



SemanticMining

NoE 507505

Semantic Interoperability and Data Mining in Biomedicine

Results 2006

Report Version: 1

Report Preparation Date: 2007.02.22

Contract Start Date: 2004.01.01

Duration: 3.5 years

Project Co-ordinator: Hans Åhlfeldt

Department of Biomedical Engineering / Medical Informatics

S-581 83 Linköping University, Sweden

<hans.ahlfeldt@imt.liu.se>



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



Table of Content

Objectives	1
Research gaps	1
Work plan of SemanticMining	2
Partnership	5
Mobility and doctoral programmes	6
Dissemination	8
Joint research programme (WP20-27)	13
Multi-lingual medical dictionary (WP20)	13
Ontology engineering (WP21)	16
SNOMED CT (WP22)	19
Health care statistics (WP23)	20
Text mining in biomedicine (WP24)	22
Terminology systems in laboratory medicine (WP25)	23
The electronic health record (WP26)	24
Laymen terminology (WP27)	30



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.

Partnership

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



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



Mobility and doctoral programmes

The objective of the third year of mobility program (WP6) is 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. This year 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.

Thirdly, as in the first year, 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. We counted 37 participants and the feedback we received from students and discussants was overall very positive. We are planning further improvements in the programme for the next one.

As the project has been extended until July 2007, we intend to continue with the activities undertaken in 2006. There will, however, be a number of extensions and changes.

Recording of student information: Will continue as is until the end of the project.

Visit grants scheme: The visit grants scheme will continue, the mobility scheme having been granted a further 10,000 Euro.

Doctoral colloquium: We intend to organize another doctoral colloquium, at the Summer School/Closing event in Barcelona, June 2007.

2006 was a successful year 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 partner sites, definitely has a very positive effect on the progress of PhD students doctoral programmes. Several PhD student in the network now have established close contacts with tutors at other partner sites.



Mobility program 2005 Name	Duration	From	To	Award
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

Mobility program 2006 Name	Duration	From	To	Award
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



Dissemination

The MERMIG platform is used as public website available at 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 2006, 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.

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.

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/](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. Number of participants was 57 - industry: 14 - national centers: 11 - academia: 32. Nations: - Germany (27) - U.K. (9) - USA (5) - France (4) - Japan (4) - Switzerland (3) - Sweden (2) - Spain (1) - Finland (1) - Taiwan (1).

Training Course in Biomedical Ontology

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

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

Potential members of the SNOMED CT SDO (Standard Developments Organization) had meetings in Copenhagen in connection to the SMCS 2006 Conference with representatives of the College of American Pathologists in order to set up a timeframe for transfer of the IPRs of SNOMED CT. Due to this development and the need to make decisions on national level about whether or not to join the SNOMED CT SDO, the SMCS 2006 Conference was offered at the right time and met a high demand of health professionals, system vendors, researchers, and health policy makers. This may explain the extraordinary response to this conference, the high level of scientific contributions and the readiness of top experts to give tutorials and keynote presentations. The feedback of participants was positive to enthusiastic. Excellent keynote presentations showed not only achievements in the SNOMED CT development, but also shortcomings, concerning the content and maintenance quality. The Danish SNOMED CT localization experience was presented and shown high interest by representatives from other European countries. The new route SNOMED CT is taking by the foundation of the SNOMED CT SDO and its implications were extensively discussed. In summary, SMCS 2006 was a highly satisfying event which took place at the right place, with the right content, at the right time.

Terminology Conference in Romania

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



Input to standardisation work

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) 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 with respect to the possible exploitation of the multilingual medical dictionary. Discussions are still ongoing. 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 in 2006.

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 remaining phase of the project, “public-friendly” material will be produced where results from the SemanticMining project will be 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 EFMI, the European Commission, and national networks. The revised list of deliverables contains deliverables summing up research results as well as available tools and services produced within SemanticMining.

Joint research programme (WP20-27)

The research activities in SemanticMining have during 2006 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 ~ WP20, WP26)

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. Work package 28 on terminology services has been terminated and relevant activities merged with WP21 and WP26. Measurements of quality indicators as well as follow-up of milestones are presented below.

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 during 2006

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.

