Heterogeneous Medical Data Integration in the Health e-Child

Kalyani P. Ugale
Student, CSE Department, SIPNA COET Amravati, India.

A. D. Gawande
HOD, CSE Department, SIPNA COET Amravati, India.

Abstract – The approach advocated in this paper surrounds the provision of an Integrated Data Model plus links to/from ontologies to homogenize biomedical (from genomic, through cellular, disease, patient and population-related) data in the context of the EC Framework 6 Health-e-Child project. Clinical requirements are identified, the design approach in constructing the model is detailed and the integrated model described in the context of examples taken from that project. Pointers are given to future work relating the model to medical ontologies and challenges to the use of fully integrated models and ontologies are identified. In particular the requirement of integrating multiple, potentially distributed, heterogeneous data sources in the medical domain for the use of clinicians has set challenging goals for the health grid community.

Index Terms – Medical informatics, data models, ontology, data integration.

1. INTRODUCTION

The Health-e-Child (HeC) project [1], [2] is an EC Framework Programmed 6 Integrated Project that aims to develop a grid-based integrated healthcare platform for pediatrics. It is hoped that using this platform biomedical informaticians will integrate heterogeneous data and perform epidemiological studies across Europe. The resulting Grid-enabled biomedical information platform will be supported by robust search, optimization and matching techniques for information collected in hospitals across Europe. In particular, pediatricians will be provided with decision support, knowledge discovery and disease modelling applications that will access data in hospitals in the UK, Italy and France, integrated via the Grid. For economies of scale, reusability, extensibility, and maintainability, HeC is being developed on top of an EGEE/gLite$_1$ based infrastructure that provides all the common data and computation management services required by the applications.

This paper discusses some of the major challenges in bio-medical data integration and indicates how these will be resolved in the HeC system. HeC is presented as an example of how computer science (and, in particular Grid infrastructures) originating from high energy physics can be adapted for use by biomedical informaticians to deliver tangible real-world benefits. The HeC project aims to develop a prototype system which will demonstrate the integration of heterogeneous biomedical data sources over a grid linking multiple hospitals in Italy, the UK and France. In this integration, particular emphasis is put on distinguishing features such as universality of information, person-centricity of information and universality of application leading to the main tenet of the HeC effort: "the integration of information across biomedical abstractions, whereby all layers of biomedical information (i.e. genetic, cell, tissue, organ, individual and population layer) are vertically integrated to provide a unified view of a child's biomedical and clinical condition" [2]. One essential element required for the integration of data across multiple layers of biomedical information is the provision of suitable models for data and information.

2. RELATED WORK

To support the HeC objectives, a set of models for representing biomedical information needs to be put in place. As a prerequisite relevant domain knowledge should be reused and at the same time the knowledge base should be aligned to existing data models (such as patient record formats, examination templates, and clinical protocols). Given the heterogeneity and diversity of the large quantity of data that makes up a complete medical record, it is far from easy to capture and align even identical concepts. By collecting common terms that appear in each of the protocols, areas of overlap have been identified. This in turn allowed us to begin to identify key concepts and the relationships between them. Where a form showed data that did not appear in other protocols a decision had to be made as to its inclusion within the model. This is a difficult balancing act since adding everything may produce a structure with a number of redundant sections. On the other hand, there is the danger of missing a piece of data that may prove essential later on. The iterative process of constant review and updating of the models allowed us the flexibility to refine our methods over time. This gradual refinement eventually gave rise to a settled, if not fully integrated, set of components. Finally, the integration of these components within a coherent structure was the last stage in building a conceptual model of HeC domain.

As an initial modelling step, a group of key conceptual entities that inhabit the domain space has been identified including person, hospital, family tree, demographics and several others. The entities chosen represent important roles in the system and also provide an efficient means of sub-dividing complex conceptual relationships into more manageable sections of the overall model. Once such overlapping aspects have been identified and captured in the model, the disease-specific concepts that can add useful extra information should be considered. It is clear that creating new concepts for each and every difference that exists between medical domains would make the conceptual model unduly complicated and perhaps even unworkable. In order to avoid this potential problem it was decided that a set of common medical terms could be produced and that each should serve several roles in the model. In effect, our results amount to identifying reusable patterns of the medical domain which fit the scope of HeC with most also being applicable well beyond. For instance, much of the clinical data that form the basis of the patients' assessments are acquired by various measurements and represented as physical quantities.

Most of these quantities are fully defined in terms of a number (the measurement value) and a suitable measurement unit. Without the (possibly implicit) knowledge of the unit, the quantities cannot be interpreted or compared. Which units are suitable for the attribute of the quantity is determined by its dimension: weight can be measured in kg or pounds but not seconds, etc. The analysis model of the physical quantities is shown on Figure 1.

