The International Journal of Digital Curation
Issue 1, Volume 1 | Autumn 2006

Scientific Publication Packages – A Selective Approach to the Communication and Archival of Scientific Output

Jane Hunter,
Professorial Research Fellow
University of Queensland

Abstract

The use of digital technologies within research has led to a proliferation of data, many new forms of research output and new modes of presentation and analysis. Many scientific communities are struggling with the challenge of how to manage the terabytes of data and new forms of output, they are producing. They are also under increasing pressure from funding organizations to publish their raw data, in addition to their traditional publications, in open archives. In this paper I describe an approach that involves the selective encapsulation of raw data, derived products, algorithms, software and textual publications within "scientific publication packages". Such packages provide an ideal method for: encapsulating expert knowledge; for publishing and sharing scientific process and results; for teaching complex scientific concepts; and for the selective archival, curation and preservation of scientific data and output. They also provide a bridge between technological advances in the Digital Libraries and eScience domains. In particular, I describe the RDF-based architecture that we are adopting to enable scientists to construct, publish and manage "scientific publication packages" - compound digital objects that encapsulate and relate the raw data to its derived products, publications and the associated contextual, provenance and administrative metadata.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. ISSN: 1746-8256 The IJDC is published by UKOLN at the University of Bath and is a publication of the Digital Curation Centre.

Introduction

Recent developments in digital technologies, experimental techniques and scientific instrumentation have changed the way that scientists work and led to an explosion in the rates of data generation in many disciplines. Simulations, observations, sensors, experiments and scientific instruments are currently capable of producing far more data than can possibly be analysed. Additionally, the range of "born-digital" research output is expanding and now includes data streams, images, video, audio, maps, complex arrays, algorithms etc. as well as traditional textual publications. Long-term accessibility to ever-increasing volumes and varieties of scientific data is essential to enable its re-use, maximize the potential derivable knowledge and reduce wasteful duplication. However many scientific communities are struggling with the challenge of how to manage the curation, archival and retention of the terabytes of data and information they are producing, often on a daily basis.

Digital librarians have been developing sophisticated technologies for indexing, storing, searching, retrieving and integrating mixed-media digital objects in both open access and access-controlled digital repositories. Digital library researchers have tended to concentrate on technologies to support digital objects at the scholarly publishing and e-learning end of the research chain, rather than the raw data being produced at the beginning of the chain. However the emerging eScience infrastructure is laying the foundation for new forms of intellectual products that require new modes of curation, publication and collaborative interaction. Scientific communities and their funding bodies, are already talking about the need for scientists to publish their raw data sets, experimental details, analytical methods and visualizations, in addition to the traditional scholarly publications. This record of the complete scientific discovery process will enable peers to review the method of conducting the science as well as the final conclusions. It will also enable greater sharing, re-use and comparison of scientific results, reduce duplication and insure against data loss because the additional contextual and provenance information will ensure the repeatability and verifiability of the results.

However these new information formats present significant challenges to digital library researchers, who are used to dealing with file-based digital objects. In this paper I present an approach that uses scientific publication packages to provide a common understanding between both digital librarians and scientists. Scientific Publication Packages (SPPs) provide a method for linking the raw data, its associated contextual and provenance metadata and the derived information, knowledge and publications within a single package, that can be treated like any other, albeit complex, digital object. They also provide an ideal mechanism: for authenticating and tracking individual contributions to scientific collaborations; for publishing and disseminating scientific results; for integrating research into teaching; and for selective archival and preservation of scientific data.

This paper describes the high level architecture and some of the tools, services and technological approaches that we are developing to enable scientists to capture, index, store, share, exchange, re-use, compare and integrate scientific results through SPPs. The aim is to analyse and support the needs of a wide range of scientific communities, in order to expedite solutions to both discipline-specific and cross-disciplinary scientific problems. The scale and dynamic nature of the problem will be tackled by determining commonalities and differences across communities and

building the tools and services on top of an underlying, extensible object-oriented infrastructure – the Semantic Grid. The aim is also to encourage the publication and sharing of SPPs by providing tools that will enable their easy construction and submission to either open access or access-controlled institutional repositories. New initiatives such as the Science Commons are expected to deliver a set of standardized licences, designed specifically to fulfil scientists' needs, that scientists can attach to their data and results. The aim of these licences is to encourage sharing of scientific data whilst also ensuring protection of the associated intellectual property. Finally, by making it easier for scientists to store their scientific data and results in institutional repositories rather than personal workspaces, then (assuming the custodial institution accepts its archival, curatorial and preservation responsibilities), the chances of long term access to the potential knowledge held within the data will be greatly improved.

Related Work

A number of researchers have proposed the use of the scientific model concept for publishing scientific data and results and for documenting the lineage of scientific theories and advances.

Hill, Crosier, Smith, and Goodchild (2001) propose a content standard for describing computational models - the Content Standard for Computational Models (CSCM), developed in response to the needs of the Alexandria Digital Earth Project (ADEPT) at the University of California, Santa Barbara (UCSB). CSCM was designed to describe computational models that have adjustable variables and parameters and includes both the modelling software plus datasets. It does not include components such as workflows, detailed provenance information, animations, simulations and visualizations, documentation or publications. It also primarily focuses on environmental models.

Cavalcanti, Mattoso, Campos, Llirbat, and Simon (2002) also developed a high-level architecture for publishing scientific models. The authors acknowledge that the large majority of scientific problems requires the construction of models by combining existing multidisciplinary models or deriving new models from a collection of shared data and models. However the wide variety of possible data types (relational, object-oriented databases, mixed-media files, spreadsheets, Web sites) and model types (probabilistic models, numerical/theoretical and empirical) raises serious interoperability issues. Cavalcanti et.al. propose an architecture that enables publication of and access to data and programs through a Scientific Publication Metamodel (SPM) that provides improved metadata support. Few details are provided of the actual metadata fields or how the interoperability issues are overcome. Like Hill et al. (2001), Cavalcanti et al. only consider the software and data associated with computational models.

The CCLRC has also developed a Scientific Metadata model (Sufi & Mathews, 2004) that aims to provide a high-level generic model to describe scientific studies and associated datasets and that can be specialized to specific scientific disciplines. It uses an XML schema to define the metadata model. Implementing it as an ontology and including an explicit high level "event" class would enhance its usability, extensibility and semantic interoperability and enable the capture of precise provenance data and the inferencing of new implicit knowledge or relationships between datasets.

Coleman (2002a) aims to define and categorize scientific models by treating them as “works” based on the IFLA (International Federation of Library Associations and Institutions) Functional Requirements for Bibliographic Records (FRBR) model (IFLA Study Group, 1998) and Smiraglia's definition, i.e. "a work is the .... entity" (Smiraglia, 2001). Coleman (2002b) collates the physical and conceptual components of scientific models. The physical components include textual works, datasets, software and services. The conceptual components are the ideas that the model expresses and include the research foci, model type, mathematical functions, instrumentation, theories or hypotheses and a record of the modification history. Coleman (2002a) goes on to define a set of metadata terms based on Dublin Core plus additional facets that describe and index models to enable their discovery. However, to date, there does not appear to be an implementation or evaluation of the proposed metadata schema.

