Metadata-Driven Software for Clinical Trials

Charles Crichton, Jim Davies, Jeremy Gibbons, Steve Harris, Andrew Tsui
Oxford University Computing Laboratory
Wolfson Building, Parks Road, Oxford OX1 3QD, UK
firstname.lastname@comlab.ox.ac.uk

James Brenton
Cancer Research UK, Cambridge Research Institute
Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, UK
james.brenton@cancer.org.uk

Abstract

The CancerGrid approach to clinical trials information systems is based on a metamodel developed from the CONSORT statement of best practice in reporting randomised controlled trials. The metamodel is instantiated with metadata elements drawn from a repository, to create a model of a particular clinical trial. The model is then used to derive automatically a trial management system customized for that trial: generating electronic case report forms, configuring randomisation and eligibility services, and parameterising the security subsystem. The key benefits of the approach are a uniform mechanism for trial registration and discovery, reduced cost and rapid implementation of information systems, and shared semantics leading to improved opportunities for meta-analysis. We describe our implementation of this approach, and outline two applications: for a breast cancer study in the UK, and for a rheumatoid arthritis study run by the US Veterans' Health Administration.

1. Introduction

1.1. The CONSORT statement

The *Consolidated Standards of Reporting Trials* (CONSORT) statement [18] is a widely adopted, successful [20] checklist of thirty-two items, designed to capture best practice in reporting clinical trials. Since an adequate report of an experiment contains sufficient information about that experiment for it to be repeated, it follows that this checklist will detail a number of design items, common to all clinical trials; these design items must be defined at the inception of the trial in the *clinical trial protocol*.

A clinical trial protocol, together with the specific standard operating procedures and case report form definitions, provides a complete specification for a clinical trials information management system; in other words, it is a *model* of the clinical trial itself. Typically, the protocol is presented in an unstructured fashion, as a textual document; were it to be expressed in a more structured manner, it could be used to *generate* elements of the information system needed to support the trial. Where a number of individual models have common features, one may model these features to produce a model of models—a *metamodel*—which can be used to author and validate models.

1.2. Model-driven software engineering

This *model-driven* approach to software engineering is becoming more widely accepted for the development of software systems in specific domains. In much the same way as for trial protocols, a relational database schema is a model of the information structure in a relational database, and may be used to support data representations, indexing, query evaluation, and so on. The metamodel specifies what constitutes a valid database schema, and determines the behaviour of schema design tools. The model-driven approach to database development has become such accepted practice today that few database programmers would think of handcrafting the artefacts that they can generate automatically; it has become simply 'programming'.

1.3. Commonalities and variabilities

By specifying common, essential requirements on clinical trials protocols, the CONSORT statement standardises elements of a clinical trials protocol, and provides the basis for the derivation of a metamodel from them, at least to the

extent of the science and the statistics underpinning the research. Such a metamodel serves two purposes: the validation of a protocol with respect to compliance with the CONSORT statement, and the translation of the protocol into a working information system with specific support for the experiment it describes. However, we face the problem of how to represent the variable areas of the protocol—the *eligibility criteria*, *stratification variables*, and the precise *data* collected at specified timepoints—if we hope to develop services that read the definition and generate or configure services to support these processes.

An eligibility criterion is a clinical or administrative variable which defines the study cohort; it is typically framed as a question with a true/false response, such as 'age at least 18' or 'not pregnant or lactating'. Stratification variables are ones which aim to avoid bias against known prognostic factors, and form series of ranges against which an eligible patient is tested; for example, 'age: $<$40; 40–49; 50–59; $>$60', or 'tumour stage: T0+T1; T2; $>$T2'. If we wish to develop a general service that calculates eligibility or the particular stratification group to which an individual study subject will be assigned, we need a simple, mathematical expression of these declarations, which in turn requires a consistent representation of the types of data involved, and the ranges to be applied. Thus the definition of *value domains*—the range of valid values a clinical or administrative variable may take—with stratification variables and eligibility criteria is an important step in rendering a particular trial protocol computable.

Clinical trials also differ in the data they collect. Data is routinely collected in forms, which are composed of clinical or administrative variables that are collected at certain dates, on certain events, or after certain periods of time have elapsed. The clear definition and and reuse of these variables, dates, events, and periods is an essential prerequisite for data sharing between different groups within the trial—clinical data management teams, pathologists, and translational researchers need to agree and share identifiers so that information may be integrated for analysis—and between groups within the community—statisticians conducting meta-analyses and regulatory authorities monitoring serious adverse events need to share definitions so that sets of data may be unified. A number of minimum datasets have been proposed in the UK over the years [26, 21] which have informal—textual—definitions of these variables and value domains. If we wish to meet our ambitions for clinical research data re-use, we must agree on standards to cover common eligibility criteria, stratification variables, serious adverse event assessments, and subject and sample identifiers. Managing such a diverse array of variables is a difficult task; a PDF document containing this data would be unwieldy and difficult to comprehend in its entirety. We need to formally define these variables, place them somewhere where they are readily accessible to the community, and find ways to focus the community upon those variables it wishes to promote.

A number of health care and research initiatives—most notably the US National Cancer Institute (NCI)—have begun to use the ISO/IEC 11179 standard for metadata registries to provide management of these variables. The *NCI Cancer Data Standards Repository* (caDSR) [8] contains over 10,000 clinical and administrative variables, documenting existing datasets and promoting approved standards for use throughout the Cancer Bioinformatics Grid (caBIG) [28]. In the standard, variables are given a structured representation and associated with terms from a controlled vocabulary—here the NCI Thesaurus [23]—to formally characterise their meaning and to facilitate the management of the resource. The structure in the representation of the data elements provides us with a metamodel against which we can program services that can process specific data elements. Thus in the stratification example, we have a standard way of representing value domains and ranges in order that a randomisation service may be developed to process the stratification specification.

Finally, common units of functionality in the management of clinical trials—generating web forms, storing collected data, randomization, unblinding, and so on—can be generated or configured following the trial protocol, encapsulated as *services*, and made available in a distributed fashion among the participating centres in a trial. Reuse of functionality in this way offers a reduction in the cost of development of specific information system support, and separates the detailed workflows involved in running a clinical trial from an abstract model of the science, statistics, and administration.

