



SemanticMining

NoE 507505

Semantic Interoperability and Data Mining in Biomedicine

D24.2

Report: Data mining and information retrieval

Report Version: 1.0

Report Preparation Date: 30-May-06

Classification: RE

Contract Start Date: 2004-01-01

Duration: 3 years

Project Co-ordinator: Hans Åhlfeldt

Department of Biomedical Engineering / Medical Informatics

S-581 83 Linköping University, Sweden

<hans.ahlfeldt@imt.liu.se>



Project funded by the European Community under the FP6 programme “Integrating and Strengthening the European Research Area” (2002-2006)



Table of Contents

1	ADMINISTRATIVE INFORMATION	4
2	INTRODUCTION	5
3	OBJECTIVES OF THE WORKPACKAGE	5
4	QUALITY INDICATORS RELEVANT TO WP24	6
4.1	Q1 Workshops and symposiums	6
4.2	Q2 Sharing of resources and use of research software tools.....	6
4.3	Q6 Short-and medium-term visits of staff members	6
4.4	Q7 Co-authoring of research papers, reports and educational materials	6
5	SUMMARY OF ACTIVITIES IN 2005 / 2006	7
5.1	Workshops and Symposiums	7
5.1.1	Workpackage Meetings	7
5.1.2	International Conference Participation	7
5.2	Co-authoring	8
5.3	Joint Research Programme	10
5.3.1	EBI's text processing engines (Cat-2)	10
5.3.2	GO categorizer (Cat-2)	12
5.3.3	Integration of text resources and biomedical databases (Cat-3)	13
5.3.4	Preparation of cross-lingual information retrieval (Cat-3).....	13
5.3.5	Identification and disambiguation of biomedical semantic types in the scientific literature (Cat-3).....	13
5.3.6	Towards a common annotation framework for semantic types (Cat-1).....	15
5.3.7	Medical image CLEF (Cat-4).....	17
5.4	Exchange Visits	18
5.5	Follow-up Grant proposal	18
5.5.1	IST funded project "BOOTStrep"	18
5.5.2	@neurIST Project	18
5.6	New collaborations established throughout 2005 and 2006	18
5.6.1	SYMBiomatics	18
5.6.2	Network of Excellence "InfoBioMed"	19



5.7	Mini-Symposium on Semantic Enrichment of the Scientific Literature	20
6	WORKPACKAGE ASSESSMENT	21
6.1	Assessment against Quality Indicators	21
6.2	Assessment Against Workpackage Objectives.....	21
6.2.1	Exchange of Methods and Resources	21
6.2.2	Integration with other Workpackages.....	21
7	TABLES AND FIGURES.....	23
7.1.1	EBIMed	23
7.1.2	Screenshots from the GO Browser/GO Categorizer	25
7.1.3	Medical Image CLEF	26
	REFERENCES	28

1 Administrative Information

Lead contractor: EBI

Responsible: EBI, UKLFR, Jena, DIM

Participants: IMT-LiU, IDA-LiU, VUM, SPIM

New participants: Medical Center Erasmus University Rotterdam

Deliverable: D24.2

Author: Dietrich Rebholz-Schuhmann

Summary

WP24 of the Network of Excellence achieved several goals over the time period 2005 and first half of 2006. Research work was focused to information retrieval and information extraction from the scientific literature in the biomedical domain. Ongoing work lead to the realisation of public services based on state of the art technology (e.g. EBIMed, GO categoriser, MorphoSaurus (mainly WP20)). Ongoing work was published in journal articles in conjunction with conference presentations. Related research work is concerned with the disambiguation of semantic types, the standardisation of the representation of semantic types in biomedical text and ongoing work in the integration of cross-lingual term normalisation for information retrieval (collaboration with WP20).

In addition members of WP24 achieved to raise fund moneys for ongoing and new research work (e.g. projects “BOOTStrep” and “@neurIST”) and established collaborations with other project initiatives funded by the European Commission in the biomedical domain (e.g. SYMBiotics SSA and Network of Excellence “InfoBioMed”).

In addition outreach activities lead in conjunction with WP14 and WP15 to the organisation of the international conferences SMBM 2005 (at the European Bioinformatics Institute) and SMBM 2006 (at the Jena University) and to the one-day Mini-Symposium on “Semantic Enrichment of the Scientific Literature”.

2 Introduction

The European Network of Excellence “SemanticMining” instantiates the collaboration between 25 organisations from 11 European countries. In this deliverable the progress in work package 24 “Data mining and information retrieval” will be reported. The report covers the research activity from month 11 (November 2004) until month 29 (May 2006) after the start of the NoE.

Target audience of this deliverable is the Commission and project partners, but not the public.

The lead contractor of WP 24 (EBI) is the leading bioinformatics research and service center in Europe and is thus complementary to the domain of medical informatics forming the core in the NoE. On the other side EBI fulfils the public demands on IT services in the biomedical domain. As a result plans have been settled between EBI, DIM, UKLFR and Jena to establish different solutions for information retrieval (IR) and information extraction (IE) engines, which provide access to Medline abstracts and eventually to full text documents. Such IR and IE solutions integrate software components available from partners in the NoE, amongst others the representation of medical terms from Morphosaurus (<http://www.morphosaurus.net>) for cross-lingual medical information retrieval.

Due to the resubmission of the deliverable, ongoing work in particular in collaboration with the NoE InfoBioMed and project partners in the SYMBiomatics project will be attached to the document. Furthermore a one-day mini-symposium (Title: “Semantic Enrichment of Scientific Literature”) has been organized in conjunction with the NoE Semantic Mining to raise consent on the automatic semantic analysis of scientific literature.

3 Objectives of the Workpackage

The work package objectives are to offer information retrieval and data mining for text and database resources (Medline, full papers, bioinformatics databases). The following objectives are stated in the project proposal:

- Common understanding and framework for mining of relations in large data volumes composed of signals, symbols, text.
- Latent semantic indexing
- To further develop non-linear CCA and related methods and to combine them with existing methods for text processing with applications to e.g. automatic classification of medical diagnosis from patient records and cross-language information retrieval.
- To share in-depth knowledge of the content and structure of EBI databases, as well as knowledge of useful methods and tools for data mining in these databases.

4 Quality Indicators Relevant to WP24

Annex I of the contract defines those quality indicators by which the consortium seeks to assess its progress. Those relevant to WP 24 are listed below.

4.1 Q1 Workshops and symposiums

Participation in national and international conferences

- Organisation of 2 international conferences by members of the WP24 (SMBM 2005, SMBM 2006)
- contributions to 5 international conferences from members of the WP24

4.2 Q2 Sharing of resources and use of research software tools

Software Resources

Goal: At least two software tools developed at one location but used by different partners

Baseline: N/A

Terminological / Lexical Resources

Goal: At least two resources exchanged between partners

Baseline: N/A

- Whatizit (EBI): components used by UKLFR and DIM
- GO categorizer (DIM): assessed by the EBI
- MorphoSaurus (UKLFR): assessed by the EBI

4.3 Q6 Short-and medium-term visits of staff members

Medium-term visits = 2 weeks or longer, for preparation of joint publications, student supervision, conference organization etc.

Short-term visits

Baseline: assumed 0

Goal: all institutions should participate at least one visit per year for WP24.

- 7 Short-term visits exchanges: members of WP24 visiting partners of the NoE
- Organisation of the workshop on textmining in Balatonfuerd in 2005 and 2006
2006: in collaboration with NoE InfoBioMed.

4.4 Q7 Co-authoring of research papers, reports and educational materials

Participants in all institutions at senior researcher, post doc and PhD student levels should be involved with participants from other institutions in the preparation of co-authored scientific papers (original research paper, reviews, conference proceedings etc.) and reports.

