



SemanticMining

NoE 507505

Semantic Interoperability and Data Mining in Biomedicine

Deliverable 26.2

Report on EHCR

Delivery date: month 21

Report Version: 2

Report Preparation Date: 2006.05.05

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

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



Table of Content

TABLE OF CONTENT	2
ADMINISTRATIVE INFORMATION	3
SUMMARY	3
1 OVERVIEW	4
1.1 Objectives	4
1.2 Milestones	4
1.3 Project meetings	4
1.4 Deviations from Plan	5
2 MAIN RESULTS	6
Introduction	6
3 PARTNER SPECIFIC ACTIVITIES	7
3.1 University College London	7
3.2 University of Manchester	8
3.3 INSERM	9
3.4 University of Linkoping	13
4 COLLABORATIVE THEMES ON SEMANTIC INDEXING	14
4.1 The role of Archetypes	14
4.2 Binding Record Structures and Terminology	16
4.3 Negation	18
4.4 Linking questionnaires in use with archetypes based on a generic model such as CEN EHRCOM or HL7 RIM	20
4.5 Applying ontologies to support interoperability between guideline management systems and electronic health records	24
5 CONCLUSION	25



Administrative information

Lead contractor/partner for WP/Deliverable
University College London

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

Contributors list

UCL: Dipak Kalra
 David Lloyd
 Thomas Beale
 Nathan Lea
UoM: Alan Rector
 Jeremy Rogers
 Rahil Qamar
INSERM: Christel le Bozec
 Bruno Frandji
 Marie-Christine Jaulent
 Patrice Degoulet
 Julie Niès
 David Ouagne
 Thierry Delbecque
CNR-ISTC: Domenico Pisanelli
Li: Hans Ahlfeldt
 Erik Sundvall

Summary

This deliverable is the second for Workpackage 26. The first, submitted after Month 12, summarised the areas of research that the partners had identified as being relevant to the semantic indexing of the EHR. This second one reports progress on the key threads of work identified by the partners during the project to contribute towards semantically interoperable and processable EHRs.

This report also provides a set of short summaries on key topics that have emerged as important, and to which the partners are able to make strong contributions. Some of these are also being extended via two new EU Framework 6 proposals that include WP26 partners: this is also a measure of the success of this Network of Excellence.



1 Overview

1.1 Objectives

<i>Objectives</i>	<i>Progress towards achieving objectives</i>
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.	Each of the partners is continuing research and development activities in line with those summarised in Deliverable 26.1 last year. Not surprisingly, some collaborative themes have emerged, and joint activities are now beginning within WP26. Many of the work threads also feed into wider international collaborations and to international (CEN, ISO, HL7) standardisation.

1.2 Milestones

<i>Milestone</i>	<i>Planned date</i>	<i>Actual date</i>	<i>Comments</i>
Exploration of the way in which archetypes can be linked to terminology and ontology resources	May 2005	May 2005	
Exploration of terminology and record structure binding issues: proposals for in depth investigation	May 2005	May 2005	
Further exploration of binding of SNOMED-CT to archetypes	November 2005		
Review of OWL representation of archetypes	December 2005		

1.3 Project meetings

<i>Milestone</i>	<i>Planned date</i>	<i>Actual date</i>	<i>Comments</i>
2 day WP26 workshop for WP26 partners: review of research threads and agreement of new activities	N/a	27 & 28 May 2005	A very successful meeting, with both days focussing on the actual work and semantic challenges relevant to WP26



1.4 Deviations from Plan

<i>Causes and Description</i>	<i>Corrective actions</i>
Nothing to report	



2 Main Results

Introduction

Workpackage 26 of the Semantic Mining NoE defines and manages a thread of research activity on semantic aspects of the electronic health record. The core challenge being addressed is to enable EHR data to be semantically “indexed” so that retrieval, decision support and population analyses can be performed on EHR data even if those data have been accumulated from diverse and dispersed clinical applications, and even if different terminology systems have been used. This is a complex challenge, typically considered to be subsumed by the phrase “semantic interoperability”. However, semantic interoperability is still a rather imprecise vision, whilst the challenges being addressed by WP26 permit a more focussed set of work-plan threads to be pursued.

In the first WP26 deliverable (Del 26.1, submitted at the end of Year 1), this general area of challenge was described, and the expertise of each partner summarised. Work had really only just begun on understanding these challenges, and the deliverable outlined a set of intended research journeys to be pursued during year 2.

This deliverable is able to report on the progress that has been made in understanding the sub-components that contribute towards semantic indexing of the EHR (and therefore contribute towards semantic interoperability). Firstly the focus of work at each partner site is summarised. Then some specific areas of collaborative work into which the semantic indexing challenge has been divided are described. These collaborative areas are listed below.

1. The potential role of Archetypes to support consistent recording and safe analysis; the requirements for safe/valid data mining; archetype representation and archetype policies.
2. Record structures and terminology binding. What are the difficulties, why are they a problem, and what work is emerging to help towards semantic interoperability. A specific example topic of current importance is the handling of negation, which is discussed in more detail in this report.
3. Representing patient questionnaire data within a generic EHR model (such as CEN/ISO 13606-1); enabling the analysis of these data in conjunction with clinician-authored record entries.
4. Applying ontologies to support interoperability between guideline management systems and electronic health records.

