

Comparing METS and OAI-ORE for Encapsulating Scientific Data Products: A Protein Crystallography Case Study

Charles Brooking, Stephen R.Shouldice, Gautier Robin, Bostjan Kobe, Jennifer L. Martin, Jane Hunter
*The University of Queensland,
St Lucia, Queensland, Australia
jane@itee.uq.edu.au*

Abstract

This paper describes the set of eResearch services developed by the eResearch Lab within the University of Queensland (UQ) for the Structural Genomics (SG) Group at UQ. The aim of these services is to enable collaborative teams of protein crystallographers in the SG group to track their experiments and to manage the plethora and diversity of data that they generate through distributed high-throughput approaches and complex scientific workflows. More specifically we describe: the secure Web-based laboratory information management system (TIMTAM) and the X-ray diffraction image archive (DIMER) used to monitor experiments and record data captured prior to structure determination and the publication of a new crystal structure in public repositories such as the Protein Data bank (PDB). We also describe the services that we have developed to relate the different products generated at each stage in the protein crystallography pipeline through OAI-ORE compound objects. We conclude by comparing the OAI-ORE approach for publishing and sharing related scientific outcomes with the METS-based approach employed by other scientific laboratories.

1. Introduction

High-throughput techniques are enabling rapid experimentation at every stage of structural biology – from cloning, expression and purification to structural data measurement and structure determination. However, these automated techniques and the distributed nature of collaborating scientific teams lead to challenges in tracking the progress of experiments, managing the wide range of research products and documenting the provenance of results. The Open Archives Initiative Object Reuse and Exchange (OAI-ORE) data model appears to offer an ideal approach for representing artifacts generated during the protein crystallography scientific life cycle and for formally expressing the relationships between the data, products, structures and publications in a formal machine-processable way.

In this paper we describe the set of Web-based services we have developed that enable distributed teams of protein crystallographers at the University of Queensland to: link data captured in a Web-based laboratory information management systems (TIMTAM) to a repository of X-Ray diffraction images (DIMER), a data bank of protein crystal structures (PDB) and a repository of publications (ACS or PubMed). Using the OAI-ORE protocol, typed relationships and unique persistent identifiers for target proteins, experiments, people and publications, scientists are able to relate the significant scientific outputs at each stage in the experimental workflow and encapsulate these individual products within re-usable compound objects – thus facilitating the validation and verification of the scientific results. Previous approaches to encapsulating the different outputs from crystallography experiments (e.g., TARDIS [1] and the UK eBank project [2]) have employed METS [6]. Hence, an additional objective of this paper is to compare the benefits and disadvantages of OAI-ORE and METS for encapsulating the products of a scientific process or workflow to enable discovery, re-use and learning.

The remainder of the paper is structured as follows: Section 2 describes background information and previous related work; Section 3 describes the overall system architecture and the different architectural components, implementation, system functionality, and user interface; Section 4 provides an evaluation and comparison between OAI-ORE and METS; Section 5 outlines future work plans and conclusions.

2. Background and Related Work

2.1. Linking Raw Data to Publications

There are a number of pre-existing projects that are archiving scientific data and images associated with crystallography experiments, in order to enable independent validation and verification of published results. These include: the TARDIS project at Monash University [1]; SPECTRa at the University of Cambridge and Imperial College London [4]; the

eBank UK project [2] and its follow-on eCrystals project at the University of Southampton [3].

TARDIS (The Australian Repository of Diffraction Images) aims to support the archival and sharing of X-ray diffraction images for the protein crystallography community. It links sets of images to the PDB structure and publication through a METS schema [22]. SPECTRa developed a set of customized software tools to enable chemists to routinely deposit experimental data, in Open Access digital repositories, but did not support “linked datasets”. The eCrystals project supports the archival of the underlying data generated during the course of structure determination from a single crystal x-ray diffraction experiment – as a set of separate files that can be downloaded. This approach, although useful, is limited with regard to discoverability of the individual components and extensibility, re-use, visualization and inferencing of relationships between data distributed across networks. The eBank UK project [2] supported the packaging and dissemination of the different artifacts associated with a crystallographic experiment, through METS packages.

Our approach is different to these previous related efforts, in that we use OAI-ORE to link the experimental data (in TIMTAM) to the X-Ray-diffraction images (in DIMER), and then to the PDB structure and publication.