All of the above approaches are limited in some way. Many consider only computational models – that comprise only mathematical formulae, programs and datasets – and neglect other important components such as the animations, visualizations, textual documents and workflows that are necessary for the validation of the model and the repeatability of the results. The majority focuses on models from a single discipline or, if an approach does consider multi-disciplinary models, it neglects the importance of semantic descriptions and semantic mediation to support the interoperability of different models both within and across disciplines. They also do not provide precise descriptions of the lineage relationships between different components of the models. “Scientific Publication Packages” incorporate those components of scientific models described above, but also encapsulate: a) lineage relationships between components and b) semantic descriptions of the components.

A recent special issue of the International Journal on Digital Libraries on complex digital objects, includes several papers that focus on technologies to support the storage, management and dissemination of complex digital objects – not dissimilar to the Scientific Publication Packages that we are proposing in this paper. Lagoze, Payette, Shin, and Wilper (2005) describe the Fedora open source digital repository service, that is designed to manage complex digital objects (and the relationships between their components). It uses an RDF-based relationship model to represent relationships among digital objects and their components, to support distributed information networks such as the National Science Digital Library (NSDL).

The aDORe system (Van de Sompel, Bekaert, Liu, Balakireva, & Schwander, 2005) developed at the Los Alamos National Laboratory research library also provides a standards-based repository for managing and accessing complex digital objects. Objects are encoded in XML using the MPEG-7 DIDL (Bekaert, Hochstenbach, & Van de Sompel, 2003) and a limited set of object relationships can be expressed using RDF.

Within this paper, the focus is on the tools required to construct and then publish scientific publication packages through ingestion within repositories such as Fedora, aDORe or DSpace. In particular, two things are essential to the construction, publishing, re-use and preservation of scientific models: a) capturing the complete set of contextual or lineage information associated with the model and b) capturing semantic descriptions of each of the individual components that comprise a scientific model and the relationships between them. These two aspects are discussed in more detail in the next two sub-sections below.

The Importance of Workflows and Lineage in Constructing Models

Workflow technologies represent an increasingly important component of the scientific process. They capture the chain (or pipeline) of processing steps used to generate scientific data and derived products. They also enable scientists to describe and carry out their experimental processes in a repeatable, verifiable and distributed way and to track the source of errors, anomalies or faulty processing. Consequently, a number of international research groups are concentrating on developing workflow specification and enactment systems that allow scientists to easily define, save, edit, share and re-use their workflows.

Although scientific workflows differ from business workflows, recent systems are based on BPEL4WS (Business Process Execution Language for Web Services) (Andrews et al., 2003) and graphical interfaces that enable users to combine and orchestrate a number of Web Services (both local and remote) in order to carry out a higher-level complex scientific task or experimental process. For each workflow instance there is a business process written in BPEL4WS and an associated WSDL (Christensen, Curbera, Meredith, & Weerawarana, 2001) file that describes the interface that the process will present to clients (plus WSDL documents that describe the services that the process will invoke during its execution). The BPEL4WS process itself is basically a flow-chart representation of an algorithm or set of processing steps. When the sub-workflows are deployed to the BPWS4J (IBM, 2004) engine, they are treated as web services and invoked accordingly.

The ability to compose web services dynamically is increasingly important as eScience becomes more collaborative and distributed, relying on geographically distributed groups of scientists working together to capture, share, correlate and analyse large-scale data sets in order to solve complex problems. As situations change and processing and analytical tools improve, scientists want to be able to discover and invoke the optimum combination of web services for their current task. Three examples of significant open source workflow systems that are based on dynamic web service composition and are designed specifically to support eScience, are the MyGrid project's Taverna (n.d.), Kepler (Altintas, Berkley et al., 2004) and YAWL (Aalst, Aldred et al., 2004).

One of the major aims of such web service-based workflow systems, is to relieve the effort required to capture the precise provenance metadata demanded by scientists in order to validate scientific results and enable their duplication. Our objective is to exploit these predefined workflow instances and the associated captured metadata to determine precisely the lineage of the data and its products, and to use this metadata to streamline the construction, description and archival of scientific models. Assuming appropriate metadata is being captured at each stage in the workflow associated with scientific model development, then many of the relationships between the components of a scientific model are either explicitly captured or can be inferred later, as required.

It is important, at this stage in the discussion, to highlight explicitly the difference between workflow and lineage. As Bose and Frew (2005) articulate, "Workflow is prospective in nature and defines plans for desired processing. Lineage on the other hand is retrospective (like an audit trail) and describes the relationships between data products and data transformations after processing has occurred." Thus in addition to source observations or information, the lineage of data product encompasses data acquisition and compilation methods, conversions, transformations and analyses, along with the assumptions and criteria applied at any stage of the data product life cycle (Clarke & Clarke, 1995). Capturing precise lineage data can be a very complex process, particularly if the metadata captured at each stage during the workflow is inadequate or ambiguous.

The ABC model is an "event-aware" model designed to enable the precise recording of life cycle events for digital objects in the library, archives and museum domains (Lagoze & Hunter, 2001). Figure 1 illustrates the class hierarchy for the ABC model. States represent the set of relevant digital objects that are input to and output from Events. The ABC model also uses the IFLA FRBR Work, Expression, Manifestation and Item concepts in order to link sets of resources (Manifestations) to Expressions of common intellectual content (a Work). Although originally developed for cultural and library resources, the ABC model, if extended, can be used to precisely capture the provenance or lineage of scientific models and provides an ideal top-level ontology for defining the classes and properties associated with scientific models. This is discussed in more detail in the section *Extending the ABC Ontology to Describe Scientific Models* below.

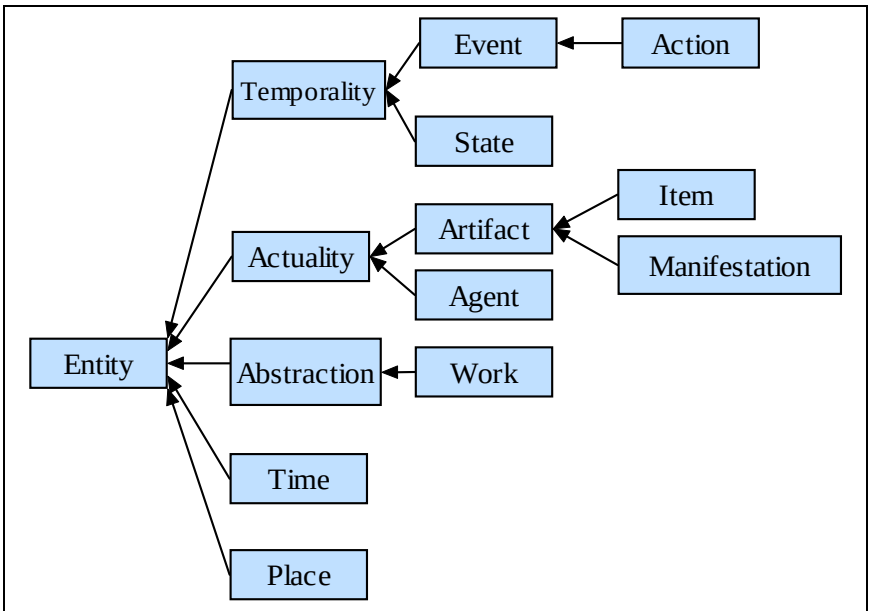

Figure 1: The ABC Model Class Hierarchy

The Importance of Semantics

Scientists need to be able to discover, re-use and compare SPPs and their components (e.g., processing, analytical, visualization services) – both within and across disciplines. They want to be able to combine models and model components to form new improved more complex models. They need to be able to detect or be notified when new data or improved processing services, of relevance to their models,