Baseline: assumed 0

Goal: all research institutions should have at least 1 and an average of 2 co-authored scientific papers per year

- 4 co-authored publications (2005, quarter Q1 and Q2 in 2006)
- 16 original publications from members of the NoE and participants of WP24

5 Summary of Activities in 2005 / 2006

5.1 Workshops and Symposiums

5.1.1 Workpackage Meetings

The following two meetings with their scope restricted to WP 24 were organized and hosted within the network:

Title	Date	Location	Network Delegates	Comments
WP 24 meeting	11 April 2005	Cambridge, Uk	6	Delegates from EBI, UKLFR, Jena, DIM
WP 24 meeting	28 June 2005	Balatonfuered, Hungary	3	Delegates from EBI, UKLFR, Jena
WP 20/24 meeting	05 Sep 2005	Paris		Delegates from EBI, UKLFR, Jena
WP 20/24 meeting	05 Oct 2005	Freiburg		Delegates from EBI, UKLFR
WP 24 meeting	Dec 2005	EBI	3	Delegates from Jena, EBI collaboration on mapping of terms to biomedical terminologies, PP disambiguation
WP 24 meeting	12 Apr 2006	Jena University	5	Delegates from UKLFR, Jena, EBI
WP 20/24 meeting				Delegates from Freiburg, IFOMIS
WP 20/24 meeting		Saarbrücken		Workshop in Saarbrücken on "Quantities, Numbers and Parts"

5.1.2 International Conference Participation

WP 24 relevant participation in international conferences include:

Title	Date	Location	Network Delegates	Comments
SMBM 2005	10-13	EBI,	U. Hahn,	Hahn: PC co-chair for

	April 2005	Cambridge, Uk	D. Rebholz-Schuhmann, H. Kirsch, M. Arregui	SMBM 2005 (organization of peer reviewing, paper selection, preparation for a special issue of "Bioinformatics" together with Alfonso Valencia, mediator at plenum sessions)
ECCB 2005	29 Sept – 01 Oct 2005	Madrid, Spain	D. Rebholz-Schuhmann, M. Arregui	Presentation of EBIMed, Whatizit and Paella to the public
3 rd Text Mining Symposium	Oct 2005	Fraunhofer Gesellschaft St. Augustin	M. Proprat	Invited paper
EACL 2006, workshop for multidimensional markup	06 Apr 2006	Trente, Italy	D. Rebholz-Schuhmann	Paper presentation
SMBM 2006	10 Apr - 13 Apr 2006	Jena, Germany	D. Rebholz-Schuhmann	Tutorial on Text mining of biomedical literature
ISMB 2006 (forthcoming)	04 Aug -09 Aug 2006	Fortaleza, Brazil	D. Rebholz-Schuhmann	1) Poster presentation on EBIMed + PCorral 2) Software demo on EBIMed, Whatizit + PCorral 3) Paper submitted to the BioLink workshop 4) Bird of Feather session on annotation of semantic types in scientific literature
ECCB 2006 (forthcoming)	10 Sep – 13 Sep 2006	Eilat, Israel	D. Rebholz-Schuhmann	Paper presentation on EBIMed

5.2 Co-authoring

- **Patrick Ruch**, Robert Baud, Christine Chichester, Antoine Geissbühler, Frédérique Lisacek, Johann Marty, **Dietrich Rebholz-Schuhmann**, Imad Tbahriti, Anne-Lise Veuthey. Extracting Key Sentences with Latent Argumentative Structuring Proceedings of MIE2005, **Vol. 116**, [Studies](#)

-
- [in Health Technology and Informatics](#), Edited by: [R. Engelbrecht](#), A. Geissbuhler, C. Lovis and G. Mihalas, August 2005, 1052 pp.
- **Patrick Ruch**, Celia Boyer, Christine Chichester, Imad Tbahriti Antoine Geissbühler, Paul Fabry, Julien Gobeill, Violaine Pillet, **Dietrich Rebholz-Schuhmann**, Christian Lovis, Anne-Lise Veuthey. Using Argumentation to Extract Key Sentences from Biomedical Abstracts. *Int J Med Inform.* 2006. (Accepted for publication)
 - **Dietrich Rebholz-Schuhmann**¹, Graham Cameron¹, Dominic Clark¹, Francesco Beltrame², Jean-Louis Coatrieux³, Eva Del Hoyo Barbolla⁴, Fernando Martin-Sanchez⁵, Luciano Milanese⁶, Ioannis Tollis⁷, **Erik van Mullighan**⁸, Johan van der Lei⁹. SYMBiomatics: Synergies in Medical Informatics and Bioinformatics – exploring current scientific literature for emerging topics. *BITS 2006, Bologna*. (submitted)
 - ¹EMBL-European Bioinformatics Institute, U.K.
 - ²Dist University of Genova, Italy
 - ³INSERM, France
 - ⁴Ministry of Education and Science, Spain
 - ⁵Institute of Health “Carlos III”, Spain
 - ⁶CNR-ITB – Institute of Biomedical Technologies, Italy
 - ⁷Foundation for Research and Technology, Greece
 - ⁸Leyden University, Leyden, Netherlands
 - ⁹Erasmus Medical Center, Netherlands
 - **William Hersh**, Henning Müller, Jeffery Jensen, Jianji Yang, Paul Gorman, **Patrick Ruch**. (2006) ImageCLEFmed: A Test Collection to Advance Biomedical Image Retrieval. *JAMIA*. 2006 (Accepted for publication)

Further publications of work funded by the NoE SemanticMining:

- **Kirsch, H**, Gaudan S, **Rebholz-Schuhmann D**. (2006) Distributed modules for text annotation and IE applied to the biomedical domain. *Int. J. Med. Inform.* 75(6), 496-500.
- **Dietrich Rebholz-Schuhmann, Harald Kirsch**, Goran Nenadic, Sylvain Gaudan, **Miguel Arregui**. (2006) Annotation and disambiguation of semantic types in biomedical text: a cascaded approach to named entity recognition. *Workshop on multidimensional markup with Xml (XMLNLP), EAACL 2006*, Trento, Italy.
- **Dietrich Rebholz-Schuhmann, Harald Kirsch, Miguel Arregui**, Sylvain Gaudan, Mark Riethoven, Peter Stoehr. (2006) EBIMed – Text crunching to gather facts for Uni-ProtKB/Swiss-Prot proteins from Medline. *ECCB 2006*, Eilat, Israel. (Accepted for publication)
- **Dietrich Rebholz-Schuhmann, Harald Kirsch**, Goran Nenadic. (2006) IeXML: towards a framework for interoperability of text processing modules to improve annotation of semantic types in biomedical text. *BioLINK, ISMB 2006*, Fortaleza, Brazil. (submitted)
- **Patrick Ruch**. (2006) Automatic assignment of biomedical categories: toward a generic approach. *Bioinformatics*, 22(6):658-64.
- N. Collier, A. Nazarenko, R. Baud, R. **Ruch**. (2006) Recent advances in natural language processing for biomedical applications. *Int J Med Inform.*, 75(6):413-7.
- **Patrick Ruch**, Imad Tbahriti, Julien Gobeill and Alan R. Aronson. (2006) Argumentative Feedback: A Linguistically-motivated Term Expansion for Information Retrieval *ACL/COLING 2006*.
- **Patrick Ruch**, Laura Perret, Jacques Savoy. (2005) Features Combination for Extracting Gene Functions from MEDLINE. *ECIR 2005*, 112-126
- **U. Hahn**, A. Valencia (Eds.) *SMBM 2005 – Proceedings of the 1st International Symposium on Semantic Mining in Biomedicine*. Hinxton, Cambridge, U.K., April 10-13, 2005. Online-Proceedings: <http://CEUR-WS.org/Vol-148/>
- **J. Wermter**, J. Fluck, J. Stroetgen, S. Geißler, **U. Hahn** (2005) Recognizing Noun Phrases in Biomedical Text: An Evaluation of Lab Prototypes and Commercial Chunkers In: U. Hahn & A. Valencia (Eds.), *Proceedings of the 1st International Symposium on Semantic Mining in Biomedicine – SMBM 2005*. Hinxton, England, U.K., April 10-13, 2005.
- **U. Hahn, J. Wermter**, J. Fluck. (2005) Noun Phrases and Named Entities in Biomedical Texts: Does Domain Change without Retraining Matter? In: *RANLP 2005 – Proceedings of the 5th*