Model driven software engineering has been greatly facilitated by XML technologies; the ability to treat programs, validation schemas and service contracts as data and write programs to generate programs is a fundamental requirement for computer aided software development.

1.4. Structure of this paper

Section 2 of this paper presents the CancerGrid approach: a service-oriented architecture for supporting clinical trials; domain modelling of randomised controlled trials, and proper curation of the metadata required for meta-analysis of trial results; automatic generation from the domain model of individual software artifacts, such as forms and services; deployment in the form of plug-ins to standard office automation software; and export of data to standard statistical tools for subsequent analysis. Section 3 discusses our experiences with the approach, touching on services, metadata, domain modelling, a case study with the US Veterans' Health Administration in re-engineering data models from

existing forms; it also discusses related and future work. Section 4 concludes.

2. Methods

Three sources of information were considered in the derivation of the CancerGrid clinical trials model: the CONSORT statement itself and its associated guidance [2], also making reference to a detailed analysis of the CONSORT statement from the perspective of whole study meta-analysis [22]; additional attributes required to register a clinical trial in the UK, including the National Cancer Research Network portfolio [19] and the Metaregister of Controlled Trials (mRCT) [12]; and a sensible base specification for clinical trials information system, drawing on the experience of the clinical members of the project team.

2.1. Service-oriented clinical trial execution

A core set of functions is required of an information system for the recording of clinical trials data. Services are required for: data input; data cleaning; data quality monitoring; clinician management; participant registration, eligibility and randomised treatment allocation; and cross-tabulation of data into sets suitable for analysis. Highly desirable features also include event reminders that help prevent protocol violations and support the timely collection of data, and special procedures for serious adverse event management. At a high level, the influence of local trials unit standard operating procedures upon these requirements is small, and may be implemented simply by varying the precise capabilities of the services supplied. We also wish to demonstrate how this approach could further reduce the complexity of trial startup by supporting automated trial registration. Our intention is to indicate that a 'single document' model for clinical trials—where the trial name and acronym is declared only once—could also ensure consistent documentation and comparison for review, approval, registration, and reporting.

2.2. Approaches to modelling

A number of approaches have been used or evaluated by the project team to achieve these requirements, including early work by the Birmingham Cancer Research UK Trials Unit using a combination of imperative programming and relational database technology, and pre-project feasibility prototyped in the Protégé Frames environment [10] and the W3C web ontology language, OWL [16]. While offering a good user experience and generating functional software, the imperative language/relational database approach does not separate model from generated code, and thus yields systems that are difficult to maintain and extend. The ontological approaches using both the Protégé Frames environment and W3C OWL suffered from the opposite problem: a rich metamodel for a clinical trial could be readily developed, but the user experience and generative support was lacking. Finally, we settled upon the Unified Modeling Language (UML) [9] as being the best fit for project requirements. Figure 1 shows a UML class diagram capturing a fragment of the CancerGrid clinical trials metamodel, showing in particular how case report forms are composed from metadata elements.

2.3. Data standards

A major goal of the CancerGrid work has been to integrate the data standards work of the US National Cancer Institute into an XML-based model-driven generative framework. Thus, where clinical variables are required, these are specified by the 'administered item identifier' from ISO 11179 in accordance with the ISO metadata registry standard used by the NCI caDSR. This standard is extended to incorporate the idea of ranges upon the data elements, in order to provide for eligibility criteria, stratification variables and branching in workflows.

2.4. Generation of software artifacts

One advantage of a UML approach is that a UML class diagram may be readily transformed into a W3C XML schema, from which a number of tools can generate a user interface for creating documents that conform to that schema. These include the W3C XForms standard [6], Adobe PDF Forms [1], and Microsoft InfoPath [17]. In the UK National Health Service, the most accessible and robust of these tools is InfoPath, by virtue of its inclusion in the Microsoft Office suite. To provide a user interface for the creation of electronic clinical trials protocols, the CancerGrid clinical trials metamodel was transformed into a W3C XML Schema, the schema was loaded into InfoPath, and a form designed by hand to guide the user in the creation of a model. Figure 3 shows some screenshots of this trial designer in action.

2.5. Plugins for office software

In order to place definitions of variables into the protocol document, we have developed a plugin for InfoPath that allows users to execute searches against terminology and metadata element resources. At each point that a user needs to reference one of these external definitions, a button is placed upon the form which invokes the plugin. This plugin accesses a search service which has been configured to access popular metadata and terminology services and return matches to any search term input. When the user selects a