Progress on each of these topics is given. Each topic summary provides an outline of the problem being addressed, a paragraph about what the problem is; the progress being made towards tackling the problem, including the contribution made through



Semantic Mining; a resume of the current approach being taken and what future work and contribution is intended by the Partners during 2005-6.

3 Partner specific activities

3.1 University College London

***openEHR* Archetype Specifications**

An *openEHR* Archetype is a knowledge artefact instance that specifies a set of constraints on a generic EHR Reference Model. An archetype specifies how a defined set of clinical data is to be represented as part of the EHR, meeting the business requirements of a clinical domain.

The archetype specification has been developed to be a generic formalism for representing constraints. Over the past year a set of requirements, information models and an XML Schema have been developed. The *openEHR* Archetype Definition Language (ADL) has been refined in partnership with Ocean Informatics (Australia). This complete set of specifications has been incorporated into the draft Release 1.0 *openEHR* specifications, due for publication in February 2006. They have also been incorporated into a draft CEN standard (prEN 13606-2) and submitted for formal CEN Enquiry (the principal ballot cycle within CEN). This work has been lead jointly by UCL and Ocean Informatics. These specifications have also been presented and reviewed extensively within the HL7 Templates SIG, and are in the process of being harmonised with the HL7 Templates draft specification, with a view to future interoperability between these specifications. This harmonisation is proceeding slowly but positively, and UCL plays a leading role in that process.

***openEHR* Archetype tools**

An ADL parser, written in Eiffel, that underpins all of the presently available archetype tools, has been updated and refined to reflect the latest ADL specifications. A Java wrapper, incorporating over a hundred separate function calls, was developed with collaborative support from the Mayo Clinic to enable work to commence on Java archetype editors. Work has progressed on the development of a Java archetype editor, and of a simplified version that binds to a new version of the UCL Record Server. The latter is completed, and will be used initially for work on the binding of archetype nodes to ontology concepts jointly with University of Manchester. The former is nearly completed, and will be used for research into binding with SNOMED-CT terms and OWL. The intention is for the authorship of archetypes, and in particular for the binding of archetype leaf nodes to SNOMED-CT terms, to be assisted semantically through OWL-based reasoning services.

Work has also commenced to explore novel ways of implementing archetype services including Java annotations, which will work in real time alongside EHR services at UCL. This work will also be written up as an MSc project.



***openEHR* Archetype governance**

An important aspect of ensuring good quality patient records and their interoperability is to have archetypes that faithfully reflect clinical evidence and/or consensus. These also need to optimally represent clinical data structures to enable analysis and clinical audit functions to be performed accurately across populations of records. A new thread of work, which is being undertaken in collaboration with Australian colleagues, will define principles for archetype governance, authoring guidelines and quality criteria. This work will interface with a new EU Framework 6 Project Q-REC.

New EHR Server

Considerable progress has been made during this year on the design and implementation of an EHR server at UCL that is compliant with the forthcoming CEN/ISO EHR Communications standard EN13606. This engineering work, undertaken in Java and using the latest tools and components of the Java computing environment, has provided valuable feedback on and verification of this draft standard. The server will be used to support a live clinical service in cardiology in north London, replacing a predecessor server that conformed to the former CEN EHR standard. It will then be used to support new research jointly with other WP26 partners and for the new WP28.

3.2 University of Manchester

Standardising Clinical Data Models to External Terminology Systems for use in Healthcare organisations

This research is concerned with the development of a methodology for the standardisation of clinical content in archetype models with respect to SNOMED-CT. The SNOMED-CT terminology is a Description Logic based terminology which helps in providing semantic structure to clinical content.

The project has evolved during 2005. The major refinements to the methodology have been:

- (i) Change of terminology from GALEN to SNOMED-CT. The reason for the change was that the project relied on the transformation of the GALEN terminology from GRAIL to OWL for processing matches. This became a tedious and problematic task when additional GALEN OWL segments were required to run more examples. SNOMED-CT enables the extraction of required segments directly from the database making it easier to run several examples conveniently.
- (ii) Change in strategy for performing semantic and syntactic searches. The Morphosaurus tool developed by Freiburg University has been replaced with a suite of applications from National Library of Medicine (NLM) and other independent suppliers.



Progress made so far:

- (i) The methodology has been divided into three main but overlapping categories – (a) syntactic procedures, (b) semantic procedures, and (c) normalisation procedures.
- (ii) Syntactic procedures have been developed using the EMT-P¹ service provided by NLM. It is a natural language processing tool that generates a list of appropriate matches of the archetype terms with UMLS concepts.
- (iii) Semantic procedures have been developed with the help of the UMLS^{SKS} service provided by NLM. SNOMED terms that belong to the same semantic type as the UMLS concept are extracted as possible mappings.
- (iv) Remaining unmatched archetype terms are sent through a series of normalisation techniques involving a mix of both semantic and syntactic methods. The GATE application and a few UMLS Lexical tools have been used in addition to a bespoke component to facilitate the normalisation task.
- (v) The final stage of this phase needs to be developed but essential involves the elimination of inappropriate resultant matches through a context and semantic search. The output will be a set of candidate terms that will be presented to the archetype modeller for inclusion in the Term Binding section as well as the EHR.