-
- International Conference on Recent Advances in Natural Language Processing*. Borovets, Bulgaria, 21-23 September, 2005.
- **M. Poprat, U. Hahn.** (2005) Enough is enough! – Estimating Upper Bounds of the Size of Training Corpora for Unsupervised PP Attachment Disambiguation. In: *RANLP 2005 – Proceedings of the 5th International Conference on Recent Advances in Natural Language Processing*. Borovets, Bulgaria, 21-23 September, 2005.
 - **J. Wermter, U. Hahn.** (2005) Finding New Terminology in Very Large Corpora. In: *K-CAP 2005 – Proceedings of the 3rd International Conference on Knowledge Capture*. Banff, Canada, October 2-5, 2005. New York/NY: ACM Press.
 - **J. Wermter, U. Hahn.** (2005) Paradigmatic Modifiability Statistics for the Extraction of Complex Multi-Word Terms. In: *HLT/EMNLP 2005 – Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing*. Vancouver, B.C., Canada, 6-8 October 2005.
 - **J. Wermter, U. Hahn.** (2005) Massive Biomedical Term Discovery. In: *DS 2005 – Proceedings of the 8th International Conference on Discovery Science*. Singapore, 8-11 October 2005.
 - **J. Wermter, U. Hahn.** (2005) Effective Grading of Termhood in Biomedical Documents. In: *AMIA 2005 Symposium – Proceedings of the Annual Symposium of the American Medical Informatics Association. Biomedical and Health Informatics: From Foundations to Applications to Policy*. Washington, D.C., USA, October 22-26, 2005. Philadelphia/PA: Hanley & Belfus.

5.3 Joint Research Programme

The joint research programme comprises (1) mining of relations in large data volumes of text, which includes semantic indexing with or without use of precompiled terminological resources, e.g. for cross-lingual information retrieval, (2) evaluation of the results of information retrieval and text mining and (3) data mining in EBI's databases including text resources.

Several results have been generated by members of the WP24 over the past year. The results can be categorized as:

- Cat-1: Pure scientific results (refer to list of publications in section 5.2)
- Cat-2: Information retrieval and information extraction solutions developed as demonstrators to ongoing research work
- Cat-3: Literature and database analyses that contribute to new scientific results
- Cat-4: Organisation of international competitions to improve progress in the research field

5.3.1 EBI's text processing engines (Cat-2)

The Rebholz group (EBI) has established online services ([Whatizit](#), [EBIMed](#)), which analyze biomedical documents. Whatizit processes any type of text data that has been submitted via cut&paste or http request. It identifies contained terminology and links it to biomedical databases [1].

EBIMed combines these information extraction capabilities with a retrieval engine based on Lucene technology [2].

Design principles for EBIMed

The information contained in Medline abstracts is conveyed to a great part by biomedical terminology such as protein names and GO terms. It is EBIMed's goal to make this information accessible by extracting, ranking and organizing these key terms.

EBIMed labels UniProtKB/Swiss-Prot protein name in the text if it co-occurs with another UniProtKB/Swiss-Prot protein name, a GO term, a drug or species name. Such co-occurrences can be interpreted as two proteins being involved in the same biological process (e.g. protein-protein interactions), as functional annotations (GO annotations), as proteins being targeted by drugs (drug-protein relations) and as proteins of model organisms. Other types of terminology could be integrated. This would require better understanding how these terms (e.g. diseases) are related to proteins and how well the terminology is supported.

The user can either submit a keyword query or a list of PubMed identifiers (PMIDs) to start the process. The query terms are not used as parameters for the analysis and therefore need not contain any protein name, drug, species or GO term. The only dependence between the query and EBIMed's analysis is the set of retrieved abstracts.

Terms that occur in the same sentence form a **pair** (co-occurrence). Pairs are uniquely identified and may appear in different sentences across the retrieved abstracts. All pairs from all sentences are gathered, sorted, unified to pairs of concepts, ranked by their frequency and presented in a table (fig. 1). For each pair in the table, a link is provided to a list of sentences containing the pair. Each sentence is linked to its original abstract.

Initially, the leftmost column of the table lists a UniProtKB/Swiss-Prot protein and all other columns to the right list the co-occurring concepts (e.g. protein names, GO terms, drug names, species names) that form a pair with the protein. A column to the right will become the leading leftmost column as soon as the user selects it with a mouse click on the header of the column. The content of the table is then reorganised to match the concepts in the leftmost column. Above the table a display shows the total number of abstracts analysed and a list with the number of pairs encountered for the different types of terms.

In principle, the number of pairs increases with the number of retrieved abstracts and the number of identified terms from the domain of molecular biology. For example, the query *Wnt* currently retrieves 4,675 abstracts¹ with 3,275 listed UniProtKB/Swiss-Prot proteins, RNAi leads to the selection of 2,511 abstracts dealing with 2,821 identified proteins, whereas the gynecological treatment 'cerclage' induces the retrieval of 1,478 abstracts with only 80 UniProtKB/Swiss-Prot entries.

EBI will assess the quality of the IR/IE engines in collaboration with the curation teams at the EBI and with the partners from the NoE [3].

For the assessment we manually evaluated results returned by EBIMed in different analyses (results not shown), e.g. precision of identified terms, precision of identified protein-protein interactions and precision of drug-protein relations. Finally we assessed the retrieval of facts for documented protein/protein interactions from the *Wnt* pathway. As representatives we chose protein pairs that have been described in KEGG as well as in STKE, where a pair is either two nodes linked by an edge or two nodes side by side. We identified 108 unique pairs and 10 pairs common to both sources. We measured how many of these pairs were reproduced by EBIMed upon the query *Wnt*. In addition we

¹ Date of retrieval: 12th March 2005

randomly selected 4 interaction pairs out of the 10 pairs confirmed in both sources (Not shown) and used both protein names in conjunction with Wnt for a combined query such as Wnt AND APC AND PP2A to measure the coverage of this relation in Medline.

To summarise EBIMed extracts protein names at a rate of more than 90% precision (accepting false boundaries of nested terms, if the term is used with the correct meaning), while 37% of the extracted protein pairs and 50% of the drug/protein pairs represent a meaningful interaction. For the interaction pair Dkk and LRP of the Wnt pathway we were able to clarify the interactions between the subtypes of Dkk and LRP, which are neither documented in KEGG nor in STKE.

Altogether, the use of EBIMed leads to a better access to statements in Medline in comparison to PubMed because the user reads mainly relevant sentences. In the case of the four evaluation examples, 110 sentences were automatically selected from 52 abstracts. 43 out of the 110 sentences carried relevant information (39% precision), which did not require the user to read the remaining information in the abstract. This is an improvement of 16% over the baseline precision of 13% that results from completely reading all 52 abstracts².

5.3.2 GO categorizer³ (Cat-2)

DIM designed a generic text categorization system to automatically assign biomedical categories to any input text (GO categorizer, cf. Ruch 2006). Unlike usual automatic text categorization systems, which rely on data-intensive models extracted from large sets of training data, our categorizer is largely data-independent. In order to evaluate the robustness of our approach we test the system on two different biomedical terminologies: the Medical Subject Headings (MeSH) and the Gene Ontology (GO). While the former is a general vocabulary for life science, the latter is particularly important in proteomics as it is used by leading online databases for molecular biology (Swiss-Prot, ENTREZ\Gene...). Our lightweight categorizer, based on two ranking modules, combines a pattern matcher and a vector space retrieval engine, and uses both stems and linguistically-motivated indexing units. Results show the effectiveness of phrase indexing for both GO and MeSH categorization, but we observe the categorization power of the tool depends on the controlled vocabulary: precision at high ranks ranges from above 90% for MeSH to less than 20% for GO, establishing a new baseline for categorizers based on retrieval methods. Significant improvements are expected for the GO categorization.

