CARMEN: a practical approach to metadata management

Mark Jessop, Mike Weeks and Jim Austin

References	**This article cites 2 articles** http://rsta.royalsocietypublishing.org/content/368/1926/4147.full.html#ref-list-1
Rapid response	Respond to this article http://rsta.royalsocietypublishing.org/letters/submit/roypta;368/1926/4147
Subject collections	Articles on similar topics can be found in the following collections e-science (31 articles)
Email alerting service	Receive free email alerts when new articles cite this article - sign up in the box at the top right-hand corner of the article or click **here**

To subscribe to *Phil. Trans. R. Soc. A* go to:
http://rsta.royalsocietypublishing.org/subscriptions

CARMEN: a practical approach to metadata management

BY MARK JESSOP*, MIKE WEEKS AND JIM AUSTIN

Advanced Computer Architectures Group, Department of Computer Science, University of York, Heslington, York YO10 5DD, UK

The Code Analysis Repository & Modelling for E-Neuroscience (CARMEN) project aims to enable broad sharing of resources, through the provision of a secure, online environment for storage and curation of data, analysis code and experimental protocols, together with the ability to execute data analysis. While the CARMEN system is initially focused on electrophysiology data, it is equally applicable to many domains outside neuroscience.

Metadata are essential for a system such as CARMEN that has the potential to store thousands of data collections and analysis codes; without metadata, resource discovery, interpretation, evaluation and re-use would be severely impeded. Therefore, when any resource (data, service or workflow) is added to the system, users must provide adequate descriptions. These descriptions form a metadata repository that is searchable to allow users to find any kind of resource held in the system, assuming that the user has appropriate access rights.

This paper discusses and explores the project's approach to implementing such a metadata repository that meets both system requirements and user expectations. Initial approaches were refined after user evaluations, and a more practical approach was followed that better aligned with the aims of the users and the project as a whole.

Keywords: portal; metadata; XML; services

1. Introduction

The CARMEN (Code Analysis Repository & Modelling for E-Neuroscience) project, see www.carmen.org.uk as described by Smith *et al.* (2007), aims to advance neuroscience research and collaboration. The project consortium consists of 20 scientific investigators at 11 UK universities bringing together neurophysiologists, neuroinformaticists and computer scientists to address the complete life cycle of neurophysiology data, of which the primary type is electrophysiology neural activity recordings, both signals and image series.

Funded by the UK Engineering and Physical Sciences Research Council (EPSRC), CARMEN seeks to enable the broad sharing of resources through the provision of a user-driven online environment to allow research groups to store and share significant volumes of data, metadata and analysis code rapidly,

*Author for correspondence (mark.jessop@cs.york.ac.uk).

securely and privately irrespective of geographical separation or computing platform. Users may choose to share resources internally, externally or publicly with the world at large.

In order to achieve the project's aims, the CARMEN system, as described by Gibson *et al.* (2008*a*), has been developed and deployed. The current system release affords registered users the ability to upload experimental data, which they can describe with extensive metadata, and apply security policies. These resources can be searched, shared with other users, downloaded, visualized and analysed by a user-supplied code. Support for data analysis has been provided by encapsulating an analysis code in Web services. The capability to join services together through a common Neuroscience Data Format (NDF) (see Liang *et al.* 2008) and workflows is currently under development.

Metadata are essential for a system such as CARMEN that stores large amounts (multiple terabytes) of data and analysis codes; without such metadata, resource discovery, interpretation, evaluation and re-use would be severely impeded. Therefore, when any resource (data, service or workflow) is added to the system, users must provide adequate descriptions. The descriptions are stored together to form a metadata repository that is searchable to allow users to find any kind of resource held in the system, assuming that the user has appropriate access rights.

2. Experimental data and metadata

For many of the scientists in the CARMEN consortium, the technology and ideas being investigated by the project form a new and unfamiliar territory. The users generally fit into three groups: neuroscientists who generate data, researchers who consume data and generate algorithms, and those who bridge both groups. Often collaborations are formed between the groups, to each other's mutual benefit.