SUITSEARCH[®]
HEALTH RECORDS

juckendes erythem

PIZ: Nachname: Vorname: Geschlecht:

4 Ergebnisse in 3 msec gefunden, 100% Genauigkeit. -- Zeige Ergebnisse 1 bis 4: [Info](#)

22990373	Presley	Elms	08.01.1935	♂
... Anamnese und Befund: Seit 01.03.06 kam es bei Herrn Presley zur Ausbildung von Erythemen mit teilweise papulösen Hautveränderungen am linken Unterschenkel welche mit starkem Juckreiz und Brennen vergesellschaftet waren. Im Laufe der letzten 2 Tage sei es dann zu einer ... Ersteller: Dr. Albert Schweitzer Erstellung: 06.03.2006 Übermittlung: 09.03.2006				
25724135	Hendrix	Jimi	27.11.1942	♂
... unserer stationären Behandlung befand. Diagnosen: 1 Pruriginöses Ekzem 2 Demenz 3 Harninkontinenz 4 Eingeschränkte Mobilität bei Hüftkopfnekrose rechts 5 Schwerhörigkeit Anamnese und Befund: Herr Hendrix lebt in einem Pflegeheim und klagte dort seit ca. 4-6 Monaten über starken Juckreiz am ... Ersteller: Dr. Albert Schweitzer Erstellung: 09.03.2006 Übermittlung: 09.03.2006				
22118404	Joplin	Janis	19.01.1943	♀
... am unteren Rücken. Ebenso sieht man am Decolletée zwei ca. 3 mal 2 cm große erythematoöse und nässende Erosionen bei Zustand nach Lasertherapie. Am Hals und prästernal sowie am rechten Unterschenkel weiterhin erythematoöse z.T. erosive z. T. hyperkeratotische Hautveränderungen Die ... Ersteller: Dr. Albert Schweitzer Erstellung: 01.02.2006 Übermittlung: 06.03.2006				
21965936	Kelly	Grace	12.11.1929	♀
... Ecural Lösung). Zusätzlich erhielt die Patientin Antihistaminika gegen den Juckreiz . Bei Aufnahme finden sich strecksseitenbetont am gesamten Integument multiple erythematosquamöse Plaques. Die Kopfhaut ist gerötet und weist eine feine weißliche Schuppung auf. Axillär beidseits inguinal und in der Rima ani ... Ersteller: Dr. Albert Schweitzer Erstellung: 02.03.2006 Übermittlung: 02.03.2006				

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.

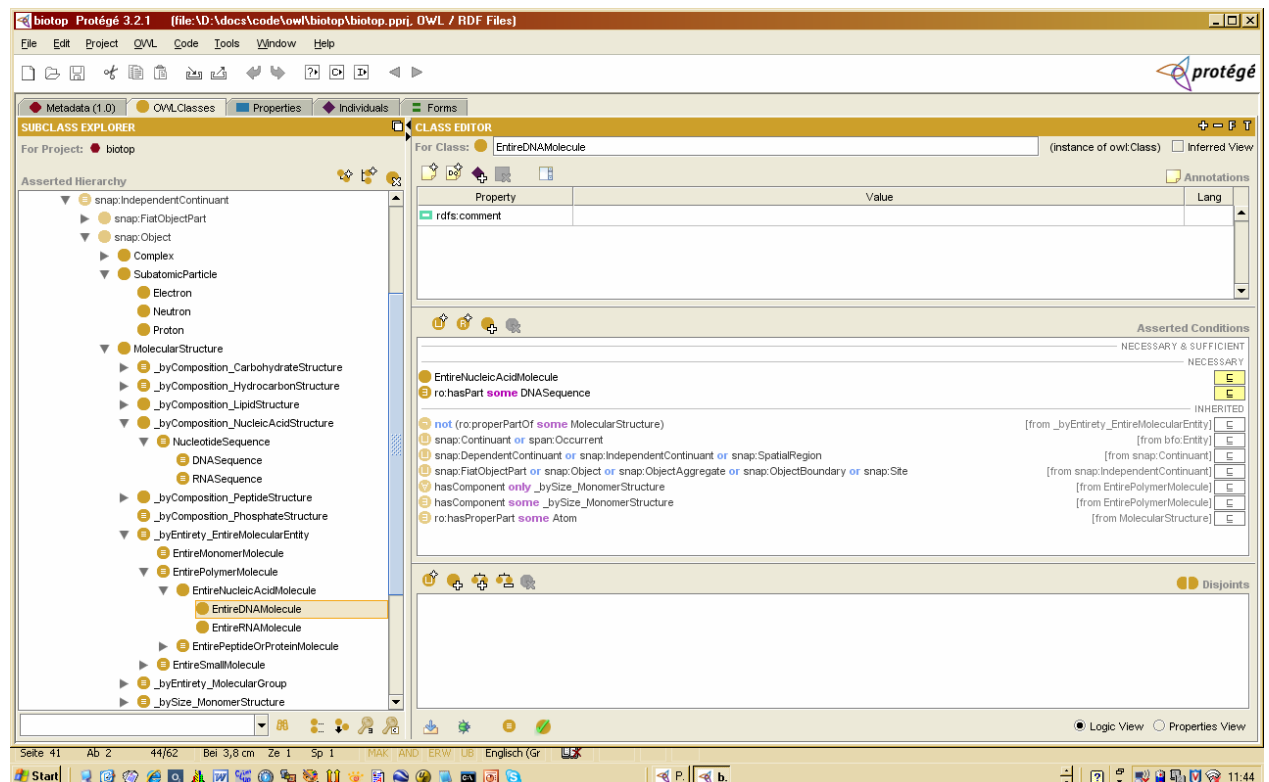
Q7 Co-authoring of research papers, reports and educational materials:

