



SemanticMining

NoE 507505

Semantic Interoperability and Data Mining in Biomedicine

Final SemanticMining Report on Tools and Services

Covering period 2004.01.01 – 2007.06.30

Report Version: 1

Report Preparation Date: 2007.09.15

Classification: PU

Contract Start Date: 2004.01.01

Duration: 3.5 years (42 months)

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 Content

FINANCIAL/ADMINISTRATIVE CO-ORDINATOR	1
RESEARCH ACTIVITIES	2
PARTNERSHIP	3
SERVICES AND TOOLS DEVELOPED BY SEMANTICMINING	4
Description of services, tools, and research material	4
Accessibility (technical requirements, platforms etc.)	6
Maintainability and sustainability of services	6
Dissemination and exploitation aspects	8
Website	8
Educational Material Available	8
Access Methods for new Twiki website	8
Restricted Website	13
DVD	13
Dissemination through conferences and workshops	14
Input to standardisation work	16
Exploitation of results	17
Future of European Networks of Excellence	18
List of deliverables	19
Table of resources, tools and material	22
Table with information on exploitation	23
Table with research applications and funding	24



Financial/Administrative co-ordinator

Name: Hans Åhlfeldt

Address: Department of Biomedical Engineering / Medical Informatics
S-581 83 Linköping University, Sweden

Phone Numbers: +46 13 227574

Fax Numbers: +46 13 101902

E-mail: hans.ahlfeldt@imt.liu.se

Project websites: www.semanticmining.org

Editors of report: Hans Åhlfeldt and Hans Gill based on input from WP-leaders and Board members.



Research activities

The research activities in SemanticMining have been focused around the following areas (work packages):

- the construction of a multi-lingual medical dictionary (WP20)
- principles in ontology engineering (WP21)
- evaluation of SNOMED CT (WP22)
- health statistics, semantic distance and ontologies (WP23)
- text mining and information retrieval in bioinformatics (WP24)
- terminology systems in laboratory medicine (WP25)
- the concept-based electronic health record (WP26)
- medical terminology for laymen (WP27)

Moreover, a series of work packages have been focused on dissemination and outreach activities.

- web portal (WP3/4)
- mobility (WP6)
- summer schools (WP7)
- input to standardisation (WP8)
- dissemination (WP9)

For a more in-depth description of the research context of the presented services and tools in this report, see Final SemanticMining Report on Research Activities.



Partnership

SemanticMining is based on the partnership of 25 partners from 11 European countries (see list below) with approximately 100 identified researchers (25 female) and 35 associated PhD students (10 female). For further information about see www.semanticmining.org

LIU (IMT)	Biomedical Engineering, Medical Informatics, Linköping University, Sweden
LIU (IDA)	Computer Science, Linköping University, Sweden
LIU (C-NPU)	Committee Nomenclature, Properties and Units in Laboratory Medicine, Linköping University, Sweden
KI	Karolinska Institutet, Stockholm, Sweden
SU	Sahlgrenska University Hospital, Göteborg, Sweden
UGOT	Dept of Swedish, Göteborg University, Sweden
UKLFR	Dept of Medical Informatics, Universitätsklinikum Freiburg, Germany
UNIFR	Computational Linguistics Research Group, Albert-Ludwigs-Universität Freiburg, Germany
IFOMIS	IFOMIS, University of Saarland, Germany
CAU	Institute of Informatics and Applied Mathematics, Christian-Albrechts-University of Kiel, Germany
DIM	Division of Medical Informatics, Geneva University Hospital, Switzerland
UOM	Dept of Computer Science, University of Manchester, UK
UCL	Centre for Health Informatics and Multiprofessional Education, University College London, UK
OPEN	Open University, Milton Keynes, UK
INSERM	Public Health and Medical Informatics Laboratory, Broussais University Hospital, Paris, France
CNR-ISTC	Institute of Cognitive Science, Laboratory for Applied Ontology, Italy
EMBL-EBI	European Bioinformatics Institute, UK
ESKI	National Institute and Library for Health Information, Budapest, Hungary
NORDCLASS	WHO Collaborating Centre for Classification of Diseases in the Nordic countries, Uppsala University, Sweden
SOS	The National Board of Health and Welfare, Sweden
STAKES	National Research and Development Centre for Welfare and Health, Finland
KITH	KITH AS, Norway
NBH	National Board of Health, Denmark
MRI	Merrall-Ross International Ltd, UK
EDSA	European Dynamics S.A., Greece



Services and tools developed by SemanticMining

Description of services, tools, and research material

Services and tools developed by WP20

Lexicons and exact evaluation results are available on demand from WP leader (Stefan Schulz stefan.schulz@gmail.com)

Subword Indexing Demonstrators are available at www.morphosaurus.net, user account on demand.

Services and tools developed by WP21

Protege4 with OWL 1.1. Software developed in collaboration with UK funded CO-ODE project and Stanford University. Now de facto standard open source editor for OWL. Open source download available from <http://protégé.stanford.edu>. Additional material available from <http://www.co-ode.org>

Top Bio Ontology (BioTOP). Downloadable from <http://www.ifomis.uni-saarland.de/biotop>, Top Bio Ontology (BioTOP) discussion list. Downloadable from <http://groups.google.com/group/biotop>

BFO Top Level Ontology Manual and OWL Implementation. OWL-implementation and extensive manual concerning Basic Formal Ontology (BFO) with further material (<http://www.ifomis.uni-saarland.de/bfo/home.php>)

Simple Top Bio. OWL implementation of a simple tutorial top ontology illustrating “20 questions” approach to understanding high level categories. In process of being harmonised with TopBio. (<http://www.cs.man.ac.uk/~rector/ontologies/simple-top-bio/>)

MoST module for Archetype Editor. Software suite for identifying appropriate SNOMED codes for binding to Archetypes developed by Rahil Qamar, `PhD Student at University of Manchester sponsored by Semantic Mining in collaboration with LiU and using the LiU Archetype Editor.

