



SemanticMining

NoE 507505

Semantic Interoperability and Data Mining in Biomedicine

D24.1

Report: Data mining and information retrieval

Report Version: 0.4

Report Preparation Date:

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	3
2	INTRODUCTION	3
3	OBJECTIVES OF THE WORKPACKAGE.....	3
4	QUALITY INDICATORS RELEVANT TO WP24.....	4
4.1	Q1 Workshops and symposiums	4
4.2	Q2 Sharing of resources and use of research software tools.....	4
4.3	Q6 Short-and medium-term visits of staff members	4
4.4	Q7 Co-authoring of research papers, reports and educational materials	4
5	SUMMARY OF ACTIVITIES IN 2004	4
5.1	Workshops and Symposiums.....	4
5.1.1	Workpackage Meetings	4
5.1.2	International Conference Participation	5
5.2	Co-authoring.....	5
5.3	Joint Research Programme	5
5.4	Exchange Visits.....	6
6	WORKPACKAGE ASSESSMENT	6
6.1	Assessment against Quality Indicators	6
6.2	Assessment Against Workpackage Objectives.....	6
6.2.1	Exchange of Methods and Resources	6
6.2.2	Integration of text resources and biomedical databases.....	6
6.2.3	Preparation of cross-lingual information retrieval	7
6.2.4	Integration with other Workpackages.....	7

1 Administrative Information

Lead contractor: EBI

Responsible: EBI, UKLFR/UNIFR, DIM

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

Deliverable: D24.1

Author: Dietrich Rebholz-Schuhmann

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 7 (June 2004) until month 11 (November 2004) 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 only project partner in the NoE, which has its origin in the domain of bioinformatics and molecular biology in contrast to medical informatics. The first project months were used to assess available IT solutions from the main WP participants, which can be integrated into a common solution for both domains.

On the other side EBI is the only project partner, which provides IT services to the public fulfilling information needs in the biomedical domain. It is obvious that the NoE can contribute to information provision to the public, thus the developmental work in the NoE can be transformed into IT solutions to the public.

As a conclusion, plans have been settled between EBI, DIM and UKLFR/UNIFR to establish at the EBI a combined information retrieval (IR) and information extraction (IE) engine, which provides access to Medline and available full text documents, and which integrates suitable software components from the NoE, amongst others the representation of medical terms from MorphoSaurus for cross-lingual medical information retrieval.

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

WP24 should be represented with several member-institutions presenting as a unit via workshops, information meetings, and the like in at least 3 national and 1 international conference per year

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

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 = less than 2 weeks, for conferences and workshops, meetings with students

Baseline: assumed 0

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

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

5 Summary of Activities in 2004

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 Kick off meeting	7 July	Balatonfüred, Hungary	5	Delegates from UKLFR, UNIFR, DIM

WP 24 Follow up meeting	6/7 December	Copenhagen, Denmark	4	Delegates from UKLFR, UNIFR, DIM
-------------------------------	-----------------	------------------------	---	----------------------------------

Future Plans

The next WP24 internal meeting will be held at EBI, probably in April.

5.1.2 International Conference Participation

WP 24 relevant participation in international conferences include:

Title	Date	Location	Network Delegates	Comments
ISMB 2004	31 July – 4 August	Glasgow, U.K.	1	International conference; poster presented by network members
COLING 2004	23-27 August	Geneva, CH	2	International conference; poster presented by network members
CLEF 2004	5-17 September	Bath, UK	1	International workshop; paper presented by network member

5.2 Co-authoring

No co-authored papers.

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.

Every author integrates his domain knowledge into his published documents. Such domain knowledge relies on the proper use of terminology, which is different but not disjoint in the medical field in comparison to molecular biology. The Rebholz group (EBI) has established online tools ([Whatizit](#)), which analyze documents after their disposal to the online tool, which identify contained terminology and which link it to biomedical databases. In the next developmental step this information extraction engine will process a number of documents, which have been returned from the retrieval engine (under development). In collaboration with DIM and UKLFR/UNIFR the retrieval engine will be tuned to cross-lingual querying (UKLFR/UNIFR) and to optimized information retrieval (DIM).

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.

Future plans foresee the setup of a full text repository at the EBI for use in the NoE. Documents in the repository will have a standardized format and tagged sections for common use. Tag sets will follow standards defined amongst the participants of WP24.

5.4 Exchange Visits

The number and duration of exchange visits between network members is another indicator of network effort spent towards future joint research. The summer school and several conferences have provided opportunity for many informal bilateral and group discussions. Up to now there have been only short exchange visits:

- Visit of Dietrich Rebholz-Schuhmann and Harald Kirsch, EBI at Freiburg (UKLFR, UNIFR, Mar 15)
- Visit of Dietrich Rebholz-Schuhmann and Harald Kirsch, EBI at Geneva (DIM, Aug 27)
- Visit of Patrick Ruch (DIM) at the EBI, Cambridge, UK (Oct 18/19): installation of DIM's retrieval engine optimized for GO terminology

6 Workpackage Assessment

6.1 Assessment against Quality Indicators

Month 7 to 11 was scheduled for WP24 activities (in total 5 months). During this period the methods from DIM and UKLFR/UNIFR were assessed for integration into a common solution. DIM's retrieval engine based on GO terminology was installed at the EBI for evaluation purposes.

Early developmental work was mainly done at the EBI to allow more time to settle plans with the participants of WP24. The main work programme has started in November and will lead to a first prototype till end of the year.

6.2 Assessment Against Workpackage Objectives

6.2.1 Exchange of Methods and Resources

DIM's IR engine was installed at the EBI and assessed. It is one of 3 approaches, which has been presented to the Gene Ontology team. As a preliminary result the GO curators have selected the solution with the highest precision. Nevertheless DIM's IR engine for GO terms selects highly specific terms from GO. This feature is under further assessment..

MorphoSaurus has been evaluated with online queries using medical terms and generated (as expected) normalized terms. Based on this small assessment further plans have been prepared to use MorphoSaurus' cross-lingual mappings in the query frontend to the combined information retrieval / information extraction engine at the EBI.

The Whatizit server is open to DIM via http requests on a special port at the EBI. DIM transmits any type of text and can specify the pipeline of modules, which have to process the text. This approach standardizes information extraction tools and results. It takes away overhead from DIM to maintain the software modules. Furthermore, DIM profits from links, which have been added to the processed text and which provide consistent navigation into EBI's public databases

6.2.2 Integration of text resources and biomedical databases

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](#) is an online tool from the Rebholz group, which matches

this need. It provides a set of information extraction modules, which identify protein and gene names (UniProt, LokusLink), GO terminology, as well as mutations and protein-protein interactions.

In the next developmental phase this technology will be combined with an information retrieval engine to support processing of large number of documents in a short period of time to the public.

6.2.3 Preparation of cross-lingual information retrieval

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/UNIFR is assessing the mapping of MeSH thesaurus to MorphoSaurus and the advantages for information retrieval.

6.2.4 Integration with other Workpackages

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

- WP 20 (Multilingual Medical Dictionary): The MorphoSaurus 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.