Experimenters use various brands of hardware to generate data recordings that are in proprietary formats that tend not to be publicly available. In order to allow data to be processed, manufacturer-supplied tools or libraries are used to convert these data to a scientist's own format, often in Matlab. It is also usually the case that different experimenters generate incompatible formats. These data are often processed using local desktop tools and software such as Matlab, R and Python.

The metadata that go with these data tend to come as hand-written notes in laboratory books, which are linked to the physical data via long file and/or folder names. Clearly, this forms not only a difficult resource to access but also an unsearchable resource that is very difficult to share with other scientists. In the cases where the datasets are many terabytes in size, sharing becomes both impracticable and expensive.

3. Requirements for metadata

The CARMEN system has not been designed as a primary data collection tool; for example, it is not used for live streaming of experimental data as they are recorded or generated. The aim is that scientists will collect data offline and record metadata during this process. Once the data have been collected,

they are uploaded to the system and described in the metadata system. Later versions of the system may incorporate real-time data collection and computational steering.

Without good-quality metadata, the system simply becomes a large storage device that makes sharing and collaboration very difficult. Therefore, the metadata system must enable and encourage users to supply as much metadata as possible by ensuring that metadata entry is as painless as possible. For each resource type that a user or the system itself may describe, the metadata schema must be wide enough to ensure that the resource is adequately described. To promote resource discovery, sharing and collaboration, the metadata must be extensively searchable so that users can find data, services and workflows that are of interest. Finally, the metadata must be viewable. This means not simply that there must be a way of viewing metadata, but that it must be made available to as many users as possible or ideally be publicly viewable by anyone. Without visibility of metadata and data, collaboration and sharing are severely impeded.

4. Defining metadata

For each of the resource types in the CARMEN system, a comprehensive metadata definition is documented, for example, neuroscience electrophysiology experimental data are described by the 'Minimum information about a neuroscience investigation' (MINI) document (see Gibson *et al.* 2008*b*). This is analogous to the 'Minimum information about a proteomics experiment' (MIAPE) document for proteomics described by OBI (2009), which breaks the data down into sections that map onto the Functional Genomics Experiment (FuGE) data model described by Jones *et al.* (2007). FuGE models common aspects of life-science experiments such as protocols, equipment, materials and software, and is being implemented by domains such as genomics, proteomics and metabolomics. The neuroscience domain shares many components with these domains as well as containing additional components. The FuGE model can be readily extended to incorporate further components and domains, and so there is a good fit between CARMEN and this existing model.

The MINI and related documents represent the minimum information that should be recorded about a resource and its associated dataset to allow a reader to interpret and evaluate the processes performed and the conclusions reached. Although the structure of the documents map onto the FuGE model, the metadata definitions only define *what* should be recorded and not *how* they are recorded or stored by the system.

The content of the MINI was largely defined by the scientific users and then shaped by the development team. This gave the users an investment in the process that should help them to provide as full a metadata record as possible for their data. During the definition process, users were encouraged to contribute to a 'term of the week' discussion designed to help define the scientific terms and language used by the scientists. This was an invaluable process, as it soon demonstrated that, although the users operated within the same domain, they did not all use the same terms or agree on the definitions. In one particular instance the term 'electrode' defined by Gilmour (2001) as 'small piece of metal used to take an

electric current to or from a power source, piece of equipment or living body'
elicited a response of tens of emails and some heated discussions and after two
weeks had to be brought to an end.