The oreChem project [5] is a Microsoft-funded collaboration (between Cornell, University of Cambridge, University of Indiana, Penn State, PubChem and the University of Southampton) that is also developing new models for research and dissemination of scholarly materials in the chemistry community, based on OAI-ORE. The collaborators are designing a graph-based object model for the chemistry domain that is built around the central role of the “molecule” and the “chemical compound” and the underlying specifications of OAI-ORE. The collaborators are mainly focusing on inorganic crystallography - the work described here will build on the ontologies and models being developed within oreChem, but extend, refine and evaluate them for the protein crystallography domain.

2.2. The Open Archives Initiative Object Reuse and Exchange (OAI-ORE) Model

The Open Archives Initiative Object Reuse and Exchange (OAI-ORE) [7] is an international collaborative initiative, focusing on an interoperability framework for the exchange of information about Digital Objects between cooperating repositories, registries and services. OAI-ORE aims to support the creation, management and dissemination of the new forms of composite digital resources being produced by eResearch and to make the information within these

compound digital objects discoverable, machine-readable, interoperable and reusable.

OAI-ORE endorses Named Graphs as a means of publishing compound digital objects that clearly define their logical boundaries. The nodes in the Named Graph correspond to the individual aggregated resources, and the arcs correspond to typed relationships between those resources. In the terms of OAI-ORE, compound objects correspond to ORE Aggregations, and the Named Graphs that describe them to ORE Resource Maps (ReMs). The result of an HTTP access of a Resource Map URI is a serialization of the triples describing the Aggregation. This serialization may be in any of the OAI-ORE serialization syntaxes: RDF/XML [10], RDFa [11], and Atom [12].

Our objective in the work described here is to evaluate the application of OAI-ORE to the encapsulation of scientific products generated across the life cycle of a protein crystallography experiment – and to compare this approach with the METS approach being employed by other similar efforts in the protein crystallography community. Figure 1 shows the components that will be incorporated within the OAI-ORE compound objects being generated by the UQ Structural Genomics group using the services that we have developed.

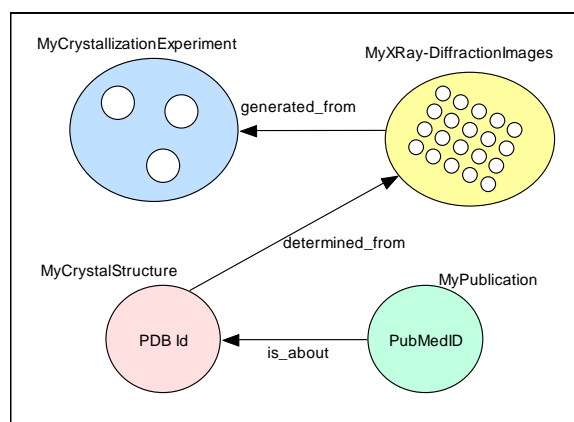


Figure 1. Compound object for protein crystallography

The work described here also builds on similar efforts undertaken by the UQ eResearch Lab in evaluating the application of OAI-ORE for authoring compound scientific objects in the disciplines of Australian Literature [16] and nano-materials optimization [17].

2.3. The Relationship between OAI-ORE and METS

The Metadata Encoding and Transmission Standard (METS) [6] is a standard developed by the library community, for encoding descriptive, administrative,

and structural metadata associated with digital library objects. It uses the XML schema language to specify an XML syntax for identifying both the content files constituting the object and the metadata describing the object and content files, and the relationships between the various component content files and metadata. Whilst OAI-ORE and METS have been designed for different objectives, they have both been used for encapsulating multiple related digital objects, in a range of contexts including protein crystallography

McDonough [14] recently published a paper describing how the tree structure of a METS document can be “aligned with” or mapped to an OAI-ORE graph. He also acknowledged the relative inflexibility of METS compared with OAI-ORE and the problems of lossiness that arise when attempting to automate the mapping between METS and OAI-ORE documents.

Habing and Cole [15] have also written a discussion paper analysing approaches for describing ORE Aggregations in METS. They determine that the most useful mapping is from the richer, more flexible, OAI-ORE Resource Map to a METS document, in order to make Aggregations more accessible to METS-based applications and tools. This is the approach that we have adopted in the context of this project. In particular, we provide an additional service that generates a METS document from the OAI-ORE Aggregation – to enable exchange and re-use of our objects with TARDIS.

