SemanticMining

NoE 507505

Semantic Interoperability and Data Mining in Biomedicine

Deliverable 26.1

Report on EHCR

Delivery date: month 11

Report Version: 1

Report Preparation Date: 2004.12.19

Dissemination level: RE

Associated work package: 26

Lead contractor:

University College London (UCL)

CHIME, Holborn Union Building, Highgate Hill, London N19 5LW

UK

Editor: Dr Dipak Kalra: d.kalra@chime.ucl.ac.uk

Table of Content

Administrative information

Lead contractor/partner for WP/Deliverable
University College London

Assisting partners for WP/Deliverable
University of Manchester
INSERM
CNR-ISTC

Author list
UCL: Dipak Kalra
              David Lloyd
              Thomas Beale
UoM: Alan Rector
              Jeremy Rogers
              Rahil Qamar
INSERM: Christel le Bozec
              Bruno Frandji
              Marie-Christine Jaulent
              Patrice Degoulet
CNR-ISTC: Domenico Pisanelli

Summary

The challenge of richly interpreting electronic health information, in order to populate EHR instances with suitable terms, to provide decision support in the care of individuals, to identify suitable patients for teaching or clinical trials recruitment, and to mine populations of records for public health or to discover new medical knowledge, all require that the heterogeneous clinical entry instances within EHR repositories can be systematically analysed and interpreted.

Achieving this requires the combination and co-operation of many different health informatics tools and technologies, underpinned by shared representations of clinical concepts and inferencing formalisms. Much of this work is at the level of R&D, and is well represented across the Semantic Mining consortium.

The challenge of WP26 is to build up a vision of the ways in which these historically independent threads of health informatics research can collaborate, and uncover the research challenges that are needed in order to deliver good demonstrations of semantically indexed and richly analysable EHRs.

The partners have begun WP26 by acquiring a better knowledge of each other's areas of endeavour, and are beginning to steer their research interests towards future areas of collaboration.

1 Overview

1.1 Objectives

Objectives	Progress towards achieving objectives
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. To support the future seamless and standards-based interaction of knowledge, record and inference services.	Partners are gradually acquiring an understanding of the relationships that need to exist between EHR services, archetype services, terminology services, concept services, inference services and guideline services within an overall collaborating middleware environment, to support a range of users and use cases

1.2 Milestones

Milestone	Planned date	Actual date	Comments
Formalisation of the archetype approach	September 2004	September 2004	
Discussion of the interfaces between archetypes and ontologies	October 2004	November 2004	
Formalisation of active research threads	December 2004		Probably realistically by Feb 2005

1.3 Project meetings

Milestone	Planned date	Actual date	Comments
Discussion during EHR workshop (organised through WP16)	October 2004	27 November 2004	

1.4 Deviations from Plan

Causes and Description	*Corrective actions*
Recruitment delay	The lead partner (UCL) has experienced some delay in obtaining work permit clearance to employ an international expert in the field, whose intended research is at the heart of WP26. This person is now in post and work has begun in earnest.
Scheduling of the EHR Workshop	This workshop was planned to be the significant forum at which the main informatics participants would be present and at which the research challenges could be discussed. As expected, it has provided the real human insights needed to get WP26 underway, but was held about two months later that was originally hoped.

2 Main Results

The vision

Electronic Health Record Systems are becoming widely available, supporting clinical data storage and retrieval, at present mainly for the benefit of the local information holder. However the capabilities of these systems are often still far from (and sometimes contrary to) what might be expected from an information system dedicated to the support of clinical care, in terms of:

- the quality of the clinical information (precision, completeness);
- its availability and understandability (semantic interoperability) by other care providers or by the patient directly;
- the ability to support knowledge-based clinical decision-support, data retrieval and aggregation.

The ability to understand and exploit richly the information that is held in the individual electronic health record of one person, and to mine populations of electronic health records, is a challenge that draws on many threads of health informatics research. However, these are largely isolated niche areas of health informatics, with limited mutual awareness of the problems being tackled, the progress made, and of the possible solutions that each niche could offer to another in support of common challenges and goals.