5. The metadata system

An early aim of the CARMEN project was to provide a system capability
quickly by bringing together many existing tools from consortium members
in order to allow scientific users to gain a real benefit from the beginning of
the project. To achieve this, early implementations of the CARMEN system
used the Systems and Molecular Biology Data and Metadata Archive (SyMBA)
package developed by Lister *et al.* (2007) as part of the CISBAN project,
see http://www.cisban.ac.uk, as the metadata system component. SyMBA
is an implementation of the FuGE data model and was one of very few
software implementations, consisting of a front-end graphical user interface
implemented in Java Server Pages (JSP) coupled to a Java back-end and
FuGE database. The MINI was encoded into an XML (eXtensible Markup
Language) document that mapped the MINI descriptions onto the FuGE data
model with appropriate extensions and ontological terms where appropriate.
This was ingested by SyMBA to allow it to provide both user interfaces and
metadata storage.

SyMBA was provided with its own storage mechanisms, which were not
compatible with the storage request broker (SRB) used by CARMEN. SRB,
described by Baru *et al.* (1998), is a distributed file system that allows remote
sites' storage capacity to be joined together irrespective of the hardware platform
used. The content stored within SRB can simply be addressed as a file path
without referring to the physical location of the data. Therefore, the system and
the JSP interface were modified to fit within the CARMEN system. During the
upload of experimental datasets, SyMBA captured metadata via forms generated
from the MINI definitions in the FuGE database. Metadata search and view
facilities were provided by an extended Java interface but displayed through the
standard CARMEN user interface.

At this time in late 2006, SyMBA was still in very early stages of development.
While the database was a true FuGE implementation and it was possible to
represent the CARMEN metadata requirements within it, the user interfaces,
usability and system performance were not at that time ready to support live
users. Early user trials with CARMEN consortium members identified issues with
both the generated metadata forms and the system performance. The users found
that the input forms were both too complicated and too long, with extended
response times. In our experience, users tend to be resistant to completing onerous
metadata forms, and therefore any obstructions that make this process slow or in
any way difficult would result in few metadata being entered. Without metadata
the CARMEN system would become little more than a large disk storage resource
and hence not promote the sharing and collaboration that the project is designed
to foster.

In response to this, the metadata requirements were revisited, with an emphasis
on supporting users in order to encourage maximum metadata capture. A set of
primary requirements were developed: encapsulation of metadata schemas for

different resources as with FuGE; easy-to-navigate user interfaces with which users could work; fast and responsive user interfaces that made light work of metadata entry, which was considered a 'chore' by users; search capabilities across all fields in the definitions and across the different definitions; and the ability to re-use previous metadata entries so that users could generate new metadata documents without having to re-enter them in their entirety.

These requirements led to the rapid development of a new CARMEN metadata system that separated the data model from the user interface. The system was designed to be able to store multiple metadata descriptions such as the MINI in one database. The system keeps descriptions of the metadata documents known as schemas and instances in separate database tables. From the document schemas the system generates two XML form descriptions. The first contains a description of the structure of a metadata document and the second contains a description of the contents of the document.

The XML schemas that represent a metadata document, such as the MINI, separate the document into subdocuments, which are analogous to the components within the FuGE model. Each subdocument may contain many sections that represent the contents of the metadata document. This arrangement allows the user interface to be managed more effectively and to be broken down into smaller forms that are more acceptable to the user.

The XML schemas allow the user interfaces to be built in almost any conceivable way, and by separating structure from content they allow the system to be very responsive. User interface tools only need to request document content descriptions as required.

The CARMEN system is a standard three-tier Web architecture as described by Eckerson (1995); see figure 1. The top tier is a Web portal built using the Google Web TOOLKIT (GWT; see Google 2009), which allows the interface to be built in Java and then compiled to JavaScript. The middle tier consists of a collection of Java servlets and the bottom tier a MYSQL database.

When a user needs to enter metadata, the portal requests the XML document that describes the structure of the appropriate metadata schema from the middle tier. From this it builds the underlying structure of the metadata document and sufficient interface elements to allow the user to start navigating the document. As the user visits subdocuments and sections of the metadata document, the portal requests the content description XML from the middle tier and then generates the appropriate interface elements.