The future tasks that will form a part of the research include:

- (i) Development of a reference ontology that will include medical logic for determining the most appropriate set of SNOMED-CT codes to choose from in the given context.
- (ii) At the outset, to incorporate the strategy as mentioned in the previous point for inclusion in archetype models that have “free text” facilities for a given clinical entry. To later plan an extension of this facility for runtime modelling.

The aim of this work is to help in achieving a degree of standardisation of clinical content held within archetype models to the SNOMED-CT terminology. This research is limited in scope to the number of case studies that can be investigated due to time restriction.

3.3 INSERM

Context

The healthcare professionals and patient themselves need to share medical data & information for diagnosis and treatment (e.g. healthcare networks or French DMP), decision support, clinical research etc. Medical information is being available within Electronic Healthcare Records (EHR) managed by Hospital, GP, Regional Network, Patient information systems.



When exchanging EHR data & information between two or more systems, several levels of interoperability can be achieved.

- Functional interoperability allows the exchange of extracts from EHRs between two or more systems. This involves agreeing on the common network protocols, the transport binding and the message format. EHR standards have been developed for many years, proposing message formats.
- Semantic interoperability implies mechanisms allowing two or more EHR systems to exchange meaning of information. The meaning could be available to the systems users (semantic interoperability at the user level), or at the systems themselves for information automated processing like decision support systems (semantic interoperability at the system level)

Supporting semantic interoperability for medical data is much challenging due to the diversity, contextual dependence and complex nature of the clinical information, the strict ethical and legal requirements of good clinical practice, the current base of existing healthcare record systems, all developed according to their own data architectures. The exchange of meaning implies the usage of an universal language (shared concepts and relations in the medical field), or at least through high level languages that are translatable. The challenge is the consensual operationalization of such languages (e.g. usable by a computer) in order to:

- on one hand enable systems to talk comprehensively to each other;
- on the other hand to ensure a sufficient expressivity (e.g. allowing to encompass all information at the level of an individual at a specific time).

Current responses from ongoing work are based on the definition of a single healthcare information reference model (architecture) providing with a shared logical structure of information elements described within their context. Such reference models include the HL7 RIM version 3 and the CEN TC251 13606 part 1 (EHRCOM). These models are (more or less) independent from the clinical content of the information exchanged, providing only the logical structure allowing to organize, classify and position the information elements.

For a human to correctly understand the clinical content of the extracts, additional features are required to identify and describe the information elements from the professional, clinical point of view. This is addressed by the archetype approach [Beale T]. An archetype is a syntactically and semantically structured aggregation of vocabulary or other data that is the basic unit of clinical information [Elkin HL7 Templ SIG]. Archetypes provide a consensual description of healthcare professional business objects such as a Blood Cell Count or a Surgery Intervention Protocol. Archetypes are typically stored separated from the Information Model in archetype repositories. In order to support the full interoperability at the system level the previous mechanisms have to be linked to domain ontologies (like ICD 10, SNOMED CT, UMLS...), providing a common understanding of the organisation, classification and description of the professional concepts included within the extract.

Approach to the research



More and more hospitals have already implemented EHR and designed hundreds of models/templates of templates of clinical documents, relying on a local medical dictionary (own local languages (mostly not exportable)) and used in daily practice for structured data entry. EHR are often relying on proprietary information/data model. For these hospitals, the issue is to propose methods and/or tools for:

- Sharing documents, templates of clinical documents, terminological resources with others hospitals relying on shared and validated templates, archetypes and reference terminologies maintained into a repository
- Designing new local templates making use of shared validated professional/business objects models i.e archetypes or HL7 templates (as they are or after specialization or customization).
- Aligning their medical dictionary on standardized terminological resources
- Contributing to archetype & template authoring (proposing existing questionnaires and/or parts of questionnaires to be candidate to become archetypes).
- Contributing to reference terminology update

The hypothesis to verify is that theoretical results from research and standardisation efforts are today mature enough to offer the suitable background and tools supporting the evolution of existing local EHRs into elegant, standard and interoperable EHRs.

HEGP context

The component-based information system of HEGP (Georges Pompidou University Hospital, “Assistance Publique - Hôpitaux de Paris” (AP-HP) includes computerized orders, and results, notes. Decision support for computerized ordering of medications is under development. A hospital policy was voted in May 2004: to enter computerized notes in the EHR as structured questionnaires for most encounters. For hospitalisations, reports built from computerized notes will be sent to the referring physician before at least 8 days after outcome. The migration of pre existing specialized records into the structured questionnaires of the medical record component is on the way.

Current issue is the progressive sharing of documents with affiliated institutions, practitioners and patients.

This experience has already given evidence or problems to be solved like the distinction between question concepts and response concepts, how to infer response concept from binary (yes/no), numerical or free text values and how to link both sets to terminologies. Professional/business objects models (i.e. archetypes) provide partly solutions to these issues.

