A2-D7

GoPubMed: ontology-based literature search for the life sciences

---

Project title:	Reasoning on the Web with Rules and Semantics
Project acronym:	REWERSE
Project number:	IST-2004-506779
Project instrument:	EU FP6 Network of Excellence (NoE)
Project thematic priority:	Priority 2: Information Society Technologies (IST)
Document type:	D (deliverable)
Nature of document:	R (report)
Dissemination level:	PU (public)
Document number:	IST506779/Dresden/A2-D7/D/PU/b
Responsible editors:	Gihan Dawelbait, Heiko Dietze
Reviewers:	Albert Burger, Michael Schroeder, Uwe Assmann
Contributing participants:	Dresden
Contributing workpackages:	A2
Contractual date of deliverable:	30 August 2007
Actual submission date:	30 August 2007

Abstract

With the ever increasing size of scientific literature, finding relevant documents and answering questions has become even more of a challenge. Recently, ontologies — hierarchical, controlled vocabularies — have been introduced to annotate genomic data. They can also improve the question answering and the selection of relevant documents in the literature search. Search engines such as GoPubMed.org use ontological background knowledge to give an overview over large query results and to answer questions.

Here we give an overview over GoPubMed. We show how it can answer questions using the GeneOntology and the Medical subject Headings as background knowledge. We also demonstrate that GoPubMed is general by applying it to the problem of associating genes, tissues, and developmental stages, as described in the Edinburgh Mouse Atlas. GoPubMed builds on background knowledge in the form of ontologies, which are given for the previous two applications. We describe a method to automatically generate the vocabulary for ontologies and compare our method to 3 other approaches in the context of a lipid metabolism ontology.

The deliverable comprises three sections. GoPubMed is described in Section 1, MousePubMed in Section 2 and ontology generation in Section 3.

Keyword List

Text minning, ontology-based literature search, ontology engineering.

Project co-funded by the European Commission and the Swiss Federal Office for Education and Science within the Sixth Framework Programme.

GoPubMed: ontology-based literature search for the life sciences

Dimitra Alexopoulou[Dre], Michael R. Alvers[TI], Bill Andreopoulos[Dre], Liliana Barrio-Alvers[TI], Gihan Dawelbait[Dre], Heiko Dietze[Dre,TI], Andreas Doms[Dre,TI], Cecilia Eyre[Uni], Jörg Hakenberg[Dre], Jan Mönnich[TI], Laura Pickersgill[Uni], Conrad Plake[Dre,TI], Andreas Reischuck[TI], Loïc Royer[Dre], Michael Schroeder[Dre] Thomas Wächter[Dre,TI], Matthias Zschunke[TI],

[Dre] Technische Universität Dresden, Germany, [TI] Transinsight GmbH, Germany, [Uni] Unilever, UK

30 August 2007

Abstract

With the ever increasing size of scientific literature, finding relevant documents and answering questions has become even more of a challenge. Recently, ontologies — hierarchical, controlled vocabularies — have been introduced to annotate genomic data. They can also improve the question answering and the selection of relevant documents in the literature search. Search engines such as GoPubMed.org use ontological background knowledge to give an overview over large query results and to answer questions.

Here we give an overview over GoPubMed. We show how it can answer questions using the GeneOntology and the Medical subject Headings as background knowledge. We also demonstrate that GoPubMed is general by applying it to the problem of associating genes, tissues, and developmental stages, as described in the Edinburgh Mouse Atlas. GoPubMed builds on background knowledge in the form of ontologies, which are given for the previous two applications. We describe a method to automatically generate the vocabulary for ontologies and compare our method to 3 other approaches in the context of a lipid metabolism ontology.

The deliverable comprises three sections. GoPubMed is described in Section 1, MousePubMed in Section 2 and ontology generation in Section 3.

Keyword List

Text minning, ontology-based literature search, ontology engineering.

Contents

1 GoPubMed: Ontology-based literature search

1.1 Introduction

Which techniques use the Prominin-1 (CD133) marker? Which proteins are related to Alzheimer's disease? Which hormone is Autistic Disorder associated with? Is apoptosis a hot topic? Which are leading centers and scientists for liver transplantation? Where is the main research done for dengue and leprosy? What treatments does the web discuss for Alzheimer?

The scientific literature and the web hold answers to all of these questions, but it is difficult to obtain them with classical search engines, as they merely present possibly long lists of search results. In contrast, ontologybased search engines can use their hierarchical background knowledge to provide an intelligent filing system, which categorizes results. The categorization gives an overview over large result sets and can be used to answer questions. For example to find the techniques associated with CD133, a query for CD133 will return many documents as a long list in a classical search engine. In contrast, a search engine with ontological background knowledge will identify flow cytometry as a technique and categorize the documents accordingly. The user can then use this hierarchical filing system to select the few articles mentioning techniques and even fewer ones mentioning flow cytometry.

Key to this new search paradigm is the background knowledge, which is used to categorize documents. With efforts such as the GeneOntology[Ashburner et al., 2000] and MeSH, the needed knowledge is readily available. MeSH contains for instance the fact that flow cytometry is a technique and the GeneOntology contains that apoptosis is also known as programmed cell death and that caspases are part of the apoptotic programme.

The central problem of ontologybased search is the mapping of ontology terms to text, this task, known as term extraction, is difficult, as authors do not write their abstracts with an ontology in mind. For instance the mapping must be flexible and map the ontology term "transcription factor binding" to the text "... a transcription that binds ...", although it does not appear literally.

In the remainder of this chapter we give a brief introduction into ontologies, finding ontology terms and entity recognition in text. We show how GoPubMed.org, a search engine which uses the GeneOntology and MeSH to index PubMed, can answer the introductory questions and more. Furthermore we present GoWeb an application using the GoPubMed features to introduce an ontological knowledge base for web search. We conclude by comparing several other search engines, including other PubMed search engines and ontology-based search engines.

1.2 Ontology-based Text Mining

1.2.1 Ontologies

A fundamental aspect for the work of researchers is the need to share knowledge. In the beginning this was often done without the help of a controlled vocabulary or nomenclature. This is in particular applicable for the biomedical area and life sciences. There are many genes and proteins that have multiple names or identifier. An example is Hnrpa1 which is also known as Tis, Fli-2, heterogeneous nuclear ribonucleoprotein A1, helix-destabilizing protein, single-strand-binding protein, hnRNP core protein A1, HDP-1, and topoisomerase-inhibitor suppressed.

More over there seems to be also in some cases a competition for creative gene names like Cleopatra, Ariadne, groucho, lost in space, brokenheart, hairy, superman and many more. Of

course there have also been efforts to standardize names or at least to reach a consensus for naming. For instance in the context of yeast research and for human genes there are widely used standards, even if they are not always adhered to in literature.

Similar issues arise, if the task is to annotate genes and their function within the categories biomedical process, molecular function, and cellular components. You can find that

- Cellulose 1,4-beta-cellobiosidase is also known as exoglucanase,
- superoxide-generating NADPH oxidase as cytochrome B-245,
- thiamin as vitamin B1,
- pyrexia as fever,
- heme as haem, and
- Apoptosis as cell death.

The aim of ontologies is to reduce this problem. They include concepts, synonyms and their relation ships.

One prominent example for a widely used ontology is the GeneOntology [Ashburner et al., 2000]. In the beginning it was developed for the annotation of the fruitfly genome. Later the GeneOntology was adapted and expanded for mouse and other genomes and covers now biomedical processes, molecular functions, and cellular components. It uses two kinds of relationships to model the dependencies between the concepts: `isa` and `partof`. Today the GeneOntology is part of the Open Biomedical Ontology (OBO) effort, which houses over 60 ontologies covering many areas of interests. This includes anatomy, chemical compounds, development, experimental conditions, phenotype, taxonomy and more.

The second example are the Medical Subject Headings (MeSH). The MeSH thesaurus is developed by the U.S. National Library of Medicine (NLM). Its main purpose is to provide an index for the articles, books and other media in the National Library of Medicine. It tries to cover all relevant topics for the medical area this includes disease, anatomy but also others like geographic locations and experimental techniques.

There are other medical ontologies, e.g. GALEN, SNOMED and UMLS [Bodenreider, 2004]. An overview of all presented Ontologies is available in Table 2. The Unified Medical Language System (UMLS) has a different approach. It tries to integrate as much relevant ontology as possible. The UMLS consists of three parts: a metathesaurus, a semantic network and the specialist lexicon. Whereas the metathesaurus represents the concepts including the synonyms, the semantic network corresponds to categories and the specialist lexicon acts as a kind of index.

A non-trivial aspect is the design and later on the evolution of ontologies. With many thousands concepts and definitions how does one keep it all including the relations consistent. Although this starts with question: How is consistence defined in the first place? The GeneOntology follows an informal approach. The transitive closure still has to hold. This means, if a concept A `is-a` B and B `is-a` C then A `is-a` C has to be true. These inferred redundant relationships are not kept directly in the ontology. This helps to ease the maintenance of the ontology as corrections, modifications and additions only need to check if their direct relations are still valid.

Even though this consistency definition is a pragmatic solution there are more formal approaches. One such idea is the usage of description logics to formally define concepts and their relations. This was used for instance in the GALEN and SNOWMED ontologies. The

advantage of the formal definitions is the chance to automatically check for inconsistencies in the ontology. Imagine that one adds the new fact heparin `is-a` glycosaminoglycan, but it was not yet stated that heparin biosynthesis `is-a` glycosaminoglycan biosynthesis. Because of the formally defined relations and concepts, this additional relation can be inferred with this new fact in the knowledge base.

1.2.2 Finding Ontology Terms in Text

The ontologies presented above have been designed to annotate data or to be used as classification schemes. But they were not designed for the purpose to build novel search engines. Therefore the identification of ontology entities in free text remains a challenging task. For instance, a recent assessment for extracting GeneOntology terms revealed performances around 20% success rate only [Ehrler et al., 2005]. The difficulties of automating manual annotation is evident from the fact that only as few as 15% of manually annotated terms appear literally in the associated abstracts. Biomedical text mining uses various techniques and algorithms, e.g. natural language processing, information retrieval and machine learning, to identify the relevant entities [Jensen et al., 2006] and have to deal with groups of problems.

Ad-hoc Variations of Names To begin with, terms in vocabularies and labels of concepts in ontologies appear in many, slight or severe, variations in natural language texts.

- orthographic: IFN gamma, Ifn-γ
- morphological: Fas ligand, Fas ligands
- lexical: hepatitic leukaemia, liver leukemia
- structural: cancer in humans, human cancers
- acronyms/abbreviations: MS, Nf2
- synonyms: neoplasm, tumor, cancer, carcinoma
- paragrammatical phenomena/typographical errors: cerevisae, nucleotid

Some of the terms encountered in texts are rather ad-hoc creations, which cannot be found in any term lists.

Synonymity of Ontological Terms As mentioned before, terms in a vocabulary or ontology might not appear literally in a text, but authors rather use synonyms for the same concept. First of all, this complicates proper searches: When searching for "digestive vacuole", results should also contain texts that mention "phagolysosome"; mentionings of "ligand" refer to the concept "binding"; an "entry into host" might occur as an "invasion of host". In the Plant ontology for example, many synonyms exist for the same structure in different species. "Inflorescence" is referred to as "panicle" in rice, and as "cob" in sorghum, and "spike" in wheat, for instance. We note that there are also intra-ontology synonymities: "eye" in AnoBase can refer to the eye spot or the adult compound eye.

Ambiguity of Ontological Terms Terms can have a very specific meaning in biomedical research, but mean other things in other contexts. Examples are "development", "envelope", "spindle", "transport", and "host". Protein names such as "Ken and Barbie", "multiple sclerosis" or "the" that resemble common names, diseases, or common English words are especially hard to disambiguate. The same problems arise from drug names like "Trial" or "Act".

Stemming and Missing Words Some aspects for finding terms in text refer to the actual processing of natural language and appear rather technical. Very often, words will appear in different forms, such as "binding" and "binds". These refer to the same concept, which can be solved by resolving words to their stem ("bind"). However, the analogous reduction of "dimerisation" to "dimer" is more questionable. The former talks about the process, the latter about the result. A similar example is "organization", where a transformation into "organ" is invalid.

Texts contain additional words that are missing in the ontological term. This happens, for instance, when a text contains further explanations that describe findings in more detail. An example is "tyrosine phosphorylation of a recently identified STAT family member" that should match the ontology term "tyrosine phosphorylation of STAT protein." In general, matching is allowed to ignore words such as "of", "a", "that", "activity", but obviously not "STAT".

Additional background information on term variations is needed to know that a "family member" can refer to a protein. Formatting of terms represents another source for potential matching errors. Terms in ontologies contain commas, dashes, brackets, etc., which require special treatment. For "thioredoxindisulfide" the dash can be dropped, for "hydrolase activity, acting on ester bonds" the clause after the comma is important, but unlikely to appear as such in text. Terms containing additions such as "(sensu Insecta)" may have important contextual information, but are also less likely to appear in text.

Ontology Specific Issues

Term overlaps — some concepts can overlap in their labels or synonyms: in many cases there is a difference between what authors write and what they actually mean to express. Unfortunately, researchers do not have strict and formal ontologies or nomenclatures in their minds when composing a scientific article; in most of the cases they might use parent terms to refer to a child term, or viceversa. For example, many people are treating the MeSH terms 'cardiovascular disease' and 'coronary artery disease (CHD, CAD)' the same, although the latter is a child of the first.

Descriptive labels — in most of the cases, the labels in an annotation ontology cannot be used directly for text mining, often due to their explanatory nature. For example, it is unlikely that the Gene Ontology term "cell wall (sensu Gramnegative bacteria)" will appear as such in text. Terms like "positive regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism" and "dosage compensation, by inactivation of X chromosome" are almost complete sentences and are also unlikely to be found as such in text.

Ambiguity — results either from identical abbreviations for different terms, or, in general, tokens that can refer to terms that might or may not be of our interest. An example of an ambiguous *abbreviation* is "CAM" that can stand for "constitutively active mutants", "cell adhesion molecule", or "complementary alternative medicine". The second category of ambiguities — and the most difficult to handle — is that of terms that (in the context

of anatomy) can refer to different species. An example of such ambiguities is "embryo", which can be a chicken, mouse, human, or even zebrafish embryo.

1.2.3 Entity Recognition

Although finding ontological concepts in free text is important, there are many more relevant things to find in the text for instance:

- proteins,
- genes,
- species, or
- mutations.

The task to find these entities is called entity recognition. The identification of ontology terms can be seen as a sub species of the more general task of entity recognition. As a consequence many of the techniques and problems described above are also valid for entity recognition. But for instance for protein and gene name identification there some other difficulties [Finkel et al., 2005].

One challenge is the increased ambiguity and synonymity of names. Often the gene name and the protein are used by the authors as synonyms or a gene has the same name in different organisms. Another task is to deal with the number of entities one can find. As an example, the UniProtKB/TrEMBL protein database contains over 4,500,000 entries. The real number to match is even higher as one has to integrate all the synonyms and variants of the protein names and genes. For the case of identification on which species an article talks about, a reoccurring problem is, that the species is sometimes never mentioned in the text. For mentionings of point mutations one has to recognize the mutations and the related proteins to have a useable result [Lee et al., 2007, Baker and Witte, 2006]. Moreover many cases of ambiguities and missing concepts can only be resolved if one tries to use any information available from the text. For example if a gene name was found, that lists several possible proteins. To reduce the list of candidates, one can try to find the species or a point mutation in the text and than verify if they match with any of candidate proteins.

he importance of entity recognition and their relations has been acknowledged by the scientific community. There have been efforts to establish benchmarks and competions to advance the research. Examples for this are the "bioentity recognition task at JNLPBA" [Kim et al., 2004] or the "Critical Assessment of Information Extraction in Molecular Biology" (BioCreAtIvE) [Hirschman et al., 2005]. In the BioCreAtIvE II for the gene mention task the best systems [Hakenberg et al., 2007] could achieve a precision of 78.9% and a recall of 83.3% as an example for the current state of the art.

All of the above problems mean that extracting entities from literature will not be errorfree. However, despite all of these problems, ontologybased literature with text mining can answer questions as posed in the introduction. Next, we introduce GoPubMed and illustrate how they help to answer questions.

