



## **SemanticMining**

*NoE 507505*

### **Semantic Interoperability and Data Mining in Biomedicine**

## **Periodic Activity Report 3**

**Covering period 2006.01.01 – 2007.06.30**

Report Version: 1

Report Preparation Date: 2007.09.15

Classification: RE

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

<b>FINANCIAL/ADMINISTRATIVE CO-ORDINATOR</b>	<b>1</b>
<b>EXECUTIVE SUMMARY</b>	<b>2</b>
Objectives	2
Research gaps	2
Work plan of SemanticMining	3
Dissemination and exploitation	5
Management	5
Joint Research	6
Partnership	7
<b>PROJECT OBJECTIVES AND MAJOR ACHIEVEMENTS</b>	<b>8</b>
General objectives	8
Risk or problem anticipated	8
Deviations from Plan	9
<b>PROGRESS OF THE WORK PROGRAMME</b>	<b>9</b>
Milestones	9
<b>Progress report</b>	<b>11</b>
Management	11
Mobility and doctoral programmes	12
Dissemination	16
Joint research programme (WP20-27)	21
<b>Deliverables</b>	<b>42</b>
<b>Performance indicators</b>	<b>45</b>
<b>CONSORTIUM MANAGEMENT</b>	<b>47</b>
Changes of the consortium	47
Project Meetings	48
Sharing of resources, tools and material	51



---

<b>Mobility</b>	<b>52</b>
-----------------	-----------

<b>ANNEX – PLAN FOR USING AND DISSEMINATING THE KNOWLEDGE</b>	<b>53</b>
---	-----------

<b>A.1 Conferences, workshops, demonstration etc. attended/organised/foreseen by the project</b>	<b>53</b>
--	-----------

<b>A.2 Articles published, development web sites etc.</b>	<b>55</b>
---	-----------

<b>A.3 Participation in standardisation work</b>	<b>60</b>
--	-----------

<b>A.4 Patent applied for, contact and agreement for the exploitation</b>	<b>62</b>
---	-----------

<b>A.5 Research applications and funding</b>	<b>63</b>
--	-----------



---

## **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](mailto:hans.ahlfeldt@imt.liu.se)

*Project websites:* [www.semanticmining.org](http://www.semanticmining.org)

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



---

## Executive Summary

### Objectives

The general objective of the Network of Excellence entitled Semantic Interoperability and Data Mining in Biomedicine [SemanticMining] funded by the European Sixth Framework Programme, is to establish Europe as the international scientific leader in medical and biomedical informatics. The long-term goal of the network will be the development of generic methods and tools supporting the critical tasks of the field; data mining, knowledge discovery, knowledge representation, abstraction and indexing of information, semantic-based information retrieval in a complex and high-dimensional information space, and knowledge-based adaptive systems for provision of decision support for dissemination of evidence based medicine.

### Research gaps

Biomedical informatics is the emerging field where data from lower levels of molecules and cells are integrated and put into a common framework with higher level data originating from persons or populations. Biomedical informatics is a multidisciplinary discipline, which could be described as being formed at the cross-road between problems and challenges put forward by life sciences and the potential of technology as to problem solving. Indeed, there is a great potential for synergy between bioinformatics and medical informatics with a view on continuity and individualisation of healthcare, allowing for all the derived benefits to the population. Main objectives of biomedical informatics are the improvement of health and quality of life of the individual as well as to reduce the overall cost for the health care system.

An overall objective of the European research programmes is identification and filling of gaps in the European research infrastructure, to facilitate cross-fertilisation between scientific disciplines and to establish a durable structure for such as collaborative approach at a European level. Traditionally academic departments in the domain of biomedical informatics have their roots either in computer science, system engineering (including a variety of engineering disciplines) or in a biomedical or clinical context. A collaborative effort between the disciplines is suggested as a way to bridge the current gap between them, so that interdisciplinarity and synergies are exploited to the maximum effect.

Another bridging activity addressed is knowledge transfer and co-operation between academia and organisations in the health and welfare sector, including standardisation bodies and the different public and private institutions involved in health care delivery and management. The national institutes and organisations responsible for policy making and quality management with a regulatory and normative function will have an important role to play in the exchange of ideas and experiences. We believe that co-operation between these organisations and those involved in research departments needs to be strengthened, both in the early phase of research programme identification and in the later phases of implementation and large-scale evaluation of results and impact.

Another obvious gap concerns the language barriers in Europe. Although English is a de facto international language, there is a gap between the large corpus of scientific and health related text written in English and the non-native English population.



Another language barrier, relates to the difference in laymen terminology versus health care professional language.

Interdisciplinary gaps could also be identified. There are still gaps between different sub-disciplines such as computer linguistics and text mining and “structured” database applicants. Different views exist on principles for ontology construction from philosophy, computer science and practical users. Technically, semantic interoperability gaps exist when it comes to communication and pooling of data between and from different information systems.

In the following section, the work plan of SemanticMining is described in response to these identified gaps.

## Work plan of SemanticMining

Improved information handling within the health care system is considered as one of the key factors for the further development of cost-effective and high-quality health care services. The challenge of reuse and pooling of information is often addressed, and sometimes expressed as the problem of *semantic interoperability*, which simply means that semantics is preserved in communication between information systems, a condition which should be natural but has proven to be very hard to achieve, especially in the complex application area of health care and at a time when combined advances in life sciences and information technologies are increasingly modifying the practices of the domain. Thus, a main concern of SemanticMining is semantic interoperability.

**Main goal of  
SemanticMining**

It is well known that the health care system is faced with a series of challenges concerning quality and cost-effectiveness. The distribution of health care services in ways which allow the patient to take an active part in relevant decisions and the provision of evidence-based medicine at all levels in the system and the effective use and reuse of information are all key issues for the organisation of health care delivery in Europe. The information and communication technology infrastructure should reflect a view of the health care system as a seamless system where information can flow under the necessary forms of regulation, across organisational and professional – and national – borders.

The need for cross-referencing between biological and clinical information provides a grand challenge. The vast amount of data available in bioinformatics databases together with the growing volume of electronically available clinical information calls for automated (or at least semi-automated) methods for high-quality indexing, annotation, and cross-referencing through discovery of patterns and relationships. Thus there is a need for harmonisation and resources for the integration of data derived from divergent sources of the sort which ontology can provide.

**Terminology  
systems in  
laboratory  
medicine – WP25**

Text mining may play a vital role in ontology design. By exposing relationships between terminology entities in biomedical text, it can assist in the construction, refinement and validation of ontologies. Ontologies in turn can support text mining by providing a framework

**Principles in  
ontology  
engineering –  
WP21**



for clustering synonyms and structuring terminologies, and defining the types of entities and relations that text mining aims to discover.

Control over semantic overlap between terminology systems is a major challenge. Representation by a reference ontology provides a foundation for discovery of such overlaps, but, several large-scale medical terminologies still fall outside of any formal representation. However, valuable insight into the content of the terminology systems may be obtained through text mining; statistics on occurrence and co-occurrence of words and phrases can assist the semantic analysis and highlighting of potential semantic overlap.

**Text mining in  
bioinformatics –  
WP24**

Research carried out with language technology in the network address the need for approaches in Europe which will bridge language barriers and facilitate access for non-English native persons to the large scientific corpus of texts written in English. Because patients reports are written in national language all over Europe, such cross-language abilities are needed to promote a unified and ubiquitous health care system across Europe.

**The construction  
of a multi-lingual  
medical  
dictionary –  
WP20**

In some countries, patients already have or soon will have access to their own health records over the Internet, and hence there is a growing need for online facilities that can help patients without medical knowledge to access relevant information in the health records. In some cases it is even required that the records not only be made available as-is, but also that the patients should be able to receive their records in a generally understandable form.

**Patient  
empowerment  
through language  
technology –  
WP27**

A central problem in ontology engineering, although not specific to the medical domain, is the so-called *boundary problem*. Boundary problems arise when more than one model is used at the same time for a specific purpose and the source models overlap semantically. An example might be when an information model of the overall structure of the electronic health record (e.g. HL7) is used together with a terminology model (such as SNOMED CT). This situation is ubiquitous in medical informatics where models to represent instances of care phenomena (information models), e.g. a specific service request, may (and often do) conflict with models to represent types of care phenomena (terminologies), e.g. the type of service requested.

**Principles in  
ontology  
engineering –  
WP21**

**SNOMED CT –  
WP22**

Electronic Health Records (EHRs) are becoming widely available, supporting clinical data storage and retrieval, at present mainly for the benefit of the local health care provider. However the capabilities of these systems are often still far from what might be expected from an information system dedicated to the support of clinical care, in terms of completeness and precision of the clinical information, and the ability to support knowledge-based clinical decision-support, data retrieval and aggregation.

Considerable effort has been invested over the years by the standardisation community of CEN TC251 (and the HL7 community in USA) in advancing the formalism of the EHR, specifically addressed in EN13606, a forthcoming CEN standard for EHR architecture. A specific contribution of EN13606 is a standard for *archetypes*, which have been pioneered by the *openEHR* foundation. The combination of the EN13606 information model describing sections and rubrics in the EHR, and the different terminology systems

**The electronic  
health record –  
WP26**



used when specifying the instances of these rubrics for a particular patient, offer the principal boundary problem described above.

Health and health care are not only important for each individual but also important indicators of the state of a society. Therefore statistics about health are an important part of the information system. Issues in focus are the scope of health and health care statistics, the tools used for coding and classification, as well as problems of quality and comparability of data. A basic research question is how the move from traditional classifications to reference terminologies may improve the quality of health statistics. While several coding systems are utilised in health care domains such as diagnoses, health problems, and interventions, the challenge is to allow aggregation according to different aspects and to assure high information quality on all levels of data abstraction.

**Health care  
statistics – WP23**

**Long-term goals**

The long-term goal of SemanticMining will be the development of generic methods and tools supporting the critical tasks of the field of biomedical informatics: data mining, knowledge discovery, knowledge representation, abstraction and indexing of information, semantic-based information retrieval in a complex and high-dimensional information space.

## **Dissemination and exploitation**

The joint research programme has resulted in a series of national and European research applications, all with the goal of achieving funding for continued joint research. The application addressing patient safety as part of the FP7 Call 1 has been approved. Five partners from SemanticMining form the core of this new project which will start in January 2008.

During the last period, a series of 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. 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.

## **Management**

The managerial structures of the network are a Management Office, a Board as operative steering body, and an Assembly as general decision making body. The work programme is organised in work packages, lead by an appointed work package leader. The Management Office and the Board is responsible for strategic planning and follow-up of the progress of the network (WP1-4). Work package leaders are organising the integrating and dissemination activities (WP5-9), transfer of knowledge through workshops and joint research according to plan. Joint research activities are set up in WP20-WP27.

The Description of Work has annually been updated based on feedback from the annual review process. General concerns of the review panel have been dissemination and outreach activities, integration aspects with other NoEs, and description of the work program in relation to European research gaps.



---

The revised Description of Work contains new sections with Identification of gaps in the European Research Infrastructure and Vision and Goal of SemanticMining. New material is regularly published on the web site. To ensure long-term sustainability of the web and archiving facilities after the project ending, a new Wiki-based web platform has been developed hosted in an academic environment and to be maintained in a collaborative manner.

A request for extending the project period with six months into 42 months has been accepted by the Commission. The project ended July 2007, and a closing event was organised together with INFOBIOMED in Barcelona, June 26-29.

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.

## Joint Research

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)
- impact of ontologies on health statistics (WP23)
- concept systems in laboratory medicine (WP25)
- text mining and information retrieval in bioinformatics (WP24)
- the concept-based electronic health record (WP26)
- medical terminology for laymen (WP27)

In the Description of Work (DoW) milestones and quality indicators are defined by which the consortium seeks to assess its progress. Results from the joint research programme show evidence of substantial improvement in these quality indicators, e.g. in terms of joint research publications, contributions to international workshops and conferences, sharing of resources and tools, and a series of tools and services available e.g. as open source. Measurements of quality indicators as well as follow-up of milestones are presented below.

## 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](http://www.semanticmining.org)



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



## Project objectives and major achievements

### General objectives

<i>Objectives</i>	<i>Progress towards achieving objectives</i>
<p>Main objective of the NoE is to bridge gaps in the European research infrastructure and to facilitate cross-fertilisation between scientific disciplines. Progress and evaluation of Management, Integration, and Joint Research Activities are reported below.</p>	<p><b>Management:</b> The decision making structures are according to the Consortium Agreement in operation. The Management Office handles administrative matters and communication with project officers in Brussels. The Board consisting of six members have regular meetings where the progress of the NoE is followed-up. Assembly meetings where all partners are represented are held twice a year, when decisions on strategies and resource allocation between activities and partners are taken.</p> <p><b>Integration:</b> Successful continuation of integrative events such as summer school, seven Assembly meetings, and a number of workshops on different research topics. During 2006 and 2007, the mobility program has funded a series of exchange visits by PhD students and researchers. A new Wiki-based web and archiving infrastructure has been developed to ensure long-term sustainability after the project life time.</p> <p><b>Joint Research Activities:</b> Work package leaders are organising the joint research activities and integrating activities according to plan. Evidence of improvement in quality indicators.</p>

### Risk or problem anticipated

<i>Causes and Description</i>	<i>Possible Impact</i>	<i>Corrective actions</i>
<p>No major problems are to be reported. Minor delay of some deliverables during period 2. All scheduled deliverables now published.</p>		<p>The management structure of the network is designed to monitor threats and weakness (as well as strengths and opportunities), and on a regular basis through Board and Assembly meetings take corrective actions if necessary.</p>



## Deviations from Plan

<i>Causes and Description</i>	<i>Corrective actions</i>
Termination of WP28 during 2006.  No other major deviation from plan is encountered.	Re-organisation of WP28 research programme – parts of WP28 merged with WP21 and WP26.  The Board is actively working together with the work package leaders to monitor the progress of the work plan.

## Progress of the work programme

### Milestones

In the Description of Work (Annex I) milestones are defined by which the consortium seeks to evaluate its progress.

<i>Milestone</i>	<i>Planned date</i>	<i>Actual date</i>	<i>Comments</i>
M1 Technical infrastructure supporting communication and integration	m8	2004.04.28	Release of MERMIG as communication platform.
M2 Kick-off meeting	m2	2004.01.26-27	Successful kick-off meeting with all partners in Linköping, 60 participants in total.
M2 Summer school	m8	2004.07.03-09	Over 80 participants in first years Summer School. Successful realisation of summer schools 2004-2006. First joint European Summer School in Biomedical Informatics with INFOBIOMED and BIOPATTERN 2006. Second joint summer school being organised June 2007.
M3 Co-authored research papers	m12 >1 m24 >10 m36 >40	during 2004 during 2005 during 2006	Seven co-authored papers 2004 Over 70 co-authored papers 2005-2007
M4 Artefacts/models as result of joint research	m18 >1 m30 >5 m42 >10	2005-2006	Interchange format for entities in a multi-lingual medical dictionary (July 2005). (see WP20 description) Multi-lingual medical dictionary in at least five languages with more than 20,000 entries July 2005, and over 100.000 entries Dec 2005. Prototype platform for exchange of biomedical text corpora.  Protégé/OWL – software package for



			<p>ontology engineering (see WP21 description)</p> <p>WhatizIt, EbiMed, Pcorral, Paella – software packages for information retrieval implemented at EBI. (see WP24 description)</p> <p>BFMed, EAGL, SUITSEARCH – tools for information retrieval developed by WP20 and WP24</p> <p>Archetype editor, MoST terminology search engine, TermViz terminology browser – software components as part of openEHR. (see WP26 description)</p> <p>CNPU coding schema/database (see WP25 description)</p> <p>BFO top level ontology with OWL implementation (see WP21 description)</p> <p>BioTop – an ontology in the biomedicine domain</p>
M5 Evaluation by Scientific Advisory Committee (SAC) Cross-WP-analysis	m18	April 2005 July 2005	Meeting with SAC in Rome, April 29 – May 2. Cross-WP-analysis
M5 Mobility program	m18 >1	April - Dec 2005	First and second round of mobility program with 13 PhD student exchange visits plus a series of short-time visits by researchers.
	m30 >10 m42 >20		Third round of mobility program with 12 exchange visits. 20 grants for non-NoE PhD students for participation in summer school.
M5 New joint research applications	m18 >1 m30 >5 m42 >10		12 EU applications + a series of national applications with references to SemanticMining. New FP7 project (DEBUGIT) has been approved.