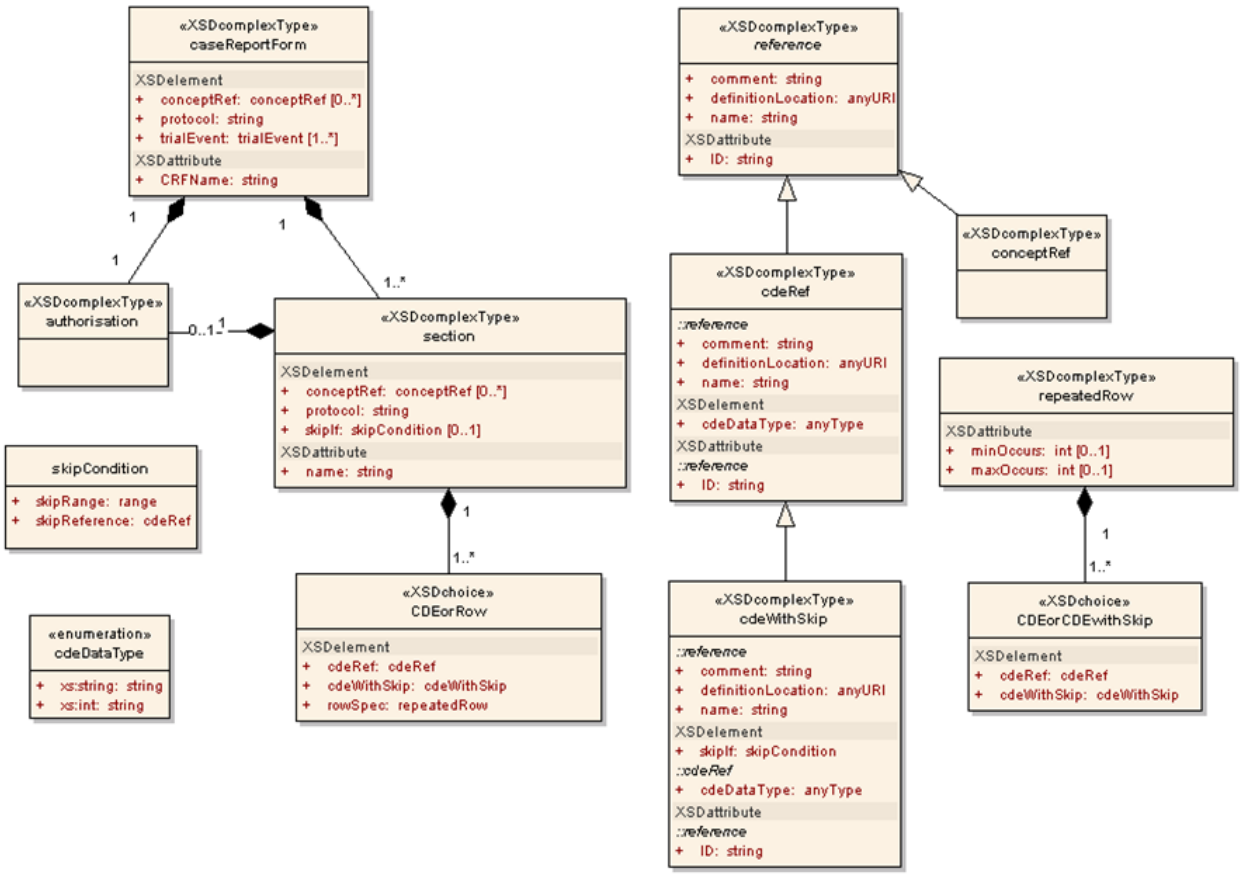

Figure 1. A fragment of the CancerGrid clinical trials metamodel, showing how case report forms are composed from metadata elements

particular term or metadata element, its identifier is inserted into the protocol, together the name, description, and—for metadata elements—the specification of its value domain. This plugin is also shown in Figure 3.

2.6. Deployment

These documents—and each of their revisions—are stored and controlled in an XML database, and accessed by the generating software for transformation and deployment. In transformation, the definition of a case report form is transformed into an XML Schema; this is in turn transformed into a web form for the creation of data, and into a web service that accepts and stores data created and edited by the user. Similarly, the definition of eligibility is transformed into a schema for collecting raw data and a second schema that rejects data describing an ineligible patient. Other transformations produce further configuration files, web service definitions, schemas, and web forms as required, and these are deployed in a *portal* to provide a usable system. This workflow is illustrated in Figure 4.

2.7. Data analysis

Finally, data collected and cleaned through the portal can be imported into popular statistical tools such as SAS and SPSS for analysis. Noting that each metadata element used in a CancerGrid protocol is analysed against a subject to which the data is applied and a predicate describing the property of the subject measured, each data point obtained becomes the object of a set of RDF [15] triples, ideally suited for joining and merging with other datasets for meta-analysis. We have developed a tool that can take the model of a trial dataset and the definitions of its data elements and use it to generate RDF triples so that semantic web tools for storing, querying, and cross-tabulating RDF can be used to support meta-analysis, regulatory reporting, trial registration, and research portfolio management.

