Ontology-driven Development of a Clinical Research Information System

John J Chelsom
Centre for Health Informatics, City University, London, UK.
john.chelsom.1@city.ac.uk

Ira Pande
Nottingham University Hospitals NHS Trust
ira.pande@nuh.nhs.uk

Ron Summers
Systems Engineering Innovation Centre, Loughborough University, Loughborough, UK.
R.Summers@lboro.ac.uk

Ian Gaywood
Nottingham University Hospitals NHS Trust
ian.gaywood@nuh.nhs.uk

Abstract

Clinicians at Nottingham University Hospitals NHS Trust have developed a clinical information model that enables detailed patient data to be gathered in clinics, stored in a patient record, linked with samples in a Biobank and then used to identify and select patient cohorts for research studies. The objective has been to enable the assembly of cohorts based on any plausible combination of clinical and laboratory features. The model is represented as an ontology, coded in OWL/XML and is itself built upon an ontology-based information architecture. The model can be used to generate the runtime configuration and operational data structures for a clinical information system, which has been implemented using an open source toolkit developed at City University, London.

1. Introduction

Collecting patient data for use in clinical research is a crucial activity that directly enables the testing of hypotheses central to all research studies. Rheumatology specialists at Nottingham University Hospitals NHS Trust developed their own clinical information system that enabled them to gather data during routine clinic encounters and to search the data to identify cohorts of patients for their clinical studies.

Although it met the needs of the clinicians who developed it, this system had a number of disadvantages, which are common in local, clinician-led developments:

- it was not built on a scaleable enterprise technology platform
- it made little use of internationally recognized open standards
- it was not supported by the local IT services, or by the wider clinical faculty

These issues were highlighted when the Trust began the deployment of a Biobank and was looking at ways to combine data in the Biobank with routinely gathered clinical data used in research studies. Although the system used in Rheumatology met many of the clinical and research requirements, it was not scalable to support the organization at an enterprise level.

At the same time, the original developers were looking at more sophisticated ways to classify diagnoses and use these as a key tool in identifying patient cohorts. They found that ICD-10 coding [1] did not have sufficient scope to code to the level of detail required and did not support the hierarchical classifications required. Similar issues have led to the development of localized variants of ICD-10 and the onward push towards ICD-11 [2].

Similarly, SNOMED CT coding [3] did not provide a consistent way to code diagnoses in accordance with the requirements of local clinicians, which is also an issue observed and documented elsewhere [4].

Hence the search began for ways to support the local requirements for classification of diagnoses and the selection of patient cohorts based on data gathered in clinic encounters, whilst remaining aligned to internationally recognized standards for clinical coding and data representation.

2. Objectives

The primary objective has been to provide a solution to the clinical need for a coding system for research studies, meeting the requirements to:

- identify patient phenotypes to the required degree of detail
- produce untainted patient cohorts
- include or exclude individual disease characteristics
- include or exclude specific treatment details,
- record and assess outcomes
- enable search in real time, to answer any plausible research question

The resulting scheme for the classification of diagnoses and the core data sets with characterize them has been named ORCHID. An additional design requirement is for the mapping of the terms in each ORCHID-based clinical research information system to terms from existing clinical coding systems such as ICD-10 and SNOMED-CT.

The ORCHID model has been used to create an open source based solution that is extensible to more than one clinical specialty, with the potential to be used throughout the Trust, as well as at national and international levels of engagement.

3. Clinical Coding and the ORCHID Model

The ORCHID classification of Diagnoses is consists of three levels, represented as an ontology and maintained using the Protégé tooling [5].

The isTypeOf relationship is used to classify individual Diagnoses in ORCHID. The three levels of hierarchical classification of Diagnosis are defined by restrictions as described below.

Level 3 Diagnosis. This is a detailed diagnosis at the lowest level of classification. A diagnosis at this level has an isAssociatedWith relationship with one or more Specialties and one or more isTypeOf relationships with Level 2 or Level 1 Diagnoses.

Level 2 Diagnosis. Lies in the intermediate levels in the hierarchical classification. A diagnosis at this level has at least one isTypeOf relationship with other another Diagnosis and at least one hasType relationship with another Diagnosis i.e. it is both a parent and a child in the hierarchy.

Level 1 Diagnosis. The top of the hierarchical classification i.e. it is a parent, but not a child in the hierarchy. A diagnosis at this level has at least one hasType relationship with a Level 2 or Level 3 Diagnosis and no isTypeOf relationships with any other another diagnosis (although it is not necessary to specify this axiom in order to classify a Level 1 Diagnosis).

Because of the Open World Assumption, a Level 3 Diagnosis is classified through its association with a Specialty; the Level 2 and Level 1 Diagnoses can then be classified based on their position in the classification hierarchy, as defined by the hasType/isTypeOf axiom.

Although any Diagnosis is classified as being at Level 1, Level 2 or Level 3 in ORCHID, there may be multiple Level 2 nodes encountered when traversing a branch from a root node (Level 1) to a leaf node (Level 3). In some instances a Level 3 Diagnosis is connected directly to a Level 1 parent, without any intermediate Level 2 nodes.

When recording the diagnosis for a patient, a single Level 3 Diagnosis is used; when specifying criteria for assembling patient cohorts a Diagnosis can be specified from any level.

A Core Data Set is a set of findings associated with a Level 3 Diagnosis which describe Characteristics of the Diagnosis that are of interest to clinicians and/or researchers.

Each Characteristic takes the value of yes, no or unknown for a particular patient. Once a Characteristic has been set as yes, then it is remains set as yes . There are no restrictions/relationships between Characteristics in a Core Data Set - i.e. each Characteristic is an independent finding.