3. System Architecture

Figure 2 illustrates the sequence of steps in protein crystallography experimentation that lead to the determination of a new crystal structure. In addition, Figure 2 shows the corresponding information systems that have been designed to manage the data being generated at each step in the protein crystallography pipeline:

- TIMTAM – the laboratory information management system that captures output from target selection to crystallisation
- DIMER – the repository for X-ray diffraction images that are processed to determine the crystal structure.
- Protein Data Bank (PDB) – the repository for 3D protein crystal structures
- ACS/PubMed Journal Repository – the online repository of publications describing new protein crystal structures and function.

The following sub-sections provide an overview of the TIMTAM and DIMER systems that we have developed and describe an approach for publishing their contents as compound objects using OAI-ORE.

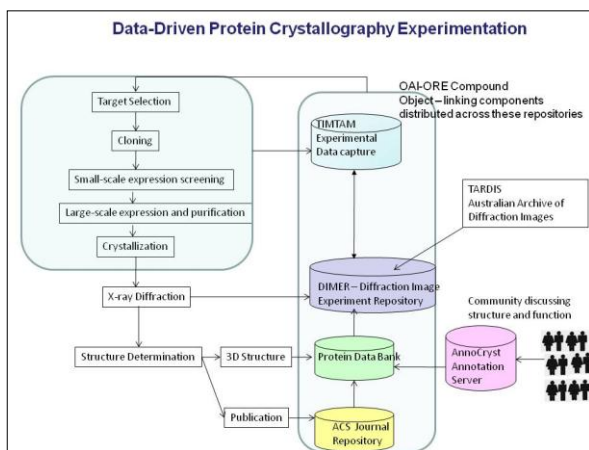


Figure 2. Overview of the system components

3.1. The TIMTAM Laboratory Information Management System (LIMS)

TIMTAM provides distributed teams of collaborating structural biologists with a secure Web application for target information management and tracking. These are the first steps in the protein crystallography pipeline. TIMTAM enables teams to record and share the results of target selection, protein preparation and protein analysis experiments. It has been designed to be readily adaptable and to support a wide variety of experimental workflows and parameters.

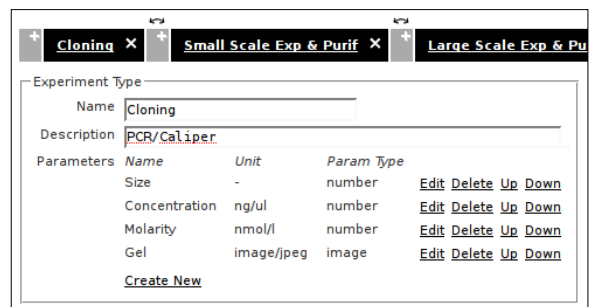


Figure 3. Defining workflows in TIMTAM

The first step in using TIMTAM involves the creation of a project by a user (the project owner) and the granting of read-only or read/write access to specific team members. Following creation of a project, the project owner defines a sequence of experiment types representing a workflow. Examples include 'Cloning', 'Expression and Purification' and 'Crystallization'. Each experiment type is defined by a set of parameters that can be numerical, textual, files or images. For example a 'Purification' experiment might be defined by 5 parameters: 3 numbers (yield, purity and molecular weight on SDS-PAGE) and 2 images (size exclusion chromatogram and SDS-PAGE image). The definitions of experiment types and their parameters are project-specific but can be modified during the course of a project, via a simple user

interface. Figure 3 shows the user interface being used to define a new experimental workflow.

The second step involves specifying the actual target protein and constructs that the project aims to solve. Target information can include sequences from multiple organisms, external identifiers that link to public databases (e.g., NCBI, Ensembl) as well as notes on target selection or additional file attachments. A "person in charge" is assigned to each target to reflect the project team's responsibilities. Constructs are created by specifying the full sequence or sub-sequences of the target protein, including primer information. Data entry for targets and constructs is assisted by automatic conversion between protein and DNA sequences, calculation of molecular weight, pI and crystallizability scores and by retrieving data from public databases.

A graphical interface based on microplate format enables users to drag-and-drop constructs into each well and define associated vectors. Experimental results are entered via online forms that are dynamically generated from the experiment type definitions. A Java client is also available that supports bulk upload of microplate experiment data from files and images on the desktop. Figure 4 illustrates a protein purification experiment whose results comprise two images and several analysis parameters. The header in Figure 4 specifies the originating well ("A5") and the microplate ("Plate #5") of the sample. The tick in the top right hand corner indicates that this experiment has been tagged as successful. The scientist has also provided a comment at the bottom interpreting the results.

