BIOMEDICAL ONTOLOGIES AND GRID COMPUTING
As New Resources for Cancer Registries

Giulio Napolitano[1,2], Alejandra González Beltrán[3], Colin Fox[1]
Adele Marshall[2], Anthony Finkelstein[3] and Peter McCarron[1]
[1]*Centre for Clinical and Population Sciences, School of Medicine and Dentistry, Queen's University Belfast, U.K.*

[2]*Centre for Statistical Science and Operational Research, School of Maths and Physics, Queen's University Belfast, U.K.*

[3]*Department of Computer Science, University College London, UK*

Keywords: Cancer registries, Ontologies, Biomedical grid systems, Semantic web.

Abstract: Cancer registry information systems need to deal with several data sets annotated with different coding systems. Designing, maintaining and linking these datasets involves dealing with semantic issues, tackling the shortcomings exhibited by coding systems as well as considering an appropriate computing infrastructure. We argue that biomedical ontologies and a Grid service infrastructure, together with a clear separation between semantic and coding models, can prove beneficial to cancer registries in terms of accuracy of knowledge modelling, interoperability and knowledge sharing with other registries and related data sources, automation of information retrieval. A real-life example is illustrated and a brief review of related projects is provided. We conclude that a formal semantic layer, which is the basis of large scale meaning-oriented projects such as the Semantic Web, is the key to the provision of a uniform, science-based view across cancer registries and related systems.

1 INTRODUCTION

The biomedical domain is characterised by the use of a variety of information and the continuous generation of new data. Thus, it is fundamental to be able to design, maintain and link different data sets. These data sets could not only reside in distinct locations but also use a variety of syntaxes, even when data items have the same meaning or data items may differ in terms of meaning. Information systems for cancer registries are no exception with respect to these characteristics.

Cancer registration constitutes the source of data on cancer incidence, prevalence and survival rates (UKACR). In the UK, cancer registries have the important role of implementing and monitoring the national initiatives that aim to improve the quality of care and survival prospects for cancer patients (UKACR). A consortium of UK cancer registries recently underwent a re-design of their information systems (PRAXIS). Within this consortium there is a desire to standardise the registration practices and systems. The system is now capable of capturing generic code items from any coding system for each tumour record. Presently, the main coding or terminological systems used for clinical and pathological information that must be adopted by the registries include ICD, SNOMED-CT, READ, OPCS, TNM, plus several other tumour stage and grade coding systems. These coding systems enable the recording of all the information the cancer registries are supposed to collect for epidemiological purposes. Licensed versions of these coding systems are defined, distributed and revised by the World Health Organisation (WHO) and other organisations and are adopted by healthcare institutions (hospital, cancer centres, diagnostic units and laboratories).

On the other hand, these coding systems present several issues. First, there is not a universally accepted medical terminology that could support a widespread development of electronic medical records (Fung, 2005). As terminologies need to fulfil a wide range of functions (such as direct patient care, billing, statistical reporting, automated decision support and clinical research), it is unlikely that a single terminology could ever be suitable for all the

Napolitano G., González Beltrán A., Fox C., Marshall A., Finkelstein A. and McCarron P. (2009).
BIOMEDICAL ONTOLOGIES AND GRID COMPUTING - As New Resources for Cancer Registries .
In *Proceedings of the International Conference on Health Informatics*, pages 512-517
DOI: 10.5220/0001781705120517

purposes (Fung, 2005). Moreover, given the basic requirements of capturing healthcare information electronically and enabling secondary uses of the data for different purposes, the mapping between standard terminologies is a necessity (Fung, 2005).

If we focus on the activity of cancer registration, additional issues related to coding systems also include:

- The meanings of codes and terms created for particular purposes (such as billing) may be unclear in other contexts (such as epidemiology) (Cimino, 1996).
- New versions of the systems are periodically released, and adopted at different times by different organisations. Cancer registries need to cope with these inconsistencies.
- Sometimes the new coding systems are not backward compatible (e.g. SNOMED 3 is not backward compatible with SNOMED2), for instance due to the reallocation of existing codes to different meanings (e.g. T57000 is *Stomach, NOS* in SNOMED3-T and *Gallbladder, NOS* in SNOMED2-T) or because new codes are created for new kinds of disease (such as MALT-Lymphoma, M-9699/3 in SNOMED3-M). Then, semantic mappings between coding systems are required (Fung, 2005).
- Linkage to repositories using different terminologies requires inter-terminology mappings, which is a complicated process not easily generalisable.

