Scientific Models – A User-oriented Approach to the Integration of Scientific Data and Digital Libraries

Jane Hunter
Professorial Research Fellow
University of Queensland
jane@dstc.edu.au

Abstract:
Many scientific communities are struggling with the challenge of how to manage the terabytes of data they are producing, often on a daily basis. Scientific models are the primary method for representing and encapsulating expert knowledge in many disciplines. Scientific models could also provide a mechanism: for publishing and sharing scientific results; for teaching complex scientific concepts; and for the selective archival, curation and preservation of scientific data. As such, they also provide a bridge for collaboration between Digital Libraries and eScience. In this paper I describe research being undertaken within the FUSION project at the University of Queensland to enable scientists to construct, publish and manage scientific model packages that encapsulate and relate the raw data to its' associated contextual and provenance metadata, processing steps, derived information and publications. This work involves extending tools and services that have come out of the Digital Libraries domain to support e-Science requirements.

1. 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. Long term accessibility to ever increasing volumes of scientific data is essential to enable its re-use, maximise 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 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. However, 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. Already, scientific communities and their funding bodies are talking about the need for scientists to publish their raw data sets, experimental details, analytical methods and visualisations, 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. It will reduce duplication and insure against data loss because the additional contextual and provenance information will improve 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 based on the *scientific model* paradigm that provides a platform for common understanding by both digital librarians and scientists. The “scientific model” concept provides 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.

Scientific models provide a primary method for representing and encapsulating expert knowledge in many domains, including particle and high energy physics, astronomy, biological sciences, earth sciences, chemistry and nano-materials. They provide simplified representations of real world objects, systems or phenomena – often in mathematical terms - enabling scientists to better predict how complex systems behave under different conditions or parameters, or change over time, in order to solve problems associated with a particular research focus. 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 some of the tools, services and technological approaches that we are developing within the FUSION project that will enable scientists to capture, index, store, share, exchange, re-use, compare and combine different scientific models. Section 2 describes the background and overall objectives of the project. Section 3 describes related work. Section 4 describes the results of our requirements analysis. Section 5 describes the

outcomes to date. Section 6 describes the major challenges and future work plans. The paper concludes with Section 7.

2. Background and Objectives

The FUSION project is a collaboration between the e-Research Group at the School of ITEE, the Centre for Microscopy and Microanalysis and the Australian Institute of Bioengineering and Nanotechnology at the University of Queensland, which is investigating the application of Semantic Web/Grid technologies to e-Science. Although the FUSION project started in 2004, the investigation of scientific model packages, as a means of selectively managing scientific data and publishing results, only began in 2005 and is at a relatively early stage.

The aim of the project is to analyse and support the data management requirements of a range of scientific communities, by developing innovative transparent information technologies that 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 (De Roure, Jennings & Shadbolt, 2005).

The aim of the project is multifold but involves developing tools and services that:

1. Capture the precise provenance data associated with the generation of scientific results and the scientific discovery process;
2. Make it easier for scientists to store and share their scientific data and results within institutional repositories rather than in personal workspaces;
3. Enable the easy construction of composite “scientific models” by linking and encapsulating related digital objects;
4. Enable the submission of “scientific models” to either open access or access-controlled institutional repositories and the attachment of high quality metadata;
5. Provide secure access through authentication, authorisation and access control mechanisms based on current standards;
6. Provide innovative search, retrieval and presentation interfaces that graphically illustrate the scientific workflow and relationships between data and derived products;
7. Enable the attachment of licenses to data and results that encourage sharing of scientific data whilst also ensuring protection of the associated intellectual property.
8. Provide curatorial and preservation services that will ensure long term access to the data and results, even as formats and software versions change over time.

Within this paper we focus on the work being undertaken to support items 3 and 4 above – the construction, storage and publication of scientific model packages.

3. Related Work

A number of researchers have proposed the use of the scientific models for publishing scientific data and results and for documenting the lineage of scientific theories and advances. In Section 3.1, I provide an example illustrating a typical scientific model.

Hill et.al. (Hill et al. 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 visualisations, documentation or publications. It also primarily focuses on environmental models.