Services and tools developed by WP22

Mapping tables from SNOMED CT to NCSP and ICF available in deliverable D22.3.

Experience from translation platform used in Denmark available on demand from Arne Kverneland <ark@sst.dk>

Services and tools developed by WP23

Workbench for simulation of inter-rater agreement is available on demand from WP leader (Håkan Petersson <Hakan.Petersson@imt.liu.se>).

Services and tools developed by WP24

Whatizit Web Services: This IT solution is accessible from servers running at the EBI since 2005. In its latest version, Whatizit delivers its services based on the SOAP standard. A number of modules are part of Whatizit that annotated different semantic types (e.g., genes/proteins, chemical entities, diseases, other). Whatizit received about 1 million requests from 520 distinct hosts during January – June 2007 and 5 GBytes of data have been delivered to user on a monthly basis in this period.



EBIMed: This Web portal retrieves and annotates Medline abstracts. The user turns in his query terms, the retrieval uses the query terms to seek the corresponding Medline abstracts from EBI's in-house Medline installation, and then EBIMed processes all documents using a selection of Whatizit modules. Altogether, EBIMed extracts the co-occurrences of annotations in the sentences of the Medline abstracts. EBIMed sorts all results, ranks them and then presents them in a table to the users. In addition to the central table, the user can access other collections of information (e.g., lists of identified proteins) to explore the content of the retrieved documents. EBIMed has received 130,000 requests from 2,350 distinct hosts over the reporting period.

PCorral: This application integrates syntactical patterns to identify protein-protein interactions from the scientific literature. Furthermore, PCorral also applies cooccurrence and tri-cooccurrence (agent, patient and the verb) apart from syntactical patterns, to identify protein-protein interactions with higher recall and lower precision. Again the application sorts, ranks and delivers the results to the user.

BFMed: a cross-language French-to-English search engine for MEDLINE

GOCat (Gene Ontology Categorizer): an automatic text categorizer to automatically annotate proteins using GO categories (molecular functions, subcellular locations and biological processes).

EAGLi ("eagle eye"): an advanced search engine for MEDLINE with terminology-powered navigation skills (Gene Ontology, Swiss-Prot keywords).

EAGL: a set of services: Gene ontology and MeSH categorizers, argumentative classifiers

EAGLb: another set of categorizers for biomedical texts

Services and tools developed by WP25

The C-NPU coding schema is available from <http://dior.imt.liu.se/cnpu/>. The standard CEN EN1614 is available from <http://www.centc251.org>.

Services and tools developed by WP26

The Archetype tools developed by WP26 provide the ability to:

- define models of clinical data (concept domain) structure and their interrelationships (e.g. the recording of 'adverse reaction') by clinical and other domain specialists;
- document a particular clinical observation, evaluation, instruction or action in an agreed, formal, interoperable and re-usable way;
- connect data stored in the EHR in a valid way to terminologies;
- create semantic queries based on paths extracted from archetypes.

The ADL workbench is an open source software tool published by the *openEHR* Foundation (www.openehr.com)

Archetype editors available from the *openEHR* Foundation (www.openehr.com) and Linköping University (www.imt.liu.se/mi).

Archetype Template builder available from the *openEHR* Foundation (www.openehr.com)

The MOST software suite for identifying appropriate SNOMED codes for binding to Archetypes available from University of Manchester



Toolkit for visualisation of EHR-data available from Linköping University (www.imt.liu.se/mi).

The WP27 resources

- (1) Proof-of-concept simple patient report generation system producing output in English, French and Swedish;
- (2) Corpus of English medical texts in the cancer domain, with expert and non-expert components;
- (3) Corpus of Swedish medical texts (all sub-domains), with expert and non-expert components.

Accessibility (technical requirements, platforms etc.)

Technical requirement for WP20-services is standard web browsers and text editors.

All WP21 material is open source (LGPL license), written in Java, and runs on all standard platforms. Most requires at least 1GB memory with current operating systems (Windows XP/Vista, Mac OS X or Linux). For details of download sites see description above.

Whatizit, EBIMed and PCorral run in a Web browser. They are accessible from any computer that is linked to the internet and that has an installation of a Web browser. Every software developer can access Whatizit through SOAP Web services. The Web pages of Whatizit contain a sample software solution to call Whatizit Web services.

Technical requirement for WP25-services is standard web browsers.

Each of the WP26 tools has been documented in greater detail in Deliverable 26.3. During the early development phases the source code has usually remained within the authoring team, but been shared informally on request within the consortium. Many of them are now at an advanced stage and the source code and execution code modules have been contributed to the *openEHR* Foundation which houses them under open source licence in a public domain accessible source server.

Technical requirements for the WP27 resources listed above are:

- (1) Java program and SQL database;
- (2) No technical problems, but IPR will be an issue for wider use of the corpus;
- (3) Character coding needs attention (ISO 8859-1 vs. UTF-8), but is not problematic in practice. However, IPR will be an issue for wider use of the corpus.

Maintainability and sustainability of services

WP20 tools and demonstrators are available at www.morphosaurus.net assured by AVERBIS GmbH.

The members of WP21 are involved in a wide variety of further research projects, both EU and Non-EU funded, which ensure the survival of the key resources described above. The project has increasingly engaged with OBO and the European Bioinformatics Institute which is a major centre with continuing commitment to development. Likewise, it has engaged with the US National Center for Bio-ontologies. The software tools are currently supported



through a combination of project grants (EU, UK JISC and NIH), industrial collaborations and volunteer open source contributions.

