GeS 2007 German e-Science

Available online at http://www.ges2007.de

StemNet: An Evolving Service for Knowledge Networking in the Life Sciences

Udo Hahn[1], Joachim Wermter[1], David S. DeLuca[2], Rainer Blasczyk[2], Michael Poprat[1], Asad Bajwa[3], Peter A. Horn[2]

[1] Friedrich-Schiller Universität Jena, Computerlinguistik – JULIE Lab, 07743 Jena
[2] Medizinische Hochschule Hannover, Inst. für Transfusionsmedizin, 30625 Hannover
[3] Clarity AG, 61352 Bad Homburg
Web: http://www.stemnet.de
email: {hahn,wermter}@coling-uni-jena.de
phone: (+49 3641) 944 320 *fax:* (+49 3641) 944 321

Abstract

Up until now, crucial life science information resources, whether bibliographic or factual databases, are isolated from each other. Moreover, semantic metadata intended to structure their contents is supplied in a manual form only. In the StemNet project we aim at developing a framework for semantic interoperability for these resources. This will facilitate the extraction of relevant information from textual sources and the generation of semantic metadata in a fully automatic manner. In this way, (from a computational perspective) unstructured life science documents are linked to structured biological fact databases, in particular to the identifiers of genes, proteins, etc. Thus, life scientists will be able to seamlessly access information from a homogeneous platform, despite the fact that the original information was unlinked and scattered over the whole variety of heterogeneous life science information resources and, therefore, almost inaccessible for integrated systematic search by academic, clinical, or industrial users.

1 State of the Art in Accessing Life Sciences Information

The life sciences, i.e., medicine, biology, chemistry and pharmacology, experience a dramatic growth of the amount of available data. This can be observed, e.g., in the area of genomic and proteomic research in which we witness an exponential growth of available sequence databases. Another source of evidence for this trend is the ever-increasing number of life science publications, i.e., scientific journal articles, patent reports as well as the growing proportion of free-text comments in biomedical databases.

At this point, the sheer volume of biomedical literature makes it almost impossible for biologists, clinical researchers and medical professionals to retrieve all relevant information on a specific topic and to keep up with current research. For example, in the world's largest bibliographic database for the life sciences,

PubMed[1], the current number of entries (as of March 2007) already amounts to over 16 million entries, with up to 4,000 new ones being added each day. As a result, in recent years PubMed has attained a truly (and still growing) global impact as the most widely used and queried bibliographic database in the life sciences. This is exemplified by the number of query requests which is steadily rising (see Figure 1).[2]

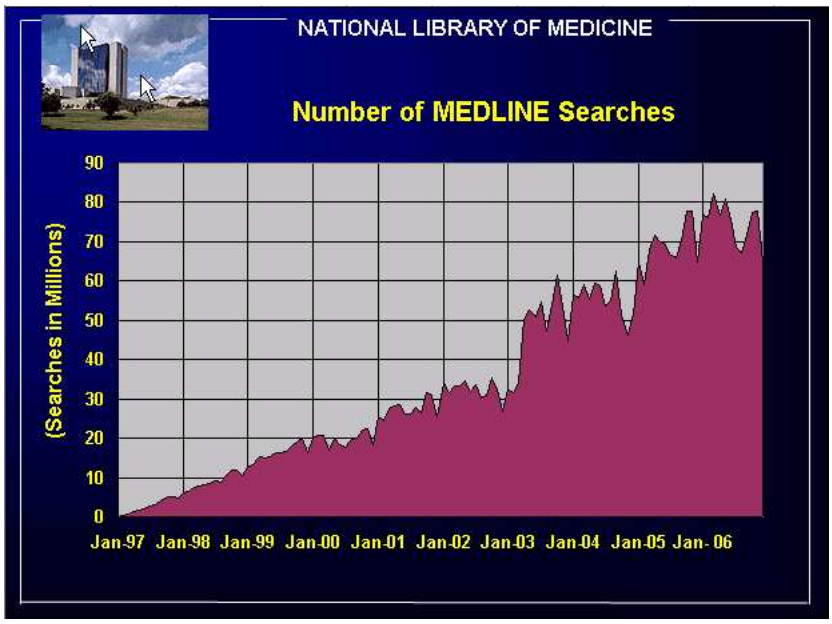

Figure 1: Development of PubMed Search Statistics

Large-scale and fully-interlinked semantic access to this vast amount of knowledge, however, is hampered mainly for two reasons (as illustrated in Figure 2):

