

# EXPRESSION OF INTEREST

## Network of Excellence

### **Semantic Interoperability and Adaptive Methods for Data Mining – Applications within Biomedicine**

#### *Prepared by*

Ass Prof Hans Åhlfeldt  
Dept Biomedical Engineering / Medical Informatics  
S-581 83 Linköping University, Sweden  
E-mal: [hans.ahlfeldt@imt.liu.se](mailto:hans.ahlfeldt@imt.liu.se)

#### In cooperation with

Dept Biomedical Engineering / Medical Informatics, Linköping University, Sweden  
Dept Learning, Informatics, Management and Ethics, Karolinska Institutet, Stockholm  
Sahlgrenska University Hospital, Göteborg, Sweden  
Dept Medical Informatics, Computational Linguistics, Freiburg University, Germany  
Dept Medical Informatics, Geneve University Hospital, Switzerland  
Dept Computer Science, University of Manchester, England

National Board of Health and Welfare, Center for Epidemiology, Sweden  
WHO Collaborating Centre for Classification of Diseases in the Nordic Countries

*This Expression of Interest was submitted in response to Call EOI.FP6.2000*

7<sup>th</sup> June 2002

## **1. Aim of the Proposed Work**

**Networks of Excellence:** The aim of this proposal is to establish Europe as the international scientific leader in medical and biomedical informatics. The field has its methodological roots in computer science, information theory and biomedical engineering with real world problems from biomedicine and the health care sector as application area. The long-term goal of the network will be development of generic methods and tools supporting the critical tasks of the field; data mining, knowledge discovery, information retrieval, and decision support for establishment and dissemination of evidence based medicine.

### **1.1 Contribution to Priority Thematic Area of Framework 6**

The network will contribute to the fields Biotechnology for Health and Information Society Technologies as described in the priority thematic areas of research in FP6. The network will specifically contribute to join the efforts of medical informatics and bioinformatics through development of generic methods and tools for managing and interpreting the vast amount of digital data produced by functional genomics and the emerging electronic patient record of clinical medicine. The network will contribute to the development of technological platforms for generation and distribution of evidence-based medicine, taking full advantage of the potential of information and knowledge-based services offered by the information and communication technologies.

### **1.2 Contribution to the European Research Area**

The proposed effort will contribute to the European research area through the establishment of a multidisciplinary network composed of researchers, industry and health care professionals. The network will establish close links to the standardization work in health informatics organized by CEN TC251 and its four working groups (I) Information Models, (II) Terminology and Knowledge Bases, (III) Security, Safety and Quality, and (IV) Technology for Interoperability. Cooperation will be performed with WHO collaborating centres and national boards for health care system administration and epidemiological studies. Support from FP6 for establishment of a Network of Excellence covering the areas of medical informatics, bioinformatics and the standardization body will work against the fragmentation and the under financing of the field. We strongly believe that a unified effort based on research in the scientific disciplines of medical- and bioinformatics with application of emerging standards and technological infrastructures will be necessary for the development of a cost effective distributed health care system.

## **2. Background to the Proposed Work**

It is well known that the health care system is faced with a series of challenges concerning quality and cost-effectiveness. The distribution of cost-effective health care allowing the patient to take active part in the caring process, provision of evidence-based care on all levels in the system and effective use and reuse of information are key issues for the health care organization. The information and communication technology infrastructure should therefore reflect the view of the health care system as a seamless system where information can flow, although under strict regulation, across organizational and professional borders. The complex application area of distributed health care provide in this perspective a series of general research problems which we believe must be studied to allow successful systems to be implemented. A basic problem is semantic interoperability, which simply means that semantics is preserved in communication between health care actors using information systems, a condition which should be natural but has proven to be very hard to achieve. The need for a unified framework for data mining in large data repositories of different data types (signals, symbols, text etc.) and semantic interoperability is closely related to the need for a unified framework for managing and analyzing data of both genotype and phenotype.

## **3. Expected Results from the Proposed Work**

The results of the combined efforts in the network will be described in relation to the following problem areas.

### **Terminology issues in health care**

Terminological systems used in health care are systems supporting communication of information through standardized use of language. These systems include thesauri, nomenclatures, classifications, vocabularies and also formally and non-formally represented health-care models. Traditionally the systems have been disseminated through paper-based publications. Now however, the integration of terminological systems with health care information systems is becoming more viable. Thus, the focus of this research is on terminological

systems within health care and their use in health care information systems such as documenting and retrieval systems, and for cross-referencing between clinical and pre-clinical terminologies.