GO categorizer is currently tested at the EBI for its performance. Application notes to advertise the service are being prepared. Screenshots of the system's output are given in section 7.1.2.

As future work, the browser will be tuned to navigate two new terminologies: the SNOMED CT (Systematized Nomenclature of Medicine – Clinical Term) and the WHO-

² On average a Medline abstract contains 6.16 sentences (191 sentences in 31 abstracts, tbl. 1), which leads to 320 sentences in 52 Medline abstracts. The user has to read the complete abstract to have 100% recall on all UniProtKB/Swiss-Prot protein co-occurrences and interactions. This leads to the result that the user has to read 320 sentences, to find all 43 identified sentences. This results to 13% precision.

³ 129.194.97.165:8081/EAGL/

ART (World Health Organization – Adverse Drug Reactions Terminology). This work will benefit from cooperation and synergies between WP24 and WP22.

5.3.3 Integration of text resources and biomedical databases (Cat-3)

EBI offers access to a large number of databases. Each database makes use of domain terminology and even ontologies. Researchers in the biomedical domain and curators of biomedical databases seek support in cross-linking facts from the scientific literature to the scientific databases for immediate information retrieval. [Whatizit](#), [EBIMed](#) and [Paella](#) are online services from the Rebholz group, which match this need. In the same vein, the GO browser designed by the DIM, allow to directly link textual contents to the Gene Ontology. Both services provide a set of information extraction modules, which identify protein and gene names (UniProt, LokusLink), GO terminology, as well as mutations and protein-protein interactions and links the extracted named entities to the database entries.

5.3.4 Preparation of cross-lingual information retrieval (Cat-3)

After the preliminary assessment of the Morphosaurus representation of medical terms, the decision was taken to use Morphosaurus for cross-lingual information retrieval on MEDLINE abstracts. Morphosaurus will be used to normalize the index to MEDLINE and any incoming query.

UKLFR and Jena is assessing the mapping of MeSH thesaurus to Morphosaurus and the advantages for information retrieval. DIM will aim at delivering a multilingual retrieval MEDLINE demonstrator, using UMLS and WP20 resources; in particular French and English resources provided by the DIM.

5.3.5 Identification and disambiguation of biomedical semantic types in the scientific literature (Cat-3)

Identification of named entities (NEs) in a document can be viewed as a three-step procedure. In the first step, single or multiple adjacent words that indicate the presence of domain concepts are recognised (term identification). In the second step, recognised terms are classified into broader domain classes (e.g. as genes, proteins, species) called term categorisation. The final step is mapping of terms into referential databases. The first two steps are commonly referred to as *named entity recognition (NER)*.

NER in the biomedical domain profits from large existing terminological resources, which are either provided as ontologies (e.g. Gene Ontology, ChEBI, UMLS) or result from biomedical databases containing named entities (e.g. UniProt/Swiss-Prot [4,5,6]). However, terminological variation, recognition of boundaries of multiword terms, identification of nested terms and ambiguity of terms are the difficult issues when mapping terms from the literature to biomedical database entries [7,8].

Combining sets of terms from different terminological resources (e.g. biomedical databases) leads to naming conflicts such as homonymous use of names and terminological ambiguities. The most obvious problem is when the same span of text is

assigned to different semantic types (e.g. ‘*rat*’ denotes a species and a protein). In our case we have to consider three types of ambiguities:

(Amb1) A name is used for different entries in the same database, e.g. the same protein name serves for a given protein in different species [9].

(Amb2) A name is used for entries in multiple databases and thus represents different types, e.g. ‘*rat*’ is a protein and a species.

(Amb3) A name is not only used as a special biomedical term but also as part of common English in contrast to biomedical terminology, e.g. ‘*who*’ and ‘*how*’, which are indeed used as protein names.

The first type of ambiguity is not resolved in our text processing solution. Instead, for a given biomedical term links to all entries referring to this term in the same database are kept.

One approach to the disambiguation of **Amb2** and **Amb3** would be to integrate all terms into one massive dictionary, identify the strings in the text and then disambiguate between n semantic types. This would require that the disambiguation module be trained to distinguish all semantic types. If a new type is added, the disambiguation module has to be retrained.

In our modular approach, all terms of a semantic type are kept in a separate dictionary. After identification of a term in the text, disambiguation only decides whether the term is of that type or not. If it is not, the annotation is removed and allows downstream modules to tag the term differently. While this requires n disambiguation steps, adding new types is independent of modules already present.

This approach to finding names can create three types of ambiguities mentioned above in Section 2.

In the current implementation, **Amb1** is not resolved. Rather, the links to *all* entries in the same database are maintained. **Amb2** and **Amb3** are partially resolved for protein/gene names as explained below. Note that **Amb2** is resolved on first-come first-serve basis, meaning that an annotation introduced by one module is not overwritten by a subsequent module.

Many protein names are indeed or at least look like abbreviations and it has been proved that ambiguities of abbreviations and acronyms found in Medline abstracts can be automatically resolved with high accuracy [10]. Three possible results are produced: name is an acronym and can be resolved as (a) a protein, (b) not a protein, and (c) name cannot be resolved. In case of (c), the frequency of the name in the British National Corpus (BNC) is compared with a threshold. If the frequency is higher than the threshold, the name is assumed not to be a protein name. The threshold was chosen not to exclude important protein names that have already entered general English (such as *insulin*).

The first module performs the matching and indiscriminately assumes each match to be a protein name. The name is marked up as a protein. The markup also contains the frequency of that name in the BNC. The next module marks up all known acronym expansions in the text and combines the two pieces of information: a marked up protein

name is looked up in the list of abbreviations. If the abbreviation has an expansion that is marked up in the vicinity **and** denotes a protein name, the abbreviation is verified as a protein name (case (a) above) by adding an attribute with a suitable value to the NE tag. Similarly, if the expansion is clearly not a protein name, the same attribute is used with according information.

The annotation includes the normalised form of the acronym, which serves as an identifier for further database lookups.

Only the module M-5 removes the protein name markup if the name is either (b) clearly not a protein, or in case (c) has a BNC frequency beyond the threshold.

5.3.6 Towards a common annotation framework for semantic types (Cat-1)

Prepared for discussion at the ISMB 2006 in Fortaleza, Brazil.

We propose a task-oriented data exchange framework that facilitates interoperability for annotation of semantic types, and provides a basis for other more complex annotations (e.g. event annotations). It includes the basic principles for marking up biomedical text to enable interoperability through a 'common exchange format', where various levels of mark-up are gathered within the document.

5.3.6.1 Basic principles for annotations provided by TM/IE modules

We have identified a set of tasks (and corresponding modules) that may be used to support annotation of semantic types. These include: sentence finder, tokeniser, POS tagger and term tagger. The modularisation used here may be more fine-grained than encountered in some TM/IE systems today. This supports a precise specification of tasks as well as input/output behaviours, and should not hinder merging modules (e.g. tokenisation and POS tagging) as long as they match the requirements.

Our framework is based on a basic (minimal) set of conceptual tags and attributes that each component should provide. However, this should not preclude any module from producing other tags or attributes, nor should the tags represent any rigid meaning. As an example, we propose how a token should be annotated to allow for the reuse of tokenisers, but we do not attempt to define what a token is, so not to restrict the choice in available tokenisers.