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).



Plans for 2007

We plan work in four categories:

- The interaction of Ontologies/Terminologies with Medical Records
- Ontology/Terminology quality with specific reference to SNOMED CT
- Convergence and description of upper ontologies
- Education and outreach

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 SemanticMining 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.

Ongoing research work includes mapping of SNOMED CT terms to legacy classification systems such as ICD10 (International Classification of Diseases) and NCSP (Nordic Classification for Surgical Procedures). Initial 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 has also started in aligning SNOMED CT with the C-NPU format and the European standard EN 1614 on laboratory requests and reports. 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 WP22 and WP25.



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.

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

- The first SemanticMining conference on SNOMED CT was highly appreciated
- SNOMED CT is recognised as a very interesting use case in the Danish EHR implementation
- Translation of SNOMED CT within the Danish EHR-project is ongoing and is providing valuable experiences for future translational projects
- Processes for future maintenance and quality assurance of the SNOMED CT content is crucial.

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.

It 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 test set is designed to enable separate analysis of different kinds of relationships, e.g. to check whether physicians are more likely to agree on (the ranking of) generic relationship than partitive relationships. A subset of the material will be used for analysing properties of the rankings in order to examine their validity as a metric.

The study is Web-based and a questionnaire application has been built and is running on a server in Linköping. After a successful pilot study, we are currently collecting data for the real study. Physicians in different countries around Europe have been recruited and are taking the survey. Since not all data are available, only preliminary results exist at the moment. However, based on what we have seen this far there seems to be evidence for agreement on (the ranking of) semantic distances.

The study will be completed during the spring of 2007 and will result in a scientific paper co-written by the work package team. If possible, the results will also be presented at the Semantic Mining closing meeting in Barcelona. 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 ([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.

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.

The screenshot shows the EBIMed search interface. A search box contains the term 'wnt'. Below the search box are links for 'Advanced Search' and 'Query Syntax'. The search results are displayed in a summary view, including a 'Summary' section with a 'HitPair table' icon. A small image of a book and a bee is shown next to the text '3656 Abstracts'. A table summarizes the results by type, and a 'HitPair table' section shows a grid of results for the query 'wnt'.

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

Uniprot	Uniprot	Cellular component	Biological process	Molecular function	Drug	Species
beta-catenin (score: 6853)	APC or APCs (240/428) GSK-3 beta or glycogen synthase kinase-3 beta (154/198) Axin or axins (145/259) E-cadherin (97/162) 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.