become available. Intelligent integration of the highly heterogeneous data and services, described using multidisciplinary metadata vocabularies, requires Semantic Web technologies such as the Resource Description Framework (RDF) and ontologies to provide the necessary semantic mediation.

The Resource Description Framework (RDF) is ideal for representing, navigating and querying highly interlinked networks of resources. It is ideal for specifying the semantic relationships between agents (human and software), data, resources and services, that comprise provenance logs. It provides an explicit unique identification system for resources (through URIs). It uses a graph-based model for relating resources which is more realistic than the tree model of XML and it provides a well-defined association to ontologies. Zhao, Wroe et al. (2004) demonstrate how RDF can be used to create a Semantic Web of Provenance Data.

Ontologies provide the semantic agreement necessary to enable information to be integrated across communities. They provide a machine-processable way of representing the *meaning* of a model or its components so it can be more easily discovered and re-used. OWL (Ontology Web Language) (McGuinness & Harmelen, 2004) descriptions of models and model components will be necessary to enable: semantic interoperability and comparisons between models; the detection of relationships, overlaps, conflicts or inconsistencies between models; and the amalgamation of models to generate better discipline-specific models or multi-disciplinary models. More specifically they will be required to describe, relate and enable interoperability between:

1. different types of scientific models (e.g., computational, logical, stochastic, deterministic, conceptual, graphical (2D and 3D));
2. discipline-specific models e.g., environmental models, chemical models, hydro-dynamic models;
3. the full range of Grid resources (agents/people, data, hardware (computers), scientific instruments, software and grid/web services, networks, storage systems etc) used to generate and refine the models.

Given the ontological descriptions of models and their components, combined with machine-processable inferencing rules (such as RuleML (Rule Markup Initiative, n.d.) and SWRL (Horrocks et al., 2004)), we have an infrastructure capable of advanced knowledge mining and reasoning services. Examples include obsolescence detection services, notification agents, discovery agents and invocation agents that automate the semantic matching, composition and invocation of services required to maintain, preserve, combine and reproduce the scientific models and associated data sets. Moreover, ontology-based browse interfaces, such as the Haystack semantic web browser (Zhao, Goble, Stevens, & Bechhofer, 2004) or the graph-based open source data visualization package, JUNG (Java Universal Network/Graph Framework, n.d.), will enable visualization of the semantic relationships between the components within SPPs. These visualizations will illustrate the provenance of the components, reveal contributions by individual team members and enable easier comparison between SPPs.

Requirements Analysis

An Illustrative Example

In this section, we describe a typical example of a scientific publication package and the process of developing this package. We use an example that crosses both the library domain and the scientific domain. Libraries and archives that are responsible for the long-term preservation and accessibility of electronic records, are very interested in factors that affect the longevity of data stored on particular electronic storage media, such as CD-ROMs.

Aim/research focus.

A pilot study is established to evaluate the life expectancy of pre-recorded compact discs (CD-ROMs). The aim of the study is to understand the factors that influence the Life Expectancy (LE) of CD-ROMs and, it is hoped, predict the average LE of CD-ROMs under different conditions.

Related and prior work.

Temperature and humidity are well known to be key factors that affect the LE of CD-ROMs. Their effect can be modelled using various techniques including acceleration models. Acceleration models predict "time to fail" as a function of operating stresses. Two common acceleration models, derived from chemical kinetics, are the Arrhenius and the Eyring models. These provide good potential starting points for developing a model. References to publications describing these models should be included in the scientific publication package – either as direct links or possibly as an EndNote file. As more scientific publication packages are published, then new models will refer to these or the associated raw data, rather than the traditional scholarly publications.

Experimental design, processes and data capture.

Two hundred pre-recorded compact discs were randomly sampled. The CDs were subjected to environmental stress conditions (temperatures of 60, 70 and 80 °C, relative humidity (RH) from 55-85%) over a time period of 500-1000 hours. The rate of deterioration of each specimen was determined by measuring block error rate (BLER) and by carrying out microscopic and chemical analyses of the CD-ROMs.

Different groups and individuals were responsible for different aspects of the study. One group conducted the experiments and captured the experimental data. Another group performed the microscopic image analysis and spectrometry. A third group was responsible for the data analysis and model fitting. An experimental design and workflow instance is defined using graphical workflow specification tools. The output is a BPEL4WS representation of the processes.

The microscopic images are captured using a Zeiss STEMI Apo binocular microscope with a Media Cybernetics Evolution camera. The images are analysed using ImagePro MC image processing software. Microchemical analyses of the degraded areas are carried out using FTIR (Fourier Transform Infrared) and Mass Spectrometry.

The complete process generated a database containing the following data for each of the 200 CD-ROMs: Identifier, Temperature, Relative Humidity, TimeUnderStress, BLER. In addition, for each CD-ROM there is a corresponding TIFF image showing the degradation plus spectrometry data and chemical analyses of the degraded region (Figures 2 and 3 below).