1.3 Question Answering with GoPubMed

Traditional keyword based searching gives a possible very long list of results. But finding the relevant documents is only the start; the user has to check if the results are relevant to him.

Often there is a question behind query. GoPubMed can answer all the introductory questions, as it uses the ontological background knowledge, namely the GeneOntology and MeSH to index search results. This allows GoPubMed to categorize the search results, identify relevant terms in the result set and to summarize trends for a topic. This topic can either be a term and its children or the result set of a query. For ontology enhanced web search the GoWeb systems is available. Figure 1.4 shows a screen shot of GoPubMed. The main panel contains the search results and the panel on the left the relevant categories from the ontologies in a summary and as a tree.

Now let us consider the questions and more importantly the answers in detail. Please consider that these questions were answered with GoPubMed in July 2007, due to the increasing number of publications the results may vary in the future.

Question: Which techniques use the Prominin-1 (CD133) marker?

Answer: Search in GoPubMed for "CD133" and open Techniques and Equipment in "top five & more" on the left. Listed as first is "Flow Cytometry". If you hover with the mouse above this term, you will see the description in a tool tip. The listed articles for flow cytometry contain statements like:

> "CD133+ and CD34+ cells were analyzed by flow cytometry to assess expression of cell division antigens" (Denner L. et al., Cell Prolif., 2007).

Other interesting terms are "Cell Separation" and "Immunohistochemistry" There you can find a statement like:

> "Microarray screening, single and dual-label immunocytochemistry and RT-PCR were performed to detect embryonic and neuronal stem cell markers, such as Oct3/4, Nanog, CD133, and Musashi-1." (Seigel GM et al., Mol. Vis., 2007).

A follow up question might be "Which types of cells are often targeted with these techniques?" The answer is already present "Stem Cells"; it is the top term for the query.

Question: Which proteins are related to Alzheimer's disease?

Answer: Type in Alzheimer and open chemicals and drugs in "top five & more" on the left. Among others there are "Amyloid", "Amyloid beta-Protein" and "Cholinesterase Inhibitors" listed as related proteins. By clicking on Amyloid beta-Protein, we can reduce from 1000 to 60 relevant articles and get the following definition:

> "A 4-kDa protein, 39-43 amino acids long, expressed by a gene located on chromosome 21. It is the major protein subunit of the vascular and plaque amyloid filaments in individuals with Alzheimer's disease and in aged individuals with trisomy 21 (DOWN SYNDROME). The protein is found predominantly in the nervous system, but there have been reports of its presence in non-neural tissue."

The article with from Ohyagi Y et al. from 2007 mentions e.g. "Inhibition of aggregation of amyloid pprotein (AP) ... are known as potent therapeutic tools for Alzheimer's disease (AD)." Another article (Chiarini A. et al., Ital J. Anat. Embryol., 2006) states "Reportedly, betaamyloid peptides (Abeta40 and Abeta42) induce the neurodegenerative changes of Alzheimer's disease (AD) ...".

Question: Was Abeta42 already used in a clinical setting?

Answer: Enter "Abeta42 drug" into the GoPubMed system and go on the result page to hierarchy of content on the lower left. Open first the category "Chemicals and Drugs" and than "Organic Chemicals". By clicking on "hydrocarbons" you reduce the result set to only 41 articles. A quick skimming over of the abstracts reveals statements like

- "... minocycline treatment did not alter the cerebral deposition of Abeta ..." (Fan R et al., J Neurosci, 2007), or
- "... naproxen that do not lower Abeta42 ..." (Cole GM. et al., Ann. NY Acad. Sci., 2004).

Select the next category "Carboxylic Acids", this will display 36 articles. On top of the results you can find again the definition of the term but also a link the Wikipedia article about carboxylic acids. The list of articles includes statements such as

- "... ibuprofen possess preferential Abeta42-lowering activity ..." (Leuchtenberger S. et al., Curr. Pharm. Des., 2006).

The GoPubMed system provides also links for mentioned protein names in an article, e.g. APP. This links opens the EBISwissprot database showing a list of all proteins related to APP.

Question: Which hormone is Autistic Disorder associated with?

Answer: Submit Autistic Disorder as query in GoPubMed. In "hierarchy of content" open "Chemicals and Drugs", then "Hormones, Hormone Substitutes and Hormone Antagonists", and then select "Hormones", which reduces the number of relevant articles to 49. For more details on special hormones you can browser in the "Gonadal Hormones" category which also has the term "Testosterone". Selecting testosterone, the result now shows 5 articles, with sentences like

- "... that high fetal testosterone levels could play a role in the aetiology of autism." (de Bruin EI. et al., Dev. Med. Child Neurol., 2006),
- "Fetal testosterone and sex differences in typical social development and in autism" (Knickmeyer RC. et al., J Child Neurol, 21 (10): 825-45, 2006), or
- "... high levels of testosterone influences some autistic traits and that hormonal factors may be involved in vulnerability to autism." (Knickmeyer R. et al., Horm. Behav., 2006)

For more example questions and answers have look at Table 2.

1.3.1 Hot Topics

Despite the overall growth of literature, some topics are hot and take-off while others are stagnant or are in a cool down phase. Bibliometric analyses aim to shed light on such developments and help to identify emerging trends. Such analyses data back to the 1960s [de Solla Price, 1965] and typically focused on research topics [Garfield and Melino, 1997], specific journals [Boyack, 2004], or the researchers themselves [de Solla Price, 1965, Newman, 2004]. The Hot topic feature of GoPubMed features views on ontology terms from the knowledge base. It

considers a term and all its children as one topic. For each topic a bibliometric analysis is provided.

The hot topic page for an ontology term includes two graphs showing the absolute number of publications per year for a topic. The second graph shows the relative share compared to the total number of publications per year in PubMed. An increase in the share indicates that the topic is growing faster than overall number of publications. Both graphs can be used to check whether the publication activity in a topic is decreasing, stagnant, or growing. In addition the publication count you can find a list of the most active authors, the list of journals with the most publications for this topic and a list of cities and countries with the most publications. To visualize coauthorship, which author publishes together with which other authors, we provide a coauthor network image. Publications between authors are denoted as edges between the author nodes. If no edge exists then the authors did not yet publish together, according to the publications listed in PubMed for this topic. The last feature is a world map where red dots indicate where all the publications are located for the current topic. All these features of the hot topics page are precalculated using the list of authors and affiliation of an article and the annotations from the GoPubMed system for all 16 Million PubMed articles.

To check the hot topics in GoPubMed for a term there are two options. The first way is to just search for the term in the normal search field and select the link from the list after "Show statistics for term:". Or second option, one can directly use the "Hot Topics" mode by selecting it in top bar. There you could also choose to use the advanced search, use the help page, a contact form or see the content of your clipboard.

Question: Is apoptosis a hot topic?

Answer: Use the hot topics to search for apoptosis. There are two apoptosis entries available, one from the GeneOntology and the other from MeSH. Select one of them by clicking on it. To answers the questions about trends have a look at the two graphs in Publications over time. They both reveal that the topic has been growing since the early 1990's. This is in line with Garfield and Melino's [Garfield and Melino, 1997] investigation of the field. But the second graph with the relative research interest shows also, that in the last 3 years the growth was not faster than the average growth of the whole PubMed literature.

Question: Which are leading centers and scientists for liver transplantation?

Answer: Query GoPubMed for "liver transplantation" and open the hot topics statistics for this term (see also Figure 1.4). Among the top authors is "Neuhaus P" and among the top cities is "Berlin". Prof. Peter Neuhaus works at the Charité Hospital Berlin, Germany. He is a leading specialist in the field. A look in the coauthor graph reveals with whom Peter Neuhaus has worked and published with.

Question: Where is the main research done for dengue and leprosy?

Answer: Retrieve the term statistics for Dengue. You will find that in the list of top cities there are Bangkok and Rio de Janeiro as the two top cities. In the top countries Brazil, Thailand and India are in the top 4.

For the term Leprosy you will find in the countries section India is the top country. This is also reflected in the list of important cities, where one can find several cities located in India. Both terms show that the local occurrence of diseases can be shown in GoPubMed.

All the examples for the usage of hot topics where based on the precomputed statistics using the ontology terms from the knowledge base as topics. But the result set of a given query may also be seen as a topic. This dynamic hot topics feature of GoPubMed offers you a bibliometric analysis of any result set of a query. This analysis contains the graphs about the publications over time, the lists of top authors, journals, cities and countries. It also includes the world map for the visualization of the geographic locations.

The dynamic hot topics can of course also be used to answer questions, for instance:

Question: Who are the top authors for Abeta42 Protein?

Answer: Use the GoPubMed site to search for "Abeta42". This query finds currently 767 articles. In the query summary field above the articles there is a link saying "Show statistics for these 767 articles". Clicking on this link will lead to the dynamically created hot topics. After the two graphs for the publications over time, there is the list of top authors. Listed there you can find for instance "Bennow K" as top author. The number of shown authors can be increased by clicking on the "more" link below the table. The publications for the author can be retrieved by using the provided link with the author name from the table.

Question: Who publishes most at the Max Planck Institute of Molecular Cell Biology and Genetics, Dresden (MPI-CBG)?

Answer: Search for "Dresden[AD] Planck[AD] Genetics[AD]" and click on the link "show statistics for these 305 articles". Currently the top author is Kai Simons with 41 publications, but this will probably change when new articles are published. In addition to the people one can also easily retrieve all publications in the Science journal from the MPI-CBG by clicking on the provided link "Science" in the top journals list.

This example can be extended to be used with any institution mentioned in the affiliations of PubMed articles. One might also consider to use date ranges (e.g. years) to check for changes in the publication profiles over time.

1.3.2 GoWeb

Sometimes the search with PubMed is not enough and the user wants to use normal general purpose search engines like Google or Yahoo. With GoWeb (gopubmed.org/goweb/) we offer internet search with ontological background knowledge. Some of the resources you can search with are for instance full text articles not included in PubMed, nonscientific sources like wikipedia or web based patent databases, commercial sites and vendors for equipment, special interest sites like the alzforum.org, or even news sites.

GoWeb uses standard web search engines and categorizes the results with its annotation algorithms. Normally web searches return not only the url but also the title and a short text snippet from the result page containing your searched keywords. These texts are textmined and the resulting terms are used in the same way as in GoPubMed to present the results of your search. You can use the ontological background knowledge to answer questions and reduce the result in a fast an efficient way without the need to read all the presented results. It includes, if available, also wikipedia links and protein names. Some example questions and answers are:

Question: Are there antibodies for ADDL?

Answer: GoPubMed can also search the web. Go to gopubmed.org/goweb and type ADDL antibody. Open "Chemicals and Drugs" and click on "Antibodies, monoclonal". The search results are now reduced from 100 to 8. Besides many pages of the Alzforum, there is the news that "Acumen and Merck Enter Into Alzheimer's Collaboration" which talks about: "... exclusive rights Acumen's ADDL technology monoclonal antibodies ... million development approval milestones first antibody product is commercialized. ..."

Question: What treatments does the web discuss for Alzheimer?

Answer: Go to gopubmed.org/goweb and type "Alzheimer treatment". Go to "Chemicals and Drugs" there you can find the term Memantine and also the term Vitamins. For more information on Vitamins click on the term. This will reduce the result set from 100 to 2 documents. In the result snippets you can find a statement like: "... vitamin may also be an ideal natural treatment for Alzheimer's disease too. ... Over the course of a small study, researchers at the University of Wisconsin ..."

1.4 Comparison and Conclusion

Currently, there is a lot of interest in literature search as is evidenced by the recent engines such as Google Scholar or Microsoft's Windows Live Academic (see Table 1.4). This includes also publishers like Elsevier with Scopus. These Engines offer a more comprehensive or different document base, than the classical PubMed search does, but they currently do not include intelligence to answer questions.

GoPubMed [Doms and Schroeder, 2005] indexes PubMed search results with ontological background knowledge, such as GeneOntology and MeSH. As shown above, this novel approach to search can help to answer questions. In particular the summary of important terms in "top five & more" is a most helpful feature for answering questions or reducing the big initial result to a smaller set of relevant articles in one click. With GoWeb the ontological background knowledge can also be applied to normal web search and be able to nonPubMed sources to answer questions.

GoPubMed's hot topics feature additionally allows users to get an overview of research trends, relevant journals, key authors and regional research interests. This feature is not provided by any of the other engines so far. GoPubMed is scalable and the system currently handles a user's search result up to 10.000 documents. It provides also additional useful features like links to wikipedia pages and mentioned proteins in SwissProt.

As GoPubMed is not the only PubMed search engine we give here a brief comparison of other tools (see also Table 1.4):

HubMed [Eaton, 2006] is direct front end to PubMed. It offers tools for the citation management of found PubMed articles. It also provides options for expanding the query or clusters the results in categories. This is all based on the MeSH terms directly provided by PubMed. If there are no MeSH concepts available for an article, than this features do not work, because no term matching is done by HubMed itself. This is usually the case the more recent articles. As an alternative they offer a tagging system where you can add your own tags to an article.

iHOP [Hoffmann and Valencia, 2004] uses genes and proteins as hyperlinks between sentences and abstracts. It converts the information in PubMed into one navigable resource.

The navigation along the gene network allows for a stepwise and controlled exploration of the information space. Each step through the network produces information about one single gene and its interactions.

eTBLAST [Lewis et al., 2006] is quite different approach to search the PubMed articles. It is based on text similarity and allows you to search for related articles using a relevancy ranking different from PubMed. Input a paragraph/abstract which is relevant for your search and eTBLAST returns a list of articles. For a search result one can list relevant authors, journals and a timeline.

PubFinder [Goetz and der Lieth CW von, 2005] Simliar to eTBLAST it can find related articles from a set of abstracts. It derives a list of discriminating words, which is subsequently used for scoring all defined PubMed abstracts for their probability of belonging to the defined scientific topic.

Textpresso for C. elegans [Mller et al., 2004] has been developed as part of the Wormbase effort. It offers currently about 100 concepts such allele, anatomy, association, characterization, clone, comparison, consort, developmental stage, disease, drugs, effect, entity feature, gene, involvement, life stages, mutants, nucleic acid, organism, pathway, phenotype, purpose, regulation, reporter gene, restriction enzyme, sex, spatial relation, strain, time relation, transgene, transposon, vector and including also a subset of GeneOntology concepts. It searches only abstracts and full text articles relevant for C. elegans. Textpresso does not offer an ontology tree for the exploration of a result set.

Vivisimo ClusterMed does not use existing ontologies, but clusters documents hierarchically, although it distinguished between categories like title and abstract, authors, affiliation, or publication date.

From document clusters, it derives representative terms. This automated hierarchy generation inevitable merges concepts of different nature, as the algorithm is only guided by the given documents, thus missing a lot of background knowledge a human uses in the creation of an ontology. Since Vivisimo clusters documents on the fly there is a limit to its scalability.

XploreMed [Perez-Iratxeta et al., 2003] filters PubMed results by the eight main MeSH categories and then extracts topic keywords and their co-occurrences. Abstracts can be retrieved for co-occurring keywords. The topic keywords are single words, usually occurring with a high frequency. Thus multi word concepts such as “Stem Cell” are not proposed as keyword. Currently XploreMed has a limited scalability and searches are restricted to 500 documents.

The combination of text mining and ontology-based background knowledge holds the possibility for intelligent search either in literature or in the web. With a new generation of emerging search engines, biomedical researchers can answer questions and get an overview over a topic.