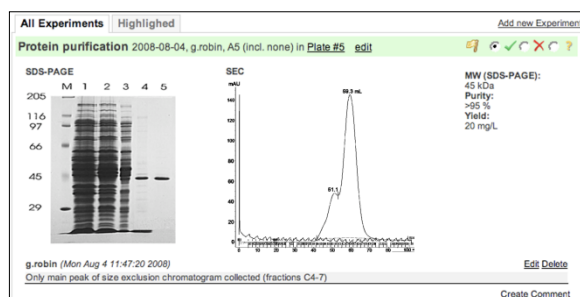


Figure 4. Experiment results in TIMTAM

Each construct is tracked by TIMTAM as it progresses through the experimental workflow. The user interface for monitoring progress is shown in Figure 5. Individual experiments are marked with a tick, cross, or question mark to indicate success, failure, or undetermined outcome. Initially, the table displays a single icon indicating the most successful outcome among all the experiments for each target and each experiment type. However this can be expanded to show individual outcomes for each vector type (e.g., see the "Cloning" column in Figure 5). An important feature of the monitoring interface is the "E-value".

This is a measure of similarity between the target and sequences published in the PDB/ohPDB. It is updated weekly by downloading local copies of each database and using BLAST to measure similarities. If the E-value is below the threshold (default 1.0×10^{-10}), an email is sent to the target owner to notify them that a homologous target has been solved.

Target ^{YA}	Construct ^{YA}	Person ^{YA}	External gene ID ^{YA}	E-value ^{YA}	Cloning ^{YA}		Small Scale Exp & Purif ^{YA}
					pMCSG7	pMCSG19C	
R02	R02m1	Gautier		1.6	✓	✓	?
R02	R02m2	Gautier		7.0	✓	✓	✓
R03	R03m2	Gautier	Cd274	5.0e-91	✓	✓ ? X	✓
R03	R03m3	Gautier	Cd274	1.0e-85	✓	✓	✓
R04	R04h1	Gautier		0.022	X	X	X
R04	R04h2	Gautier		0.02	X X X X	X X X X	X
R04	R04h3	Gautier		0.021	✓	✓ X X X	✓
R04	R04m3	Gautier		0.19	✓	✓	✓
R05	R05h1	Gautier		1.0e-09	✓	✓	✓

Figure 5. Monitoring progress of constructs

TIMTAM documents precise experimental protocols via textual descriptions or by uploading external documents. These protocols describe the exact laboratory procedures for each experiment type.

Finally, a major advantage of TIMTAM over similar LIMS is the provision of a toolkit of software services to assist with target selection and construct design. Four tools are currently available via the TIMTAM toolbar: calculation of crystallizability scores; BLAST sequence matching to public databases; in silico PCR; and protein-to-DNA translation. These services ensure that the experimenters are not wasting time or expensive resources on targets that have already been solved or that are unlikely to crystallize.

3.2. DIMER – Diffraction Image Experiment Repository

Following the protein production and crystallisation steps that are recorded through TIMTAM, successfully generated crystals are then transported to the X-Ray diffraction lab for structure determination.

The Diffraction Image Experiment Repository (DIMER) is an online archive for raw diffraction images. It provides a secure online indexed storage of the images, prior to analysis, structure determination and publication. It also makes them accessible so they can be linked to from publications, searched by researchers, and integrated into other online databases.

A fundamental aspect of managing experimental data in DIMER is registration of users and the creation of groups and projects to which users are assigned permissions. DIMER distinguishes three roles: managers (e.g., group leaders) who can modify projects, experiments, groups and permissions; writers (e.g. group members) who can modify experiments; and readers (e.g., external collaborators) who can view experimental data but not modify it. Figure 6 shows the DIMER user interface; a project named DsbA is displayed with users assigned to the roles of manager

and writer, and a group representing users working in the Dsb Project are assigned read access. The project's experiments are also listed.

DIMER Logged in as jenny (Logout)

Home Projects Experiments Datasets Citations Users Groups Search

DsbA Project

Home » projects » dsba

View Edit

DsbA, a 21-kDa protein from *Escherichia coli*, is a potent oxidizing disulfide catalyst required for disulfide bond formation in secreted proteins. The active site of DsbA is similar to that of mammalian protein disulfide isomerases.