Cavalcanti et al (Cavalcanti et al. 2002) also developed a high-level architecture for publishing scientific models. The authors acknowledge that a large majority of scientific problems require 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 (Hill et al 2001), Cavalcanti et. al. only consider the software and data associated with computational models.

In (Coleman 2002a) Coleman aims to define and categorise scientific models by treating them as “works" based on Smiraglia’s definition i.e., “A work is the intellectual content of a bibliographic entity”. In (Coleman 2002b) Coleman 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 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 (Coleman 2002a). However, there does not appear to be any implementation that evaluates the usefulness of the proposed metadata schema.

In addition, a number of XML markup languages have been developed or are under development for describing and exchanging mathematical models. CellML (Nielsen & Halstead 2004) is an open markup language designed to describe and exchange mathematical models of biological cellular and sub-cellular processes. The Predictive Model Markup Language (PMML) (Raspl 2004) is an XML-based language which provides a way for applications to define statistical and data mining models and to share models between PMML compliant applications.

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, visualisations, textual documents and workflows, that are necessary for the validation of the model and the repeatability of the results. The majority focus on models from a single discipline or if they do consider multi-disciplinary models, they neglect the importance of semantic descriptions and semantic mediation to support the interoperability of different models both within and across disciplines.

Our research to date has determined that 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.

3.1 The Importance of Workflows and Lineage

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 Processing Language for Web Services) (Andres & Cubera 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 (W3C 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 2002) engine, they are treated as web services and invoked accordingly.

The ability to dynamically compose web services 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 SCUFL (Taverna 2005), Kepler (Altintas et al 2004) and YAWL (Aalst et al 2004).

One of the major aims of such web-service based workflow systems, is to relieve the effort required by scientists 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 precisely determine 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 explicitly highlight the difference between workflow and lineage. As Bose and Frew (Bose & 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, by extending the ABC model , it 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 Section 5.3.

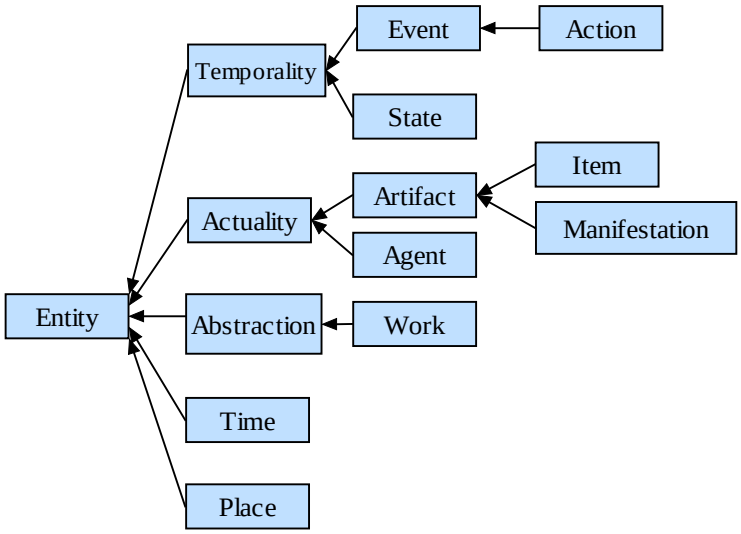

Figure 1: The ABC Model Class Hierarchy

3.2 The Importance of Semantics

Scientists need to be able to discover, re-use and compare scientific models and model components (e.g., processing, analytical, visualisation 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) (W3C 2004a) 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. In (Zhao et al 2004a), Zhao et.al. 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) (W3C 2004b) 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 (RuleML 2004) and SWRL (W3C 2004c)), 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 et al 2004b) will enable visualisation of the semantic relationships between the components within scientific models, illustrating their evolution over time and enabling comparisons between models and present individual collaborators' contributions.

4. Requirements Analysis

4.1 An Illustrative Example

In this section, we provide an example of a scientific model and describe the process by which this model was developed. We use an actual example that crosses both the library domain and the scientific domain.