Ontologies	
geneontology.org	Ontology with $\geq$ 20.000 terms on biomedical processes, molecular functions and cellular component
nlm.nih.gov/mesh	Medical Subject Headings created by the U.S. National Library of Medicine, taxonomy with $\geq$ 150.000 terms
opengalen.org	formal medical ontology, with $\geq$ 70.000 terms
snomed.org	commercial medical ontology, which contains $\geq$ 350.000 terms
nlm.nih.gov/research/umls/	Unified Medical Language System created by the U.S. National Library of Medicine, contains $\geq$ 1.000.000 terms
obofoundry.org	Open Biomedical Ontology, collection of over 60 specialized biomedical ontologies
Search engines	
pubmed.org	NIH's literature search engine
hubmed.org	"PubMed rewired"
invention.swmed.edu/etblast/	Text similarity: an alternative way to search Medline
glycosciences.de/tools/PubFinder/	PubFinder
www-tsujii.is.s.u-tokyo.ac.jp/medie	MEDIE answering questions
ihop-net.org	iHOP, gene network for navigating the literature
scholar.google.com	Google's literature search engine
academic.live.com	Microsoft's literature search engine
scopus.com	Elsevier's literature search engine
clustermed.info	Document clustering on the fly with Vivisimo
Ontology-based literature search engines	
gopubmed.org	Explore PubMed with ontological background knowledge
textpresso.org	Wormbase full texts with many ontologies
xploremed.org	Classification with high-level MESH headings and word co-occurrences

Table 1: URLs for ontologies, literature search engines and ontologybased literature search engine.

Which diseases are associated with HIV? **Answer:** Type "HIV" and wait for the tree on the left to appear. Go to "top five & more" and click on "disease". Among others hepatitis and tuberculosis are mentioned. Clicking on tuberculosis retrieves the relevant articles including statements such as "HIV and parasitic co-infections in tuberculosis patients".
Which anatomical structure is affected by the bacterium helicobacter pylori? **Answer:** Type "helicobacter pylori", go to "top five & more" and open "anatomy" Among the terms listed is "gastric mucosa". Hovering the mouse over the term reveals an explanation, which mentions that gastric mucosa is the lining of the stomach.

Which biological process is the protein Rab5 involved in and where is located in the cell? **Answer:** Type "rab5" and wait for the tree on the left to appear. Go to "top five & more". Click on biological process shows "endocytosis" and clicking on "cellular component" shows "endosomes". Hovering over the terms displays brief explanations of what endocytosis and endosomes are.
In which organisms is toluene degradation studied? **Answer:** Type "toluene degradation" and wait for the tree on the left to appear. Go to "top five & more" and open "organisms". The bacteria pseudomonas is listed first. A click retrieves the relevant articles.
Which enzymes are inhibited by aspirin? **Answer:** Type "aspirin" and wait for the tree on the left to appear. Go to "hierarchy of content" and then "chemicals and drugs" and "enzymes and coenzymes". From there always click the top child until you reach "cyclooxygenase 1" and "cyclooxygenase 2". Clicking reduces the articles to a few which mention that aspirin inhibits cyclooxygenases.
Which enzymes are important for congenital muscular dystrophy? **Answer:** Type "congenital muscular dystrophy" and wait for the tree on the left to appear. Go to "hierarchy of content" and then "chemicals and drugs", "enzymes and coenzymes", "enzymes", "transferases". There are a number of articles with statements such as "glycosyltransferases has revealed a novel mechanism for congenital muscular dystrophy."
Which techniques are frequently used to study zebrafish development? **Answer**: Search for "zebrafish development". Under "top five & more" open "techniques and equipment". In situ hybridization is listed first. Clicking the term retrieves relevant articles.
Which process are osteoclasts involved in? **Answer:** Search for "osteoclast". Under "top five & more" open "biological process". The first entry is "bone resorption".
What are common histone modifications? **Answer:** Search for "histone modification". Under "top five & more" open "biological sciences" and find methylation and acetylation.
Which diseases are associated with wnt signalling? **Answer:** Search for "wnt signalling". Under "top five & more" open "disease" and find "carcinoma" and many other cancer terms.
Were there clinical trails focusing on Abeta42 and were any side effects observed? **Answer:** Search for "Abeta42 clinical trail". In "top five & more" open Diseases and click on "Meningoencephalitis". The result now shows 4 articles, with titles like "Subacute meningoencephalitis in a subset of patients with AD after Abeta42 immunization". So, yes there where clinical trials, but there were also severe side effects like brain inflammation.
Which molecular function is Autistic Disorder associated with? **Answer:** Search for Autistic Disorder. Under "top five & more" open "Molecular Function" and find "neurexin binding".
Which disease is Autistic Disorder associated with? **Answer:** Search for Autistic Disorder. Under "top five & more" open "Diseases" and find for instance "Fragile X Syndrome" as a related disease.

Table 2: More example questions answered with GoPubMed

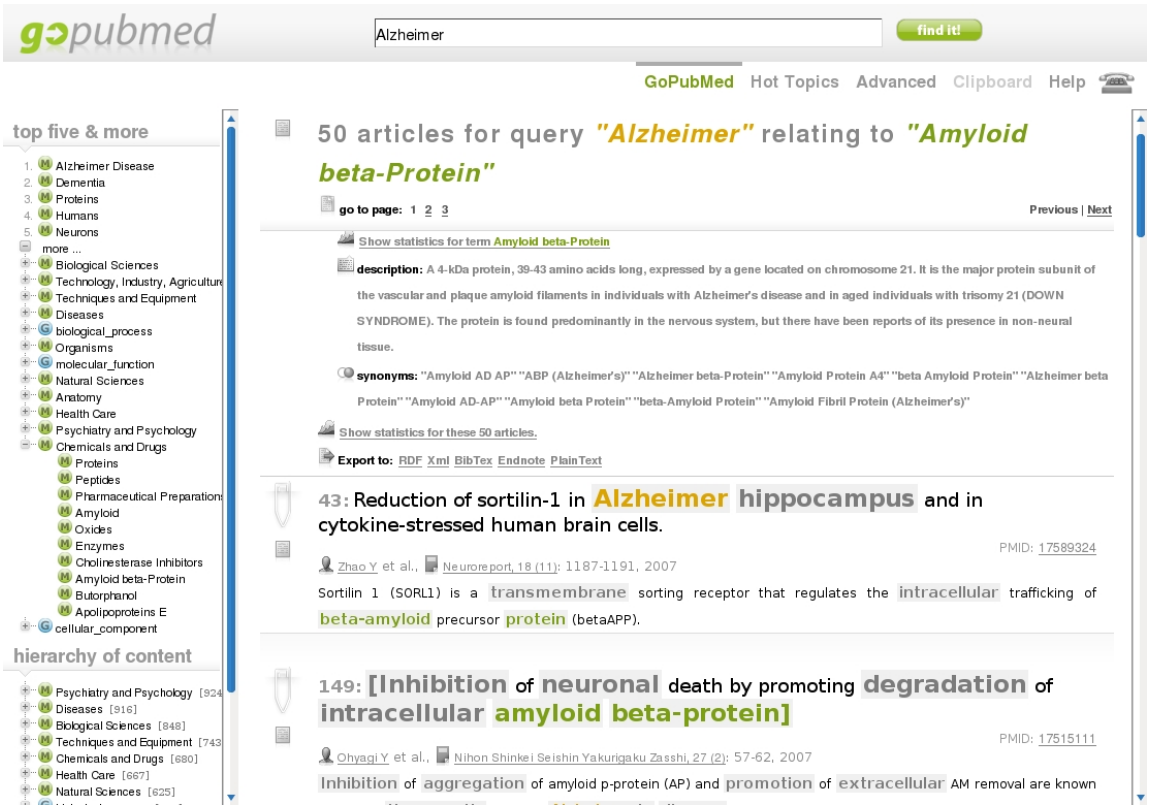

Figure 1: **Which proteins are related to Alzheimer's disease?**
GoPubMed uses its ontological background knowledge to index search results according to the GeneOntology and MeSH. The interface consists of three parts. The top most part contains the input field for the query, in this example it is "Alzheimer". You can submit a query by using the "find it!" button. The panel below comprises the results for the query and is split into to a left and a right part. The left panel contains the ontological background knowledge relevant to your query. A summary over all identified terms in your result is presented in "top five & more". If you open the category "Chemicals and Drugs" you can find also proteins. In "hierarchy of content" the complete induced ontology tree is available for browsing all concepts found.
On the right side, you can browser the found articles. The articles are shown with title, authors, journal, abstract and affiliation, also Wikipedia links and links to proteins identified from the text are offered if available. In the picture shown here is the abbreviated version for articles for faster browsing. You may switch between the full and short variant with the provided buttons. On top of the articles there is a summary with details for the query. This may include a link to dynamic Hot Topics and if your query matched an ontological concept a link to the corresponding term Hot Topics. There also links to export the results to citation mangers.
After selecting a term from the left side, here "Amyloid beta-Protein", the result view is updated. It shows now the articles containing the selected concept. This includes also all child terms of the selected term. Please remark that the initial result set size of 1000 articles was reduced down to 50 relevant articles in two clicks. In the summary field additionally there are now the term description and term synonyms listed. In case of "Amyloid beta-Protein" there are currently 10 synonyms listed. To select an interessiting article into the buildin clipboard use the paper clip icon provided directly next to the each article. To export a single articled you can use the export icon. To view the content of your clipboard select Clipboard link in the top bar. There you can also find the link to Hot Topics, Advanced Search, Help and a contact form.

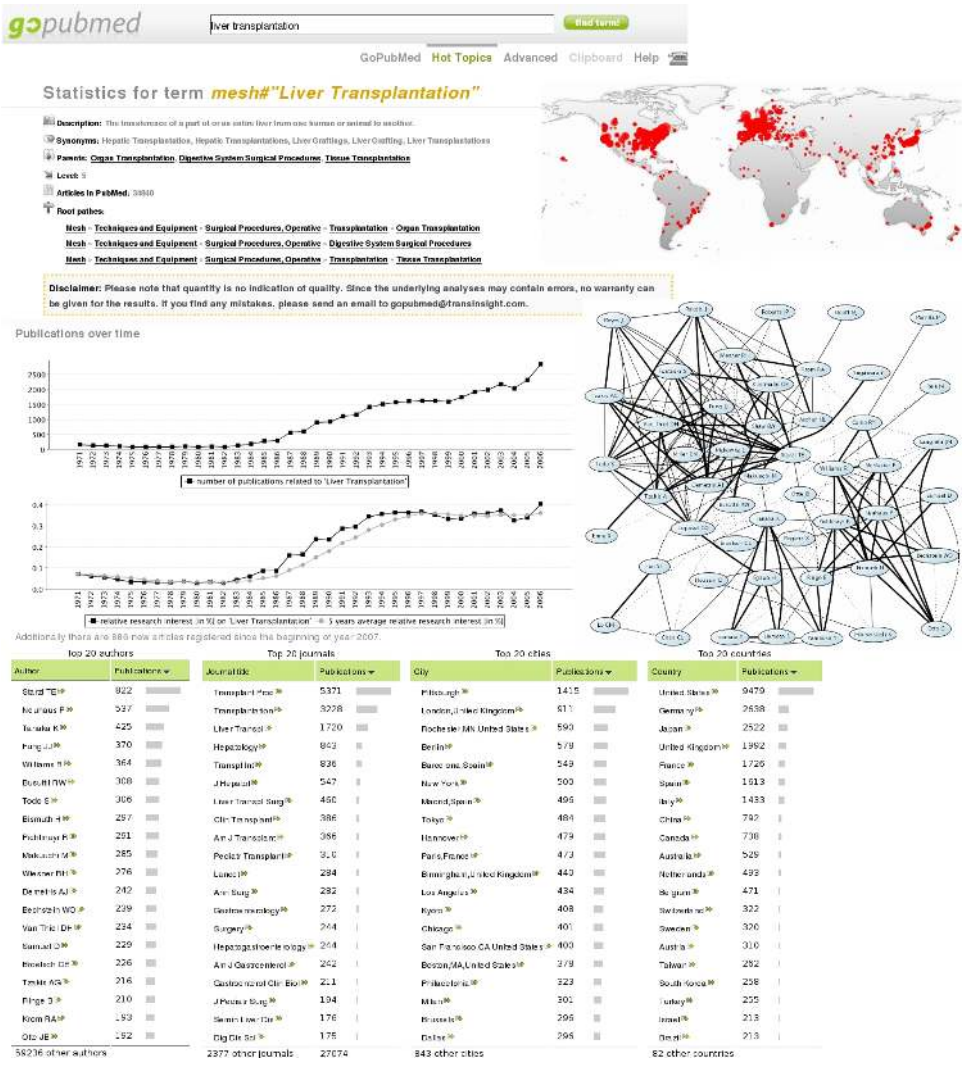

Figure 2: **Hot topics for "Liver Transplantation"**
The result page for a query to Hot Topics starts with a summary for the selected concept including a description, synonyms or the number of all publications in PubMed. Under Publication over time you can find two graphs. The first graph displays the number of publications related to this term per year. The second graph visualizes the fraction of publications on the topic over the total number of publications in that year. For "Liver Transplantation" the first graph displays growing number of publication, but the second graph denotes over the last years stagnation in comparison to the overall publication growth in PubMed. The top authors, journals, cities, and countries are presented as tables. All table entries are links and retrieve the related articles. If you would click on "Neuhaus P" you can retrieve all the publications which have an author with this name. The coauthor graph shows which author did publish together with whom. The more thick a line, the more articles contain their names as co-authors. The world map shows the regional distribution of the articles.

2 MousePubMed: Searching Biomedical Literature with Anatomy Ontologies

2.1 Introduction

Ontologies and vocabularies such as the Gene Ontology [Ashburner et al., 2000], UMLS [Bodenreider, 2004], Mesh[1], OBO[2], Snomed[3] GALEN[4] are widely used for annotation of biomedical data. They typically contain thousands of terms and cover broad subject areas of biomedical research. Additionally, many species specific vocabularies for anatomy have been designed covering among others plant [Jaiswal et al., 2005], C. elegans [Altun and Hall, 2006], drosophila [Grumbling and Strelets, 2006], mouse [Baldock et al., 2003, Bard et al., 1998], and human [Rosse and Mejino, 2003] anatomy. These vocabularies are used to facilitate communication between scientists in different communities and inter-operability between databases. Annotators, who are usually human, assign terms from such terminologies for example to genes. These assignments are ideally based on direct evidence from literature. Therefore, it is an important problem to automatically identify terms from ontologies in literature to support and even partly automate the annotation process. However, if terms from ontologies can be found in text, then ontologies can serve directly in literature search. Recently, a number of such knowledge-based search engines have emerged such as for example Textpresso [Mller et al., 2004], XplorMed [Perez-Iratxeta et al., 2003], and GoPubMed [Doms and Schroeder, 2005]. The ontological background knowledge can serve to answer questions with such tools. Consider for example a researcher interested in the Pax6 gene. He/she might have the following questions:

- Which processes is Pax6 involved in?
- Which diseases is Pax6 involved in?
- At which developmental stages is Pax6 active in mice?

Literature holds answers to these questions, but a classical literature search cannot answer the questions directly, as articles will not mention terms like gene, disease or process, but rather specific instances such as Pax6, Aniridia, or eye development. Since ontologies contain knowledge that e.g. Pax6 is a gene, Aniridia is a disease, and eye development is a process, they can help to answer questions.

Here, we will show how ontology-based literature search with GoPubMed can answer questions as the ones above. To accommodate the specifics of anatomy we will also discuss the use of specialised background knowledge. In particular, we will devise an algorithm and a system, called MousePubMed, to work with genes, tissues, and developmental stages as used in the Edinburgh Mouse Atlas [Baldock et al., 2003]. We evaluate MousePubMed's automated annotation on PubMed abstracts with the handcurated annotations of the Edinburgh Mouse Atlas. Before we go into the details of ontology-based literature search we will discuss the general problem of identifying ontology terms in text with a specific emphasis on anatomical and developmental terminology.

[1] nlm.nih.gov/mesh
[2] obo.sourceforge.net
[3] snomed.org
[4] opengalen.org

2.2 Databases and text mining

Curating Databases The large amount of species-specific databases today helps researchers to easily access various kinds of information on many organisms. Most such databases are manually curated by domain experts and constantly improved in terms of quantity and quality with input from the respective research communities. This manual curation process guarantees high quality and degree of reliability of the data. Annotations, for instance of genes and gene products, are stored in structured manners (associated functions, phenotypes, etc.), so that they can easily be queried by a researcher. Controlled vocabularies and ontologies designed for specific types of annotations reduce the amount of ambiguity for both curation and later access.

Database curators constantly scan the relevant literature to find evidence for new annotations related to their domain. These annotations are standardised terms from controlled vocabularies, often referred to as ontologies. For genes and gene products, annotations reflecting functions, locations, and processes are sought [Ashburner et al., 2000]. For drugs, it is interesting to find known digestive pathways and respective (desired and undesired) targets. Such facts often are reported in the literature, spread over a large variety of journals and other publication formats.