Access to IT services to the public developed by WP24 is part of the mission of the EBI. As a result, all development teams at the EBI are urged to develop software that is compliant to standards that ensure long-term maintainability and sustainability of IT services. The Web interfaces of EBIMed, PCorral and Whatizit are based on modules that are embedded into template components for all applications at the EBI. All processes integrated in all three applications are centralized in one server machine and profit from the same task queuing system, the same document repository and the same information retrieval engine. The solutions based on this infrastructure have been publicly available for about 2 years now and have been maintained without any severe problems. Currently the EBI investigates on the resources required to maintain this infrastructure as part of the services team at the EBI.

The main requirement for maintenance of the Archetype tools is funding. Fortunately it already appears that the UK NHS will provide some follow-on funding for some of these tools, and the partners are continuing to seek additional such avenues.

At the George Pompidou Hospital (HEGP) in Paris, methods and services will be further tested and developed in close cooperation between the WP26 partner and the providers of EHR information systems (DxCare®, Medasys©) with the objective to capture data for both patient care and research in the cardiovascular radiology domain.

In Sweden, the WP26 partners (Linköping University, Karolinska Institute, National Board of Health and Welfare) will continue their cooperative work with development of pilot EHR implementations as part of national projects. A particular objective of this project is to further test the archetype approach in conjunction with SNOMED CT with the objective to capture data for both patient care and research in child and maternity care.

The partners are in a good position to promote these tools, since they are a unique resource for those wishing to develop or exploit clinical data structures in an interoperable way i.e. by using archetypes. They are also amongst the only tools that presently conform to EN13606-2. SemanticMining have through the WP26 partners made significant contributions to national strategy plans regarding EHR implementation. In Sweden, the WP26 partners have during spring 2007 been organising four seminars devoted to in-depth presentations and discussions on EHR modelling (e.g. EN13606 and *openEHR*), and terminology binding (e.g. with SNOMED CT). These seminars have contributed to the increased interest of uptake of “the semantically well-defined EHR” among both EHR vendors and health care professionals

As with text corpora in general, IPR issues are difficult. For the time being regarding the WP26 resources (2) – (3), only the collecting institutions (The Open University and Göteborg University, respectively) can use the corpora in their research, but not share them with others. Göteborg is negotiating with the IPR holders of the Swedish texts for permission to distribute a part of the corpus for research and education purposes.



Dissemination and exploitation aspects

A significant role of this NoE is to provide educational material based on both workshops and the research. It is hoped that this educational material available through the web site will be useful, not only to students associated with SemanticMining, but also for the general public, and other interested parties. An important way of sharing research results is also through scientific publication.

During the course of SemanticMining a large amount of educationally-orientated material has been developed and presented at Summer Schools and conferences.

The following types of material will be presented.

- Summer school presentations
- Deliverables
- Educational material overviews, specifically for the website

The following routes of dissemination will be considered:-

- Website (both public and restricted areas)
- Semantic Mining DVD
- Doctoral colloquia
- Conferences and workshops
- E-mailing lists

Website

Semantic Mining has both a public, and a restricted (partners-only) website sections, based on the Mermig platform, hosted by European Dynamics. From July, 2007, a new Twiki-based site hosted by Jena University has gone live. The layout and contents are based on the current www.semanticmining.org site. The domain name will remain the same. All material derived from Semantic Mining will, therefore, be archived for future requirements on this Twiki website.

Educational Material Available

As many presentations as possible, from the Summer Schools, have been uploaded. Where external speakers were involved, permission for uploading presentations was sought, though not always obtained – these presentations have not been made public.

In addition, there are some overview articles and lists of relevant links. It was considered that the best summaries of the work carried out by Semantic Mining are the deliverables for public dissemination, which have been uploaded in PDF format.

Access Methods for new Twiki website

Until 1st July 2007, the material on the public website was accessed via a menu and submenus. The new Twiki site will contain the basic menu, a graphic display of the available material (based on the mind map design), a full and detailed index with the index entries hyperlinked to the relevant article, and a search engine.

Standard Menu of Twiki-based Website (www.semanticmining.org)



