



SemanticMining

NoE 507505

Semantic Interoperability and Data Mining in Biomedicine

Final SemanticMining Report on Contributions to the European Research Area Covering period 2004.01.01 – 2007.06.30

Report Version: 1

Report Preparation Date: 2007.09.15

Classification: PU

Contract Start Date: 2004.01.01

Duration: 3.5 years (42 months)

Project Co-ordinator: Hans Åhlfeldt

Department of Biomedical Engineering / Medical Informatics

S-581 83 Linköping University, Sweden

<hans.ahlfeldt@imt.liu.se>



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



Table of Content

FINANCIAL/ADMINISTRATIVE CO-ORDINATOR	1
1 OBJECTIVES OF SEMANTICMINING	2
1.1 Research gaps addressed	2
1.2 Work plan of SemanticMining	3
1.3 Partnership	6
2 SCIENTIFIC RESULTS AND CONTRIBUTION TO THE EUROPEAN RESEARCH AREA	7
2.1 Creation of a multi-lingual medical dictionary	7
2.1.1 Main activities and scientific results	7
2.1.2 Contribution to overall objectives of NoE	8
2.1.3 References to deliverables, quality indicators and milestones	9
2.1.4 Future opportunities, continued collaboration, new joint research programmes	9
2.1.5 Conclusions	9
2.2 Principles in ontology engineering	10
2.2.1 Main activities and scientific results	10
2.2.2 Contribution to overall objectives of NoE	15
2.2.3 References to deliverables, quality indicators and milestones	15
2.2.4 Future opportunities, continued collaboration, new joint research programmes	15
2.2.5 Conclusions	15
2.3 SNOMED CT	17
2.3.1 Main activities and scientific results	17
2.3.2 Contribution to overall objectives of NoE	18
2.3.3 References to deliverables, quality indicators and milestones	18
2.3.4 Future opportunities, continued collaboration, new joint research programmes	19
2.3.5 Conclusions	19
2.4 Health statistics, semantic distance and ontologies	20
2.4.1 Main activities and scientific results	20
2.4.2 Contribution to overall objectives of NoE	21
2.4.3 References to deliverables, quality indicators and milestones	21
2.4.4 Future opportunities, continued collaboration, new joint research programmes	21
2.4.5 Conclusions	21
2.5 Text mining in biomedicine	22
2.5.1 Main activities and scientific results	22
2.5.2 Contribution to overall objectives of NoE	23
2.5.3 References to deliverables, quality indicators and milestones	24
2.5.4 Future opportunities, continued collaboration, new joint research programmes	25
2.5.5 Conclusions	25
2.6 Terminology systems in laboratory medicine	27
2.6.1 Main activities and scientific results	27
2.6.2 Contribution to overall objectives of NoE	27



2.6.3 References to deliverables, quality indicators and milestones	27
2.6.4 Future opportunities, continued collaboration, new joint research programmes	27
2.6.5 Conclusions	28
2.7 The semantically well-defined EHR	29
2.7.1 Main activities and scientific results	29
2.7.2 Contribution to overall objectives of NoE	39
2.7.3 References to deliverables, quality indicators and milestones	39
2.7.4 Future opportunities, continued collaboration, new joint research programmes	40
2.7.5 Conclusions	40
2.8 Laymen terminology	41
2.8.1 Main activities and scientific results	41
2.8.2 Contribution to overall objectives of NoE	44
2.8.3 References to deliverables, quality indicators and milestones	44
2.8.4 Future opportunities, continued collaboration, new joint research programmes	45
2.8.5 Conclusions	45
2.9 Mobility program	46
Descriptions of PhD work and dissertations	46
Visit Grant Scheme	46
Doctoral Consortium at Summer School	48



Financial/Administrative co-ordinator

Name: Hans Åhlfeldt

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

Phone Numbers: +46 13 227574

Fax Numbers: +46 13 101902

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

Project websites: www.semanticmining.org

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



1 Objectives of SemanticMining

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.

1.1 Research gaps addressed

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.

1.2 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.



1.3 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

2 Scientific results and contribution to the European Research Area

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

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

2.1 Creation of a multi-lingual medical dictionary

2.1.1 Main activities and scientific results

The main objective of WP20 was the creation, standardization and pooling of multilingual medical language resources, i.e. lexicons and thesauri.

We considered the main objective of WP20 as achieved. A common standard for the exchange of lexical data including morphological and semantic information has been formulated. Several techniques for automatic lexicon population and mapping have been created. Using these methods, the lexicon has been populated by medical terms in four languages. The quality of the automated mapping was finally evaluated. A wealth of collateral scientific work, mainly embedded into PhD theses has been carried out. Software tools have been developed and tightened up. A spin-off company was founded at one site.

The WP20 activities were mainly characterized by the following strands of collaborative work, research and developments:

- Common interchange format for lexical information
- Common link format for semantic lexeme mapping
- Common interchange format corpora.
- Domain-specific lexical sources at different locations
- Pooling of sources on a common platform
- Use of multilingual lexicons in prototypical applications and research scenarios.
- Corpus-based cognate mapping techniques, method and evaluation
- Corpus-based semi-automated lexeme acquisition, method and evaluation
- Corpus-based acronym learning techniques
- Word alignment techniques on parallel corpora
- Concept similarity measurement techniques
- Swedish primary health care corpus
- Evaluation of cross-language retrieval
- Evaluation of different word alignment techniques
- A semantic cross-mapping of existing lexical sources using two alternative methods
- Evaluation of semantic role extraction from a medical document collection in French
- Re-engineering of the subword indexing tools (MorphoSaurus)
- Web based editing tool for collaborative editing of subword lexicons
- Platform for corpus exchange

- Evaluation of correctness and the completeness of the multilingual dictionary.
- A spin-off company was founded at UKLFR 2007, after successful application for a two year grant (German government)
- Contact to Elsevier Health Sciences Division established.

The general objective of the NoE, the cross-fertilization between scientific disciplines has been addressed by WP20 by promoting numerous joint activities involving computer scientists, biomedical domain experts, and linguists, all of them covering several European languages.

The common repository of medical terms in different languages was iteratively fed by new entries and a final version was released in November 2006. 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 an exploitable resource will widely exceed the resources available.

Accordingly, the evaluation results (coverage and correctness) for the intra- and cross-lingual term mapping showed that fully automated bootstrapping methods are only suited to provide raw data that inevitably have to be curated manually by domain experts. 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. Results of WP20, are being re-used in the FP6 BootSTREP project, and will be used in the FP7 DEBUG-IT project. There are also contacts to a big international medical publisher who showed interest into the WP20 methodologies, tools, and resources.

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. A seamless follow-up of the WP20 activities has not been embarked on, because the current FP7 calls are not specifically addressing the development of multilingual resources in a generic way. However, part of WP20 partners participated in the elaboration of a new EU proposal, called DEBUG-IT. Addressing patient safety, this proposal suggests the use of text mining for a semantic analysis of unstructured content in electronic health records. To this end, important results and resources of WP20 will be re-used.

2.1.2 Contribution to overall objectives of NoE

WP 20 addressed a central objective of EU research and cooperation, namely multilinguality. This brought together partners from linguistics and medical informatics from different countries, sharing and developing resources and scientific results. This interaction was also supported by student mobility and informal co-tutoring.

However, the sustainable development of freely available terminological resources on a large scale for a vast domain as the life sciences (from cell to population across Europe) should constitute an objective on its own, requiring an effort that is at least one order of magnitude higher than what can be supported by a EU NoE. We therefore regret that the development of multilingual terminological resources is currently not founded by the EU, so that a promising strand of cooperative work as in WP20 cannot be followed up.

2.1.3 References to deliverables, quality indicators and milestones

An assessment of the WP20 activities against the quality indicator defined in the Description of Work (DoW) of SemanticMining, 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 are mainly restricted to internal meetings. WP20 participated actively in the all four SemanticMining summer schools and doctoral consortia, the last being held in Barcelona in June 2007. There were numerous participations in national but predominantly in international conferences, presenting activities performed by WP20.

Q2: As mentioned above, the sharing of resources is a central objective of WP20. This has been accomplished by the joint development of language resources and tools.

Q5: A formal co-tutoring of PhD students involved in WP20 related research has not been carried out in this project, mainly due to incompatibilities between the national academic systems. However, informal tutoring support could be registered in some cases.

Q6: There were a total of 12 short- and medium-term visits between WP20 partners

Q7: There were a total of 30 research papers co-authored by WP20 partners.

2.1.4 Future opportunities, continued collaboration, new joint research programmes

Some WP20 partners are participating in a new FP7 proposal. The spin of Company AVERBIS GmbH, that grew out of the Freiburg partner, is further developing and bringing to market several tools developed with WP20 resources and is maintaining joint and continuing activities together with several WP20 partners.

2.1.5 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 outcome assessment of the multilingual dictionary has provided evidence that the creation of a common standard is highly valuable for bringing together heterogeneous linguistic sources. The cross-dictionary matching experiments have clearly shown that a high amount of raw material can be produced by adapted automated techniques. This has shown that lexicon acquisition and mapping processes can be considerably accelerated, but that nevertheless a high amount of work – exceeding, by large, the dimensions of this project – for cleansing, curation, and maintenance would still be required.

2.2 Principles in ontology engineering

2.2.1 Main activities and scientific results

The objectives of the work package were:

- *Education and dissemination* – to share understanding across the three ontological disciplines, especially those issues unique to or especially important to the biomedical domain, and to coordinate future research efforts so as to achieve coherent divisions of labour and to avoid duplication of effort. This effort also included sponsoring students, exchanges, and workshops to increase the knowledge and skills of both partners and others in the EU community in ontologies and their application
- *Input to standards* – to coordinate input into standardisation activities relevant to biomedical ontologies, including the emerging semantic web and its associated ontology authoring and delivery environments as well as established international medical informatics standards activities such as ISO, CEN, IEEE and HL7.
- *Improved consensus on upper ontologies* – to contribute to an understanding of the differences and commonalities in the various upper ontologies proposed and work towards their harmonisation.
- *Convergence of medical and biological ontologies* – to foster the convergence of ontologies originally developed independently for clinical medical and molecular biology in order to contribute to the convergence of the disciplines and translational biomedical research.
- *Contributions to evaluation of SNOMED-CT* – to provide an ontological and terminological framework for understanding the issues in using SNOMED CT along with WP 20.
- *Interworking of Ontologies and Electronic Health Records* – to develop principled methodologies for standardising the use of terminologies with EHRs, particularly SNOMED-CT and Archtypes/CEN 13606.
- *Tools for ontology development and management* – to contribute both requirements and practical developments to key ontology engineering and tools efforts in an open source framework.

Education and dissemination

The work package has been active in developing training and education programmes both for the participants and for wider groups in the EU and beyond, by:

- Providing modules for each of the Semantic Mining Summer Schools
- Providing a series of tutorials on ontologies at major conferences
- Sponsoring two tutorial events at Dagstuhl Castle for a Europe-wide audience
- Taking the lead in the first conference on Foundations of Clinical Terminologies and Classifications under the sponsorship of EFMI in 2006 and provided two of the keynote speakers
- Sponsoring two tutorials on the new Web Ontology Language, OWL, in 2004 and 2006.