Any IE/TM module adds or changes tags in an input document to produce a result document by generating at least a set of basic elements for that type. We distinguish between the following basic types of annotation elements that need to be annotated:

- **tokens** (*w* elements) are minimal individual constituents of text; apart from words, numbers etc., these include punctuations, references, links, inline formulas, dates, measurements, etc. They do not contain any of the other three elements, and have *c* attributes that specify their category (e.g. POS information).
- **terms** (*e* elements) comprise one or more tokens and have a possibly ambiguous semantics denoting concepts, objects or entities. This includes names, terminology and nomenclature strings or sequence mutations. Attributes are used to attach semantics (*sem* attribute) and POS information (*c* attribute). They can contain token and term elements only.

- **chunks** (*c* elements) denote syntax units, such as noun, verb phrases or prepositional phrases. They can contain token, term and chunk elements only, and their type is assigned via a *t* attribute.

- **sentences** (*s* elements) refer to sentence elements. They can contain any of the above types.

Given the above requirements, a brief summary of the descriptions of the identified tasks is as follows. A class of sentence finders provides *s* elements to the content of the input document. Tokenizers are responsible for providing *w* elements to an input document, while NER and term recognisers annotate documents with *e* elements and specify a mandatory *sem* attribute that points to semantic types (see also below). A POS-tagger generates *c* attributes for any *w* and *e* elements found in a document.

Furthermore, we demand that the original document should be completely recoverable from any result document. This enables client applications to display IE/TM annotations as semantic enrichment of documents.

5.3.6.2 POS and term tagging – an annotation example

Although there are notable differences between *terms* and *named entities*, they are typically used uniformly in further processing steps and are referred here as *terms*. Therefore, we suggest a common element (*e*) to denote terms, named entities and any other domain-specific sequences (such as mutation notations) that need a semantic attribute. Similarly, we refer to ATR (automatic term recognition) and NER modules as *term taggers*. If we apply the above mentioned principles in detail to a software component that is classified as a term tagger, then the software component has to adhere to the following requirements: (1) Domain-specific terms are marked up as *e* elements with a mandatory attribute *sem* to specify the semantics of the term. (It is our intention to allow a yet unspecified set of possible values for the *sem* attribute.) The term tagger may assign POS-information as a *c* attribute to a term. (2) If the term tagger is applied to tokenised text (text containing *w* elements), each term must contain at least one token. If it is applied to un-tokenised text, each term must contain tokenisable text that would result in at least one token if tokenised. (3) The input document may already contain *e* elements. Terms may be nested.

In the case of a POS-tagger, the input to a POS-tagger is a tokenised document, and POS information is encoded with the attribute *c* assigned to each *w* element. This attribute is also added to every term element. The POS-tagger may add POS information to tokens that are inside terms. Also, the POS-tagger may change POS information in tokens or terms that have already assigned POS information. The attribute value may be the empty string to signal that the POS tagger does not know which POS applies.

In general, this framework allows for modification of the order of applied text processing modules; only in the case of the POS-tagger we require that it receives a tokenised input document. This does not exclude modules that merge a tokenizer with a POS-tagger at the same time. Furthermore, our framework leaves enough flexibility to allow, for example, introduction of *c* attributes before the POS tagger is applied, to have tokens with or without POS information inside of terms and to transform multiple tokens into a single token representation (“multi word tokens”).

5.3.7 Medical image CLEF⁴ (Cat-4)

Information retrieval is not restricted to textual data only. Image data is increasingly important to the bioinformatics domain and had high importance to medical informatics in the past. Currently initiatives are on its way to allow standardized information retrieval including image data. This requires reliable assessment of results and in the best case public competitions on this information retrieval task.

WP24 (DIM, Patrick Ruch) has an ongoing collaboration with William Hersh from the Oregon Health Science University (funded by the NSF). This collaboration led to the successful organization of the medical ImageCLEF challenge. A sample of participants with the methods used to perform the multilingual and multimodal runs is given in the following table 2. Summary based on Hersh and al. 2006:

Background: Biomedical users of information retrieval systems are increasingly interested in searching for images. However, image retrieval is much less advanced than text (document or Web page) retrieval. We developed an image retrieval test collection for use by ourselves and other research groups participating in the image retrieval task of Cross-Language Evaluation Forum. The analysis in this paper assessed the results obtained from 13 different research groups who employed a variety of image retrieval techniques.

Methods: After individual research groups obtained and assessed results from their systems, we analyzed the results of all groups, looking for common themes and trends. In addition to overall performance, results were analyzed on the basis of topic categories (those most amenable to visual, textual, or mixed approaches) and run categories (those employing queries entered by automated or manual means as well as those using visual, textual, or mixed indexing and retrieval methods). We also assessed results on the different topics and compared the impact of duplicate relevance judgments.

Results: A total of 13 research groups participated. Analysis was limited to the best run submitted by each group in each run category. The best results were obtained by systems employing both visual and textual methods. There was substantial variation in performance across topics. Systems employing textual methods were more resilient to visually oriented topics than those using visual methods were to textually oriented topics. The primary performance measure of MAP was not necessarily associated with other measures, including those possibly more pertinent to real users, such as precision at 10 or 30 images.

Conclusions: We developed a test collection for image retrieval amenable to assessing visual as well as textual methods for image retrieval. Future work must focus on further assessment of topic and run types. Users studies are necessary to determine the best measures for image retrieval. A call for the 2006 edition of the event has been published and 36 participants are already registered, including prestigious corporate research groups such as Microsoft China.

⁴ ir.ohsu.edu/image/

5.4 Exchange Visits

Further visits are planned between collaboration partners DIM, UKLFR and Jena in October.

5.5 Follow-up Grant proposal

5.5.1 IST funded project “BOOTStrep”⁵

EBI and Jena have prepared a grant proposal to the EC’s IST program. The project proposal is called BOOTStrep and is a Strep with 8 partners including EBI, Jena and UKLFR. The project has been accepted by the CEC and started in April 2006.

The project proposal has been supported by the collaborative work done between EBI, Jena and UKLFR as part of the NoE SemanticMining and the WP24, WP14 and WP15. Furthermore the project will induce benefits to the WP24 of the NoE.

Mission:

”BOOTStrep (Bootstrapping Of Ontologies and Terminologies STRategic REsearch Project) is funded in the EC’s 6th Framework Programme. The project will pull together already existing biological fact databases as well as various terminological repositories and implement a text analysis system which continuously increases their coverage by analysing biological documents.”

5.5.2 @neurIST Project⁶

The @neurIST project is a project amongst UKLFR and DIM. It brings together different data resources to support disease management of cerebral aneurysm.

Mission:

”@neurIST seeks to provide channels for the integration of all data sources on cerebral aneurysm. It has three work packages dedicated respectively to the collection, processing and integration of these data. (...) The primary theme of @neurIST is to develop vertical integration across data structures and across length scales, but horizontal integration at every level of abstraction, from access to information sources, to complex information processing, knowledge representation, structuring and fusion will cement the collaboration between the disciplines. (...) Furthermore the approach will be extendable to other disease processes and scalable to federate a large number of clinical centres and public databases.”

5.6 New collaborations established throughout 2005 and 2006

5.6.1 SYMBiomatics⁷

This project is a Specific Support Action (SSA) funded by the European Commission.

Participating members are:

- (1) EMBL-European Bioinformatics Institute, U.K.
- (2) Dist University of Genova, Italy

⁵ www.bootstrep.eu

⁶ www.aneurist.org/

⁷ www.symbiomatics.org

-
- (3) INSERM, France
 - (4) Ministry of Education and Science, Spain
 - (5) Institute of Health “Carlos III”, Spain
 - (6) CNR-ITB – Institute of Biomedical Technologies, Italy
 - (7) Foundation for Research and Technology, Greece
 - (8) Leyden University, Leyden, Netherlands
 - (9) Erasmus Medical Center, Netherlands

Mission statement of Symbiomatics (shortened):

“Bioinformatics and medical informatics are both rapidly advancing fields. (...) The SYMBIOmatics Specific Support Action (SSA) (...) seeks to identify and exploit synergies between bioinformatics and medical informatics as well as identifying addressable challenges for the medium term future. The project will document the state-of-the-art in biomedical informatics in Europe and identify areas of new opportunity. This will be done by systematically identifying European expert (...). Simultaneously, bibliometric and data-mining methods will identify and analyse the content of the relevant scientific literature. Areas of opportunity will then be documented and prioritised. (...) A White Paper summarising the findings will be completed by Nov 2006 and will provide input to future European scientific and funding policy.”

