Evolving Health Informatics
Semantic Frameworks and Metadata-Driven Architectures

Jim Davies*, Jeremy Gibbons*, Steve Harris*, and Denise Warzel†

* *Oxford University Computing Laboratory*

† *NCI Center for Biomedical Informatics and Information Technology*

Advances in technology allow us to communicate large amounts of information, almost instantaneously, between any two points on the globe. Advances in analysis and imaging techniques, and progress in genomic, proteomic, and metabonomic science, allow us to obtain detailed information about the health of an individual. Advances in the computerisation of social and business infrastructure allow us to obtain similarly detailed information about other aspects of our lives. The automatic integration of this data, based upon a computable representation of its meaning or semantics, will revolutionise both medical and clinical research, and the impact on healthcare delivery will be dramatic. Not merely in terms of personalised medicine, informed by the new biology, but also in the very nature of national and international healthcare systems. In this position paper, we briefly explore two problem domains, explain the importance of semantic frameworks and the software engineering opportunities that they present, and discuss current progress in two linked initiatives.

1 Problem domains

Clinical studies Systematic studies of large populations are expensive, but essential to progress in medicine. We might expect that the data gathered by studies addressing the same disease, or simply asking similar questions, could be usefully combined: the resulting, larger population may be all that is needed to detect a pattern or prove a hypothesis. For *Tamoxifen*, data from 400 different trials allowed researchers to identify the subset of the population responsive to the drug, and indicated the optimum period of treatment. This evidence changed clinical practice in the UK, and reduced mortality from operable breast cancer by 24%.

However, such integration is often impossible: different studies make different observations, record the same information in different ways, or leave the precise meaning of the data unclear. A detailed, computable account of the meaning of data, together with improved coordination in experimental design, can guarantee successful, automatic integration, with significant benefits for trials operation and analysis. For example, automated integration could have allowed earlier assessment of adverse effects in the case of *Vioxx*, and could have reduced the delay between study completion and subsequent licensing in the case of *Herceptin*.

Infectious disease control Increasing travel, global trade, and changing climate have meant that infectious diseases are emerging, and spreading, faster than ever before. The *World Health Report 2007* argues strongly that effective data sharing across organisational and national boundaries is essential to a responsive national health system, and hence to global public health security. Recent examples of significant threats include: *SARS*, for which the economic impact in Asia, excluding that of human sickness and death, is estimated at $30 billion; pandemic influenza, in particular, strains of avian flu with death rates of up to 58%; *BSE*, for which the economic impact in the UK was in excess of $39 billion; multi-drug resistant tuberculosis, now confirmed in 37 countries; and regular epidemics of *meningococcal serogroup A disease* in sub-Saharan Africa.

The *International Health Regulations 2005* address exactly this issue. By June 2012, each country should have a system in place for timely and reliable notification, consultation, and verification; in most cases, a period of 24 hours is specified. It is unlikely that this can be achieved without advances in methods and tools for automatic, metadata-driven data sharing and integration. And of course there are also many local outbreak issues (such as food poisoning) in which time is an issue; these may be less dramatic, but are an ongoing 'real-life' issue.

2 Semantic frameworks

The basis of effective data sharing and re-use is a common semantics: we must agree upon the meaning of the data being shared. Natural language is sufficient for such a semantics only when the concepts are straightforward, the community is small or homogeneous, and the period of time over which understanding must be maintained is short. For complex problems, large, diverse communities, or long-lasting collaborations, a more formal approach is required.

At a bare minimum, we require something equivalent to a dictionary, providing standard definitions for the community to use. A definition might explain the meaning of a word, the purpose of a piece of data, or the intended interpretation of a message or action. An informal explanation may be supplemented by precise statements categorising the elements being defined, formalising relationships between them, and explaining how they may be used in combination.

Information-driven health requires a significant degree of formalisation and *computerisation* of semantics. The size of the community, the volume and variety of information, the rate of evolution, and the importance of documentation make it essential that the semantics can be accessed, maintained, and incorporated into delivered systems without extensive, costly, error-prone manual intervention. A practical, semantic framework can be defined in terms of constructs at three different levels:

- *terminology services* provide access to a collection of defined terms, structured in a way that suits one or more possible applications. An exemplary implementation is the NCI Thesaurus, a reference terminology that provides rich textual and ontological descriptions of more than 60,000 key biomedical concepts, linked to the National Library of Medicine's Unified Medical Language System.
- *metadata registries* contain collections of structured 'metadata elements', each describing a measurement or observation. Each element may be related to one or more terms in the underlying terminology, and additional semantic information is provided by informal explanations of intended purpose and an association with a domain of possible values.
- *model repositories* present re-usable models for the definition of information artifacts, such as database schemas, service descriptions, forms, queries, mappings, transformations, and reports. The fields on the forms, the service descriptions, and the columns on the spreadsheets are defined, and given computable semantics, by linking them to the elements in a metadata registry.

Examples of existing, partial semantic frameworks include: the NCI Cancer Data Standards Repository (caDSR), the Metadata Online Registry (METeOR) for the Australian Institute of Health and Welfare, the Canadian Institute for Health Information, the National Information Exchange Model (NIEM) from the US Departments of Homeland Security and Justice, and the US Health Information Knowledgebase. Of these, caDSR is the most complete.

3 Software engineering opportunities

Semantic frameworks of the kind described above, whether for information-driven health research and practice or for any other domain, present significant software engineering challenges. These challenges fall into two categories: those of *data semantics* in representing models, and those of *model-driven development* in generating system artifacts from these models. There are solid foundations on which to build—notations such as XML, RDF, and SAWSDL—but much remains to be done. In particular, a continuously evolving context forces a raising in the level of abstraction. It is a tenet of model-driven development that system artifacts are inherently transient, and should be generated automatically from models rather than laboriously crafted by hand. The current state of the art in model-driven development entails automatic generation of system artifacts, but manual construction of the transformations that generate them. In a continuously evolving context, even these transformations are transient, and thus should be constructed by automatic refinement of models into code.