Through the appropriate use of caching, the portal can give the user a seamless route through the document while dynamically building user interfaces as it goes. This is responsive enough for the user not to see the joins between subdocument and sections. By only requesting the documents and sections that are actually visited by the user, the interface is fast and responsive, while simultaneously reducing network connections and traffic.

At the end of the metadata entry process, the portal generates an XML description of the metadata in the correct format for the document schema, which it passes to the middle tier. The middle tier Java servlets parse the XML into the metadata storage tables ready for later searching, viewing and processing.

A form designer user interface was developed within the portal that allows metadata documents such as the MINI to be turned into an XML document schema. The designer allows for interface elements such as edit boxes, text
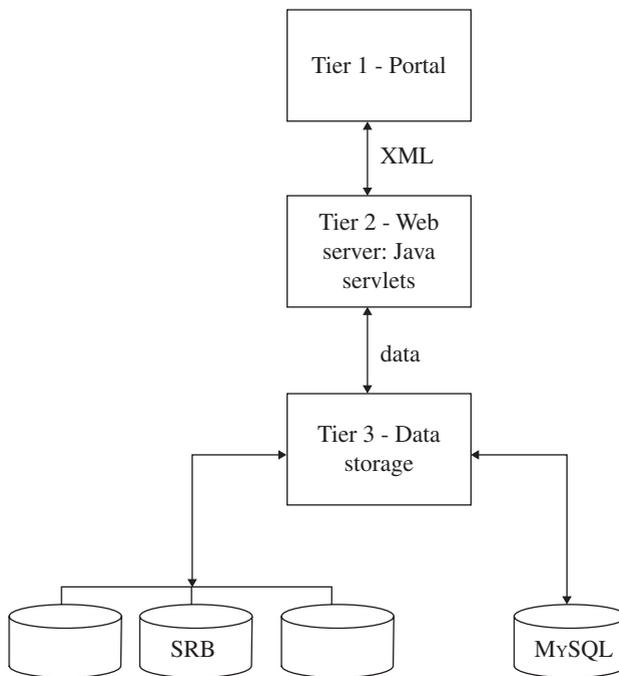
M. Jessop et al.

Figure 1. CARMEN system architecture.

areas and selection lists to be placed on a form. Where an element represents a value that has a unit of measurement, a unit selector can be added. Each element can be specified as either 'required' or 'optional', whether or not the field can be included in any metadata search operations and whether the field is an editable field or simply read-only. The metadata system allows constraints to be specified for each field in a form. This can be used to ensure that a field only contains values in a certain range or that it conforms to a particular format.

The form designer generates an XML document that is processed by the middle tier to extract the schema structure and content into the metadata system database tables. The XML schema is stored separately as a complete document for the purposes of system regeneration.

The CARMEN system uses life science identifiers (LSID) as defined by the LSID Resolution Project (2007), to uniquely identify every kind of resource in the system, such as users, metadata schemas and metadata instances. LSIDs encapsulate version information within the identifier, and this allows metadata schemas and instances to be changed, thus creating new versions while still being linked to the original item.

(a) User-driven design

Given the adverse reaction of users to the first attempt to build a metadata system, it was important that users were involved in the design of the user interfaces of the new system. A series of usability tests were conducted where

users were put in front of a metadata system and asked to complete a set of tasks while explaining their actions, thoughts and expectations, which were recorded by an observer.

An initial prototype was built, which provided basic metadata manipulation. A group of 20 users was gathered and tasked to interact with the system, but without instruction or guidance. During the task the users talked about what was right or wrong, how they expected the system to behave and what was missing. At the end of the session the users were asked to draw what they thought the system should look like.

The usability notes and users' own ideas were used to draw a series of story boards that demonstrated how the system worked and what it looked like. These were then presented back to the users for comment. From the story boards the prototype was amended and the usability trials repeated. This process was conducted through a number of iterations to arrive at a metadata system that the users could use and of which they felt an ownership.

Over the course of the usability trials the different experiences and backgrounds of the users led to conflicting requests. In addition to the system, the content of the MINI specification itself also came under scrutiny. Although the users had been largely responsible for the content of the MINI, many still found it confusing or did not understand the meanings of some sections.