However, the value of a Characteristic may be derived on the basis of other findings related to the patient.

4. The Wider Ontology Model

Whilst the ORCHID model for Diagnosis and Core Data Sets is useful for research purposes, it is not sufficient to define a complete clinical information system.

A larger ontology uses core concepts from the ISO 13606 model of EHR [6], to create a data dictionary of clinical findings that can be combined together into sets that define the clinical data entry forms and summary data views required in a clinical information system.

This model includes the Composition, Section, Entry and Element components of IS0 13606. Folders are not used and Clusters are only formed dynamically in views of longitudinal data sets over a specified time period.

This wider ontology also supports preferred terms and synonyms for any concept and clinical coding from SNOMED-CT and/or ICD-10 that can attach to any concept.

5. Ontology-driven Development

The ontology is used to drive the configuration and runtime data structures of the clinical information system, which has been implemented using open source software components assembled as part of the Open Health Informatics research programme at City University, London.

The four-stage approach to system design follows the conceptual approach of the openEHR system of archetypes [7].

The basic ontology model is predefined as a clinical Information Architecture; this includes the ORCHID model and the additional components based on ISO 13606.

Clinicians then use the Protégé tool to create the specific Information Model for their domain, including specification of the Diagnoses, Core Data Sets, Data Dictionary (of clinical findings - observations, lab results, procedures, medications, etc), data entry forms (corresponding to a Composition in the ISO 13606 model) and summary views (which are views and/or reports based on search of the recorded clinical data).

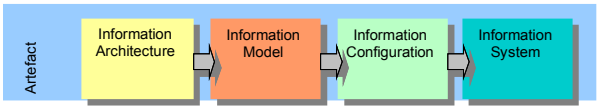

Figure 1. Four Stages of System Design

The common standard used throughout the architecture, model and runtime system is XML [8]. A series of automated transformations (XSLT [9]) drive the generation of the Information Configuration and persistent data structures for the runtime Information System, which uses HL7 CDA [10] as its main representation of clinical data.

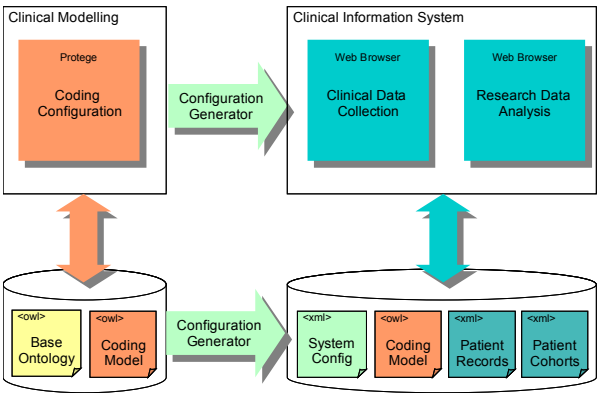

Figure 2. Generating the Runtime System

The database created as a result of this system design contains XML documents in a variety of different vocabularies: HL7 CDA for clinical data, OWL/XML for the ISO 13606-based data dictionary and the ORCHID classification of diagnoses.

The implementation of this database uses an open source native XML data store that can be queried using the standard XQuery language [11].

7. Conclusions

The use of an ontology architecture and model has enabled clinicians to develop their own clinical information system using open standards and scalable enterprise architecture, whilst focussing on the clinical requirements for data gathering and research.

Using the ontology-driven approach allows clinicians to incorporate their own conceptual organization of clinical data and diagnoses, whilst retaining compatibility with recognized standards for data representation and clinical coding.

8. References

[1] International Classification of Diseases (ICD). World Health Organisation. Available at: http://www.who.int/classifications/icd/en/

[2] Jetté, Nathalie, et al, “The Development, Evolution, and Modifications of ICD-10: Challenges to the International Comparability of Morbidity Data”, *Medical Care*, December 2010 - Volume 48 - Issue 12 - pp 1105-1110 doi: 10.1097/MLR.0b013e3181ef9d3e.

[3] Systematized Nomenclature of Medicine-Clinical Terms (SNOMED CT). International Health Terminology Standards Development Organisation (IHTSDO). Available at: http://www.ihtsdo.org/snomed-ct/

[4] James E. Andrews, Rachel L. Richesson and Jeffrey Krischer, "Variation of SNOMED CT Coding of Clinical Research Concepts among Coding Experts", *J Am Med Inform Assoc.* 2007;14:497–506. DOI 10.1197/jamia.M2372.

[5] Protégé Knowledge Acquisition System. http://protege. stanford. edu

[6] ISO 13606-1:2008 Health informatics - Electronic health record communication - Part 1: Reference model. Available at: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=40784

[7] Beale T, Heard S: openEHR Architecture: Architecture Overview. openEHR specification, 2007. Available at: http://www.openehr.org/releases/1.0.1/architecture/overview.pdf.

[8] World Wide Web Consortium. "Extensible Markup Language (XML) 1.1 Recommendation." Available at: http://www.w3.org/TR/2004/REC-xml11-20040204/

[9] World Wide Web Consortium. "XSL Transformations (XSLT) Version 2.0." Available at: http://www.w3.org/TR/xslt20/

[10] Dolin RH, Alschuler L, Boyer S et al. "HL7 Clinical Document Architecture, Release 2." *J Am Med Inform Assoc.* 2006;13(1):30-9.

[11] World Wide Web Consortium. "XQuery 1.0: An XML Query Language." Available at: http://www.w3.org/TR/xquery/