The vision encapsulated by Workpackage 26, as part of a Network of Excellence project, is

- to enable the various Semantic Mining partner teams who work in each of these niche areas (threads) of health informatics to come to a better understanding of each other's areas of expertise and of the capabilities within each other's domains;

- to build up the research activities at each partner site that contribute to the interfaces and commonality that must be developed within and between each thread of health informatics;

- to develop a shared understanding of a "big picture" in which these threads of informatics, when expressed through middleware components, can form part of a mutually interoperable rich computational resource to enable the understanding of EHR data in support of individual and population health, research and learning;

- to work towards future demonstrators and pilots, by bidding together for further research funding, that enables this bigger picture to be realised and evaluated;

- through all of the above to enable the participating research teams to become individually and collectively world leaders in this aspect of health informatics.

The research challenges

In the first few months of this Workpackage, the partners have been developing and refining aspects of their individual research activities that will contribute to this broader vision, whilst learning more about the work of others and starting to recognise limitations and strengths in each research thread. An important opportunity to share that understanding was afforded by the first Semantic Mining EHR workshop, organised through Workpackage 16, in November 2004 (and reported in Deliverable 29).

Through this workshop, and a preceding satellite conference on ontologies, it became clear that each thread of health informatics research presently has limitations that will make it difficult for its formalisms and tools to be really useful to some of the other threads.

Europe has accumulated over thirteen years of Commission-funded research and is developing its third generation of CEN standards on the structural representation of EHR information, towards realising the vision of interoperable federations of clinical systems and EHR servers. Implementations of generic EHR servers are still limited to academic and small vendor products, but the large vendor systems are increasingly capable of aggregating significant proportions of a person's healthcare journey. However, historic work on EHR representation has focussed on preserving faithfully the original meaning of the author and the original representations within the clinical feeder systems that contribute to a federal EHR. This inevitably means that these EHRs retain some internal heterogeneity: similar kinds of clinical information may be represented using different data structures and clinical schemata, if the original authors and systems chose different ways of representing that information.

Over the same period of time terminology systems have evolved from simple hierarchical code sets to poly-hierarchies capable of rich combination (pre- and post-coordination) and expressiveness. These however now permit multiple ways of representing the same concepts and ideas – with some but presently imperfect ways of mapping these alternatives to each other.

Text mining algorithms and tools are challenging the traditional assumption that clinical information must be coded to be computable. However, subtle and complex constructions are still difficult to interpret and few will yet regard the field as advanced enough to use text-mined data for clinical decision-making.

Decision support agents and guideline systems are starting to show how evidence based care can be delivered in practical everyday health care, reducing mortality and morbidity. However, in order to be widely adopted these must interoperate with EHRs, to avoid clinicians having to enter data into a guideline that the EHR already "knows", and each guideline system must be able to leave a medico-legal trace of its advice in the EHR as part of the record of care.

Population queries and the data mining of dedicated clinical data warehouses are starting to yield novel results and conclusions that are capable of advancing medical

knowledge in unexpected ways, complementing clinical trials. However, these are almost always software programmes that are bound to particular databases and clinical content – little of this work is generic.

We know that much of what is needed to analyse EHRs, in order to answer research questions, cannot easily be expressed as a simple EHR query because the data is not formally recorded as a single data point, but rather needs to be inferred from across multiple clinical documents and entries. Formalising this kind of inference is still a long way off.

Ontologies are now being developed to represent the systems of concepts of many medical knowledge domains, with much energy currently being devoted to bioinformatics areas. However, these are attempts to systematise a knowledge domain from the perspective of the knowledge providers, to assist in database curation and to ensure it grows in a logical and coherent form. However, for each ontology, this perspective is only one of the possible ways in which the information might be categorised – the field has seemingly so far not been good at envisaging the alternative viewpoints that "customers" of each ontology may require, or developing a wide range of use cases for exploiting its contents.

All of these formalisms and advances have been developed largely in isolation from each other. If we are to take full advantage of each, and really exploit the health record information that is increasingly being captured electronically, these approaches need to evolve closer together: to develop common formalisms to share clinical meaning, and service interfaces that are of use to the other components in this healthcare computing middleware environment.