Members of the WP24 have contributed to the SYMBiomatics project in terms of a literature analysis to identify synergies between the research domain medical informatics and bioinformatics. The results from this analysis have been submitted as a publication to the BITS 2006 conference in Bologna. This collaboration has as well contributed to the newly established collaboration with members of the NoE InfoBioMed (see below).

5.6.2 Network of Excellence “InfoBioMed”⁸

A collaboration between WP24 of the NoE “SemanticMining” and members of the NoE “InfoBioMed” has been established, i.e. between EBI (D. Rebholz-Schuhmann) as member of the NoE “SemanticMining” and the Medical Center at the Erasmus University of Rotterdam (J. van der Lei, Erik van Mullighan). Collaborative efforts are concerned with the exchange of data, for example curated data on nuclear receptors and transcription factors (to be integrated into the knowledge base of InfoBioMed) and drug related information extracted from the scientific literature.

Further collaborative effort has been established as part of the co-organisation of the summer school at lake Balaton, Hungary. Both members from both NoEs and in particular members of the WP24 are involved in organizing tutorials for attendees to the summer school. For example Udo Hahn, Dietrich Rebholz-Schuhmann and Erik van Mullighan organize the text mining tutorial.

Mission of InfoBioMed

“(…) The EC-funded BIOINFOMED Study (EC-IST 2001-35024) has recently carried out a prospective analysis of the relationships and synergy between Bioinformatics (BI) and Medical Informatics (MI). (...) Biomedical Informatics (BMI) is the emerging discipline that aims to put these two worlds together so that the discovery and creation of novel diagnostic and therapeutic methods is fostered.”

“The specific objectives of INFOBIOMED are:

⁸ www.infobiomed.org

-
1. To enable systematic progress in clinical and genetic data interoperability and integration.
 2. To advance the exchange and interfacing of methods, tools and technologies used in both MI and BI.
 3. To enable pilot applications in particular fields that demonstrate the benefits of a synergetic approach in BMI.
 4. To create a European BMI community that extends beyond the proposed core network to serve as an open forum for dialogue between the actors involved.
 5. To widely spread the knowledge acquired and developed in the framework of the network to the scientific community, healthcare professionals, citizens, industry, authorities and other stakeholders.
 6. To enable a robust framework for education in BMI, as well as training and mobility of involved researchers that allows for the creation of a solid European BMI research capacity.
 7. To create a long-lasting, self-sustainable structure in the European BMI field.”

5.7 Mini-Symposium on Semantic Enrichment of the Scientific Literature

Results from information retrieval and information extraction solutions improve significantly under the conditions that good terminological resources are available, that the use of terms in the scientific literature follows known standards and that in any case of potential ambiguity the contextual information or other additional information contained in the text document supports proper disambiguation. Unfortunately scientific literature does not adhere yet to any of the three conditions leading to demands in improving the current state of scientific documents to better standards. Such a standard could be any semantic enrichment of scientific literature.

Semantic enrichment is the integration of metadata into text. It allows facts within text to be annotated and linked to electronic resources such as the biological databases maintained by the EMBL-EBI and its collaborators. This makes free text better suited for automatic exploitation in text mining, information retrieval and presentation.

As part of the WP24 activities we organised the one-day mini-symposium on “Semantic Enrichment of Scientific Literature”. The symposium took place at the EBI on February 20th. It was attached to the regular meetings of the Industry Programme at the EBI.

We invited key players in the text-mining field, representatives from STM publishing, representatives from other industries who share an interest in making the most of semantic enrichment, researchers and curators to exchange their views on automatic annotation of scientific text and to better understand their requirements. The meeting was used to exchange views on automatic annotation of scientific text and to help understand the requirements of curators, researchers, authors, publishers and other interested fields.

Presentations:

- Michael Ashburner: The impact of biomedical terminologies on curation and publishing.⁹
- Olivier Bodenreider: NLM Resources for Semantic Enrichment.¹⁰
- Matthew Cockerill: Practical benefits of semantically-enriched articles for authors and readers.¹¹
- Henning Hermjakob: IntAct: Covering Molecular Interaction Space.¹²

⁹ www.ebi.ac.uk/~rebholz/SemanticEnrichmentLit/MichaelAshburner_SESL.pdf

¹⁰ www.ebi.ac.uk/~rebholz/SemanticEnrichmentLit/OlivierBodenreider_SESL.pdf

¹¹ www.ebi.ac.uk/~rebholz/SemanticEnrichmentLit/MatthewCockerill_SESL.pdf

¹² www.ebi.ac.uk/~rebholz/SemanticEnrichmentLit/HenningHermjakob_SESL.pdf

-
- Barend Mons: Semantic support technology for on-line discovery and distributed annotation.¹³
 - Peter Murray-Rust
 - Martin Romacker: Navigating the Knowledge Space: Concept Identification in Texts and Referencing across Data Sources.¹⁴
 - Jasmin Saric: Lessons learned from collecting kinetic data.¹⁵
 - Stefan Geissler: Recognition of Chemical Entities in Biomedical Literature.¹⁶
 - Alfonso Valencia: Linking database information and textual sources.¹⁷
 - Anne-Lise Veuthey: Metadata in the biomedical literature: UniProtKB curators' requirements.¹⁸

6 Workpackage Assessment

6.1 Assessment against Quality Indicators

The contribution to national and international conferences is high. Members of WP24 contributed to the organisation of 2 international conferences that were organized through the NoE SemanticMining (SMBM 2005 and SMBM 2006). In addition submitted manuscripts describing ongoing work in the NoE have been accepted to important international conferences (EACL 2006, ECCB 2006). Apart from the conferences, 4 publications have been co-authored amongst members of WP24 and 15 publications have been produced and submitted, of which 13 publications have already been accepted.

As part of the work in WP24 several software solutions have been made available to the public. All solutions will be further assessed and improved in the future.

Altogether EBI, DIM, UKLFR and Jena University show significant and highly beneficial exchange of research work, know-how and collaborative efforts.

6.2 Assessment Against Workpackage Objectives

6.2.1 Exchange of Methods and Resources

Resources developed as part of WP24 have been exchanged between the work package partners. The Whatizit server has been reused by DIM. DIM and EBI are assessing the MorphoSaurus for integration into their solutions. GO categorizer is currently assessed by the EBI for integration.

This exchange of techniques improves progress in the different research teams and allows integration of state of the art technology amongst all members.

6.2.2 Integration with other Workpackages

Although not specified in the original work program, we see interesting prospects of linkage to the following workpackages:

¹³ www.ebi.ac.uk/~rebholz/SemanticEnrichmentLit/BarendMons_SESL.pdf

¹⁴ www.ebi.ac.uk/~rebholz/SemanticEnrichmentLit/MartinRomacker_SESL.pdf

¹⁵ www.ebi.ac.uk/~rebholz/SemanticEnrichmentLit/JasminSaric_SESL.pdf

¹⁶ www.ebi.ac.uk/~rebholz/SemanticEnrichmentLit/StefanGeisslerTemis_SESL.pdf

¹⁷ www.ebi.ac.uk/~rebholz/SemanticEnrichmentLit/AlfonsoValencia_SESL.pdf

¹⁸ www.ebi.ac.uk/~rebholz/SemanticEnrichmentLit/AnneLiseVeuthey_SESL.pdf

-
- WP 20 (Multilingual Medical Dictionary): The Morphosaurus (<http://www.morphosaurus.net>) indexer is, principally, neutral with regard to specific retrieval environment or search engines. However, search engines should be optimized to best support document retrieval mediated by subword identifiers. An agreement was made with WP 24 to use Morphosaurus for indexing Medline abstracts. Evaluation of the Morphosaurus is planned for October 2005 at the EBI.
 - WP 22 (SNOMED CT): cooperation with WP22 has started to deliver the SNOMED browsing tool. Cooperation with WPs 21 and 26 are investigated for the delivering and testing of the WHO-ART browser.