- Taking a major part in the lead for the first workshop on Knowledge Representation in Medicine in November, 2006.

In addition the project sponsored a series of exchanges of PhD students and sponsored one PhD student working on the interface between ontologies and electronic health records.

Input to Standards - Interworking of Ontologies and Electronic Health Care Records

The work on standards has focused on the interworking of ontologies/terminologies and Electronic Health Care Records. The project was active in both CEN and HL7. It put particular, Manchester devoted effort into the early phases of the Terminfo effort of HL7, which aimed at standards for coordinating SNOMED-CT and HL7 messages. This effort was coordinated with consulting with the UK National Health Service on mechanisms for using OWL to specify such standardisation.

The work on the work has involved three joint workshops with WP26 and resulted in a major piece of software developed by the PhD student sponsored by the project, which extends the Archetype Editor developed by LIU to aid developers to bind archetypes to SNOMED-CT. Deliverable D21.5 on quality assurance of biomedical ontologies was heavily influenced by work on SNOMED, although the more open licensing available from the SNOMED IHSDO did not become available until to late in the project to allow the level of evaluation which the project would have liked to have achieved.

A series of papers have been produced documenting this work [1, 2][3] which is continuing both in collaboration with the UK NHS and through the Semantic Health Roadmap Project. It represents a major new approach to understanding the relationship between ontologies, coding systems and information models which is bearing fruit in further collaborations and industrial partnerships. The approach has been prototyped with the UK NHS in a separately funded study binding NHS messages to SNOMED codes.

The approach has been most elaborated with respect to the Archetype models of electronic health records used and developed in WP26 and SNOMED-CT. The basic approach is regard both codes and Archetypes as being about information structures, whereas the ontology proper is about patients and their conditions. The codes in the information model are derived from the ontology and bound to the information model. This process can be implemented either in special tools or by specialised user interfaces to the generic OWL infrastructure. The overall pattern is shown in Figure 1.

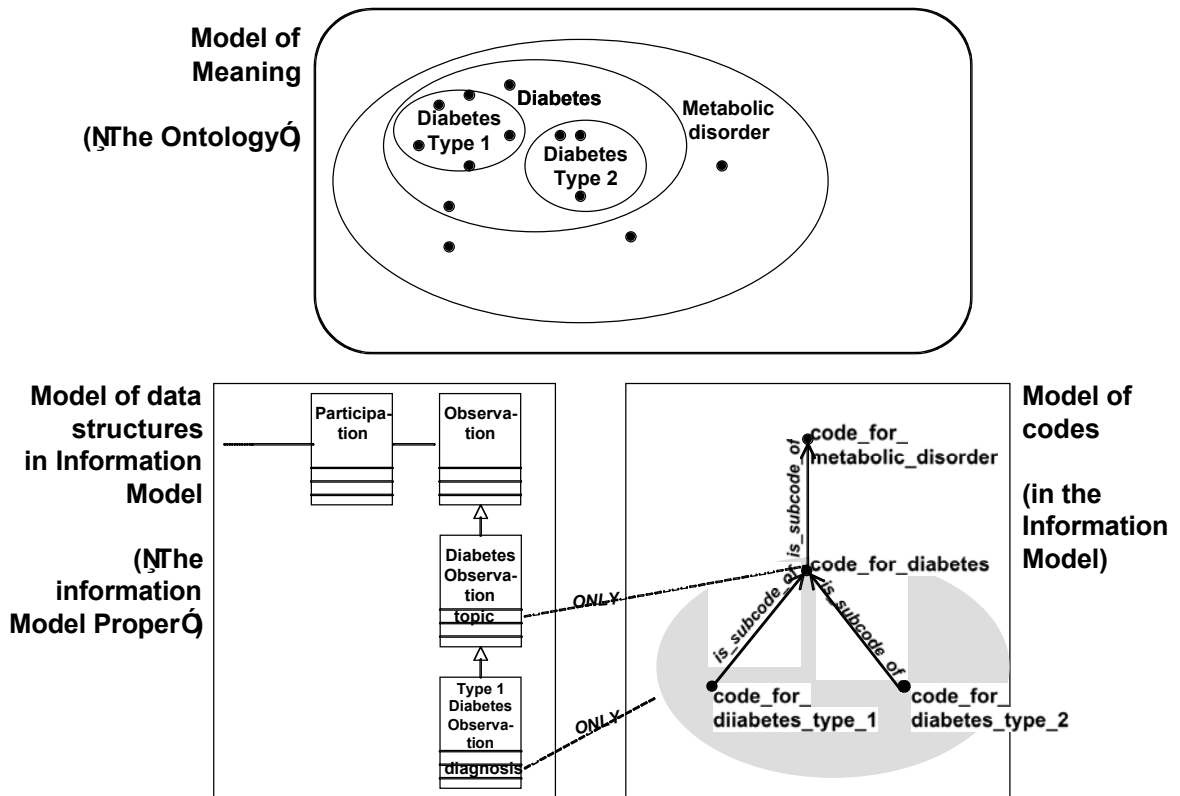


Figure 1: Binding Ontologies to Medical Records or Archetypes (See Rector et al. 2006)

Finally the group has been highly influential in stimulating the development of the new OWL 1.1 pro-standard, which is designed to address issues raised by users in their experience with the initial OWL 1.0 standard. The work package leader (Rector) will lead the Education and Requirements taskforce in the new W3C working group on OWL 1.1.

The approach leads to an interpretation of codes which, amongst its other advantages, handles negation entirely naturally within a DL framework. The results of the approach are shown with the classic example of skull fracture with or without various kinds of bleeds in figure 2

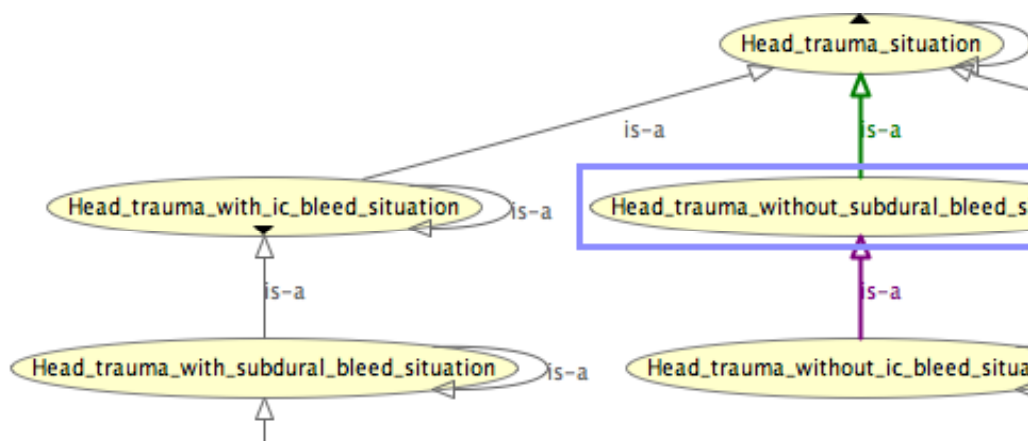


Figure 2: Portion of classification of head trauma with and without bleeding showing correct inversion of negatives. (See Rector 2007)



Improved Consensus on Upper Ontologies - Convergence of Medical and Biomedical ontologies

The period of the project has been a very active time in ontology development in biomedicine. Major effort has been invested by partners with important consortia in the molecular biology including with the Gene Ontology consortium and the Open Biological Ontologies (OBO) group which emerged from it, with the BioPax consortium, and with the emergence of the Healthcare and Life Sciences Special Interest Group from W3C. Joint papers emerging from this effort are listed in Section 3.2. In addition there has been a large number of publications by partners in the consortium, of work that has grown, at least in part, out of the consortium efforts.

The project as developed a sound theoretical base for understanding the issues amongst the various upper ontologies and ontology engineering methods. Deliverable D21.2 concerned itself with the ontological foundations, and Deliverable D21.3 with the computer science and logical foundations of ontologies.

A major focus for IFOMIS and University of Manchester has been to engage with the European Bioinformatics Institute (EBI) and its Open Biomedical Ontologies effort. Much of the work of IFOMIS is now being made available through the OBO Foundry Web site.

The project is also engaging with the US National Center for Bio-ontologies, in which members participate and sit on the strategic advisory board, and with the US National Cancer Institute (NCI) and UK National Cancer Research Institute (NCRI).

The venues for these efforts have been the summer schools and workshops sponsored and supported by the project. During 2005, following on from the initial Summer School, the partners in the project collaborated on a major paper on Relations in Biomedical Ontologies [4]. The workpackage has resulted in a series of bilateral and multilateral meetings, with a gradual convergence of the BioTop ontology [5], first between IFOMIS and University of Freiburg, and then with the increasing involvement of University of Manchester and CNR.

The theoretical foundations for this work are contained in Deliverables D21.2 and D21.3 on the ontological and computer science foundations of ontologies, respectively.

For a complete list of joint publications see Section 3.

Contributions to the Evaluation of SNOMED-CT

The work on interworking between ontologies, terminologies, and Electronic Health Records has been largely directed at SNOMED-CT as the most likely emerging international standard.

The PhD student sponsored by the project has developed a system, MoST, for binding the OpenEHR Archetypes used by WP26 to express medical records models to SNOMED-CT [1]. The student's thesis is in the final preparation stages, but provides clear indications of deficiencies and criteria for quality for both SNOMED and Archetypes if they are to be bound and used together repeatably. A screenshot is shown in Figure 3.