Created	12 March 2009
Published	3 August 2009
Updated	6 August 2009

Collaborators

- Jenny Martin (manager) Dsb Program (reader)
- Stephen Shouldice (writer)

This project has been published and is publicly accessible. Unpublish this project

Experiments

- DsbA His32 Mutants
- DsbA Wild Type

Displaying 1-2 of 2 [view all] [create new]

Figure 6. DIMER Web page showing a project

As shown by the green box in Figure 6, this particular project was published on August 3. Prior to that date, users without permission would not see the project in any search results. Once a project is marked as published, it is visible through the public search interface. Publishing can also be done at the level of individual experiments within projects – so that not all of the experiments within a project are exposed.

DsbA His32 Mutants Experiment

Home » projects » dsba » experiments » his32

Despite the dramatic changes in stability, the structures of all three oxidized DsbA His32 variants are very similar to the wild-type oxidized structure, including conservation of solvent atoms near the active-site residue.

Created	14 April 2009
Published	5 August 2009
Updated	6 August 2009

PDB IDs: 1FV, 1ACL, 1ACV

This experiment has been published and is publicly accessible. Unpublish this experiment

Citations

- Structural analysis of three His32 mutants of DsbA

DsbA, a 21-kDa protein from *Escherichia coli*, is a potent oxidizing disulfide catalyst required for disulfide bond formation in secreted proteins.

Guddat LW, Bardwell JC, Glockshuber R, Huber-Wunderlich M, Zander T, Martin JL

Datasets

- DsbA H32L Mutant
- DsbA H32S Mutant
- DsbA H32V Mutant

Displaying 1-3 of 3 [view all]

Figure 7. Diffraction image experiment

Figure 7 shows one of the experiments, 'DsbA His32 Mutants', within the DsbA project. This experiment groups together three diffraction datasets relating to different mutants of the DsbA protein. At the level of an experiment, users can specify the PDB IDs for deposited structures and cite journal articles that document the experiment. Typically an experiment will be published only after the related paper and structures have been published. As shown in Figure 7, the PDB IDs are used to generate links to the structure summaries in the Protein Data Bank; in addition, a Jmol applet is displayed, showing the 3D structure for the first PDB ID in the list. Citations are entered using their PubMed ID, which are retrieved using the Entrez

EFetch service [19] to obtain complete bibliographical information. Title, abstract, and author information is stored and indexed.

Images captured during the X-ray diffraction experiment are selectively uploaded from a temporary workspace to the appropriate experiment folder. Figure 8 shows the dataset for diffraction of the H32L mutant of DsbA. Metadata is automatically extracted from the headers of the raw images (see the table in Figure 8) and JPEG previews are generated (see the thumbnails in Figure 8). The extraction of metadata and JPEG images is achieved using the CCP4 Diffraction Image Library [20] and is handled by a background process on the server that listens for new files and is executed automatically.

DsbA H32L Mutant Dataset

Home » projects » dsba » experiments » his32 » datasets » h32l

Metadata

Created	15 April 2009
Updated	6 August 2009

Name	Value
Collection date	05 Jun 2005
Exposure time	1.551 s
Oscillation range	82.000° to 112.000°
Two theta	0.000°
Detector distance	200.000 mm
Beam centre	102.060 mm, 100.880 mm
Image size	209.715 mm, 209.715 mm
Pixel size	0.102 mm, 0.102 mm
X-ray wavelength	0.980 Å
Image type	SMV

Images

Four small thumbnail images of diffraction patterns are shown.

Figure 8. Diffraction image dataset

File transfer is a challenge for diffraction datasets, which may contain several thousand images. This was a major reason for choosing the Jackrabbit JCR implementation [21] as a data storage technology: Jackrabbit includes support for WebDAV, a protocol similar to FTP, and exposes repository nodes as directories and files. Permissions are implemented in Jackrabbit's access management layer, meaning the same restrictions on reading and modifying content are enforced in both WebDAV and the HTML interface. A Java applet is included on the dataset page that enables transfers within a browser environment, but users can also connect with other WebDAV client software.

3.3. Generating OAI-ORE Compound Objects

Both TIMTAM and DIMER use proprietary storage technologies: the former using a relational database, and the latter using a JCR repository. Exposing the data in these repositories for external consumption requires an independent and standard data model, data formats, and protocols for data exchange. Web-based formats and protocols provide an ideal architecture for linking the distributed data components within compound objects. The OAI-ORE data model and specifications (see Section 2.2) was deemed an ideal

candidate for integrating data generated via the protein crystallography pipeline.