- Current retrieval methods are insufficient since they are not geared toward getting deeply at the semantics of the biomedical text. Apart from bibliographic meta-information such as author names and publication years, the user interface of PubMed basically supports keyword-based (Google-like) querying. Due to the terminological and semantic complexity of the life sciences domain, the retrieval results for such queries are typically incomplete and suboptimal [1]. More semantically focused searches (e.g., for certain proteins or biological processes in which such proteins are involved) are only marginally supported, e.g., by manually assigned

[1] http://www.ncbi.nlm.nih.gov/entrez/

[2] http://www.nlm.nih.gov/bsd/medline_growth.html

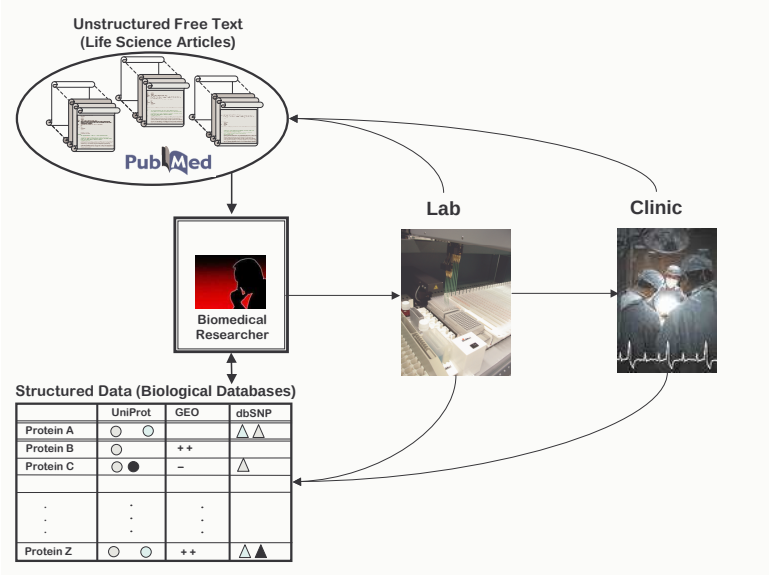

Figure 2: Current Practice: Hampered Access to Life Sciences Information

document-level indices as provided by the biomedical MeSH thesaurus[3] which, however, has proven to be rather inconsistent and incomplete. Furthermore, PubMed needs to service *all* aspects of the global life science community and thus needs to retain a high degree of semantic generality. This, however, stands in contrast to very specific search topics often formulated by researchers and clinical users.

- Apart from being locked in unstructured free text, a substantial amount of biomedical knowledge is housed in structured biological databases. These databases focus on specialized biomedical data, such as (species-specific) sequence information for defined genes and proteins, gene expression information in certain tissues, etc. Unfortunately, the knowledge found both in unstructured text and in structured databases is not linked. Thus, if a biomedical researcher finds information on a certain protein in a scientific article, linking this information to the respective database entry for this protein in a specialized database is usually not supported. An additional knowledge management problem immediately occurs since protein identifiers differ from one database to the other in an unpredictable way. Similarly, links from free-text fields of a database (which contain manually supplied annotations of the data in verbal, i.e., unstructured form) to relevant publications are not supplied on a larger scale. In any case, such linkings or mappings from free-text sources to unique biomedical database entries are hampered by the enormous degrees of ambiguity of biomedical terms and names [2].