This resulted in refinements to help text and field descriptions and a couple of additional elements in the metadata schemas. Where the conflicts related to the system design, a path had to be steered between the issues. In many cases these related to some users wanting the system to look and feel like a Windows application or more like a Web page for others. Often users expected forms of interface interaction that were not usual in a Web-based application.

For many users, their day-to-day experience was that of keeping hand-written laboratory notes alongside experimental data stored on local laptops or CDs/DVDs. These users found the experience of trying to gather metadata, data and experimental protocols via a more formally governed method quite alien, finding it hard to describe or qualify their requirements or expectations.

6. Entering metadata

When a user adds a new resource to the CARMEN system, the appropriate form description for the resource type is retrieved from the middle tier and used to automatically generate input form interfaces in the portal, as shown in figure 2. The user can fill in the forms afresh or use previous instances known as templates to pre-populate the forms. New templates can be created at any point in the process. In the simplest case, a user may use a standard template for each new experiment and simply change one or two values. At its most flexible the system allows the user to create templates for each kind of experiment, item of equipment, study subject, process, etc., and then to pull one or more templates together to describe a new experiment.

The portal automatically generates a sequential path through the metadata forms, allowing the user to visit as much or as little of the forms as they require, the only caveat being that they must complete those sections that are regarded as mandatory.
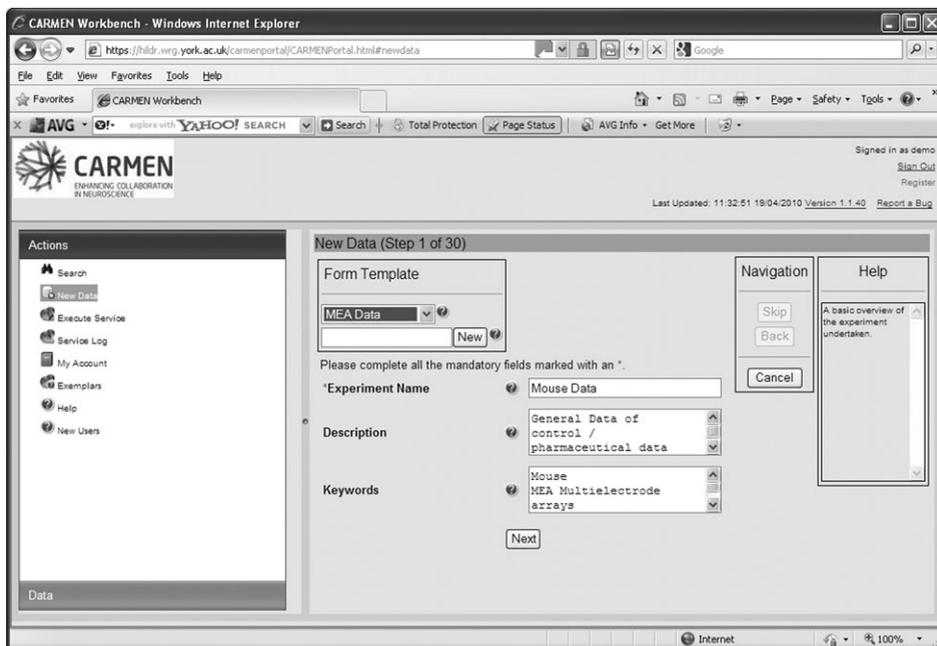
Figure 2. Entering metadata.

Once the metadata are complete, from the user's perspective, they are converted to an XML document for transmission and storage in the metadata system. Once stored, the metadata document is referenced against the resource being added to the system. The owning user is able to apply security constraints to the metadata and resource independently. They can be kept private, shared with other users or groups in the system, or made public so that they are viewable or downloadable by anyone visiting the CARMEN system.

Users who have appropriate access privileges are able to edit the metadata. The metadata system employs versioning so that any changes result in a new version being recorded in the system.