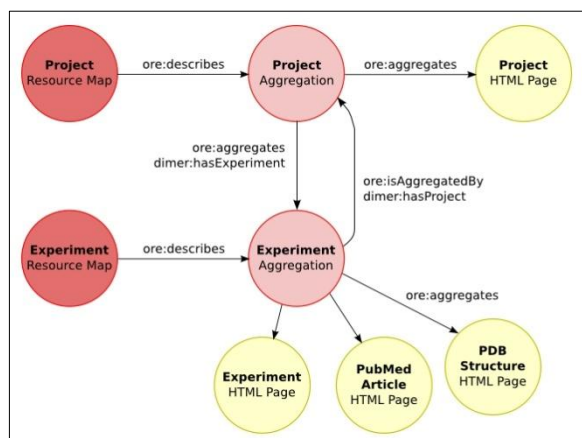


Figure 9. Project and experiment in OAI-ORE

Figure 9 shows the OAI-ORE representation for a DIMER project containing an experiment. The project Aggregation is described by its Resource Map that aggregates the project's HTML page. Similarly, the experiment Aggregation also has a Resource Map that aggregates HTML resources for the experiment and links PubMed articles and PDB structures. The project and experiment Aggregations are related using nesting: the Resource Map for the project includes an `ore:isDescribedBy` statement indicating that the aggregated resource for the experiment is itself an Aggregation described by the linked Resource Map. Similarly, the experiment's Resource Map includes an `ore:isAggregatedBy` statement that points in the reverse direction. Composing nested aggregations in this way allows large compound objects to be formed while providing for the identification and reuse of its individual components. Representation of *experiments*, *datasets* and *files* use a similarly nested structure.

The link between *projects*, *groups*, and *users* diverges from this hierarchical model. *Users* are represented using Aggregations and aggregated by *groups*, which define non-exclusive sets of users, and projects, which record the roles of collaborators. This exploits the graph-based nature of ORE.

DIMER and TIMTAM use concepts/terminology from the CCLRC scientific data model [18]. In addition, we extended the OAI-ORE vocabulary/data model to support the specific terminology needs of our protein crystallography users. For example, `dimer:Experiment` and `dimer:hasExperiment` were defined as a subclass and subproperty of `ore:Aggregation` and `ore:aggregates`. This enables external software to specifically process DIMER objects - or to fall-back on treating them as ORE aggregations. The use of formal OWL ontologies also supports general reuse: for example, if a Semantic Web application is aware of an `acme:Dataset` class and is supplied with an

`owl:equivalentClass` linking this with `dimer:Dataset`, then it can use inferencing to also integrate DIMER datasets. In the future we plan to ensure our ontologies are compliant with `oreChem` by refining/extending the ontologies generated by the `oreChem` project.

3.4. Generating METS

In order to support exchange and reuse of our data with METS-based crystallography applications, we have also developed a service to generate a METS encoding for experiments using the TARDIS schema [22]. This encoding contains metadata for not only the experiment, but also its datasets and diffraction images. Metadata sections are encoded as XML according to XML Schema definitions specified by the TARDIS project, with relationships contained in the standard `mets:structMap` element. Due to each dataset containing hundreds of images, the resulting XML file is several megabytes in size. There is no means of retrieving metadata for individual objects without processing the entire file; however, this encoding is suited to the bulk transfer of experiments, such as the harvesting performed by the TARDIS repository.

4. Evaluation

4.1. User Feedback

TIMTAM and DIMER have both been developed in collaboration with and extensively tested by members of the Structural Genomics team at UQ.

The user feedback to TIMTAM was highly positive. The SG team members liked the simple Web-based user interface (Figure 3 **Error! Reference source not found.**) that enables users with little or no programming skills to define new workflows and parameters that reflect new experimental protocols. The Web interface enables team leaders to monitor progress of experiments, any time and any where. Data input was streamlined and data quality was enhanced through the incorporation of additional services that assist with target selection and construct design. The services that were available through the TIMTAM toolbar included: calculation of molecular weight, pI and crystallizability scores; BLAST sequence matching to public databases; *in silico* PCR; and protein-to-DNA translation. Automated fetching of PDB E-values was seen as so valuable that users requested an email alert system to notify them as soon as the structure of a current or similar target is deposited in the PDB.

Improvements that the users did request included: the ability to upload and view TIFF and BMP images that represent experimental data; support for cyclical and parallel workflows (not just sequential workflows); the ability to print the details of an experiment, to stick in the lab notebook; support for multiple workflows

within a single project. These changes are currently being implemented.