Defining a pragmatic methodology to make use of validated and shared archetypes, templates and reference terminologies for structured data entry i.e for designing local templates (questionnaires in use at HEGP), for sharing documents among healthcare networks (e.g. French DMP), enhancing the quality of data entries and queries.

Recent progress and future plans



During the fourth trimester we have designed and established plans for the following methods:

a) Analyse differences between models

- ☞ In the structure & content

Between data model and reference information models

Between local templates and validated shared templates

Between local medical dictionary and reference terminologies

- ☞ In links

Between local templates and local medical dictionary

Between medical dictionary and medical data in the underlying data model

- ☞ In functionalities between local editing tool (templates and medical concepts) and archetype & ontologies editing tool (PROTÉGÉ)
- ☞ Of the local medical dictionary and the PROTÉGÉ data model (ontology based database)

b) Develop a mediator to automate the model analysis described above and propose solutions for bridging consensual archetypes and local templates

c) Author local templates from shared archetypes & templates

- ☞ Select templates to be designed in different medical areas (Generic Medical Observation ,Oncology, Cardio Vascular Radiology, AVK, etc)
- ☞ Search and select validated and shared archetypes corresponding to the same medical domain in the repository (= Analysing the current concept dictionary to identify the relevant archetypes /template in the domains of interest)
- ☞ Translate and import archetypes in the local editing tool (=Developing an environment to redefine questionnaires using the archetypes repository)
- ☞ Map the terminology to the local medical dictionary
- ☞ Design local templates including the archetypes

d) Author archetypes & templates from local templates

- ☞ Select templates to be shared in different medical areas (Generic Medical Observation ,Oncology, Cardio Vascular Radiology, AVK, etc)
- ☞ Author archetypes from the local templates (=Developing the new archetypes using existing tools (reusing if possible)
 - Selecting a subset of items
 - Expressing constraints
- ☞ Map the related local terminology to reference terminologies (=Linking the archetypes to SNOMED CT)
- ☞ Translate archetypes in OWL and exporting them to the repository

e) Align terminological resources

Define a methodology to align local medical dictionary(ies) to reference terminologies

f) Align templates repositories



Define a methodology to align local templates repository to reference archetypes & templates

g) Perform tests to rate the “added-value” of archetypes and templates defined that way

h) Define queries to see how information is shared between Cardiology and HTA.

3.4 University of Linköping

Activities at LiU within WP26 concern the study of the general problem of storage, retrieval visualisation and navigation of patient data and terminology systems, together with the realisation of software components. A focus of the research is the so-called boundary problem arising when an information model such as EHRcom (prEN13606) is combined with a terminology model such as SNOMED CT. The activities have been carried out in the form of student projects and PhD work.

Archetype Editor

A joint master thesis project aims at the realisation of an archetype editor compliant with the openEHR specification. The editor handles archetypes described with the Archetype Definition Language, ADL, and uses a connection to the UMLS database for convenient handling of terminology bindings inside the archetypes. The archetype editor has therefore the potential to be a useful tool for analysis of the terminology binding problem. The editor will be released within WP26 during the first quarter of 2006.

EHR-Portal for patients

Another joint master thesis project aims to develop a web-based portal for retrieval and visualisation of archetype-based patient information. The portal is aimed to be a test bench for studies of the use of structured patient information. The project also test and evaluate the usefulness of openEHR’s reference implementation and ADL parser in Java. The project will be finalised during the first quarter of 2006.

Archetype repository

In a software engineering project course students have designed and implemented a web-based portal for collecting, storing and retrieving of archetypes. The project is finished and the result is a well-designed and implemented portal with useful base functionality including access and version control. The portal can be used as is for basic archetype handling and also serve as a base for further development for more advanced archetype portals, including visualising and editing of archetypes. An analysis of the portal will be carried out during spring 2006.

Visualisation and navigation of terminology systems

Free-text queries are natural entries into the exploration of complex terminology systems. The way search results are presented has impact on the user’s ability to grasp the overall structure of the system. Hierarchies like the one used in SNOMED CT,



where nodes have multiple parents and other relationship types, makes visualisation challenging.

A prototype has been developed applying well-known methods like “focus+context” and self-organizing layouts from the fields of Information Visualisation and Graph Drawing to terminologies like SNOMED CT and ICD-10. The aim is to simultaneously focus on several nodes in the terminologies and then use interactive animated graph navigation and semantic zooming to further explore the terminology systems without losing context. The prototype, demonstrates how a number of Information Visualisation methods can aid the exploration of medical terminologies with millions of elements and can serve as a base for further development.

A future integration of the visualisation components and the archetype editor will presumably ease the selection of appropriate terms and concepts to be used in archetype authoring. Research regarding visualisation and navigation of longitudinal patient data will be intensified during 2006, partly based on results from the EHR-portal and the terminology visualisation.

