

Vision and Goal 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 focus 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.

Challenges

The need for cross-referencing between biological and clinical information provides a grand challenge. The vast amount of data available in laboratory databases, together with the growing volume of electronically available clinical information call 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 –
WP25¹**

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

**Principles in
ontology
engineering –
WP21**

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**

¹ Joint research in SemanticMining is organised in work packages (WP20-WP28)

Research carried out with language technology in the network also address the need for approaches in Europe which will bridge language barriers and facilitate access for non-English native persons to the large scientific corpus of texts written in English. Because patient 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 used when specifying the instances of these rubrics for a particular patient, offer the principal boundary problem described above.

The Electronic Health Record – WP26

Terminology services – WP28

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

Health care statistics – WP23

interventions, the challenge is to allow aggregation according to different aspects and to assure high information quality on all levels of data abstraction.

**Long-term goals
of Semantic-
Mining**

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.

Project title: Semantic Interoperability and Data Mining in Biomedicine
[SemanticMining]

Project Identifier: FP6-2002-IST-507505

Instrument: Network of Excellence

SemanticMining is based on the partnership of 23 partners from 11 European countries with approximately 100 identified researchers (25 female) and 35 associated PhD students (10 female).

Project co-ordinator: Hans Åhlfeldt, Linköping University, Sweden

Email: hans.ahlfeldt@imt.liu.se

Website: www.semanticmining.org