7 Tables and Figures

7.1.1 EBIMed

EBIMed | | | [Advanced Search](#) | [Query Syntax](#) |

Summary | 153.543 seconds

3656 Abstracts

Type	Hits	HitPairs
Uniprot	2708	39814
Cellular component	111	2932
Biological process	334	7177
Molecular function	56	1183
Drug	105	1073
Species	233	7043
Total	3547	59222

HitPair table

- You can explore a total of 39814 permutations for this HitPair table arrangement. Click on the secondary columns' headers to rearrange the table.
- Rows 1 to 5 (out of 2629).

first << 1/526 >> last

Uniprot	Uniprot	Cellular component	Biological process	Molecular function	Drug	Species
beta-catenin (score: 6653)	APC or APCs (240/428)	nucleus (132/176) cytoplasm (81/89)	Transcription (341/449) development (201/247)	binding (183/241) DNA binding (19/22)	Lithium (23/32) thyroid (9/30)	cancers (253/423) humans or man or Homo sapiens (210/270)
	GSK-3 beta or glycogen synthase kinase-3 beta (154/198)	intracellular (61/72) plasma membrane or cell membrane or cytoplasmic membrane (39/51)	phosphorylation (157/238) localization (129/182)	kinase activity (4/4) cadherin-binding (3/4)	chondrocytes (9/22) retinoic acid (7/7)	Xenopus (117/149)
	Axin or axins (145/259)	membrane (37/49) adherens junction (27/34)	transduction (102/117) cell adhesion (67/80)	protein binding (3/3) mitogen-activated kinase (2/2)	anti-inflammatory drugs or indomethacin (5/12)	Armadillo (107/150)
	E-cadherin (97/162)	apoptosis (41/71)	cell-cell adhesion (45/49)	E2 (2/2)	modular or monomeric (4/6)	mouse or nude mice or transgenic mice or Mus musculus (106/146)
	cyclin or cyclins (89/142)	extracellular or extracellular regions (16/18)	cell proliferation or cells proliferation (35/42)	MMP-9 or MMPs-9 (2/2)	etodolac or Sulindac or Ibuprofen (3/9)	axis (85/124)
	Wnt-1 or Wnts 1 (73/133)	cytoskeleton (13/13)	pathogenesis (23/24)	GPCR (2/2)	caffeine or aspirin (3/6)	Drosophila (75/79)
	Lef or Lefs (64/94)	transmembrane (12/12)	embryogenesis (20/22) morphogenesis (18/22)	PKG (1/2) SAPK (1/2)		

Fig. 1 (EBIMed summary display): The keyword query ‘Wnt’ leads to the retrieval of 3,656 abstracts referring to 2,708 Uniprot proteins and covering 39,814 hit-pairs of Uniprot proteins, i.e. co-occurrences of Uniprot proteins in single sentences.

The hit-pair table lists ‘beta-catenin’ in the left-most column, called the primary column. All other columns are secondary columns and each of their entries refers to the entry in the primary column (hit-pair), e.g. “APC” refers to “beta-catenin” as a hit-pair. Selecting the header of any secondary column induces that it becomes the primary column including rearrangement of all hit-pairs.

The two numbers after the hit-pair term indicate how many abstracts and sentences, respectively, contain this hit-pair. The numbers link to the list of sentences, where each sentence is shown with the abstract’s PMID, its first author and year of publication and again links to the full abstract. Abstract and sentences are marked up with identified terminology, which also links to the terminological or database resource from which the terms were taken.

EBIMed: <http://www.ebi.ac.uk/Rehholz-srv/EBIMed>

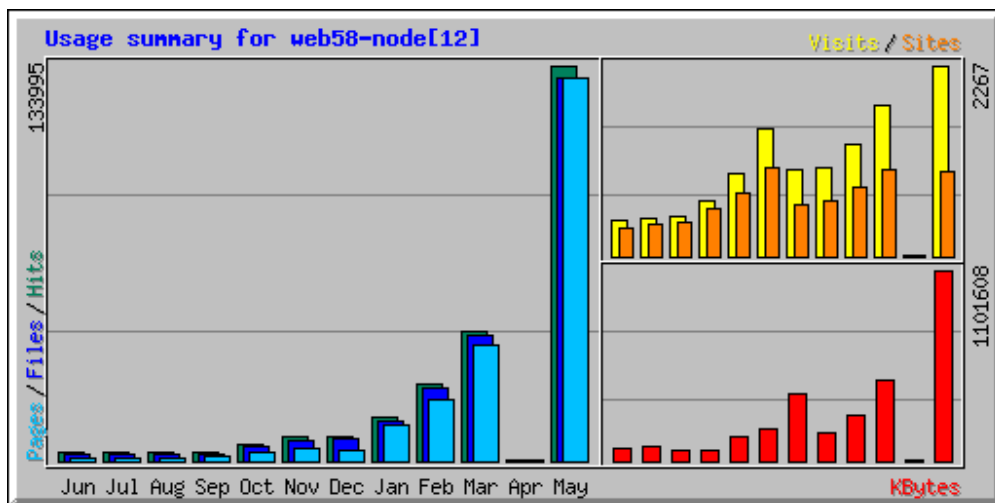


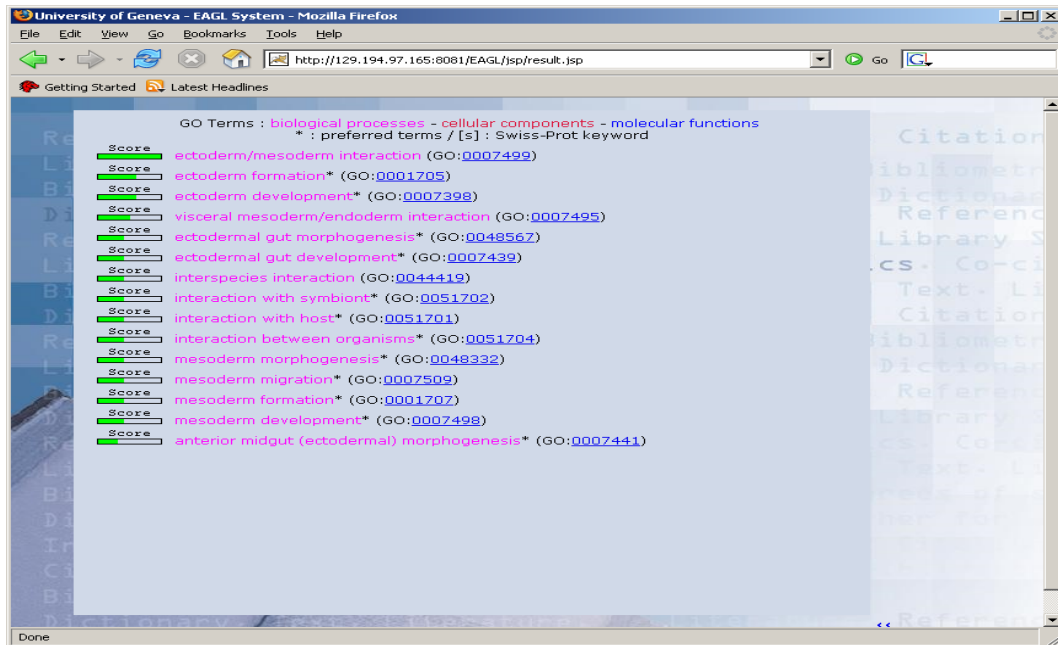
Fig. 2 (EBIMed and Whatizit usage over time): Requests to the information extraction infrastructure of the EBI is coming from roughly 1,000 different sites. Over the past 2 months the amount of submitted data has strongly increased (April not shown due to a technical fault in the analysis of the Log files).