4 Collaborative themes on semantic indexing

4.1 The role of Archetypes

An introduction to the archetype concept was given in Deliverable 26.1. The problem that archetypes seek to address is to provide a systematic and potentially standardisable framework for defining and sharing the data structures within an EHR that are needed to represent each distinct clinical information entity, at an appropriate level of granularity. Archetypes are an evolution of results from previous EU-funded projects (in particular, Synapses) and other international (notably Australian) work. The specification is held and maintained by the *openEHR* Foundation, an international not-for-profit company of which one of the Semantic Mining partners (UCL) is a founding shareholder. The archetype approach is now also in the process of being standardised at a European level by CEN TC/251.

Unlike fixed-content healthcare messages that have hitherto been defined to support patient registration, claims, episode and HRG 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 is 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, and forthcoming standards from CEN and ISO (EHRcom: CEN/ISO13606) will enable EHR data to be communicated internationally 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, this approach also risks 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.

It is widely accepted that archetypes, when used to constrain a generic EHR Reference Model such as the *openEHR* RM or that defined by prEN13606, provide a first layer of semantic coherence and thereby offer a basic foundation on which fine grained semantic interoperability can be built. A major activity undertaken in the past year has therefore been to refine the archetype approach, both within *openEHR* and within CEN TC/251, led by UCL. The work undertaken includes the drafting of archetype requirements, a complete and machine process-able UML logical representation for archetypes, refinements to an Archetype Definition Language, and the creation/refinement of archetype authoring tools, validation tools and a first version of an archetype repository (UCL in partnership with Ocean Informatics in Australia). Much of this work is now incorporated into a CEN Enquiry draft of prEN13606-2: Archetype Interchange Specification, and will be taken forward in an accepted New Work Item Proposal: Archetype Knowledge Framework. Much of this work has been shared with HL7 (via the Templates Special Interest Group) and this group is presently reviewing the archetype specifications (starting with requirements) with a view to potential adoption or adaptation for use within HL7 version 3.

Other Semantic Mining work that is at an earlier stage of research includes investigations of optimal ways to bind archetypes to terminology systems, the cross-representation of parts of an archetype in OWL to support reasoning queries, the mapping of archetypes to patient questionnaires and the representation of knowledge constructs within guidelines (via ontologies). These work items have required cross-partner collaboration within WP26.

The next phases of work to be undertaken within WP26 on archetypes are: to explore further the mapping of archetype semantics within OWL and establish clearly the value of achieving this; mapping archetypes within a repository to an ontology (using Protégé) to permit indexing and selection of relevant archetypes; mapping archetype concepts to key terms that can in turn be linked to educational resources so as to permit patients to understand more about their own health through the EHR. A recent archetype-based query component has been implemented as part of a UCL EHR server, and will be extended over the coming year to support quite complex research (data mining) queries.

A key consideration that will determine the value of archetypes in promoting coherence of EHR organisation is the quality and consistency of the archetypes themselves. This requires formal policies and quality criteria, and a governance authority, to ensure that archetype definitions are rigorous, meet appropriate clinical evidence, avoid unnecessary duplication or overlap, and are managed in a regulated way. The set of governance policies and a governance body is being established through the *openEHR* Foundation, and a new EU Framework 6 project (Q-REC) will be developing quality criteria and exploring accreditation mechanisms for archetypes. Semantic Mining WP26 partners are involved in both activities.



4.2 Binding Record Structures and Terminology

The Problem

Medical information systems differ from most other information systems in that the terminology model and information or messaging model are developed separately, e.g. HL7 develops messaging models to be used with terminology models developed by LOINC, SNOMED, ICD9-CM, etc. Similarly, the same set of Archetypes in an *openEHR/CEN 13606* EHR model might be used with any of the above or various national variants or other terminologies.

This poses three problems:

Joint meaning: The meaning of any statement in a message or EHR can only be understood by combining the meanings of the information/message model and the terminology model. This means that for any given Message/EHR model-Terminology pair, there needs to be a transformation of combined statements to a single joint canonical form for querying.

Partitioning and mutual constraints: The scope of the information/message model and the terminology model overlap, so that there are things that might be said in either or both. For example, there is a family history marker in most EHR/Information models and there are also terms for “Family history of...” in most terminologies. (SNOMED has an entire section on ‘context’ to deal with such issues.). This means that there must be a principled partitioning of information between any given terminology system and any given EHR/Record model, represented by mutual constraints.

Negation and Null values: A particular problem of context is negation. Negation has a different logical status from context markers such as “Family history of” because it has a well defined logical semantics that must be respected. Negative findings such as “Absent pedal pulses” or “Loss of consciousness” present special problems.

This problem is becoming more acute with the use of more expressive terminologies such as SNOMED-CT. It is made still more acute by the decision of the UK national Health Service, now the largest user of the combined HL7 version 3 and SNOMED-CT systems that all clinical information should be represented in SNOMED.

Many of these issues focus on the relationship between the SNOMED-CT ‘Context model’ – that portion of the terminology that determines ‘Family history of...’, ‘Past history of...’, ‘Risk of...’ etc and related attributes in the HL7 RIM and EN13606 data model.

In addition to the above three issues of principle, there are many more specific issues relating to each information/messaging model – terminology pair, e.g.



The exact meaning of the HL7 ‘Code’ and ‘Value’ attributes, and whether a finding – e.g. Diabetes – should appear in one, the other, or both, possibly with different wrappings.