---

## Progress report

### Management

WP1 Management, WP2 Assessment/planning, WP5 Planning of meetings

The managerial structures of the network are a Management Office, a Board as operative steering body, and an Assembly as general decision making body. The work programme is organised in work packages, lead by an appointed work package leader. The Management Office and the Board is responsible for strategic planning and follow-up of the progress of the network (WP1-4). Work package leaders are organising the integrating and dissemination activities (WP5-9), transfer of knowledge through workshops and joint research according to plan. Joint research activities are set up in the areas listed above (WP20-27). Seven Assembly meetings have been held with representatives from all partners. The Board have had regular meeting (physical or tele-meetings).

The executive summary of the annual review March 2006 expressed ideas for improvement of the SemanticMining work plan, which have been addressed. The main concerns of the review panel were:

- description of work program in relation to European research gaps
- description of integration aspects e.g. with other NoEs
- dissemination and outreach activities
- inclusion of conclusions in deliverables of workshops and conferences
- coherent follow-up of efforts reporting in relation to budget and work plan.

The revised DoW contains new sections with Identification of gaps in the European Research Infrastructure and Vision and Goal of SemanticMining. Follow-up of under- and overspending partners have specifically been addressed in Assembly meetings, where revised budget for the last phase of the project have been discussed and decided upon. New material is published on the web site (see [www.semanticmining.org](http://www.semanticmining.org) under NoE Description and News). To ensure long-term sustainability of the web and archiving facilities after the project ending, a new Wiki-based web platform is under development hosted in an academic environment. A series of specific outreach activities have been organized during 2006 and will be reported below. An inherent problem with the NoE as an instrument for achieving higher level of integration in the European research arena is the short funding period. In order to avoid an abrupt termination of the network, a request for extending the project period with six months into 42 months have been put forward and been accepted by the Commission. The project ended July 2007, and a closing event was organised together with INFOBIOMED in Barcelona, June 26-29. The joint research programme has resulted in a series of national and European research applications, all with the goal of achieving funding for continued joint research. An application to FP7 Call1 on patient safety has been approved (DEBUGIT) and is currently in a contract preparation phase. Five partners from SemanticMining form the core of this new project, scheduled to start January 2008.

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 doctorial consortium (PhD student centred), tutorial and workshops, a “best of” scientific paper session, a NoE cluster meeting, and NoE specific administrative meetings. The objective of the NoE cluster meeting was to share experiences of management, examples of good practise, and to discuss future initiatives from the biomedical informatics community in Europe. The following aspects were addressed by the coordinators from the three NoEs: integration, contractual, management, dissemination, and



---

future approaches. As a result of the discussions, the following areas were identified as a common priority list for the NoEs.

- Information sharing (deliverables, events etc.)
- Student mobility
- Sharing of tools and resources
- Event organisation and participation
- Sharing of best practice management, gender balance, collaborative research etc.
- New funding and collaborative opportunities
- Shared research activities

Discussions were held on future opportunities focused on the upcoming FP7 call and its potential priority areas. The SYMBIOMatics priority list presented at the ICT for BIO-Medical Sciences Conference, June, 2006, were discussed, as well as different opportunities for joint applications in the form of NoEs, IPs or STREPs.

As a continuation of the first joint summer school, an 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.

Internal SemanticMining strategic discussions have focused on opportunities for cross-WP-integration. Resource allocation between different types of activities in the network is continuously discussed by the Board and decided by the Assembly meeting where all partners are represented. Concrete results of these strategic discussions for the last phase of the project are an increased resource allocation on research and dissemination in relation to infrastructure and internal knowledge sharing, and on “high integrative” work packages (WP20, WP21, WP24, WP26, WP27).

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.

### **Mobility and doctoral programmes**

The objective of the third year of mobility program (WP6) was to build on and extend the mobility activities initiated in the first two years. To contribute to these objectives, mainly two activities were undertaken.

Firstly, we have been keeping a record of the mobility activities of PhD students in the network. We have been keeping track of mobility activities, primarily through the visit grants scheme application and decision process. Currently, this information is made accessible via the MERMIG platform to all network participants. In collaboration with WP4 (Public Website), a subset of this information will also be made available to the general public.

Secondly, we administered the visit grants. During the last period both students and researcher could apply for grants (maximum per person 1500 euros). Awards were administered by the Open University. Decision on the awarding of grants was made by the



---

Board upon receiving recommendations from the WP6 administration team. We sent out announcements to the semantic mining mailing list to draw attention of network members to the scheme, we advertised the scheme on the WP6 homepage and at the Semantic Mining Summer School. Twenty four mobility grants and twenty grants to non-NoE PhD student for attending the summer schools have been awarded. A total of 29 visit grants ranging from one week to three months have been approved.

Thirdly, as in the first period, WP6 contributed to the organization of the annual Summer School of the network. We also organised and successfully run a doctoral consortium which took place during the Summer School. We had discussion sessions, a presentation session, a poster sessions and collective activities.

The last period was successful for the mobility programme. The visit grants scheme stimulated new research collaborations among network partners involving PhD students, junior researchers and senior researchers at host institutions. Both the doctoral colloquium and the gender panel at the summer school were well attended and feedback on these activities was very positive. Reports with descriptions of research activities are compiled and made available on the SemanticMining web. The exchange visits are listed in the tables below. Lessons learned regarding doctoral programmes are that formal co-tutoring is still difficult due to administrative regulations, but that scientific cross-fertilisation among partners, definitely has a very positive effect on the progress of PhD students doctoral programmes. Several PhD students in the network now have established close contacts with tutors at other partner sites.



<b>Mobility program 2005</b>				
<b>Name</b>	<b>Duration</b>	<b>From</b>	<b>To</b>	<b>Award</b>
Imad Tbahriti	3 months	Geneva	EMBL-EBI	1500
Louise Deleger	1 week	INSERM, Fr	Linkoping	1500
Mikael Nystrom	1 week	Linkoping	Freiburg	1500
Karim Nashar	1 month	Manchester	EMBL-EBI	1200
Mikel Egana	1 month	Manchester	EMBL-EBI	1200
Anna-Karin Hermansson	1 week	Goteborg SU	Freiburg	1390
Gaston Burek	3.5 week	OU	INSERM, Fr	1500
Alexander G. Castro	1 month	EMBL-EBI	Manchester	1200
Philip Daumke	2 weeks	Freiburg	EMBL-EBI + Jena	1475
Michael Poprat	1 week	Jena	EMBL-EBI	1000
Eric Sundvall	1 week	Linkoping	CHIME, UCL	1342
Vincent Claveau	1 week	INSERM, Fr	Freiburg	1000

<b>Mobility program 2006</b>				
<b>Name</b>	<b>Duration</b>	<b>From</b>	<b>To</b>	<b>Award</b>
Felix Balzer (student)	2 days	University of Freiburg	University of Jena	220
Audrey Baneyx (student)	9 days	INSERM	CNR-ISTC	960
Rahil Qamar	10 days	University of Manchester	Linkoping University	1480
Daniel Schober	12 days	EMBL-EBI Cambridge	IFOMIS University of Saarland	1400
Olivier Steichen (student)	7 days	INSERM	Linkoping University	1400
Julie Nies (student)	7 days	INSERM	Linkoping University	1400
Ines Jilani (student)	6 days	INSERM	Univeristy of Genève	1500
Marie-Christine Jaulent	6 days	INSERM	Univeristy of Genève	1500
Irena Spasic	12 days	University of Manchester	EMBL-EBI	1480
Natalia Grabar	10 days	Univeristy of Genève	The Open University	1500
Felix Balzer (student)	2 days	University of Freiburg	University of Jena	220
Audrey Baneyx (student)	9 days	INSERM	CNR-ISTC	960



---

<b>Mobility program 2007 Name</b>	<b>Duration</b>	<b>From</b>	<b>To</b>	<b>Euro</b>
Lars Borin	2 days	Göteborgs University	Open University	864
Dimitrios Kokkinakis	2 days	Göteborgs University	Open University	864
Maria Toporowska Gronostaj	2 days	Göteborgs University	Open University	864
Dana Dannélls	2 days	Göteborgs University	Open University	570
Attila Nagy	5 days	Debrecen University	CNR-LAO	790
Sandra Williams	2 days	Open University	INSERM UMR_S 729	839



## Dissemination

SemanticMining web site (WP3-4), Summer School (WP7) and Dissemination (WP8-9)

The MERMIG platform is used as public website available at [www.semanticmining.org](http://www.semanticmining.org), and as internal network communication platform and repository. The web site is regularly updated and populated with new material. During the last year material from the major workshops and conferences has been made available. Site performance and statistics are monitored.

A significant role of this NoE is to provide educational material based on both the 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 last two years, there has been a significant increase in co-authored research papers within the network.

To ensure a sustainable web platform after the ending of the project, a new web platform based on the Wiki-technology, hosted by an academic partner, is under development.

During the last period 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 itself 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.

### *International Symposium on Biomedical Informatics in Europe*

As a continuation of the first joint summer school, an 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. We offered four tutorials on Sunday, April 9. The scientific program started on Monday, April 10th and ended on Wednesday, April 12th. It included three keynote talks, panel discussions and an industry exhibition. All papers (except those that were selected for a journal reviewing process) and posters were published online at CEUR-WS (<http://ftp.informatik.rwth-aachen.de/Publications/CEURWS/Vol-177/>). Five papers were selected for a second reviewing process in order to be published in the BMC Bioinformatics journal.

### *Training Course in Biomedical Ontology*

In May 2006, a three-day training course 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.

The course was set up in groups of block lectures by the speakers Barry Smith: Introduction to Biomedical Ontologies; Werner Ceusters: Biomedical Ontologies and the Electronic Health Record; Olivier Bodenreider: On Mapping, Aligning and Integrating Biomedical Ontologies; Mark Musen: Case Studies in Ontology Development. It also included a discussion session on the final day.

The quality and composition of the training course pleased most participants as reflected in the feedback we collected. Due to large number of interested people who were unable to participate due to the limited numbers and due to the positive feedback from the participants, we intend to carry out another similar event at Dagstuhl in June 2007. The preparations for this are ongoing.

### *Swedish Terminology Conference*

The objective of the Swedish Terminology Conference, held in Kalmar, September 28-29, 2006, 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 through 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, 2006, 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, in April 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**

One important way of disseminating the scientific results of the Network partners has been the interaction with relevant standards organisations for health information. This is co-ordinated by WP8 but include participation of most of the research work packages.

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](http://www.ehcs.org)) with information on all major eHealth standards and activities.
- The further work on finalising the EN 13606 Health Informatics - Electronic Health Record Communication series (in co-operation with WP26) now also being balloted as an ISO standard and in close co-operation with HL7. 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.
- The finalisation of the EN 12967 Health Informatics - Service Architecture (HISA) standard.
- The work on the EN 1614 model for representing a Structure for nomenclature, classification, and coding of properties in clinical laboratory sciences. After intensive discussions at SemanticMining meetings with WP25, a new version was established. During 2006, EN 1614 has been finalised and approved as European standard: "Health Informatics — Representation of dedicated kinds of property in laboratory medicine". The standard provides a metrology and terminology framework for Laboratory Medicine developed within the NoE and the Committee on Nomenclature Properties and Unit of the IFCC and IUPAC. The new EN 1614 is now being used as input to the LOINC–C–NPU SNOMED CT mapping discussion.
- 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. During 2006 the standard was sent out for enquiry, receiving valuable comments from, among others, NoE participants.
- A CEN Technical Specification for medical knowledge resource metadata descriptions has been developed, Clinical knowledge resources - Metadata (MetaKnow).
- 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).

#### **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 was 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, public-friendly material has been produced where results 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.

## Joint research programme (WP20-27)

### Multi-lingual medical dictionary (WP20)

The main objective of WP20 is the creation, standardization and pooling of a multilingual medical dictionary. The WP20 activities are mainly characterized by the following strands of collaborative work:

- Joint elaboration of a common interchange format for lexical information and for corpora.
- Elaboration and maintenance of lexical sources at different partner sites (at LIU, UKLFR, UGOT, DIM). Export of these sources to a common platform, according to the interchange specifications
- Semi-automated lexeme acquisition
- Use of multilingual lexicons in prototypical applications and research scenarios.

### Main activities and results

The WP20 activities were characterized by the continuing population and maintenance of lexical sources at different locations, using manual and automated lexicon acquisition methods, and the export of these sources to a common platform, according to the interchange specifications.

In particular, the following activities have been reported by WP20 partners:

- Word alignment techniques on parallel French-English medical corpora are used to extract pairs of French-English and Swedish/English words and terms (INSERM / LiU)
- An LREC 2006 satellite workshop was organized. *Acquiring and representing multilingual, specialized lexicons: the case of biomedicine. LREC workshop Genova, Italy, 2006. Edited by P. Zweigenbaum et al.*
- Concept similarity is measured using latent semantic indexing, both between thesaurus relations and predications extracted from unstructured text (OPEN)
- Electronic health record texts are used to compile of a Swedish primary health care corpus (UGOT).
- The Gothenburg MedLex database was enhanced with lexical information, using semi-automatic acquisition techniques of medical vocabulary from corpora and alignment of medical lexicons (UGOT).
- LiU has started an evaluation of which kinds of word alignment techniques that are best suited for word alignment of rubrics from medical terminology systems.
- Jointly with WP27, text was used for differentiating and distinguishing between expert and layman medical vocabulary, for further addition of this information to the MedLex database and later to the multilingual medical dictionary (UGOT, SU)
- The 2006 OntoLex workshop was organized, bringing together lexicographers, ontologists and computational linguists: *Proceedings of OntoLex 2006, 27 May 2006, Genoa. Edited by A. Oltramari et al.*
- A common link format to represent semantic links between lexicon entries was proposed by DIM. After the first cross-mapping experiments the format was modified by LiU.
- A semantic cross-mapping of the existing lexical sources was done in accordance with the common link format, using morphosemantic indexing (UKLFR).
- New entries were added to the MorphoSaurus lexicon. A framework for auditing the lexicon acquisition process was developed. Italian was included as a new language for the MorphoSaurus lexicon (UKLFR).
- A new version of the MorphoSaurus vocabulary editor for was developed.