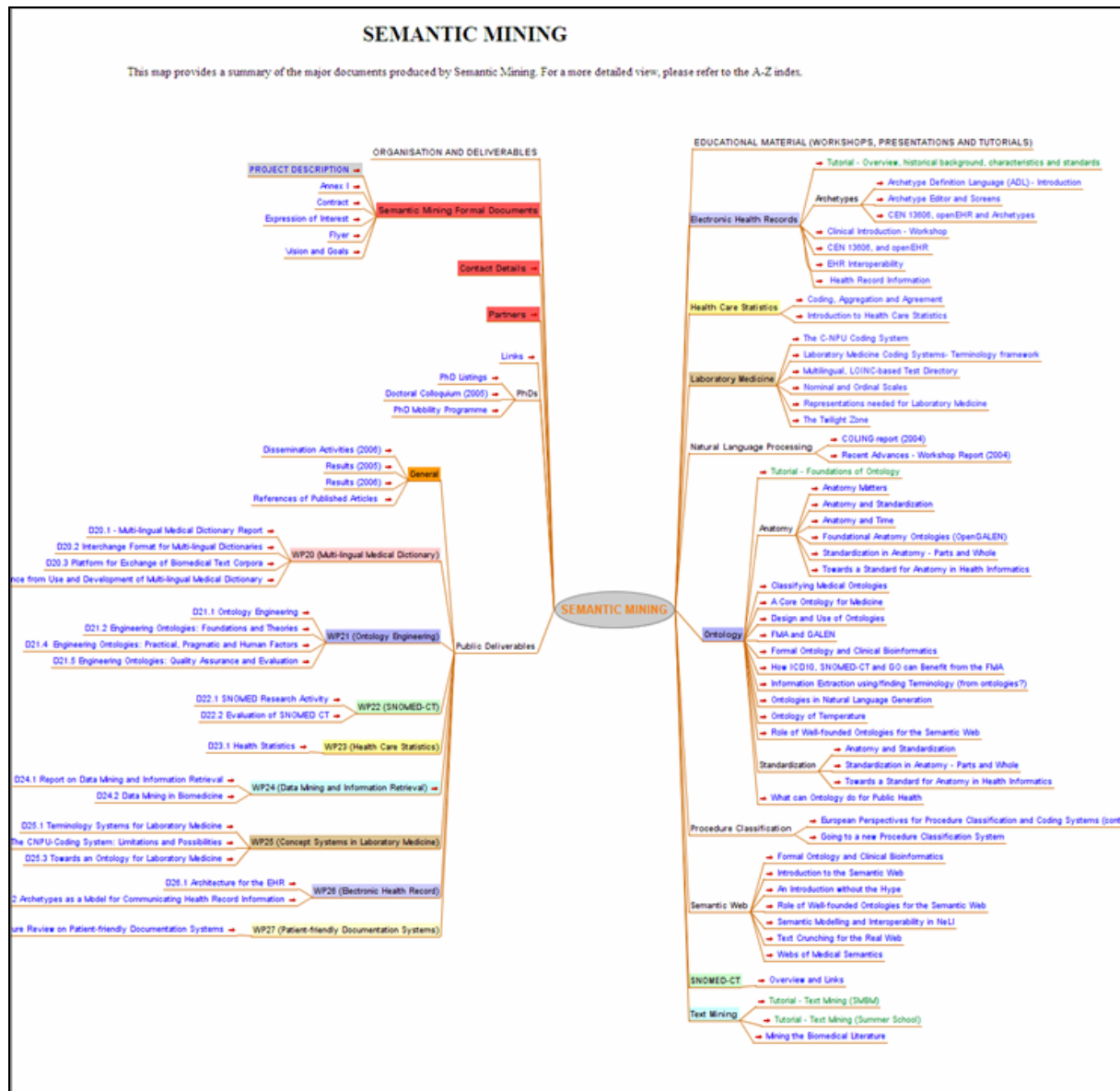
The screenshot shows a web browser window displaying a Twiki page. At the top right, there is a search box and a 'Jump' button. Below the header, a navigation menu is visible with 'External' and 'Internal' sections. The main content area features the title 'Semantic Mining (NOE 507505)' in red, followed by three logos: the IST logo, the European Union flag, and the Information Society Technologies logo. The text describes the network's aim to establish Europe as a scientific leader in medical and biomedical informatics. A red arrow points to the 'Internal' link in the navigation menu, which is labeled 'Menu'.

Twiki-based Website hosted by Jena University



Menu with 'Mind map-style' format

This graphical presentation lists the various deliverables and other resources at heading level i.e. it provides a graphical contents list. It links, by colour, the deliverables and other presentations for each WorkPackage.



Menu using Mind-map-format



Index

All the educational resources from Semantic Mining have been indexed, in some detail. The index entries are hyperlinked to the relevant webpages, or to each presentation, or deliverable. This index enables retrieval at the greater level of granularity than other methods of retrieval. Both the contents listing and the graphic map only provide retrieval at the level of headings, but there is a huge amount of educational information provided by Semantic Mining that would otherwise be less easy to find. For instance, some of the Summer School tutorials extended for approx. 3 hours, yet without an index specific aspects of the presentation would not be easily retrievable e.g. *Foundations of Ontology*, by Barry Smith, which included such concepts as SNAP and SPAN.

The index has approx 880 entries, which includes cross-references. There are two types of cross-references used:-

see references, which act as 'go to' commands for synonyms, from the non-preferred term to the preferred term

see also references, which provide 'related terms', for additional retrieval.

The cross-references add to the value and richness of the index. Synonyms, and abbreviations, have been included to ensure full retrieval of data.

The index has been constructed based on the International Standard, ISO 999. 1996 *Guidelines for the content, organization and presentation of indexes*. (International Organization for Standardization, 1996). This standard is primarily for indexes to printed and electronic documents, but the rules are also applicable to website resources.



Index

Note: This index covers the activities of Semantic Mining, with special emphasis on the educational material and the public deliverables. All hyperlinks are to the **start** of the individual presentations or tutorials i.e. you may need to scroll down, or through the slides, in order to view the required information. Index entries in **blue** are hyperlinked to these pages or presentations; heading without hyperlinks are in **orange**.

This index is deliberately very detailed in order to show the extent of the data available, and to improve retrieval. For a more concise overview, please use the graphical map.