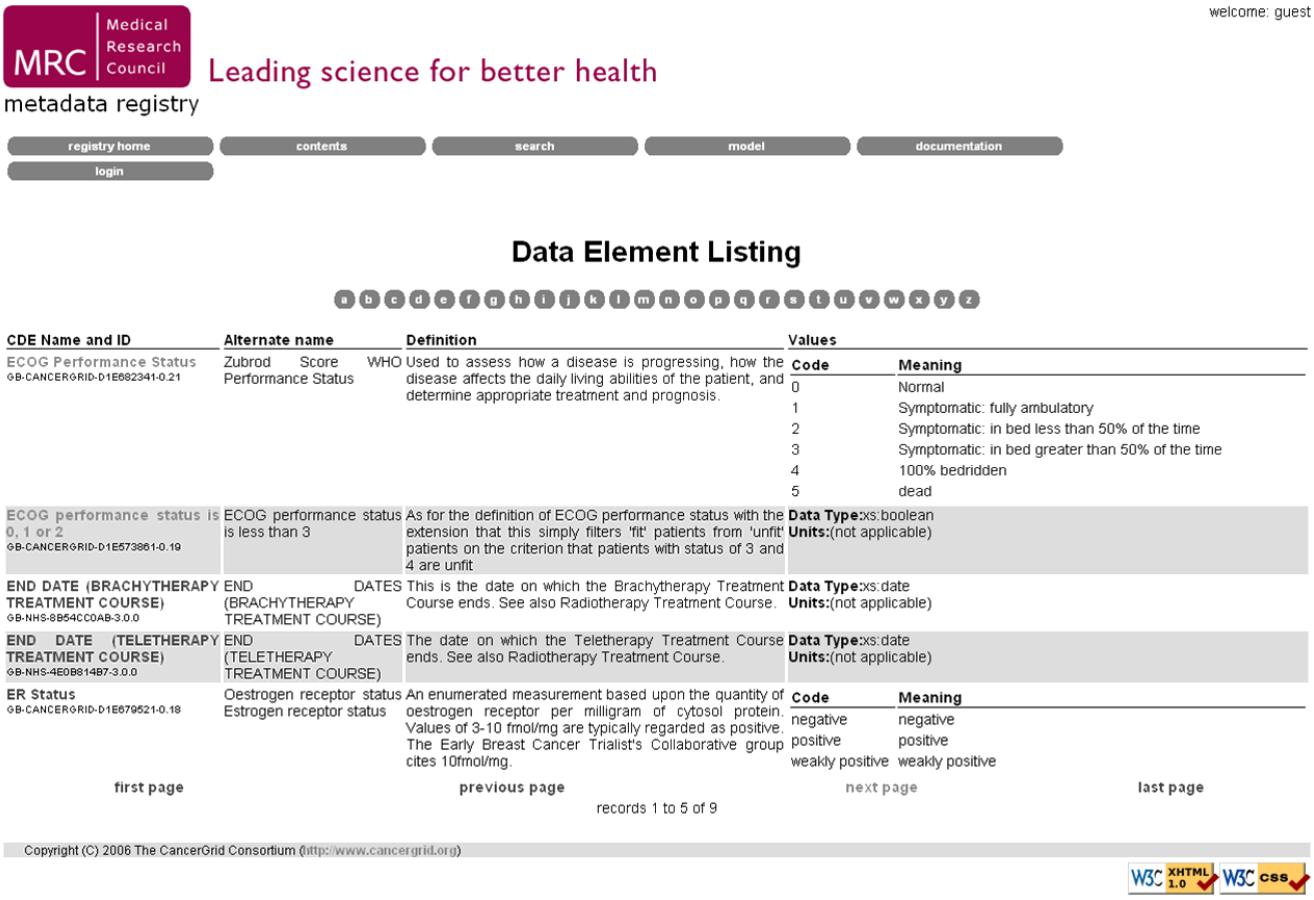

Figure 2. The CancerGrid metadata registry, showing trial-specific and NHS cancer dataset data elements

2.8. Modelling eligibility

One of the problems with any attempt to develop a representation of clinical trials protocols is in the meaning and representation of eligibility. Much of the diversity of clinical trials protocols is in the precise wording of eligibility questions, with many wordings having similar or identical semantics that are difficult to address; a common, interoperable and computable expression of eligibility is key to our goals. Eligibility is assessed against a number of criteria: each criterion is the combination of an observation or assessment and a condition upon the result of that observation. Expressing each eligibility criterion as a metadata element which has a true or false outcome goes part of the way to improving data sharing in clinical trials: one can imagine a data community standardizing upon a set of eligibility questions, from which those required for any particular trial are drawn. However, it also leads to an explosion in the number of data elements—such as a collection of elements varying only in the *number* of weeks elapsed since surgery—and the loss of important data, or the duplication of effort—the exact date of surgery is either recorded elsewhere or lost.

In adopting the ISO 11179 standard for metadata registries as our encoding of our clinical and administrative data standards, we always know the type of the data we are dealing with and how the information on the type of data is represented when it is communicated between components of the CancerGrid system. Taking the example of 'time elapsed since surgery' as an eligibility criterion, instead of a data element 'no more than 3 weeks since surgery', we can specify an open question—'how many days have elapsed since surgery?', with units of days and a datatype of integer—and define a simple range—'$<$21 days'—as the criterion, as shown in Figure 5. This specification is transformed into a restrictive XML Schema that takes the base type defined for the data element and adds a further restriction that will cause the schema to reject XML documents where the number of days elapsed is $\geq$21. In this way we

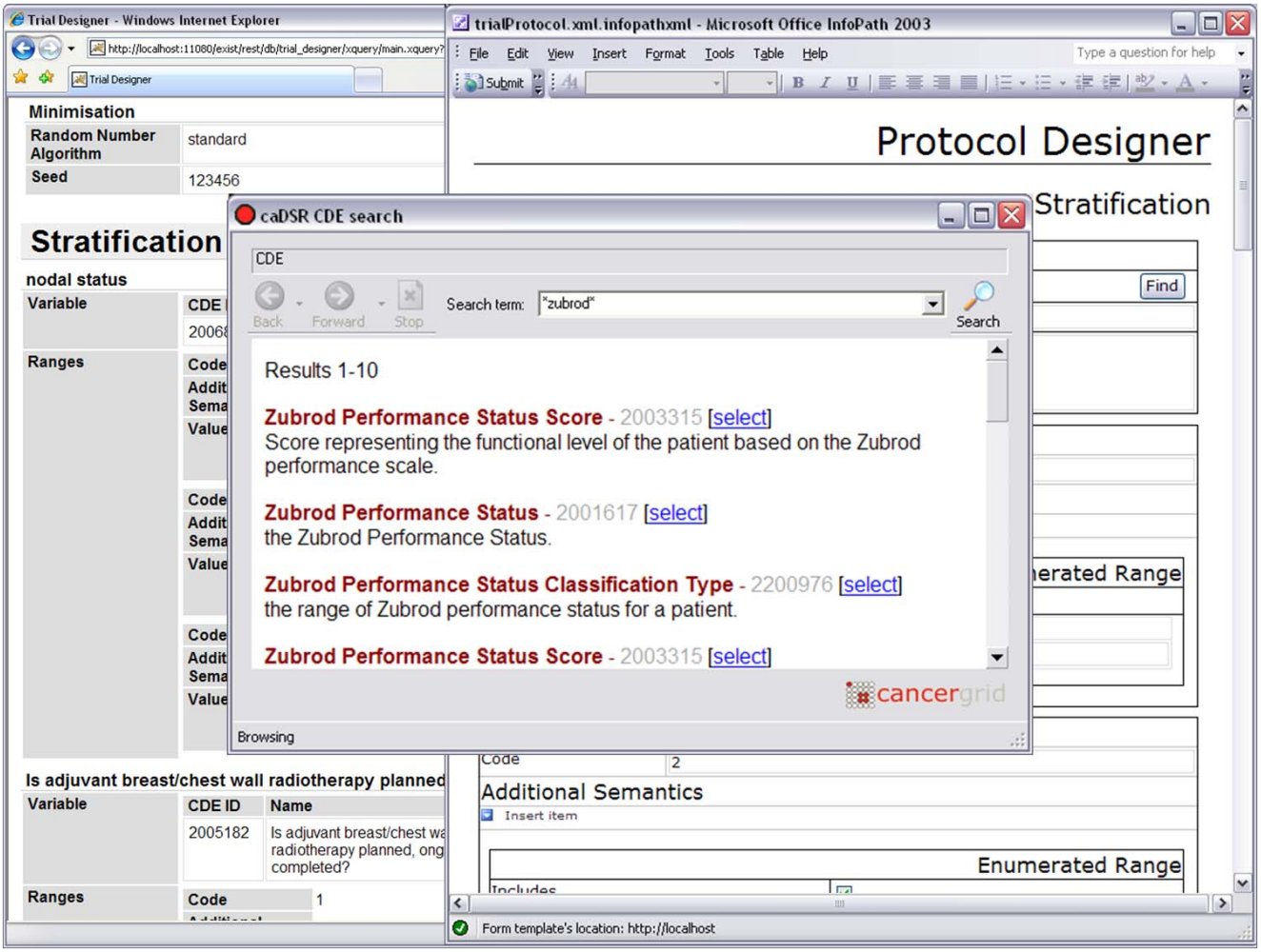

Figure 3. The CancerGrid trial designer, showing the use of a plugin to browse metadata elements from the US NCI cancer data standards repository

can automate the software process of eligibility checking, making more efficient use of a clinician's time and record better data simultaneously. Similar considerations apply for stratification: the clinician completing the form is asked to supply the actual values of the stratification variables, and the system works out what population the study subject should be randomized into.