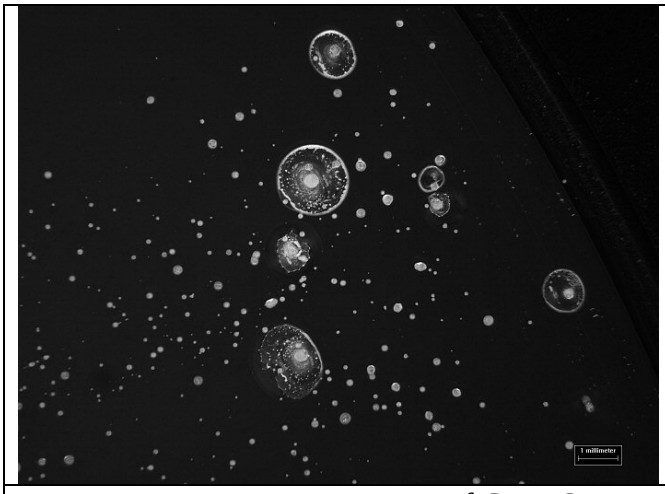

Figure 2: Microscopic Image of CD-ROM

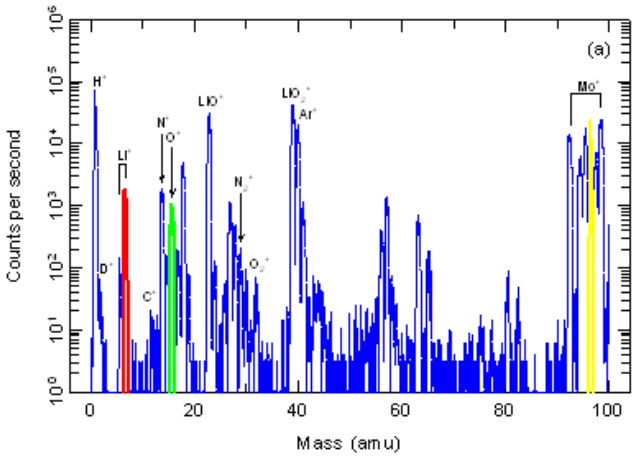

Figure 3: Spectrograph of CD-ROM

Model Fitting and Refinement.

The data was then analysed and plotted using the R statistical analysis package. The graphical results were saved as GIF images. The estimated "time to fail" for each disc subjected to a particular stress condition was compared against the two existing relevant models (the Arrhenius and Ehring models) (Figure 4). It was determined that the Ehring model provided the best fit to the experimental data:

Average LE = $1/T \exp^{-(A-B/T)}$

(where A and B are model parameters determined from the actual empirical results.)

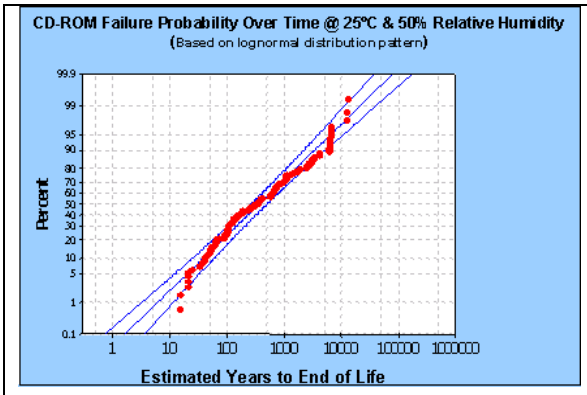

Figure 4: Life Expectancy of CD-ROMs Under Stress[1]

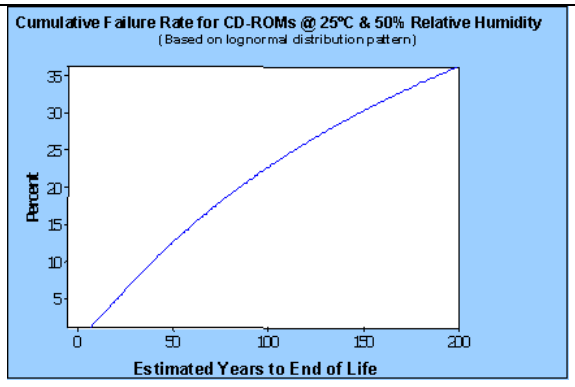

Figure 5: Model Predictions of Life Expectancy

Applying the refined Ehring model enabled the prediction of end of life estimates for the CDs at different temperature and RH conditions.

Testing the Model.

[1] Figures 4 and 5 courtesy of Chandru Shahani and Vivek Navale (See acknowledgements for source)

A further series of tests were carried out in order to compare real empirical data against predicted data generated using the refined Eyring model. This generated further sets of data, graphical results and slight refinement of the model parameters, A and B.

Publishing the Results.

A paper was then published outlining the results of the study:
Slattery, O., Lu, R., Zheng, J., Byers, F., Tang, X. "Stability Comparison of Recordable Optical Discs- A study of error rates in harsh conditions," Journal of Research of the National Institute of Standards and Technology, 109, 517-524, 2004

The Scientific Discovery Process

The example above illustrates the typical set of steps associated with the scientific discovery process - and the development of a scientific publication package. In many cases, the sequence of events can be generalized and simplified to the following:

- Inception of the idea;
- Discovery, retrieval and analysis of prior, related work;
- Experimental design;
- Capturing the empirical and observational data;
- Uploading the data to a database and recording associated descriptive metadata;
- Analysing, processing, interpreting and annotating the data;
- Formulating an hypothesis and/or constructing conceptual and/or numerical models (that are analogous or predictive and often take the form of a mathematical relation);
- Verifying and refining the hypothesis and/or model by capturing further experimental data and comparing it with data predicted using the model;
- Documenting and publishing the findings (with links to the data, hypothesis and model).

It should be noted that many scientific studies are not comprised of a single sequential pipeline and involve a complex web of parallel, linked, cyclical and intersecting workflows.

At each step in the workflow, there are different inputs, outputs, tools, assumptions, constraints, conditions and participants. Ideally the precise details are recorded by the associated metadata capture and digital curation tools that are part of the established workflow. In addition, when capturing all of the relevant information associated with each step, it is essential also to capture the relationships between each of the components. In the next two sections we:

- describe the range of components that comprise a Scientific Publication Package;
- propose an approach to enable the relationships between components to be inferred;
- construct an SPP whose internal structure reflects the relationships between its components.

The Components of a Scientific Publication Package

Scientific Publication Packages are complex, composite digital objects which encapsulate a variety of related heterogenous components. As illustrated in the

example in the Section *An Illustrative Example*, they may contain any of the following components or references to them:

- Pre-existing data, models, hypotheses or publications;
- Large datasets generated from experiments, observations and instruments. These may include: numerical data, survey data, questionnaires, images, video, audio, maps, spectral data, real-time sensor data;
- Experimental and instrumental conditions, settings and parametric ranges or constraints;
- Assumptions made and criteria applied;
- Formulas, rules, hypotheses, numerical models, mathematical functions;
- Conceptual models - paradigmatic, explanatory information or ideas in the form of axioms, models and metaphors;
- Software tools and services – that perform the analysis, interpretation, transformation, visualization, simulation and modelling of the data. This includes actual source code or executables, applets or links to web services as well as documentation describing the software;
- Hardware specifications – the instruments used to generate the data, the instrumental settings, and the computers that execute the analysis, processing, integration and visualization of the data;
- Workflows – steps involved in transforming the raw data into knowledge;
- Visualizations – 2D, 3D imagery, graphs, tables, charts, diagrams, animations;

Extending the ABC Ontology to Describe Scientific Models

The ABC model (Lagoze & Hunter, 2001) was developed for the library, museum and archival domains to capture the events that a digital object undergoes during its lifecycle. We believe that the ABC model can be extended in order to capture the provenance or lineage of scientific output. It also provides an ideal top-level ontology for defining the classes and properties associated with scientific outputs and their components. Figure 6 illustrates the class hierarchy for the extended ABC model that we have developed to support eScience provenance. The new classes are shaded in yellow. Associated with each of these new sub-classes are a set of properties specific to that sub-class.