- Information on lexicon interchange format was provided to participants of the new BootSTREP project, in which both UKLFR and JENA are partners
- A platform for corpus exchange was proposed and prototypically implemented (JENA).
- An evaluation study was initiated, measuring the correctness and the completeness of the multilingual dictionary. The rather poor results were partly due to an error in the mapping algorithm. As a consequence, the dictionary was re-generated and the manual evaluation will be re-done in January. As a consequence a delay in the delivery of the evaluation report may occur.
- At UKLFR, a spin-off company (AVERBIS) is being founded in March 2007. The purpose of this spin-off will be to bring the multilingual MorphoSaurus system to the German market. The spin-off will be subsidized for the first two years by a grant from the German federal state Baden Württemberg.
- WP 20 was contacted by one of the leading medical publishers interested in the representation formats and lexicon acquisition techniques, which might be useful for a further development of their multilingual lexical databases. It is planned to resume the contact in 2007.
- In interaction with the SemanticMining board, WP 20 developed different plans for embedding the dictionary activities into a new FP7 project. This activity is being led by INSERM.

In detail, the activities are demonstrated in the publications which were written during the reporting period.

The screenshot displays the SUITSEARCH HEALTH RECORDS web interface. At the top, the search term 'juckendes erythem' is entered in the search bar. Below the search bar, there are fields for PIZ, Nachname, Vorname, and Geschlecht. The search results are displayed in a table with four entries:

ID	Name	Address	DOB	Gender
22990373	Presley	Elms	08.01.1935	♂
25724135	Hendrix	Jimi	27.11.1942	♂
22118404	Joplin	Janis	19.01.1943	♀
21965936	Kelly	Grace	12.11.1929	♀

Each entry includes a brief description of the medical condition and the date of creation and transmission.

Screenshot 1. Web-based document retrieval supported by the multilingual dictionary (WP20).

### **Progress towards achieving objectives**

The general objective of the NoE, the cross-fertilization between scientific disciplines has been addressed by this work package by promoting joint activities involving computer scientists, biomedical domain experts, and linguists, all of them covering several European languages. During the reporting period, the main event was a satellite workshop of the LREC 2006 conferences: “Acquiring and representing multilingual, specialized lexicons: the case of biomedicine” which was held on May 23, 2006 in Genoa. The goals of this workshop were to present and disseminate WP20 work to a Computational Linguistics audience on the one hand, and discussions with researchers involved in lexicon representation standards and distributed lexicon development on the other hand. 10 papers have been submitted among which 8 were selected. 2 additional presentations were invited: one on the Lexical Markup Framework, an ISO initiative to design a standard for lexicon representation, and one on the *Papillon* Project, which has been organizing distributed work on a general multilingual dictionary for five years. The workshop was attended by 35 participants during the whole day, with lively discussions which extended beyond the scheduled time.

The common repository of medical terms in different languages was iteratively fed by new entries and a final version was released in November. However, the lexicon still remains largely heterogeneous in terms of coverage and granularity. It became obvious that the input necessary to upgrade the growing repository toward a exploitable resource will widely exceed the resources available. This is not a surprising result, compared to the high effort necessary for traditional lexicon maintenance. However, in order to warrant the sustainability of the present work, steps will have to be taken toward new partnerships and projects. For the time being, our impression is that the upcoming EU FP7 calls will not be well suited for realistically acquiring funding for this purpose. Other partnerships will therefore have to be analyzed. Thanks to the intermediation of Janine Ross, a contact to the Chief Publishing Officer of Elsevier Health Sciences Division has been initiated. This contact may be interesting, since Elsevier publishes the Dorland Medical dictionary and is willing to enhance it to become a computable lexical database.

Due to the importance of domain-specific corpora, the decision had been taken to add an additional task to the work plan, the pooling of biomedical corpora as they exist at different locations. A prototype of a corpus interchange platform was prototypically implemented, following the specifications of WP20.3, submitted within this reporting period. However, the WP decided not to use this platform in this project due to other priorities.

### **References to quality indicators and milestones**

The lexicon data from different partners were collected on a common platform and disseminated as deliverable 20.2. According to milestone four in the work program, a multilingual medical lexicon is principally available. An assessment of the WP20 activities against the indicator yields the following results:

- Q1: Workshops and symposiums. According to the characteristics of WP20 as a technical work package, meetings and symposiums organized by this group during the first two years are mainly restricted to internal meetings. Three events were noticeable in the reporting period.
  1. Work package meeting in Geneva (January)
  2. International workshop in Genoa (May)
  3. Participation of WP20 at the kick-off meeting of the new work package WP27 which uses tools and resources developed in WP20 (February)
  4. Work package meeting in Jena (September)



- Q2: As mentioned above, the sharing of resources is a central objective of WP20 and the common repository provides evidence for this.
- Q5: Informal tutoring support could be registered in several cases.
- Q6: There were several short visits between WP20 partners
- Q7: There were a total of 14 research papers co-authored by WP20 partners.

### **Conclusions**

WP20 has met or even exceeded its target in terms of joint research, resource creation and dissemination. The joint scientific production provides good evidence that the network is meeting its higher level goal of integration and cross-fertilization. The assessment of the usefulness of the multilingual dictionary is ongoing.

### **Ontology engineering (WP21)**

The main objectives of WP21 are to share understanding on principles for ontology engineering, to collaborate in research and to give input to standardisation bodies. In summary, WP21 contributions are:

- The Workshop on the Foundations of Terminologies and Classifications was organised as part of the European Federation of Medical Informatics Special Topics conference in Timisoara, Romania. One invited key note was given by Alan Rector, work package leader.
- Participation in the Swedish Terminology Conference, with presentations and demonstrations of methods and tools developed in WP21 and WP26.
- The second and third of a series of joint workshops with WP26 on the interaction of terminologies and ontologies with electronic healthcare records was held in February and November, respectively in Paris. The work has resulted in a major set of developments now awaiting publication, the first fruits of which were published as KR-MED 2006 and will be republished in the journal Applied Ontology. Further publications are in preparation.
- Out of these workshops also grew exchanges held between the Universities of Manchester and Linköping for the development of tools for binding terminologies to EHR Archetypes.
- Additions to the Protégé-OWL toolset to accommodate specific requirements for medical Ontologies included in the new OWL 1.1 specification have been undertaken and are now incorporated in the new 4-Alpha release.
- The work package participants contributed to the SemanticMining Conference on SNOMED in Copenhagen in October at which the work package leader, Alan Rector, was keynote speaker.
- Work on top level ontologies has been undertaken jointly by the Universities of Freiburg and Jena and by University of Manchester. Merging and joint development are now in progress. IFOMIS provided an OWL-implementation of Basic Formal Ontology (BFO) together with an extensive manual and further material. This material is being used by partners developing FuGO. Joint publications between the universities of Freiburg, Saarland, and Manchester are in progress



- UKLFR, IFOMIS, and JENA drafted BioTop, a domain top level ontology for biology, using the whole range of OWL-DL constructors aiming at precise definitions of basic classes of the domain of Biomedicine (see screenshot 2).
- IFOMIS offered a Training course in Biomedical Ontology at Schloss Dagstuhl in May at which many participants from the network were present. A full report is available in deliverable D40.
- The ontology section of the annual Summer School was delivered in the first week of July, which this year was a joint event with other related Networks of Excellence.
- Joint work on ontology quality assurance is being undertaken with the Knowledge Web Consortium.
- Manchester, UCL, and Linköping all participated in a series of workshops sponsored by the SemanticHealth Roadmap project.

#### **Progress towards achieving objective**

Sharing understanding across disciplines: The joint summer school with two other networks of excellence was highly successful. Members of the work package have been highly active in general biological Ontologies in both anatomy and disease, through collaboration with the US National Center for BioOntologies.

Input to Standardisation: In collaboration with WP26, the work package has focused on the interaction between Ontologies and medical records. Major contributions have been made to both CEN/OpenEHR in the binding of terminology and OpenEHR Archetypes. A tool based on this work has been developed by a PhD student at Manchester as a module for the Archetype editor developed at University of Linköping. There has also been a close collaboration with WP27 on quality of SNOMED and quality metrics.

To contribute to the consensus on biomedical “upper ontology”. There are now six variant upper Ontologies linked in various degrees to different members of the consortium. Efforts to delineate the reasons for difference and harmonisation are scheduled for 2007 and beyond the end of the project.

#### **References to quality indicators and milestones**

Deliverable 21.3 on Computer Science Foundations has been delivered and is undergoing quality assurance. Deliverable 21.4 on human engineering is in final preparation following agreed delay because of staffing changes.

Annex I (DoW) of the contract defines those quality indicators by which the consortium seeks to assess its progress. A subset of these indicators is listed below:

Q1 Workshops and symposiums:

participation in five international conferences.

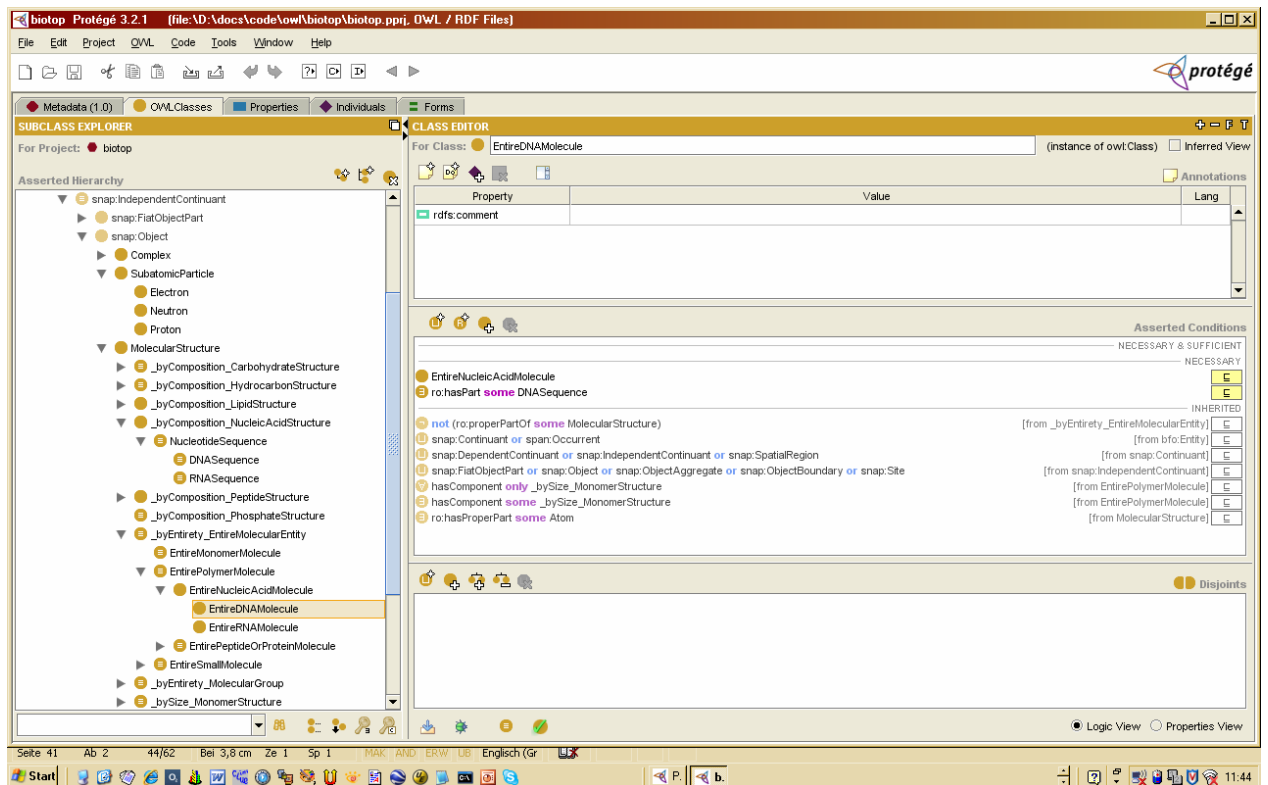
Q7 Co-authoring of research papers, reports and educational materials:

12 cross-partner co-authored papers in 2006.

Q8 Participation in standardisation work:

input to CEN TC251 and HL7 (e.g. EN 1614, EN13606, openEHR)

Q9 Jointly executed research programmes: Several research exchange visits



Screenshot 2. Ontology engineering with the Protégé/OWL tool set - segment of top-level biomedical domain (WP21).

## Conclusions

Significant contributions have been, and are continuing to be made to the standardisation process and the understanding of the interaction between terminologies, ontologies and medical records. Significant contributions have been made to tools and to the convergence of upper level ontologies.

The education and outreach efforts have been a major success with important conferences sponsored by the project – Foundations of Clinical Terminologies and Classifications, and Semantic Mining Conference on SNOMED – and to other workshops and conference, notably a key role in organising the first workshop on medical knowledge representation, KR-MED 2006.



## SNOMED CT (WP22)

A major event during 2006, was the Semantic Mining Conference on SNOMED CT in Copenhagen, October, 1-3 (SMCS 2006). This event was intended to be the first European fora for health policy makers, clinicians, nurses, system developers, computer scientists, terminologists and translators with a special focus on SNOMED CT. It should embrace both scientific presentations and invited presentations which provide an overview of current efforts and developments in the context of SNOMED CT. The feedback of the 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. SNOMED CT was recognized as a framework suitable for the content of a multilingual terminology necessary for true semantic interoperability in healthcare. Summing up SMCS 2006 was a highly satisfying event which took place at the right place, with the right content and at the right time.

In the Semantic Mining NoE we have undertaken to evaluate SNOMED CT from a number of different perspectives:

The basic structure of SNOMED in relation to Ontology research. This was done together with WP21, and resulted in a critical review of the Description Logic approach and as implemented in SNOMED CT.

SNOMED in relation to application development. A series of studies have been performed where the requirements from clinical applications for various purposes were compared with the SNOMED content. This also included some studies on its usefulness for recording nursing. Generally it was concluded that most of the terms needed existed in SNOMED CT but the relationship to the information models needed to record information remains an important issue.

SNOMED was also studied in relation to the EHR models and archetype work, particularly openEHR and EN 13606 together with WP26.

The result of the evaluation studies of Semantic Mining as summarized in deliverable D22.2 has been published and discussed also in a number of publications and in the large conference on SNOMED CT that was organised by the work package and reported in full in deliverable D33.

Research work includes mapping of SNOMED CT terms to legacy classification systems such as ICD10 (International Classification of Diseases), NCSP (Nordic Classification for Surgical Procedures) and ICF (International Classification of Functioning). Experiments of health statistics based on the SNOMED CT hierarchies and ICD10 coded data from the National Danish Patient Registry indicate the potential of using SNOMED CT as aggregating tools for producing health care statistics. Work on aligning SNOMED CT with the C-NPU format and the European standard EN 1614 on laboratory requests and reports has also been conducted. Modelling issues regarding results of examinations as attached to procedures representing the activities performed to obtain the result (in SNOMED CT the *Procedure* hierarchy) versus attached to kinds-of-property (in SNOMED CT the *Observable Entity* hierarchy) is particularly worked on. Based on these considerations active collaboration is now established with the SNOMED CT concept model working group and the C-NPU community.



SNOMED CT is currently being translated to Danish in preparation for implementation of EHR systems that can utilize a medical terminology. Since classification systems as the Nordic Classification of Surgical Procedures (NCSP) are currently used for reporting of medical events to national patient registers, a map to the NCSP could facilitate comparable statistics over time and a continued use of the currently implemented DRG systems. Translation of SNOMED CT into other languages is also planned (see description of IHTSDO below).