Ranges of this type are also used in the model's simple concept of workflow: a clinician can model the sequence and timing in which one expects case report forms to be completed, so that the system can remind clinicians and trial coordinators when data is due. A branching condition in a workflow is defined in the same way as an eligibility criterion or a stratification variable: the designer specifies the set of observations to be considered and the conditions upon those observations that cause a branch to be selected, and the software evaluates those conditions when the calculation is required.

3. Discussion

3.1. Service-oriented, model-driven engineering

We have developed an approach to the modelling of clinical trials that allows a group, institution or community of researchers to reuse metadata elements across studies to improve opportunities for data sharing in the comparison, reporting and meta-analysis of clinical trials, and we have shown how this approach can be used to develop functional information systems that faithfully implement the model, avoiding transcription and programming errors during a separate configuration or development process. Obviously

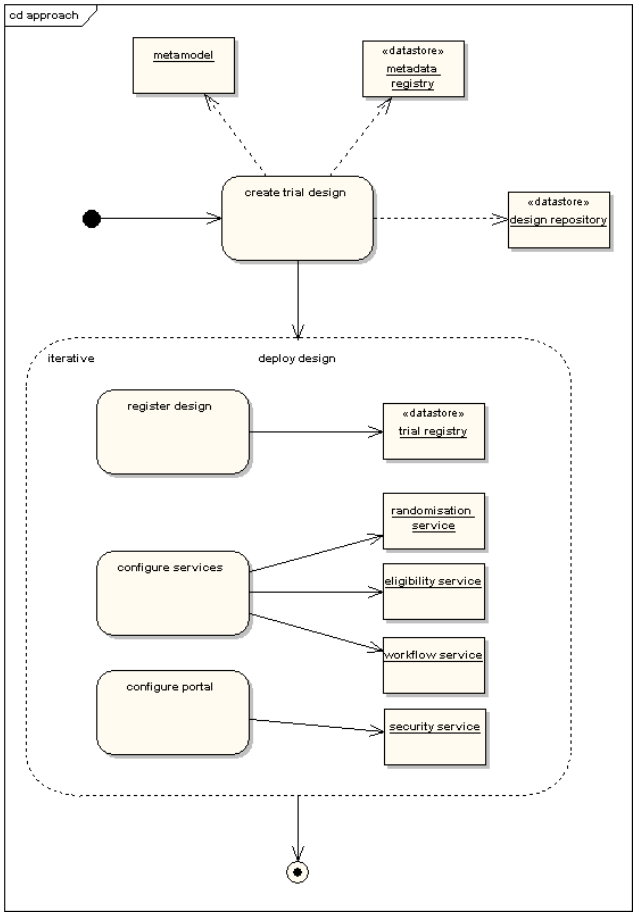

Figure 4. Creating and deploying a trial design

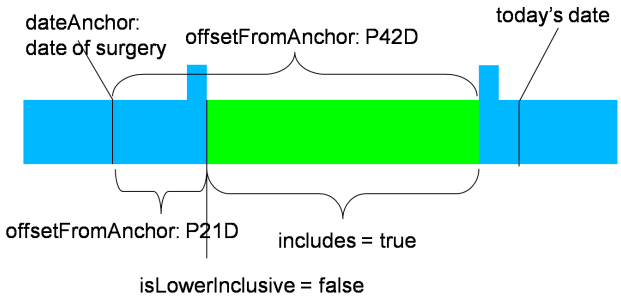

Figure 5. Specifying an eligibility criterion as a restriction on an atomic variable

this offers immediate advantages both in time and cost of implementation of information systems; it also offers clear potential for improvements in functionality. The service-oriented approach to software development provides better factorisation of functionality than traditional, modular, imperative software; and since the model that orchestrates the services is based on the underlying experimental technique, there are clear opportunities for a community to cooperate on the development of common functionality, and to share services where appropriate. For instance, if a single 24-hour unblinding service supported by human operators were required by the community, this could be created and integrated into a service-oriented architecture without the requirement for a central clinical trials management system or office, preserving a healthy diversity vital for innovative research.

3.2. Metadata elements

The declaration of metadata elements has obvious benefits for meta-analysis: data of a common type, or that may be transformed into a common type, may be readily identified. Beyond that, the registration of clinical and laboratory metadata elements has additional benefits. By standardising upon metadata elements, as well as whole case report forms, flexibility is maintained to support innovation within the community. Generated interfaces can link back to the definitions, or compose the definitions into a 'help file' for the form, to support improved data quality in clinical research. Each research study reusing a metadata element definition can rely upon the curators of the registry to provide the comprehensive definitions, standard procedures, specifications, and calibration requirements, reducing the cost of training centres and offering the ability for detailed comparison of data collection practices for each metadata element at a centre across the whole trial portfolio.