Consistent use of language, or more generally any piece of information used for communication purposes, is a prerequisite for high-quality classification and for semantic interoperability. An ideal terminological system should thus serve as an inter-lingua in communication in-between health care workers, and permit data comparison across organizational, professional, and even national borders. Concept representation systems based on formal logics is a key technology. Description logics, as a knowledge representation formalism based on formal logics, offers a precise language for the engineering of domain ontologies, and provides taxonomic subsumption as a basic inference mechanism. Off-the-shelf implementations exist, and considerable experience has been gained in previous projects in the medical domain, e.g. through the GALEN-project. The health care domain is rich of terminological systems targeted towards different areas and specialties, and several models are proposed as basic reference terminology, possibly providing a common framework for further integration and mapping between specific terminological systems. A great challenge would be to combine the massive coverage offered by informal medical terminologies (such as ICD10, ICF, ICPC etc.) with the high level of formally solid description logics formalisms. Of specific interest for realization of a reference terminology based on formal logics is the new release of SNOMED CT.

### **Data mining and information retrieval**

Considerable progress has been made in the field of text processing and information retrieval during the last decenium but these achievements have not yet reached a level of maturity or reached the health care sector as ready-to-use products. Successful information retrieval rely on consistent use of an indexing language which, depending on the actual domain, consists of terms from any of the terminological systems. This research will also address the need of multilingual approaches in Europe, to bridge language barriers between countries and to facilitate access to non-English native persons to large scientific corpus of text written in English.

Information retrieval techniques include mathematical modeling of the relation between document content, which in our context is the patient record, and the end-user query. Generally, the set of terms used for indexing and thereby also for retrieval is a flat list without internal structure apart from what is given lexically. An interesting research perspective is how the structure of conceptual models imposed on the indexing system could be utilized to improve the performance of search engines.

A common problem in text mining and information retrieval is the lack of a suitable metric to measure similarity between documents and a user query. An approach to handle synonymy is latent semantic indexing in which a low rank approximation of the term-document-matrix is obtained using singular value decomposition. Previous research has shown that canonical correlation analysis (CCA) is superior to traditional principal component analysis and that CCA can effectively discover and model relevant relations between different sources of information by maximization of mutual information, a concept well known from information theory. Thus, our objective is to further develop non-linear CCA and related methods and to combine them with existing methods for text processing with applications to e.g. automatic classification of medical diagnosis from patient records and cross-language information retrieval. The overall goal is to develop a common framework for mining of relations in large data volumes composed of signals, symbols, text etc. Such a framework will be critical for knowledge extraction from clinical and pre-clinical databases including the information explosion of genomics and proteomics. Searching automatically for interactions between genes and proteins names on one side, and diagnosis, signs and symptoms and related body parts on another side, is ready to strongly sustain the experts of the field.

Additional problems for medical language processing include automated indexing vs. manual indexing vs. semi-automatic manual indexing, construction of lexicons for capturing meaningful terms (morphemes, words, noun phrases, acronyms) according to different languages, algorithms for word stem extraction and noun phrase recognition. Evaluation of information retrieval and text mining requires extensive annotated test corpora which do not yet exist for the medical domain. A main question is the degree of content abstraction we aim at. On the one hand it must be expressive enough in order to measure the semantic distance, on the other hand it must be viable with regard to lexical and conceptual coverage as well as computational cost.

### **The semantic web**

High requirements on semantic interoperability is put forward by the notion of 'the semantic web'. The web search engines have become extremely powerful in terms of processing speed and information storage, but there

still exist fundamental problems concerning web navigation and retrieval. By the expression the semantic web is meant a system where requirement of exact match between search terms is relaxed while more functions are offered to support semantic equivalence and other types of associative relations between expressions or concepts. We understand our work as an active contribution to the semantic web initiative, which offers a useful framework for adding semantics to web documents by using metadata based on common standards, such as XML for document markup and DAML+OIL for expressing the underlying knowledge. Different ways to generate semantic web compatible documents should be analyzed, e.g. the knowledge-based authoring of documents, semantic tagging of unstructured text, etc.