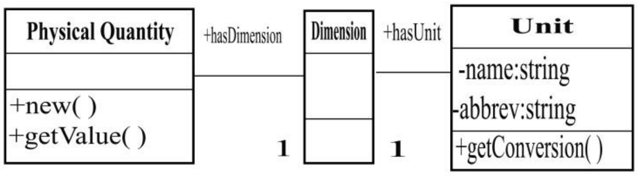

Figure 1 Physical Quantities

A new physical quantity is created using a numerical value and an (existing) Unit instance where the latter automatically establishes the dimension of the quantity (e.g. Joe's height measured [145,<cm>]). Then the measurement can be queried using other units of the same dimension, in this case one can ask for the value of Joe's height in centimetres or inches. For this to work, units of the same dimension must get converted to each other; this is also represented on the class diagram (see Figure 1).

As physical quantities do not cover all attributes that we need to model in the medical domain similar patterns have been identified for classifications (i.e. corresponding attributes take values from a finite set of discrete possible values, for instance, *Yes/No/Unknown* or *Mild/Moderate/Severe*), clinical observations (for instance, a collection of observations of medical signs with various), free-text annotations etc.

The following section presents how the conceptual models can be harnessed to create a HeC integrated data model and demonstrates how the features captured in the conceptual models are reflected in the data model.

3. PORPOSED MODELLING

One crucial factor in the creation of integrated heterogeneous systems dealing with changing requirements is the suitability of the underlying technology to allow the evolution of the system [4]. A 'reflective' system utilizes an architecture where implicit system descriptions are instantiated to become explicit so-called "metadata objects" [5]. These implicit system aspects are often fundamental structures and their instantiation as metadata objects serves as the basis for handling changes and extensions to the system, making it somewhat self-describing. Metadata objects are the self-representations of the system describing how its internal elements can be accessed and manipulated.

The ability to dynamically augment and re-define system specifications can result in a considerable improvement in flexibility. This leads to dynamically modifiable systems which can adapt and cope with evolving requirements [6]. In this way we can separate the system description in terms of metadata from the particular physical representations of the data and thereby promote ease of integration and querying of the data whilst retaining the ability for the semantics of the system to evolve.

The complexity which arises from the use of diverse distributed data sources in HeC and the anticipated evolution of its medical information led us to the decision to adopt a modelling approach which heavily relies on metadata. In addition the model is enhanced with a semantic layer to facilitate the semantic coherence of the integrated data and to allow linking and reuse of the external medical knowledge. The metadata reveals the structure of the underlying heterogeneous medical data allowing consistent queries across populations of patients and disease types. The semantic layer adds knowledge to this metadata thereby facilitating the resolution of queries that bridge between related concepts. It is this combination of descriptive metadata with system semantics that provides the HeC data model with the ability to be both reactive in terms of the queries generated by user applications and to have the richness to enable integration across heterogeneous data sources. The resulting HeC Integrated Data Model (IDM) constitutes the structures for the representation of data, information and knowledge for the biomedical domain of the HeC (see Figure 2).

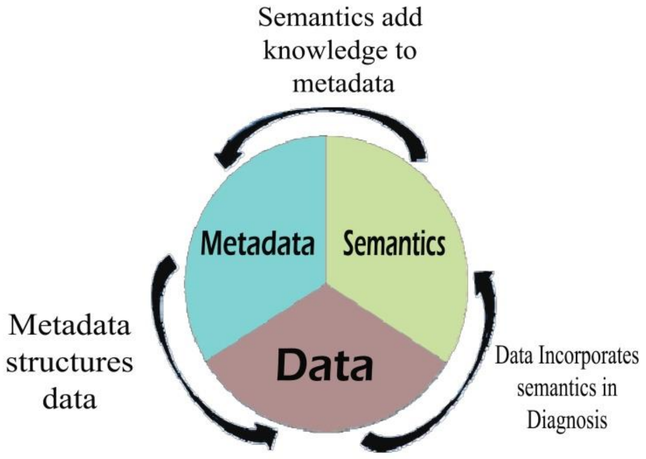
Semantics add knowledge to metadata
Metadata
Semantics
Data
Metadata structures data
Data Incorporates semantics in Diagnosis

Figure 2 A high-level overview of the HeC IDM

4. RESULTS AND DISCUSSIONS

However, there are many ways of storing the same data, and at this simple level there is no semantic annotation, although this method does suffice for getting data quickly into the database. In our model there are two places where semantic tags can be applied – firstly from the metadata, annotating the description with concepts relating to the nature of the data being measured (e.g. a heart rate measurement linking to the concept of heart rate) and secondly from the data, to instantiate concepts in patient data (e.g. a patient having arthritis in the elbow, requiring a link to the concepts of ‘arthritis’ and ‘elbow joint’).