Matching and equivalences of complex post-coordinated SNOMED terms and similar information expressed in either EN13606 or SNOMED by relations between multiple statements.

Current Developments and Activities

Currently four are at least three active streams of work internationally on the problem of binding terminology and EHR/Message structures:

The HL7 Terminology working group, which is explicitly concerned with the use of SNOMED-CT within HL7-V3 messages

The work on Archetypes within the OpenEHR/CEN EN/13606 (Electronic Healthcare Record Architecture) development stream

PhD research by Rahil Qamar sponsored by the Semantic Mining project developing mechanisms for Archetype authors to bind their work to specific terminologies and produce accompanying conformance testing.

Contract work with the NHS National Program for IT on a common representation for SNOMED and HL7 v3 in OWL to support a common reference resource and conformance testing.

All three groups involve many of the same individuals. Key documents are available from SNOMED under their Concept Modelling Working Group. The group is semi-open, but the documents are normally considered proprietary and not circulated. These include detailed textual rules for constraints.

State of the Art

For both CEN EN13606/*openEHR* and HL7-v3, the issue of a computable, machine readable binding for terminology and message/EHR models remains an open question. Progress is being made in setting the requirements for such expressions in text form.

The feasibility study of a joint representation in OWL, partly based on work within Semantic Mining, has been very well received in both the NHS and HL7 and larger scale development is under active consideration.



4.3 Negation

The problem

The meaning of negation in clinical records is often ambiguous. In addition, mechanisms for negation exist both in most message/EHR models and in most terminologies. To make matters more difficult still, many terminologies do not indicate negation consistently, so that a single code may indicate “No intracranial bleeding” or “Not pregnant”.

There is no clear stated policy in most systems about the difference between failure to mention a finding and the patient definitely not having the finding. For example, is it safe to assume because there is no mention of diabetes in a patient’s record that they do not have diabetes? When retrieving all patients with kidney disease but without diabetes from a database should patients only patients for which it is explicitly with kidney disease for whom it is explicitly stated that they do not have diabetes be included, or should all patients for whom there is no entry for diabetes be included?

Secondly, whereas most findings are abnormal when present and normal when absent, some such as ‘pedal pulses’ are normally present and abnormal if absent. Such negative findings add particular confusion.

Furthermore, the scope of negation is often ambiguous in the formal representation. Although the natural language is normally clear, the formal representation often makes it difficult whether a given combination of attributes means “No family history of diabetes”, “Family history of non-Diabetes” or “Diabetes but not a family history of diabetes”.

Thirdly, many classification terms, particularly in ICD related systems are of the form “Head injury without intracranial bleed” or “Peptic ulcer with haemorrhage but without perforation”.

Finally, to confuse matters further, until recently, the computationally tractable formalisms used for complex terminologies such as SNOMED or GALEN did not support negation properly. Newer representations such as OWL can support negation properly but are not yet in widespread use.

There are therefore two issues:

Clarity on the intended meaning and scope of negation in any given

1. The distinction between explicit negation and non-mention
2. Simple findings
3. Compound findings using ‘with’/‘without’ or similar constructs
4. ‘Negative findings’ such as ‘pedal pulses’
5. Negation involving context

A formal representation which provides a faithful model of the intended meanings and distinctions.



Intended meaning

Of the above five issues concerning the intended meaning

1. *The distinction between explicit negation and non-mention*: It is generally agreed that this distinction must be maintained in such a way that query mechanisms can query for either depending on the circumstances.
2. *Simple findings* – Present few problems in themselves as to meaning, but see formal representation below
3. *Compound findings* using ‘with’/‘without’ or similar constructs – again provide little argument about meaning but much about representation
4. *‘Negative findings’* such as ‘pedal pulses’ – There is a growing consensus in both HL7 and CEN, that terminology for negative items needs to be divided into two cases:
a) Those in which the negation genuinely means the absence of some condition: e.g. ‘no endocrine disease’, and b) those in which the negation is best regarded as a linguistic artefact best treated as an ‘observable’ (analogous to a laboratory test’ with value of ‘absent’ or ‘zero’. These two cases can be distinguished by two criteria: i) Whether the best interpretation is a) “The patient does not have any X” or b) “The patient does not have some X”. ii) Whether the only possible values are present/absent or whether there might be a scale, e.g. absent, reduced, normal, augmented. In each case, the finding should be treated as a ‘simple’ finding if the first criterion applies; as an observable with a possible value of ‘absent’ or ‘zero’ in the second.

Formal representation

The issue of formal representation is discussed in a draft paper being prepared for preparation. The issues can be focused on the following.

It is never the finding itself that is being negated: i.e. the meaning is not Patient has ‘non-diabetes’

To achieve a uniform representation, it is more effective to ‘wrap’ all findings and observables as aspects of a ‘Clinical Situation’. One can then represent negation easily:

Clinical situation that includes head injury and does not include intracranial bleed.