A related issue, regarding the information systems, is that validation and Quality Assurance (QA) rules are hard-coded in the cancer registry information systems. There is currently no facility to represent them in an abstract fashion and they even change between registries. In spite of this, it is noted that some efforts have been made in the UK to create a modular 'tumour matching box' to be used and possibly customised by individual registries using the new PRAXIS system.

For all the reasons stated above, this paper advocates building a semantic layer on top of all the cancer registries' subsystems requiring to use coding systems such as ICD9 and ICD10. This semantic layer combines Semantic Web and Grid computing technologies. The rationale behind the proposed approach is explained in the next section. Section 3 introduces a motivating example, while Section 4 presents related work. We conclude with a discussion and a plan for future work.

2 APPROACH

As the biomedical field is a knowledge-based discipline, Semantic Web technologies governing knowledge representation are suitable to tackle transparent search, request, manipulation, integration and delivery of information (Baker, 2007). The vision of data sharing in biomedical systems also requires common standards for data storage and novel frameworks for cross-referring terms and their biological contexts, expressed as controlled vocabularies or ontologies, between different types of data (Nature, 2004). In this context, an ontology is a formal specification of the shared conceptualisation for a domain of interest, which includes the definition of the types of objects occurring in the domain, their attributes and relations between them (Staab, 2004). Ontologies are defined by means of a logical formalism (Baader, 2004).

On the other hand, Grid computing refers to a distributed computing infrastructure supporting resource-sharing in wide-area networks for advanced science and engineering (Foster, 2004). Grid computing deals with system and syntactic heterogeneity, i.e. it tackles the coexistence of several hardware platforms and operating systems as well as different protocols and encodings. This infrastructure is suitable to support scientific practice in biology (Goble, 2001), characterised by distributed collaborations.

In this paper, we argue that combining a Grid service infrastructure with biomedical ontologies could prove beneficial to several aspects of cancer registries' activity. While Grid computing enables collaboration in a distributed and heterogeneous environment (in terms of software and hardware platforms, protocols, etc.), ontologies enable a common understanding of the domain knowledge by providing a formal specification of the conceptualisation shared by experts. The benefits that these technologies can provide to cancer registries are analysed in the following sections.

The approach of building common specifications on the basis of a common semantic model is already strongly encouraged, and proved successful, by Semantic Web projects and Grid services and technologies, as presented in the related work section.

2.1 Ontologies and Cancer Registration

Rector et al. (Rector, 2006) analyse the relation between coding systems and ontologies and distinguish between 'information models' and 'meaning models', respectively. While 'information models' specify data structures for healthcare records and messages, 'meaning models' or 'ontologies' specify human conceptualisations of reality. Thus, 'information models' are metamodels of the 'meaning models' and are used to specify validity conditions for data structures used by coding systems. On the other hand, ontologies are used to test the accuracy of the representation of the world. Consequently, if we take 'disease' as an example, corresponding individuals in the two models represent individual illnesses (John's flu) and classes of illnesses (conditions). This decoupling makes it possible to reason separately about the two models, which are about separate realities. Beside the trivial observation that a code is not a condition or a patient, in practice coding systems and the models behind them are usually based on no or flawed meaning models.

In the context of cancer registries, we advocate to follow the same distinction between coding systems and ontologies to obtain the following benefits.

Firstly, the development of a specific ontology for the domain of cancer registration would be an accurate model, independent from information and coding models and their original intended purposes. It would also encourage the standardisation of operational processes.

Secondly, the ontology would subsume all the relevant notions agreed upon by data managers and medical experts, in the light of the current knowledge (concepts, relationships and restrictions).

Additionally, formal ontologies are specified in suitable logic languages (typically in Description Logic languages, which are a decidable fragment of first order classical logic), by means of specialised tools. This enables the use of automatic reasoners for the computation of satisfiability of individual instantiations of the concepts. Thus, validation rules can be abstracted from the registry implementation, to facilitate sharing and maintenance and rules manipulation and updating would be accessible to domain experts not necessarily versed in ICT.

Finally, an ontology developed on solid theoretical principles, shared by the larger healthcare and research community, would bridge the gap between cancer registries and other repositories of relevant data, such as tissue banks, clinical administration systems, specialised registries for related morbidities and screening databases among the others, provided the ontology, the meaning model, is wide enough. This could be a longer term achievement, although newly conceived repositories, such as tissue and imaging banks, may be more up to date with Semantic Web technologies and ready to share a semantic model.