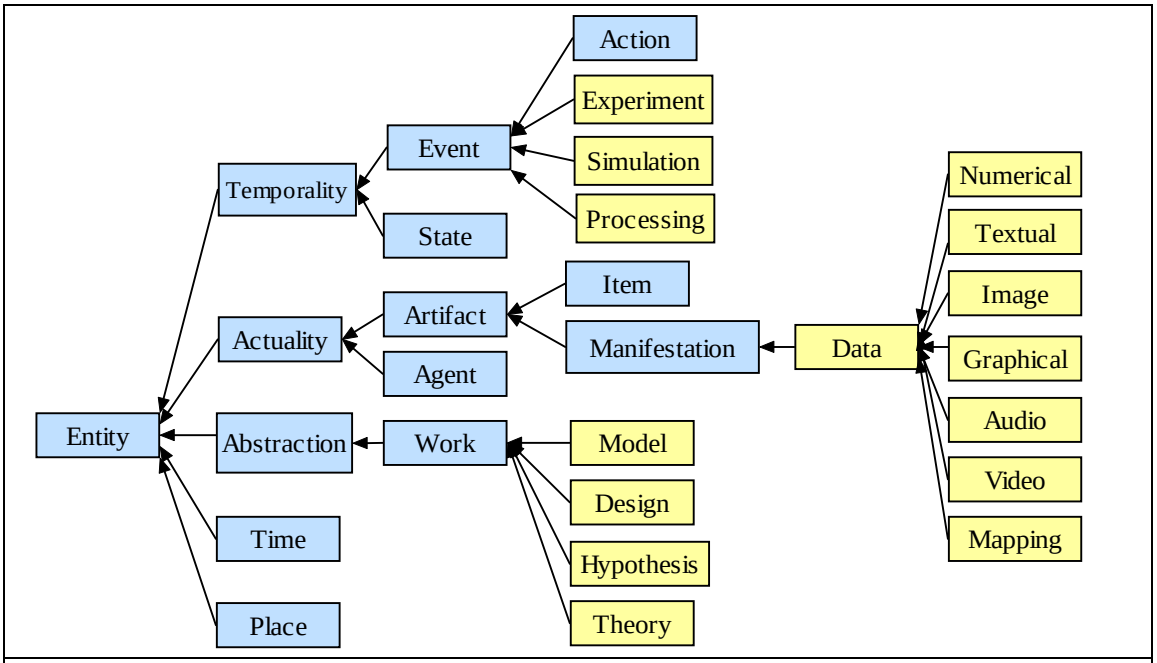

Figure 6: Extensions to the ABC Model Class Hierarchy for eScience Provenance

The Resource Description Framework (RDF) provides a number of advantages for representing scientific model packages and for recording the relationships between the components. RDF instance data provides XML-based descriptions of both the complete set of components (uniquely identified via URIs) within a scientific model package as well as the lineage (e.g., derivation, temporal, spatial, containment) and semantic relationships between these components. Alternative XML-based representations such as METS (Metadata Encoding and Transmission Standard, n.d.) and the MPEG-21 DIDL (Bekaert et al., 2003) provide syntactic interoperability, but do not provide the necessary semantic interoperability or the ontology-based reasoning that can be applied to objects described using OWL. The self-describing nature of RDF and OWL models also enable flexible descriptions for data collections, suiting those whose schemas may evolve and change, or whose data types are hard to fix, like knowledge bases of scientific hypotheses, provenance records of in silico experiments or publication collections (Zhao, Wroe et al., 2004).

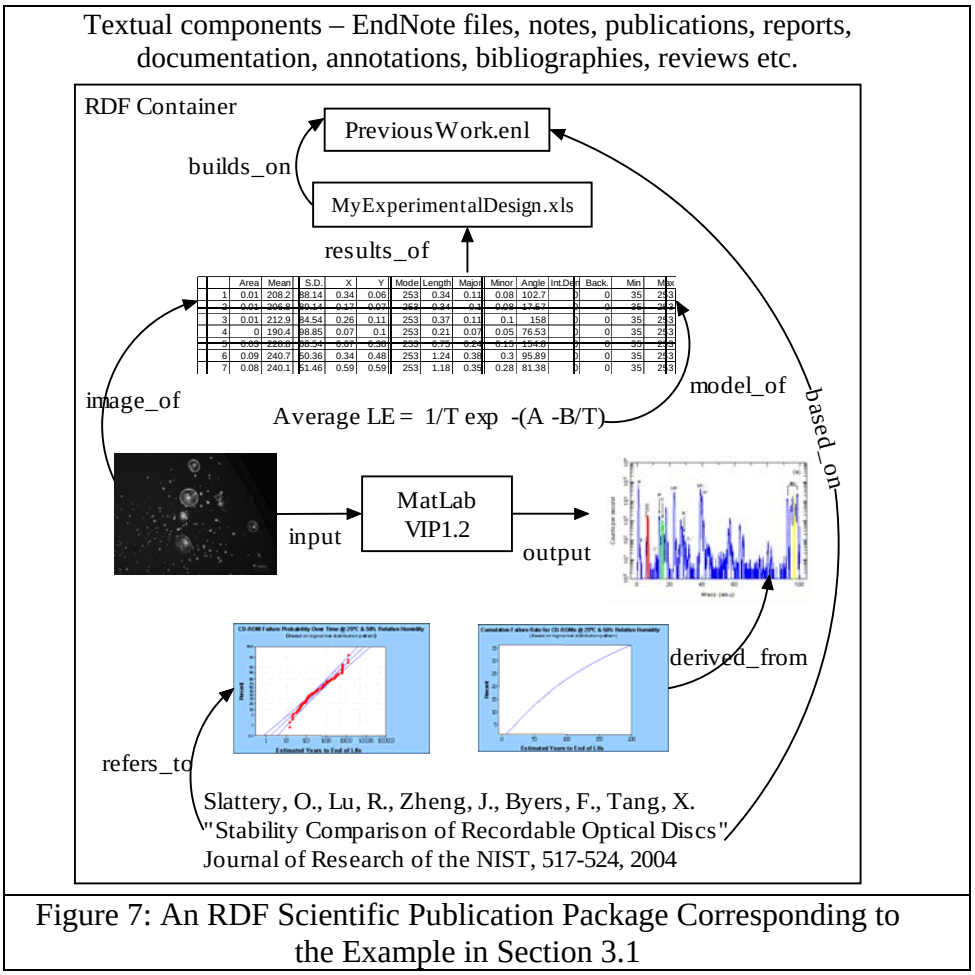

Figure 7: An RDF Scientific Publication Package Corresponding to the Example in Section 3.1

Constructing and Publishing a Scientific Publication Package

Let us consider the steps and tools required to enable a scientist to construct and then publish a new Scientific Publication Package (SPP):

1)At the start of the project, the scientist requires a logical collection area within his/her own private Workspace area in which to put all of the working data (numerical data, spreadsheets, notes, drawings, images, spectrometry, graphs, tables and publications) generated as part of a particular project. This step involves the creation of a "Project" folder within the scientists' Workspace area as well as a parallel folder in a shared workspace area. Observational and experimental data may be stored in distributed databases capable of handling both small datasets generated by scientists and large scale datasets generated by sensors or instruments (e.g., SRB (Storage Resource Broker)). If the experimental data is stored in remote databases, then the scientist needs methods for referring to subsets of these databases (i.e., specific rows, columns, tables) from their local resources and annotating these.