Type of analysis	Type of term	Term used in correct sense	Term used in correct sense	Term incomplete	Term has other sense	Total
(No. of abstracts containing sentences / No. of unique sentences analysed)		Nested terms admitted	Nested terms excluded			(100%)
Analysis-1: Wnt query. Analysis of first 31 abstracts (31/191)	Protein	280 (92%)	165 (54%)	3 (1%)	23 (7%)	306
	Species	91 (76%)	89 (72%)	0	29 (24%)	120
	GO term	83 (95%)	47 (54%)	4 (5%)	0	87
Analysis-2: Wnt query. Evaluation of protein/protein relations (37/94)	Protein	250 (94%)	214 (81%)	12 (5%)	4 (2%)	266
Analysis-3: Wnt query. Evaluation of drug/protein relations (37/118)	Protein	114 (90%)	99 (78%)	5 (4%)	8 (6%)	127
	Drug	99 (74%)	62 (47%)	6 (5%)	28 (21%)	133
Analysis-4: Evaluation of protein/protein relations from the Wnt pathway (60/110)	Protein	242 (100%)	231 (95%)	0	0	242

Tbl. 1. We assessed the performance of EBIMed to estimate the precision of term identification in 4 different analyses. In analysis-1 (31 abstracts, 191 sentences) the precision for UniProtKB/Swiss-Prot proteins was 92% (recall 94%, not shown). If we did not count nested terms as correct (e.g. HZF-1 nested in 'HZF-1 ortholog', grey column), the precision was 54% (recall 55%, not shown). Overall precision for the identification of protein names varied between 90% and 100%. The correct species has been identified at 76% precision. GO terms were identified at 95% precision, mainly due to the fact that exact term matching was applied [1]. (Note: Results in analysis-2 sum up to 101% due to rounding.)

7.1.2 Screenshots from the GO Browser/GO Categorizer

The below figure shows the output of the GO browser, given queries such as «interactions between ectoderm and mesoderm» or «ectodermal and mesodermal relationships »:



The below figure shows the output of the GO categorizer (given the abstract in PMID=12604344). The hyperlink to QuickGO, the EBI visualisation tool for GO, is available to access the ontology from a hierarchical perspective. This example illustrates the ability of the system to attribute categories (“response to uv”, in green at the bottom of the figure), although the term does not appear explicitly in the input abstract.



7.1.3 Medical Image CLEF

Institution	Group Code	Country	Runs	Brief description of runs submitted
CEA [19]	CEA	France	5	All runs automatic with two visual only and three mixed runs. Techniques used included the PIRIA visual retrieval system and a simple frequency-based text retrieval system.
U.Concordia - Computer Science [20]	CINDI	Canada	1	One visual run containing a query only for the first image of every topic using only visual features. Technique applied was an association model between low-level visual features and high-level concepts mainly relying on texture, edge, and shape features.
U. and U. Hospitals Geneva [21]	GE	Switzerland	19	All automatic runs, including two textual, two visual runs, and 15 mixed runs. Retrieval relied mainly on the GIFT (visual) and easyIR (textual) retrieval systems.
Inst. Infocomm Research	I2R	Singapore	7	All automatic visual runs; first they manually selected visually similar images to train the features and then used a two-step approach for visual retrieval.
Institute for Infocomm Research [22]	i2r	Singapore	3	All runs visual with one automatic and two manual. Main technique applied was the connection of medical terms and concepts to visual appearances.
IPAL-CNRS (Institute for Infocomm Research) [23]	IPAL	Singapore	6	Submitted a total of 6 runs, all automatic with two being text only and the other a combination of textual and visual features. For textual retrieval they map the text onto single axes of the MeSH ontology. They also use negative weight query expansion and mix visual and textual results for optimal results.
Daedalus & Madrid U. [24]	MIRA	Spain	14	All runs automatic, with 4 runs visual only and 10 mixed. Mainly used semantic word expansions with EuroWordNet.
National Chiao-Tung U. [25]	NCTU	Taiwan	16	All runs automatic, with 6 visual only and 10 mixed. Used simple visual features (color histogram, coherence matrix, layout features) as well as text retrieval using a vector-space model with word expansion using Wordnet.
Oregon Health & Science U. Medical Informatics [26]	OHSU	USA	3	Two manual and one automatic runs. One of the manual runs combined the output from a visual run using the GIFT engine. For text retrieval, the Lucene system was used.
RWTH Aachen - Computer Science [27]	RWTH CS	Germany	10	Two visual only runs with several visual features and classification methods of the IRMA project.
RWTH Aachen - Medical	RWTH MI	Germany	2	Submitted runs included two manual mixed retrieval, two automatic textual retrieval, three automatic visual retrieval and three automatic mixed retrieval runs. The Fire image

Informatics [28]				retrieval system was used with varied visual features and a text search engine using English and mixed-language retrieval.
U. of Jaen - Intelligent Systems [29]	Sinai	Spain	42	All automatic runs, with 6 textual only and 36 mixed. GIFT was used as a visual query system and the LEMUR system was used for text retrieval in a variety of configurations to achieve multilingual retrieval.
U. Buffalo SUNY - Informatics [30]	UB	USA	6	Submitted one visual and five mixed runs. GIFT was used as a visual retrieval system and SMART as a textual retrieval system, with mapping of text to UMLS Metathesaurus terms.

Tbl. 2: The table shows the research groups that took part in the competition, the number of submitted runs and a brief description of the type of runs performed.

REFERENCES

- [1] Kirsch,H., Gaudan,S. and Rebholz-Schuhmann,D. (2005) Distributed Modules for Text Annotation and IE applied to the Biomedical Domain. *Int. J. Med. Inform.*, (epub), doi:10.1016/j.ijmedinf.2005.06.011.
- [2] Hatcher,E. and Gospodnetic,O. (2004) *Lucene in Action*. Manning (Publisher).
- [3] Dietrich Rebholz-Schuhmann, Harald Kirsch, Miguel Arregui, Sylvain Gaudan, Mark Riethoven, Peter Stoehr. (2006) EBIMed – Text crunching to gather facts for Uni-ProtKB/Swiss-Prot proteins from Medline. *ECCB 2006*, Eilat, Israel. (Accepted for publication)
- [4] A. Bairoch, R. Apweiler, C.H. Wu, W.C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M.J. Martin, D.A. Natale, C. O’Donovan, N. Redaschi and L.S. Yeh. Jan 1, 2005. The Universal Protein Resource (UniProt). *Nucleic Acids Research*, **33**(Database issue):D154-9.
- [5] A.C. Browne, G. Divita, A.R Aronson and A.T. McCray. 2003. UMLS language and vocabulary tools. *AMIA Annual Symposium Proceedings*, p. 798.
- [6] GO Consortium. Jan 1, 2006. The Gene Ontology (GO) project in 2006. *Nucleic Acids Research*, **34**(suppl_1):D322-D326.
- [7] 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.
- [8] P. Durusau and M.B. O’Donnell. 2002. Concurrent Markup for XML Documents. Proceedings of XML Europe. Atlanta, Georgia.19
- [9] L. Chen, H. Liu and C. Friedman. 2005. Gene name ambiguity of eukaryotic nomenclature. *Bioinformatics*, **21**(2):248-56
- [10] Gaudan,S., Kirsch,H. and Rebholz-Schuhmann,D. (2005) Resolving abbreviations to their senses in Medline. *Bioinformatics*, **21**(18), 3658-64.

¹⁹ <http://www.idealliance.org/papers/xml02/dx_xml02/papers/03-03-07/03-03-07.html>