User testing and feedback to DIMER generated the following list of requests that have since been added:

- The ability for users to selectively batch upload large numbers of images to the repository using a WebDAV plugin displayed within the browser;
- A simplified user interface for defining individual and group privileges and permissions;
- An improved search interface that enables simple and advanced searches for individual entities and/or compound objects, based on metadata fields (e.g., user/creator, projects, proteins, experiments, datasets, PDB IDs, citations/publications).

Users also requested the ability to visualize and edit the graphical structure (nodes and arcs) of protein crystallography compound objects, via an interactive visualization interface. We are currently adapting and customizing the LORE compound object authoring tool for this purpose.

4.2. Comparison of OAI/ORE and METS

OAI-ORE encompasses an abstract data model [8] (including machine-readable vocabulary [9]) for representing compound objects that comprise multiple related components as a Resource Map/Named Graph. OAI-ORE objects can be serialized in several data formats including: RDF/XML [10], RDFa [11], Atom [12]. The specification applies the principles of Linked Data [13] as the basis for HTTP implementation. The ability of OAI-ORE objects to dereference aggregation URIs and to reuse aggregations through nesting provide significant advantages, including enhanced discoverability and re-use. Because the underlying specification is based on Semantic Web principles, we maximize interoperability and semantic richness.

METS, on the other hand, is based on the concept of XML *packages* that express the hierarchical structure of complex digital library objects as a collection of files and associated metadata. These packages support the transmission and exchange of complex digital objects between digital library repositories. METS defines only a schema, leaving the question of transfer and discovery entirely up to other protocols such as OAI-PMH. This notion of a package is distinct from *resource maps* in OAI-ORE. Resource maps describe an object using a Named Graph in which each resource in the graph has its own identity, and hence is rediscoverable and re-usable. Whilst a METS document has an *object ID* and other files are identified using XML IDs in the context of the document, resources used in ORE are available for direct referencing through URIs.

To summarize, the advantages of OAI-ORE over METS include:

- The graph structure of OAI-ORE supports much more complex and richer structures than the more limited tree structure of METS;
- The URIs of nodes within OAI-ORE Named Graphs enable greater discoverability of atomic components within the compound object
- OAI-ORE objects offer greater extensibility and re-use – compared to the hard-wired rigidity of METS XML files
- The range of tools that enable the display and editing of RDF Named Graphs
- The ability to infer new relationships and knowledge based on RDFS/OWL ontologies used in OAI-ORE Resource Maps
- The separation of metadata from structure, components and presentation. METS objects embed the metadata, structure, components and display of the object in a single XML file.
- The comparative size of files – OAI-ORE representations are human-readable and of a reasonable size due to the ability to nest links to other OAI-ORE files. METS files on the other hand can be unwieldy because they contain everything within the one XML file.

5. Conclusions and Future Work

In this paper, we have presented a set of services and distributed repositories that we have developed to enable distributed teams of protein crystallographers to manage, monitor and “link” the increasing volumes of experimental data, derived data and publications they are generating via high-throughput experimental, robotic and imaging techniques.

We have described in detail:

- The secure, Web-based and flexible laboratory information system (TIMTAM) that we have developed - for recording and tracking the details and results of: target selection, cloning, expression, purification and crystallization steps in the wet lab.
- The DIMER repository for storing and publishing X-ray diffraction images generated from protein crystals.
- The OAI-ORE object creation services that enable the linking and publication of: experimental data, diffraction images, 3D crystal structures and publications, through well-structured, machine-processable and semantically rich compound objects.

We have also compared our OAI-ORE-based approach with similar approaches based on METS (e.g., TARDIS and eBank), and concluded that OAI-ORE provides a more flexible, extensible basis for representing networked scholarly and scientific

aggregations. However the OAI-ORE approach does not preclude exchange and interoperation with other “container” methods such as METS. In fact, it facilitates easy translation to METS representations to enable the exchange of data and interoperability with METS-based repositories and tools.

Future plans include developing an interactive compound object authoring interface for the protein crystallography community that enables users to selectively author and publish their own compound objects in searchable open-access repositories. In parallel with this, we plan to also collaborate with the oreChem project by extending and refining the ontology that they are developing for modelling artifacts of chemistry scholarship, so that it can also support the specific needs and vocabularies of protein crystallographers.