2.2 Grid Computing and Cancer Registration

While an ontology can bridge the semantic gap between several resources, Grid computing enables collaboration and resource-sharing by providing a suitable middleware infrastructure. As regards cancer registries, implementing a sound strategy based on Grid services can facilitate data-sharing in the epidemiology domain as well as providing other potential advantages enumerated below.

Firstly, the integration with other cancer research resources, as those mentioned above, is possible. Current examples include projects like the cancer Text Information Extraction System (caTIES), whose ultimate goal is to integrate seamlessly heterogeneous resources that provide annotations for individual tissue samples. Cancer registries already contain several items of information manually extracted from clinical notes and reports, in the form of cancer registrations, which can complement other sources of clinical and pathological data for tissue banks.

Secondly, it is possible to provide services that can be used across cancer registries. For instance, the core of the caTIES system is an Information Extraction engine for pathology reports. Some of the authors are exploring the possibility of customising such a service for cancer registration purposes (Napolitano, 2008). Additionally, these services can be combined into workflows that can support the business logic.

Last but not least, the requirement to operate with Grid-compliant services will act as a stimulus to develop the desiderata list mentioned in the Approach section, in terms of accuracy and standardisation. In particular, agreed information and meaning models which can form a common platform for the definition of shared, principled cancer registration rules and mapping between coding systems.

3 MOTIVATING EXAMPLE

This section illustrates how the information and meaning models (Rector, 2006) can be interfaced and how to exploit the expressivity of formal ontologies to automatically enrich the information recorded in a Cancer Registry.

An example of a coding system created with sound clinical motivations is provided by the TNM staging system (UICC). Smaller tumours, tumours confined to the primary site and not involving regional lymph nodes or distant organs, commonly indicate better prognosis for the patients. Based on this observation, a TNM stage provides a staging code as a combination of a T value (mainly based on the size of the tumour), an N value (based on the presence of metastasis in lymph nodes close to the tumour site) and an M value (based on the presence of distant metastasis: M0=no distant metastases, M1=distant metastases present, MX=information not available). Thus, is it possible to automatically infer a TNM staging code for a patient's disease, if this is not explicitly provided by the pathologist?

3.1 Inferring a TNM Staging Code using an OWL Ontology

Let us focus on the M value of the TNM stage. It is assumed that the main Information System of a model Cancer Registry is equipped with a component projecting the incoming information onto a formal ontology. Also, it is assumed that some incoming records from a Patient Administration Systems and Tissue Pathology Reporting Systems cause the creation, in this additional knowledge base, of the individuals *john*, *john_Cancer* and *john_Metastasis* as instances of the classes *Person*, *Cancer* and *Metastasis* respectively, together with relevant relationships:

$$\text{john has\_disease john\_Cancer} \quad (1)$$

$$\text{john\_Cancer has\_metastasis john\_Metastasis} \quad (2)$$

Finally, assume that the ontology also contains the following axioms, embodying relevant knowledge of this domain (in Manchester OWL syntax (Horridge, 2006)):

$$\text{Patient} \equiv \text{Person AND has\_disease SOME Disease} \quad (A1)$$

$$\text{Cancer} \subseteq \text{Disease} \quad (A2)$$

$$\text{M1\_Cancer} \equiv \text{Cancer AND has\_metastasis SOME Metastasis} \quad (A3)$$

$$\text{M1\_Cancer} \subseteq \text{has\_TNM\_M SOME M1} \quad (A4)$$

After classification by a reasoner, John is inferred to be a *Patient* with an instance of an *M1_Cancer*. This information can be easily extracted from the semantic layer, by querying the inferred ontology, and then, transferred to the registry database.

The key advantage of this approach derives from the possibility to keep separate the semantic model, represented by the first two axioms (A1-A2), from the information model, represented by the last two axioms (A3-A4). These information model axioms provide the actual interface between the TNM staging system and the meaning model (A1-A2), which models the reality as described by scientists. These models may reside in two distinct ontologies, eventually combined as modules of the working system. One will be 'authoritative' from a scientific point of view, in the sense that the widest agreement will be sought around it by the domain experts. The other will be 'ephemeral', in the sense that it will be subject to constant and more frequent revision, in the light of the more dynamic (and more arbitrary) nature of the coding systems.