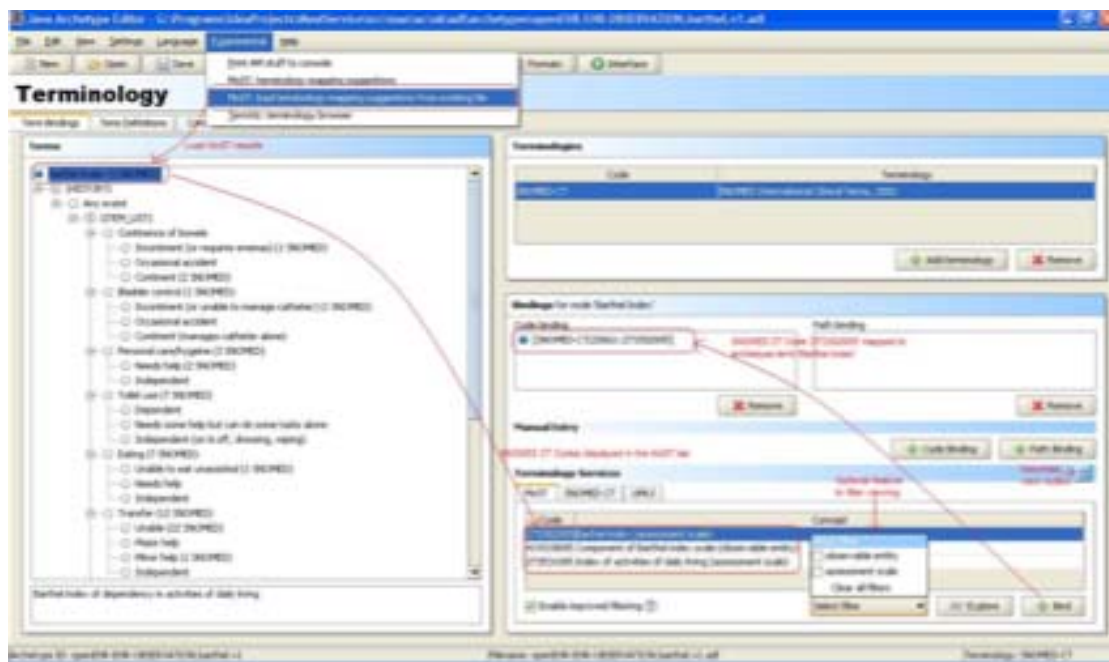


Figure 3: Screenshot of the MoST system

Over the same period members of the workpackage from Linköping and Manchester have been active in both the UK and European workshops on SNOMED, and the workpackage leader, Alan Rector, gave the keynote address at the Semantic Mining Conference on SNOMED in Copenhagen on October 2006. Earlier that year the workpackage organised the first workshop on the Foundations of Clinical Terminology and Classification (FCTC) under the aegis of the European Federation of Medical Informatics (EFMI), much of whose content and discussion was devoted to aspects of the analysis and development of SNOMED-CT.

The foundation of this work is contained in Deliverable D21.5 on quality assurance of ontologies.

Tools for Ontology Development and Management

The technology for implementing and delivering ontologies has matured rapidly over the life of the project. At the beginning of the project the initial new Web Ontology Language, OWL, had just been published by W3C. By the end of the project a proposed revision to OWL 1.1 had occurred and the W3C was in the process of setting up a working group to extend the definition. Protege-OWL was developed in a collaboration focused on University of Manchester and Stanford University focused on user requirements and developing tools to fit those requirements. Semantic Mining contributed to the development of the initial design and to the revisions to implement new features for OWL 1.1 particularly required by biomedicine, including improved support for constructs required for anatomy and the approach to “qualities” used in the Basic Formal Ontology (BFO) from IFOMIS. The tools were used by many partners in the project and in the course of the tutorials and workshops.

A separate but related development was the MoST tool, which is an extension to the editor for Archetypes developed at Linköping University (LiU), primarily by the PhD student sponsored by Semantic Mining at Manchester. MoST is a tool for assisting engineers to bind Archetypes to terminologies, specifically SNOMED-CT. It has undergone a preliminary evaluation as part of the PhD process and will be further evaluated beyond the life of the project.



The methodological foundations for this work are contained in the key project deliverables on the ontology engineering and human factors aspects of ontology development, D21.4.

2.2.2 Contribution to overall objectives of NoE

The issue of common ontologies and, particularly, of the binding of ontologies to electronic health records has come to the top of the international health informatics agenda during the course of the project.

There is a serious shortage of informaticians with the necessary skills in ontology development and application to address these issues. The project has provided a major fraction of the tutorials available in Europe on these issues, both through its own work and in cooperation with other EU funded activities.

It has developed new methodologies which have been tested in practice outside the project, and contributed to the requirements and practical implementation of key tools for ontology development.

2.2.3 References to deliverables, quality indicators and milestones

Workshops and Symposia

The workpackage has held a series of workshops and conferences on biomedical ontologies which have been widely attended by both members and non-members of the consortium. Tutorials have been presented by members of the workshop at virtually all major relevant conferences.

It has contributed to the regular summer schools both of the Semantic Mining NoE and to the Reasoning Web summer school sponsored by the REWERSE Network of Excellence.

Sharing of resources and software tools

The workpackage has provided the Protégé-OWL tool and MoST tools to other members of the consortium and in return made extensive use of the Archetype editor developed at Linköping University.

It has provided tutorial material and ontologies widely throughout both the consortium and the larger biomedical community (see 2.2.4).

Deliverables

For a list of deliverables see attached table.

2.2.4 Future opportunities, continued collaboration, new joint research programmes

Members of the Network are participating in at least five funded projects under Framework 7, and further projects are under review. The workpackage leader serves on the Strategic Advisory Board of the US National Center for BioOntologies (NCBO). The tools developed have a wide open source following and are being actively extended by other projects. (See 3.4)

Two ongoing industrial collaborations have grown out of the project and related activities, one with Informatics CIS, a UK SME, one with Siemens Health Systems. (See 3.3)

2.2.5 Conclusions

During the course of the project, ontologies have become established as key infrastructure for interoperability and integration in healthcare and biomedical research. It is a new area, with major needs for training, tools, and resources. Because it is new, the techniques are evolving rapidly, particularly in the area of the interworking of ontologies and health records.



The Semantic Mining Network of Excellence has contributed to the convergence and standardisation, to education, to tools, and methodologies. In combination, WP21 and WP26 have been particularly important in contributing to methodologies for binding ontologies and healthcare records into specifications for systems that are truly interoperable.

2.3 SNOMED CT

2.3.1 Main activities and scientific results

The objective of the WP22 research activity is to share experiences and understanding of the use of large scale reference terminologies in general and SNOMED CT in particular. Moreover, the objective is to encourage sharing of methods and tools for translation, and to co-ordinate the evaluation of SNOMED CT as a reference terminology e.g. for structured data entry applications in different domains.

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

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

SNOMED in relation to application development.

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

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

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

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

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



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

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

Translation of SNOMED CT into Danish is ongoing. Founded on the experiences from the translation form 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.

2.3.2 Contribution to overall objectives of NoE

The interest of the SemanticMining Network in SNOMED CT is rather obvious and the network has both through the joint research plan and outreach activities (e.g. conferences, workshops, summer schools) created a meeting place for system developers, health care professionals, logicians, philosophers, and managers, where different aspects of SNOMED CT such as content coverage, formal structure, quality assurance, multi-language and multi-cultural issues have been dealt with.

2.3.3 References to deliverables, quality indicators and milestones

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. Participation in joint summer schools and organiser of international conference on SNOMED CT in 2006.

Q6: There were a total of 12 short- and medium-term visits between partners.

Q7: There were over 15 research papers with reference to SNOMED CT co-authored.

Q9 Jointly executed research programme

Joint research programme on SNOMED CT together with WP21, WP22, WP23 and WP26.



2.3.4 Future opportunities, continued collaboration, new joint research programmes

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

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

2.3.5 Conclusions

The vision of a universal clinical terminology, covering a broad range of health-related domains and meeting the needs of all health professionals has stimulated numerous health informatics research activities in the last two decades.

During this period, SNOMED grew from a pathology-centered vocabulary to a comprehensive clinical terminology. SNOMED Clinical Terms (CT), is the result of a joint development between the English NHS and the College of American Pathologists (CAP).

There is really no competitor with regard to comprehensiveness and in recent years the interest has grown with in November 2007 eight national governments having decided to join the new international organization International Health Terminology Standardization Development Organisation (IHTSDO), being formed to manage this huge terminology. These are in addition to the founding US and UK, Denmark, Sweden, the Netherlands, Lithuania, Australia and Canada.

Discussions have been intense during 2006 with the WHO and there has been a mutual agreement on co-operation. The SNOMED organization has also been discussing with the European Commission and the member state representatives of the i2010 eHealth group has been discussing the possible role of SNOMED CT for pan European interoperability.

Still there are only few prototypical implementations of SNOMED CT in clinical settings, and there has been some concerns raised about the feasibility of such a comprehensive terminology as the basis for the whole health delivery process. While many appreciate the ambitious approach of the SNOMED CT development, there has also been some voices raised questioning some of the basic design choices and the quality management procedures.



2.4 Health statistics, semantic distance and ontologies

2.4.1 Main activities and scientific results

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. Specific aspects 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 the last years 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 was completed during the spring of 2007 and will result in a scientific paper co-written by the work package team. Lately, a number of papers dealing with issues related to this topic have been published, but none of them have used such an extensive and empirical approach as in this study. We are confident that our study will generate a great deal of interesting topics for further exploration by us and others.

Our conclusion is that there is great interest in the topic of semantic distance and that the efforts of work package 23 will result in an interesting paper that will illuminate the topic of semantic distance. In turn, this will be useful in various applications of information retrieval and statistics.

2.4.2 Contribution to overall objectives of NoE

A bridging activity addressed by this NoE is knowledge transfer and co-operation between universities and public organisations in the health and welfare sector, including standardisation bodies. The national institutes and organisations responsible for policy making and quality management with a regulatory and normative function have had an important role in the network. This mix of researchers and public health care actors have been very fruitful in WP23, where key persons from classification centres responsible for the compilation of health statistics through use of traditional classification systems have met and worked together with researchers bringing new services which ontologies can offer into the joint working plan. Thus, WP23 has contributed to the filling of one important gap between the research community and the public health care system.

2.4.3 References to deliverables, quality indicators and milestones

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

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.

2.4.4 Future opportunities, continued collaboration, new joint research programmes

The new international organization International Health Terminology Standardization Development Organisation (IHTSDO) will offer an arena for several of the SemanticMining partners (NORDCLASS, SOS, LIU, NBH, UKLFR, DIM, INSERM, UOM) with an interest in further research, development and assessment in relation to the use of SNOMED CT as a reference terminology.

2.4.5 Conclusions

An extensive report on challenges in European health statistics has been written (D23.1). A cross-European study on agreement on semantic distance between medical concepts have been undertaken. Cross WP-relations established with WP21 and WP22.

2.5 Text mining in biomedicine

WP24 of the Network of Excellence has been focused to research work information retrieval and information extraction from the scientific literature in the biomedical domain. This led to continued research work on the identification of semantic types from the biomedical literature, on multilingual information retrieval and on retrieval of multimodal type of information (images vs. text). Related research work is concerned with the disambiguation of semantic types, the standardisation of the representation of semantic types in biomedical text and ongoing work in the integration of cross-lingual term normalisation for information retrieval (collaboration with WP20).

Collaborations initiated in WP24 have been extended in other research projects (e.g. projects “BOOTStrep” and “@neurIST”). Furthermore, new collaborations established in 2006 have already led to ongoing projects (e.g. SYMBiotics SSA, Network of Excellence “InfoBioMed”, Network of Excellence “Nutrigenomics”).

Latest developments in the project are concerned with the harmonisation of annotations in the scientific literature. This led into a number of initiatives that harmonise annotation services amongst partners (IeXML). Furthermore, solutions developed from members of the Network of Excellence are reused amongst partners, e.g. GoCAT, Whatizit, MorphoSaurus.