Ontologies as semantic frameworks for cross-database queries Efforts are under way to design ontologies suited not only for a single species, but rather a range of organisms. Some of these ontologies have already reached advanced stages and are widely used for annotations by many databases. One example is the Gene Ontology (GO), a hierarchy of concepts related to biological processes, molecular functions, and cellular components of genes and gene products. Many of the databases curating data on genes and proteins use GO for their annotations such as UniProt and EntrezGene. Another example is the Plant Ontology, a controlled vocabulary reflecting plant structures and developmental stages [Jaiswal et al., 2005]. It is used by the TAIR, Gramene, MaizeGDB, and other databases [Berardini et al., 2004, Jaiswal et al., 2006, Vincent et al., 2003]. The use of such common ontologies that are applicable to disparate databases, which may be species-centred like SGD or gene-centred like EntrezGene, alleviates cross-database queries. An example is a query across multiple species to find similarly annotated genes, possibly restricted to a common type of tissue. The proper design of exhaustive ontologies and/or controlled vocabularies to annotate, for instance, genes and gene products with structures, functions, processes, stages, or phenotypes, and their installment in relevant databases present major tasks towards facilitating comprehensive annotations and queries.

Databases vs. literature Queries across disparate databases are quite useful. However, a lot of data are not yet stored in such a structured form. This is due to two main reasons. For one, there is no immediate interest for researchers to submit their findings to (one or more) relevant databases, as scientific publications function as the main instrument for making information accessible and gaining reputation. The second reason comes with the necessary process of manual curation of database entries and annotation to maintain a certain quality standard. Another resource of data are aforementioned scientific publications themselves. Fairly often, these provide insight into more recent findings than databases. In addition, more background information, descriptions of experimental settings, etc. can be found in texts, showing broader context as well as in-depth details. Natural language often is more suitable to express facts than the structured form of any database. Moreover, many annotations in databases come in the

form of free text, for instance functions and diseases in UniProt, or phenotypes in MGI. This shows that scientific publications and other textual descriptions present important resources to be considered when searching for certain information. So, how can ontological terms be found in text?

Text mining In biomedical text mining, researchers use techniques from natural language processing, information retrieval, and machine learning to extract desired information from text [Jensen et al., 2006]. Even when the concepts to extract are available in a structured form, such as a controlled vocabulary or ontology, finding them in free text is not always an easy task. For instance, a recent assessment for extracting Gene Ontology terms revealed performances around 20% success rate only [Ehrler et al., 2005]. The difficulty of automating manual annotation is evident from the fact that only as few as 15% of manually annotated terms appear literally in the associated abstracts.

Ad-hoc variations of names To begin with, terms in vocabularies and labels of concepts in ontologies appear in many, slight or severe, variations in natural language texts.

- orthographic: IFN gamma, Ifn-γ
- morphological: Fas ligand, Fas ligands
- lexical: hepatitic leukaemia, liver leukemia
- structural: cancer in humans, human cancers
- acronyms/abbreviations: MS, Nf2
- synonyms: neoplasm, tumor, cancer, carcinoma
- paragrammatical phenomena/typographical errors: cerevisae, nucleotide

Some of the terms encountered in texts are rather ad-hoc creations, which cannot be found in any term lists.

Synonymity of ontological terms As mentioned before, terms in a vocabulary or ontology might not appear literally in a text, but authors use synonyms for the same concept. First of all, this complicates proper searches: When searching for "digestive vacuole", results should also contain texts that mention "phagolysosome"; mentionings of "ligand" refer to the concept "binding"; an "entry into host" might occur as an "invasion of host". In the Plant ontology for example, many synonyms exist for the same structure in different species. "Inflorescence" is referred to as "panicle" in rice, and as "cob" in sorghum, and "spike" in wheat, for instance. We note that there are also intra-ontology synonymities: "eye" in AnoBase can refer to the eye spot or the adult compound eye. In a similar manner, the Edinburgh Mouse Atlas contains unspecific mentions such as "cavity" or "body" for the mouse.

Table 3: Some anatomical terms that have other meanings in different domains. Some misinterpretations occur only when certain spelling variations are allowed, for instance, ignored capitalisation or plural forms.

Term	Other meaning
rod	common English
iris	species: plant; common English
axis	species: deer; common English
chin	common English
beak	common English
pons	protein: Serum paraoxonase/arylesterase 1 (PON)
penis	protein: Penicillinase repressor (penI)
sigma	common English
patella	species: limpet
cicatrix	disease: scar
nephrons	drug: bronchodilator (Nephron)
hemocytes	drug: iron supplement (Hemocyte)
chondrocytes	drug: cartilage cells for implantation
hippocampus	species: seahorse

Ambiguity of ontological terms Terms can have a very specific meaning in biomedical research, but mean other things in other contexts. Examples are "development", "envelope", "spindle", "transport", and "host". Protein names such as "Ken and Barbie", "multiple sclerosis" or "the" that resemble common names, diseases, or common English words are especially hard to disambiguate. The same problems arise from drug names like "Trial" or "Act". Table 3 lists some anatomical terms that have other meanings in different domains. Especially where cross-ontology or cross-database queries are needed, one has to consider ambiguity, for instance when applied to different organisms: "gametogenesis" (sexual reproduction) in plants is different from "gametogenesis" in metazoans.

Stemming and missing words Some aspects for finding terms in text refer to the actual processing of natural language and appear rather technical. Very often, words will appear in different forms, such as "binding" and "binds". These refer to the same concept, which can be solved by resolving words to their stem ("bind"). However, the analogous reduction of "dimerisation" to "dimer" is more questionable. The former talks about the process, the latter about the result. A similar example is "organisation", where a transformation into "organ" is invalid.

Texts contain additional words that are missing in the ontological term. This happens, for instance, when a text contains further explanations that describe findings in more detail. An example is "tyrosine phosphorylation of a recently identified STAT family member" that should match the ontology term "tyrosine phosphorylation of STAT protein." In general, matching is allowed to ignore words such as "of", "a", "that", "activity", but obviously not "STAT". Additional background information on term variations is needed to know that a "family member" can refer to a protein.

Formatting of terms represents another source for potential matching errors. Terms in

an ontology contain commas, dashes, brackets, etc., which require special treatment. For "thioredoxin-disulfide" the dash can be dropped, for "hydrolase activity, acting on ester bonds" the clause after the comma is important, but unlikely to appear as such in text. Terms containing additions such as "(sensu Insecta)" contain important contextual information, but are also less likely to appear in text.

2.3 Ontologies and text mining

Three main key dimensions of ontologies have been defined by Uschold: formality, purpose, and subject matter [Uschold, 1996]. The degree of formality by which a vocabulary is created and meaning is specified varies among different ontologies. The purpose refers to the indented use of an ontology. Domain ontologies (such as medicine or anatomy), problem solving ontologies, and representation ontologies comprise examples for different subject matters an ontology is characterising.

In contrast to ontologies designed primarily for annotating biological objects, there is a clear distinction to ontologies designed for text mining. We will describe this distinction and its impact on text mining strategies as well as on the redesign of dedicated ontologies. In the case of a text mining ontology, there must be some compromises on the relationships and on the labels used. The ontology should be easily used by the search engine in order to locate in the text important ontological terms/parts but also be easily used and edited by the domain experts. Therefore, it must not be very formal in terms of containing many different relationships between terms (such as 'derives from', 'causes', 'part of', etc.) or of distinguishing between 'classes' and 'instances'. It should rather be a structured vocabulary containing only *'is_a'* relationships between terms (only child-parent relationships). In this way, any search engine can identify each class in the biomedical literature. In general, there has to be a compromise to obtain a correct ontology with valid relations and still get the best possible results from text mining. The most prominent topics considering ontologies designed for text mining are the following.

- *Term overlaps* — some concepts can overlap in their labels or synonyms: in many cases there is a difference between what authors write and what they actually mean to express. Unfortunately, researchers do not have strict and formal ontologies or nomenclatures in their minds when composing a scientific article; in most of the cases they might use parent terms to refer to a child term, or vice-versa. For example, many people are treating the MeSH terms 'cardiovascular disease' and 'coronary artery disease (CHD, CAD)' the same, although the latter is a child of the first.

- *Descriptive labels* — in most of the cases, the labels in an annotation ontology cannot be used for text mining, usually due to their explanatory nature. For example, it is unlikely that the Gene Ontology term "cell wall (sensu Gram-negative bacteria)" will appear as such in text. Terms like "positive regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism" and "dosage compensation, by inactivation of X chromosome" are almost complete sentences and are also unlikely to be found as such in text.

- *Ambiguity* — results either from identical abbreviations for different terms, or, in general, tokens that can refer to terms that might or may not be of our interest. An example of an ambiguous *abbreviation* is "CAM" that can stand for "constitutively active mutants" , "cell adhesion molecule" , or "complementary alternative medicine" . The second category of ambiguities — and the most difficult to handle — is that of terms that (in the context

of anatomy) can refer to different species. An example of such ambiguities is "embryo", which can be a chicken, mouse, human, or even zebrafish embryo. Therefore, if we are interested in the different developmental stages of the mouse embryo nervous system, we need to retrieve articles focusing on studies on mouse embryos and not on any other study organism. If the term "embryo" is inserted in the Mouse Anatomy ontology as such, then the search engine will return articles on all kinds of embryos. If the term "mouse embryo" is inserted in the ontology, the number of articles retrieved will not be the real number of articles mentioning the term "mouse embryo", since not all of them will mention the term as such. A similar example is that of organs/tissues common to different species, such as "eye" or "lens".

- *Generic and specific labels* — when using the ontology for text mining in a specific biomedical sub-domain (anatomy, disease, glucose metabolism, etc.), the ontological concepts must be specific for that domain. The articles retrieved must be anatomy-specific or disease-specific or glucose-metabolism-specific. Therefore, we need a vocabulary specific enough to distinguish between relevant and irrelevant articles, but general enough to not exclude potentially relevant articles. If the concepts are too generic, they could be referring to many other domains. For example, during the design of a glucose-metabolism ontology, we might need to include information on kinetics. "Kinetics" as such is too generic to be used as a term, as it can refer to different kinds of kinetics (kinetics of phase transition, hydrolysis kinetics, kinetics of equilibrium reactions). On the other hand, the term "glucose kinetics" might be too specific, as it might seldom appear as such in a text. The decision on which terms should be used in the ontology ideally should only be made after exhaustive searches with different variations of terms.

We can derive some simple rules from all these observations, which can be used for (re-)design of ontologies when they should serve as resources for text mining applications.

- Avoid descriptive labels and synonyms: they should be likely to appear in texts as such – avoid "and", "of" and the like;
- Avoid improper spelling variations: capitalisation, noun plural forms, verb flexions;
- Use common names as labels or include them as synonyms;
- Add structural and lexical variations wherever possible;
- Keep the nomenclature consistent, precede terms with superstructure name;
- Use different representations of a concept in the ontology.

For a proper extraction of terms and subsequent term disambiguation in case of homonyms, the occurrence of parents helps to decide on the exact term. As, especially in anatomical ontologies, terms can have multiple representations, such multiple hierarchies should also reflected by the ontology. Examples are spatial and systemic representations of a tissue — "lung" is a "body part", and also a specific "organ system". Depending on the context in which "brain" is found, parent terms below "head" might not be found in the text at all, but rather terms related to "organ system." An ontology should therefore cover at least the most likely paths to subsume a tissue.

All of the above problems mean that extracting terms from literature will not be error-free. However, despite all of these problems, ontology-based literature with text mining can answer

questions as posed in the introduction. Next, we introduce three such engines, GoPubMed, MeshPubMed, and MousePubMed and illustrate how they answer questions.

2.4 GoPubMed and MeshPubMed

GoPubMed [Doms and Schroeder, 2005], MeshPubMed and MousePubMed, which is discussed in the next section, index articles provided by PubMed with ontology terms from GO, Mesh, and Mouse anatomy/development respectively. As an example consider Fig. 3, which shows a screenshot of MeshPubMed when queried for Pax6. The key difference to a classical search are the relevant ontology terms on the left. They list frequently occurring terms and the complete hierarchy of relevant terms found in any document mentioning the given keywords. Clicking on any of these terms reduces the result set and allows users to quickly filter large result sets to the necessary documents needed to answer their question.

Let us consider the three questions about Pax6 from the introduction:

- Which processes is Pax6 involved in? A query in GoPubMed for Pax6 shows that the most frequent process mentioned is development. Opening the development branch furthermore reveals the processes of brain and eye development as well as organ morphogenesis including pancreas development. And indeed the corresponding articles support this essential role of Pax6 as transcription factor and master control gene in development of eye, brain and pancreas [Kleinjan et al., 2006].

- Which diseases is Pax6 involved in? A query in MeshPubMed for Pax6 shows that the most frequent disease mentioned is aniridia. Hovering the mouse over the term gives an explanation that it is "a congenital abnormality in which there is only a rudimentary iris. This is due to the failure of the optic cup to grow. Aniridia also occurs in a hereditary form, usually autosomal dominant." A click on aniridia shows articles mentioning both the disease and the gene such as for example [Brinckmann et al., 2006], which confirm the answer.

- At which developmental stages is Pax6 active in mice? A query in MousePubMed for Pax6 shows that Theiler stages up to 14 (9 dpc, days post conception) are frequently mentioned supporting Pax6 role in early development. Clicking on a stage reveals e.g. the statement "In the early development of the vertebrate eye, Pax6 is required for..." in [Azuma et al., 2005]

Indeed, Pax6 is the most researched gene of the family of Pax genes and appears throughout the literature as a 'master control' gene for the development of eyes and is of medical importance because heterozygous mutants produce a wide spectrum of ocular defects such as aniridia in humans. We can now further check in MeshPubMed whether aniridia is a 'hot topic' and who the most active authors publishing on aniridia are. Consider Fig. 4. It turns out that V. van Heyningen is the number one publishing author having the most collaborations, especially together with A. Seawright, as shown on the co-authorship network in Fig. 4.

2.5 MousePubMed

To use ontology-based literature search for developmental biology, we built MousePubMed using vocabularies for mouse anatomy (EMAP), human anatomy (EHDA), mouse genes

Figure 3: MeshPubMed query for "Pax6". On the left, the five frequent terms, frequent terms by category and all relevant terms are shown. The most frequently mentioned disease is aniridia. Clicking the term and retrieving the articles mentioning aniridia confirms that Pax6 is involved in aniridia

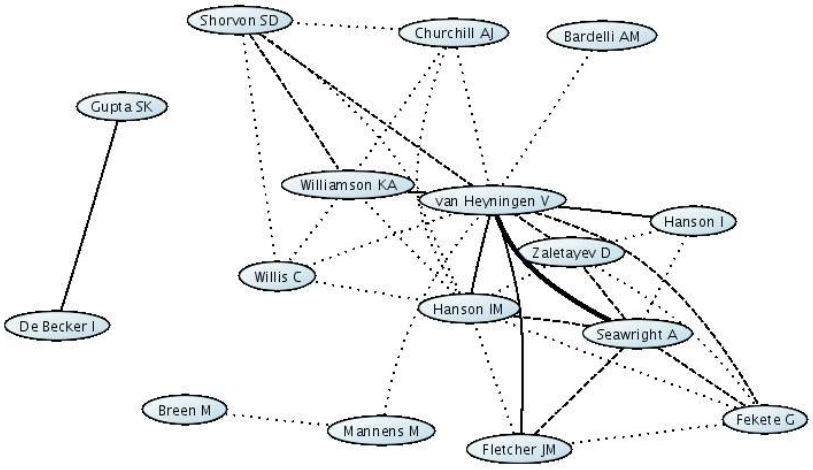

Figure 4: Part of the co-authorship network for "aniridia" in MeshPubMed showing V. van Heyningen and A. Seawright as the authors active in this area.

(from EMAP), and mouse developmental stages (Theiler) as resources. To demonstrate MousePubMed's usefulness, we evaluate it against tissue and developmental stage annotations in the edinburgh Mouse Atlas. Before we discuss this evaluation, we introduce the matching algorithm developed.