3.3. Trial metamodel

By incorporating metadata elements into a relatively lightweight metamodel of common scientific and administrative elements of a clinical trial, we lay the foundation for an open-standards framework for clinical research informatics that will connect researchers to high quality, interoperable data. We make explicit the implicit standardisation that is the foundation for meta-analysis, and provide a mechanism for managing an expansion of the range of data that may be joined and integrated. We show how a single document model can automate and enhance routine tasks such as trial registration and information system configuration and generation: this could be extended to support the preparation of an application for ethical approval, enhanced peer review of proposed protocols, and—with incorporation

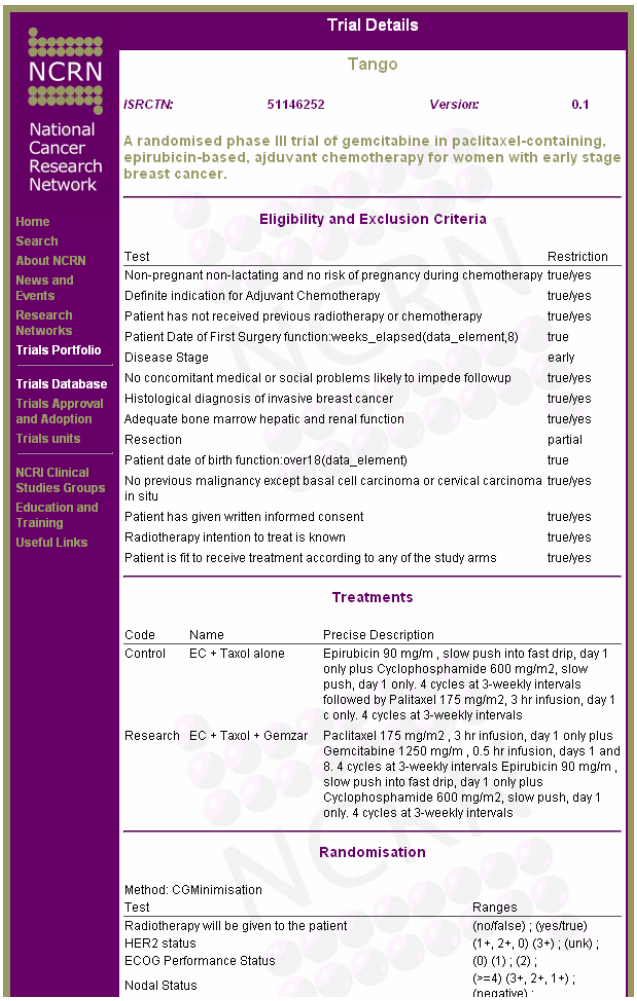

Figure 6. An example of a trial design transformed into a registration webpage. This transformation uses an earlier, functional representation for elapsed periods

of the detailed modelling of clinical trial analysis from the Trial Bank work [22]—standard analysis and visualisation of the results.

The use of simple ranges afforded by the incorporation of a standard for the representation for value domains inherited from ISO 11179 reduces data loss or duplication in the determination of eligibility or stratification. This offers further opportunities for data integration—particularly in the support of virtual tissue banking—and allows us to see how eligibility criteria from a number of clinical trials could be combined into a decision tree that would allow a clinician to match a patient with clinical trials [7].

The use of metadata elements to compose eligibility criteria, stratification variables and case report forms also allows us to offer a degree of automated case report form completion: identifying, registering, and mapping metadata elements from the National Health Service Data Dictionary and the local hospital information system onto those specified by the trials community would allow the clinician to fetch, transform, and insert data from the patient health record into the case report form. Unlike direct mining of clinical records, this approach maintains the correspondence between the clinician and the statistician about the precise representation and quality of the clinical information, and preserves the essential ethical and legal requirements for the individual to opt in to clinical research.

3.4. Capturing and reusing existing implicit common data elements

The US Veterans' Health Administration's *Cooperative Studies Program* (VACSP) [27] has an existing system for the electronic collection of case report forms, built upon the Microsoft SharePoint platform, and using Microsoft's InfoPath form-based data entry system. Each time the VACSP needed to create a system for a new trial, they built the custom XML Schemas which define the structure of the forms. Although these XML Schemas are similar for each trial, they obviously have to vary according to the individual details of the trial. Once the XML schemas are built, the process of creating the forms themselves has to be completed. Essentially this is a matter of specifying the controls that are to appear on the form together with their correponding text. This results in a professional appearance and good usability; however, it is difficult to re-use the effort for the next trial, as InfoPath is not designed to allow schemas to be changed. There are mechanisms to allow sets of controls to be re-used, but these are tied to particular namespaces.

As a case study, we took a Serious Adverse Event form from a VACSP study concerning rheumatoid arthritis, and built a tool to 'scrape' the data from their existing InfoPath forms. The combination of the text in the user interface controls and the type information in the XML Schemas contained within the InfoPath form was used to create a set of candidate common data elements (CDEs). For various technical reasons, it is not always possible to match the text describing a control with the type information—usually because the text and controls have been formatted in a table and it is difficult to know whether the text matches the control to the left, right above or below it. However, some simple heuristics usually get it right, and where there are multiple possibilities, both are output as candidate CDEs: the CDEs are then curated in the normal way using the CancerGrid Metadata Registry (cgMDR) technology.

The normal CancerGrid process of building a form was then conducted to build a model of the form from the CDEs. This model was then used to generate the XML

Schema for the Serious Adverse Event form. The generated XML Schema was then used as part of the definition of the InfoPath form. The text accompanying controls was drawn manually from the CDE definition in cgMDR. Whilst we would like to be able to generate the whole InfoPath form—as we are able to do with other technologies such as XForms—this was still a step forward with respect to the VACSP's existing approach, as it provided a consistent and methodical way of producing the schemas from the CDEs in a way that was easily modifiable for new trials. The XML Schemas we were able to provide back to the VACSP were also better, as they included extra SAWSDL [14] markup indicating which CDE each generated XML Schema Element was derived from. This extra markup information can be used to help analyse the data collected on the actual forms.

3.5. Virtual tissue banking as meta-analysis

A primary clinical goal in the CancerGrid project was to facilitate translational research by extending case-based meta-analysis into the genomic domain. Given the genetic element of cancer and the increasing capability of high-throughput genomic techniques, it seems clear that the collection of tissue in cancer clinical research will be seen as essential in the near future. If we have standards for key clinical and tissue metadata, we should be able to assemble large, well-specified retrospective observational cohorts from a variety of different prospective clinical studies where each study subject has had tissue collected at comparable timepoints, and requisition the corresponding tissue collection. The CancerGrid trials metamodel supports this activity by: declaring a common, treatment-centric event sequence for study subjects; facilitating the reuse of standard metadata elements for clinical and administrative variables, which one presumes will include key standards associated with eligibility, known prognostic factors and outcome; reducing data loss by calculating eligibility and stratification from the actual value of the underlying metadata element; and allowing the researcher to discover these cohorts by comparing trial designs rather than querying actual patient data. The latter point is particularly important: tissue sets can be proposed, ethical issues examined, and funding assessed before sensitive, personal data on study subjects is exchanged.