Strings extracted from the form elements can be queried for in ontologies, but this can only ever be a semi-automatic matching process. Although the machine can present best matches, a qualified person must select the correct ones. For querying purposes the semantic part of the database must be sufficiently complete to provide the correct concepts for all of these cases. It would be desirable to link concepts together to a certain degree, but the extent to which this would be useful is, as yet, unknown.

The greater the degree of specification in the semantic section, the more powerful the queries that can be written against it, but performance can suffer if it becomes too large. Data migration into a better designed metadata structure is more difficult since text fields from the form must be matched with imported concepts.

During the population of the description side of the IDM, overlap between semantic and metadata from other areas of the model produces a useful web of links between different fields (see Figure 4).

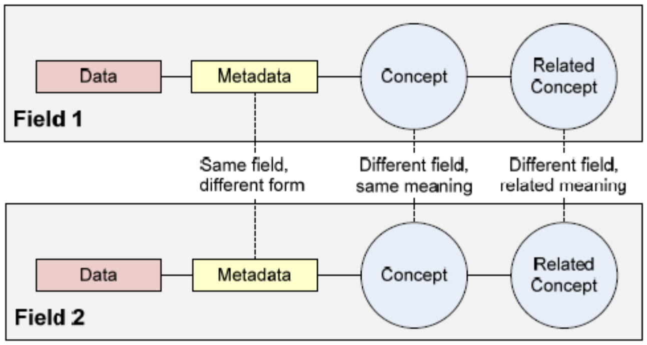
Data
Metadata
Concept
Related Concept
Field 1
Same field, different form
Different field, same meaning
Different field, related meaning
Data
Metadata
Concept
Related Concept
Field 2

Figure 3 The significance of shared metadata and semantic data

In cases where fields are reused between forms, the overlap will occur in the metadata. For different fields that describe the same concept, they will each link to the same concept. Similarly several metadata elements can be annotated with the same concept (for instance, “heart rate” measurement and “heart murmur” symptom can be related to the “Heart” concept) capturing the fact that, though the fields appear on the different forms, they describe the same concept. Finally, different metadata annotated with different concepts can share a (set of) common related concept(s), facilitating the semantic coherence of the data from different forms.

5. CONCLUSION

The data model ensures that all the information recorded can be stored and reused. The metadata model ensured the abstraction required to integrate pieces of data into a coherent whole and to define sufficient description of data elements so that they can be properly interpreted and compared; it is the primary means for data creation and access. To make full use of the information that is captured, one needs sufficient formal semantics associated with the data. To exploit the full potential of the IDM the metadata and semantics should be populated. The semantic structures add flexibility and descriptive power to the IDM, and, as a consequence, the semantic annotation of clinical protocols becomes crucial. Semantic annotation requires the tagging of data with conceptual knowledge which can be formally represented as, for example, an ontolog.

REFERENCES

[1] The Information Societies Technology Project: Health-e-Child Framework6 Integrated Project EU Contract IST-2004-027749 Description of Work.

[2] J. Freund et al., “Health-e-Child: An integrated biomedical platform for grid-based pediatrics”. In Proc of HealthGrid 2006, volume 12 of Studies in Health Technology and Informatics, pages 259-270, Valencia, Spain, 2006. PMID: 16823144.

[3] A.Anjum et al., “The Requirements for Ontologies in Medical Data Integration: A Case Study”. Proceedings of the 11th International Database Engineering & Applications Symposium (Ideas2007). IEEE

Press ISBN 0-7695-2947-X, pp 308-314. Banff, Canada September 2007

[4] F. Estrella et al., "Handling Evolving Data Through the Use of a Description Driven Systems Architecture". Lecture Notes in Computer Science Vol 1727, pp 1-11 ISBN 3-540-66653-2 Springer-Verlag, 1999

[5] F. Estrella et al., "Meta-Data Objects as the Basis for System Evolution". Lecture Notes in Computer Science Vol 2118, pp 390-399 ISBN 3-540-42298-6 Springer-Verlag, 2001

[6] B. Foote and J. Yoder, "Metadata and Active Object Models", Fifth Conference on Pattern Languages of Programs (PLOP 98), Illinois, USA, August 1998.

[7] K.Munir, M. Odeh & R. McClatchey "Ontology Assisted Query Reformulation using Semantic and Assertion Capabilities of OWL Domain Ontologies". Submitted to the 25th British National Conferences on Databases (BNCOD08). Cardiff, UK July 2008.

[8] S. Zillner et al. "Semantic visualization of patient information". Accepted for publication at CBMS 2008.

[9] R. Berlanga et al., "Semantic Annotation of Medical Protocols in Health-e-Child: A Case Study". Accepted for publication at CBMS 2008.

Authors

Kalyani P.Ugale

Student, CSE Department, SIPNA COET Amravati, India.