Aim/Research Focus

Organisations (such as libraries and archives) that are responsible for the long-term preservation and accessibility of electronic records are concerned with factors that affect the longevity of data stored on electronic storage media, such as CD-ROMs. Hence a pilot study was established by the US National Institute of Standards and Technology (NIST) to understand the factors that influence the life expectancy of CD-ROMs and to predict the average life expectancy of CD-ROMs under different conditions.

Experimental Design, Processes and Data Capture

Two hundred pre-recorded compact discs were randomly sampled. Because temperature and humidity are well known to be key factors that affect the life expectancy of CD-ROMs, 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 was defined using graphical workflow specification tools. The output was a representation of the

processes in the BPEL4WS (Business Processing Expression Language for Web Services) format.

The microscopic images were captured using a Zeiss STEMI Apo binocular microscope with a Media Cybernetics Evolution camera. The images were analysed using ImagePro MC image processing software. Microchemical analyses of the degraded areas were performed using FTIR 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, a TIFF image (showing surface degradation) as well as spectrometry data and chemical analyses were captured (see Figures 2 and 3 below).

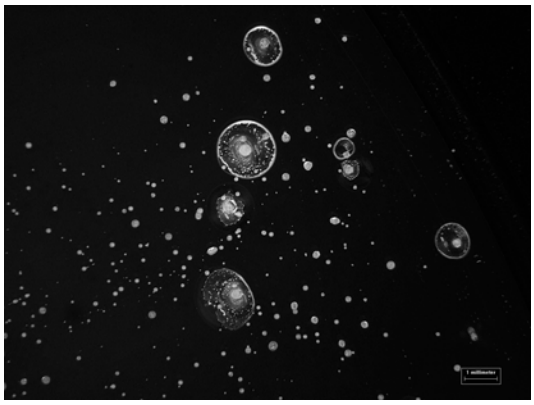

Figure 2: Microscopic Image of CD-ROM

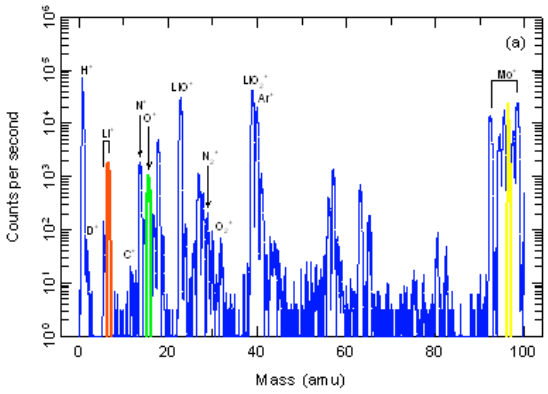

Figure 3: Spectrograph of CD-ROM

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 Eyring models) (Figure 4). It was determined that the Eyring model provided the best fit to the experimental data:

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

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

Applying the refined Eyring model enabled the prediction of end of life estimates for the CDs at different temperature and relative humidity conditions.

Testing the Model

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, new graphical results and further refinement of the model parameters, A and B.

Publishing the Results

Finally a paper was 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

4.2 The Scientific Discovery/Modelling Process

The example above illustrates the typical set of steps associated with the development of a scientific model. More generally, this involves the following sequence of events:

- Inception of the idea;
- Discovery, retrieval and analysis of prior, related work;
- Experimental design;
- Capturing the empirical and observational data;
- Analysing, processing, interpreting and annotating the data;
- Formulation of an hypothesis and/or the construction of conceptual and/or numerical models. The models may be analogous or predictive and often take the form of a mathematical relation;
- Verification and refinement of the model and/or hypothesis by capturing further experimental data and comparing it with data predicted using the model;
- Documenting and publishing the findings (with links to the data and model).

At each of these stages, there are different inputs, outputs, tools, assumptions, constraints, conditions and participants. Ideally the precise details are recorded by the associated metadata capture tools that are part of the established workflow. Significant progress has been made in developing scientific workflow systems that can capture precise provenance data (Altintas 2004). However, in addition to capturing all of the relevant information associated with each step, it is essential to also capture the relationships (or be able to infer relationships) between each of the components. In the next section we:

- describe the range of components that comprise a scientific model package;
- propose an approach to enable the relationships between components to be inferred;
- describe the process of constructing a Scientific Model Package (SMP) whose internal structure reflects the relationships between its components.