### Information quality

The documentation of clinically useful, patient-specific information is fundamental for management of health care related problems. Over the last years, the structure and content of a multi-professional patient record have gained much interest, one reason being the introduction of electronic patient records (EPRs), and another being organizational issues originating from the objective of using the patient record for communication over organizational and professional borders and as an active instrument in planning and outcome analysis. The EPR has to face the requirements of quality assurance and evidence-based practice, but at the same time preserve the richness of the patient-centered story. In this perspective, it is well known that the traditional paper-based patient record suffers from a series of limitations.

Consequently, demands of information handling within the health care sector range from clinically useful, patient-specific information to a variety of aggregation levels for follow-up and statistical reporting. Several coding systems are for this purpose put into use in domains such as diagnoses, health problems, interventions, and procedures. The challenge is to assure high information quality on different levels of abstraction, and to allow aggregation according to different aspects or viewpoints.

### Expected results

Expected results from the proposed work can be summarized as follows:

Expected Result	Users of the Results
Adaptive methods for data mining – a unified framework for signals, symbols and text	Researchers, health care professionals
A reference terminology for clinical medicine and biomedicine	Researchers, system developers
Methods and tools for information retrieval in clinical and biomedical databases	Researchers, health care professionals, system developers
Standardized software modules for realization of the virtual patient record	System developers, health care professionals
Standardized software modules for terminology services	System developers, health care professionals
Framework for distribution of evidence-based medicine	Health care professionals, system developers, citizens
Methods and tools for realisation of ‘semantic web functionality’ within biomedicine and health care	Health care professionals, system developers, citizens
Increased quality of aggregated data for follow-up and quality control	Health care professionals, politicians

## 4. Activities to Achieve the Proposed Objectives

### 4.1 Integration Activities

The integration activities include training for all partners in the state of the art of the individual scientific fields, workshops on the definition of the integrated approach, joint training courses, electronic communication networks, mobility of personnel, establishment of shared databases, and websites.

## 4.2 Research Activities

Specific research activities will be formed within the problem areas described above based on the input from all partners in the network.

## 5. Expertise Needed to Achieve Objectives

The network will need a critical mass of European scientists in the field working together in a multidisciplinary context with health care professionals, policy makers, standardization bodies and industry. Currently the network is composed of the following partners, but the network is open for additional members with interest and competence in the field.

	<b>Organisation</b>	<b>Country</b>	<b>Contact persons</b>	<b>Area of Excellence</b>
<b>1</b>	Dept Biomedical Engineering / Medical Informatics, Linköping University	Sweden	Hans Knutsson Hans Åhlfeldt	Adaptive methods for data mining, information theory Knowledge representation Terminological systems
<b>2</b>	LIME, Karolinska Inst. Stockholm	Sweden	Gunnar Klein Gunnar Nilsson	Standardization work (TC 251), electronic patient records and terminology
<b>3</b>	Gothenburg Univ Hospital	Sweden	Anders Thurin	Standardization work (TC 251), terminological systems
<b>4</b>	Dept Medical Informatics, Freiburg Univ Hospital	Germany	Rudiger Klar Stefan Schulz	Medical language and knowledge engineering Computational linguistics
<b>5</b>	Computational linguistics lab, Freiburg University	Germany	Udo Hahn	Computational linguistics Ontology engineering
<b>6</b>	Medical Informatics, Geneve Univ Hospital	Switzerland	Robert Baud Anne-Marie Rassinoux	Computational linguistics Natural language processing
<b>7</b>	Medical Informatics Group, Computer Science, Univ of Manchester	England	Alan Rector Jeremy Rogers	Medical language and knowledge engineering Description logics
<b>8</b>	National Board of Health and Welfare, Epidemiology	Sweden	Lars Berg	Medical classifications, epidemiology
<b>9</b>	WHO Collaborating Centre Nordic countries	Nordic countries	Björn Smedby Martti Virtanen	Medical classifications, epidemiology
<b>10</b>	To be continued ...			

## 6. Promotion of Results Outside of the Consortium

Results will be shared through the following mechanisms: scientific publications, contribution to international standards (especially CEN TC251), input to EU and national policy committees, input to educational courses, direct to users (workshops targeted at the above associations, web communities, email groups, training courses, e-learning initiatives, guidelines, best practices) and input to regional development organisations.

## 7. Project Management

The network will be composed of a core group of main partners supported by assistant partners with specific tasks. Professional expertise in project management will be utilized when setting up the network. Calls for proposals will be launched by the consortium to invite SMEs in the areas of ICT and biotechnology to participate in the network as well as new academic and professional partners.