[3] `http://www.nlm.nih.gov/mesh`

2 Goals of the StemNet Project

The goals of the StemNet[4] project respond to these two shortcomings. On the one hand, we aim at providing truly *semantic* access to the vast amount of knowledge found in the unstructured free texts of the PubMed bibliographic database. On the other hand, we aim at *linking* this knowledge encoded in free texts to respective knowledge stored in structured biomedical databases. In StemNet we plan to improve the semantic interoperability of currently disconnected information in the life sciences.

The biomedical subdomains the StemNet project focuses on are Hematopoietic Stem Cell Transplantation (HSCT) and Immunology. Both lie at the center of the fast-growing and crucial interface between genomic/proteomic research, on the one hand, and medical/clinical application, on the other hand. HSCT is used for a variety of malignant and nonmalignant disorders to replace a defective host marrow or immune system with a normal donor marrow and immune system. In many cases, the clinical treatment of patients with leukemia and other malignant hematological tumors is only successful, if a HSCT with a genetically different (allogeneic) donor is carried out, thus triggering the therapeutic effect of tumor cell elimination, known as the graft versus leukemia (GVL) effect.

However, besides this desirable effect, numerous unintended immunological side effects might occur. Graft versus host disease (GVHD) is a very common immunological complication of allogeneic transplantation. It is an immune response of donor T lymphocytes against host cells. Current research focuses on improving our understanding of the pathophysiologic pathways of GVHD to design targeted therapies and genetic modifications of donor T-cells in order to prevent and treat GVHD.

The high risk of HSCT is due to the complex genetic differences of both HLA-genes and non-HLA genes between stem cell recipients and donors [3], which can only be controlled for through a complex and interactive analysis of numerous parameters. Since the GVHD and GVL effect are closely interrelated, the severity of GVHD is inversely related to the risk of relapse and strategies aiming at reducing GVHD may increase relapse rates. Currently, new strategies are being developed to separate these two effects in order to decrease the incidence and severity of GVHD without increasing the risk of relapse.

The overall goal of the StemNet project is thus to increase the clinical success rate for HSCT by taking advantage of the comprehensive, but highly dispersed and heterogeneous data available in Internet-based textual and database repositories.

3 Resources for StemNet's Knowledge Network

Using PubMed as a starting point, the following knowledge resources are essential for the StemNet knowledge network under construction (see Figure 3):