2.5.1 Extracting gene names, anatomy terms and developmental stages

Ontology based text mining is not restricted to finding words or word groups in texts. The structure of the ontology can be used to state the relation between a term and a document by finding the children of the term. This task is reasonably well solvable for the Gene Ontology where its term labels are self descriptive. Many terms in GO are contained in their children terms [Ogren et al., 2004]. As an example, the term "envelope" is refined into "organelle envelope" and further to "organelle envelope lumen". The ontology for the Abstract Mouse contains anatomical concepts in the mouse embryo at different embryonic developmental stages. The vocabulary is used to annotate images of mouse embryos. It unifies the vocabulary needed to describe the different parts throughout 26 Theiler stages. Concepts like organs or body parts are further refined into tissue types, unspecific loci such as "cavities", "left", "upper", as well as general terms such as "node" or "skin". Ontologically spoken the Abstract Mouse ontology defines specific individuals rather than general classes. It contains taxonomic relationships between specific anatomical entities which cannot be generalised at the level of classes. For example, "chorion" has the children "mesoderm", "ectoderm" and "mesenchyme". "Amnion" and "yolk sac" have children sharing the same labels. Searching for documents related to "chorion" will retrieve very similar document sets to searching for "amnion", only because the documents mention "mesoderm", in this case with meaning "mesoderm specific to amnion". Different anatomical concepts share the same term label. For instance, there exist 171 individuals with label "epithelium". These all refer to different body parts at a specific stage in development.

Ontology based text mining relies on the assumption that unique or similar types of directed non-cyclic relationships exist which can be unified in the hierarchical relationships creating a taxonomy. For the Abstract Mouse ontology this assumption does not hold. There does not always exist a path to the common root supported by only one type of hierarchical relationships. Therefor in our analysis, a document is annotated with a term from the Abstract Mouse ontology, only taking the term label and its synonymous labels into account. In the Abstract Mouse Ontology the term labels follow various creation patterns. Sometimes a child term contains information of the parent term (for example, "cavities" has the child "amniotic cavity"). In other cases a term like "umbilical vein" has the children "left" and "right", rather than "left umbilical vein" and "right umbilical vein", respectively. These short and common sense labels make the text annotations arbitrary.

For our experiments we slightly adapted the ontology. For the terms "left", "right", "upper", "lower", "common", "anterior" and "posterior" we expanded the term labels with its parents labels. "Eyelids" thus became "upper eyelids" and "lower eyelids", for instance, and we removed the children terms "upper" and "lower" accordingly. To distinguish between common terms such as "skin" occurring — for instance, for different organs — the matching algorithm took text annotations for ancestor terms into account. Terms with the same label were grouped according to the number of text annotations for their ancestors in the same document. Only annotations of the top ranked group were confirmed. Figure 5 shows an example for the term "skin". There were multiple possibilities to resolve this term to a specific tissue. Only when a parental term (shoulder, upper arm, etc.) was found, the text was annotated with the specific

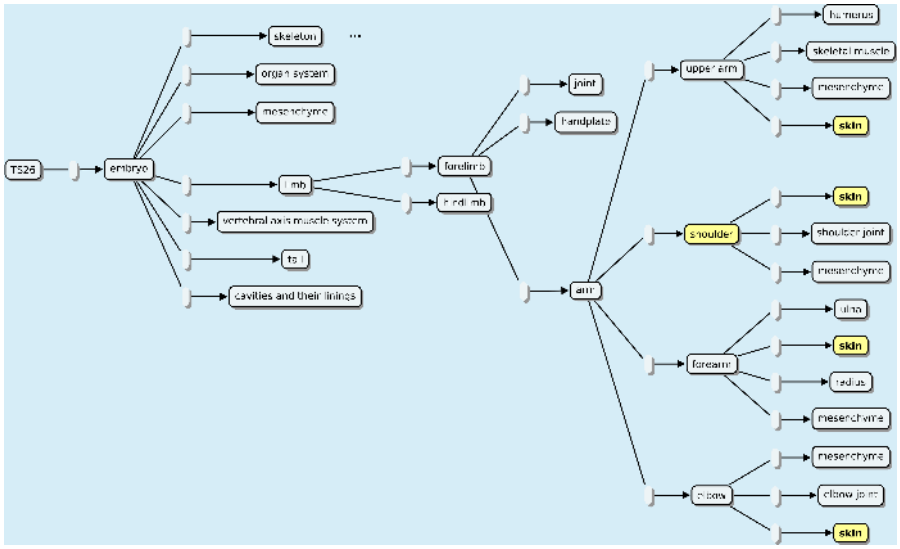

Figure 5: Excerpt from the anatomy ontology, for different types of skin. Occurrences of the term "skin" (yellow concept nodes) in a text were resolved using the hierarchical dependencies. Only when a parental node was also found, for instance, "shoulder", we annotated the text with "skin."

skin.

Finding gene names in documents is done using exact matching against gene names contained in EMAP. We enriched this set using additional names and synonyms for each gene taken from the MGI database[5]. We tested all 1437 genes mentioned in EMAP for their annotations with tissues and Theiler stages in PubMed.

We analysed 123,074 abstracts retrieved from PubMed with the query "mouse AND development". This amounted for approximately 0.7% of all documents listed in PubMed. Based on the document annotations with ontology terms we issued in total 36,358 statements on relations between genes, tissue and developmental stages, which we extracted from EMAP. Cases with multiple Theiler stages from EMAP were split into separate statements. We evaluated the tissues mentioned using EMAP's Abstract Mouse ontology and the anatomy part or MeSH. For path descriptions like "embryo.ectoderm" in EMAP we required the matching document to be annotated with the terms "embryo" and "ectoderm". For MeSH, as in MeshPubMed, we also included descending terms. A document was annotated with the term "embryo" if annotations for its descendants, for example, "germ layers" or its children "ectoderm", "endoderm" or "mesoderm" were found.

To find mentions of Theiler stages in texts, it was not enough to search for them directly, as they seldom occur as such in abstracts ("Theiler stage 12", "TS12", etc.). We therefore compiled a set of regular expressions based on two main notions, the mentioning of embryonic days (E) and of days post coitum (dpc). These expression had to capture occurrences like

- "embryonic day 10.5",

[5]See http://www.informatics.jax.org

- "day 9 mouse embryos",
- "between E3.5 (E = embryonic day) and E8.5",
- "12.5 days post coitum", and also
- "7.5-13.5 days post-conception."

As mentionings of Theiler stages do not often occur, but rather general time spans are given ("early embryonic development"), we decided to assign Theiler stages one to 14 to "early development", and stages 20 to 27 to "late development," respectively. Every mention of an "early developmental stage" thus was treated as a match for stages one through 14. Both assignment were based on statements found in PubMed relating days to general time spans.

2.5.2 Experiment designs

To assess the potential of ontology-based literature searches, we designed to experimental scenarios. For the first, we manually collected two sets of queries and detailed answers. For the second scenario, we evaluated the complete EMAP data. Using the methodology described in the previous section, we tried to find textual evidences for all sets in PubMed. This means that we searched PubMed for abstracts that shared annotations for each collected triple consisting of a gene, tissue, and Theiler stage.

Manually curated test set We first selected set of questions manually to study results in detail. The idea was to send simple keyword queries to MousePubMed, asking for mouse abstracts that discuss a certain tissue and embryonic day. MousePubMed should then identify all genes mentioned in the top-ranking abstracts. Questions and retrieved answers were as follows.

- Which genes play a role in the development of the nervous system in Theiler stage 14? A query for "mouse development nervous system 9 dpc" finds the genes Adamts9, Hoxb4, Otx3, and EphA4 within the first eight abstracts[6]. In addition, the genes EphA2, A3, A7, B1, B2, and B4 are found, which are not yet annotated in the EMAP database.
- Which genes play a role in sex differentiation during murine embryo development? A corresponding query for "mouse sex 10 dpc" results in a set of eight genes within the first fifteen abstracts: Fgf9, Asx11, Sry, Sox9, Usp9x, Maestro/Mro, Wt1, Amh1 and Fra1[7]. Only half of the genes can be found in EMAP so far.
- Which genes play a role in the development of the murine embryonic liver? A query for "mouse 'liver development'" results in a set of several genes, most of which can be found in EMAP as well: Shc, Pxn, Grb2, PEST/Pcnp, GATA6, HNF4a, Foxa1/2, Zhx2, HNF6, Mtf1, SEK1, Nfkb1, c-Jun, Itih-4, and Hex. To answer this question exactly, however, too few abstracts mention particular Theiler stages or days post congestion. They rather refer to "early stages of development", and the exact time span might be presented in the full text article only.

[6]Important for answering this query are returned PubMedIDs 12736215, 12055180, 11403717.
[7]Important are PubMedIDs 16540514, 16412590, 14978045, 14684990, 14516667, 12889070, 9879712, 9115712.

All the results, in particular where genes and exact Theiler stages are concerned, are highly dependent on the ordering of abstracts as provided by PubMed. Whenever a new publication appears containing the same search keywords, it will displace abstracts potentially more informative regarding the original question. Abstracts answering the original question might not appear among the first few and be immediately present to the user. However, text mining methods will still extract all the data, even from older publications, and still the right set of articles can easily be found.

The abstracts resulting from a keyword search occur in the same ordering as provided by PubMed. That is, in general, the most recent articles occur first. However, querying for species, tissues, and stages still returns the abstracts that discuss the interesting genes. Although corresponding expression patterns might first have been described in older publications, even in recent publications the desired genes reappear quite often.

Reconstructing outcomes of large-scale screening Thut et al. provided a list of 62 genes found expressed during eye development in mice, together with developmental stage and substructure [Thut et al., 2001]. Of the 62 genes, 26 were not previously reported (as of 2001); to 16 genes, novel valuable information could be added; 20 genes were fully reported before. Expression patterns were summarised for E12.5, E13.5, E14.5, E16.5, E18.5 and P2. Using MousePubMed, we tried to reconstruct the result of this large-scale screen of 1000 genes.

As Table 4 shows, nine PubMed abstracts contained the full information as stated by Thut et al., mentioning gene, tissue, and specific stages (days). For most cases, however, not all data were contained in one single abstract. In three cases, we were not able to automatically spot the gene name (left column), in all cases this was due to synonyms lacking in EMAP and MGI. Note that the assessment of recognising genes was based only on genes mentioned in EMAP. The tissue could be found in almost all of the cases; from most abstracts, even the specific part of the eye could be extracted.

Complete EMAP test set To evaluate capabilities of automated searches against the complete EMAP data, the experimental setting was as follows. Genes in EMAP have annotated tissues, in which they were detected at various stages of embryo development. Thus, we queried MousePubMed with each gene and checked which tissues were mentioned in the resulting PubMed abstracts. This was based on co-occurrence of the gene considering, a tissue, and a Theiler stage (day) in the same abstract. Currently, there are 1437 genes in the EMAP database annotated with (sometimes multiple) tissues and stages. All in all, we identified 18,179 such triples — gene, tissue, and stage — in EMAP. Many of the annotations consist of general annotations for tissue, like "mouse", "embryo", "left", "female", "node". We removed such trivial instances, because they would very frequently found. 12,782 triples referred to specific tissues, and we tried to find these triples using the aforementioned term extraction (also see Table 5).

As Table 6 shows, we were able to reconstruct 31% of the gene-tissue associations in EMAP using PubMed abstracts. Only 13% of the full information (gene, tissue, exact stage) was contained in abstracts. All in all, the data recovered from PubMed included information on about 37% of the EMAP genes. We noted that in many cases, abstracts do not mention specific time points during development. Sometimes, "early" and "late development" are mentioned, which we resolved as described previously in this section. On the other hand, mentions like "in early liver development" could not be resolved to specific overall-stages without background

Table 4: Expression patterns identified by MousePubMed in articles derived from [Thut et al., 2001]. Often, an abstract does not mention a (specific) developmental stage; —: MousePubMed did not find this particular fact; otherwise: facts as identified by MousePubMed. Given are only tissues related to the murine eye.

Gene	Tissue	Stage	PubMedID
Sparc	retina, RPE, eye	E4.5, E5, E10, E14, E17	9367648
Sparc	lens	embryonic day (E)14	16303962
Stat3	retina, RPE, eye	-no specific stage-	12634107
Stat3	lens	E10.5	14978477
Pedf	RPE	-no specific stage-	7623128
Pedf	retina	E14.5, 18.5	12447163
Runx1	inner retina	embryonic day 13.5	16026391
Col15a1	conjunctiva, cornea	E10.5-18.5	14752666
Otx2	outer retina	-no specific stage-	15978261
Edn1	retina	-no stage-	11413193
IGF-II	eye, cornea, retina, scleral cells	E14	2560708
Wnt7b	anterior eye, cornea, optic cup, iris	-no specific stage-	16258938
CDH2	—	-no stage-	9210582
—	lens	-no stage-	9211469
Col9a1	eye, lens vesicle, neural retina, ciliary epithelial cells, cornea	13.5, 16.5-18.5 d.p.c.	8305707
Tgfb2	cornea, lens, stroma	-no specific stage-	11784073
Thra	retina	-no specific stage-	9412494
BMP4	retina	E5	17050724
Bmp4	optic vesicle, lens	-no specific stage-	15558471
BMP4	lens, optic vesicle	-no specific stage-	9851982
—	eyes	N/A	15902435
Sox1/2	lens	-no stage-	15902435
—	retina, eye axis	E2, E3, E5	15113840
Notch1	eye	-no specific stage-	11731257
Notch2	eye	-no specific stage-	11171333

Table 5: Types of information and quantity contained in EMAP.

Type of information	Amount of data
Genes with tissues, stages	1437
Genes with at least one non-trivial tissue, stages	1346
Triples of gene, tissue, stage	18,179
Triples of gene, non-trivial tissue, stage	12,782
Tuples of gene, non-trivial tissue	8653

Table 6: Number of tuples/triples consisting of gene and tissue or gene, tissue and stage found in PubMed abstracts retrieved by the query "mouse AND development."

Type of information	Amount of data
Triples of gene, non-trivial tissue, stage	1637 (12.8%)
Tuples of gene, non-trivial tissue	2667 (30.8%)
Genes with at least one tissue and stage	537 (37.4%)

information. Cross-checks revealed that indeed much of the necessary information was only mentioned in the full text of references annotated by EMAP for a specific association.

2.6 Conclusion

Ontologies are widely used for annotation. They are also useful for literature search, but the extraction of terms from text is a difficult problem due to the complexity of natural language. Here, we demonstrated the use of the ontology-based literature engines GoPubMed, MeshPubMed, and MousePubMed to answer questions in the context of development. We discussed the specific extraction algorithms needed for MousePubMed and evaluated them small scale on examples relating to eye development and large scale on gene-tissue-stage triple from the Edinburgh Mouse Atlas. We were able to reconstruct 37% of genes, 31% of gene-tissue associations and 13% of gene-tissue-stage associations from PubMed abstracts. These figures are encouraging as only abstracts are used.

3 Ontology design for text-mining: guidelines and automatic term recognition in the lipoprotein metabolism domain

3.1 Introduction

The engineering of ontologies is still a new research field. There does not yet exist a well-defined theory and technology for ontology construction. This means that many of the ontology design steps remain manual and a kind of "art" and intuition ([Soldatova and King, 2005]; [Sowa, 2000]; [Castro et al., 2006]). There exists a variety of different ontologies, constructed for different purposes and projects.

As far as the biomedical ontologies are concerned, during the last years there have been major efforts in the biological community for organizing biological concepts in the form of controlled terminologies or ontologies ([Cantor et al., 2005]; [Ashburner et al., 2000], [Evsikov et al., 2004]). There have also been developed tools to provide interoperability among different ontologies ([Bodenreider, 2004]; [Cantor et al., 2005]) in order to provide a common frame of reference among the different research communities. Examples of ontologies are the Gene Ontology ([Ashburner et al., 2000]) that provides a controlled vocabulary to describe gene and gene products in any organism, the Mouse Anatomy (MA) ([Evsikov et al., 2004]), the Cell Ontology (CL) ([Bard et al., 2005]) and SNOMED ([Spackman, 2004]). The Open Biological Ontologies (OBO)[8] consortium hosts over 50 open source ontologies associated with phenotypic and biomolecular information. Baker et al. ([Baker et al., 1999]) give an overview on biomedical ontologies.