Acknowledgements

This research has been supported through an Australia Research Council (ARC) Discovery Project grant DP0770465, “Macrophage Proteins: Structure, Function and e-Science”. We would also like to acknowledge the valuable contributions made by Ronald Schroeter and Gregor Guncar to the development of TIMTAM. We would also like to acknowledge Karl Byriel for his assistance and for providing access to the UQ ROCX Diffraction Facility.

6. References

- [1] S. Androulakis et al, 'Federated repositories of X-ray diffraction images,' *Acta Cryst D*, June 18, 2008.
- [2] The UKOLN eBank UK project. <http://www.ukoln.ac.uk/projects/ebank-uk/>
- [3] M. Patel and S. Coles, "A Study of Curation and Preservation issues in the eCrystals Data Repository and proposed federation", September 2007 [http://www.ukoln.ac.uk/projects/ebank-uk/curation/eBank3-WP4-Report%20\(Revised\).pdf](http://www.ukoln.ac.uk/projects/ebank-uk/curation/eBank3-WP4-Report%20(Revised).pdf)
- [4] A. Tonge, P. Morgan, *Project SPECTRa*, Report, March 2007. <http://www.lib.cam.ac.uk/spectra/FinalReport.html>
- [5] C. Lagoze, "The oreChem Project: Integrating Chemistry Scholarship with the Semantic Web,". *WebSci'09*, Athens, Greece, 18-20 March 2009.
- [6] METS Editorial Board, *Metadata Encoding and Transmission Standard: Primer and Reference Manual*, Digital Library Federation, September 30, 2007. <http://www.loc.gov/standards/mets/METS%20Documentation%20final%20070930%20msw.pdf>
- [7] C. Lagoze et al, *ORE User Guide - Primer*, Open Archives Initiative, October 17, 2008. <http://www.openarchives.org/ore/1.0/primer.html>
- [8] C. Lagoze et al, *ORE Specification - Abstract Data Model*, Open Archives Initiative, October 17, 2008. <http://www.openarchives.org/ore/1.0/datamodel.html>
- [9] C. Lagoze et al, *ORE Specification - Vocabulary*, Open Archives Initiative, October 17, 2008. <http://www.openarchives.org/ore/1.0/vocabulary.html>

- [10] Beckett, D. and McBride, B. *RDF/XML Syntax Specification (Revised)*. W3C, 2004. <http://www.w3.org/TR/rdf-syntax-grammar/>
- [11] B. Adida, M. Birbeck, S. McCarron, S. Pemberton, *RDFa in XHTML: Syntax and Processing. A collection of attributes and processing rules for extending XHTML to support RDF*. W3C, 2008. <http://www.w3.org/TR/2008/PR-rdfasyntax-20080904/>
- [12] Nottingham, M. and Sayre, R. *The Atom Syndication Format*. Network Working Group, IETF, 2005. <http://tools.ietf.org/html/rfc4287>
- [13] C. Bizer, R. Cyganiak, T. Heath, *How to Publish Linked Data on the Web*, June 27, 2007, <http://www4.wiwiw.fu-berlin.de/bizer/pub/LinkedDataTutorial/>
- [14] J.P. McDonough, "Aligning METS with the OAI-ORE Data Model," *JCDL '09*, ACM, Austin, TX, USA, June 15-19, 2009, pp. 323-330.
- [15] T. Habing and T. Cole, "Candidate Approaches for Describing ORE Aggregations in METS", Preliminary Discussion Draft, 7 Jan 2009 <http://ratri.grainger.uiuc.edu/oremets/>
- [16] A. Gerber, J. Hunter, "LORE: A Compound Object Authoring and Publishing Tool for the Australian Literature Studies Community", *11th International Conference on Asia-Pacific Digital Libraries (ICADL) 2008*. Bali, Indonesia, December 2 - 5, 2008.
- [17] Cheung, K., Hunter, J., Lashtabeg, A., Drennan, J.: SCOPE - A Scientific Compound Object Publishing and Editing System, In: 3rd International Digital Curation Conference, Washington DC, 2007
- [18] S. Sufi, B. Mathews, *CCLRC Scientific Metadata Model: Version 2*, CCLRC, August, 2004. <http://epubs.cclrc.ac.uk/bitstream/485/>
- [19] Entrez EFetch Utility, Technical Overview. <http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi>
- [20] CCP4 Diffraction Image Library, Technical Manual. <http://www.ccp4.ac.uk/html/DiffractionImage.html>
- [21] Apache Jackrabbit. <http://jackrabbit.apache.org/>
- [22] S. Androulakis, *XML Schema*, March 18, 2009. <http://tardis.edu.au/schema/>