[A](#) [B](#) [C](#) [D](#) [E](#) [F](#) [G](#) [H](#) [I](#) [J](#) [K](#) [L](#) [M](#) [N](#) [O](#) [P](#) [Q](#) [R](#) [S](#) [T](#) [U](#) [V](#) [W](#) [X](#) [Y](#) [Z](#)

A [return to top of page](#)

- Abnormality theory, disease (Reznek)
- Activities, of Semantic Mining members
- Aggregation, in primary health care (health statistics)
 - agreement
 - levels
- Ambiguities, in text mining
- Anatomy
 - applications (including knowledge-based)
 - definitions, *link 1*
 - definitions, *link 2*
 - Foundational Model, *see* Foundational Model of Anatomy
 - foundational ontology
 - foundational principles
 - NCICB terminology
 - ontology
 - ontology, standardization and
 - ontology, time and
 - patient-specific model
 - standard development by CEN
 - standardization and
 - standardization, for health informatics
 - standardization, parts and wholes
 - standardization, reasons
 - surgery and, ontology
 - taxonomy
 - terminology standard
 - time and
- Annex 1
- Archetype definition language (ADL)
- Archetypes, electronic health records
 - 13606-1 profile for
 - as model for communicating health information (D26.2)
 - definition and repositories
 - deliverable on
 - editor, and screens from
 - examples from workshop
 - formalisms
 - integration with ontologies and terminologies
 - model
 - repositories
 - requirements
 - tutorial
 - value of
- Aristotelean definitions
- ATA (Anatomical Transformation Abstraction)
- Audit Commission Report on Patient Health Records (1995)
- Authoring, of ontologies

B [top of page](#)

- Binary results, in laboratory medicine
- Bio-informatics
 - convergence limitations vs medical informatics
 - text mining tasks
 - see also* Medical informatics
- Biological structures, standardization, parts and wholes
- Biomedical databases, text mining vs
- Biomedical informatics
 - Joint European Summer School
 - see also* Bio-informatics
- Biomedical literature, text mining
- Biomedical ontologies
 - development
 - electronic health record and
 - quality assurance
 - reference ontology, *see* Biomedical reference ontology
 - training course
- Biomedical reference ontology
 - FMA presentation
- Biomedicine, natural language generation in
- Biostatistical theory, health (Boorse)

Small sample of Index for Semantic Mining



Restricted Website

Mermig has provided a communication and management tool for partners of Semantic Mining. All NoE content has been uploaded and is accessible, by password, to all partners. This website will now be hosted as the Internal Twiki website hosted by Jena University, and remains accessible to all partners.

In addition to the material available on the public website, this restricted-access site contains

- all deliverables (both public and restricted circulation),
- all formal documentation,
- templates for submission of reports and
- conference presentations for which permission for public release was not available

Since the educational material is available on the public website, this sector of the website will not be considered further.

DVD

Semantic Mining is providing a DVD to all partners. The DVD will contain **all** educational material i.e. deliverables and Summer School presentations (including those presentations not made public), in addition to formal documents. It will provide all participants with an archive of material presented and developed. The DVD is down-loadable from the Jena University website.



Dissemination through conferences and workshops

During the last phase of the project, a series of specific dissemination and outreach activities have taken place. The dissemination activities were designed to facilitate uptake of results (knowledge, services, tools etc.) by targeted groups, in particular public health care policy and decision makers, ICT system vendors, and the research community. The series of events do also facilitate knowledge transfer between partners and thereby facilitate cross-fertilisation between scientific sub-disciplines within the network. Since all events are a joint responsibility between several partners, the dissemination programme in it self fosters cooperation. Dissemination through scientific publication is reported below in section A2.

First Joint European Summer School in Biomedical Informatics

As a result of NoE cluster meetings with INFOBIOMED and BIOPATTERN, the first European Summer School in Biomedical Informatics was successfully held in Hungary, July 2006. The one week program contained a doctoral consortium (PhD student centred), tutorial and workshops, a “best of” scientific paper session, a NoE cluster meeting, and NoE specific administrative meetings. The topics for the tutorials Text and web mining, Data visualisation, and Ontology engineering were chosen as they represent core knowledge areas of the three networks. Experts from all networks were involved in the preparation and presentations. The results of this part of the summer school were twofold; sharing of in-depth knowledge between participants, and insight into different application areas of the three networks. The summer school was attended by nearly 100 participants, representing all three NoEs, and a good mix of PhD students and senior researchers.

As a continuation of this event, a International Symposium on Biomedical Informatics in Europe, was jointly organised by SemanticMining and INFOBIOMED in June 2007. The scientific program included tracks on Systems Biology Insight into Human Disease, Biocomputation and Knowledge Management in Drug Discovery, Data and Knowledge Mining in Biomedicine, Modelling and Simulation in Biomedical Research, Standards and Ontologies in Biomedical Informatics, and Case studies. A Doctoral Consortium was also held, where PhD students and senior researchers met and shared experiences on formulation of research questions, research methodology, and the strength of research results.

Semantic Mining in Biomedicine

Partners of SemanticMining have taken the initiative to start a new conference series called “Semantic Mining in Biomedicine” (SMBM). SMBM 2005 (Hinxton, Cambridge, U.K.) was the first international conference event world-wide which combined such diverse research areas as biomedical ontologies and terminological infrastructure as well as biomedical data and text mining and other forms of content-oriented document processing and text analysis from biomedical data bases. The second SMBM was held at Jena University, April 2006. Both conferences have gained additional recognition by the scientific community based on their policy to publish the best papers in special issues of high-impact journals e.g. BMC Bioinformatics.

The scope of SMBM 2006 included the following topics applied to the domain of Biomedicine: Information extraction, information retrieval, text mining, knowledge discovery and data mining, term engineering, named entity recognition and interpretation, evaluation standards, ontological foundations of molecular biology and related areas, automated corpus/lexicon construction for Biomedicine.



Training Course in Biomedical Ontology

In May 2006 and June 2007, a three-day training course respectively was designed to provide a basic introduction to the field of biomedical ontology and to enhance awareness of current developments and best practices in ontology in the life sciences.

It hoped to gain the interest of participants with the following backgrounds: developers and users of biomedical ontologies, terminologies and coding systems, developers and users of electronic patient record systems, biologists and physicians interested in the possibilities of modern ontologies, and targeted advanced doctoral students, but also interested post-doc and industrial participants or people from hospitals for synergetic effects. The number of participants was to be restricted to about 30 to maximize possibilities for intense discussion. All participants should receive from their attendance in this tutorial hands-on training in ontology design and use.