Conclusion

The lead contractor of WP24 (EBI) has its origin in the domain of bioinformatics and molecular biology in contrast to medical informatics. EBI is the only project partner, which provides IT services to the public fulfilling information needs in the biomedical domain. It is obvious that the NoE can contribute to information provision to the public, thus the developmental work in the NoE can be transformed into IT solutions to the public.

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 have also been pursued during 2006.

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.

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.

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.

In recent months a new collaborative has been formed jointly between HL7 and *openEHR*, known as the Detailed Clinical Models (DCM) Group. The DCM aims to build up empirical experience of EHR Archetype development and to host a library of good practice Archetypes.



Archetype tools

Archetypes provide the ability to:

- define models of clinical content (e.g. the recording of ‘adverse reaction’) by clinical and other domain specialists;
- connect data stored in the EHR in a valid way to terminologies;
- create semantic queries based on paths extracted from archetypes.

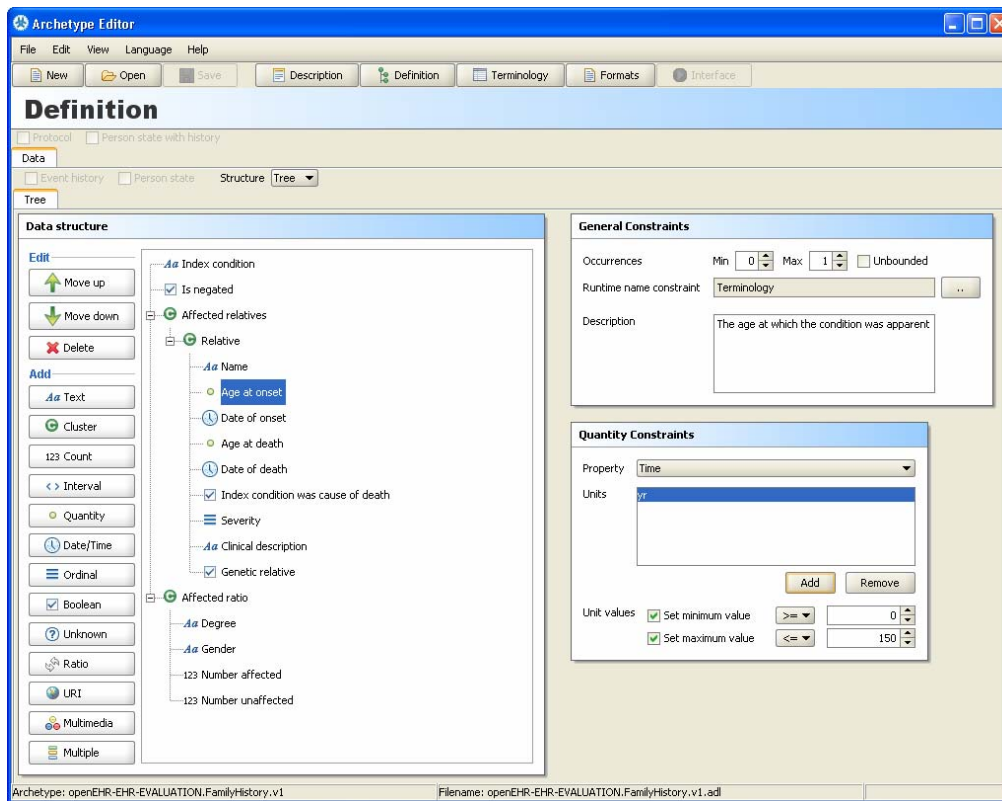
One of the major areas of semantic research in which UCL and Ocean Informatics are engaged is the design and use of Archetypes and ‘Templates’ for modelling clinical content outside of the models used to build software (i.e. UML models etc). In parallel, the Karolinska Institutet and the University of Linköping have been maturing a Java Archetype parser and an editor tool; functionally this editor is very close to the Ocean one, with some improvements in the way archetype internal terminology is manipulated and displayed on the screen. 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.