The specific research challenges in the area of data semantics involve metadata, ontologies, models, security, and analysis. To support diverse and dynamic data communities, techniques are needed for the construction, federation and versioning of metadata, and the establishment and maintenance of multiple, purposed ontologies over the same vocabulary and mappings between them. In order to model sophisticated standards such as the CONSORT [4] statement faithfully, progress is required in expressing and composing workflow fragments—the dynamic aspects of a model, in addition to the static aspects. To derive the full benefit from the semantic framework, the facility is needed for semantics-driven data analysis: metamodels of analytic techniques, enabling the automatic configuration of analytic software. Finally, sensitive domains such as healthcare have stringent security requirements, particularly regarding privacy, and these could be significantly enhanced by exploiting the semantics of data, rather than mere syntax.

In the area of model-driven development, the main challenge is the aforementioned raise in the level of abstraction, to allow the automatic generation of artifact-producing transformations as well as the artifacts themselves. These transformations of transformations are most naturally expressed in terms of higher-order and datatype-generic functional programming, but those techniques need further development to be practically deployable.

4 Addressing the challenge

The US *caBIG*$^{\text{TM}}$ initiative [3] and the UK *CancerGrid* [2] project have had considerable success in constructing semantic frameworks based upon ISO 11179, the international standard for metadata registries. Both groups have developed implementations of the standard, taking complementary approaches. The caBIG implementation, *caDSR*, is intended to serve an entire research community, providing high availability and sophisticated features, and requires a central, highly skilled curation team. The CancerGrid implementation, *cgMDR*, is a lightweight, distributed solution aimed at workgroups and individuals, augmenting content imported from other metadata registries with data elements appropriate to local concerns.

caBIG and CancerGrid are working together to increase the usability and the interoperability of the semantics-based approach: by providing alternative methods of generating web services, accessible from a range of architectures, facilitating semantic interoperability with .NET-based infrastructures such as the VA Cooperative Studies Program (CSP). *CancerGrid* has delivered packaged implementations of registry software that can be used to connect new users to the *caBIG* resources, together with plug-ins to bring semantics to the desktop—for creating and re-using metadata definitions in applications such as *Excel*, *Word*, and *Enterprise Architect*. Work is underway to integrate cgMDR and its plugins into the caCORE infrastructure, to offer these facilities to the caBIG community. Key use-cases facilitated by the collaboration include: deferred public registration of data definitions during experiment design, helping to avoid a proliferation of interim or redundant definitions; the consistent, compositional extension of semantics using local definitions that remain compatible with caBIG tools, useful for research organisations that do not wish to expose or constrain their methodology; the ability to work with local copies, subsets, or reorientations of central resources, increasing the effectiveness of mobile workers and new collaborations.

CancerGrid has developed technology for automatic data integration, and for the automatic generation of software artifacts [1]. These have been been demonstrated in: an Anglo-Canadian clinical study, in which clinical data and tissue sample metadata from five centres was integrated automatically on the basis of declared semantics in cg-MDR, allowing an analysis across 4000 samples that would have been impractical under existing approaches; and in a proof-of-concept usage on the VA CSP, where Microsoft InfoPath forms have been generated for serious adverse event reporting that allow the incorporation of metadata directly from the NCI Thesaurus in caBIG. Components of the semantic frameworks have enabled caBIG collaborators to work independently designing systems in which the metadata elements can be automatically harmonised and re-used, linking models to terminology and allowing discovery services to find potentially combinable data.

The CancerGrid team are working with colleagues in the Centre for Clinical Vaccinology and Tropical Medicine at the University of Oxford, with funding from the UK Research Councils, to demonstrate the application of the semantic frameworks approach to infectious disease control: specifically, on the automated information specification, gathering, and visualisation for the discovery and monitoring of infectious diseases. The intention is to explore a scenario in which a new infectious disease is suspected, confirmed, tracked, and contained: where vocabularies may be extended to describe new diseases; new items of data defined to monitor extent and severity; revised models for reporting, containment and treatment will need to be deployed to both the vet or doctor in the field and their instruments and information systems. We will explore how the monitoring of the epidemic and the actions taken in the field can be integrated into a real-time activity with continual process improvement. This exercise will inform the development of prototypical systems and the formulation of standard messages that direct data capture and decision workflow.

References

[1] Radu Calinescu, Steve Harris, Jeremy Gibbons, Jim Davies, Igor Toujilov, and Sylvia Nagl. Model-driven architecture for cancer research. In *Software Engineering and Formal Methods*, September 2007.

[2] Steve Harris and Jim Davies. CancerGrid: Model-driven and metadata-driven clinical trials informatics. In *Microsoft eScience Workshop*, John Hopkins University, 2006.

[3] George A. Komatsoulis, Denise B. Warzel, Francis W. Hartel, Krishnakant Shanbhag, Ram Chilukuri, Gilberto Fragoso, Sherri de Coronado, Dianne M. Reeves, Jillaine B. Hadfield, Christophe Ludet, and Peter A. Covitz. caCORE version 3: Implementation of a model driven, service-oriented architecture for semantic interoperability. *Journal of Biomedical Informatics*, 41(1):106–123, 2008.

[4] David Moher, Kenneth F. Schulz, and Douglas G. Altman. The CONSORT statement: Revised recommendations for improving the quality of reports of parallel-group randomised trials. *The Lancet*, 357:1191–1194, 2001.