PMID

InstanceOfs: Article

Base Names: ArticleTitle

Ocorrencias: Abstract, PMID, MedlinePgn

MedlineCitation

Year

Tipo: MedlineCitation-Year

Label: ublished in a Year

Role: was written on

Ref: Year

Role: is the publication yeai

Ref: Year

1

Fig. 4. PubMed's XS4TM Specification

from the XML tree. The topic characteristics are defined in the first column and the association characteristics are defined in the last column.
With the complete XS4TM specification, Oveia[3] can processes it. Its behavior can be described in four steps: (1) reads the XS4TM specification, (2) extracts the topics and associations from the query result set, (3) creates the topic map, and (4) stores it in the repository.

4.2 Browsing the topic map

When it will be browsing the semantic network obtained from PubMed local database, Ulisses gives the user an interface to navigate inside any of the stored topic maps. It allows the following interfaces:

Topic Maps : is the browser entry point and shows a list of all stored topic maps.

Ontology Index : gives you a structured view of a topic map showing the abstract concepts: topic types, association types, occurrence types, and association role types.

Individuals Index : lists all non-type topics in alphabetical order.

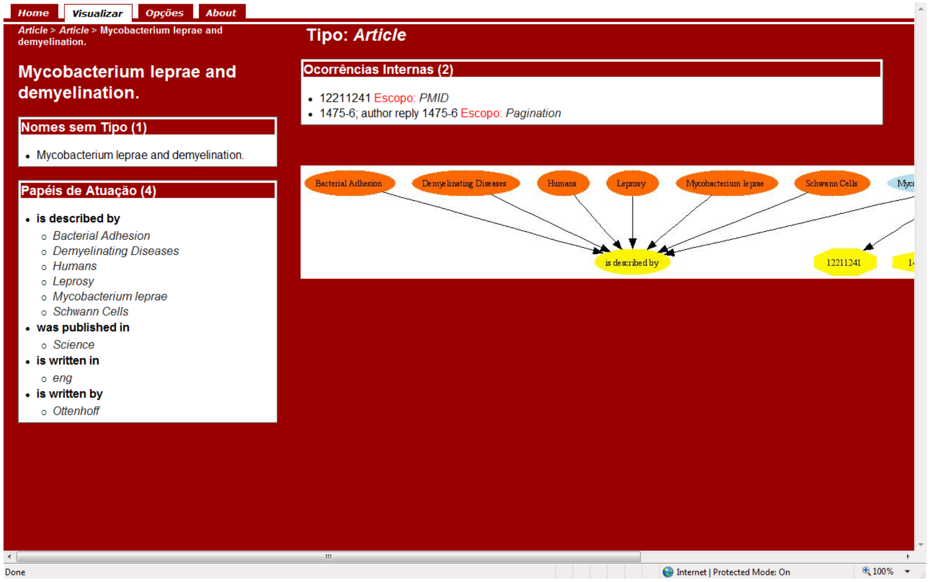
Full Index : lists all named topics.

Topic View : lists a subset of the available information about a topic; for the moment: the basenames, its type, all the associations it participates in together with the other members and their roles, internal occurrences and external occurrences.

[3] Oveia is a Metamorphosis' module

- **Association View** : lists the names associated with the association and all its descendants.

Figure 5 a view to the topic of type article called Mycobacterium leprae anddemyelination. This page display every topic characteristics and its associations in a Web way, as well in a graph view.



Article > *Article* > Mycobacterium leprae and demyelination.

Mycobacterium leprae and demyelination.

Nomes sem Tipo (1)

- Mycobacterium leprae and demyelination.

Papéis de Atuação (4)

- **is described by**
  - *Bacterial Adhesion*
  - *Demyelinating Diseases*
  - *Humans*
  - *Leprosy*
  - *Mycobacterium leprae*
  - *Schwann Cells*
- **was published in**
  - *Science*
- **is written in**
  - *eng*
- **is written by**
  - *Ottenhoff*

Tipo: *Article*

Ocorrências Internas (2)

- 12211241 Escopo: *PMID*
- 1475-6; author reply 1475-6 Escopo: *Pagination*

Fig. 5. Ulisses topic view

Creating a virtual map of the information enables us to keep the information systems in their original form, without changes. It is also possible to create as many virtual maps as the user wants generating multiple semantic views for the same sources.

5 Conclusion

This paper described the integration of data from PubMed information system using the ontology paradigm, in order to generate an homogeneous view of this resources. PubMed is a searchable compendium of biological literature that is maintained by the National Center for Biotechnology Information (NCBI).
The proposal uses Metamorphosis for the automatic construction of Topic Maps with data extracted from the various data sources, and a semantic browser to navigate among the information resources.
Topic Maps are a good solution to organize concepts, and the relationships between those concepts, because they follow a standard notation – ISO/IEC 13250 – for interchangeable knowledge representation.

In this paper we claimed that the semantic integration of PubMed documents is possible to achieve with Metamorphosis. In order to achieve this we proposed the following methodology:

1. Look at the information resources and decide how your conceptual view should look like;
2. Choose what information bits must be extracted in order to produce that conceptual view;
3. Specify the extraction task using Oveia;
4. Upload the generated Topic Map into MMRep;
5. Browse it with TMDiscovery and use this interface to access the information resources.

With this methodology the original information resources are kept unchanged and we can have as many different interfaces to access it as we want. We just have to create/generate/specify a Topic Map for each one.
As a future work we aim the integration of Topic Maps and MeSH headings minimizing *false hits* and saving time in the searches. Another project is to identify other useful – but frequently overlooked – features of the PubMed database.

References

[GM05] Lars Marius Garshol and Graham Moore. Topic Maps – Data Model. In *ISO/IEC JTC 1/SC34*. `http://www.isotopicmaps.org/sam/sam-model/`, January 2005.

[LRH04] Giovani Rubert Librelotto, José Carlos Ramalho, and Pedro Rangel Henriques. Ulisses: Um Navegador Conceptual para Topic Maps. In *XXXI Conferencia Latinoamericana de Informática*, pages 783–794, 2004.

[LRH06] Giovani Rubert Librelotto, José Carlos Ramalho, and Pedro Rangel Henriques. Metamorphosis - A Topic Maps Based Environment to Handle Heterogeneous Information Resources. In *Lecture Notes in Computer Science*, volume 3873, pages 14–25. Springer-Verlag GmbH, 2006.

[oM06] U.S. National Library of Medicine. MEDLINE – Fact Sheet. http://www.nlm.nih.gov/pubs/factsheets/medline.html, 2006.

[oM07] U.S. National Library of Medicine. PubMed. http://www.ncbi.nlm.nih.gov/sites/entrez?db=PubMed, 2007.

[Ont02] Ontopia. The Ontopia Omnigator, 2002. `http://www.ontopia.net/omnigator/`.

[vAMMS07] Mark van Assem, Véronique Malaisé, Alistair Miles, and Guus Schreiber. A Method to Convert Thesauri to SKOS. http://thesauri.cs.vu.nl/eswc06/, 2007.