It should be noted that none of the existing coding systems conform to these criteria, and that the SNOMED-CT context model poses particular problems, since it includes a ‘qualifier’ for negation and then elaborate rules for treating it differently from other ‘qualifiers’. The fact that in SNOMED the ‘context’ is parallel to rather than wrapped around the finding, also means that, even were the negation treated properly in a stronger representation, the result of negating, for example, ‘family history of diabetes’ would be ambiguous and not give the intended results using conventional



semantics. There are straightforward transformations that overcome these problems, but they have not yet been accepted nor implemented.

4.4 Linking questionnaires in use with archetypes based on a generic model such as CEN EHRCOM or HL7 RIM

For hospitals having already implemented the EHR, where hundreds of models of questionnaires have been designed and are used in daily practice, the issue is to propose methods and/or tools for:

- ☞ designing new questionnaires making use of shared validated professional/business objects models i.e archetypes (as they are or after specialization or customization) based on generic EHR models such as *openEHR*, CEN EHRcom or the HL7 RIM.
- ☞ contributing to archetype authoring (proposing existing questionnaires and/or parts of questionnaires to be candidate to become archetypes)

The collection of questionnaires used at large scale represents an interesting basis for the design of archetypes that could be consensual at least within a domain of application such as hospitals, hospital groups, healthcare regional networks.

- ☞ sharing documents and/or questionnaires with others hospitals relying on archetypes linked with reference terminology/ontologies maintained into a repository
- ☞ developing and sharing Decision Support modules relying on questionnaires and concepts

In order to address these issues there is a need to link the professional/business objects models currently based on proprietary data models from Electronic Healthcare Record Systems in use and the available generic EHR models.

Current developments and activities

HEGP

The component-based information system of HEGP (Georges Pompidou University Hospital, “Assistance Publique - Hôpitaux de Paris” (AP-HP) includes computerized orders, results and notes. A decision support system for computerized ordering of medications is under development.

A hospital policy was voted in May 2004: to enter computerized notes in the EHR as structured questionnaires for most encounters. For hospitalisations, reports built from



computerized notes will be sent to the referring physician before at least 8 days after outcome. The migration of pre existing specialized records into the structured questionnaires of the medical record component is on the way.

This experience has already given evidence or problems to be solved like the distinction between question concepts and response concepts, how to infer response concept from binary (yes/no), numerical or free text values and how to link both sets to terminologies.

The current issue is the progressive sharing of documents with affiliated institutions, practitioners and patients.

UCL

UCL is a joined founder of the *openEHR* foundation and works closely with its co-founding partners Ocean Informatics. In particular, The Ocean Archetype Editor is a tool that supports the authoring of archetypes as part of the *openEHR* initiative and the CEN EHR standardisations. UCL has done a lot of work on the development of archetypes. It brings significant experience on the design and use of archetypes in live clinical EHR settings. A recent project incorporates placeholders on the integration of medical knowledge (in the ontology domain) and protocol services.

CEN TC251

As part of WG1 activities, after the previous releases of Pre- European Normative Vouchure (1995) and European Normative Vouchure (1999) a European Normative document is under development to deal with Electronic Healthcare Record Communications (EHRcom).

This document includes a generic model of Electronic Healthcare Records supporting European requirements for EHR management and communication. It also defines precisely the archetype paradigms and provides the means to design and use archetypes within EHR systems. Some archetypes are also to be referenced within the document together with how to address the underlying security/confidentiality issues that inevitably emerge when trying to address the issue of EHR interoperability.

HL7

HL7 provides many threads of activities in this field, based on the RIM generic model whose scope is broader than EHRs and on the CDA Clinical Document Architecture. Numbers of RIM and CDA based professional/business objects are being designed.

HL7-France

Within a HL7-France specific working group, templates of RIM-based CDA are designed in two areas : general medical observation and pathology reports in oncology. They are professional/business objects models including other reusable professional/business objects like “allergies”, “antecedents”, etc.



These models of CDA are to be used for cross-healthcare enterprise sharing in two frameworks : the French personal medical EHR (DMP project from the ministry of health) and the Communicating EHR in Oncology (promoted by the Cancer Institute).

ARTEMIS

Within the Artemis project, ebXML Registry semantic constructs were used for annotating, storing, discovering and retrieving archetypes.

To realize archetype based interoperability, the healthcare systems need to identify the accurate existing archetypes based on their semantics; In order to achieve this, archetypes are annotated using ontologies and this annotation is used for archetype retrieval; The authors also described how archetype data can be retrieved from clinical information systems by using ebXML Web services.

Since healthcare systems exploit different ontologies based on the domain knowledge through different standard bodies like HL7, CEN TC/251, ISO TC/215, interoperability is not easily achieved. Within the Artemis project a Web Service architecture, adapted from the Semantic Web solutions has been developed. This architecture does not propose globally agreed domain models (HL7 RIM, CEN models...) but a mediator component enabling healthcare institutes to reconcile their semantic differences. The mediator component uses ontologies based on prominent healthcare standards (HL7, CEN) as references to facilitate semantic mediation among involved institutes.

MEDIQ

MEDIQ (Danish consultancy company) is involved in projects developing clinical guidelines, care management plans and medical decision support systems based on reusable semantic building blocks. As part of their work professional/business EHR objects are being designed.