An exciting opportunity is represented by the potential to register fragments of clinical trials in the design portal, and incorporate them into the protocol being edited. Aside from the obvious reuse of case report form and form section definitions, we envisage that detailed models of control treatments could be made available to the protocol author, enhancing opportunities for virtual tissue banking.

3.6. Access control with metadata

Privacy issues are clearly very important in clinical trials. The CancerGrid trials metamodel allows the specification of fine-grained user access policies, describing who should have access to data and what they can do with that data. The metadata-based approach we have taken greatly assists in this regard, by allowing access policies to be specified according to combinations of particular data elements. For example, one can specify that a member of the statisticians' user group should be unable to access the name or address data elements of a patient in an anonymized trial.

3.7. Comparison with other approaches

Modelling a field of knowledge has an inevitable pitfall: as models become more useful—in detail or extent—so they become more specific and limited in their application. Alternatives to the CancerGrid model can be broadly categorised into two types: ones expressed as an ontology, and ones expressed in notations such as UML.

3.7.1. Ontological approaches

The most important published ontological model of a clinical trial is the work of Sim [22]. This has taken the CONSORT statement and provided a detailed analysis in support of a 'Global Trial Bank'—a resource that would allow users to discover trials and assess their quality and compatibility. This limited scope—support for a specific resource—is its greatest strength: by relying upon the conceptual framework of the CONSORT statement and arranging this framework into a model to support a resource, it has wide applicability but does not require universal acceptance. The Trials Bank work is primarily aimed at retrospective analysis of publications: many of its attributes support the description of the data collected and of the analysis performed; significant extension and rearrangement would be required if one wished to generate information system artifacts from it. Additionally, our model requires less expertise from the user: it is simpler, while still addressing fundamental requirements; it inherits a clearer view of the encoding of clinical variables from the ISO 11179 standard; and designs can be created by including or modifying previously defined clinical variables, case report forms, and ultimately by cutting and pasting whole sections of protocols. However, the common approach that Sim and the CancerGrid team have taken means that a significant subset of the Trials Bank dataset could be generated automatically from a CancerGrid trial design.

3.7.2. Domain modelling approaches

A common approach to establishing common semantics is through domain modelling. In this approach, a group of

knowledge workers attempt to identify and order all of the things of interest to their community within a single model, expressed in a notation such as UML. While UML class models are good at representing the kind of complex relationships frequently required of a domain model, the approach does not scale: the effort required to negotiate agreement, derive the model, communicate its intent, police its application, and maintain its currency is out of proportion with the additional benefits obtained over more focussed harmonisation projects.

One endemic problem is accounting for local differences in practice: for example, in version 1.49 (January 2007) of the BRIDG Model [4], the class *Person* has the required attribute *educationLevelCode* with US-centric valid values of 'less than High School Diploma', 'High School Diploma', 'Some College', and so on. Similar context dependence is evident in the values of other required attributes such as *ethnicGroupCode* and *raceCode*. Similarly, the SNOMED clinical terminology [13] models—among many other things—occupations, but the military occupations (at least in the January 2008 version) are all UK-specific, despite SNOMED being nominally an international standard. A second problem is the common confusion of a model of the domain with one of a simulation of that domain; while it is a fundamental tenet of object orientation that these two are closely related, nevertheless (in Korzybski's dictum) 'the map is not the territory'. For example, Smith and Ceusters [24] roundly criticize the HL7 [11] Reference Implementation Model for its 'unsure treatment of the distinction between information about an action on the one hand and this action itself on the other'.

Domain modelling in an ontology can meet with more success [3], simply because it is possible to develop a more general model without problematic relationships and assertions. However, this generality precludes their use as models to support software generation.

3.8. Future work

3.8.1. Other CONSORT-style standards

The CONSORT standard is only one of a related set of checklists, each of which could provide the basis for a metamodel and a service-oriented approach to information gathering and analysis to provide extensive information system support. Where elements of the models coincide—in the model of case report forms, for instance—then services developed against that element of the metamodel could be reused to support elements of the required information system. This approach works equally well where the experimental design commissions the collection of new data, or where existing data is reused in a uniform fashion—definitions for case report forms in an observational study are replaced with the schema of the source data in *Diagnostic Accuracy* [5] and *Meta-analysis of Observational Studies* [25].

3.8.2. Clinical record system design

Given that the UK National Health Service is committed to developing a data dictionary, and that it is desirable to place the definition of clinical record gathering in the hands of clinicians, the CancerGrid approach has implications for the patient health record. An ISO 11179 metadata registry offers a flexible, standard, and transparent mechanism for declaring metadata elements, offering a bridge between terminologies such as SNOMED-CT [13] on the one hand and models such as UML diagrams or trial designs on the other. If it is possible to declare a metamodel for a medical record form, then we have shown that it is a simple task to provide a knowledge worker with a user interface that allows them to assemble terms and metadata elements into a design from which the electronic form, web service interfaces, and validation schemas may be automatically generated. We have also shown that one may simply declare ranges upon data elements by virtue of a standard representation of its value domain, and use these ranges to define fixed and variable workflows that route data through processes. Finally, we have shown how such designs may be deployed into low-cost open- and closed-source portal frameworks, and how the setting of standards through metamodels and metadata elements can ensure that a distributed community collects interoperable data to the extent that interoperability is required. All of these elements, when brought together, could provide local solutions to health record projects that guarantee interoperability on a national level, and reduce the need for information system developers to understand the intricacies of healthcare. It is interesting to note that this is also the goal of the HL7 version 3 messaging standard [11], achieved without recourse to a large, unwieldy, poorly understood and less-than-optimal domain model.