2.5.1 Main activities and scientific results

In ongoing research work, the group has defined new information extraction methods for gene ontology terms. It is in general difficult to identify GO terms in the scientific literature and more research is required to tackle this problem. The newly defined solution considers several parameters: specificity of terms, proximity of terms in the text and similarity of terms in the text and in the ontology. The solution applies an information theory based approach to determine which candidates of gene ontology terms are suitable to characterize a piece of text. The output is a list of candidates that contains all candidates ranked according to the score that measures the similarity between a term and the text. Furthermore, the solution identifies gene candidates and links the gene candidates to the GO annotation.

For the evaluation, we used the corpus of Task 2.2 in the BioCreAtIve I competition. In addition, we also compared our results to the results of the GO annotation from DIM (annotations were performed and provided by DIM). The novel approach outperformed the solution by DIM and another solution available from the University of Lisbon (Francisco Couto). We submitted the publication on the results of this work to a journal of computational biology.

In another experiment, we used the automatic GO annotation from the scientific literature to predict protein functions. We selected “conserved” proteins, i.e. proteins that participate in protein-protein interaction networks from different species. For some of these proteins the molecular function is yet unknown (according to the GOA database). Based on literature analysis, we identified evidence for the molecular function of the before-mentioned proteins. We checked all annotations with the help of a professional curator and could confirm all of them (100% precision). The publication is in the review process.

The annotation of pieces of text with concepts from the gene ontology is a difficult but rewarding task. On the one side, we are still lacking knowledge how professional annotators themselves fulfil this task and what their expectations are with regard to any automatic means that supports their work. On the other side, any successful annotation gives new clues about the specialty of a gene or protein. Since the information in the literature represents secondary information that could be called hypothetical, any inferred annotation as well raises the notion of being a suggestion rather than pure truth.



In our next research work we applied the proposed novel solution for the identification of ontological terms to terminology from Snowmed CT. We compared the results from our identification of medical diseases to the state of the art solution MetaMap. The evaluation is still ongoing.

For any of the before-mentioned solutions, we defined an information extraction module that is now available to the public. The modules are part of the Whatizit infrastructure. In addition, we designed a new information extraction solution for the identification of chemical entities. We selected existing solutions and combined them to achieve this goal. The solution available as part of Whatizit integrates the terminology from the ontology ChEBI, terminologies from the ENZYME database, and OSCAR3, which is the state of the art solution for the identification of chemical entities. Collaborators at the EBI and at the European Patent Office have used this information extraction module to process documents. Similarly a new solution

Further collaborative work between Eric van Mullighan (Erasmus Medical Center, Rotterdam, NoE InfoBioMed) and Olivier Bodenreider (NLM) has lead to the specification of an annotation schema that will be used in the future to exchange automatic annotations in biomedical scientific literature to compare and evaluate competitive and complementary approaches (biomed-MTC, “biomedical multi-tagged corpus”). Furthermore, a sample corpus is now available that is correctly annotated according to the specification. Further work will explore on the challenges arising from the annotation of scientific literature with different techniques and make the results of this research available to the public. The goal of this study is to prepare a common annotation schema that will be used on repositories available from the EBI for public use in the text mining community.

2.5.2 Contribution to overall objectives of NoE

WP24 has contributed to the overall objectives of the NoE “SemanticMining” for the following reasons:

(1) As part of the project work, the members of the WP24 have shown new solutions how to integrate terminological and ontological resources into information extraction and information retrieval approaches. A number of these solutions are available to the public as Web services. The research done in conjunction with these developments has shown the limitations and the needs arising from the ontological resources. The solutions contribute to the understanding how to bridge from ontologies to text mining.

(2) The results from WP24 are relevant to the NoE’s objective aiming at better use and reuse of information for health professionals and patients to achieve overall better health care delivery. The Work done as part of WP24 links bioinformatics data resources with medical data resources and helps to overcome the gap between both domains. The demonstrators available to the public lead to immediate benefit. It is worth stating that the cross-referencing between biological and medical information is another objective of the NoE.

(3) Another objective of the NoE is the automatic support for the cross-referencing of biomedical data resources. The work done in the WP24 on automatic integration of electronic data resources and automatic processing of the literature resources fulfils this demand. Altogether, several solutions have realized this goal, are described in scientific publications and contribute to the understanding of the scientific community into new ways how to better access relevant information.

(4) The initiative “biomed-MTC” is an initiative to standardize the annotations in the biomedical scientific literature and contributes to the objective “standardisation of resources”. It is too early to say whether this initiative will affect the distribution of scientific literature in



the future, but currently it will have an impact on the documents delivered by the UKPMC content at the EBI.

(5) WP24 did not have the task to develop new semantic resources, but WP24 profited from the use of semantic resources for example in the annotation of scientific literature with gene ontology concepts or the annotation of genes/proteins with such concepts. Such approaches contribute to semantic interoperability, since any other data resource can make better reference to the annotated electronic resources. From a more global perspective, the annotations contribute to the formation of a heavily interlinked network of biomedical data resources. The information contained in the data resources and the use of links between these data resources offers health professionals and patients new ways to explore available information. In the best case, the richness of the existing information in conjunction with the means to explore it could serve as the appropriate means to support communication between different professions in the health care environment. As overall result, the infrastructure improves the communication between patients and health-care professionals. This leads to cost effectiveness of the healthcare system, since the health professional spends less time in retrieval of information and the patient can turn to the same public tools and services to improve his understanding of his medical conditions (another objective of the NoE).

(6) The NoE had the objective to mine relations in the data resources. WP24 contributed to this objective by the provision of services that mine relations (e.g., EBIMed, refer to Rebholz-Schuhmann et al., Bioinformatics, 2007).

(7) WP24 delivered IT solutions that provide their share of ubiquitous computing, since the solution bring together information automatically that could not be analysed so easily otherwise.

2.5.3 References to deliverables, quality indicators and milestones

The deliverable 24.1, 24.2 and 24.3 have been provided throughout the project.

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 seminars

- Participation in the ISMB conference 2005, 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 joint summer school in 2006 in collaboration with NoE InfoBioMed.

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

Several co-authored research papers.

2.5.4 Future opportunities, continued collaboration, new joint research programmes

IST funded project “BOOTStrep”¹: EBI and Jena have prepared a grant proposal to the EC’s IST program. The project proposal is called BOOTStrep and is a Strep with 8 partners including EBI, Jena and UKLFR. The project has been accepted by the CEC and started in April 2006.

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

@neurIST Project2: The @neurIST project is a project amongst UKLFR and DIM. It brings together different data resources to support disease management of cerebral aneurysm.

2.5.5 Conclusions

WP24 has contributed to several objectives of the NoE. Certainly, the biggest contribution results from the fact that the members of WP24 were bringing electronic resources from the biological and the medical domain together in such a way that semantic interoperability is achieved for selected use cases. This work will inspire other research teams to propose similar and by the nature of science, more advance solutions. Altogether, the two domains of medical informatics and bioinformatics have grown together a bit.

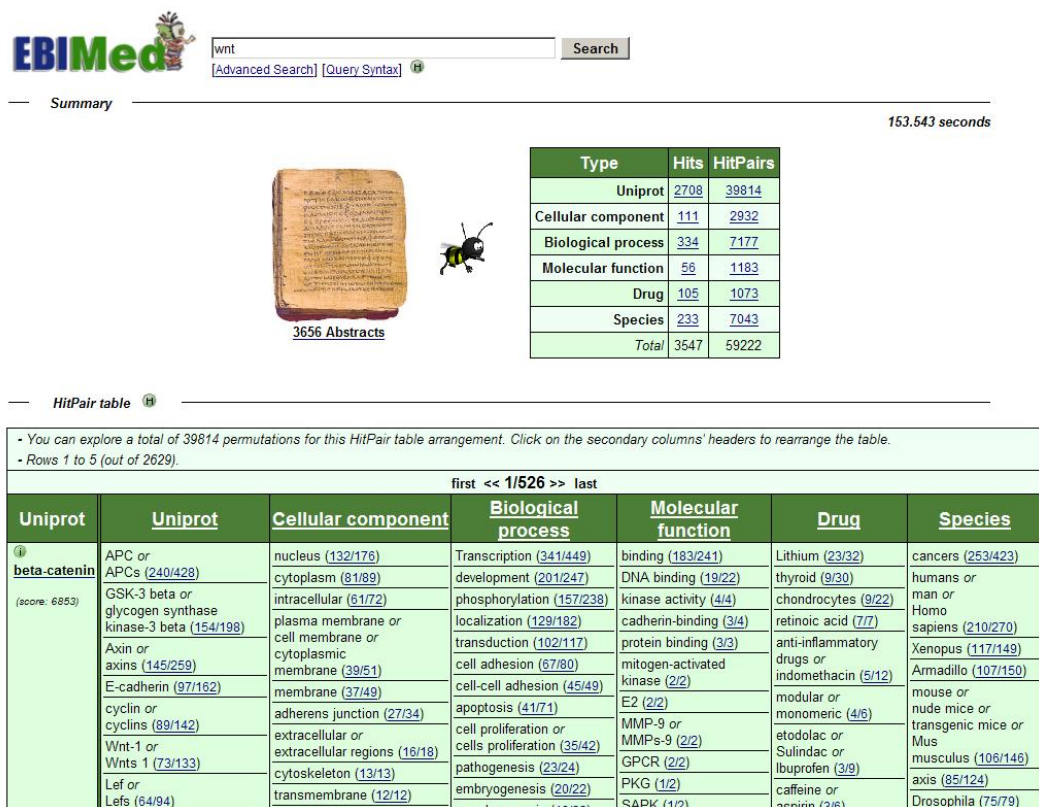


Figure 4. EBIMed summary display of search for abstracts referring to Uniprot proteins.

¹ www.bootstrep.eu

² www.aneurist.org/

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

2.6 Terminology systems in laboratory medicine

2.6.1 Main activities and scientific results

The objective of this work is standardization for the sake of interoperability between medical laboratories and health care provider electronic systems of result of examination in the laboratory.

A main activity of WP 25 has been to consolidate the work done on the prEN 1614 into a permanent European standard. The standard is now published as CEN EN 1614 “Health informatics — Representation of dedicated kinds of property in laboratory medicine”

The wider issue is that of connectivity between medical laboratories and clinical care to ensure that reports for medical diagnosis and treatment between stake holder’s are transmitted correctly. One scope is to provide a standardised way of communicating kinds-of-property in laboratory requests and reports. This is especially important in communication with Electronic Health Care Record systems (EHR systems) While one possible 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. A major break through in our perception on how to achieve this 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. Preliminary work was done in Copenhagen at the Semantic Mining partner Danish National Board of Health during 2006 (Ulla Magdal and others), explored during the Semantic Mining conference in Copenhagen October 2006 and further pursued at the satellite meeting “Workshop on representation of results within SNOMED CT (C-NPU)”. Active collaboration is now established with the International Health Terminology Standards Development Organisation and SNOMED CT concept model working group (Daniel Karlsson member of the group). This work is at the time of writing very active and fruitful.

2.6.2 Contribution to overall objectives of NoE

WP25 has contributed in a major way in that the foundation for connectivity between different language and professional domains in laboratory medicine has been explored and given a European framework (CEN standard).