[4] `http://www.stemnet.de`

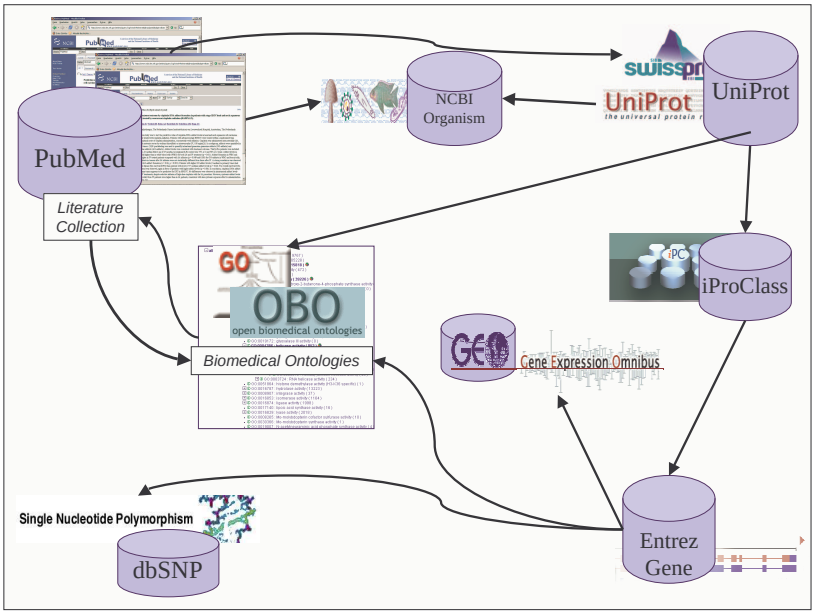

Figure 3: StemNet Knowledge Resources Network

- **OBO – Open Biomedical Ontologies.** OBO[5] is an umbrella organization for ontologies and shared terminologies for use across all biological and biomedical domains. In particular, the *Gene Ontology* (GO) [4] provides a community-wide accepted semantic framework to describe and annotate biomedical knowledge found both in unstructured free text and in biomedical databases. In order to grant semantic access to the scientific literature kept in PubMed, it is essential to annotate textual data with OBO/GO-based biomedical terms (in particular, the molecular function and location of genes and proteins, as well as the biological processes they are involved in). Once a significant sample of free text is manually annotated using this vocabulary, such an annotated corpus can be exploited to automatically train entity and relation taggers in a supervised way [5]. After successful training, these text analysis engines will perform large-scale semantic annotation of textual data in a fully automatic way.
- **UniProt and iProClass.** The Universal Protein Resource (UniProt) provides the life-science community with a single, centralized, authoritative resource for protein sequences and functional information. It is a comprehensive, fully classified, richly and accurately annotated protein sequence knowledge base with extensive cross-references [6]. Each protein entry is associated with its respective organism (e.g., human, mouse, bacteria, viruses, etc.) and provides a link to the NCBI taxonomy organism

[5] `http://obo.sourceforge.net`

database.[6] The iProClass mapping database[7] links UniProt to over 90 biological databases, including databases for protein families, functions and pathways, interactions, structures and structural classifications, genes and genomes, disease information, etc.

- **Entrez Gene** is provided by the U.S. National Center for Biotechnology Information (NCBI)[8] to organize information about genes, and serves as a major node in the nexus of genomic map, sequence, expression, protein structure, function, and homology data. Entrez Gene records are established for known or predicted genes, which are defined by their nucleotide sequence or map position. This database serves as a hub of information for databases both within and external to NCBI.
- **dbSNP – Single Nucleotide Polymorphism.** In collaboration with the National Human Genome Research Institute, the NCBI has also set up the dbSNP database to serve as a central repository for both single base nucleotide substitutions and short deletion and insertion polymorphisms, which are key to genetics research in associating sequence variations with heritable phenotypes (particularly diseases).
- **GEO – Gene Expression Omnibus.** In genomics research, the examination of gene expression patterns using high-throughput techniques has become a core technology in the recent years. Microarray hybridization and serial analysis of gene expression (SAGE) allow for the simultaneous quantification of tens of thousands of gene transcripts. The *Gene Expression Omnibus* (GEO) is a public repository that archives and freely distributes high-throughput gene expression data submitted by the scientific community. GEO currently stores some billion individual gene expression measurements, which are derived from over 100 organisms, addressing a wide range of biological issues.

Up until now, these resources remain, by and large, unconnected. One major goal of the StemNet project will be to interlink the underlying terminological resources used to describe the biological data in the databases and thus develop a conceptual foundation for interoperability based on a carefully designed formal ontology infrastructure (the rationale and progress of this work is described in [7]). Once this link has been fully established, these resources will be integrated into the StemNet system as its conceptual backbone.

4 Semantic Knowledge Networking and Semantic Access

Using state-of-the-art text mining technology [8, 9], we automatically annotate a sample of the PubMed textual data with terms from the OBO/GO ontologies. Adding this semantic metadata to documents empowers and further facilitates *semantic* retrieval of biomedical knowledge [10, 11] beyond the traditional keyword-based search [1]. In this respect, we also annotate the molecular

[6] http://130.14.29.110/Taxonomy

[7] http://pir.georgetown.edu/iproclass

[8] http://www.ncbi.nlm.nih.gov/

functions of genes and proteins, the key players of biological processes at the molecular level. In preliminary experiments, we achieved an F-score[9] of about 90% in automatically annotating the immunologically relevant *cytokine function* of proteins and about 80% F-score in annotating mentions of *cytokine receptor functions*. Similar results were also obtained for *variation events* (i.e., polymorphisms), *organisms*, *immune cells* and *antigens*.[10] This evaluation data, still in a very early stage of the project, already compares with the performance level that has been reported at BioCreaTive, the latest major BioNLP software competition [12].