4 RELATED WORK

This section presents related work using Grid and Semantic Web technologies in the context of biomedical systems. These systems follow a service-oriented architecture and use semantic resources, such as controlled vocabularies or ontologies, to integrate heterogeneous and distributed data resources.

The cancer Biomedical Informatics Grid (caBIG™) (Fenstermacher, 2005) is a virtual informatics infrastructure building a federated environment to connect data, research tools, scientists and organisations in the cancer research community. It is an initiative of the National Cancer Institute in the United States. The underlying Grid middleware is called caGrid and it extends a basic Grid infrastructure by focusing on data modelling and semantics. caGrid adopts a model-driven architecture requiring that all the data types are formally described, curated and semantically harmonised. In the United Kingdom, the National Cancer Research Institute Informatics Initiative is developing the Oncology Information eXchange (ONIX), an informatics platform to facilitate access to and movement of data generated from cancer research. ONIX key requirement is to be interoperable with caBIG™. Within this project, some of the authors are working on the ONIX

Semantic Federated Query Infrastructure (González Beltrán, 2008), which allows to perform queries over distributed resources in terms of concepts from the domain ontology. By following the approach presented in this paper, cancer registries could also exploit this additional functionality.

caBIG™ consists of several projects and software tools. In particular, one tool that is relevant for cancer registries is the cancer Information Extraction System (caTIES), which is designed to populate data structures with information extracted and coded from surgical pathology reports. It uses controlled terminologies and provides an interface for querying, browsing and acquiring annotated data. The main text process functionality is managed using GATE (Cunningham, 2002), which is a widely-used open-source natural language processing framework.

DartGrid (Chen, 2006a-b-c) is a project that started in 2002 for the Traditional Chinese Medicine (TCM) community, aiming at integrating heterogeneous and distributed legacy relational databases. DartGrid builds an RDF-view of the relational databases, even considering incomplete information, and supports RDF queries over these views.

The project Advancing Clinical-Genomic Clinical Trials (ACGT) (Tsiknakis, 2006) also aims at building a biomedical Grid for data resources in Europe. ACGT supports data integration based on ontological representation of clinical and genomic/proteomic data taking into account standard clinical genomic ontologies and metadata.

The use of OWL-DL reasoners for tumour grading has already been investigated in (Dameron, 2006). In particular, some limitations of the OWL language (modelling negative restrictions and continuous numeric ranges among the others) are discussed.

(Puleston, 2008) focuses on the balance of distributing the representation of biomedical knowledge and rules between a declarative (ontological) model and a programmatical (software) model. They conclude that different models are appropriate to different tasks and actors in the development of systems and applications in the Health Care realm. This additional distinction is also crucial and we believe it should be considered carefully when implementing meaning and information models in a production Information System.

5 DISCUSSION AND FUTURE WORK

Cancer registry information systems need to deal with a variety of data annotated with several different coding systems. The development of a unique coding system is an elusive goal, as it is unlikely that a single terminology could consider the great range of functions needed. Thus, in the scenario of many co-existing coding systems, issues such as interpreting the ambiguous terms from distinct coding systems, coping with different versions of the terminologies and providing mappings between codes must be considered.

In this paper, we have argued that the most efficient solution to this problem for cancer registries is to combine two cutting-edge technologies such as Semantic Web and Grid computing. Thus, we propose to build a semantic layer on top of cancer registries sub-systems using a myriad of coding systems to annotate data. This semantic layer includes a domain ontology providing a uniform view across sub-systems and it is in charge of providing mappings to the used terminology systems.

We have shown that the additional benefits of using a logic-based representation of information for cancer patients can be exploited to automate coding procedures. This automation is based on the interaction between the scientific knowledge, embedded in the semantic layer, and the coding conventions guiding the structuring of this knowledge for specific purposes.

The projects revolving around the Semantic Web and the Grid infrastructure show that the use of formal ontologies is the preferred route to semantic interoperability. In this paper, we have shown that this approach in the extraction and coding of pathological information for cancer patients has the potential to produce immediate, beneficial effects. As the next step, we will identify other aspects of cancer registration that would benefit in the short term from such a strategy. It is expected that consequent positive results would generate a 'dragging effect', stimulating the extension of research and applications to the remaining aspects of cancer registry activities, where benefits seem to be less obvious or less immediate. Thus, we claim that the use of ontologies in conjunction with the exploitation of relevant Grid services is a route that should be pursued in cancer registration.