2.6.3 References to deliverables, quality indicators and milestones

Three deliverables are published (D25.1, D25.2, D25.3). The work package has contributed in several workshops, taken part in the mobility program, and published results in scientific journals and conferences. A major milestone was the publication of CEN EN 1614 “Health informatics — Representation of dedicated kinds of property in laboratory medicine”.

2.6.4 Future opportunities, continued collaboration, new joint research programmes

Continuation of this work is participation in work of the International Health Terminology Standards Development Organisation and SNOMED CT. A potential forthcoming application will be the FP7 proposal DebugIT, aiming at control of adverse events in the domain of infectious diseases.



2.6.5 Conclusions

The foundations for connectivity between different language and professional domains in laboratory medicine has been explored and given a European framework (CEN standard) in that properties examined are rigorously defined.

2.7 The semantically well-defined EHR

2.7.1 Main activities and scientific results

The objectives of WP26 were to define an interoperable means of specifying classes of data within the EHR with sufficient granularity and precision that clinical applications, decision support systems and other tools can create or retrieve data values or sets of patients that precisely match any given clinical criteria. Moreover, the objective was to support the future seamless and standards-based interaction of knowledge, record and inference services.

Each of the partners has undertaken research and development activities as reported in Deliverables 26.1, 26.2 and 26.3. Common themes have been pursued through inter-site collaborations. Difficult challenges have been explored through theoretical research, giving rise to some world-leading contributions to the field. Some of the practical solutions have been embodied as tools, mostly published open source and shared across the sites. New research projects are being fostered (e.g. FP7). Many of the work threads also feed into wider international collaborations and to international (CEN, ISO, HL7) standardisation.

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 EN/ISO 13606, or HL7 CDA Release 2). The original description of this work package implied a goal of semantic indexing the EHR. It became 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 broadened the interpretation of the original WP26 tasks in order to respond to the goals that have emerged, and re-focused on specific challenges that became 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.

The main theoretical aspects of these challenges and outcomes, including formal contributions to international standards, are summarised in this section; tools arising from the work are summarised in the SemanticMining report on tools and services.

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. This work is informing the development of archetype design guidance and educational materials.

Binding EHRs and Archetypes to Terminology and ontology

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 at UCL, and will be published later. This challenge is significant, and much other work in Work package 26 has focussed on the binding of archetypes to SNOMED CT.

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 work package 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 George Pompidou European Hospital (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.

Building of these achievements, the INSERM team made progresses in the following directions.

- Binding of terminology and data structures in the domain of the management of patient antecedents in the EHRs.
- Redesign of templates based on updated knowledge (ontologies). Any structured entry form for clinical records should be kept up to date with regard to physicians’ needs during the clinical encounter and to the evolution of medical knowledge and practice. We updated the computerized medical record form of a hypertension clinic according to the study of its previous use and of clinical guidelines. Statistical analysis of previously completed forms pointed to unnecessary items almost never used by clinicians. Terminological analysis of guidelines and of free text answers in previously completed forms revealed relevant topics for actual clinical practice. Accordingly, new items were added in the structured part of the form and some topics previously recorded as free text were structured to update the form. Collaboration



with clinicians was necessary to interpret the results of statistical and terminological analyses taken as a starting point and guide for the update process

- Designing a virtual EHR, a mirror of the HEGP EHR, based on a standard information model - the HL7 Reference Model – to enhance interoperability and data analysis. The short term objective is to enhance the implementation of hospital quality indicators (QI) defined by the national EHR Quality Insurance project (COMPAQH project).

Methods: We followed a three step methodology a) design of the database of the virtual EHR, based on the Reference Model; b) identification of relevant data required for QI calculation and alignment of these data with the Reference Model; c) definition and automation of feeding procedures of the virtual EHR from the HEGP EHR. We conducted a preliminary evaluation of the architecture of virtual EHR comparing QI extraction from the virtual EHR and from the EHR.

Results: The RIM-based Entity-Relationship (ER) model of the virtual EHR is composed of three tables corresponding to classes inherited from the Act class, one table corresponding to the Participation class, one table corresponding to the Role class and four tables corresponding to classes inherited from the Entity class. The remaining tables correspond to the HL7 data type classes (19 tables) and to the classes of the Common Terminological Services (CTS) dedicated to link terminologies to coded entries of structured documents. The virtual EHR was supplied by the metadata of 181.661 textual documents and 78.394 structured or semi-structured documents produced in 2005. EHR quality indicators (“Presence of hospitalization report” and “Presence of the initial clinical evaluation”) were produced relying on structured entries of the administrative data of these clinical documents.

Conclusion: We showed that it is possible to design a RIM-based ER model for a virtual EHR to feed it from the HEGP EHR and to enhance the production of EHR quality indicators (QI) as defined by the national COMPAQH project. Indeed, even while taking into account only the metadata of the clinical documents, we produced the QI. Our perspectives are to extend the evaluation using clinical structured data in addition to administrative data. We also plan to implement within the CTS solutions to align our local dictionary of concepts to reference terminologies.

- Sharing information between EHR and Clinical Data Management Systems (CDMS). In both areas - patient care and biomedical research- considerable efforts were realized for the standardization and communication of medical information, but in independent ways. In the healthcare domain, the HL7 organization and CEN TC 251 create standards for the exchange, management and integration of electronic clinical information. HL7 has developed over about ten years a version 3 of its messages, usually implemented in XML, the coherence of which is guaranteed by the Reference Information Model (RIM), to be adopted by the ISO. The first real use of this HL7 RIM was the development of a standard format for clinical documents (Clinical Document Architecture (CDA)). In the biomedical research domain, the CDISC (Clinical Data Interchange Standards Consortium) organization defines platform-independent standards that support the electronic acquisition, exchange, submission and archiving of study data and metadata for pharmaceutical companies and Food and Drug Administration (FDA). The CDISC Operational Data Model (ODM) was developed especially for data exchange in clinical trials between different study software. As part of the Cancer Biomedical Informatics Grid (caBIG) project, CDISC initiated the Biomedical Research Integrated Domain Group (BRIDG) to harmonize the semantics from available clinical trials information models into a shared model.



Although there is an issue to connect patient care and biomedical research, nowadays, in practice, biomedical research applications often require specific manual re-entry of data, some of which already reside in the EHR. Our objective consisted in implementing a solution to share data between EHR and CDMS.

Material and methods: Nowadays, in the G.Pompidou Hospital (HEGP), radiologists enter clinical data of arteriographies through a form of the EHR (DxCare®, Medasys©) and then fill another form of a national CDMS for clinical research in cardiovascular radiology (AIR On Line (AOL)®, Kika Medical©). We used a two steps methodology to implement a solution to capture data for both patient care and research in the cardiovascular radiology domain.

1) Designing a multi purposes “care-research” form

Items required for patient care and for clinical research were analyzed and compared. A single multi purpose form was designed including all the items required for cardiovascular radiology. Since we could only make changes in the EHR forms, to avoid semantic interoperability issues the items that were required in both contexts (patient care and research) were represented according to the research context. We took into account the data types, the modalities of answer and the data type controls of the CDMS for clinical research. The multi purposes form was implemented on the one hand in DxCare® (“Care-research” CV form (a)) and on the other hand using the standard XForms (“Care-research” CV form (b)).

2) Exporting data from EHR to clinical data management system (CDMS)

We analyzed the differences between the HL7 CDA and CDISC ODM models from both structural and organizational points of view (with regard to the context of data capture). We implemented two approaches to export data from EHR to clinical data management system (CDMS):

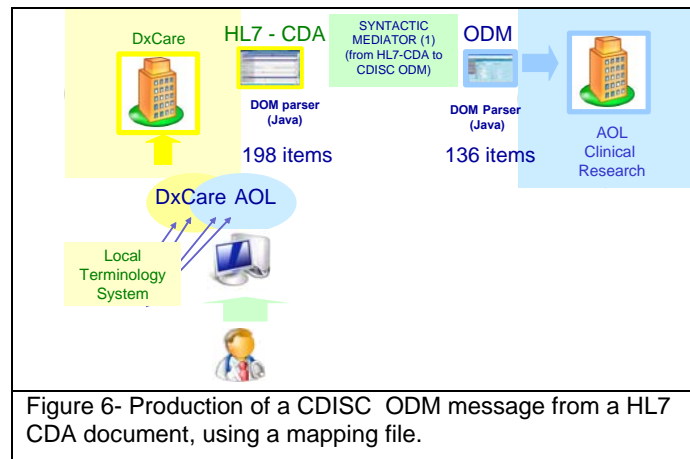
The first approach consists in generating for each instance of the “Care-research” CV form (a) implemented in DxCare® the corresponding HL7 CDA document and aligning XML structures of HL7 CDA document and CDISC ODM messages to export relevant data from EHR to CDMS (figure 6).

The second approach, based on the IHE integration profile RFD, consists in displaying the “Care-research” CV form (2) implemented using XForms technology within the EHR and generating both HL7 CDA document and CDISC ODM message to fill directly both EHR and CDMS (figure 7).

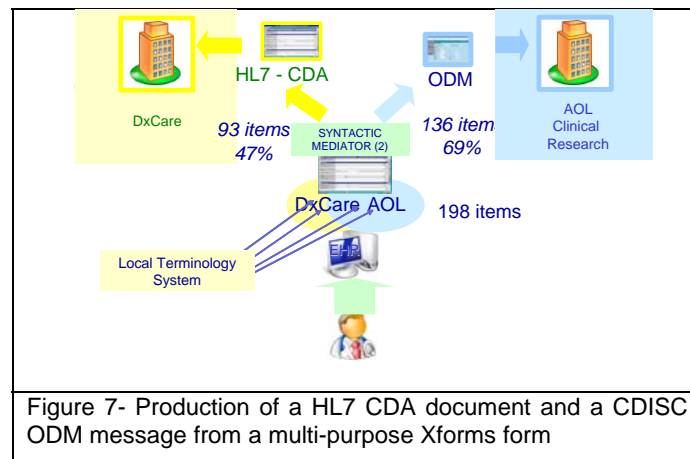
Results: The comparative analysis of clinical data captured in patient care and clinical research showed that, in the domain of cardio-vascular radiology, 93 clinical items are relevant for patient care and 136 for clinical research. The mixed “Care-Research” CV form contains 198 items.

The comparative analysis of both standards HL7 CDA and CDISC ODM showed that some differences are due to different contexts of data capture in patient care and clinical research. Tools were developed to implement two approaches to export data from EHR to clinical data management system (CDMS).

The tool developed according to the first approach uses the mixed “Care-Research” CV form implemented in DxCare (198 items) and produces a CDISC ODM message including 136 items for biomedical research (figure 6).



The tool developed according to the RFD profile produces an HL7 CDA document including 93 items dedicated to patient care and a CDISC ODM message including 136 items dedicated to biomedical research from the same single “Care-Research” implemented in XForms (198 items) (figure 7).