Medical information systems and standards are increasingly based on principled models of at least three distinct sorts of information - patient data, concepts (terminology), and guidelines (decision support). Well-defined interfaces are required between the three types of model to allow development to proceed independently. Two of the major issues to be dealt with in the defining of such interfaces are the interaction between ontological and inferential abstractions - how general notions such as 'abnormal cardiovascular finding' are abstracted from concrete data - and the management of the meaning of information in guidelines in different contexts. A principled approach is also required to decide which information belongs in each model based on the nature of the queries or inference to be performed: necessary or contingent, open or closed world, algorithmic vs heuristic.

Specific activities and research interests

UCL has invested considerable effort over this year in advancing the formalisms to support the use of archetypes, in partnership with others through the *open*EHR Foundation and through work within CEN/TC 251 towards prEN13606. In parallel, INSERM has been developing a system of concepts that will permit a pilot of archetype-like specifications in the design and population of patient functional health

questionnaires. The University of Manchester has been exploring how ontologies and inferences can be used to interrogate archetype and EHR repositories.

What are archetypes?

When understanding the nature and role of archetypes, it is important to remember that these are intended to be used alongside a Reference Model. In the case of the EHR, the Reference Model represents the global characteristics of health record entries, how they are aggregated, and the context information required to meet ethical, legal and provenance requirements. The model defines the set of classes that form the generic building blocks of the EHR, and reflects the characteristics of an electronic health record that apply to all clinical domains.

Such a generic information model for the EHR needs to be complemented in the knowledge domain by a formal method of communicating and sharing the named hierarchical structures within EHRs, the data types and value ranges that actual record entries may take, and other constraints, in order to ensure interoperability, data consistency and data quality.

Archetypes each define (and effectively constrain) allowed combinations of the building-block classes defined in the Reference Model for particular clinical concepts by specifying particular record component names, data-types, and may constrain values to particular ranges.

Archetypes can each be viewed as a statement of the rules by which terms and other data may be constructed into a meaningful health information representation. Archetype instances themselves conform to a formal model, known as an Archetype Model and are expressed in various forms. The standard form accepted by *open*EHR is Archetype Description Language (ADL). Although the ADL and Archetype Model are stable, individual archetype instances can be revised or succeeded by others as clinical practice evolves. Version control ensures that new revisions do not invalidate data created with previous versions.

Since any instance of EHR data can either refer directly or be mapped to an archetype to which it conforms, each archetype effectively identifies a set of EHR instances that meet a particular clinical (business) purpose, and contain particular EHR data items and candidate values. From an EHR perspective, the archetype therefore provides the most fundamental level of semantic indexing of the structural organisation of EHRs, and is the logical interface for richer systems of concepts (ontologies) and other services that need to mine the EHR.

Progressing the work on archetypes

Since the start of Semantic Mining, UCL has worked with others to produce a first published set of requirements for archetypes, and a generic Archetype Model. Early versions of ADL have been refined, and colleagues working in Australia have developed a new archetype editor. This evolving work has been shared with selected

Semantic Mining colleagues, who are now in a stronger position to critique, utilise and extend the work as part of shaping their own research work plan.

As part of the MRC funded CLEF project, UCL has also developed a query formalism and computational interface to permit population queries to be performed on its EHR server.

The next phase of work within UCL will be to develop a design for an archetype ontology, in partnership with University of Manchester and INSERM, to permit a broader range of queries to be performed on the EHR. This ontology will need represent a systematised "view" of EHR domain knowledge that can be held in common with other medical knowledge ontologies and inference services.

Integrating archetypes with terminologies

This work fits well with research interests and activities at the University of Manchester, dealing with the integration of archetypes with existing terminologies., Archetypes, besides referencing a specific Reference Model, can be bound to defined terminologies in standardised terminology engines such as SNOMED-CT, LOINC, GALEN and others.

One important issue that is critical to ensure that archetypes are used widely within existing health care systems is to enable automated integration of the Reference Model, so that the archetypes fit in well with existing policies and requirements of the health system. Since terminology engines are being adopted increasingly within health care and archetypes attempt to bind to these terminologies within the model, it is imperative that some sort of validation mechanism is made available to ensure that archetypes and terminologies, alike, are being used in the same context within the domain. This work will be taken forward over the coming year as part of a PhD.