2) The scientist then goes through the set of steps described in Section 3.2. At each stage in the model development, the RDF metadata store captures specific metadata associated with each event, including the agents (human or software), inputs, outputs, tools, instruments, settings, constraints, time, place etc.

3) At the end of the scientific discovery process, the scientist decides to publish his/her SPP.

4) The components to be incorporated within the model must be specified. These can either be included as references (to the unique identifier) or actual bitstreams incorporated within the package. The scientist is provided with tools that allow him/her to specify the precise components including:

a. Data: database values, images, visualizations, graphs;

b. Mathematical functions represented in MathML: input variables, output variables, constants, constraints;

c. Software specifications (source code, executables, applets or links to web services);

d. Textual documents (EndNote files, notes, reports, documentation, annotations, publications)

5) The Scientific Publication Package (SPP) is then generated. It is a compound digital object represented as an RDF package. The relationships between the atomic objects within the compound object are either explicitly defined during the metadata capture, inferred from the rules associated with the ontology, or defined by the scientist during the SPP specification.

6) Descriptive metadata for the SPP is input and validated. It is envisaged that this metadata set could be based on the extensible CCLRC Scientific metadata model:

- Identifier
- Title
- Research focus/Topic
- Study
- Model type (drawn from a hierarchical thesaurus)
- Creator/Investigator – name and contact details, organization etc.
- Date Created
- Date Published

7) The creator/author attaches a Science Commons (Science Commons Initiative, n.d.) licence (selected from a menu of licence templates) to the SPP

8) The SPP object can then be ingested and saved to a DSpace (Dspace Federation, n.d.) or Fedora (Fedora, n.d.) digital library/institutional repository.

Figure 8 illustrates the various different storage areas envisaged in an ideal scientist's environment, and the relationships between them.

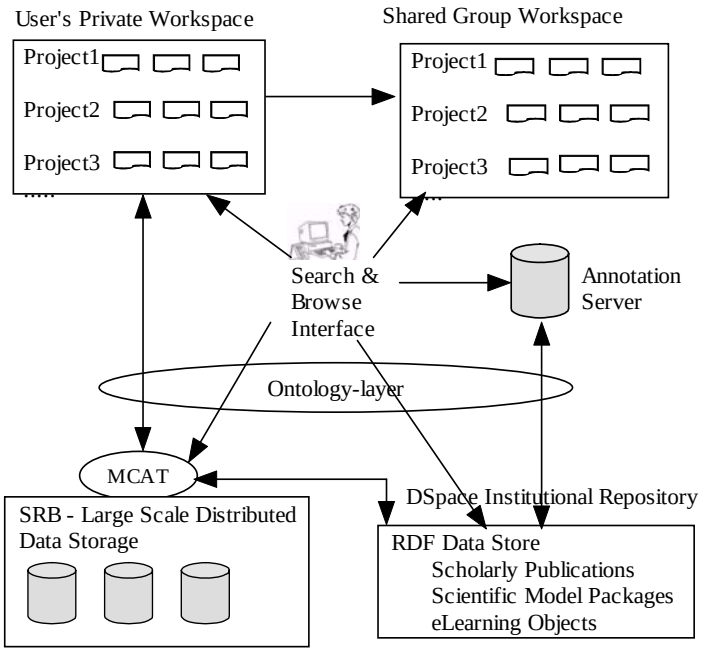

Figure 8: Envisaged Storage Area Requirements for eScience

The Preservation of Scientific Publication Packages

In (Hunter & Choudhury, 2005) we describe the PANIC (Preservation webservices Architecture for Newmedia and Interactive Collections) system in detail. It is an integrated, semi-automated preservation system based on a Semantic Web services architecture. As illustrated in Figure 9, PANIC comprises three main software components:

1. Preservation Metadata Capture. This comprises tools which enable the generation of preservation metadata for either atomic or composite mixed-media digital objects. Details of the metadata schema and input tool are provided in the next section. The preservation metadata can be saved in either an extended METS schema (Metadata Encoding and Transmission Standard, n.d.) or an MPEG-21 Digital Item Declaration Language (DIDL) schema (Bekaert et al., 2003).
2. Obsolescence Detection and Notification. This software component periodically compares each object's/sub-object's preservation metadata with software and format registries (e.g., PRONOM) which store information about the latest available authoring, rendering or viewing software and recommended formats. When there is incompatibility between an object's/sub-object's format and the latest available software or format recommendation, a notification is sent to the relevant agent (human or software). The EU CASPAR project (Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval, n.d.) is implementing a similar tool based on the UK DCC Representation Information Registry. Quantitative risk assessment methodologies such as VRC (Virtual Remote Control) or INFORM (INvestigation of FOrmats based on Risk Management), could also easily be incorporated to quantify the risk and trigger the notification.

3. Preservation Service Discovery and Invocation. When preservation action is required, the system allows the collections manager to specify the attributes of the required preservation service. A Discovery Agent then dynamically discovers the most appropriate preservation service by matching the specified attributes against descriptions of available preservation services. This is implemented by making preservation software modules available as Web services and describing them semantically using a machine-processable ontology (OWL-S, n. d.). Collections managers then have the option to choose from the ranked list of atomic or composite services retrieved by the Discovery Agent. Service Selection and Invocation Agents then select (and possibly compose) and invoke the most appropriate preservation services for that sub-object and update the provenance metadata.