Conclusion : We showed that it was possible to exploit medical information of the EHR within the framework of clinical research. We used a method of alignment between XML formats of the clinical documents (HL7/CDA) model and the clinical research (CDISC/ODM) model. We also tested an alternative approach based on generation of both HL7/CDA documents and CDISC/ODM messages from a single XForms form. Tools implementing these two approaches were developed and tested. The benefit for healthcare providers consists in avoiding time consuming and error prone double data entry and allowing automatic transfer of administrative and medical data relevant for biomedical research towards a national server. Moreover, since relevant clinical data are stored in a structured manner in the EHR, they remain available for patient care and are both readable by healthcare professionals and exploitable by computers for decision support and patient safety (alerts, etc).

Perspectives: We are working with Thales-Medasys© and TelemedecineTechnologies©, a company developing CDMS for clinical research to implement this workflow in HEGP. Since our work only makes it possible to ensure syntactic interoperability between two applications (DxCare® and CDMS), we plan



to take into account semantic interoperability, in order to ensure that these two applications have a common understanding of the exchanged data even when these data are expressed differently in both applications. This future experiment will be conducted making use of the results of the BRIDG efforts. We will focus on implementing solutions for semantic harmonization between various source models and/or reference terminologies such as MedDRA used for the designation of the medical devices, the adverse effects, LOINC used for the laboratory tests, ICD-10 or SNOMED-CT used for the procedures and diagnoses. Comparing manually semantically connected domain models is a tedious task. Defining the efficient use of assistance for model alignment such as receiving recommendations of merging options is a challenging issue.

- Further work has started to explore the potentialities of these semantic interoperability approaches in the domain of vigilances and patient safety.

Research on methods and tools for terminology binding in archetypes have been performed in cooperation between Linköping and Manchester University, Karolinska Institutet and UCL. A 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. The archetype editor has support for manual or semi-automatic creation of bindings between archetypes and terminology systems. Lexical and semantic methods are applied in order to obtain automatic mapping suggestions. Information visualisation methods are also used to assist the user in exploration and selection of mappings. The methods and tools are general, but only bindings between SNOMED CT and archetypes based on the *openEHR* reference model have been reported in detail.

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.

Contributions to EHR formal specifications and standards

UCL has a fifteen year history of R&D in EHR requirements, information architectures, and implementation, largely funded through successive EU Health Telematics programmes. UCL is a founder and sponsor of the *openEHR* Foundation, alongside Ocean Informatics, and has in recent years also led work on the development of a European (CEN) and in International Standard (ISO) for EHR Communications (ISO/EN 13606), some parts of which have recently been published and others of which are now close to final publication. The *openEHR* Foundation holds the primary IP for the EHR Archetype approach and its formalisation, which originates from over ten years of research and clinical demonstrators in Europe and Australia. This work has been contributed into the international standards arena (as Part 2 of the 13606 standard).

Over the past year 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.



There is a strong synergy between the research activities on the semantic content of EHRs, formalisms to define semantic structures within EHRs (Archetypes) and the need to define and facilitate good standards in this area. It has been of great value that pioneers in the field, such as Ocean Informatics and UCL, expert research groups such as the partners in WP26, and the developers of EHR standards, have had close working collaborations during the lifetime of Semantic Mining. Indeed, some persons are representative of all three categories.

Specific contributions to standardisation have come from all members of Work package 26, by:

- direct participation in the development of standards through membership of Project Teams and attendance at standards meetings, which have been listed in each Semantic Mining annual report;
- reviewing and performing formal evaluations of draft specifications, for example by building tools and proof-of-concept implementations, outlined in the tools section of this report and in more detail within previous WP26 deliverables;
- conducting research and exploring more complex semantic representation topics that have informed the standards, and also helped to define limitations to what can be considered fit to standardise at this time.

The openEHR foundation has developed an innovative design for interoperable and future-proof Electronic Health Record (EHR) systems based on a dual model approach with a stable reference information model complemented by archetypes for specific clinical purposes.

As a cooperation between WP26 partners (Karolinska Institute, Linköping University, UCL) all the stable specifications have been implemented in the Java programming language. The implementation was adopted as the openEHR Java Reference Implementation in March 2005 and released under open source licenses. This encourages early EHR implementation projects around the world and a number of groups have already started to use this code.

The early Java implementation experience has also led to the publication of the openEHR Java Implementation Technology Specification. A number of design changes to the specifications and important minor corrections have been directly initiated by the implementation project over the last two years. The Java Implementation has been important for the validation and improvement of the openEHR design specifications and provides building blocks for future EHR systems.

The software from the Java Reference Implementation project is released under open source licenses, namely Mozilla Public License (MPL), Gnu Public License (GPL) and Less Gnu Public License (LGPL). The users can choose any one of these three that suits them best. It is worth to mention that among these three, MPL is the least restrictive and GPL is the most prohibitive for commercial use. This means that the even commercial suppliers can use this software as part of their offerings without any commitment to *openEHR*. The software is freely accessible through Internet and all changes are public and accessible through the source repository.

Although contributions from Semantic Mining have informed a broad spectrum of EHR standards and specifications, the particular focus of much the WP26 research has been on *EHR Archetypes*.

Both the *openEHR* Archetype approach and its adoption into a European Standard have occurred during the Semantic Mining NoE project, and the international acceptance of EHR Archetypes has benefited significantly from Work package 26 research outputs.



Overview of the openEHR / 13606 Archetype approach

Unlike fixed-content healthcare messages that have hitherto been defined to support patient registration, claims, episode and Health Resource Group returns, population screening and disease registers, the support of clinical shared care and longitudinal care require the communication of fine grained and diverse clinical data structures within a coherent and harmonised framework.

The first step in EHR interoperability has been to define a generic framework, often called an EHR Reference Model. Considerable experience now exists of the ideal characteristics and requirements for such a model. The most comprehensive such model has been published by the *openEHR* Foundation. Recent standards from CEN and ISO (ISO/EN 13606) include a simpler interoperability Reference Model to enable EHR data to be communicated internationally between heterogeneous EHR systems in ways that preserve their internal structure and medico-legal provenance properties.

However, by being generic such models are deliberately devoid of clinical domain (clinical semantic) concepts. Whilst permitting flexible and faithful representation of the underlying clinical data, the use of a generic Reference Model alone risks the adoption of heterogeneous approaches to how particular clinical data structures are represented. An EHR Reference Model therefore enables faithful EHR communication but only plays a limited role in supporting semantic interoperability. What is additionally needed is a way to specify how particular aspects of clinical (EHR) data are to be structured, and represented using the Reference Model, such that different clinical teams in different settings, using different clinical applications and EHR systems, can share and richly (computably) interpret each others EHR entries.

Clinical data structures are presently specified in a variety of formats, on paper and electronically, such as forms and templates, clinical guidelines, standardised data sets and message definitions. None of these formats are interoperable, and many of these specifications are insufficiently rigorous to permit EHR instances to be interoperably specified.

EHR Archetypes provide a formal, rigorous, and interoperable specification of an agreed clinical data structure. Libraries of EHR Archetypes may be built up from accumulated best practice, as a means of systematising the way that data is organised within EHRs, or as a mirror of the de facto data structures that are already being stored and/or shared. It is now widely accepted that EHR Archetypes, when used to constrain a generic EHR Reference Model such as that defined by *openEHR* or by ISO/EN13606, provide a first layer of semantic coherence and thereby offer a foundation on which fine grained semantic interoperability can be built.

Contributions to the EHR Archetype Specifications

During the Semantic Mining project lifetime significant revisions have been made to the *openEHR* and ISO/EN 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 Object Model (AOM) has proved to be a solid formalism for building tools. A serialised form, Archetype Definition Language (ADL), has been improved in a number of 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>



- updated list of units of measurements to make a more complete list available.

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.

EN 13606 Part 2 has now passed its Final Vote, and is anticipated to be a published European Standard by autumn 2007. ISO 13606 Part 2 is at the stage of a New Work Item proposal, with the European draft to be balloted as an ISO Draft International Standard (DIS).

Over the past year 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.

The DCM Group has agreed to begin its work using the *openEHR* Archetype specification and tools, and will also review ongoing work on the OWL representation as a means to validate Archetype content and as a means to compare and index Archetypes. Semantic Mining partners from UCL and Manchester are members of the DCM, together with representatives from the NHS Connecting For Health and several US organisations including vendors.

HL7 is supportive of the DCM project, and is exploring how the EHR Archetype approach can best dovetail with its emerging (and more technical) Template formalism. Semantic Mining partners have regularly contributed to HL7, and in particular its Templates group.

Tools and services

The ADL workbench is an open source software tool published by the *openEHR* Foundation and developed by Ocean Informatics. It contains an ADL parser written in Eiffel. It has been radically improved through Semantic Mining, and is capable of individual archetype analysis as well as automated testing of an entire repository of archetypes. The ADL Workbench remains the reference tool for the ADL language and is widely used. New features include path analysis, which extracts sets of paths for use in archetype-based querying.

The Ocean Archetype editor is also an open source tool published by the *openEHR* Foundation. It uses the same Eiffel parser, in a .Net build, as the ADL workbench. The main functional differences between the Archetype Editor and the Workbench are:

- the Editor allows editing of archetypes, the Workbench only allows viewing and testing ;
- the Editor is targeted to the *openEHR* reference model, i.e. it knows all the semantics of the published *openEHR* reference model, which it uses to ensure archetypes it builds are *openEHR*-compliant.

Numerous improvements have been made to the editor over the last 18 months, many in response to errors or problems reported by WP26 users. The Ocean Archetype Editor is the main tool in use around the world by clinical people building archetypes, including in the US-based Detailed Clinical Modelling (DCM) activity and in the NHS.

In parallel, the Karolinska Institutet and Linköping University 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. Through a co-operation between Manchester and Linköping University, the archetype editor has support for manual or semi-automatic creation of bindings between archetypes and terminology systems. Lexical and semantic methods are applied in order to obtain automatic



mapping suggestions. Information visualisation methods are also used to assist the user in exploration and selection of mappings. The methods and tools are general, but only bindings between SNOMED CT and archetypes based on the openEHR reference model have been reported in detail.

Ocean Informatics has developed a tool for building templates based on *openEHR* archetypes. A template is a structure of archetypes that can be used for building a top-level piece of content, such as an *openEHR* Composition. This tool allows the definition of the logical template, and then the construction of screen forms based on it. A template tool and a SNOMED-CT server and query builder have been built by Ocean Informatics.

The last 18 months have seen significant improvements to both the methodology and the above 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 has 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.

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 Linköping group in 2006 (whose editor was outlined above). 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 R&D, now almost operational, 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.

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 EHRBank 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.

2.7.2 Contribution to overall objectives of NoE

It is now recognised that, since clinical care is more distributed between multiple healthcare organisations and the quality management of long-term illnesses requires access to long-term records, there is a need to integrate EHR data right across a national health system (and more widely internationally). Such integration needs to be robustly computable, and not just human readable, in order for guidelines, care pathways, alerting and decision support components to function effectively and safely across EHRs that have been combined from heterogeneous systems. This goal is now termed “semantic interoperability” and new EU Strategic Support Action, Semantic Health, is helping to develop an EU Roadmap of challenges and approaches towards realising this goal. Some members of this work package are also members of that SSA, and the investigations performed through WP26 of Semantic Mining have already made an invaluable contribution towards that Roadmap.