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 in the following ways:

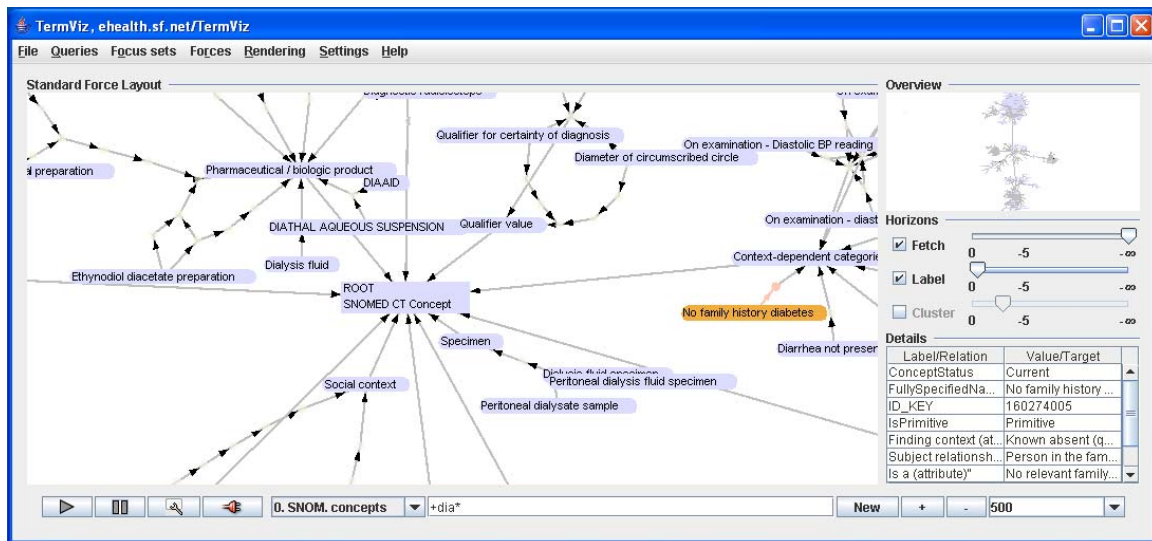
- exposure of other participants to archetypes, including those from Manchester University (doing research on semantic web including in medicine), and Santé Publique et Informatique Médicale (SPIM) has increased the user base, providing better critical feedback;
- interaction with the Manchester group in particular has provided valuable insights into how better to model the connection between archetypes and terminology.

These tools are all starting to be used by modelling groups around the world, including in the NHS and in HL7.

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.



Screenshot 4. Archetype editor as part of openEHR software components (WP26).



Screenshot 4. Terminology browser (TermViz) as part of openEHR software components (WP26).

Binding EHRs and archetypes to terminologies

The University of Manchester has made use of ongoing advances with archetypes to refine their theory of how to use terminology in information systems. 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. There have been three major activities:



- Binding of terminology and data structures: analysis of the formal relation between ontologies and data structures, including both Archetypes and the HL7 RIM.
- Matching of terminology to data archetypes: development, implementation and evaluation of methodologies and tools, integrated with the Archetype Editor developed by LiU, for finding and filtering the best matches from SNOMED for binding to the ontology section of an Archetype.
- Analysis of issues of quality in SNOMED.

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”). This work was demonstrated to the whole of WP26 during a hosted session at the hospital in December 2006.

In newer research, INSERM are 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 its reference 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. The ontology of "blood pressure", for instance, will account for the physical phenomenon, its measurement (a process), the outcome of such a process (data) and the participants to the process (physicians, instruments, etc.). An archetype has a different scope and its aim is not to feature such a semantic representation. This is the reason why an evolution is needed from a 'classic' openEHR architecture to an ontology-based patient record, whose information elements are mapped into a reference ontology of medicine.

Archetype-based EHR implementations

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. This early work was done with much advice from UCL but later this implementation was influenced by early adopters from a number of other institutions, not the least the LiU group in 2006 (whose editor was outlined above). The software build is managed by Maven from the Apache Software Foundation. It provides comprehensive support of software projects and handles software library dependency particularly gracefully. The Java implementation projection software is currently change controlled by a Subversion repository hosted on the *openEHR* site. The Java implementation experience has been summarized and written up as the Java Implementation Specification (ITS) of *openEHR* design which is now included as part of the official release of the *openEHR* specifications.