Semantic meta-information provided in the form of ontologies has proven useful in order to search ([Doms and Schroeder, 2005]) or index large collections of documents (e.g. MeSH for indexing MEDLINE). Meta-information found for text documents is often general (keyword list) or still too complex for an automated evaluation (article abstract). Finding terms of controlled vocabularies in text overcomes this shortage, while relations between terms provide the necessary navigation structures.

Ontological background knowledge can serve to answer questions with knowledge-based search engines ([Mller et al., 2004]; [Perez-Iratxeta et al., 2003]; [Doms and Schroeder, 2005]). In the domain of lipoprotein metabolism, for example, a search for "analphalipoproteinemia" will retrieve articles for Tangier's disease, which is actually a synonym. In case of a syndrome, such as the "metabolic syndrome", in a properly designed ontology the articles retrieved will contain symptoms and other characteristics for it (e.g. type II diabetes, hypertension, insulin resistant, low HDL, hypertension, all of them being parts of the metabolic syndrome). Researchers explore literature on different parameters that can affect the lipoprotein metabolism, such as the phenotype, genotype and age of the patients/animals tested, environmental factors and lifestyle, specific lipoprotein and enzyme concentrations and others. Questions like:

- What is the activity of cholesterol ester transfer protein in diabetes?
- Which cells/tissues is apoE expressed in?
- What is the impact of a fish oil diet on metabolic syndrome individuals?
- Which genes/proteins/metabolites are hypertension-specific?

[8] http://obo.sourceforge.net/

can be answered with the use of a well designed ontology on lipoprotein metabolism, containing terminology found in literature with semantically interconnected terms.

The GoPubMed search engine ([Doms and Schroeder, 2005]) allows users to explore PubMed search results with the Gene Ontology (GO) ([Ashburner et al., 2000]) and Medical Subject Headings (MeSH, [Bodenreider, 2004]). GoPubMed/MeshPubMed retrieves PubMed abstracts for a search query, detects terms from the GO and MeSH in the abstracts, displays the subset of GO and MeSH relevant to the keywords and allows for browsing the ontologies and displaying only articles containing specific GO and MeSH terms. The search engine is developed in a way that any ontology (e.g. a Lipoprotein Metabolism Ontology) can be easily integrated and used for a domain-specific literature search. One of the benefits of such an ontology-based literature search is the categorization of abstracts according to a specific ontology, allowing users to quickly navigate through the abstracts by category and providing an overview of the literature. It can also automatically show general ontology terms related to the original query, which often do not even appear directly in the abstract.

In this paper, we introduce design principles for ontologies used for textmining. A key problem in this context is the generation of terms, which is corroborated by ([Castro et al., 2006]), who compared different ontology design methods and tools all of which lacked automated term recognition. The paper is organized as follows. We first introduce the design principles followed when designing the lipid metabolism ontology and turn to the question how to automate the generation of terms. We introduce two methods to identify terms and evaluate them together with two existing tools for this task.

3.2 Ontology design principles

Three main key dimensions of ontologies have been defined by Uschold: *formality*, *purpose* and *subject matter* ([Uschold, 1996]). The degree of formality describes whether the ontology is expressed loosely, in a more structured/restricted form of a natural language or in an artificial formally defined language. The purpose refers to the intended use of an ontology which can be *communication between people* (sharing a common structured vocabulary to understand each other and making decisions), *inter-operability among systems* (used as an interchange format to translate between different modelling methods, languages and software tools) and *systems engineering benefits* (re-usability, automation of consistency checking), etc. An ontology can be generic or specific, depending on its purpose and the level of detail it contains. The more generic, the more applicable it can be and the higher its re-usability. Finally, domain ontologies, problem solving ontologies and representation ontologies comprise examples for different subject matters an ontology is characterizing.

There are some basic steps that should be followed during the design of an ontology:

Identification of the range of intended users: people who will use the ontology for their research projects, others who will use it as part of another ontology, knowledge engineers, technical people, biomedical researchers, et al. The ontology must be easy for them to understand and use.

Decision on the purpose and main research area of the ontology: take time to think why you need this ontology and how sufficient is your knowledge on the specific domain.

Definition/prediction of further possible applications. Apart from its initial purpose, it could also serve a different one. For example, the Gene Ontology

(GO)[Ashburner et al., 2000] has also been used by the search engine GoPubMed[9] [Doms and Schroeder, 2005] and by GoMiner[10] for gene expression data evaluation.

Formulation of questions: construct motivating scenarios and define a complete set of competency questions to express different reasoning problems. Examples of such questions are: "What is the activity of enzyme A in disease B?", "Which cells/tissues is protein X expressed in?", "What is the substrate specificity of Z?", etc. Start with general questions to insert general terms first.

Reuse existing ontologies that may cover to some extent the ontology under design. Be aware that ontologies might differ in granularity, completeness, development stage and format. There exist some ontology development tools like OntoMerge [Dou et al., 2002], SAMBO [Lambrix and Tan, 2006] and others described by Duineveld et al., 1999 [Duineveld et al., 2000] and Lambrix and Edberg, 2003 [Lambrix and Edberg, 2003] that can merge parts from two different ontologies.

Literature scanning for deciding on the basic concepts: you might need to consult several corpora of knowledge. Do some brainstorming to come up with a list of relevant concepts and try to group them into semantically similar categories.

Definition of the relationships between concepts: they can be simple `is_a` and `part_of` relationships or even more complex ones. They can even be whole sentences, but at the end this will be on expense of re-usability and applicability. The simpler the relationships the easier to re-use the ontology.

item[Add a definition for each term:] either during editing or at the end, it is always important to keep the definitions in mind to come up with a well-structured, useful ontology. The definition must not be cyclic, meaning that it must not contain the term itself (e.g. GO:0050896 'response to stimulus' definition: A change in state or activity of a cell or an organism ... as a result of a stimulus, GO:0016788 'hydrolase activity, acting on ester bonds': Catalysis of the hydrolysis of any ester bond); it must be very clear to other users what the ontology term is about.

Decide on a label for each concept. This is one of the most crucial steps during the structuring of the ontology. This task is difficult for humans as it requires good knowledge of the domain of interest so as to group concepts on the hierarchy in a semantically meaningful way. It is even more difficult for machines to do this automatically. There has been previous work on automatic labeling of document clusters [Popescul and Ungar, 2000] by using the most frequent and most predictive words in clusters of documents, but there is still work to be done on that. One must firstly concentrate on the semantics of a term, decide what is really needed to be expressed with that term and then choose the appropriate name.

Last, but not least, and perhaps one of the very first steps of the designing procedure is the selection of a suitable *ontology editor*. We used Protégé[11] and CmapTools[12]. Ontology visualization is crucial when the knowledge engineer and the domain expert are two different persons and need to agree on the different versions of the ontology.

[9] http://gopubmed.org/
[10] http://discover.nci.nih.gov/gominer/
[11] http://protege.stanford.edu/
[12] http://cmap.ihmc.us/

3.2.1 Further guidelines for the design of a text-mining ontology (best practice)

With GO we experienced some limitations for text-mining. For example, it is unlikely that a descriptive label such as "cell wall (sensu Gram-negative bacteria)" will literally appear in text. A comprehensive overview of such problems is provided by Smith et al. [Smith et al., 2004]. There often exist ontology terms that are unlikely to appear as such in text but are rather of a structuring nature. For example, the terms "hydrolase activity, acting on ester bonds" (GO:0016788) or "hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds" (GO:0016810) include several different types of information: activity (hydrolase), type of bond affected (ester or carbon-nitrogen) and exception (but not peptide) (see Figure 6).

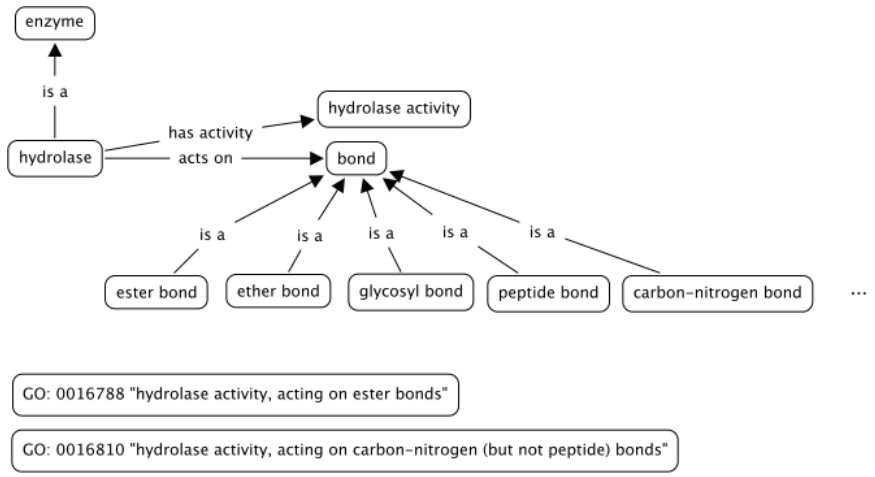

Figure 6: **Problematic terms – the hydrolase activity example.** Terms like hydrolase, hydrolase activity, bond, ester bond and relations between them (e.g. acts on) can be easily found in text, whereas full GO terms such as "hydrolase activity, acting on ester bonds" are unlikely to appear literally in an article.

These should be 3 different branches of the tree, combined with relations, therefore structuring "logical formulas". For example, in the case of the second term (GO:0016810), the exception could be expressed as a certain condition: the protein has a hydrolase activity and is acting on carbon nitrogen bonds, but not in all bonds (peptide bonds are excluded). Aranguren et al., 2007 [Aranguren et al., 2007] provide a simple and indicative example of the problem: a Person is a Man or a Woman, a Man has Testis, a Woman has no Testis, but what happens in the case of a Eunuch (who is actually a man without Testis)? There is a need for distinguishing between relations that are strict "always" rules and "normally" or "usually" relations that can also allow for exceptions. Biomedical terms are usually connected with "usually" relations between them. Another example is the definition of mammals: a simple definition[13] can be "warm-blooded vertebrate animals belonging to the class mammalia, including all that possess hair and suckle their young". Therefore, one can say that all mammals give birth to and suckle their young. But there exists the exception of the monotremes, which are mammals that lay eggs instead of

[13]taken from http://www.biology-online.org/dictionary/Mammals

bearing live young. The definition here would be "mammals are animals that normally bear live young and suckle them" and the exception "monotremes are mammals that lay eggs".

Compositional structure of terms is a major bottleneck for ontology design, especially when it comes to text mining, as the relations between terms must be as simple as possible. Ogren et al. [Ogren et al., 2004, Ogren et al., 2005] have performed an analysis of the term names in the GO to investigate substring relations between terms and revealed that 65.3% of all GO terms contain another GO term as a proper substring. These terms can be categorized into two groups: GO terms that contain other GO terms as proper substrings (e.g. "hydrolase activity, acting on acid sulfur-sulfur bonds" (GO: 0016828) and "hydrolase activity" (GO: 0016787)) and GO terms that contain strings that seem to recur frequently (e.g. "regulation of" in GO, "predominance of" in the Lipoprotein Metabolism Ontology).

In contrast to ontologies designed primarily for annotating biological entities, there is a clear distinction to ontologies designed for text mining: there have to be made some decisions and compromises on the relationships and on the labels defined.

Decisions that need to be made during the ontology design

Keep or dismiss a term When using the ontology for text-mining over a specific biomedical domain (e.g. disease, glucose metabolism, lipoprotein metabolism), the ontological concepts must be specific for that domain. The articles retrieved must be disease-specific or glucose-metabolism-specific or lipoprotein-metabolism-specific. For example, including information on 'kinetics' during the design of a "glucose metabolism ontology" is crucial. But 'kinetics' is too general as a term, as the distinction between different kinds of kinetics is important (e.g. when querying PubMed for 'kinetics', there are retrieved articles referring to 'kinetics of phenols' or a 'reconstruction kinetics well', irrelevant to the domain of interest). On the other hand, the term 'glucose kinetics' is too specific and documents mentioning it do not cover all essentials known in glucose kinetics. Searches for "glucose kinetics", "glucose" and "kinetics" and retrieval of relevant articles (e.g. PMID: 17003244 'In the current investigation we studied the effects of TZD treatment on insulin-stimulated fatty acid and glucose kinetics in ...') lead to the decision that the best term to use for 'glucose kinetics' is the exact term. There already exist previous efforts on automatic labeling of document clusters and identification of on-tology components, based on Natural Language Processing techniques or hierarchical and suffix-tree clustering [Stefanowski and Weiss, 2003, Lame, 2004, Popescul and Ungar, 2000].

Decide on ontology design/relations The ontology must not be very formal in terms of containing many different relationships between terms (such as 'derives from', 'causes', 'part of', etc.) or of distinguishing between 'classes' and 'instances'. It should rather be a structured vocabulary containing only child-parent relationships.

Decide on synonyms researchers do not have strict and formal ontologies or nomenclatures in their minds when composing a scientific article and therefore use terminology of differing granularity. They often use parent terms to refer to a child term, or vice-versa (e.g. 'coronary artery disease (CHD, CAD)' is child of 'cardiovascular disease', but in many cases authors are treating them the same). Again literature scanning, for both child and parent term, will help to clarify how researchers refer to different terms. Another problematic case is that of the different lipoprotein subclasses (based on particle size, buoyant density, composition, etc.)

where there do not exist clear limits between them. Depending on the way of measurement and the difference in surface lipid content, they can be expressed in different ways. For example, in the case of LDL, there are 5 different subclasses based on particle size (LDL I-V), but there are also references such as 'small dense LDL' or 'buoyant LDL' that are very often found in text but could contain a mixture of different subclasses. Since we need to keep only a simple hierarchy with parent-child relationships, we do not incorporate any "compositional" information (e.g. that 'small dense LDL' consists of a mixture of LDLIII and LDLIV). In these cases, we put the synonyms according to the authors' definitions, for example 'small dense LDL' as a synonym for LDL III and 'buoy-ant LDL' or 'large LDL' as synonyms for LDL I [Berneis and Rizzo, 2004]. A similar example from the GO is that of 'transporters and carriers'. In every day language 'transporter and carrier' is the same as 'transporters or carriers', but logically they are different.

Compromises that need to be made, problems, inconsistencies There must be made some compromises to retain a correct ontology (meaning that it contains valid relations) and still get the best possible results from text-mining:

Ambiguities resulting either from identical abbreviations for different terms (e.g. 'CAM' can stand for 'constitutively active mutants', 'cell adhesion molecule', or 'complementary alternative medicine'), or, incomplete term labels (e.g. 'embryo' can be referring to a chicken, mouse, or human embryo, 'male' can be referring to human patients or rats). For example, we are only interested in experiments performed in human patients and need to distinguish between human- and animal-referring articles. One option is to insert into the ontology only human-specific terms, such as 'ex-perimentee', 'patient', 'man', 'boy', etc. 'Male' cannot be in the ontology, since it could also be referring to animals. Another op-tion is to maintain a list of human- and animal- specific words or expressions and then transform the algorithm in a way that one could make a Boolean selection (e.g. AND human, NOT animal) in the query and finally include or exclude the results for the spe-cific selections.

Try to avoid any possible inconsistencies. To illustrate their 'reasoning' nature, let us describe the following example: a researcher needs to build a 'lactose metabolism' ontology and is interested in the tolerance of different ethnic groups to lactose. He needs to know the geographical location and the race/color of lactose tolerant/intolerant people, since they are both important factors affecting lactose intolerance. Combination of geographical information as well as racial information in one part of the ontology is, therefore, needed. Many articles refer to "African-Americans" as "blacks", so the term must be included under 'ethnic group'. Then the following must be valid: define 'Caucasian', 'African' and 'Asian' as 'ethnic group', 'American' is a 'Caucasian', 'African-American' is a 'African', 'African-American' is a 'American', 'African-American' is 'black' (synonym), 'Caucasian' is white (synonym) but 'African-American' cannot be 'Caucasian' or 'white' (although he is 'American'). This is similar to the case of mammals that lay eggs or the 'Man, Woman, Eunuch' example described earlier; people very often formulate rules such as "normally is-a", as there are always exceptions.