There exist map tables from the concepts in SNOMED CT to codes in various classifications. However, a map table to NCSP has never been published. This report describes a method for creation of a map from the concepts in SNOMED CT to the classification codes in the NCSP. New versions of both SNOMED CT and NCSP are released on regular basis which mean that the requirements for the method should include support for updating of the map tables.

The method takes offset in the possibility to generate subsets of SNOMED CT concepts via indexes. A table containing the approximately 7.000 NCSP codes assigned with 23.000 attribute-value pairs was generated according to the rules set by SNOMED CT. This NCSP-relationship table was then queried against the SNOMED CT relationship table and it was thereby possible to generate a map table from SNOMED CT to NCSP. The resulting map tables contain 17.000 rows representing maps from SNOMED CT procedures to NCSP. The NCSP relationship table can be updated when new releases of the NCSP are published and the map table can subsequently be updated when new releases of SNOMED CT are published. WP22 has also conducted a mapping project between SNOMED CT and International Classification of Functioning (ICF). ICF is relatively new classification, published in English by WHO in 2001, and with publication in Swedish in 2003 and in Finnish 2004. On the basis of the experiences from two pilot studies in Sweden a joint pilot study concentrated on a complete mapping of the ICF *structure* dimension (s) and on certain areas in the *activity/participation* dimension (d). The different structures used in the dimensions of ICF and the hierarchies of SNOMED CT make mapping problematic, but a mapping between the two systems is nevertheless desired by the health care organisations.

Translation of SNOMED CT into Danish is ongoing. Founded on the experiences from the translation from US English to Spanish the translation is done in to steps. In the first step bilingual resources were joint in translation in a concept based manner i.e. translation of terms with respect to their relationships in the SNOMED CT terminology. In the second step, Danish clinicians validate the concepts and add synonyms. This method is essential for the preservation of the medical knowledge that is comprehended in SNOMED CT.

Finally, in collaboration with WP24, a SNOMED text categorizer has been set up. The system accepts any document as input and output a set of SNOMED categories. Every SNOMED category is provided with an estimate of its relevance for the given input text. The tool can be tested online - select "SNOMED Categorization" (<http://129.194.97.165:8081/EAGLb/>).

#### **Assessment against relevant quality indicators**

Q1 Workshops and symposiums

Participation in three major terminology conferences (Denmark, Sweden, Romania).

Q9 Jointly executed research programme

Cross-WP activities with WP21, WP23, WP25 and WP26..

Q10 Key characteristics of partners

Work package contributing to increased cooperation between public health organisations and research department.



---

## Conclusions

On the international arena, further development of SNOMED CT is now the responsibility of the newly established organisation International Health Terminology Standardization Development Organisation (IHTSDO). As part of the IHTSDO organisation, four committees have been established (Content, Technical, Quality, Research and Innovation). As a result of recognition of the SemanticMining network, seven persons from SemanticMining have been selected for these committees (Lars Berg, Marie-Christine Jaulent, Mikael Nyström, Jeremy Rogers, Erik Sundvall, Stefan Schulz, Hans Åhlfeldt). Moreover is the IHTSDO management office in Copenhagen run by persons with an active background in SemanticMining.

SNOMED CT-related activities are also reported by WP21 (e.g. the TERMINFO project dealing with the terminology binding problem between HL7 and SNOMED CT), WP23 (use of SNOMED CT as aggregating tool for health statistics), and WP26 (where EHR archetypes are instantiated with SNOMED CT terms). All these issues will be further addressed by the new IHTSDO organisation, and by several national research initiatives.

### Interaction with other work packages

SNOMED CT-related activities are also reported by WP21 (e.g. the TERMINFO project dealing with the terminology binding problem between HL7 and SNOMED CT), WP23 (use of SNOMED CT as aggregating tool), WP25 (mapping between SNOMED CT and CNPU), and WP26 (where EHR archetypes are instantiated with SNOMED CT terms).

### Health care statistics (WP23)

The main objective of this research activity is to share experience, understanding and development of statistical methods for measuring information quality, ontologies for health indicators, and methods for quantification of semantic distance. Moreover, the objective is to encourage sharing of data material (e.g. quality registries and coded patient data) applicable for development and evaluation.

The first phase of this WP has been devoted to compilation of background and baseline material in the field of health statistics, with a natural focus on the situation in Europe. Issues in focus are the scope of health and health care statistics, used tools for coding and classification, problems of comparability and quality of data. A basic question is how the move from traditional classifications to reference terminologies may improve the quality of health statistics. A specific aspect of this is the use of SNOMED CT as aggregating tool in the production of reliable health statistics. Documentation of problems in European health statistics was completed in the report submitted as Deliverable D23.1, which also contained examples of the connection between classification, terminology and ontology.

During 2005 and 2006 activities have mainly been centred on the WP's third task (Task 23.3), namely proposal for methods for measuring reliability and semantic distance. We have commenced work on a MATLAB work bench in which to test statistical approaches to reliability measurement under various simulated and controlled circumstances.

During 2006, activities have been focused on the planning and realisation of a cross-European study on semantic distances. The aim of this study is to examine whether physicians agree on semantic distances between pairs of words or phrases. The study is based on a set of 118 pairs compiled by the work package participants, where test subjects use visual analogue scales to rate the perceived semantic similarity in two different ways. Agreement is then measured as the rank correlation between judges' ratings.

The study has been completed and will result in a scientific paper co-written by the work package team. Lately, a number of papers dealing with issues related to this topic have been



---

published, but none of them have used such an extensive and empirical approach as in this study. We are confident that our study will generate a great deal of interesting topics for further exploration by us and others. Our conclusion is that there is great interest in the topic of semantic distance and that the efforts of work package 23 will result in an interesting paper that will illuminate the topic of semantic distance. In turn, this will be useful in various applications of information retrieval and statistics.

#### **Assessment against relevant quality indicators**

Q1 Workshops and symposiums

Participation in joint NoE and WHO meetings (Reykjavik, Uppsala, Stockholm).

Q9 Jointly executed research programme

A cross-European study on agreement on semantic distance between medical concepts are under way.

Q10 Key characteristics of partners

Work package contributing to increased cooperation between public health organisations and research department.

#### **Conclusion**

An extensive report on challenges in European health statistics has been written. A cross-European study on agreement on semantic distance between medical concepts are under way. Cross WP-relations established with WP21 and WP22.

#### **Text mining in biomedicine (WP24)**

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.

The WP24 group has established online services at EBI ([Whatizit](#), [EbiMed](#), [PCorral](#)), which analyse biomedical documents. Whatizit accepts text data via cut&paste, identifies contained terminology and which links it to biomedical databases. EbiMed combines these information extraction capabilities with a retrieval engine based on Lucene and generates summaries from retrieved Medline abstracts. PCorral identifies protein-protein interactions from Medline abstracts, which are again retrieved upon keyword query. 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. Furthermore, the group has realized a software component for the annotation of full text documents (PDF and Html), which is called [Paella](#). It is available for public use upon download. All three services are suitable to mine the biomedical literature and integrate links into the databases at the EBI, which contribute as linking element named entities to database entries. A more extensive description of WP24 work plan is found in deliverable D24.3.

In parallel, the Geneva team has developed a set of services for automatic categorization of any input texts (abstracts are full-article contents) into most popular biomedical terminologies and ontologies, such as SNOMED, Gene Ontology, MeSH, powered with a passage selector, which drive the user to the important passages. The EAGL (Engine for Answering questions



in Genomics Litterature) toolkit (<http://www.natlang.hcuge.ch/Resources/online-tools.html>) provides an information retrieval engine and various services and interfaces developed in WP24: ranked documents can be sent to a set of information extraction and terminology filters, which highlight content-dense passages of the selected documents and propose a GRID-based summary of the selection. The framework also gathers developments involving other WPs: BFMED is a cross-language French-to-English search engine developed in collaboration with WP20 and WP24 (Screenshot 4, 5), and a SNOMED categorizer developed in cooperation with WP22.

#### **Assessment against relevant quality indicators**

Annex I (DoW) of the contract defines those quality indicators by which the consortium seeks to assess its progress. A subset of these indicators is listed below:

##### **Q1 Workshops and seminars**

- Participation in the ISMB 2006 (Fortaleza, Brazil): software demo, bird of feather session, paper presentation in the SIG BioLink meeting
- Participation of TREC Genomics, with the National Library of Medicine
- Participation in BioCreative II: Gene Normalization, Protein-Protein Interaction (Fall 2006)
- Participation in the MIE 2006 (Maastricht, NL)

##### **Q2 Sharing of resources and tools**

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

##### **Q6 Short- and medium term visits**

- One medium-term visit exchange: members of WP24 visiting partners of the NoE
- Organisation of the workshop on text mining in Balatonfuerd in 2006 in collaboration with NoE InfoBioMed.
- Q7 Co-authoring of research papers, reports and educational materials**
- Three co-authored research papers

##### **Q10 Key characteristics (research funding, future prospects etc.)**

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 eight 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.

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

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

A collaboration between WP24 of the NoE "SemanticMining" and members of the NoE "InfoBioMed" has been established, i.e. between EBI (D. Rebholz-Schuhmann) and the Medical Center at the Erasmus University of Rotterdam (J. van der Lei, Erik van Mulligen). 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.

#### **Conclusion**

Current achievements of WP24 are inline with our plans. Several demonstrators are now available online for large-scale tests. Interoperability between the different tools developed within the network has been successfully achieved via a pipeline of HTTP services, such as the interface between EAGLi (DIM) and EBIMed (EBI). Regarding sharing of knowledge

resources, the WP has also been able to seamlessly integrate the multilingual lexicon developed in WP20. Next efforts will include fostering work packages integration: cross-language information retrieval user evaluation, and developments of refined categorizer for clinical contents, in particular SNOMED CT and ICD.



wnt

[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
<a href="#">beta-catenin</a> <small>(score: 6853)</small>	APC or APCs (240/428) GSK-3 beta or glycogen synthase kinase-3 beta (154/198) Axin or axins (145/259) E-cadherin (97/182) cyclin or cyclins (89/142) Wnt-1 or Wnts 1 (73/133) Lef or Lefs (64/94)	nucleus (132/176) cytoplasm (81/89) intracellular (61/72) plasma membrane or cell membrane or cytoplasmic membrane (39/51) membrane (37/49) adherens junction (27/34) extracellular or extracellular regions (16/18) cytoskeleton (13/13) transmembrane (12/12)	Transcription (341/449) development (201/247) phosphorylation (157/238) localization (129/182) transduction (102/117) cell adhesion (67/80) cell-cell adhesion (45/49) apoptosis (41/71) cell proliferation or cells proliferation (35/42) pathogenesis (23/24) embryogenesis (20/22) morphogenesis (18/22)	binding (183/241) DNA binding (19/22) kinase activity (4/4) cadherin-binding (3/4) protein binding (3/3) mitogen-activated kinase (2/2) E2 (2/2) MMP-9 or MMPs-9 (2/2) GPCR (2/2) PKG (1/2) SAPK (1/2)	Lithium (23/32) thyroid (9/30) chondrocytes (9/22) retinoic acid (7/7) anti-inflammatory drugs or indomethacin (5/12) modular or monomeric (4/6) etodolac or Sulindac or Ibuprofen (3/9) caffeine or aspirin (3/6)	cancers (253/423) humans or man or Homo sapiens (210/270) Xenopus (117/149) Armadillo (107/150) mouse or nude mice or transgenic mice or Mus musculus (106/146) axis (85/124) Drosophila (75/79)

Screenshot 3. EBIMed summary display of search for abstracts referring to Uniprot proteins.



Screenshot 4. BFMed (“Bibliothèque de la Faculté de Médecine”) is a MEDLINE search engine which accepts French natural language queries and output a list of MEDLINE records. Once translated via resources developed in WP20 (Multi-lingual lexicon) and using tools developed in WP24, queries can be redirected to EBIMed (Screenshot 3), EAGLi (Screenshot 5) or PubMed.



Screenshot 5. EAGLi (“eagle eye”) queries are normalized for expansion, so that synonyms are also added to the original query. Each document is ranked by a statistical estimate expressed by a progress bar on the left. A *Semantic Summary* can be obtained by clicking the upper right GRID button. A category-driven passage summary is also proposed upon request (not displayed here).



## Terminology systems in laboratory medicine (WP25)

During the NoE summer school, Tihany July 2005, a workshop was conducted on reports for medical diagnosis and treatment and how to ensure connectivity between stakeholder's, one scope being to work out a standardised way of communicating kinds-of-property in laboratory requests and reports. One possible such approach (C-NPU) is based in metrology, another approach (LOINC) is determined by practical consideration of communicating in a HL7 environment. However, as the real world properties to be represented are of the same kind throughout the world consensus regarding representation should be achievable. Work on this matter was initiated during 2006 (researcher Martin Berzell). The current phase, exploring ontologies and representations, is work intense and has taken one year to finish before moving on to representation issues. Collaboration with the WP20 and WP21 has also been pursued.

A major break through in our perception on how to achieve such connectivity came in 2006 when it was realized that it probably is possible to represent results within SNOMED CT incorporating or mapping the C-NPU format. The former proposal of SNOMED modelers that results of examinations should be attached to procedures (activity) representing the activities performed to obtain the result (i.e. in the Procedure hierarchy) was challenged. The view based on the C -NPU and EN 1614 is that results should be attached to kinds-of-property (in SNOMED CT this is in the *Observable Entity* hierarchy). Observable Entity axis is added to be able to include observations where there were no relevant observation procedures (method) stated. This is in many cases true for laboratory observables as well. Technology development has over a short stretch of years produced many ways of examining the same observables. E.g. the count of red blood cells in blood has been measured over the years with microscopy techniques and several different kinds of flow cytometry. Nobody is in doubt that the patient observable, i.e. the kind-of-property, is the same and the technique used to measure is never reported although procedures (activity) are important in documenting the performance of an examination they should not be used as the carrier of results.

It is important to realize that the procedure (activity) performed, according to which method it was performed and the result(s) (intended or actual) are quite different kinds of concepts. The procedure (activity) is an action whereas the result (value) is the product of the action - information about the patient. In short the difference between a 'laboratory procedure' and any other examination procedures is arbitrary and the difference between 'procedure' and 'observable' is not. The 'observable-value pair' is the product of the 'procedure'. Based on these considerations active collaboration is now established with the SNOMED CT concept model working group (Daniel Karlsson member of the group). This work is at the time of writing very active and fruitful.

### Assessment against relevant quality indicators

Annex I (DoW) of the contract defines those quality indicators by which the consortium seeks to assess its progress. A subset of these indicators are listed below:

#### Q1 Workshops and symposiums

Successful realisation of workshops at Summer school, in Estonia and in collaboration with US LOINC-groups.

#### Q8 Participation in standardisation work

Active participation in standards work on laboratory requests and reports (prEN 1614)

#### Q9 Jointly executed research programmes

Submission of full application to FP6 Call 4 (BioMeld – data models and terminology for biobanking) together with four other NoE partners and eight non-NoE partners.

Growing collaboration with partners (IFOMIS, KI) on ontological principles in laboratory medicine.



## The electronic health record (WP26)

This summary report should be read in conjunction with Deliverable 26.3, which was published in December 2006. This quite detailed report documented the technical areas of work being tackled by the partners, largely in collaboration, and pointed to future directions of research that were intended. To avoid repetition, that material has not been repeated here. This section is therefore a brief summary of that deliverable.