5. Defining and Constructing Scientific Models

5.1 The Components of a Scientific Model Package

Scientific models are complex, composite digital objects which encapsulate a variety of related heterogenous components. As illustrated in the example in Section 4.1, they may contain any of the following components or references to them:

- Pre-existing data, models, hypotheses or publications;
- Large data sets generated from experiments, observations and instruments. This 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, visualisation, 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 visualisation of the data;
- Workflows – steps involved in transforming the raw data into knowledge;
- Visualisations – 2D, 3D imagery, graphs, tables, charts, diagrams, animations;
- Textual components – EndNote files, notes, publications, reports, documentation, annotations, bibliographies, reviews etc.

5.2 Constructing and Publishing a Scientific Model Package

Within the FUSION project we are currently developing the infrastructure and tools that will enable a scientist to construct and then publish a new Scientific Model Package (SMP):

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 large scale datasets generated by scientists or instruments. 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 4.2. At the end of each step in the workflow, the software captures as much metadata as possible automatically. If necessary the software will request manual data input or selection from pull-down menus by the scientist. An RDF triple store saves the metadata associated with each event e.g., the agents (human or software), inputs, outputs, tools, instruments, settings, constraints, date, time, place etc.
3) At the end of the scientific discovery process, the scientist decides to publish a particular view of his/her model. In the majority of cases, the complete view of the scientific discovery process is excessive and overly complex. A simplified or coarser-grained view of the process is often required for teaching or dissemination purposes.
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. Tools are under development that enable the scientist to specify the precise components that may include:
    a. Data: database values, images, visualisations, 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 Model Package (SMP) 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 (defined using SWRL), or defined by the scientist during the SMP specification.

6) Descriptive metadata for the SMP is input and validated. The core metadata fields are:
    - Identifier
    - Title
    - Research focus
    - Model type (drawn from a hierarchical thesaurus)
    - Creator – name and contact details, organisation etc.
    - Date Created
    - Date Published
7) The creator/author attaches a ScienceCommons (ScienceCommons 2005) license (selected from a menu of license templates) to the SMP
8) The SMP object can then be ingested and saved to a DSpace (DSpace 2005) or Fedora (Fedora 2005) digital library/institutional repository.

Figure 4 illustrates the various different storage areas envisaged in an ideal scientist's environment, and the relationships between them.

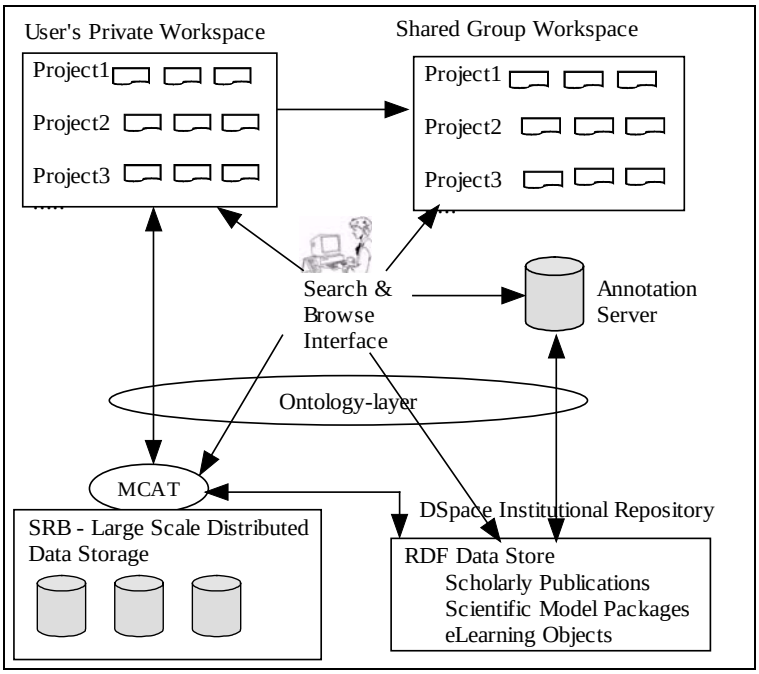

Figure 4: Envisaged Storage Area Requirements for eScience

5.3 The Ontological Foundation

Although the ABC model (Lagoze & Hunter 2001) was developed specifically for the library, museum and archival domains, we believe that it can be extended to capture the provenance or lineage of scientific models. It also provides an ideal top-level ontology for defining the classes and properties associated with scientific models and their components. Figure 5 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 subclasses are a set of properties specific to that subclass.