3.3 Results

The Lipoprotein Metabolism Ontology (LMO) was manually built in collaboration with domain experts from Unilever for the purpose of document retrieval. It consists of 653 terms (including

synonyms), with an average term length of 16 (2.2 tokens of 7.3 characters). For Automatic Term Recognition (ATR), a 'lipoprotein metabolism'-specific corpus was created, consisting of 300 abstracts collected from PubMed. Five different ATR methods were tested on that corpus, namely Text2Onto, OntoLearn, Termine [Cimiano and Vlker, 2005, Navigli and Velardi, 2004, Frantzi et al., 2000] and two methods developed in-house, one considering the relative frequency (RelFreq) of a term in the corpus and the other (TFIDF) additionally using the document frequency derived from all phrases contained in NCBI's PubMed database. OntoLearn was excluded from further analysis, as it only generated a few terms so that a meaningful comparison would be possible, see Table 7. We performed a bipartite analysis. We tried to automatically reconstruct the manually created LMO terminology, compared the terms predicted by the four methods to the current LMO terms and also evaluated manually the top 1000 retrieved terms.

	Methods									
		TFIDF		RelFreq		Termine		Text2Onto	Ontolearn
1	x	metabolic syndrome		review	x	low-density lipoprotein	x	patient	mutation
2	x	HDL	x	metabolic syndrome	x	cardiovascular disease	x	disease	fish oil
3	x	athero-sclerosis	x	diabetes	x	metabolic syndrome		risk	hyper-cholester-olaemia
4		review	x	athero-sclerosis	x	risk factor		effect	serum
5	x	LDL	x	HDL	x	cardiovascular risk		study	progression of atherosclerosis
6	x	cardiovascular disease	x	LDL	x	high-density lipoprotein		level	apheresis
7	x	diabetes	x	cardiovascular disease	x	low-density lipoprotein cholesterol	x	athero-sclerosis	omega-3
8	x	dyslipidemia	x	cholesterol	x	high-density lipoprotein cholesterol	x	cholesterol	treatment of hypertriglyc-eridemia
9	x	high-density lipoprotein		type	x	fatty acid	x	lipoprotein	reductase inhibitor
10	x	cholesterol		article	x	coronary heart dis-ease	x	statin	triglyceride
11	x	low-density lipoprotein	x	fatty acids	x	coronary artery disease		role	adhesion molecule
12	x	cardiovascular risk	x	high-density lipoprotein		clinical trial		syndrome	evolution

13	x	fatty acids		role	x	ldl cholesterol	x	diabetes	purification process
14		article	x	dyslipidemia	x	heart disease	x	trial	prescription omega-3
15	x	insulin resistance	x	low-density lipoprotein	x	diabetes mellitus		protein	omega-6
16		type	x	cardiovascular risk	x	omega-3 fatty acid	x	risk factor	hiv-infected
17	x	statin	x	hypertension		blood pressure		treatment	marker of inflammation
18	x	hypertension		combination	x	oxidative stress		event	strong evidence
19	x	inflammation	x	insulin resistance		increased risk		therapy	attractive target
20	x	VLDL		protein		density lipoprotein		review	accelerated atherosclerosis
21	x	lipid metabolism	x	disease	x	cardiovascular risk factor		type	internalization
22		combination		studies		coronary artery		mechanism	scenario
23		role	x	inflammation	x	statin therapy		evidence	protease inhibitor
24	x	oxidative stress		association	x	plant sterol		development	inflammatory cell
25	x	obesity	x	plasma	x	reverse cholesterol transport		use	inflammatory marker

Table 7: **Top 25 predicted terms per method.** Listing of the top 25 predictions for TFIDF, RelFreq, Termine, Text2Onto and OntoLearn. Terms relevant to the lipoprotein metabolism domain are marked with x.

3.3.1 Reconstruction of LMO terminology

Consider Table 8, which shows the percentage of terms that can be generated by the four methods. The first table lists the results for LMO alone, the second for LMO and terms considered relevant after manual inspection. Furthermore, we distinguish precision and average precision. The latter takes the ranking of terms into account:

$$average\ precision = \frac{\sum_{r=1}^{N}(P(r) \times rel(r)}{number\ of\ retrieved\ terms}$$

with

$$rel(r) = -\frac{2}{N^2}(r-1) + \frac{N}{2}$$

where r is the rank of retrieval and $P(r)$ is the precision at a cut-off rank. For each of the four methods we list the percentage of rele-vant terms for the top 50, top 200, and top 1000

	LMO							
	Precision				Average Precision			
Top	TFIDF	Termine	Text2Onto	RelFreq	TFIDF	Termine	Text2Onto	RelFreq
50	**35%**	19%	17%	**35%**	**65%**	54%	38%	54%
200	20%	10%	12%	**22%**	**42%**	28%	23%	37%
1000	**8%**	4%	5%	**8%**	**21%**	12%	12%	20%
	LMO + Domain expert							
	Precision				Average Precision			
Top	TFIDF	Termine	Text2Onto	RelFreq	TFIDF	Termine	Text2Onto	RelFreq
50	**75%**	67%	33%	35%	86%	**89%**	52%	70%
200	**55%**	40%	46%	22%	**74%**	65%	38%	60%
1000	**29%**	20%	14%	8%	**51%**	40%	25%	45%

Table 8: Precision and Average Precision (rank dependent) for top 50 / 200 / 1000 predictions for 4 methods (TFIDF, Relative Frequency, Termine, Text2Onto) in terms of coverage of LMO and relevant vocabulary. The key finding is that Among the top 1000 predictions there are up to 51% terms, which are in the LMO or considered good terms by expert, implying that automated term recognition can play an important role in semi-automated ontology design.

	LMO terminology predicted by TFIDF		LMO terminology literally contained
	1000	all	
300 review abstracts for "lipoprotein metabolism"	8.75%	15.23%	20.98%
3,066 abstracts for "lipoprotein metabolism"	14.99%	38.25%	53.00%
50,000 abstracts containing "lipoprotein"			71.22%

Table 9: Coverage of LMO terminology in selected document sets. The table sets the upper limit of terms that can be found with text-mining: Even a large text base with 50,000 documents contains only 71% of LMO terms. TFIDF can predict up to 38% of LMO terms.

predictions. The results show that the precision for the top 50 predictions for LMO ranges from 17-35% and 4-8% for the top 1000 predictions. Using LMO and the expert terms leads to better results of up to 75% for the top 50 predictions and up to 29% for the top 1000. Considering the average precision and thus the ranking of terms, results for the top 50 predictions go up to 89% and for the top 1000 up to 51%. Generally, Termine which favours long terms performs well for the top 50, because long terms are a good indicator of a relevant term. However, there are many short terms, which are relevant, too. The TFIDF and RelFreq method can pick up these terms, as they include background knowledge, i.e. frequencies of terms in PubMed. By and large, Text2Onto does not perform so well, as it neither includes background knowledge nor the ranking persued by Termine. Overall, the results are encouraging, as they indicate that a large part of the terminology can be generated automatically

Concerning recall, consider table 9. 3066 documents contain only 53% of the LMO terms literally. TFIDF manages to predict up 39%, which is an encouraging result. Increasing the document base to 50.000 only 71% of the LMO terms are included indicating possible upper limit.

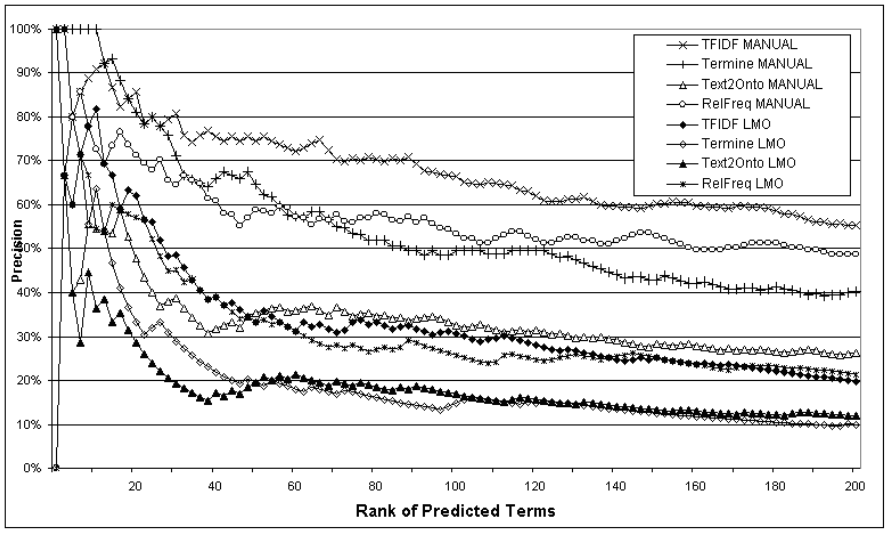

Figure 7: **Overlap with manually curated LMO and manual evaluation.**
Precision at a certain rank r represents each method's capability to recognize domain relevant terms within the top r retrieved terms. The chart shows the overlap within the top r predicted terms with LMO and the manual evaluation (MANUAL). For example, from the top 50 predicted terms by Text2Onto, 20% are in LMO and 36% are correct according to the manual evaluation.

Figure 7 provides an overview of the results we acquired from these comparisons. Figures 8 and 9 provide zoom-ins of Figure 7, describing the performance of each method in the top 50 predicted terms.

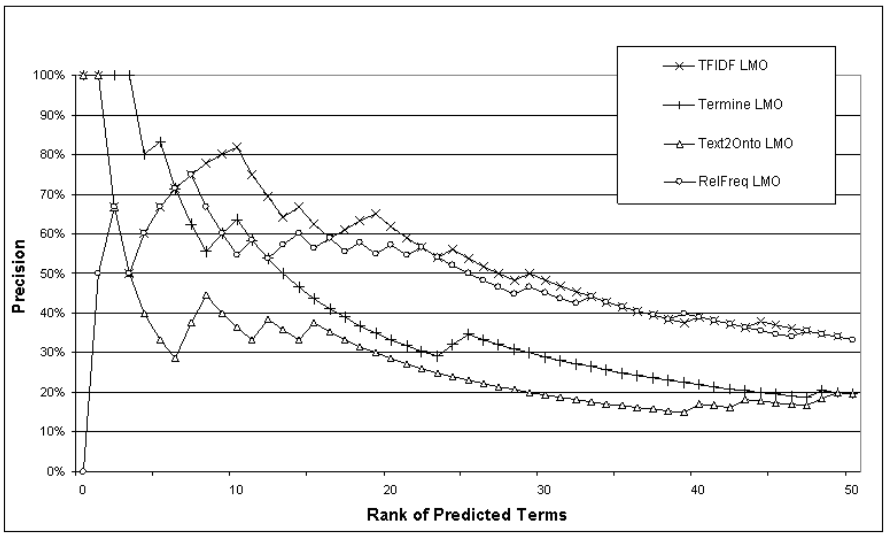

Figure 8: **Overlap with LMO.**
Precision at a certain rank r represents each method's capability to recognize domain relevant terms within the top r retrieved terms. The chart shows the overlap within the top r predicted terms with LMO. For example, from the top 10 predicted terms by Text2Onto, 40% are in LMO.

3.4 Discussion

The low coverage of the LMO in the data sets calls in question the document set selected and the suitability of the manually built LMO itself. The straightforward approach to select relevant documents from PubMed (review articles in "lipoprotein metabolism") did not return enough documents to cover all of the LMO.

The LMO terms that were absent from the 50,000 PubMed ab-stracts were grouped in five categories: rarely occurring terms, rarely occurring variants of terms, very long terms, combinations of terms/variants and, finally, terms that should normally be easily found. Terms such as 'experimentee' (2) , 'obesive' (2), 'test person' (76) and 'central fatness' (9) are LMO terms, but rarely used by authors and, therefore, rarely appearing in PubMed. The second group contains variants of terms that appear rarely in PubMed, such as 'Apo-F' (14), 'apolipoprotein c-3' (4), 'IDL I' (1), 'VLDL chol' (34), 'diabetis' (37, instead of 270177 occurrences for 'dia-betes'), 'free chol' (0, instead of 2622 for 'free cholesterol'), 'hypolipoproteinaemia' (5, "ae" spelling is rare), 'insuline resistant' (0, instead of 3912 for 'insulin resistant'), 'slo syndrome' (36) and 'sphingomyelinase deficiency disease'(0, MeSH synonym for 'Niemann-Pick Disease'). The third category contains terms that are too long and, therefore, unlikely to appear as such in text: 'receptor-mediated extra-hepatic cellular uptake' (0), 'macrophague cellular uptake' (0), 'predominance of large low-density lipoprotein particles' (0) and 'apob100 containing particles' (2). However, given the initial purpose of the LMO for document retrieval, these terms were included to be recognized by the ontology-based text-mining methods [Doms and Schroeder, 2005]. The fourth group is a combination of the previous two, i.e. LMO terms that are long terms and con-

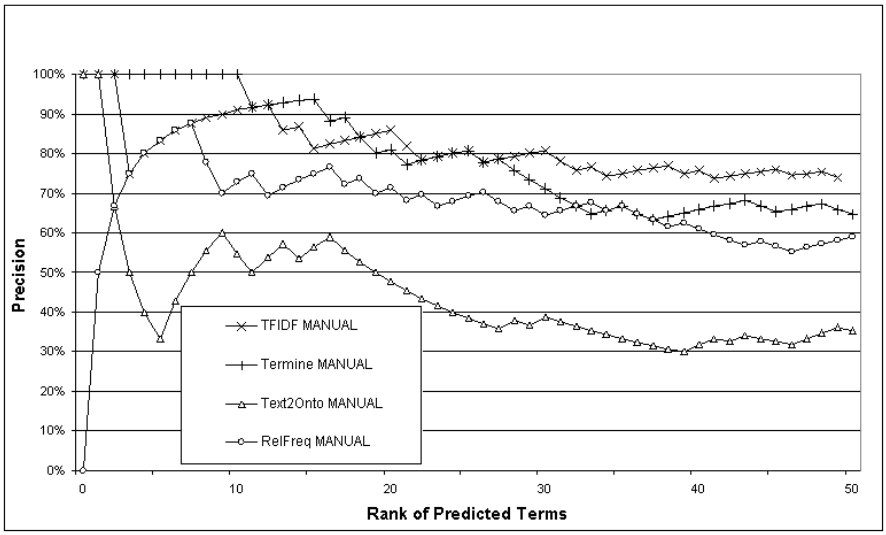

Figure 9: **Overlap with controlled lipoprotein metabolism vocabulary and additional manual evaluation (makes sense/makes no sense).**
Precision at a certain rank r represents each method's capability to recognize domain relevant terms within the top r retrieved terms. The chart shows the overlap within the top r predicted terms with the manual evaluation. For example, from the top 10 predicted terms by Termine, 100% are relevant to lipoprotein metabolism.

tain rare variants of LMO terms, such as 'elevated plasma-tg level' (0), 'increased total chol' (0, instead of 116 for 'increased total cholesterol'), 'long-lived test person' (0), 'apoprotein b100 kinetics' (0), 'elevated plasma tg concentrations' (0), and 'decreased hdl-chol' (4). The last group contains LMO terms that appear often in PubMed and should normally be identified, but are probably absent from the document set, due to its size or specificity. Such terms are 'diabetes type I' (126), 'acetyl-coa c-acyltransferase' (430), 'apolipoprotein-c' (1585), 'type-II diabetic' (1132), 'long-lived population' (23), 'middle-aged adult' (81), 'human body composition' (95), and 'lipid poor HDL' (12).

The third and fourth groups of terms belong to the same category as the hydrolase activity example described earlier. Composite terms like 'receptor-mediated extra-hepatic cellular uptake' and 'predominance of large low-density lipoprotein particles' could be easily broken into several semantic parts (e.g. receptor-mediated/ extra-hepatic/ cellular uptake, or more) and handled by an algorithm that could later compose them and still keep their semantics.