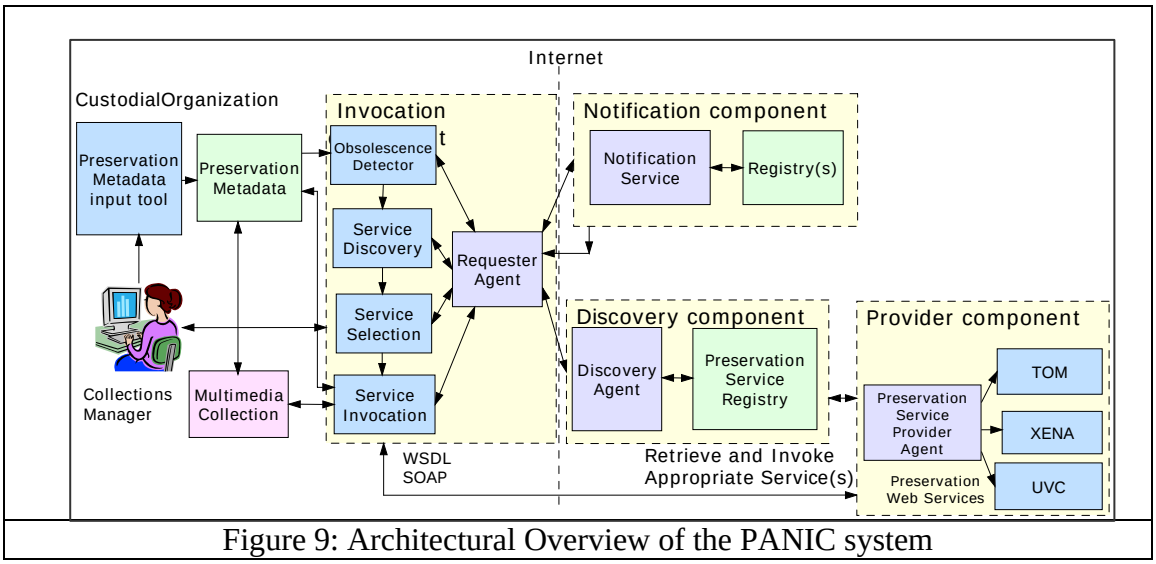

Figure 9: Architectural Overview of the PANIC system

Because we are using RDF to represent the SPPs, the process of monitoring their accessibility and discovering and invoking the optimum preservation service when required, is greatly simplified. The composite objects can be considered independently of the sub-objects and their formats, because we are using RDF only to specify the relationships between the atomic sub-objects. Hence, the system periodically considers the accessibility and preservation of each of the atomic sub-objects, prior to monitoring and processing the composite object. Our approach is to capture the preservation metadata for each of the sub-objects first, prior to defining the structure and capturing the preservation metadata for the composite object. Similarly with the obsolescence detection and notification and the preservation service discovery and invocation steps. We consider the atomic sub-objects first and only after each of the sub-objects has been dealt with is the preservation of the overall composite object considered. For example, all atomic JPEG images may first need to be migrated to JPEG-2000, after which the composite objects which contain these images may need to be migrated from RDF 1.0 to RDF 2.0.

Research Issues and Future Work

The work described in this paper is still at a relatively preliminary stage of development. We are currently in the process of working with a number of scientific communities (in the environmental, nanomaterials and molecular biology domains) in

order thoroughly to understand their scientific research procedures and associated data management requirements and the commonalities across communities.

Preliminary results of the user analysis indicate that some of the assumptions and generalizations that we have made in designing our architecture are an over-simplification of what really happens. Deployment and adoption within real communities will require modifications that take into account the complex reality of scientific discovery processes and associated issues such as non-sequential intertwining and cyclical workflows, highly heterogeneous datastreams, small scale data stored in local proprietary databases rather than SRB. In addition, a change of attitude of conventional publishers will be required in order to gain their acceptance for supporting new types of publications that include standardized raw and derived data formats.

In parallel with the detailed analysis of user needs and processes, we are also working on the:

- Development and evaluation of the extended ABC Ontology for eScience provenance logs;
- Implementation and evaluation of an eScience provenance logging system and database;
- Methods for determining or inferring relationships between selected components of an SPP, from the provenance logs;
- Development of the SPP construction and description tools;
- Development of a search, browse and retrieval interface to a repository of SPPs;
- New data citation methods that enable fine-grained references to raw data within SPPs and new citation monitoring and analysis tools that value re-use of data and workflows as highly as references to traditional publications;

In addition, we are continuing to track the outcomes of the Science Commons Initiative (Science Commons Initiative, n.d.). Science Commons is an exploratory project (focussing on three project areas: Publishing, Licensing and Data) that aims to apply the philosophies and activities of Creative Commons to the realm of science. In particular, the Science Commons Licensing sub-project is exploring standard open agreements to facilitate licensing of intellectual property and the exchange of research materials. Our aim is to provide tools to enable scientists easily to attach the emergent Science Commons licences to SPPs and their components when they want to share them - without sacrificing intellectual property rights.

Conclusions

Scientific progress depends on speedy and open access to the full spectra of scientific data and derived products. A recent OECD report on the scientific publishing industry (Houghton & Vickery, 2005) recommends that governments make publicly funded research findings more widely available in order to boost innovation and get a better return on their investment. Consequently scientists are under increasing pressure to publish their experimental and evidential data together with the related traditional scholarly publication(s). But the infrastructure required to support these new forms of scientific publishing is still immature and currently relies on an ad hoc assemblage of software that is inadequate for the task. The approach that I have proposed above

involves leveraging existing tools developed by digital librarians for atomic digital objects, but extending them to support the unique requirements of scientists and their new forms of scientific data and research output. Tools that precisely capture the provenance of resources generated during the scientific discovery process ensure the validity and repeatability of scientific results. At the same time, they provide a way of encapsulating the different components associated with a particular scientific advancement within a single compound document (i.e., a Scientific Publication Package) that can be published in an open institutional repository. This approach provides an efficient, integrated and sustainable science communication system that encompasses all forms of research output, and hence maximizes its re-use, dissemination and potential socio-economic benefits.

Acknowledgements

Figures 4 and 5 were reproduced from V.Navale “Predicting the Life Expectancy of Modern Tape and Optical Media” RLG DigiNews August 2005, where it was originally adapted from W.P. Murray and C. Shahani, unpublished Library of Congress draft report on the life expectancies of CD-ROMs. Used with permission and the kind generosity of Vivek Navale and Chandru Shahani.

References

Aalst, W. M. P. van der, Aldred, L., Dumas, M., & Hofstede, A. H. M. ter (2004). Design and implementation of the YAWL system. In: A. Persson & J. Stirna (Eds.), *Advanced Information Systems Engineering: 16th International Conference, CAiSE 2004, Riga, Latvia, June 7-11, 2004: Proceedings* (pp 142-159). Lecture Notes in Computer Science 3084. Berlin: Springer.

Altintas, I., Berkley, C., Jaeger, E., Jones, M., Ludascher, B., & Mock, S. (2004). Kepler: An extensible system for design and execution of scientific workflows. In *Proceedings of the 16th International Conference on Scientific and Statistical Database Management (SSDBM'04), Santorini Island, Greece, June 21-23, 2004* (pp. 423-424). Washington, D.C.: IEEE Computer Society.

Andrews, T., Curbera, F., Dholakia, H., Goland, Y., Klein, J., Leymann, F., et al. (2003). *Business Process Execution Language for Web Services, version 1.1*. IBM, BEA Systems, Microsoft, SAP AG, and Siebel Systems. Retrieved September 24, 2006, from http://www-128.ibm.com/developerworks/library/specification/ws-bpel/

Bekaert, J., Hochstenbach, P., & Van de Sompel, H. (2003). Using MPEG-21 DIDL to represent complex digital objects in the Los Alamos National Laboratory Digital Library. *D-Lib Magazine*, *9*(11). Retrieved September 24, 2006, from http://www.dlib.org/dlib/november03/bekaert/11bekaert.html

Bose, R., & Frew, J. (2005). Lineage retrieval for scientific data processing: a survey. *ACM Computing Surveys, 37*(1), 1-28. Retrieved September 24, 2006, from http://homepages.inf.ed.ac.uk/rbose/pubs/bose_2005_ACM_CS.pdf

CASPAR: Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval. (n.d.). Retrieved September 24, 2006, from http://www.casparpreserves.eu/

Cavalcanti, M. C., Mattoso, M., Campos, M. L., Llirbat F., & Simon, E. (2002). Sharing scientific models in environmental applications. In *Proceedings of the 2002 ACM symposium on Applied computing, Madrid, Spain* (pp. 453 - 457). Retrieved September 26, 2006, from the ACM Digital Library.