4. Summary

We have described the CancerGrid approach to clinical trials information systems. This approach is model-driven, in the sense that it is based on a metamodel of randomised controlled trials that is itself derived from the CONSORT statement of best practice in trial reporting. The metamodel is instantiated to yield a model of a particular trial, and this model is used as the basis from which to generate the software artifacts to manage the trial. (The generative technology used to acheive this will be described in a companion paper.) An important part of the instantiation process is to annotate the trial model with metadata characterising the data to be collected and recorded, so as to support subse-

quent reuse and integration; this metadata is preserved and propagated throughout the life of the data it describes. We have described two applications of the approach: to a breast cancer trial in the UK, and to a rheumatoid arthritis study run by the US Veterans' Health Administration.

The authors would like to thank the other members of the CancerGrid team, especially Carlos Caldas in Cambridge for his guiding role, and David Rose at the US Veterans' Health Administration for his help with the joint case study. The research reported here was supported by the UK Medical Research Council and Engineering and Physical Sciences Research Council, and by Microsoft Research.

References

[1] Adobe XML Forms Architecture. `http://partners.adobe.com/public/developer/xml/index_arch.html`, October 2007.

[2] D. G. Altman, K. F. Schulz, D. Moher, M. Egger, F. Davidoff, D. Elbourne, P. C. Gtzsche, and T. Lang. The revised CONSORT statement for reporting randomized trials: Explanation and elaboration. *Ann Intern Med*, 134:663–694, 2001.

[3] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene Ontology: Tool for the unification of biology. *Nature Genetics*, pages 25–29, 2000.

[4] Biomedical Research Integrated Domain Group. BRIDG model, version 1.49. `http://www.bridgmodel.org/`, 2007.

[5] P. M. Bossuyt, J. B. Reitsma, D. E. Bruns, P. P. Gatsonis, P. P. Glasziou, L. M. Irwig, J. G. Lijmer, D. Moher, D. Rennie, and H. C. De Vet. Standards for reporting of diagnostic accuracy. toward complete and accurate reporting of studies of diagnostic accuracy: The STARD initiative. standards for reporting of diagnostic accuracy. *BMJ*, 226:24–28, 2003.

[6] J. M. Boyer, editor. *XForms 1.1 W3C Working Draft*. W3C, February 2007.

[7] R. Calinescu, S. Harris, J. Gibbons, and J. Davies. Cross-trial query system for cancer clinical trials. In T. Sobh, editor, *Advances in Systems, Computing Sciences and Software Engineering—CISSE 2006*, pages 395–390. Springer, 2007.

[8] P. A. Covitz, F. Hartel, C. Schaefer, S. D. Coronado, G. Fragoso, H. Sahni, S. Gustafson, and K. H. Buetow. caCORE: A common infrastructure for cancer informatics. *Bioinformatics*, 19(18):2404–2412, 2003.

[9] M. Fowler. *UML Distilled*. Addison Wesley, 3rd edition, September 2003.

[10] J. H. Gennari, M. A. Musen, R. W. Fergerson, W. E. Grosso, M. Crubzy, H. Eriksson, N. F. Noy, and S. W. Tu. The evolution of Protégé: An environment for knowledge-based systems development. *International Journal of Human-Computer Studies*, 58(1):89–123, January 2003.

[11] Health Level Seven. `http://www.hl7.org/`.

[12] R. Horton and R. Smith. Time to register randomised trials. *BMJ*, 319(7214):865–866, 1999.

[13] International Health Terminology Standards Development Organisation. Systematized Nomenclature of Medicine–Clinical Terms. `http://www.ihtsdo.org/snomed-ct/`, 2007.

[14] J. Kopecký, T. Vitvar, C. Bournez, and J. Farrell. SAWSDL: Semantic annotations for WSDL and XML Schema. *IEEE Internet Computing*, 11(6):60–67, 2007.

[15] F. Manola and E. Miller, editors. *RDF Primer*. W3C, February 2004.

[16] D. L. McGuinness and F. van Harmelen, editors. *OWL Web Ontology Language Overview*. W3C, 2004.

[17] Microsoft Office InfoPath 2007. `http://office.microsoft.com/en-us/infopath/FX100487661033.aspx`, October 2007.

[18] D. Moher, K. Schulz, and D. G. Altman. The CONSORT statement: Revised recommendations for improving the quality of reports of parallel-group randomised trials. *The Lancet*, page 357, April 2001.

[19] NCRN Trials Portfolio. `http://www.ncrn.org.uk/portfolio/about.asp`, October 2007.

[20] A. C. Plint, D. Moher, K. Schulz, D. G. Altman, and A. Morrison. Does the CONSORT checklist improve the quality of reports of randomized controlled trials? a systematic review. *Firth International Congress of Peer Review and Biomedical Publication*, September 2005.

[21] Royal College of Pathologists. `http://www.rcpath.org`, October 2007.

[22] I. Sim and D. E. Detmer. Beyond trial registration: A global trial bank for clinical trial reporting. *PLoS medicine*, 2(11):365, 2005.

[23] N. Sioutos, S. de Coronado, M. W. Haber, F. W. Hartel, W.-L. Shaiu, and L. W. Wright. NCI Thesaurus: A semantic model integrating cancer-related clinical and molecular information. *J. of Biomedical Informatics*, 40(1):30–43, 2007.

[24] B. Smith and W. Ceusters. HL7 RIM: An incoherent standard. In A. Hasman, R. Haux, J. van der Lei, E. D. Clercq, and F. Roger-France, editors, *Ubiquity: Technologies for Better Health in Aging Societies*, volume 124 of *Studies in Health Technology and Informatics*, pages 133–138. IOS Press, 2006.

[25] D. F. Stroup, J. A. Berlin, S. C. Morton, I. Olkin, G. D. Williamson, D. Rennie, D. Moher, B. J. Becker, T. A. Sipe, and S. B. Thacker. Meta-analysis of observational studies in epidemiology: A proposal for reporting. *JAMA*, 283(15):2008–20012, 2000.

[26] UK Cancer Dataset. `http://www.datadictionary.nhs.uk/data_dictionary/messages/national_cancer_data_set/cancer_registration_data_set_fr.asp?shownav=1`, October 2007.

[27] Veterans' Health Administration's Cooperative Studies Program. `http://www.csp.research.va.gov/`.

[28] A. C. von Eschenbach. A vision for the National Cancer Program in the United States. *Nat Rev Cancer*, 4(10):820–828, 2004.