Swedish Terminology Conference

The objective of the Swedish Terminology Conference, held in Kalmar, September 28-29, was to establish a forum and meeting place for health care professionals, system developers, informaticians and researchers with an interest in the further development of documentation and sharing of patient information, the multi-professional health record, and follow-up and quality assessment of health care. Both national and international perspectives should be presented, as well as perspectives of the patient, the health care professional, and the system provider. Ongoing research within SemanticMining were presented though seminars and demonstrations, in particular the work related to the semantic-based EHR, openEHR, the terminology binding problem between the EHR and reference terminologies such as SNOMED CT (WP21, WP22, WP26). The conference was attended by 120 participants, representing the major health care regions in Sweden, the major providers of electronic health record systems (EHRs), and public health care organisations such as National Board of Health and Carelink (national network of health care providers), and academic departments. As result of the Swedish conference, workshops are being planned for 2007, where the models and tools of openEHR (archetype editor, repository, terminology binding etc.) and SNOMED CT (conceptual framework, browsers) will be further presented and scrutinised.

SemanticMining Conference on SNOMED CT

October 1-3, the first European conference on experience from SNOMED CT was held in Copenhagen. The objective was to organize an international forum for discussing achievements and actual experiences with reference terminology, framework, terminology contents and organizational issues in relation to SNOMED CT. A broad range of topics were to be addressed, including formal and ontological aspects of SNOMED CT, mapping between SNOMED CT and legacy terminologies and classifications, SNOMED CT and the Electronic Health Record, SNOMED CT support for Coding and epidemiology, viewers and browsing tools.

This event, called Semantic Mining Conference on SNOMED CT (SMCS 2006), was intended to be the first of several European fora for health policy makers, clinicians, nurses, system developers, computer scientists, terminologists and translators. It should embrace both scientific presentations and invited presentations which provide an overview of current efforts and developments in the context of SNOMED CT.

Potential members of the SNOMED CT SDO (Standard Developments Organization) had meetings in Copenhagen in connection to the SMCS 2006 Conference with representatives of the College of American Pathologists in order to set up a timeframe for transfer of the IPRs of SNOMED CT. Due to this development and the need to make decisions on national level about whether or not to join the SNOMED CT SDO, the SMCS 2006 Conference was offered



at the right time and met a high demand of health professionals, system vendors, researchers, and health policy makers. This may explain the extraordinary response to this conference, the high level of scientific contributions and the readiness of top experts to give tutorials and keynote presentations. The feedback of participants was positive to enthusiastic. Excellent keynote presentations showed not only achievements in the SNOMED CT development, but also shortcomings, concerning the content and maintenance quality. The Danish SNOMED CT localization experience was presented and shown high interest by representatives from other European countries. The new route SNOMED CT is taking by the foundation of the SNOMED CT SDO and its implications were extensively discussed. In summary, SMCS 2006 was a highly satisfying event which took place at the right place, with the right content, at the right time.

Terminology Conference in Romania

A terminology conference has also been organized by SemanticMining in Timișoara, Romania, during 2006. The choice of its location underlined the organizers' concern to integrate Eastern European Medical Informatics researchers into the ongoing discussions on biomedical terminology systems. The conference had a scientific focus, with an international call for papers and a scientific program committee. In the call for papers we emphasized ongoing research involving specialists from different disciplines, such as Medicine, Computer Science, Philosophy, and Linguistics as well as the existence of different genres of biomedical terminology systems and competing approaches mainly centered on the question of whether to represent the world of reality or the world of language.

Input to standardisation work

The SemanticMining NoE has had explicit participation standardization activities as one of its major dissemination methods and an important opportunity to influence interoperable eHealth in large scale in a medium time frame. Of course the scientific results from this NoE is only one aspect influencing the standardization but this NoE has through a number of experts had a very strong influence on key aspects of interoperability standardization. In addition, the WP leader has had the opportunity to participate in important strategic planning and promotion bodies for interoperability standardization as part of the WP8 activities such as the eHSCG, the CEN eHealth Standardization Focus Group and the eHealth Stakeholders group working as advisory bodies to the Commission.

Among the many activities that the NoE has contributed to are:

- The establishment of the eHealth Standardization Co-ordination Group which in co-operation with WHO and ITU now includes CEN from Europe, ISO, IEEE, HL7, DICOM and OASIS. A web site was established (www.ehcs.org) with information on all major eHealth standards and activities
- The finalisation in March 2005 of the eHealth Standardization Focus Group strategic report to the European Commission and the member states and since December 2005 the follow up eHealth Stakeholders Group which is to give specific advice to the CEC on standards for a Patient Summary Record
- The further work on finalising the EN 13606 Health Informatics - Electronic Health Record Communication series (in co-operation with WP 26).
The first essential part one "Reference Model" and also Part four: "Security" have been published in all of the 30 European CEN countries and the remaining three parts are in final stages of approval in CEN. In addition all parts are being processed as International standards through ISO/TC 215. As a special subtopic of this a joint

CEN-HL7 project on an Archetype Framework Standard in five parts was started with NoE partners in the lead which is also conducted in co-operation with HL7.