The terms that were predicted by most of the methods but were not in the LMO were further examined and grouped. These were either wrongly predicted ones (e.g. 'review', 'type', 'article', 'role', 'event', 'use') or vocabulary that could extend the current ontology. This would include disease-specific terms such as 'atherosclerosis', 'cardiovascular risk' and 'atherogenic dyslipide-mia', drugs or other chemicals such as 'statins', 'ezetimibe' and 'torcetrapib', or even method and therapy related terms like 'dose' and 'lipid lowering therapy'.

3.5 Conclusion

As pointed out in [Castro et al., 2006], automated term recognition is missing from many ontology design methologies. In this paper, we manually created a ontology for lipid metabolism with 653 terms, we derived design principles and systematically evaluated four methods for automated term recognition.

Automated predictions of up to 1000 terms generate in the order of 40-50% useful terms. Considering only the top 50 terms, the results improve up to 89%. This suggests that Automatic Term Rec-ognition (ATR) methods can aid and speed up the process of ontology design by providing lists of useful domain-specific terms, they cannot (yet) replace the manually designed term lists. The key problem to further improve these results are composite terms which do not appear literally in text, like GO's 'hydrolase activity, acting on ester bonds' or LMO's 'receptor-mediated extra-hepatic cellular uptake'.

Overall, our results show that ontology design can be performed in a semi-automatic way. The domain expert must have as initial input the output from an automatic term recognition method and proceed with enriching the ontology by following the guidelines described. These can serve as restrictions as well as decision points for including, excluding and reforming ontology terms. Once the domain expert acquires the list of candidate terms, he/she needs to decide on the relations between them. Formulation of questions is one of the most important steps in the ontology design process, helping to step from a list to an ontology.

We proposed principles for development of an ontology with text-mining as intended use, based on our personal experience from the manual development of the Lipoprotein Metabolism Ontology and GoPubMed. We related these principles to the performance of four different ATR methods and their agreement with the manually built LMO. To our knowledge, there have not yet been proposed ontology development principles with text-mining in focus. Open problems, relate to the choice of suitable text bodies for term recognition as well as generation of composite terms from basic ones.

References

[Altun and Hall, 2006] Altun, Z. and Hall, D. (2002–2006). Wormatlas. http://www.wormatlas.org.

[Aranguren et al., 2007] Aranguren, M., Bechhofer, S., Lord, P., Sattler, U., and Stevens, R. (2007). Understanding and using the meaning of statements in a bio-ontology: recasting the gene ontology in owl. *BMC Bioinformatics*, 8:57.

[Ashburner et al., 2000] Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., Davis, A., Dolinski, K., Dwight, S., Eppig, J., Harris, M., Hill, D., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J., Richardson, J., Ringwald, M., Rubin, G., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, 25(1):25–9.

[Azuma et al., 2005] Azuma, N., Tadokoro, K., Asaka, A., Yamada, M., Yamaguchi, Y., Handa, H., Matsushima, S., Watanabe, T., Kida, Y., Ogura, T., Torii, M., Shimamura, K., and Nakafuku, M. (2005). Transdifferentiation of the retinal pigment epithelia to the neural retina by transfer of the pax6 transcriptional factor. *Hum Mol Genet*, 14(8):1059–68.

[Baker and Witte, 2006] Baker, C. J. and Witte, R. (2006). Mutation mining–a prospector's tale. *Information Systems Frontiers*, 8(1):47–57.

[Baker et al., 1999] Baker, P., Goble, C., Bechhofer, S., Paton, N., Stevens, R., and Brass, A. (1999). An ontology for bioinformatics applications. *Bioinformatics*, 15(6):510–520.

[Baldock et al., 2003] Baldock, R., Bard, J., Burger, A., Burton, N., Christiansen, J., Feng, G., Hill, B., Houghton, D., Kaufman, M., Rao, J., Sharpe, J., Ross, A., Stevenson, P., Venkataraman, S., Waterhouse, A., Yang, Y., and Davidson, D. (2003). Emap and emage: a framework for understanding spatially organized data. *Neuroinformatics*, 1(4):309–25.

[Bard et al., 1998] Bard, J., Kaufman, M., Dubreuil, C., Brune, R., Burger, A., Baldock, R., and Davidson, D. (1998). An internet-accessible database of mouse developmental anatomy based on a systematic nomenclature. *Mech Dev*, 74(1-2):111–20.

[Bard et al., 2005] Bard, J., Rhee, S., and Ashburner, M. (2005). An ontology for cell types. *Genome Biology*, 6(2):R21.

[Berardini et al., 2004] Berardini, T., Mundodi, S., Reiser, L., Huala, E., Garcia-Hernandez, M., Zhang, P., Mueller, L., Yoon, J., Doyle, A., Lander, G., Moseyko, N., Yoo, D., Xu, I., Zoeckler, B., Montoya, M., Miller, N., Weems, D., and Rhee, S. (2004). Functional annotation of the arabidopsis genome using controlled vocabularies. *Plant Physiol*, 135(2):745–55.

[Berneis and Rizzo, 2004] Berneis, K. and Rizzo, M. (2004). Ldl size: does it matter? *Swiss Med Wkly*, 134(49-50):720–4.

[Bodenreider, 2004] Bodenreider, O. (2004). The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Res*, 32(Database issue):D267–70.

[Boyack, 2004] Boyack, K. (2004). Mapping knowledge domains: characterizing pnas. *Proc Natl Acad Sci U S A*, 101 Suppl 1:5192–9.

[Brinckmann et al., 2006] Brinckmann, A., Rüther, K., Williamson, K., Lorenz, B., Lucke, B., Nürnberg, P., Trijbels, F., Janssen, A., and Schuelke, M. (2006). De novo double mutation in PAX6 and mtDNA tRNA (Lys) associated with atypical aniridia and mitochondrial disease. *J Mol Med*, 85(2):163–8.

[Cantor et al., 2005] Cantor, M., Sarkar, I., Bodenreider, O., and Lussier, Y. (2005). Genestrace: phenomic knowledge discovery via structured terminology. In *Pac Symp Biocomput*, pages 103–114.

[Castro et al., 2006] Castro, A., Rocca-Serra, P., Stevens, R., Taylor, C., Nashar, K., Ragan, M., and Sansone, S. (2006). The use of concept maps during knowledge elicitation in ontology development processes–the nutrigenomics use case. *BMC Bioinformatics*, 7:267.

[Cimiano and Vlker, 2005] Cimiano, P. and Vlker, J. (2005). Text2onto - a framework for ontology learning and data-driven change discovery. In Montoyo, A., Munoz, R., and Metais, E., editors, *Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems (NLDB)*, volume 3513 of *Lecture Notes in Computer Science*, pages 227–238, Alicante, Spain. Springer.

[de Solla Price, 1965] de Solla Price, D. J. (1965). Networks of scientific papers. *Science*, 149(3683):510–515.

[Doms and Schroeder, 2005] Doms, A. and Schroeder, M. (2005). GoPubMed: exploring PubMed with the Gene Ontology. *Nucleic Acids Res*, 33(Web Server issue):W783–W786.

[Dou et al., 2002] Dou, D., McDermott, D., and Qi, P. (2002). Ontology translation by ontology merging and automated reasoning. In *Proc. EKAW2002 Workshop on Ontologies for Multi-Agent Systems*, pages 3–18.

[Duineveld et al., 2000] Duineveld, A. J., Stoter, R., Weiden, M. R., Kenepa, B., and Benjamins, V. R. (2000). Wondertools?: a comparative study of ontological engineering tools. *Int. J. Hum.-Comput. Stud.*, 52(6):1111–1133.

[Eaton, 2006] Eaton, A. D. (2006). HubMed: a web-based biomedical literature search interface. *Nucleic Acids Res*, 34(Web Server issue):W745–W747.

[Ehrler et al., 2005] Ehrler, F., Geissbhler, A., Jimeno, A., and Ruch, P. (2005). Data-poor categorization and passage retrieval for gene ontology annotation in swiss-prot. *BMC Bioinformatics*, 6 Suppl 1:S23.

[Evsikov et al., 2004] Evsikov, A., de, V. W., Peaston, A., Radford, E., Fancher, K., Chen, F., Blake, J., Bult, C., Latham, K., Solter, D., and Knowles, B. (2004). Systems biology of the 2-cell mouse embryo. *Cytogenet Genome Res*, 105(2-4):240–50.

[Finkel et al., 2005] Finkel, J., Dingare, S., Manning, C., Nissim, M., Alex, B., and Grover, C. (2005). Exploring the boundaries: gene and protein identification in biomedical text. *BMC Bioinformatics*, 6 Suppl 1:S5.

[Frantzi et al., 2000] Frantzi, K., Ananiadou, S., and Mima, H. (2000). Automatic recognition of multi-word terms:. the c-value/nc-value method. *International Journal on Digital Libraries*, V3(2):115–130.

[Garfield and Melino, 1997] Garfield, E. and Melino, G. (1997). The growth of the cell death field: an analysis from the isi-science citation index. *Cell Death Differ*, 4(5):352–61.

[Goetz and der Lieth CW von, 2005] Goetz, T. and der Lieth CW von (2005). Pubfinder: a tool for improving retrieval rate of relevant pubmed abstracts. *Nucleic Acids Res*, 33(Web Server issue):W774–8.

[Grumbling and Strelets, 2006] Grumbling, G. and Strelets, V. (2006). Flybase: anatomical data, images and queries. *Nucleic Acids Res*, 34(Database issue):D484–8.

[Hakenberg et al., 2007] Hakenberg, J., Royer, L., Plake, C., Strobelt, H., and Schroeder, M. (2007). Me and my friends: gene mention normalization with background knowledge. In *Proceedings 2nd BioCreAtIvE Challenge Evaluation Workshop*.

[Hirschman et al., 2005] Hirschman, L., Yeh, A., Blaschke, C., and Valencia, A. (2005). Overview of biocreative: critical assessment of information extraction for biology. *BMC Bioinformatics*, 6 Suppl 1:S1.

[Hoffmann and Valencia, 2004] Hoffmann, R. and Valencia, A. (2004). A gene network for navigating the literature. *Nat Genet*, 36(7):664.

[Jaiswal et al., 2005] Jaiswal, P., Avraham, S., Ilic, K., Kellogg, E. A., McCouch, S., Pujar, A., Reiser, L., Rhee, S. Y., Sachs, M. M., Schaeffer, M., Stein, L., Stevens, P., Vincent, L., Ware, D., and Zapata, F. (2005). Plant ontology (po): a controlled vocabulary of plant structures and growth stages: Research articles. *Comp. Funct. Genomics*, 6(7‐8):388–397.

[Jaiswal et al., 2006] Jaiswal, P., Ni, J., Yap, I., Ware, D., Spooner, W., Youens-Clark, K., Ren, L., Liang, C., Zhao, W., Ratnapu, K., Faga, B., Canaran, P., Fogleman, M., Hebbard, C., Avraham, S., Schmidt, S., Casstevens, T., Buckler, E., Stein, L., and McCouch, S. (2006). Gramene: a bird's eye view of cereal genomes. *Nucleic Acids Res*, 34(Database issue):D717–23.

[Jensen et al., 2006] Jensen, L. J., Saric, J., and Bork, P. (2006). Literature mining for the biologist: from information retrieval to biological discovery. *Nature Reviews Genetics*, 7(2):119–129.

[Kim et al., 2004] Kim, J. D., Ohta, T., Tateishi, Y., and Tsujii, J. (2004). Introduction to the bio-entity recognition task at jnlpba. In *Proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA-2004)*, Geneva, Switzerland.

[Kleinjan et al., 2006] Kleinjan, D., Seawright, A., Mella, S., Carr, C., Tyas, D., Simpson, T., Mason, J., Price, D., and van, H. V. (2006). Long-range downstream enhancers are essential for pax6 expression. *Dev Biol*, 299(2):563–81.

[Lambrix and Edberg, 2003] Lambrix, P. and Edberg, A. (2003). Evaluation of ontology merging tools in bioinformatics. In *Pac Symp Biocomput*, pages 589–600, Department of Computer and Information Science, Linkpings universitet, 581 83 Linkping, Sweden.

[Lambrix and Tan, 2006] Lambrix, P. and Tan, H. (2006). Sambo - a system for aligning and merging biomedical ontologies. *Journal of Web Semantics*, 4(3):196–206.

[Lame, 2004] Lame, G. (2004). Using NLP techniques to identify legal ontology components: Concepts and relations. *Artificial Intelligence and Law*, 12(4):379–396.

[Lee et al., 2007] Lee, L., Horn, F., and Cohen, F. (2007). Automatic extraction of protein point mutations using a graph bigram association. *PLoS Comput Biol*, 3(2):e16.

[Lewis et al., 2006] Lewis, J., Ossowski, S., Hicks, J., Errami, M., and Garner, H. R. (2006). Text similarity: an alternative way to search MEDLINE. *Bioinformatics*, 22(18):2298–2304.

[Mller et al., 2004] Mller, H.-M., Kenny, E. E., and Sternberg, P. W. (2004). Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol*, 2(11):e309.

[Navigli and Velardi, 2004] Navigli, R. and Velardi, P. (2004). Learning domain ontologies from document warehouses and dedicated web sites. *Comput. Linguist.*, 30(2):151–179.

[Newman, 2004] Newman, M. (2004). Coauthorship networks and patterns of scientific collaboration. *Proc Natl Acad Sci U S A*, 101 Suppl 1:5200–5.

[Ogren et al., 2005] Ogren, P., Cohen, K., and Hunter, L. (2005). Implications of compositionality in the gene ontology for its curation and usage. In *Pac Symp Biocomput*, pages 174–85.

[Ogren et al., 2004] Ogren, P. V., Cohen, K. B., Acquaah-Mensah, G. K., Eberlein, J., and Hunter, L. (2004). The compositional structure of gene ontology terms. In *Pac Symp Biocomput*, pages 214–225, University of Colorado at Boulder, Dept. of Computer Science, Boulder, CO, USA.

[Perez-Iratxeta et al., 2003] Perez-Iratxeta, C., Prez, A., Bork, P., and Andrade, M. (2003). Update on xplormed: A web server for exploring scientific literature. *Nucleic Acids Res*, 31(13):3866–8.

[Popescul and Ungar, 2000] Popescul, A. and Ungar, L. (2000). Automatic labeling of document clusters.

[Rosse and Mejino, 2003] Rosse, C. and Mejino, J. (2003). A reference ontology for biomedical informatics: the foundational model of anatomy. *J Biomed Inform*, 36(6):478–500.

[Smith et al., 2004] Smith, B., Khler, J., and Kumar, A. (2004). On the application of formal principles to life science data: A case study in the gene ontology. In *Proceedings of DILS 2004 (Data Integration in the Life Sciences)*, volume 2994 of *Lecture Notes in Bioinformatics*, pages 79–94, Berlin. Springer.

[Soldatova and King, 2005] Soldatova, L. and King, R. (2005). Are the current ontologies in biology good ontologies? *Nat Biotechnol*, 23(9):1095–8.

[Sowa, 2000] Sowa, J. F. (2000). *Knowledge Representation: Logical, Philosophical, and Computational Foundations*. Brooks Cole Publishing Co.

[Spackman, 2004] Spackman, K. (2004). Snomed ct milestones: endorsements are added to already-impressive standards credentials. *Healthc Inform*, 21(9):54, 56.

[Stefanowski and Weiss, 2003] Stefanowski, J. and Weiss, D. (2003). Carrot^2 and language properties in web search results clustering. In Ruiz, E. M., Segovia, J., and Szczepaniak, P. S., editors, *Proceedings of AWIC-2003, First International Atlantic Web Intelligence Conference*, volume 2663 of *Lecture Notes in Computer Science*, pages 240–249, Madrid, Spain. Springer.

[Thut et al., 2001] Thut, C., Rountree, R., Hwa, M., and Kingsley, D. (2001). A large-scale in situ screen provides molecular evidence for the induction of eye anterior segment structures by the developing lens. *Dev Biol*, 231(1):63–76.

[Uschold, 1996] Uschold, M. (1996). Building ontologies: Towards a unified methodology. In *16th Annual Conf. of the British Computer Society Specialist Group on Expert Systems*, Cambridge, UK.

[Vincent et al., 2003] Vincent, P., Coe, E., and Polacco, M. (2003). Zea mays ontology–a database of international terms. *Trends Plant Sci*, 8(11):517–20.