7. Viewing metadata

Users are able to retrieve metadata to which they have access rights either by browsing the portal to locate resources and using the interface to view metadata or via the search engine. The search engine matches search terms against values found in the metadata repository and categorizes the results according to resource type. This enables users to locate multiple resources of different types that contain specific terms in their associated metadata, without necessarily knowing for which kind of resource they are looking. As an example, a user could find both data and a service to operate on it in a single search operation.

The search function operates across all versions of all documents so that it may produce results where the match was in a previous version. Where multiple versions of a metadata document exist, the user is able to select which version he/she wishes to view.
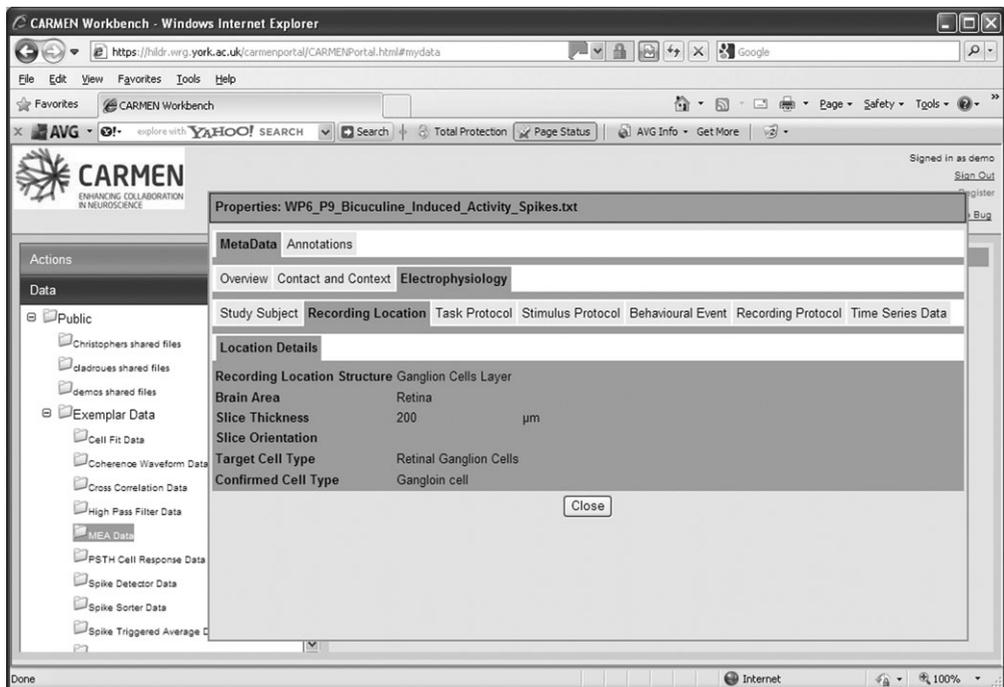
Figure 3. Viewing metadata.

In the case of services and service outputs, the metadata contain the appropriate links so that a user can trace the path from an output file to the service that created it to the inputs, both parameters and data. Where workflows are concerned, the user is able to view the complete process from start to finish and see the intermediate parameters, inputs and results.

By looking at the metadata associated with a service output or workflow, it is possible to re-run the service or workflow with the same set of parameters.

When a user selects metadata to view, as shown in figure 3, the portal requests the metadata from the middle tier. The metadata are delivered to the portal as an XML document that contains both the document structure and the content. The portal uses the structure information to generate forms that it populates with the content.

8. Metadata for services

A Web service is described by the Web Service Description Language (WSDL), as defined by Christensen *et al.* (2001), which contains a formal description of the programmatic service interface and its input and output parameters. This is enough information such that a generic tool could invoke any service from the WSDL, but insufficient information to provide a meaningful interface to a human user. The WSDL does not contain descriptive or semantic information that tells a human user what the service does and what its parameters mean. In common with

all resources in the CARMEN system, services are associated with a metadata document, in this case the 'Minimum information about a service' (MIAS). This document fulfils three purposes, firstly to document and describe what the service does, secondly to enable the system to provide a meaningful user interface and thirdly by replacing the WSDL to provide a way for the system to be able to invoke the service.