During this project it has become clear that the greatest areas of challenge in semantic interoperability for the EHR lie firstly in the definition and sharing of suitable data structure definitions (known as EHR Archetypes) and secondly in the binding of record structure nodes (effectively, Archetype nodes) to terminology. This latter challenge is particularly complex if co-ordinated terminology systems are used, of which the prime contemporary example is SNOMED-CT.

Many of the issues tackled in WP26 have been 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 clear that the work done by the WP26 partners is recognised internationally as a set of strong contributions. From the starting point of a loosely coupled set of university sites, the WP26 team gained momentum as an affiliated set of experts conducting world-leading research. New EU and other funded projects are starting to enable specific aspects of this challenge to be addressed to a greater depth. The range of publications, open source tools and proof-of-concept implementations is impressive. A key contribution from Work package 26 has been to international EHR standards, enabling these to be better underpinned by research evidence and cutting-edge understandings achieved by Semantic Mining.

It is undoubtedly recognised that these European University teams are now world leaders in this field.

2.7.3 References to deliverables, quality indicators and milestones

Workshops and symposia

WP26 has held internal workshops once or twice per year which have been very well attended by partner teams. These have provided a forum to share research results, identify common threads and to actually undertake some elements of the research. Partners have represented

the NoE and/or its research threads at many international workshops, symposia and conferences over the past four years, listed in successive annual reports. Overall, the Semantic Mining work and expertise has attracted a high profile through peer reviewed submissions and invited presentations. The innovative and complex nature of the work is such that many more publications are expected from the work after the NoE has officially ended.

Sharing of resources and use of research software tools

As described in Section 2.2, several tools and pre-products have been developed during the project. These have all benefited from multi-site contributions, some have had concrete multi-site software contributions, and most have been used by more than one site.

Co-tutoring of PhD-students

It has not been possible for formal supervisory roles to be conferred on senior staff from other universities, but an informal cross-supervision has occurred for a couple of WP26-related PhDs.

Co-authoring of research papers, reports and educational materials

Many papers have been published as joint author papers across the WP26 partners, listed in various annual reports. Many more are in draft and will be published during 2007-2008.

Participation in standardisation activities

This major contribution has been documented earlier in this report.

Jointly executed research programmes

During Semantic Mining, some of the UK partners were also in collaboration on UK eScience research projects, CLEF and CLEF-Services. Two WP26 partners are part of the SemanticHealth roadmap project, funded through FP6, as a result of their collaboration and reputation enhanced through Semantic Mining.

2.7.4 Future opportunities, continued collaboration, new joint research programmes

Several WP26 partners are collaborators in an FP7 IP proposal (DEBUGIT) that was successfully short listed for a hearing.

The tools that have been developed during Semantic Mining are still being refined, and the partners will continue to collaborate through the *openEHR* Foundation, to which many of the tools have been granted the distribution licence by the authors.

Some of the research threads will continue through unfunded collaborations until new funded opportunities can be found.

2.7.5 Conclusions

Many of the issues tackled in WP26 have been 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 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.

2.8 Laymen terminology

2.8.1 Main activities and scientific results

The objectives of WP27 were

- (1) to reach a deeper understanding of the differences between the two types of text involved (“professional” and “popular” medical text), on the basis of empirical investigations of corpora collected or created for this purpose;
- (2) to formulate requirements on non-expert text in a general, if possible language-independent way;
- (3) to explore different methods of generating patient friendly versions of medical information, including multilingual text and multimodal presentations;
- (4) to integrate educational materials in a standardized EHR repository architecture.

The main objectives of WP27 have been attained, to varying degrees: Very good progress has been made on objectives (1) and (2), which were the subject of the two WP deliverables. With objectives (3) and (4), our work has by necessity been more preliminary and exploratory. It has however allowed us to glimpse the outlines of a research program which we would like to develop further.

Literature review on patient-friendly documentation systems

The OU team coordinated this deliverable and contributed a substantial amount of the written content. The review was completed successfully. The outcome was a document published as an Open University Computing Science Department Technical Report.

Corpus study

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.

At UGOT, collection is ongoing of three Swedish (sub)corpora from one and 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.

At INSERM, web corpora of Japanese and Russian expert and non-expert text were investigated with respect to a number of linguistic variables.

These three studies formed the basis of the second deliverable of WP27, a report entitled “Empowering the patient with language technology” (see below).

Later, the UGOT team made a study of cancer-domain subcorpora of the Swedish medical text corpus, correlating the findings with those of the OU team. The results of the new contrastive corpus study have been presented in a joint conference publication (see section 3.2).

Anonymization and pseudonymization of medical text

A prerequisite for accessing genuine patient records for research purposes is at least that the records be de-identified. The explicit identifying personal data are straightforward to remove, but the free-text portions of the records (which are potential targets for language technology based tools helping the patient to understand the expert language in the records) can still contain names and other identifying expressions, which will need to be removed or replaced.



The UGOT and SU teams have been doing experiments on adapting a general Swedish named entity recognition (NER) system to the purpose of anonymizing/ pseudonymizing the free text portions of Swedish patient records. This work has so far resulted in two joint publications (see section 3.2).

Requirements analysis: Empowering the patient with language technology

The UGOT team coordinated this deliverable, to which OU, INSERM and UGOT contributed the underlying research, in the form of empirical corpus studies of English, Japanese, Russian and Swedish expert and non-expert medical texts. The resulting report correlates the empirical findings of the per-language studies, in order to capture in general terms the differences between the two types of medical texts. Some useful generalizations could be made and conclusions drawn on the basis of those generalizations on how language technology could be utilized in order to create patient-friendly medical documentation systems. The report was delivered on time, and is now being reworked into a journal submission, which will incorporate the results of the second corpus study as well..

Multilingual simple patient report generator

Louise Deléger (INSERM) and Dana Dannélls (UGOT) have developed proof-of-concept French and Swedish versions of the OU CLEF simple patient report generator. This application generates a simple text summary of (simulated) medical case histories contained in an SQL database. The database information is derived from and isomorphic to a description logic representation of the information. This work has resulted in a joint conference presentation (see section 3.2).

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.

EHR server redevelopment and linkage of Archtypes to educational resources

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 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 has been developed to provide a uniform and authorisation-managed access to the anticoagulant and other web applications for cardiovascular care. This will be used 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 the detailed design or implementation could not be scoped within the time frame of WP27.

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.

A PhD student (not funded on Semantic Mining) has commenced to develop a formal methodology whereby selected patients will be trained to manage their own anticoagulation therapy with the aid of near-patient testing equipment and access to a web based application developed by UCL to provide decision support in dosing and monitoring frequency. An extension to this project will be to provide some of the necessary educational resources on-line by direct linkage from relevant parts of the anticoagulation web application. The key to this is the relevance of the chosen material to the user's context.

The goal is for this is to be achieved using the archetype nodes to which each portion of a data entry or review screen correspond, and linking this to an indexed data base of educational web pages.

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.

A high-level architectural approach has been agreed whereby archetype nodes may reference educational tag names, and a UCL template editor has been modified to permit such tag names to be stored within the archetype service. A completely new archetype tree has been authored for a next-generation anticoagulant application, with greater coherence of archetype nodes to support linking to educational materials. Much of the software engineering to redesign the services (EHR and decision support) needed by the web application has been accomplished, including the design of an application generator that can permit rapid development of web screens is in progress. A security policy architecture has been implemented, that will permit patients to log into the system via a web portal and gain access only to their own anticoagulant record.

This is proposed to be a real implementation of a mechanism whereby patient-friendly educational information can be accessed from a live anticoagulant system - possibly limited to 1-2 screens within the overall application. This will be tested by a small group of patients in the north London area. It now appears likely that this live pilot will be feasible, but the practicalities of this mean that the systems and the patient selection and training not yet

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 Framework 7 proposal.

2.8.2 Contribution to overall objectives of NoE

Recent years have seen a growing interest in the application of language technology in the domains of biology and medicine, and this trend is reflected in several of the SemanticMining NoE work packages. This research, while yielding many useful results in all the involved disciplines, has focused on the needs of researchers in these fields, however, and in medicine, also on medical professionals (and most strands of research in the SemanticMining NoE form no exception in this regard), to the exclusion of the interested lay public. In medicine this category notably includes those on the receiving end of medical care, who have a legitimate and strong interest in informing themselves on all issues involved in their treatment. 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 which can help patients without medical knowledge to access relevant information in the health records.

In the same way that language technology and document processing such as information retrieval and extraction can help medical researchers and professionals locate and process the information they need, patients, too, could be helped by these technologies. In the case of the latter, however, domain knowledge cannot be assumed, so that information sources and connections between them which suffice for the medical professional will not be of immediate help to the layperson/patient. One significant difference between the medical researcher/professional and the patient, at least in many parts of the world, and certainly in Europe, is that of “working language”; the bulk of all (leading-edge) medical research is published in English, whereas patient records in most countries will be in some other language than English. Thus, multilinguality enters the picture in a way which it does not in systems aimed at professionals and researchers.

WP27 has, for its relatively short duration, successfully started to address these issues, specifically by laying some of the theoretical ground, with the literature review and empirical corpus study deliverables, for concrete technical applications. Work on some of the latter also started in WP27, but on a modest scale, and there is much more to be done in that area.

2.8.3 References to deliverables, quality indicators and milestones

Assessing WP27 activities by the quality indicators, we find (only applicable indicators shown):

Q1: Workshops and symposiums. Because of the short duration of WP27 (it started in the third year of the SemanticMining network), we have organized only internal meetings. WP27 has been represented at the summer schools, including doctoral consortia, in 2006 and 2007. WP27 work has been presented at a number of international conferences and workshops.

Q2: The sharing of resources has been central to WP27, seen mainly in the adaptation of OU's patient summary generation system to French and Swedish. .

Q6: There were a total of 6 short- and medium-term visits between WP27 partners

Q7: There were a total of 6 research papers co-authored by WP27 partners.

Q9: Jointly executed research programmes. (1) The corpus study and; (2) cross-language adaptation of the patient summary generation system, constitute the embryo of a jointly



executed research programme, which shows promise for the future, provided that suitable funding can be found.

2.8.4 Future opportunities, continued collaboration, new joint research programmes

At the time of writing, few concrete plans for continued collaboration between WP27 partners have been formulated. However, there is interest to pursue further some of the strands of inquiry that were initiated during WP27, provided that suitable funding can be ascertained.

UGOT and SU will be continuing their collaboration on patient record anonymization and pseudonymization.

The methodology of generating text in more than one natural language from an underlying description logic representation is promising and potentially applicable in many areas, not only medicine, especially since description logic (in the form of OWL) is the knowledge representation formalism of choice for the Semantic Web. In particular, the UGOT team, jointly with the Language Technology Group at Chalmers University of Technology, Göteborg, have started investigating cultural heritage as a possible application area.