The partners of Semantic Mining WP26 are all involved in research that explores various aspects of how clinical meaning is carried within a generic EHR model (such as prEN/ISO 13606, or HL7 CDA Release 2). The original description of this workpackage implied a goal of semantic indexing the EHR. It is now clear that this was perhaps one particular way of achieving the broader goal of semantic interoperability. As fitting with this NoE as a research-based endeavour, the partners have broadened the interpretation of the original WP26 tasks in order to respond to the goals that have emerged, and to re-focus on specific challenges that have become recognised along the journey:

- to enrich the generic specification of EHR archetypes to enable these to be fuller specifications of clinical domain knowledge, and to foster global harmonisation of efforts in this area through collaboration with international standardisation;
- to develop ontological representations of archetype content, in order to permit this archetype content to be more rigorously validated and to permit sets of archetypes to be compared and organised;
- to explore the options for binding archetypes to co-ordinated terminology, in order to identify ways in which consistent representations can be found for the use of such terminologies within structured records;
- to develop tools for authoring and managing archetypes and templates and their binding to ontology and terminology resources;
- to implement or adopt some of this research within operational clinical systems, in order to seed the potential for empirical evaluations in the future.

### Main activities and results

- ♦ The second and third of a series of joint workshops with WP21 on the interaction of terminologies and ontologies with electronic healthcare records was held in February and November, respectively in Paris.
- ♦ A 3 day WP26 workshop in Paris with WP22 partners: review of research threads and agreement of new activities
- ♦ *EHR archetype specifications*. Both the *openEHR* Archetype approach and its adoption into a draft standard have occurred during the Semantic Mining NoE project, and the international acceptance of EHR Archetypes has benefited from Workpackage 26 research outputs.
- ♦ The last 12 months have seen significant improvements to both the methodology and tooling for archetypes and templates. This has been aided by the Semantic Mining project by exposure of other participants to archetypes and interaction with the Manchester group in particular has provided valuable insights into how better to model the connection between archetypes and terminology.
- ♦ The group has focused on the interaction of terminologies and ontologies with medical records in close co-ordination with WP26 and the new workpackage on SNOMED : 1) analysis of the formal relation between ontologies and data structures, including both Archetypes and the HL7 RIM, 2) matching of terminology to data archetypes, 3) Analysis of issues of quality in SNOMED.

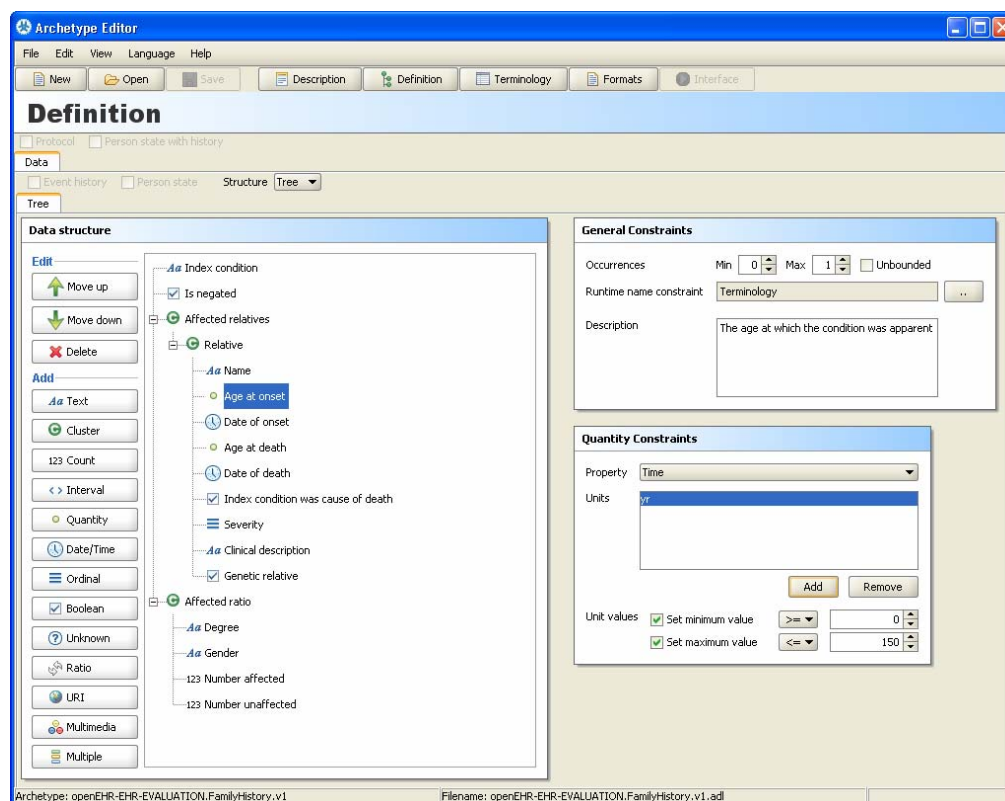
### Progress towards achieving objectives

Since the last WP26 report minor revisions have been made to the *openEHR* and 13606-2 EHR Archetype specifications in the light of feedback from tools development, WP26 research, and from a growing community of Archetype authors.

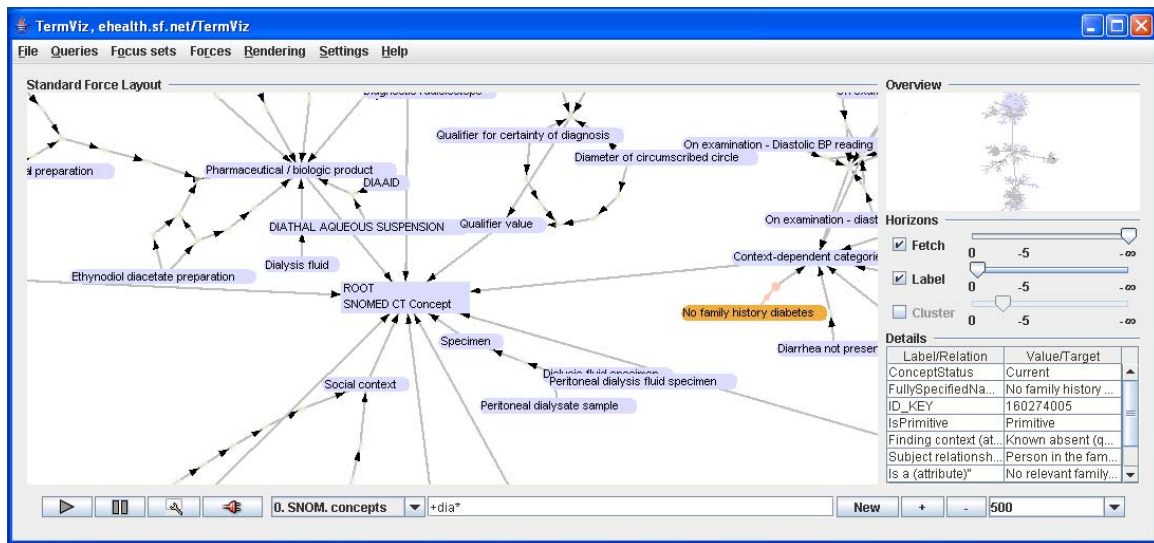
The Archetype Definition Language (ADL) has proved to be a solid formalism for building tools. It has been improved in a number of minor ways, including

- better representation of dates and times and durations
- improved grammar for assertion expressions in archetypes
- support for generic types, i.e. type names of the form Interval<Quantity>

Work on representing Archetype constraints using OWL, in Manchester, has also highlighted some corrections and improvements. A model of *openEHR* templates has now been written, although still being tested. One of the major areas of semantic research in which UCL and Ocean Informatics are engaged is the design and use of Archetypes and ‘Templates’ for modelling clinical content outside of the models used to build software (i.e. UML models etc). In parallel, the Karolinska Institutet (KI) and Linköping University have been maturing a Java Archetype parser and an editor tool; functionally this editor is close to the Ocean one, with improvements in the way archetype internal terminology is manipulated and displayed. The Java based archetype editor is now close to reaching feature completeness so that all kinds of archetypes based on the *openEHR* Reference Model can be edited. A template tool and a SNOMED-CT server and query builder have been built by Ocean Informatics.



Screenshot 6. Archetype editor as part of openEHR software components (WP26).



Screenshot 7. Terminology browser (TermViz) as part of openEHR software components (WP26).

Specific contributions of the work package include:

- ♦ INSERM have also conducted a comparison between the questionnaire editor of HEGP with the archetype editor of the Linköping University and proposed an evaluation framework for “template” editors.
- ♦ The University of Manchester has made use of ongoing advances with archetypes to refine their theory of how to use terminology in information systems.
- ♦ INSERM has conducted research on practical bindings of medical information and medical terminologies within HEGP, and performed an evaluation of the HEGP Terminology Server, comprising reference terminologies as well as a local shared vocabulary, (called the “local dictionary of concepts”).
- ♦ In newer research, INSERM is developing a methodology for establishing the link between information, inference and terminology models in order to share Information between Computerized Clinical Decision Support Systems and Electronic Healthcare Records.
- ♦ The Laboratory of Applied Ontology of CNR-ISTC has focused on extending the scope of the inference ontology in medicine (ROME) in order to represent complex concepts related to the patient folder. Ontologies should not be regarded as an alternative to archetypes, but as a useful complementary approach. This work investigates an evolution from a 'classic' openEHR architecture to an ontology-based patient record, whose information elements are mapped into a reference ontology of medicine.
- ♦ Following some previous work at KI on a supplementary form based EHR system called Julius the KI group initiated a development of a JAVA based implementation of the *openEHR* specifications in mid 2004. Later this implementation was influenced by early adopters from a number of other institutions, not the least the LiU group in 2006. The Java implementation experience is now included as part of the official release of the *openEHR* specifications.
- ♦ As part of WP26, UCL has undertaken research to investigate the feasibility of different operationalised formalisms to embody the dual-model (Reference Model plus Archetype) principle but with both models described in the UCL EHR group’s preferred development language, Java. In this approach, Archetypes extend the generic model classes (similarly to more conventional one-model systems),



attempting to combine the flexibility of dual model with the potential performance of one-model systems.

- ♦ In a project regarding EHR overviews and navigation the Linköping team has been using *openEHR* based components for storage, retrieval and processing of EHR data. Google Earth (GE) has been used for handling display and interaction of clinical information stored using *openEHR* data structures and ‘archetypes’.
- ♦ In a previous master thesis project at Linköping University a web based generic GUI for showing *openEHR* based data was developed using server based Java. A more feature generic GUI based on Java swing to be used for archetype based structured data entry is now being developed.
- ♦ UCL has been involved in the development of EHR Archetype instances in cardiology, cancer and to conform to specific NHS data sets, in order to validate the present archetype formalism and tools and to build up experience of best practice in archetype authorship.

#### **Assessment against relevant quality indicators**

##### Q1 Workshops and symposiums

Work package 26 has held two major workshops during 2006, and one more is scheduled for 2007.

##### Q2 Sharing of resources and use of research software tools

Open source software tools for archetype authorship, terminology binding and OWL ontology binding. These kinds of tools are developed by several partners, with a future view of sharing them and of interfacing them to each other.

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

Visit of member of LiU to UCL: one week, in November 2005

Visit of the full Manchester team to UCL: one day workshop, December 2005

Visit by two PhD students from INSERM to LiU, one week in November 2006

Visit by PhD student from UoM to LiU, two weeks, May 2006

Planned visit by three LiU members to UCL and UoM, spring 2007

##### Q7 Co-authoring of research papers

Two co-authored research papers.

##### Q8 Participation in standardisation activities

Major presentations of work, and discussions of the semantic links of archetypes to templates, terminologies and ontologies have taken place at international standardisation events.

##### Q9 Jointly executed research programmes

Q-REC: EU FP6 (UCL, started 1/1/06)

SemanticHealth: EU FP6 (UCL, UoM, started 1/1/06)

#### **Conclusion**

Many of the issues being tackled in WP26 are complex conceptual problems, requiring deep understanding of the semantic challenges involved in systematically representing diverse and evolving clinical concepts and data structures. The partners and their research teams have progressed considerably in both individual site activities and inter-site collaborations. It is difficult to capture in the form of a report the growing richness of this mutual understanding, or to project clearly how these innovative threads of work will contribute to next-generation solutions to the semantic indexing of EHRs. It is clear, however, that the work being done by the WP26 partners is being recognised internationally as a set of strong contributions within standardisation, in the development of an EU Semantic Interoperability Roadmap, and in



---

next-generation research proposals into the EU Seventh Framework Programme. A stream of high-impact publications is starting to flow and will continue beyond the funded period of this project.

### **Laymen terminology (WP27)**

In the Assembly meeting in December 2005, a decision was made to start up a new work package devoted to terminology barriers between health care professionals and laymen. The new WP was created as a result of cross-WP activities between WP20 and WP26. A kick-off meeting was held in February 2006.

- ♦ *Literature review on patient-friendly documentation systems.* The OU team coordinated the deliverable on literature review of patient-friendly documentation systems (D27.1) and contributed a substantial amount of the written content. The outcome was also published as an Open University Computing Science Department Technical Report.
- ♦ *Corpus collection and analysis*  
The OU team has completed work on corpus collection and analysis, investigating parallel and single corpora of patient-to-patient, doctor-to-patient, patient-to-doctor and doctor-to-doctor documents. In addition, statistical investigation of the British National Corpus for genre detection by David Hardcastle revealed that the material was too heterogeneous for use as a training set. The results of this corpus study were published in report form as an internal WP27 deliverable (September 2006).  
At UGOT, collection is ongoing of three Swedish (sub)corpora from the same medical subdomain, namely “cardiovascular disorders”. These corpora have formed the basis of a contrastive study of two (postulated) medical language varieties, conducted by Dimitrios Kokkinakis and Maria Toporowska Gronostaj. The first results of this ongoing investigation have been presented at an international conference (see publications, below).  
The Swedish subcorpora are the following: the first subcorpus, the “non-expert corpus”, derives from a number of Swedish daily newspapers and other online health information sources targeted to consumers; the second subcorpus, the “expert corpus”, derives from two Swedish medical resources intended for professionals and specialists across a broad spectrum of medical professions: the third subcorpus are made of texts from various “ask-the-net-doctor” sites.
- ♦ The UGOT and SU groups are conducting negotiations with the Göteborg University Hospital about the use of free-text portions of de-identified and anonymized patient records for research. The INSERM team have conducted work on comparing medical-content web pages in different languages with a view to identify linguistic and other indicators that could be used to distinguish between specialist and non-specialist documents.

### **Progress towards achieving objectives**

- ♦ On the basis of the corpus work conducted at OU, UGOT and INSERM, INSERM and OU together have identified a preliminary set of requirements to be placed on patient-friendly documents (“Some recommendations for the creation of patient-friendly documents”; November 2006).
- ♦ The results of the two corpus studies and the requirements analysis are now being merged into the second deliverable of WP27, an external report entitled “Empowering the patient with language technology”.