Integrating archetypes with ontologies

CNR-ISTC has been exploring the ontological status of archetypes. This includes reviewing the extent to which individual archetypes do contain ontology fragments, and what kinds of external ontologies are needed to map onto a library of archetypes (for example, within an archetype repository). This work is at an early stage, and will be taken forward in partnership with the other WP26 members in the New Year of 2005.

Chronicles

Within the MRC funded CLEF project, work is underway at Manchester to define a semantic view of the content of a typical medical record, and to determine how this view relates to the traditional electronic record and how it might be reconstructed from information in that record.

Consider that several different clinicians, on different days over a possibly long time period, may choose to record the specific diagnosis of a given patient (but perhaps not always with the same degree of precision), for example in the introductory paragraphs of several clinic letters. One caricature of the EPR is that it serves a primarily medico-legal requirement to faithfully record 'who said what and when'. A different view of the story of the patient illness is 'what is the most precise statement of what was known of the patient and how was it known'. This view seeks to hide repetitious statements as well as to make explicit things that were regarded as obvious but never recorded. Expressed as a semantic network, this view of the patient story is termed the 'Chronicle'.

The terminology boundary problem

Another longstanding challenge, recently brought into the limelight by the arrival of, but by no means limited to, SNOMED CT, is how to use terminology models and information models together to represent clinical statements and biomedical information in electronic health records for the purposes of querying, retrieval, and decision support. The problem is that sophisticated terminologies permit a range of different pre- and post-coordinated expressions to represent the same idea, and that EHR architectures contain many contextual attributes that overlap with qualifiers and modifiers within such terminologies. Expertise within the Semantic Mining consortium makes this Workpackage well placed to contribute clinical examples and candidate contributions to the resolution of this problem. At least one partner has already attended and contributed to international meetings on this topic.

Developing systems of concepts

INSERM ERM202 is evaluating the design and implementation of shared clinical concept systems in the context of the deployment of the main French industry Electronic Healthcare Record System within large leading healthcare institutions. The tasks undertaken to date include:

- To survey existing accomplishments exploiting controlled clinical conceptual systems within Electronic Healthcare Records (EHRs)
- To design a controlled concept system based on well accepted existing international clinical concept systems and adapted to the constraints from the EHR system used within our healthcare organisations
- To design EHR data retrieval and aggregation systems dedicated to some precise clinical objectives
- To evaluate the benefit of this approach

The first two steps have been completed. A state of the art survey has been performed, although there is still some editing work to do to bring it into a publishable state.

In collaboration with the HEGP team an evolution of the previous dictionary of question-concepts has been proposed towards a mono axial classification of concepts

built according to the SNOMED top classes. This evolution is now in place and used within HEGP and diffused to another large French university hospital (DIJON)

A three-month study work (July, August and September 2004) was directed to evaluate the interest of this evolution and results of this study are already available (part of them were presented during the EHR workshop in Brussels, November 2004).

In HEGP 463 types of questionnaires have been designed, including 4700 questions based on 2087 question-concepts. Patient's data include 87 636 questionnaires, 51 947 filled up by medical doctors, 35 789 filled up by other healthcare providers. Depending on the specialty, 45 to 100% of patient's EHR include the set of questionnaires that have been targeted for medical information exhaustiveness.

The next step includes a study of mechanisms to convert questionnaires and question-concepts into CEN/TC 251 prEN13606 Archetypes and other architectural elements, methods for formal linkage of question-concepts and responses-concepts with domain terminologies, the design of EHR data retrieval and aggregation systems and their evaluation.

Next steps

The partners of this Workpackage have so far put most effort into refining individual threads of research in the light of growing awareness of the work of colleagues and a growing sense of a "bigger picture" in which their work can usefully be combined. The November EHR workshop provided a valuable opportunity for the main partners to share this understanding, and in the New Year they will begin to formalise the ways in which these individual activities can be interfaced to launch new research activities within the funding budget of Semantic Mining. This blueprint for new research activities will be reported in Deliverable 26.2, in around six months time.