After having identified a protein name in a text, its entry in the UniProt database must be located. This is a challenging task because protein names are highly ambiguous on several layers of meaning [2]. Annotating the respective organism and linking it to NCBI taxonomy organism database aids in this disambiguation task. The UniProt identifier opens up the door to several other StemNet-relevant biological databases. Through the iProClass mapping database, a knowledge link to Entrez Gene is established, and from there, additional links to GEO and dbSNP can be constructed. Moreover, both the UniProt and the Entrez Gene database entries for genes and proteins also contain (curated) GO annotations. We then come full circle, as these descriptive items, in turn, may serve as additional semantic metadata for the original PubMed text and thus facilitate semantic access and retrieval for the user.

In this way, the StemNet Knowledge Server semantically links the disparate biomedical knowledge resources and thus provides biomedical researchers with an integrated view of relevant information. In particular, the user accesses information from the homogeneous StemNet server (see Figure 4). This contrasts with the original search paradigm (see Figure 2), where bibliographic and fact databases are strictly isolated from each other and thus each must be searched separately in a complicated, expensive and error-prone manner. The results of these searches largely depend on the ingenuity, experience and time investment of the searcher, who has to battle with different query languages and large amounts of a priori knowledge related to the relation structure and other content issues specific to *each* of the databases involved.

5 Assessment of User Preferences for StemNet

In order to get potential users involved in an early phase of system design and, thus, tailor StemNet to their needs, we conducted an inquiry at three major German university hospitals (Hannover, Freiburg, and Jena). We interviewed not only researchers in the field of HSCT but also employees working in HLA laboratories as well as physicians in transplantation wards. In summary, although the types of information they need differ between each group, their search strate-

[9] The F-score is a standard evaluation metric which balances between precision and recall measurements; cf. [1].

[10] Actually, on a semantically more fine-grained level, the StemNet annotations already cover over 60 different semantic categories.

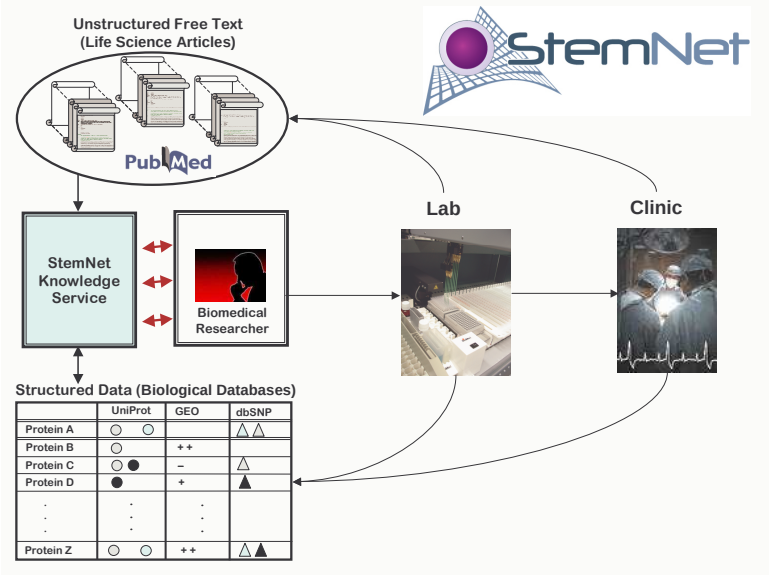

Figure 4: StemNet Scenario for Enhanced User Access to Life Sciences Information

gies showed apparent similarities.