Christensen, E., Curbera, F., Meredith, G., & Weerawarana, S. (2001). *Web Services Description Language (WSDL)*. W3C Note. World Wide Web Consortium, 15 March. Retrieved September 24, 2006, from http://www.w3.org/TR/wsdl/

Clarke, D. G., & Clarke D. M. (1995). Lineage. In S. C. Guptill & J. L. Morrison (Eds.) *Elements of spatial data quality* (pp. 13-30). Oxford: Elsevier Science.

Coleman, A. (2002a). Scientific models as works. *Cataloging & Classification Quarterly, 33*(3/4), 129-159. Retrieved September 24, 2006, from http://www.sir.arizona.edu/faculty/coleman/papers/smascrev.pdf

Coleman, A. (2002b). A classification of models. In M. J. López-Huertas & F. J. Munoz-Férnandez (Eds.), *Challenges in knowledge representation and organization for the 21st century: Integration of knowledge across boundaries: Proceedings of the Seventh International ISKO Conference, 10-13 July 2002, Granada, Spain* (pp. 86-92). Würzburg: Ergon-Verlag. Retrieved September 24, 2006, from http://www.sir.arizona.edu/faculty/coleman/papers/iskoasc.pdf

DSpace Federation. (n.d.). Retrieved September 24, 2006, from http://www.dspace.org/

Fedora. (n.d.). Retrieved September 24, 2006, from http://www.fedora.info/

Hill. L., Crosier, J., Smith. T., & Goodchild, M. (2001). A content standard for computational models. *D-Lib Magazine,7*(6). Retrieved September 24, 2006, from http://www.dlib.org/dlib/june01/hill/06hill.html

Horrocks, I., Patel-Schneider, P. F., Boley, H., Tabet, S., Grosof, B., & Dean, M. (2004). *SWRL: A Semantic Web Rule Language combining OWL and RuleML*. W3C Member Submission. World Wide Web Consortium, 21 May. Retrieved September 24, 2006, from http://www.w3.org/Submission/SWRL/

Houghton J., & Vickery G. (2005). *Digital broadband content: scientific publishing*. Organisation for Economic Co-operation and Development, Directorate for Science, Technology and Industry, Committee for Information, Computer and Communications Policy, Working Party on the Information Economy. Retrieved September 24, 2006, from http://www.oecd.org/dataoecd/42/12/35393145.pdf

Hunter J., & Choudhury, S. (2005). Semi-automated preservation and archival of scientific data using semantic grid services. In: *Proceedings of the 2005 IEEE International Symposium on Cluster Computing and the Grid (CCGrid'05), Cardiff, UK, May 9-12, 2005*, Vol. 1 (pp. 160-167). Retrieved September 26, 2006, from the IEEE Xplore database.

IBM. (n.d.). BPWS4J. Retrieved September 24, 2006, from http://www.alphaworks.ibm.com/tech/bpws4j

IFLA Study Group. (1998). *Functional Requirements for Bibliographic Records: final report*. München: K. G. Saur. Retrieved September 24, 2006, from http://www.ifla.org/VII/s13/frbr/frbr.pdf

JUNG Java Universal Network/Graph Framework. (n.d.). Retrieved September 24, 2006, from http://jung.sourceforge.net/

Lagoze, C., & Hunter, J. (2001). The ABC Ontology and Model. *Journal of Digital Information*, *2*(2), article no. 77, 2001-11-06. Retrieved September 24, 2006, from http://jodi.ecs.soton.ac.uk/Articles/v02/i02/Lagoze/

Lagoze, C., Payette, S., Shin, E., & Wilper, C. (2006). Fedora: An architecture for complex objects and their relationships. *International Journal on Digital Libraries, 6*(2), 124-138.

McGuinness, D. L., & Harmelen, F. van. (2004). *OWL Web Ontology Language overview*. W3C Recommendation. World Wide Web Consortium, 10 February. Retrieved September 24, 2006, from http://www.w3.org/TR/owl-features/

Metadata Encoding and Transmission Standard. (n.d.). Retrieved September 24, 2006, from http://www.loc.gov/standards/mets/

OWL-S 1.1 Release. (n.d.). Retrieved September 24, 2006, from http://www.daml.org/services/owl-s/1.1/

Rule Markup Initiative. (n.d.). Retrieved September 24, 2006, from http://www.ruleml.org/

Science Commons Initiative. (n.d.). Retrieved September 24, 2006, from http://sciencecommons.org/

Smiraglia, R. P. (2001). *The nature of "a work": Implications for the organization of knowledge*. Lanham, MD: Scarecrow Press.

Sufi, S., & Mathews, B. (2004). *CCLRC Scientific Metadata Model: Version 2* (CCLRC Technical Report DL-TR-2004-001). Didcot: Council for the Central Laboratory of the Research Councils. Retrieved September 24, 2006, from http://epubs.cclrc.ac.uk/bitstream/485/csmdm.version-2.pdf

Taverna. (n.d.). Retrieved September 24, 2006, from http://taverna.sourceforge.net/

Van de Sempel, H., Bekaert, J., Liu, X., Balakiereva, L., & Schwander, T. (2005). aDORe: a modular, standards-based digital object repository. *The Computer Journal, 48*(5), 514-535. Preprint retrieved September 24, 2006, from the arXiv repository: http://arxiv.org/abs/cs.DL/0502028

Zhao, J., Goble, C., Stevens, R., & Bechhofer, S. (2004). Semantically linking and browsing provenance logs for e-science. In M. Bouzeghoub, C. A. Goble, V. Kashyap & S. Spaccapietra (Eds.), *Semantics for Grid databases: First International IFIP Conference on Semantics of a Networked World, ICSNW 2004, Paris France, June 17-19, 2004: Revised selected papers* (pp. 158-176). Lecture Notes in Computer Science 3226. Berlin: Springer.

Zhao, J., Wroe, C., Goble, C., Stevens, R., Quan, D., & Greenwood, M. (2004). Using Semantic Web technologies for representing e-science provenance. In S. A. McIlraith, D. Plexousakis, F. van Harmelen (Eds.), *The Semantic Web - ISWC 2004: Third International Semantic Web Conference, Hiroshima, Japan, November 7-11, 2004: Proceedings* (pp. 92-106). Lecture Notes in Computer Science 3298. Berlin: Springer.