The CARMEN system contains a module for wrapping code supplied by scientists into a Web service and from this generating a MIAS XML schema document. The metadata schema is generated from elements of the WSDL with information provided by the owner of the code. The owner of the service supplies a textual description of what the service does, how it works and what can be expected of it. They also describe what the inputs and outputs are and what they mean, along with any default parameters. The parameters can also have constraints applied to them such as being values within a given range, maximum values or for files being of a given type.

Users can discover services to which they have access through the search interface in the portal in the same way that they can find other resources or by bookmarking them in their own work area of the portal. Services can either be invoked singly or as part of a workflow. When a user decides to invoke a service, the portal requests the metadata description from the middle tier. From this description it can generate user interface elements that allow the user to provide values for the input parameters. The constraints associated with the parameters allow the portal to validate the user inputs, which could be numerical values, a selection of choices from a predefined set or an input file. In all cases a default value may have been defined in the metadata. Where a parameter is a file, the portal provides a browse button so that the user can select a file from their data folders within the portal. The constraints may only allow prescribed file types to be selected.

Before the portal attempts to invoke a service, it validates the input values against the service metadata and ensures that all parameters have been given a value. From a user's inputs, the portal generates an XML document that identifies the service and describes the inputs and passes it to the middle tier for processing.

The middle tier parses the XML and checks that the user has appropriate access rights to the service and any input files before attempting to invoke the service. Once a service has completed executing, a further metadata schema similar to the MIAS document captures information about the service execution. It records run-time statistics, the input values and files along with service outputs. The output always consists of at least one file that is associated with the new metadata instance. Where the service generated a textual output, this is captured and placed in the output file. Services may produce multiple output files and in this case each one will be associated with the same metadata.

The metadata that represent the output from a service contain enough information to allow a service to be re-run under exactly the same conditions as with the original execution. This includes both the software environment and input values and/or files. The metadata and associated output files are owned by and private to the user who executed the service, although the user is able to change these, for example to be shared or made public.

9. User experiences

When the user trials were conducted with the new metadata system, none of the adverse reactions that had occurred with the previous system were seen again. However, given that the scientists had been involved in the design of the MINI document, a significant number were still surprised by the length of the metadata and several did not know what some of the fields meant. This second reaction fits in with the behaviour described above in §4, where users could not agree on the meaning of some terms used in the domain.

The user trials showed that the new system was faster and easier to work with and, when combined with the template facilities, allowed for fast and efficient population of the metadata forms, leading to a more complete metadata record.

Some key groups within the CARMEN project quickly started using the system for sharing data not only within the project consortium but also bringing in external research groups with which they collaborate. For these users, the metadata record tended to be fully populated and they made extensive use of the templates to simplify this process. This behaviour fits in with expectations, where it is in the interests of data generators to provide a full metadata record so that data consumers can best understand the data with which they are working. However, it was discovered that some users realized that they could simply provide enough metadata, usually keywords, to allow their collaborators to find the data through the search mechanism.

A couple of users exploited an oversight in the templates so that they generated identical metadata for multiple datasets. This behaviour revealed itself through the results of using the search engine, where some search terms produced a list of what appeared to be tens of duplicate datasets. Closer examination of the metadata and datasets revealed what had been happening.

10. Future work

As previously stated, the metadata system moved away from an ontological/data model-driven approach to one that was more practical. Although the metadata documents contain terms drawn from ontologies, there is no specific ontological support in the system. The use of ontologies is still seen as a relevant and important feature, and so future work would be to incorporate ontologies into the system.