- ♦ Donia Scott and Clara Mancini (OU) have been working on a theoretical framework on which to base the flexible presentation of medical information to patients using hypertext as a medium. We have therefore started, with both theoretical and empirical work, to extend the descriptive framework of Document Structure so that it can include non-linear as well as linear documents.
- ♦ Paul Piwek, Richard Power, Sandra Williams and Catalina Hallett (OU) have been investigating possibilities for communicating EHR information to patients via a dialogue between animated agents. A short report on this work has been published as an Open University Computing Science Department Technical Report.
- ♦ The re-design and re-engineering of an EHR server to support anticoagulant care, applications that may be used directly by patients, and ongoing research on EHR data analysis.
- ♦ For WP27, the main focus of work at UCL has been, from a research point of view, to design and implement a concept look-up index for each archetype node, to permit each node to index a set of relevant educational materials, and the specification of the corresponding look-up service. The look-up service has been partially specified and those parts that can be implemented by an archetype service are being implemented.
- ♦ A portal architecture is being developed to provide a uniform and authorisation-managed access to the anticoagulant and other web applications for cardiovascular care. This will be used in WP27 to provide patient access to the anticoagulant management system and links to educational resources. However, the actual re-design of the application and the services that will request patient educational resources or display them to the patient are not yet being addressed.
- ♦ A more complex area that is also not yet being tackled is the linkage of data item (form) values to educational resources. This will prove important in the longer term, but it is not yet clear if the detailed design or implementation can be scoped within the resources or time frame of WP27. The next step will be to link this archetype-concept look up service to the ontology used to index educational resources by other WP27 partners.

#### **Plans for the remaining phase**

In parallel to the research and engineering, work has commenced at UCL on setting up a demonstrator to validate the provision of patient-friendly information for the management of anticoagulation. This work has included a number of steps:

- (a) initial exploration of the feasibility of using anticoagulation as a field demonstrator, with selected patients to trial the system;
- (b) linking with a PhD student who might be able to design and evaluate this anticoagulant demonstrator, including discussions with the supervisor;
- (c) analysis of ethical and risk framework in which EHRs may be deployed and used in this setting.

It now appears likely that a live pilot will be feasible, but the practicalities of this may mean that the systems and the patient selection and training are only just about complete by the end of Semantic Mining. The final evaluation might therefore be publishable some months after the end of the project, but using the results of Semantic Mining work. Funds are already in place (including support from a hospital trust) to ensure that the pilot is able to continue



---

beyond the end of WP27. This work is possibly also a candidate for inclusion as part of an EU FP7 proposal.

The second WP27 deliverable is due in February 2007. As mentioned above, this deliverable will be in the form of a report based on publications and internal reports prepared during 2006. We are also in the process of writing up a joint contrastive corpus study for submission to a journal.

In February 2007, two meetings will be organized by the Open University, one for planning the addition of French and Swedish to OU's prototype generation system, while the purpose of the other meeting is to come up with ideas for how to extend our present fruitful collaboration beyond Semantic Mining, e.g. in the form of new joint project proposals. A regular WP27 meeting is slated for March 2007 in Paris.



## Deliverables

<i>Deliverable Code &amp; Name</i>	<i>Planned delivery date</i>	<i>Actual delivery date</i>	<i>Comments</i>
D1.1 Detailed analysis of research	m6	2004.06.30	
D1.4 Quarterly reports, Periodic Activity and Management Reports	m3 m6 m9 m12 m15 m18 m21 m24 m30 m36	2004.05.06 2004.06.30 2004.10.29 2005.02.14 2005.04.29 2005.07.29 2005.11.01 2006.02.27 2006.07.30 2007.02.28	Rev. 2005.06.14  Rev. 2006.08.25
D1.5 Project presentation	m3	2004.02.02	<a href="http://www.imt.liu.se/mi">www.imt.liu.se/mi</a> <a href="http://www.semanticmining.org">www.semanticmining.org</a>
D3.1 Typology of shared resources	m6	2004.06.30	
D4.1 Public website	m11	2004.04.28	<a href="http://www.semanticmining.org">www.semanticmining.org</a>
D5.1 Network meetings	m6	2004.06.30	
D6.1 Report on doctoral dissertations	m6	2004.07.08	Poster session during summer school
D8.1 Participation in standardisation	m6	2004.06.30	
D7.1 Planning of summer school	m8	2004.09.30	
D10.1 Workshop on health statistics	m9	2004.10.29	Part of summer school
D11.1 Workshop on semantic web	m9	2004.10.29	Part of summer school
D12.1 Workshop on ontology engineering	m9	2004.10.29	Part of summer school
D3.2 Common database	m11	2004.12.31	
D5.2 Network meetings	m11	2004.12.31	
D6.2 PhD programmes	m11	2004.12.31	
D6.3 Mobility program	m11	2004.12.31	
D7.2 Evaluation of summer school	m11	2004.12.31	
D9.1 Dissemination of material	m11	2004.12.31	
D13 Workshop on NLP	m11	2004.12.31	Satellite to COLING
D16 Workshop on EHR	m11	2004.12.31	Satellite to EUROREC
D20.1 Multi-lingual medical dictionary	m11	2004.12.31	
D21.1 Ontology engineering	m11	2004.12.31	
D22.1 SNOMED CT	m11	2005.02.14	
D23.1 Health statistics	m11	2005.01.28	Extensive report, 36 p.
D24.1 Data mining and information retrieval	m11	2004.12.31	
D25.1 Concept system in lab.medicine	m11	2004.12.31	
D26.1 The Electronic health record	m11	2004.12.31	
D2.1 Assessment and strategic planning	m12	2005.01.31	
D14/15 Workshop on data mining and information retrieval	m14	2005.09.12	International Symposium for Semantic Mining in Biomedicine (WP14/15), April 10-13, 2005
D4.2 Report on public website	m17	2005.06.15	
D8.2 Participation in standardisation work	m30	2006.06.23	Changed delivery date in revised DoW
D9.2 Report on educational material	m17	2005.04.15	
D9.3 Report on dissemination	m17	2005.06.07	
D20.2 Prototype multi-lingual medical dictionary	m17	2005.06.30	
D22.2 Report on SNOMED CT – evaluation	m34	2007.02.26	Changed delivery date in



			revised DoW
D22.3 Experience from translation and mapping of SNOMED CT	m34	2007.08.15	Changed delivery date
D23.2 Report on information quality in health registries	m17	2005.06.14	
D24.2 Text mining and IR in biomedicine	m17	2005.09.12	
D25.2 The CNPU-coding system – limitations and possibilities	m17	2005.06.18	
D26.2 Architecture for semantic-based EHR	m20	2005.10.15	Changed delivery date in revised DoW
D30 Workshop on ontology and biomedical informatics	m20	2006.02.24	Rome meeting in conjunction with IMIA WG6
D31 Workshop on human factors and large ontologies	m20	2005.08.30	AIME Aberdeen
D32 Workshop on the boundary problem	m22	2005.12.15	Summer school 2005
D33 Workshop on SNOMED CT	m20	2005.12.15	Summer school 2005
D34 Workshop on concept system in lab.medicine	m22	2005.12.15	Summer school 2005
D35 Workshop on text mining from EHR	m22	2005.09.30	Summer school 2005
D36 Workshop on semantic web	m22	2005.12.15	Summer school 2005
D5.3 Network meetings	m23	2005.12.31	
D6.4 Mobility program	m23	2006.01.10	
D7.3 Evaluation of summer school	m23	2006.01.10	
D2.2 Assessment and strategic planning	m24	2005.12.31	
D20.3 Platform for exchange of biomedical text corpora	m24	2006.01.27	
D21.2 Engineering ontologies: foundations and theories from philosophy and logic	m24	2006.02.21	
D21.3 Engineering ontologies: foundations and theories from computer science	m33	2006.09.23	Changed delivery date in revised DoW
D21.4 Engineering ontologies: practical, pragmatic and human factor issues	m35	2007.03.23	
D21.5 Engineering ontologies: quality assurance and evaluation	m38	2007.03.23	
D9.4 Dissemination of results	m33	2006.12.20	Included in compiled report on Dissemination activities 2006
D20.4 Experience from development of multi-lingual medical dictionary	m36	2007.02.27	
D23.3 Towards measurement of semantic distance	m36	2007.08.15	
D24.3 Biomedical semantic classes for text mining	m34	2006.11.29	
D25.3 Towards an ontology for laboratory medicine	m34	2006.10.19	
D26.3 The Boundary problem – linking information and terminology models	m34	2006.12.20	
D27.1 Literature review of patient-friendly documentation systems	m30	2006.05.22	
D40 Training program in biomedical ontology	m32	2006.12.20	Included in compiled report on Dissemination activities 2006
D41 Workshops in medical terminology	m35	2006.12.20	Included in compiled report on Dissemination activities 2006
D7.4 Joint NoE summer conference 2006	m34	2006.12.20	Included in compiled report on Dissemination activities 2006
D2.3 Assessment and strategic planning	m36	2007.02.15	



D4.3 Use of SemanticMining web site	m36	2006.11.28	
D27.2 Empowering the patient with language technology	m38	2007.03.15	
D6.5 Mobility program in biomedical informatics – results and implications	m44	2007.09.15	
D8.3 Participation in standardisation work	m42	2007.09.15	
D9.5 Educational material from SemanticMining	m42	2007.09.15	
D20-27.1 Research results from SemanticMining	m42	2007.09.15	Named: Final SemanticMining Report on Contributions to the European Research Area
D20-27.2 Data and service resources from SemanticMining (services and tools)	m42	2007.09.15	Named: Final SemanticMining Report on Tools and Services



## Performance indicators

The table below with performance indicators is based on the available guidelines for an assessment methodology for NoEs, but is adapted to the specific Quality Indicators which have been defined for SemanticMining (see section 7.2 of Annex I). The progress indicators are coded as: 5=completely achieved, 4=mainly achieved, 3=partially achieved, 2=scarcely achieved, 1=not achieved. The evaluation has been performed by the Board.

<i>Quality Indicators</i>	<i>Comments</i>	<i>Score</i>
<i>Network management</i> legal structure, CA decision making structure Management office Board Assembly	The management structures are defined and in full operation, and has proven robust and efficient	5
<i>Communication inside the network</i> Q3 Use of website	The communication platform (MERMIG) and public website is implemented and in operation. The content of the website will be transferred to a Wiki-based web platform for cooperative long-term sustainability.	4
<i>Knowledge sharing</i> Q1 Workshops and seminars	Successful series of workshops	5
Q4 PhD-study courses	PhD-student activities at workshops and at Summer school, first two rounds of student mobility program completed.	4
Q5 Co-tutoring of PhD-students	Successful doctoral consortium during summer school 2005 and 2006. Three PhD students co-supervised among partners.	3
Q7 Co-authoring	Significantly increase in co-authored research papers already published and more in progress	5
Q8 Participation in standardisation	Active participation in CEN TC251 and HL7, and partially in W3C.	5
<i>Sharing of resources and tools</i> Q2 Sharing of resources/tools	Sharing of resources and tools in all research WPs (20-27)	5
<i>Mobility</i> Q6 Short- and medium visits	Several short-time visits by researchers performed. Three rounds of student mobility program completed.	4



<p><i>Continuity of the network</i></p> <p>Q9 Joint research programme</p>	<p>A Scientific Advisory Committee is established which has been used for assessment and strategic planning during 2005. Joint NoE activities have being organised by SemanticMining, INFOBIOMED and BIOPATTERN with focus on mobility and coordination of events. Evidence of increased level of integration of research teams, increased number of joint publications and research applications. Ongoing work on FP7 application.</p>	<p>4</p>
<p>Q10 Key characteristics of partners</p>	<p>Examples of spin-off from several WPs: joint research applications, exchange of students and sharing of resources. Emerging industrial collaboration or contacts by several partners.</p>	<p>4</p>



---

## **Consortium management**

The decision making structures are according to the Consortium Agreement in operation. The Management Office at Linköping University handles administrative matters and communication with project officers in Brussels. The Board consisting of six members have regular meetings where the progress of the NoE is followed-up. Seven Assembly Meetings have been held (January 2004, July 2004, December 2004, July 2005, December 2005, July 2006, and December 2006).

## **Changes of the consortium**

A request of extension of the time period up to 42 months has been put forward to the Commission. As the request has been approved, the project runs until July 2007.

Our partner The Victoria University of Manchester merged with University of Manchester Institute of Science and Technology (UMIST) on 1st October 2004 to create one new university called The University of Manchester.

No other contractual changes are to be reported.



## Project Meetings

Title	Date and Place	Comments
Board meetings	Tele-meetings  April 27-28 Feb 12-13, 2007, Paris	3 in Q1 2006 2 in Q2 2006 2 in Q3 2006 3 in Q3 2006 2 in Q4 2006 4 in Q1-Q2 2007  Follow-up of ATR, strategic planning Planning of ATR
Assembly meetings	Balaton, Hungary, July 8 Paris, December 8 Barcelona, June 27	Sixth Assembly meeting Seventh Assembly meeting NoE meeting
NoE cluster meeting	Barcelona, January 13, 2006	NoE cluster meeting with SemanticMining, INFOBIOMED and BioPattern – planning of joint NoE summer conference
WP20 meeting	Geneva, January 19-20, 2006	Five partners attending, elaboration of final draft for link interchange format and corpus interchange format
LREC Workshop	May 23, 2006, Genoa, Italy	LREC International Workshop – Acquiring and representing multilingual, specialized lexicons: the case of biomedicine
WP20 meeting	Sep, 14-15, 2006, Jena University	Discussion of the state of the art of lexicon, exploitation issues
WP21 meeting	July 7-8, Balaton	Joint meeting with Summer school
WP21 meeting	Feb 26-27, Dec 9-11, Paris	Joint meetings with WP26
Foundations of Terminology and Classification	8 April, 2006, Timosoara, Romania	Organised as a separate work package involving all WP 21 participants
Joint workshop on integration of EHRs and Terminology	26-27 February 2006, Paris, France	Joint with WP26
Joint workshop on integration of EHRs and Terminology	November, 2006	Joint with WP26. Date to be confirmed.
SemanticHealth initiation meeting	April, 2006	Joint with Semantic Health. Semantic Mining results presented
EC ICT Bio Conference	Brussels, June 28-30	Workshop on Symbiotic white paper on priority areas in biomedical informatics. EC ICT Bioinformatics Conference
Summer School	July 3-8, Balatonfüred,	First European Summer School on



	2006	Biomedical Informatics together with INFOBIOMED and BioPattern
WP22 meeting	May 29, Copenhagen	Presentation and discussion of the Danish experiences of work with SNOMED CT as part of national ICT strategy for ehealth
WP23 meeting	July 7, Balatonfüred, Hungary	5 participants
WP23 meeting	Sept 13 Stockholm, Sweden	7 participants
WP23 meeting	Dec 7 Paris, France	6 participants
WP23 meeting	Feb 15, 2007 Stockholm, Sweden	6 participants
WP23 meeting	March 23, Budapest	
WP25 meeting	SemanticMining Summer School, July 3-8	2 participants
WP25 meeting	Workshop on representation of results within SNOMED CT (C-NPU), Oct 4 (in conjunction with the SemanticMining conference on SNOMED CT	6 participants
WP25 meeting	Meeting of the committee on Nomenclature Properties and Units, Brussels, Dec 3-5	2 WP25 participants
WP26 workshop	February 26-27, 2006, Paris	Detailed review of work on archetypes, ontology and terminology binding, and tools development, including progress on project-supported PhDs.
WP 26 workshop	December 9-11, 2006, Paris	Practical approaches to terminology and record structure binding issues: reviewing results of WP26 investigations including progress on collaborative tools development and the use of SNOMED-CT.
WP26 meeting		A further workshop, the final one for WP26, is being planned for May 2007.
WP27 meeting	Open University, UK, February 22-23	Six partners attending, kick-off meeting
WP27 meeting	July 8, 2006, at joint summer school	Six partners attending
WP27 meeting	November 23-24, 2006, Gothenburg, Sweden	Five partners attending
WP26 meeting FP7 planning Board meeting	Paris 11-13 February, 2007	Six partners attending
ATR	Brussels 19 March, 2007	Annual Technical Review
NoE FP7	Brussels 26-28 March 2007	NoE future, FP7 application
Open seminars on openEHR, arketypes and SNOMED CT	Linköping March 14, April 17, May 15, May 30	