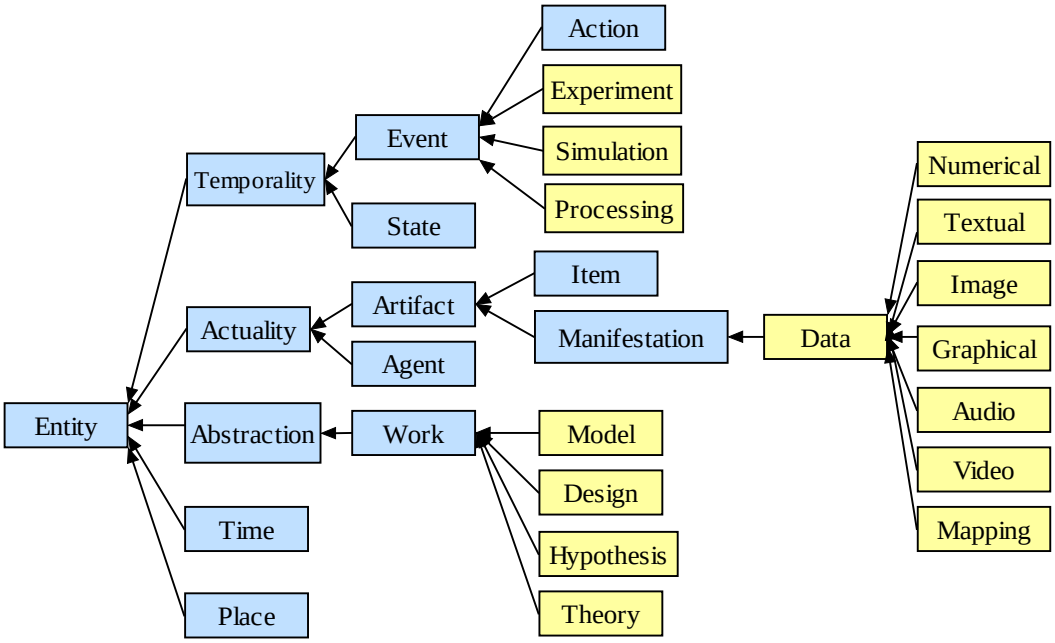

Figure 5: Extensions to the ABC Model Class Hierarchy for eScience Provenance

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 (Library of Congress 2005) and the MPEG-21 DIDL (Bekaert 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 et al 2004a). An RDF package corresponding to the example in Section 4.1, illustrating the semantic relationships between outputs of the scientific modelling process, is shown in Figure 6.