2.8.5 Conclusions

WP27 has fulfilled its goals in terms of joint research, resource creation and dissemination. There are clear indications that WP27 has met its higher-level goals of integration and cross-fertilization.

2.9 Mobility program

WP6 Mobility Programs contributed to the European Research Area through the compilation of descriptions of PhD work and dissertations, the management of a visit grants scheme and the organisation of doctoral consortia within the NoE summer schools. Below are detailed descriptions of the three activities.

Descriptions of PhD work and dissertations

The first objective of WP6 *Mobility Programs* was to compile descriptions of PhD dissertations and PhD thesis work for presentation to the NoE. Standards and procedures with respect to the production of PhD dissertations were to be summarised and compared. The compilation was to be used as a basis for a possible process towards harmonisation of procedures and for sharing of senior research staff as co-tutors in doctoral programs.

To achieve this objective, an online enquiry of all NoE participants regarding dissertations was launched in July 2004 and all semantic mining participants (especially the WP Leaders and Administrators) were invited to encourage the students to fill out the questionnaire. Students were requested to provide a brief description of their PhD project and a set of relevant key words. The interface to the questionnaire allowed students to update their information whenever necessary (see D6.1 and D6.2).

The information gathered through the questionnaire was made available online to all participants in the shared area of the Semantic Mining MERMIG platform (document manager > Top/ Mobility Programme–WP6). Out of 31 in the network, 14 students provided the requested information: the response rate was 45%, so we asked Semantic Mining coordinators to encourage the students to fill out the form (see D6.3 for details).

Through this survey, it emerged that a number of students would have been interested in visiting other participants in the Network and benefiting from external supervision or expertise (see D6.3 for details).

In addition to the questionnaire, a compilation was made of abstracts of contributions to the PhD student poster session at the NoE's 2004 summer school, between the end of June and the beginning of July. The information was originally made available in the shared area of the Semantic Mining MERMIG platform (document manager > Top/ Summer School) and it can also be found at <http://mcs.open.ac.uk/semantic-mining/>.

In total, 19 students presented their work at the poster session. Informal feedback from students was positive and showed that the opportunity to interact with both other students in the Network and Researchers was appreciated (see D6.1).

Overall, the monitoring of students' research work had a positive outcome, as it allowed us to identify students' needs such as visiting relevant partner institutions and ways of supporting their work by offering them the opportunity to travel to other universities and participate in doctoral workshops.

Visit Grant Scheme

The second objective of WP6 Mobility Programs was to identify opportunities for exchange of researchers and PhD students between partners in the NoE, and to plan and organise these as short-, medium- and long-term visits. Specifically, the objective was to keep track of, stimulate and support mobility of and interaction of PhD students and researchers across the network.

To achieve this objective, consistently with the interest in opportunities for exchange expressed by students through the online survey, in 2005 we started a visit grant scheme, to allow the interaction of students with senior researchers across the network. Grants up to €1,500 to visit another university within the NoE could be awarded by the Board upon receiving recommendations from the WP6 administration team at The Open University. To receive a grant, a student had to submit an official request including a motivation letter, with the approval of their supervisor and their hosting researcher (see Appendix). The scheme was advertised via the semantic mining mailing list and on the WP6 webpage.

In 2005, €20,000 was budgeted for the scheme. Thirteen grants were awarded for a total of €16,827, as shown in Table 1 (for more details, see Annual Report 2005).

Name	Visit Dates	Duration	From	To	Euro
Imad Tbahriti	01.05-31.07.05	3months	University of Genève	EMBL-EBI	1500
Louise Deleger	17.08.05-25.08.05	1 week	INSERM UMR_S 729	Linkoping	1500
Mikael Nystrom	09.05	1 week	University of Linkoping	Freiburg	1500
Karim Nashar	01.08.05-31.08.05	1 month	University of Manchester	EMBL-EBI	1200
Mikel Egana	05.09.05-30.09.05	1 month	University of Manchester	EMBL-EBI	1200
Anna-Karin Hermansson	06.06.05-10.06.05	5 nights	Göteborgs University	Freiburg	1390
Gaston Burek	04.07.05-08.07.05				
Alexander G. Castro	26.09.05-14.10.05	3.5 week	Open University	INSERM, Fr	1500
	01.09.05-31.09.05	1 month	EMBL-EBI	Manchester	1200
Philip Daumke	10.05	2 weeks	University of Freiburg	EMBL-EBI + Jena	1475
Michael Poprat	10.05	1 week	University of Jena	EMBL-EBI	1000
Eric Sundvall	10/11.05	1 week	University of Linkoping	CHIME, UCL	1342
Vincent Claveau	23.07.05-30.07.05	1 week	INSERM UMR_S 729	Freiburg	1000
Gaston Burek	11.05	2 weeks	Open University	CS, Manchester	1020
Total awarded					16827

Table 1. Grants awarded in 2005, by the visits grant scheme

For 2006, the decision was made to allow both students and researchers to apply for grants. This decision was taken after discussion at the fifth assembly meeting in December 2005, where it emerged that there was also a need for support to (senior) researchers for visits to network institutions in order to strengthen collaborative links. A budget of €15,000 was established for the scheme in 2006. Like in 2005, the scheme was advertised via the semantic mining mailing list and on the WP6 webpage, as well as at the Semantic Mining Summer School. Ten grants, of which five to students and five to researchers, were awarded for a total of €12,840, as shown in Table 2 (for more details, see Annual Report 2006).

Given its success also in the second year, the same scheme was continued in 2007 with the extension of the semantic mining project till the 30th of June. For 2007, a further €10,000 was budgeted for the scheme. In total six grants were awarded for a sum value of €4,971, as detailed in Table 3.

Name	Dates	Duration	From	To	Euro
Felix Balzer (student)	06.03.06-07.03.06	2 days	University of Freiburg	University of Jena	220
Audrey Baneyx (student)	02.04.06-10.04.06	9 days	INSERM UMR_S 729	CNR-ISTC	960
Rahil Qamar	08.05.06-12.05.06 29.05.06-02.06.06	10 days	University of Manchester	Linkoping University	1480
Daniel Schober	03.07.06-14.07.06	12 days	EMBL-EBI Cambridge	IFOMIS University of Saarland	1400
Olivier Steichen (student)	01.11.06-07.11.06	7 days	INSERM UMR_S 729	Linkoping University	1400
Julie Nies (student)	01.11.06-07.11.06	7 days	INSERM UMR_S 729	Linkoping University	1400
Ines Jilani (student)	30.10.06-05.11.06	6 days	INSERM UMR_S 729	Univeristy of Genève	1500
Marie-Christine Jaulent	30.10.06-05.11.06	6 days	INSERM UMR_S 729	Univeristy of Genève	1500
Irena Spasic	6.11.06-17.11.06	12 days	University of Manchester	EMBL-EBI	1480
Natalia Grabar	6.11.06-15.11.06	10 days	Univeristy of Genève	Open University	1500
Total awarded					12840

Table 2. Grants awarded in 2006 under the visit grants scheme

Name	Dates	Duration	From	Hosting Institution	Euro
Lars Borin	05.02.07-06.02.07	2 days	Göteborgs University	Open University	864
Dimitrios Kokkinakis	05.02.07-06.02.07	2 days	Göteborgs University	Open University	864
Maria Toporowska Gronostaj	05.02.07-06.02.07	2 days	Göteborgs University	Open University	864
Dana Dannélls	05.02.07-06.02.07	2 days	Göteborgs University	Open University	570
Attila Nagy	05.02.07-09.02.07	5 days	Debrecen University	CNR-LAO	790
Sandra Williams	19.03.07-20.03.07	2 days	Open University	INSERM UMR_S 729	839
Total awarded					4791

Table 3. Grants awarded in 2007 under the visit grants scheme

Over the three and a half years of WP6, we have been keeping a record of the mobility activities of PhD students in the network through the visit grants scheme. This information has been made accessible via the MERMIG platform to all Network participants as well as via the WP6 Mobility Programs webpage.

Overall, the WP6 visit grants scheme was very successful and stimulated a lot of new research collaborations among network partners involving PhD students, junior researchers and senior researchers at host institutions. The visits involved a total of 31 students and researchers, and 12 institutions across 6 European countries. The collaborations between visiting students/researchers and their hosts have produced good results, demonstrated by numerous publications and software developments. In particular, 2 posters, 2 workshop papers, 8 conference papers (1 to be submitted), 1 journal paper (to be submitted) and 5 software developments were produced as a result of the visits (for more details, see sections 3.2 of this report).

Doctoral Consortium at Summer School

A third activity (and indeed objective) emerged as a result of the student poster session held at the summer school in 2004. In line with both the first and the second objectives of the WP6 Mobility Programs, we organised a doctoral workshop, to be held during the NoE summer school in 2005, 2006 and 2007. The objective of the doctoral consortium was to support students in their research programmes following homogeneous criteria.



In 2005, the doctoral consortium was held in Tihany, Hungary, on June 30th and July 1st. A booklet with abstracts of the student presentations was compiled and made available on MERMIG. Materials of the workshop and the booklet were also made available at <http://mcs.open.ac.uk/semantic-mining/>. The doctoral consortium attracted a good number of students (21) who actively participated in various consortium activities (presentation of papers, rehearsal of brief research topic introductions, discussion groups on research methods, etc.). The consortium was chaired by Marian Petre and Paul Piwek. A number of senior members of the Network acted as discussants for the student presentations. (See Appendix for the activity programme).

The day after the doctoral consortium a gender panel was held with contributions from Donia Scott (chair), Anne de Roeck, Marian Petre and Dipak Kalra. They set of a very stimulating discussion, which ranged from the relation of gender issues and the promotion of diversity in general, to the worrying trend of ever decreasing student numbers in the field of computer science.

In 2006, the doctoral consortium was held in Balatonfüred, Hungary, on July 3rd and 4th. 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. A number of senior members of the Network acted as discussants for the student presentations and discussion groups. The consortium was chaired by Marian Petre and Clara Mancini. We compiled and distributed a booklet of presentation abstracts and other material, which were made available at <http://mcs.open.ac.uk/semantic-mining/>. (See Appendix for the activity programme).

In 2007, the doctoral consortium was held in Barcelona, Spain, on June 25th. With respect to the previous year, the programme was shorter but dense of activities nevertheless. We had short presentations on research programme topics by senior researchers, followed by moderated group and plenary discussions, and we had short student presentations (the students were also introducing one another and chairing one another's question-answer sessions). In addition, we also had a tutorial on presenting research work with PowerPoint. 17 students and several senior researchers participated very actively in the activities of the consortium with general satisfaction (according to the feedback we received). The consortium was chaired by Clara Mancini. We compiled and distributed a booklet with research abstracts and other material, which were made available at <http://mcs.open.ac.uk/semantic-mining/>. (See Appendix for the activity programme). This year, certificate of attendance were given to the participants.

Overall, the doctoral consortia organised within WP6 Mobility Programs were very successful. They were well attended and feedback on their activities was very positive.