---

Int. Symposium on Biomed.Informatics	Barcelona, June 24-27	Joint symposium including doctoral school with INFOBIOMED



## Sharing 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). Finished Nov 2006
Integration of heterogeneous lexicons using link format	Finished Nov 2006, 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 <a href="http://dior.imt.liu.se/cnpu/">http://dior.imt.liu.se/cnpu/</a>
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 work is ongoing to interface tools with each other, as clients to other partner's middleware services and as plug-ins to other partner's clients. <ul style="list-style-type: none"> <li>- Archetype editor</li> <li>- TermViz</li> <li>- MoST</li> <li>- openEHR reference model</li> </ul>
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
Archetypes	Sharing of archetypes (definition of information structure) in different medical domain
Concept corpus and work bench for simulation of agreement among raters	Sharing of tools and resources in WP23
Sharing of resources for information retrieval reported by WP24	Sharing of databases, tools for text processing and information retrieval: <ul style="list-style-type: none"> <li>- FSA Library</li> <li>- WhatIzIt</li> <li>- EbiMed</li> <li>- Paella</li> <li>- PCorral</li> <li>- BFMed</li> <li>- EAGLi</li> </ul>
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.
Biomedical Ontology	BioTop from Universities of Freiburg, Jena, and IFOMIS; Simple-Bio-Top from University of Manchester ROMA from CNR – discussions on harmonisation in progress



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

## Mobility

<i>Type / scope</i>	<i>Details / Comments</i>
Mobility program (WP6)	The visit grants scheme was launched in 2005 as part of the mobility program (WP6). So far, 24 visits have been awarded. New rounds of the mobility program, also including the other NoEs in the eHealth area, is open. Apart from the WP6 mobility program, 20 grants for non-NoE students have been awarded for participation in the Summer School.
Exchange visits reported by WP20	Several short-term visits among partners, two one-week exchange visits by PhD students
Exchange visits reported by WP21	Several short-term visits among partners, four one-week exchange visits
Exchange visits reported by WP24	Six short-term visits among partners
Exchange visits reported by WP25	Short-term visits LiU-IFOMIS, LiU-US LOINC groups, LiU-NBH
Exchange visits reported by WP26	Six short-term visits among partners
Exchange visits reported by WP27	Two short-term visits among partners
Co-supervision of PhD students	Three cases of co-supervision of PhD students
Summer school grants	Twenty grants to non-NoE PhD students to attend the Summer School



## Annex – plan for using and disseminating the knowledge

### A.1 Conferences, workshops, demonstration etc. attended/organised/foreseen by the project

<i>Date</i>	<i>Type and Title/Scope</i>	<i>Number of persons attended + other information</i>
February 7	ICMCC Conference on EHR Standards and Inter-operability The Hague, Netherlands	Open workshop by the International Council on Medical & Care Compunetics Organiser: ICMCC Presenter: UOM
April 8	Foundations of Clinical Terminologies and Classifications <i>Part of EFMI STC 2006, Timișoara, Romania</i>	Open satellite workshop within EFMI special topic conference on integrating biomedical information Proceedings will appear in <i>BMC Medical Informatics and Decision Making</i> Organiser: UKLFR, UOM, IFOMIS 50 attendees
May 10-12	eHealth 2006, Malaga Spain	Participation of Board members
May 21–24	Ontological Spring II - Training Course in Biomedical Ontology Schloss Dagstuhl, Wadern, Germany	3 day open training course providing basic introduction to biomedical ontology Organiser: IFOMIS 40 attendees
June 29-30	BMI conference, Brussels	Participation of Board members
July 2-8	Balaton, Hungary	Joint NoE summer conference (SemanticMining, InfoBioMed, BioPattern), 100 attendees
August 27-29	MIE'06, Maastricht, The Netherlands	Several presentations (scientific papers, posters) by network members
September 28-29	Swedish Terminology Conference, Kalmar	Presentation and demonstration of systems and tools from WP21, WP22 and WP26 120 attendees
November 11-14	AMIA 2006, Washington	Several presentations (scientific papers, posters) by network members
23 March, London	NIH Ontologies for Immunology Workshop	50 attendees, sponsored and paid for by NIH
30-31 May, Windermere, UK	KR 2006	200 attendees, Keynote speech
30 Sept-1 Oct, Copenhagen	Semantic Mining Conference on SNOMED	200 attendees, Keynote presentation and several scientific papers
2-5 Oct, Prague	European Workshop on Knowledge Acquisition (EKAW 2006)	200 attendees, Keynote presentation



8 Nov 2006, Washington	KR-MED 2006.	100 attendees. 3 papers and major role in organisation
August 4-9, Fortaleza, Brazil	ISMB 2006	Presentation of WP24 activities: 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
January 21-24, 2007	ECCB 2006, Eilat, Israel	Presentation of WP24 activities: Paper presentation on EBIMed Software demo on Whatizit Web services
August 29	MIE 2006	Presentation at semantic interoperability workshop: 40 persons present
September 11-14	GMDS 2006	German Medical Informatics Conference, Leipzig, Germany
October 11	WoHIT 2006	Presentation at EHR standards workshop: ~100 persons present
November 8	KR-MED 2006	AMIA WG workshop on "Ontology in Action", Baltimore, U.S.
November 9 - 11	FOIS 2006	Conference on Formal Ontologies and Information Systems, Baltimore, U.S.
		Major presentations of work, and discussions of the semantic links of archetypes to templates, terminologies and ontologies have taken place at the following international standardisation events
January 8-10	CEN TC/251 meeting	50 attendees, Working Groups I and II, The Hague, Netherlands.
April 3-6	ISO TC/215	40 attendees, Working Group 1, Korea
May 13-14	ISO TC/215	50 attendees, Working Group 4 Task Group, Newark, USA
June 1-2	CEN TC/251	30 attendees, Working Groups I and II, Lund, Sweden
September 8-12	HL7	20 attendees, Templates SIG and Detailed Clinical Models Group, Boca Raton, USA
October 8-10	ISO TC/215	60 attendees, Working Group 1) and CEN TC/251 (Working Group I, Geneva, Switzerland)
March 14, April 17, May 15, May 30	Linköping	Open seminars on openEHR, arketypes and SNOMED CT
June 24-27	Barcelona	Joint Int. Symposium on Biomed.Informatics including doctoral school together with INFOBIOMED



## A.2 Articles published, development web sites etc.

### References 2005 (co-authored papers only – authors from more than one partner)

K. Markó, S. Schulz, U. Hahn Automatic Lexicon Acquisition for a Medical Cross-Language Information Retrieval System. Proceedings of the XIX International Congress of the European Federation for Medical Informatics (MIE '05), Geneva, Switzerland. 2005: 829-834.

K. Markó, S. Schulz, O. Medelyan, U. Hahn. Bootstrapping Dictionaries for Cross-Language Information Retrieval. Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '05 ), Salvador, Brazil. 2005: 528-535.

K. Markó, S. Schulz, U. Hahn. MorphoSaurus - Design and Evaluation of an Interlingua-based, Cross-language Document Retrieval Engine for the Medical Domain. Methods of Information in Medicine. 4/2005(44): 537-545

K. Markó, S. Schulz, U. Hahn Unsupervised Multilingual Word Sense Disambiguation via an Interlingua. Proceedings of the 20th National Conference on Artificial Intelligence (AAAI '05), Pittsburgh, Pennsylvania. 2005: 1075-1080

M. Poprat, U. Hahn. Enough is Enough – Estimating Upper Bounds of the Size of Training Corpora for Unsupervised PP Attachment Disambiguation. Proceedings of Fifth International Conference on Recent Advances in Natural Language Processing (RANLP-2005)

K. Markó, S. Schulz and U. Hahn. Multilingual Lexical Acquisition by Bootstrapping Cognate Seed Lexicons. Proceedings of Fifth International Conference on Recent Advances in Natural Language Processing (RANLP-2005)

U. Hahn, P. Daumke, S. Schulz, K. Markó. Cross-Language Mining for Acronyms and their Completions from the Web. Proceedings of the 8th International Conference on Discovery Science (DS '05), Singapore. 2005.

U. Hahn, K. Markó & S. Schulz. Subword Clusters as Light-Weight Interlingua for Multilingual Document Retrieval. In: MT Summit X – Proceedings of the 10th Machine Translation. Phuket, Thailand, September 12-16, 2005. Asia-Pacific Association for Machine Translation (AAMT), 2005.

S. Schulz, K. Markó, R. L. de Andrade, E. Pacheco, P. Nohama, U. Hahn, M. Romacker. The Morphosaurus Medical Subword Lexicon. Lexicographic and Semantic Aspects. Proceedings of the 3th Workshop em Tecnologia da Informação e da Linguagem Humana (TIL '05), São Leopoldo, Brasil. 2005

Mikael Nyström, Magnus Merkel, Lars Ahrenberg, Michael Petterstedt, Håkan Petersson & Hans Åhlfeldt. Generering av ett medicinskt engelskt-svenskt lexikon med hjälp av interaktiv ordlänkning. Svenska Läkaresällskapets riksstämman 2005 (Annual Meeting of Swedish Society of Medicine).

Mikael Nyström, Magnus Merkel, Lars Ahrenberg, Pierre Zweigenbaum, Håkan Petersson & Hans Åhlfeldt. Generation of a English-Swedish medical dictionary using interactive word alignment. Submitted to BMC Medical Decision Making.

K. Marko, P. Daumke, S. Schulz, U. Hahn. Automatische Generierung einer sprachübergreifenden Akronymdatenbank. 50. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie (gmds), Freiburg 11. - 15. September



---

2005 (Annual Meeting of the German Society of Medical Informatics, Biometry and Epidemiology)

M. Poprat, K. Markó, U. Hahn. Automatische Klassifikation medizinischer Dokumente nach Sprache und Zielgruppe für Text-Retrieval-Systeme. 50. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie (gmds), Freiburg 11. - 15. September 2005 (Annual Meeting of the German Society of Medical Informatics, Biometry and Epidemiology)

R.H. Baud, M. Nyström, L. Borin, R. Evans, S. Schulz, P. Zweigenbaum. Interchanging Lexical Information for a Multilingual Dictionary. AMIA 2005 annual symposium. Washington DC. Washington, DC: AMIA. 31-35.

Stefan Schulz & Udo Hahn. Part-whole representation and reasoning in formal biomedical ontologies. *Artificial Intelligence in Medicine* (2005) 34, 179—200

Smith Barry, Ceusters Werner, Klagges Bert, Köhler Jacob, Kumar Anand, Lomax Jane, Mungall Chris, Neuhaus Fabian, Rector Alan, Rosse Cornelius. Relations in Biomedical Ontologies. *Genome Biology*, 2005, 6(5)/R 46.

Barry Smith, Jose L.V. Mejino Jr., Stefan Schulz, Anand Kumar, and Cornelius Rosse. *Anatomical Information Science*. A. G. Cohn and D. M. Mark (eds.), *Spatial Information Theory. Proc COSIT 2005 (Lecture Notes in Computer Science)* Berlin/Heidelberg/New York: Springer;:149-164.

Stefan Schulz, Philipp Daumke, Barry Smith, Udo Hahn. How to Distinguish Parthood from Location in Bioontologies. In *Proc AMIA Symposium 2005, Washington DC*;:669-673

Kumar Anand, Smith Barry, Pisanelli M, Gangemi Aldo, Stefanelli Mario. Clinical Guidelines as Plans: An Ontological Theory. *Methods of Information in Medicine*. In press  
Anand Kumar, Barry Smith, Domenico Pisanelli, Aldo Gangemi, Mario Stefanelli. An ontological framework for the implementation of clinical guidelines in health care organizations. *Stud Health Technol Inform*. 2004;102:95-107

Stefan Schulz, Suzanne Hanser, Udo Hahn, Jeremy Rogers. Semantic Clarification of the Representation of Procedures and Diseases in SNOMED CT. *Stud Health Technol Inform*. 2005;116:773-8. (Extended version also submitted to *Meth Inf Med*)

Stefan Schulz, Anand Kumar, Thomas Bittner. Biomedical Ontologies: What part-of is and isn't. *Journal of Biomedical Informatics (Special Issue)*. In press and online at [www.sciencedirect.com](http://www.sciencedirect.com)

Alan Rector, Jeremy Rogers, Thomas Bittner. Granularity scale and collectivity: When size does and does not matter. *Journal of Biomedical Informatics (Special Issue)*. In press and online at [www.sciencedirect.com](http://www.sciencedirect.com)

Christiane Fellbaum, Udo Hahn, Barry Smith. Towards new information resources for public health—From WordNet to MedicalWordNet. *Journal of Biomedical Informatics (Special Issue)*. In press and online at [www.sciencedirect.com](http://www.sciencedirect.com)

Special Issue Editors: F. Pinciroli and D.M. Pisanelli. *Ontologies in Medicine. Computers in Biology and Medicine*. In press and online at [www.sciencedirect.com](http://www.sciencedirect.com)

New Journal Launch. Editors in chief: Nicola Guarino and Mark Musen. *Applied Ontology*

Forsum U, Hallander HO, Kallner A, Karlsson D, Impact of qualitative analysis in laboratory medicine. *TrAC-Trends in Analytical Chemistry*, 2005, 24: 546 -555.

Forsum U, Karlsson D, Terminology, categories and representation of examinations in laboratory medicine. *Clin. Chem. Lab. Med*. 2005;43:344-345.



Rosse Cornelius, Kumar Anand, Mejino Jr. Jose L.V., Cook Daniel L., Detwiler Landon T., Smith Barry: "A Strategy for Improving and Integrating Biomedical Ontologies" (<http://ontology.buffalo.edu/bio/OBR.pdf>), in: Proceedings of AMIA Symposium 2005, Washington D.C., 2005, 639-643

Schulz Stefan, Daumke Philipp, Smith Barry, Hahn Udo: "How to Distinguish Parthood from Location in Bioontologies" (<http://ontology.buffalo.edu/bio/Part&Location.pdf>), in: Proceedings of AMIA Symposium 2005, Washington D.C., 2005, 669-673

Smith Barry, Ceusters Werner, Klagges Bert R. E., Köhler Jacob, Kumar Anand, Lomax Jane, Mungall Chris, Neuhaus Fabian, Rector Alan, Rosse Cornelius: "Relations in Biomedical Ontologies" (<http://genomebiology.com/2005/6/5/R46>), in: Genome Biology, 2005, R46

Smith Barry, Mejino Jr. Jose L.V., Schulz Stefan, Kumar Anand, Rosse Cornelius. "Anatomical Information Science" ([http://ontology.buffalo.edu/anatomy\\_GIS/FMA-AIS.pdf](http://ontology.buffalo.edu/anatomy_GIS/FMA-AIS.pdf)), in: Cohn A.G., Mark David M. (eds.): Spatial Information Theory. Proceedings of COSIT 2005 (Lecture Notes in Computer Science), Springer Verlag, Berlin/Heidelberg/New York, 2005, 149-164

Rosse Cornelius, Kumar Anand, Leonardo Jose, Mejino V., Cook Daniel L., Detwiler Landon T., Smith Barry: "A Strategy for Improving and Integrating Biomedical Ontologies", AMIA 2005, Washington D.C., USA

Schulz Stefan, Daumke Philipp, Smith Barry, Hahn Udo: "How to Distinguish Parthood from Location in Bioontologies", AMIA 2005, Washington D.C., USA

Patrick Ruch, Robert Baud, Christine Chichester, Antoine Geissbühler, Frédérique Lisacek, Johann Marty, Dietrich Rebolz-Schuhmann, Imad Tbahriti, Anne-Lise Veuthey. Extracting Key Sentences with Latent Argumentative Structuring. Proceedings of MIE2005, Vol. 116, 2005, pp.1052.