Most of the participants in this informal user study started a typical search using Google or PubMed. After an article of interest was found in PubMed, its built-in "related article" function was tried to find further relevant articles. However, the results of this function were often considered to be of minor quality. Furthermore, the participants in this study pointed out that they would prefer a customizable function by being able to parametrize relatedness (e.g., by selecting the biological or medical terms of interest that should occur in other abstracts, as well). As part of the inquiry, already existing semantic search engines that link gene and protein names from text to (NCBI) databases were also demonstrated. Although information linking, in general, was found to be helpful in the search process, neither the results nor the usability of these search engines were considered adequate. So there is room left for improvement by deploying StemNet.

6 Conclusions and Outlook

The StemNet Knowledge Server links disparate biomedical knowledge resources on a semantic layer and thus enables biomedical users to access and search for relevant information in an integrated manner. Starting from the vast amount of life science documents in the PubMed literature database, it provides the user with a semantic view on these documents in terms of annotations (semantic metadata added to the texts). The annotated documents are interlinked with external knowledge resources, such as biomedical ontologies and databases.

While the focus of the StemNet project is on the clinically relevant biomedical subdomain of Hematopoietic Stem Cell Transplantation (HSCT), the underlying methodologies which provide for semantic interoperability are designed and implemented to be easily extensible to other subdomains of the life sciences and, possibly, even translate to other science and technology domains, as well.

Acknowledgements

The StemNet project is funded by the German Ministry of Education and Research (BMBF) via its *e-Science* initiative (funding code: 01DS001A to 1C). The project started in April 2006.

References

1. Hersh WR. Information Retrieval. A Health and Biomedical Perspective. Springer, 2nd ed., 2002.
2. Hirschman L, Colosimo M, Morgan A, Yeh A. Overview of BioCreAtIvE task 1B: Normalized gene lists. BMC Bioinformatics. 6 (Suppl 1: S11) 2005.
3. Horn PA, Elsner HA, Blasczyk R. Tissue typing for hematopoietic cell transplantation: HLA-DQB1 typing should be included. Pediatric Transplantation 10(6) 2006:753–754.
4. Gene Ontology Consortium. The Gene Ontology (GO) project in 2006. Nucleic Acids Research. 34(1) 2006:322–326
5. Feldman R, Sanger J. The Text Mining Handbook. Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press, 2007.
6. Bairoch A et al. The Universal Protein Resource (UniProt). Nucleic Acids Research. 33(1) 2005:154–159.
7. Schulz S, Beisswanger E, Hahn U, Wermter J, Kumar A, Stenzhorn H. From Genia to BioTop: Towards a top-level ontology for biology. In: Formal Ontologies in Information Systems. Proceedings of the FOIS 2006 Conference, pp.103-114. 2006.
8. Hahn U, Wermter J. Levels of Natural Language Processing for Text Mining. In: S. Ananiadou and J. McNaught (Eds.), Text Mining for Biology and Biomedicine, pp.13–41. Artech House Publishers. 2006.
9. Buyko E, Wermter J, Poprat M, Hahn U. Automatically adapting an NLP core engine to the biology domain. Proceedings of the ISMB 2006 "Joint Linking Literature, Information and Knowledge for Biology and the 9th Bio-Ontologies Meeting". 2006.
10. Ferrucci D, Lally A. Building an example application with the Unstructured Information Management Architecture. IBM Systems Journal 43(3) 2004:455–475.
11. Carmel D, Maarek YS, Mandelbrod M, Mass Y, Soffer A. Searching XML documents via XML fragments. Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2003), pp.151–158. 2003
12. Yeh A, Morgan A, Colosimo M, Hirschman L. BioCreAtIvE task 1A: Gene mention finding evaluation. BMC Bioinformatics. 6 (Suppl 1: S2) 2005.