- The finalisation of the EN 12967 Health Informatics – Service Architecture (HISA) standard in three parts, finally approved for publication in Europe summer 2007 and also accepted as the basis for an international ISO standard.
- The EN 1614 model for representing a Structure for nomenclature, classification, and coding of properties in clinical laboratory sciences was published in 2006.
- Work on the new CEN standard for a Categorical structure for system of concepts for human anatomy took a completely new start during 2005 after extensive interactions with the NoE ontology experts and has been finally approved.
- Guiding standardization in CEN and ISO in the field of terminology and concept systems on the relation between the world of concepts and the real world described by ontologies (paper by Klein and Smith)
- The ISO 17115 Health informatics – Vocabulary for terminological systems has been published by ISO in april 2007 after several years of contributions from a NoE expert.
- The CEN/TS 15699 Health informatics – Clinical knowledge resources – Metadata was finalised and went for final ballot in 2007. This was produced by two NoE experts.

Exploitation of results

Discussions were held during 2006, between a major international publishing company and the WP20 lead contractor with respect to the possible exploitation of the multilingual medical dictionary. This link was made via the lead contractor of WP9. Discussions are still ongoing, and no agreement has yet been reached. Regarding the Morphosaurus sub-word dictionary, a subset of the WP20 multilingual dictionary, exploitation contracts were closed with a German medical library and a major German publisher of online content.

Discussions have also been held with publishers over the use of SNOMED CT for information retrieval from medical journals. Several European medical publishers e.g. Elsevier, Royal Pharmaceutical Society of Great Britain (for the British National Formulary), are now starting to utilize and incorporate SNOMED CT into their online medical resources. This will help to increase the pan-European use of SNOMED CT and European access to medical information. In addition, there is likely to be a need by medical publishing houses for consultancy work by members of WP22, and a preliminary approach has been made to one partner of this NoE for such help.

Description of results from the various SemanticMining work packages has been sent to ICT-vendors through available e-mailing lists. A series of contacts with industry for exploitation of results and know-how has been triggered by the joint work program of SemanticMining. Main areas of interest for exploitation are language technology as worked on in WP20, WP24 and WP27, and the EHR-related work ongoing in WP21, WP22, and WP26. Specific discussions concern the uptake of open source components Protégé OWL and openEHR modules for archetype generation and terminology binding.

During the last phase of the project, final reports have been produced where results (e.g. research results, references to scientific publications, description of services and tools) from the SemanticMining project are presented based on the rich list of deliverables. Target groups for distribution will be ICT-vendors, public health care organisations and regional health care networks through available lists with contact information provided by branch organisations for medical technology industry (e.g. EUCOMED, Swedish Medtech). The updated list of



deliverables contains deliverables summing up research results as well as available tools and services produced within SemanticMining.

Future of European Networks of Excellence

SemanticMining has taken part in a joint initiative regarding the future of European Networks of Excellence. An opinion paper has been signed by 58 NoEs, which will be presented and discussed at an open forum “The Future of European Networks of Excellence” to be held in Brussels, November 20, 2007.



List of deliverables

Deliverables available from the SemanticMining website (www.semanticmining.org) are listed below.

<i>Number</i>	<i>Name</i>
D1.1	Detailed analysis of research
D1.4	Quarterly reports, Periodic Activity and Management Reports
D1.5	Project presentation
D3.1	Typology of shared resources
D4.1	Public website
D5.1	Network meetings
D6.1	Report on doctoral dissertations
D8.1	Participation in standardisation
D7.1	Planning of summer school
D10.1	Workshop on health statistics
D11.1	Workshop on semantic web
D12.1	Workshop on ontology engineering
D3.2	Common database
D5.2	Network meetings
D6.2	PhD programmes
D6.3	Mobility program
D7.2	Evaluation of summer school
D9.1	Dissemination of material
D13	Workshop on NLP
D16	Workshop on EHR
D20.1	Multi-lingual medical dictionary
D21.1	Ontology engineering
D22.1	SNOMED CT
D23.1	Health statistics
D24.1	Data mining and information retrieval
D25.1	Concept system in lab.medicine
D26.1	The Electronic health record
D2.1	Assessment and strategic planning
D14/15	Workshop on data mining and information retrieval
D4.2	Report on public website
D8.2	Participation in standardisation work



D9.2	Report on educational material
D9.3	Report on dissemination
D20.2	Prototype multi-lingual medical dictionary
D22.2	Report on SNOMED CT – evaluation
D22.3	Experience from translation and evaluation of SNOMED CT
D23.2	Report on information quality in health registries
D24.2	Text mining and IR in biomedicine
D25.2	The CNPU-coding system – limitations and possibilities
D26.2	Architecture for semantic-based EHR
D30	Workshop on ontology and biomedical informatics
D31	Workshop on human factors and large ontologies
D32	Workshop on the boundary problem
D33	Workshop on SNOMED CT
D34	Workshop on concept system in lab.medicine
D35	Workshop on text mining from EHR
D36	Workshop on semantic web
D5.3	Network meetings
D6.4	Mobility program
D7.3	Evaluation of summer school
D2.2	Assessment and strategic planning
D20.3	Platform for exchange of biomedical text corpora
D21.2	Engineering ontologies: foundations and theories from philosophy and logic
D21.3	Engineering ontologies: foundations and theories from computer science
D21.4	Engineering ontologies: practical, pragmatic and human factor issues
D21.5	Engineering ontologies: quality assurance and evaluation
D9.4	Dissemination of results
D20.4	Experience from development of multi-lingual medical dictionary
D23.3	Towards measurement of semantic distance
D24.3	Biomedical semantic classes for text mining
D25.3	Towards an ontology for laboratory medicine
D26.3	The Boundary problem – linking information and terminology models
D27.1	Literature review of patient-friendly documentation systems
D40	Training program in biomedical ontology
D41	Workshops in medical terminology
D7.4	Joint NoE summer conference 2006
D2.3	Assessment and strategic planning
D4.3	Use of SemanticMining web site