William Hersh, Jeffery Jensen, Henning Müller, Paul Gorman, Patrick Ruch (2005). Trans-Atlantic Collaboration for Evaluating Image Retrieval Systems in the ImageCLEF Biomedical Image Retrieval Task. Workshop on NSF/FP6 cooperations at the 2005 eChallenges conference, Ljubljana, Slovenia, October 2005

M Donnelly, T Bittner, C Rosse. A formal theory for spatial representation and reasoning in biomedical ontologies. AI Medicine 2006 ; 36 : 1-28.

Website: <http://www.morphosaurus.net>, updated in August 2005

Website: <http://www.semanticmining.org>, updated regularly

### References 2006 (co-authored papers only – authors from more than one partner)

Rector, AL, Qamar R, Marley T, Binding Ontologies & Coding Systems to Electronic Health Records and Messages. Presented at KR-Med 2006.

Towards an Upper Level Ontology for Molecular Biology, Presented at AMIA 2006

Ontological and Epistemological Aspects of the Denotation of Biological Statements, Presented at KR-MED 2006

From GENIA to BIOTOP – Towards a top-level ontology for biology. Presented at FOIS 2006

Schulz S, Hanser S, Hahn U, Rogers J: The Semantics of Procedures and Diseases in SNOMED® CT. Method Inform Med, 2006, in press.



---

Schulz S, Kumar A, Bittner T: Biomedical ontologies: What part-of is and isn't. *J Biomed Inform*, 2006; 39 (3) : 350-361

Zaiss A, Hanser S, Schulz S: Mapping of ICHI to CCAM Basic Coding System. In: Reichert A, Mihalas G, Stoicu-Tivadar L, Schulz S, Engelbrecht R (Hrsg): *Integrating Biomedical Information: From e-Cell to e-Patient*. Proc European Federation for Medical Informatics Special Topic Conference, April 6-8, 2006, Timișoara, Romania. Berlin: Akademische Verlagsgesellschaft 2006; 281-292

Rector, A, Rogers, J, and Bittner, T, Granularity, scale & collectivity: When size does and does not matter. *Journal of Biomedical Informatics*, 2006. 39(3): p. 333-349

Fellbaum Christiane, Hahn Udo, Smith Barry: Towards New Information Resources for Public Health – From WordNet to Medical WordNet (<http://ontology.buffalo.edu/MFN/MWN-JBI.pdf>), *Journal of Biomedical Informatics*, 39 (2006) 3, 321-332.

Whetzel Patricia, Brinkman Ryan, Causton Helen, Fan Liju, Fostel Jennifer, Fragoso Gilberto, Heiskanen Mervi, Hernandez-Boussard Tina, Morrison Norman, Parkinson Helen, Rocca-Serra Philippe, Sansone Susanna-Assunta, Schober Daniel, Smith Barry, Stevens Robert, Stoeckert Chris, Taylor Chris, White Joe: The Development of FuGO – An Ontology for Functional Genomics Experiments, in: *Omics: A Journal of Integrative Biology*, 10(2) 2006, 199-204.

Erik Sundvall, Rahil Qamar, Mikael Nyström, Mattias Forss, Håkan Petersson, Hans Åhlfeldt, Alan Rector. Integration of Tools for Binding Archetypes to SNOMED CT. *Proceedings of SMCS2006*; 1-3 Oct, Copenhagen, Denmark, p 64-68.

Patrick Ruch, Julien Gobeill, Imad Tbahriti, Robert Baud, Antoine Geissbühler. Automatic Assignment of SNOMED Categories: Preliminary and Qualitative Evaluations. *Proceedings of SMCS2006*; 1-3 Oct, Copenhagen, Denmark.

Erik Sundvall, Mikael Nyström, Mattias Forss, Rong Chen, Håkan Petersson, Hans Åhlfeldt. Graphical Overview and Navigation of Electronic Health Records in a prototyping environment using Google Earth and openEHR Archetypes. *Proceedings of MIE'06 - Studies in Health Technology and Informatics*.

Mikael Nyström, Magnus Merkel, Håkan Petersson, Hans Åhlfeldt. Evaluating bilingual medical terminologies with word alignment methods. Submitted to *MEDINFO 2007*.

Dietrich Rebholz-Schuhmann, Graham Cameron, Dominic Clark, Francesco Beltrame, Jean-Louis Coatrieux, Eva Del Hoyo Barbolla, Fernando Martin-Sanchez, Luciano Milanesi, Ioannis Tollis, Erik van Mullighan, Johan van der Lei. *SYMBiotics: Synergies in Medical Informatics and Bioinformatics – exploring current scientific literature for emerging topics*. BITS 2006, Bologna. *BMC Bioinformatics* (Accepted for publication)

Rebholz-Schuhmann, D., Kirsch, H., Arregui, M., Gaudan, S., Riethoven, M., and Stoehr, P. Protein Annotation by EBIMed. *Nat Biotechnol* 2006; 24(8):902-3.

M. Poprat, K. Markó & U. Hahn. A Language Classifier that Automatically Divides Medical Documents for Experts and Health Care Consumers In: *MIE 2006 – Proceedings of the 20th International Congress of the European Federation for Medical Informatics*. Ubiquity: Technologies for Better Health in Aging Societies. Ed. by A. Hasman, R. Haux, J. van der Lei, E. De Clercq, F. H. Roger France. Maastricht, The Netherlands, 27-30 August 2006. Amsterdam etc.: IOS Press, 2006, pp.503-508 (*Studies in Health Technology and Informatics*, 124)



K. Markó, P. Daumke & U. Hahn Cross-Lingual Alignment of Biomedical Acronyms and Their Expansions In: MIE 2006 – Proceedings of the 20th International Congress of the European Federation for Medical Informatics. Ubiquity: Technologies for Better Health in Aging Societies. Ed. by A. Hasman, R. Haux, J. van der Lei, E. De Clercq, F. H. Roger France. Maastricht, The Netherlands, 27-30 August 2006. Amsterdam etc.: IOS Press, 2006, pp.857-862 (Studies in Health Technology and Informatics, 124)

K. Markó, S. Schulz, U. Hahn: Automatic lexeme acquisition for a multilingual medical subword thesaurus. *International Journal of Medical Informatics (IJMI)*. 2006 (ePublication)

K. Markó, P. Daumke, U. Hahn: Cross-Lingual Alignment of Biomedical Acronyms and their Expansions. *Proceedings of the XX International Congress of the European Federation for Medical Informatics (MIE '06)*, Maastricht, Netherlands. 2006: 857-862.

M. Poprat, K. Markó, U. Hahn: A Language Classifier that Automatically Divides Medical Documents for Experts and Health Care Consumers. *Proceedings of the XX International Congress of the European Federation for Medical Informatics (MIE '06)*, Maastricht, Netherlands. 2006: 503-508.

P. Zweigenbaum, S. Schulz, P. Ruch, editors. LREC Workshop Acquiring and representing multilingual, specialized lexicons: the case of biomedicine, Genova, Italy, 2006. ELDA. Program available at: <http://estime.spim.jussieu.fr/~pz/lrec2006/programme-specialized-lexicons.html>

L. Deléger, M. Merkel, & P. Zweigenbaum. Using word alignment to extend multilingual medical terminologies. Zweigenbaum et al. 5th International Conference on Language Resources and Evaluation (LREC '06): Workshop on Acquiring and Representing Multilingual, Specialized Lexicons, Genoa, Italy. 2006.

P. Daumke, S. Schulz, K. Markó: Subword Approach For Acquiring and Cross-Linking Multilingual Specialized Lexicons. Zweigenbaum et al. 5th International Conference on Language Resources and Evaluation (LREC '06): Workshop on Acquiring and Representing Multilingual, Specialized Lexicons, Genoa, Italy. 2006.

K. Markó, R. Baud, P. Zweigenbaum, M. Merkel, M. Toporowska-Gronostaj, D. Kokkinakis, S. Schulz: Cross-Lingual Alignment of Medical Lexicons. Zweigenbaum et al. 5th International Conference on Language Resources and Evaluation (LREC '06): Workshop on Acquiring and Representing Multilingual, Specialized Lexicons, Genoa, Italy. 2006.

S. Schulz, K. Markó, P. Daumke, U. Hahn, S. Hanser, P. Nohama, R. Andrade, E. Pacheco, M. Romacker: Semantic Atomicity and Multilinguality in the Medical Domain: Design Considerations for the MorphoSaurus Subword Lexicon. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC '06)*, Genoa, Italy. 2006: 1684-1687.

O. Medelyan, S. Schulz, J. Paetzold, M. Poprat, K. Markó: Language Specific and Topic Focused Web Crawling. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC '06)*, Genoa, Italy. 2006: 865-868.

K. Markó, R. Baud, P. Zweigenbaum, M. Merkel, L. Borin, S. Schulz: Towards a Multilingual Medical Lexicon. In *Proc AMIA Annual Fall Symposium 2006*, Washington, DC, November 2006, 534 - 538.

L. Deléger, M. Merkel, P. Zweigenbaum. Contribution to terminology internationalization by word alignment in parallel corpora. In *Proc AMIA Annual Fall Symposium 2006*, Washington, DC, November 2006, 185-189.

L. Deléger, M. Merkel, P. Zweigenbaum. Enriching medical terminologies: an approach based on aligned corpora. In Medical Informatics Europe, 2006. Stud Health Technol Inform. 2006;124:747-52.

M. Nyström, M. Merkel, L. Ahrenberg, P. Zweigenbaum, H. Petersson, H. Åhlfeldt. Creating a medical English-Swedish dictionary using interactive word alignment. BMC Medical Informatics and Decision Making, 2006. 6:35, doi:10.1186/1472-6947-6-35

J. Wermter, K. Tomanek, F. Balzer. Automatische Erkennung und effiziente Annotation von anonymisierungsrelevanten Begriffen in klinischen Freitexten. GMDS 2006, Leipzig. Düsseldorf, Köln: German Medical Science; 2006. Doc 06gmids128

### A.3 Participation in standardisation work

Researchers in the network play an influential role in the process of harmonisation and further development of terminology systems. Examples of areas of interaction are the Gene Ontology, the Foundational Model of Anatomy, and SNOMED CT. Part of the network objectives is also an active interaction with standardisation bodies such as CEN TC251, ISO and IMIA. The research carried out under the auspices of this NoE will also address the need for approaches in Europe which will bridge language barriers and facilitate access for non-English native persons to the large scientific corpus of texts written in English.

Examples of external relations during the first two years of SemanticMining are:

- standardisation activities performed in e.g. CEN TC251 and HL7,
- developers of the Foundational Model of Anatomy (FMA),
- developers of the Gene Ontology (GO),
- developers of SNOMED CT,
- developers of CNPU and LOINC in the area of laboratory medicine.

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](http://www.ehcs.org)) with information on all major eHealth standards and activities.
- The further work on finalising the EN 13606 Health Informatics - Electronic Health Record Communication series (in co-operation with WP26) now also being balloted as an ISO standard and in close co-operation with HL7. 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.
- The finalisation of the EN 12967 Health Informatics - Service Architecture (HISA) standard.
- The work on the EN 1614 model for representing a Structure for nomenclature, classification, and coding of properties in clinical laboratory sciences. After intensive discussions at SemanticMining meetings with WP25, a new version was established. During 2006, EN 1614 has been finalised and approved as European standard: "Health Informatics — Representation of dedicated kinds of property in laboratory medicine". The standard provide a metrology and terminology framework for Laboratory Medicine developed within the NoE and the Committee on Nomenclature Properties and Unit of the IFCC and IUPAC. The new EN 1614 is now being used as input to the LOINC–C–NPU SNOMED CT mapping discussion.
- 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. During 2006 the standard was sent out for enquiry, receiving valuable comments from, among others, NoE participants.

- A CEN Technical Specification for medical knowledge resource metadata descriptions has been developed, Clinical knowledge resources - Metadata (MetaKnow).
- 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).
- On the international arena, further development of SNOMED CT is now the responsibility of the newly established organisation International Health Terminology Standardization Development Organisation (IHTSDO). As part of the IHTSDO organisation, four committees have been established (Content, Technical, Quality, Research and Innovation). As a result of recognition of the SemanticMining network, seven persons from SemanticMining have been selected for these committees (Lars Berg, Marie-Christine Jaulent, Mikael Nyström, Jeremy Rogers, Erik Sundvall, Stefan Schulz, Hans Åhlfeldt). Moreover is the IHTSDO management office in Copenhagen run by persons with an active background in SemanticMining.



## A.4 Patent applied for, contact and agreement for the 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.

## A.5 Research applications and funding

(with relation or reference to SemanticMining)

<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	start 1/1/06
@neuroIST	UKLFR, DIM	approved
SYMBIOmatic, FP6 SSA	EBI et.al	approved
FP7 Patient Safety (DEBUGIT)	Core parts of SemanticMining	application May 2007 in contract preparation phase
IHTSDO	LiU, UKLFR, INSERM, UOM	7 persons elected as Committee members of IHTSDO
+ a series of national research applications related to SemanticMining		

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 was well perceived by the CEC and is ready for contract negotiations. 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.

The DIM is involved in the MultiMatch project and negotiations with the EU are on track. This STREP is a consortium of ten institutions, led by Carol Peter (IST-CNR, Pisa). The project is related to multimodal search strategies and benefits from the developments and resources designed in WP24. Finally, both the UKLFR and the DIM are participating in the ANEUR-IST Integrated Project, starting in January 2006.

Q-REC, a Strategic Support Action in the EU 6th Framework Programme (2006-2008, 30 months) will develop quality criteria, quality labelling and benchmarking instruments for EHR systems, the knowledge resources incorporated in them (such as terminology and



---

archetypes), EHR related standards and open source components. A methodology for evaluating EHR systems and for certification of conformant products will be piloted.

SemanticHealth, a Strategic Support Action in the EU 6th Framework Programme (2006-2007, 24 months) will review the challenges and current threads of research and development (including standards) that contribute to the goal of achieving semantic interoperability within health care. The result of this work will be a review report and a roadmap of future activities (short, medium and long-term) that could be fostered and/or funded to achieve this goal.

FP7 on patient safety has been submitted and approved (DEBUGIT). The new consortia including six SemanticMining partners, is currently in contract preparation phase.