ACKNOWLEDGEMENTS

G. Napolitano and C. Fox work in the Northern Ireland Cancer Registry, funded by the Department of Health, Social Services and Public Safety Northern Ireland (DHSSPSNI). A. González Beltrán and A. Finkelstein are grateful for the funding to Cancer Research UK and the National Cancer Research Institute Informatics Initiative.

REFERENCES

Baader F, Horrocks I, Sattler U, 2004. Description logics. In *Handbook on Ontologies*, 3-28.

Baker C.J.O., Cheung K-H., 2007, editors. Semantic Web: Revolutionizing Knowledge Discovery in the Life Sciences. Springer.

caTIES. http://caties.cabig.upmc.edu/. Accessed September 2008.

Chen H, Wang Y, Cui M, Yin A, Wang H, Mao Y, Zhou C, 2006a From legacy relational databases to the semantic web: An in-use application for traditional Chinese medicine. In *Proc. 22nd International Conference on Data Engineering (ICDE).*

Chen H, Wu Z, Wang H, Mao Y, 2006b, RDF/RDFS-based relational database integration. *In Proc. 22nd ICDE*. 20-23

Chen H, Wu Z, Mao Y, Zheng G, 2006c. DartGrid: a semantic infrastructure for building database Grid applications. *Concurrency and Computation Practice and Experience.* 18(14) 1811-1828.

Cimino JJ, 1996. Review paper: Coding systems in health care. Methods Inf Med, 273-284.

Cunningham H, Maynard D, Bontcheva K, Tablan V, 2002: GATE: A framework and graphical development environment for robust NLP tools and applications**.** In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02).*

Dameron O, Roques E, Rubin D, Marquet G, Burgun A, 2006. Grading lung tumors using OWL-DL based reasoning. In *Proceedings of the 9th International Protege Conference*, Stanford, USA.

Fenstermacher D, Street C, McSherry T, Nayak V, Overby C, Feldman M. 2005, The Cancer Biomedical Informatics Grid (caBIG™). In *Proc. 27th Annual International Conference of the Engineering in Medicine and Biology Society*.

Foster, I. Kesslman, C., 2004. The Grid 2: Blueprint for a New Computing Infrastructure. 2nd edition, Elsevier.

Fung, K.W., Bodenreider, O., 2005. Utilizing the UMLS fr semantic mapping between terminologies. In *AMIA Annual Symp. Proc. 266-270.*

Goble, C.,2001. The low down on e-science and grids for biology.*Comparative and functional genomics,2, 365-370.*

González Beltrán A, Finkelstein A, Kramer J, Wilkinson JM, 2008. ONIX Semantic Federated Query Infrastructure. In *Proc. NCI/NCRI Joint Conference*.

Horridge M, Drummond N, Goodwin J, Rector A, Stevens R, Wang HH, 2006. The Manchester OWL syntax. In *Proc.of the 2006 OWL Experiences and Directions Workshop (OWL-ED2006)*.

Puleston C, Parsia B, Cunningham J, Rector A, 2008. Integrating object-oriented and ontological representations: A case study in java and OWL. In *Proceedings of the 7th International Semantic Web Conference (ISWC)*: Karlsruhe, 26-30 October 2008.

Rector A, Qamar R, Marley T, 2006: Binding ontologies & coding systems to electronic health records and messages. In *KR-MED 2006 Proceedings "Biomedical Ontology in Action":* November 8, 2006; Baltimore, Maryland, USA.

Napolitano G, 2008. Ontology-based text mining of pathology reports for cancer patients. In *Proceedings of the 7th International Semantic Web Conference (ISWC)*: Karlsruhe, 26-30 October 2008.

Nature, 2004. Making data dreams come true. 428, 329. doi:10.1038/428239b.

Staab S, Studer R (Eds), 2004, Handbook on Ontologies, Springer-Verlag, Berlin.

Tsiknakis M, Kafetzopoulos D, Potamias G, Analyti A, Marias K, Manganas A., 2006. Building a European Biomedical Grid on Cancer: The ACGT Integrated Project. *Studies in health technology and informatics.* 120, 247-258.

UKACR. United Kingdom Association of Cancer Registries. http://www.ukacr.org/. Accessed September 2008.

UICC. TNM classification of malignant tumours. http://www.uicc.org/tnm. Accessed September 2008.