Archetypes work in conjunction with a Reference Model to fully describe all data to be stored about a particular aspect of health or health care within an EHR. UCL has historically developed and maintained a working archetype-based EHR server as a proof of concept evaluation of its research into EHR requirements and design. As part of Semantic Mining 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. The initial study was a success and has now been rounded-out into a full implementation. 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. A further current topic of research is how metadata



annotations could be used to automatically construct a visual display for the data, dramatically reducing application development time whilst still keeping all knowledge about a clinical construct and its use in a single location.

EHR data visualisation

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. In the project both the *openEHR* Java reference implementation and the EHR Bank from Ocean Informatics have been used as backend. A unified prototyping environment for visualization and navigation of electronic health records. Google Earth (GE) has been used for handling display and interaction of clinical information stored using *openEHR* data structures and 'archetypes'. The structured and multifaceted approach of handling time that is possible with archetyped *openEHR* data lends itself well to visualizing and integration with *openEHR* components is provided in the environment.

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.

Development of EHR archetypes

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. These scenarios could best be described as *de facto EHR Archetypes*, since they correspond to a precise user or system requirement with the priority being faithfulness to this requirement, with limited capacity to introduce best practice modifications.

A fine-grained record structure can be used to specify the details of a multi-part clinical statement, in which leaf nodes in the structure are populated with individual terms or qualifiers from a terminology system. However some multi-part terminological statements can alternatively be expressed as a single compound terminology object using a terminology like SNOMED-CT that supports post-co-ordination. Since ideally the fewest possible number of EHR Archetypes should be adopted for each kind of clinical information, a generic approach to minimising the number of options for dealing with term combination is needed. The approach of using the SNOMED-CT concept model as the design basis for a small set of high-level Reference Archetypes is being explored, and will be published later.

Assessment against relevant quality indicators

Q1 Workshops and symposiums

Work package 26 has held two major workshop 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:

- (a) The first subcorpus, the “non-expert corpus”, derives from a number of Swedish daily newspapers and other online health information sources targeted to consumers (e.g. the Swedish NetDoktor). Approx. 85,000 tokens (tokenised);
- (b) The second subcorpus, the “expert corpus”, derives from two Swedish medical resources intended for professionals and specialists across a broad spectrum of medical professions: *Läkartidningen*, published weekly by the Swedish Medical Association, and *Dagens Medicin*, a news site for medical professionals. Approx. 85,000 tokens (tokenised);
- (c) Texts from various “ask-the-net-doctor” sites. This subcorpus consists of questions posted by non-experts and answers provided by medical professionals. Approx. 22,600 tokens (tokenised).

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. This work has resulted in a conference presentation (see publications, below) and a manuscript (Sonia Krivine & Natalia Grabar: “Détection automatique des catégories scientifique et vulgarisée des document web médicaux. Étude comparative entre les corpus de documents en russe et en français”).

Requirement analysis

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 and consequently to be taken into account when creating systems. This work resulted in a report, an internal WP27 deliverable (“Some recommendations for the creation of patient-friendly documents”; November 2006).

Empowering the patient with language technology

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”.

Hypertext

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. Given the complex nature of such information and the differences between potential recipients, they have been exploring the use of hypertext as a medium. Hypertext lends itself particularly well to the differentiated presentation of information, allowing users to explore a text at different levels of articulation and depth, through different reading paths. 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. This work in progress has been reported in workshop, conference and journal papers.

Scripted dialogue generation

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. Starting with the OU CLEF system for summarization of patient records, the research will investigate a method for formulating technical concepts in CLEF as non-technical conceptual structures that can be expressed in everyday language in dialogues.



This is motivated by empirical studies regarding the beneficial effects of dialogue on learning. One of the effects that was found is that dialogue helps people express their own questions. A dialogue generation system that provides a patient with a dialogue about their situation, could be beneficial if shown just before the patient talks to a consultant. A short report on this work has been published as an Open University Computing Science Department Technical Report.

EHR server development and linkage of archetypes to educational material

The combination of a new EHR interoperability standard, the requirements for a high performance EHR repository to support secondary use and semantically indexed-EHR data, and the more recent agenda introduced through WP27 have all required 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. This will permit future requests from an anticoagulant application for patient educational resources about a particular form object to be mapped to a predefined set of resources. 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.