State of the art

For healthcare systems to exchange information in an interoperable manner, they need to discover the professional/business objects (archetypes and templates (CMET, etc)) communicated and their associated semantics.

ebXML Registry, through its semantic constructs, provides an efficient medium to annotate, store, discover and reuse of archetypes. More over, using Web services to support interoperability of hospital systems relying on different standard domain models (HL7, CEN) and different domain ontologies has being reported in the literature.

Currently, describing the semantics of Web services is a very active research area. OWL-based Web service ontology is a comprehensive effort designing an upper ontology for Web services. The use of artificial intelligence techniques for discovering services through OWL-based languages is also addressed in the literature.



More over, the emerging P2P paradigms could be an opportunity to improve semantic interoperability, providing a solution to achieve semantic agreements. It is argued that establishing local agreements is a less challenging task than establishing global agreements by means of globally agreed schemas or shared ontologies. Once such local agreements exist, through the “semantic gossiping” process proposed, global agreements can be achieved in a P2P manner.

SemanticWeb EnabledWeb Services (SWWS) architecture is a basis to develop solutions to provide semantic web services for archetypes sharing. It is largely stated in literature how ebXML registries can be enriched with Web service semantic.

Supporting interoperability between hospital systems relying on different standard domain models (HL7, CEN) and different domain ontologies require the use of ontology mapping components.

As regards the interoperability of points of view, many works propose knowledge integration solutions based on ontologies.

According to Wache, three approaches for the integration of various knowledge are generally distinguished :

- The single ontology approach uses a global ontology providing a shared vocabulary for the specification of the different sources.
- In the multiple ontologies approach, each source is described by its own ontology and an additional representation formalism defining the inter-mapping ontology is provided.
- Finally the hybrid approach combines the two previous ones. Each source is described by its own ontology built upon one global shared vocabulary. The main point is how the terms of the source ontologies are described by the primitives of the shared vocabulary.

The technical solutions developed to allow managing multiple ontologies can be classified according to the operations they can handle (difference, mapping, transformation, versioning management, etc.) or according to the local ontologies one has to deal with. One of the difficulties of multiple ontologies approach and the "hybrid" one is the definition of the mapping. There are several environments making operations of mappings (like the PROMPT suite, Open Source software). The environment to be used must be chosen according to the task envisaged and to the local ontologies which one has to deal with.

Intended contribution by the group

- Collect questionnaires in the following medical areas : Generic Medical Observation ,Oncology, Cardio Vascular Radiology, AVK, etc



- Author archetypes from the collected questionnaires using Protégé and OWL in collaboration with the other semantic mining teams
- Define the features that would enable a satisfactory management of the distinction between question concepts and response concepts, how to infer response concept from binary (yes/no), numerical or free text values and how to link both sets to terminologies.
- Design scenarios describing the interactions between a clinical decision support system (CDSS) and the HEGP EHR system. Establish a list of concepts which could be present and accessible to allow the CDSS to make use of the patient medical information stored within its EHR
- Evaluate the needs for archetypes repositories and organisation and describe the services they should offer within various domains of applications

4.5 Applying ontologies to support interoperability between guideline management systems and electronic health records

The problem and general approach

Nowadays health-care provision is more and more characterised by the enactment of clinical guidelines, also in the daily practice. Fundamental is their role – for instance – in the prevention of medical mistakes, acting basically as a reminder for things to be done, questions to pose, data to be taken and so on.

Although representing the kind of knowledge needed to enact clinical guidelines may seem the only issue, it is not neglectable to solve the crucial problem of interfacing in real-time the guideline management system and the electronic health record. Both systems have to speak the same language, and they usually don't. Ontologies are the "glue" needed to bridge the two worlds.

Present work in the field

Apart from its definition in the philosophical context - where it refers to the subject of existence - ontology in the medical context is "a partial specification of a conceptualization". There are ontologies in many specific domains of medicine (some of them are reported in www.openclinical.org). On the other hand, many efforts have been devoted on the design of so-called "top-level" ontologies, i.e. domain independent ontologies, like the DOLCE foundational ontology . What still lacks is a useful "reference ontology" in the field of medicine, i.e. a domain ontology with general medical concepts.



Future direction

The Laboratory of Applied Ontologies of CNR in Rome has been designing a clinical medicine reference ontology whose aim is to act as an "ontology mediator" in order to ensure semantic consistency among heterogeneous health-care information systems, including the EHR and the guideline management system. This reference ontology currently consists of 121 concepts linked by 17 different relationships. It has been implemented by means of the RACER description-logic and can be browsed by the Protégé tool.

5 Conclusion

Many of the issues being tackled in WP26 are complex conceptual problems, requiring deep understanding of the semantic challenges involved in systematically representing diverse and evolving clinical concepts and data structures. The partners and their research teams have progressed considerably in both individual site activities and inter-site collaborations. It is difficult to capture in the form of a report the growing richness of this mutual understanding, or to project clearly how these innovative threads of work will contribute to next-generation solutions to the semantic indexing of EHRs. It is clear, however, that the work being done by the WP26 partners is being recognised internationally as a set of strong contributions. From the starting point of a loosely coupled set of university sites, the WP26 team is gaining 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.