As discussed in §9, some users found ways to 'cheat' the metadata system and these loopholes need to be closed. In order to prevent the problem with duplicate metadata, the templates will reset some of the relevant fields in the forms so that they must be completed in documents derived from a template.

Although many users do complete the metadata, this is not always the case, and so a process of user education and an update to the user interface is under way. Very early evaluation sessions with users have shown this to be a potentially successful approach. However, there is no escaping the size of the MINI document.

11. Conclusions

While the ideal approach to managing metadata in the CARMEN project seemed valid, in reality for the project's computer scientists CARMEN is as much about implementing a usable system as it is about research. This led to the adoption of a practical metadata system while moving the ontology and data model-based approach into a research thread until the software was mature enough to serve the user's requirements.

The metadata system built for the CARMEN system has fulfilled the user's requirements and performed as expected, allowing the project to focus on the metadata content and appropriateness rather than on the implementation technology itself. Examination of the metadata system has shown that, in general, the users are successfully completing the metadata record and using the tools provided to achieve this.

The CARMEN system has been online for approximately one year, during which time the users have uploaded and described their data. As described earlier, for many users this was the first time they had employed a formal process for describing their experiments, and therefore they were not sure of what they required from the system. The experience has revealed the weaknesses in the user interfaces, particularly in the speed and ease of navigation through the metadata documents. As a result, a new round of evaluation and user trials to solve these problems has been initiated. It is only by following an iterative process of design, development and user evaluation that effective refinements become possible.

References

Baru, C., Moore, R., Rajasekar, A. & Wan, M. 1998 The SDSC storage resource broker. In *Proc. CASCON'98 Conf., Toronto, Canada, 30 November–3 December*, p. 5. Toronto, Canada: IBM Press. See http://www.sdsc.edu/srb/Pubs/srb.ps.

Christensen, E., Curbera, F., Meredith, G. & Weerawarana, S. 2001 Web services description language (WSDL) 1.1. W3C Note. See http://www.w3.org/TR/wsdl.

Eckerson, W. W. 1995 Three tier client/server architecture: achieving scalability, performance, and efficiency in client server applications. *Open Inform. Syst.* **10**, 3–23.

Gibson, F. *et al.* 2008*a* The CARMEN virtual laboratory: web-based paradigms for collaboration in neurophysiology. In *Proc. MEA Meeting 2008, Reutlingen, Germany, 8–11 July*, pp. 117–120. Stuttgart, Germany: BIOPRO. See http://www.nmi.de/meameeting2008/.

Gibson, F. *et al.* 2008*b* Minimum information about a neuroscience investigation (MINI): electrophysiology. *Nature Precedings*. See http://hdl.handle.net/10101/npre.2009. 1720.2.

Gilmour, L. (ed.) 2001 *Collins concise dictionary & thesaurus*, 2nd edn, p. 296. New York, NY: Harper Collins.

Google. 2009 Google web toolkit (GWT). See http://code.google.com/webtoolkit/.

Jones, A. *et al.* 2007 The Functional Genomics Experiment model (FuGE): an extensible framework for standards in functional genomics. *Nat. Biotechnol.* **25**, 1127–1133. (doi:10.1038/nbt1347)

Liang, B., Simonotto, J., Knowles, A. & Fletcher, M. 2008 The Neurophysiology Data Translation Format (NDTF) specification. See http://www.carmen.org.uk/standards/Carmen DataSpecs.pdf.

Lister, A. L. *et al.* 2007 Implementing the FuGE object model: a systems biology data portal and integrator. Computing Science Technical Report, Newcastle University, Newcastle-upon-Tyne, UK.

LSID Resolution Project. 2007. See http://lsids.sourceforge.net/.

OBI, Ontology for Biomedical Investigations. 2009. See http://purl.obofoundry.org/obo/obi.

Smith, L. S. *et al.* 2007 The CARMEN e-Science pilot project: neuroinformatics work packages. In *Proc. UK e-Science All Hands Meeting, Nottingham*, pp. 591–598. See http://www.carmen. org.uk/publications.