D27.2	Empowering the patient with language technology
D6.5	Mobility program in biomedical informatics – results and implications
D8.3	Participation in standardisation work
D9.5	Educational material from SemanticMining
D20-27.1	Research results from SemanticMining
D20-27.2	Data and service resources from SemanticMining (services and tools)



Table of resources, tools and material

<i>Type / scope</i>	<i>Details / Comments</i>
Pooling of heterogeneous lexicons according to common interchange format	Sharing of lexicons by common interchange format among WP20-partners. Multi-lingual medical dictionary with more than 100.000 entries (Dec 2005). See deliverable 20.2
Integration of heterogeneous lexicons using link format	See deliverable 20.2
OWL-Tabs for PROTÉGÉ	Development of OWL authoring tools within PROTÉGÉ, and associated tutorials, part funded by SemanticMining (WP21) (Available under open source licensing)
The CNPU coding schema	Sharing of the CNPU coding schema among WP25 partners, see http://dior.imt.liu.se/cnpu/
Open source software tools for archetype authorship, terminology binding, OWL ontology binding, SNOMED-CT binding validation, and longitudinal EHR validation	These kinds of tools are developed by several WP26 partners, and some work has already been undertaken to interface tools with each other, as clients to other partner's middleware services and as plug-ins to other partner's clients.
Sharing of resources for information retrieval reported by WP24	Sharing of databases and tools for text processing among WP24-partners: <ul style="list-style-type: none"> - FSA Library - Whatlzt - EbiMed - Paella - PCorral
MOST module for Archetype Editor	Software suite for identifying appropriate SNOMED codes for binding to Archetypes developed by Rahil Qamar, PhD Student at University of Manchester sponsored by Semantic Mining
Protégé OWL 4Alpha	Protégé-OWL has become the de facto standard editing environment for the new Web Ontology Language OWL. Ontologies developed using Protégé OWL have been developed by CNR, Freiburg, and others following the successful tutorial in December 2005 More recently, Protege4Alpha includes adaptations for new features in OWL 1.1 required for clinical applications. Protégé-OWL is a cooperative development with the UK Joint Infrastructure Services committee (JISC) and Stanford University.
Top Level Ontologies	TopBio from Universities of Freiburg and Jena; Simple-Top-Bio from University of Manchester ROMA from CNR – discussions on harmonisation in progress
BFO Top Level Ontology Manual and OWL Implementation	OWL-implementation and extensive manual concerning Basic Formal Ontology (BFO) with further material (http://www.ifomis.uni-saarland.de/bfo/home.php)



Table with information on exploitation

<i>Type / scope</i>	<i>Details / Comments</i>
UKLFR established contact with several partners for exploitation of the Morphosaurus search technology	Contact details still confidential.
UKLFR established contact to Chief Publishing Officer of Elsevier Health Sciences Division	Exploitation of the multilingual medical dictionary.
Industrial collaboration	University of Manchester/Siemens Health, 2 years. Development of Intelligent Clinical systems.
Industrial collaboration sponsored by UK DTI under the Knowledge Transfer Programme	University of Manchester / Informatics CIS, Glasgow. Development of intelligent information capture for pre-anaesthesia assessment.
Industrial collaboration (March 21)	UCL/WP26 - discussion with international EHR company about implementation of the archetype approach, and semantic interoperability
Industrial collaboration (April 20)	UCL/WP26 -presentation to NHS Connecting for Health, on archetypes
Industrial collaboration (May 9)	UCL/WP26 - meeting with commercial Knowledge Management company to discuss semantic indexing and archetypes
Industrial collaboration (September 28-29)	LiU/WP22-26 – presentation of systems and tools for major Swedish/Nordic EHR vendors
Industrial collaboration (November 11)	LiU/WP22-26 – meeting with EHR vendor on openEHR components
National ITC-strategy (November 22)	LiU/WP22-26 – meeting with Carelink, a national network of Health Care providers
Uptake of SNOMED CT	Several European medical publishers e.g. Elsevier, Royal Pharmaceutical Society of Great Britain (for the British National Formulary), are now starting to utilize and incorporate SNOMED CT into their online medical resources.



Table with research applications and funding

<i>Title of research application Research foundation (national, European etc.)</i>	<i>Partners</i>	<i>Comments (duration, funding etc.)</i>
BOOTStrep (Boostrapping Of Ontologies and Terminologies STRategic REsearch Project)	FSU-JENA (coordinator), Uni Rennes, USAL, UKLFR, EMBL-EBI, CNR-ILC, UoM, IR	proposal approved: EU funding, 2006 - 2009
Semantic Retrieval in Clinical Documentation Systems	UKLFR (coordinator), PUC-PR, UFRGS (Brazil)	approved (three years, funding of mobility)
BioMeld – data models and terminology for biobanking, FP6 Call4	LiU, UOM, UCL, KI, NTNU et.al	proposal not approved
Q-REC: EU FP6	UCL	start 1/1/06
SemanticHealth: EU FP6	UCL, UoM	start 1/1/06
ANEUR-IST	DIM, UKLFR	start 1/1/06
MultiMatch	DIM, IST-CNR	negotiations
EHR+G: EU FP6 IP	IFOMIS, EBI, EUROREC	proposal not approved
RIDE: A Roadmap ... EU FP6 CA	IFOMIS, CNR (Rome), EUROREC	approved, start Jan 2006
@neuroIST	UKLFR, DIM	approved
SYMBIOMATIC, FP6 SSA	EBI et.al	approved
FP7 Patient Safety - DEBUGIT	Five core partners of SemanticMining	approved, start Jan 2008
+ a series of national research applications related to SemanticMining		