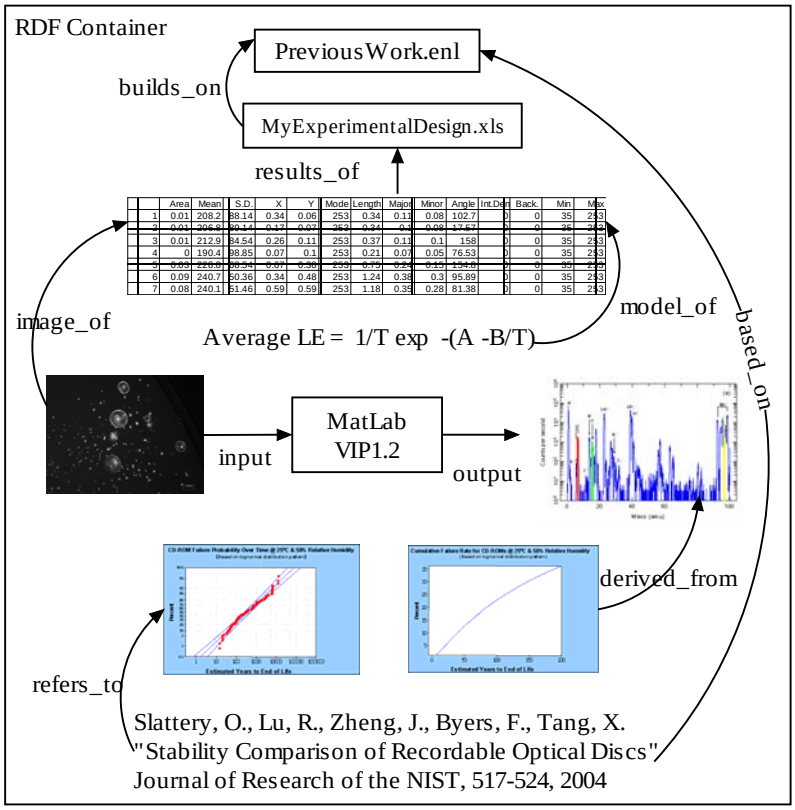

Figure 6: An example of an RDF Scientific Model Package

6. 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 particular the nano-materials and molecular biology domains) in order to thoroughly understand their modelling procedures and associated data management requirements and the commonalities across communities. In parallel with this analysis of user needs and processes, we are also working on:

- Development and evaluation of the extended ABC Ontology for eScience provenance logs;
- Implementation and evaluation of an eScience provenance logging system and database;
- Investigation of rules and methods (e.g., SWRL) for determining or inferring relationships between selected components of an SMP, from the provenance logs;
- Development of the SMP construction and description tools;
- Development of a search, browse and retrieval interface to a repository of SMPs.

In addition, we are continuing to track the outcomes of the Science Commons Initiative (ScienceCommons 2005). 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 to easily attach the emergent Science Commons licenses to SMPs and their components when they want to share them, without sacrificing intellectual property rights.

7. 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 of 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 Model 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 maximises its re-use, dissemination and potential socio-economic benefits.

References

Aalst, W., L. Aldred et al 2004, *Design and Implementation of the YAWL system* In Proceedings of the 16th International Conference on Advanced Information Systems Engineering (CAiSE 04), Riga, Latvia.

Altintas, I, C. Berkley et al, 2004, *Kepler: An Extensible System for Design and Execution of Scientific Workflows* 16th Intl. Conference on Scientific and Statistical Database management (SSDBM), Greece.

Andres, T.F. Cubera, et al, 2003, *Specification: Business Process Execution Language for Web Service Version 1.1*, BEA, IBM, Microsoft, SAP AG and Siebel System

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, vol. 9 no. 11. Available from: <http://www.dlib.org/dlib/november03/bekaert/11bekaert.html> [20 Dec 2005].

Bose, R. and Frew, J. 2005, *Lineage Retrieval for Scientific Data Processing.* ACM Computing Survey, Vol 37, No. 1, pp. 1-28. Available from: <http://homepages.inf.ed.ac.uk/rbose/pubs/bose_2005_ACM_CS.pdf > [20 Dec 2005].

Cavalcanti, M.C., Mattoso, M., Campos, M.L., Llirbat F., Simon, E. 2002, *Sharing scientific models in environmental applications***,** Proceedings of the 2002 ACM symposium on Applied computing, Madrid, pp. 453-457. Available from: <http://portal.acm.org/citation.cfm?id=508876> [20 Dec 2005].

Clarke, D.G., and Clarke D.M., 1995, 'Lineage' in *Elements of Spatial Data Quality*, S.C.Guptill and J.L.Morrison, Eds., Elsevier Science, Oxford, pp. 13-30.

Coleman, A., 2002, Scientific Models as Works. Cataloging & Classification Quarterly 33 (3/4), pp.129-159. Available from: <http://www.sir.arizona.edu/faculty/coleman/papers/smascrev.pdf> [20 Dec 2005].

Coleman, A. 2002, A Classification of Models. In Lopez-Huertas, Maria J. Ed. *Challenges in Knowledge Representation and Organization for the 21st century. Integration of Knowledge across Boundaries*. Proceedings of the Seventh International ISKO Conference, Granada, pp. 86-92. Available from: <http://www.sir.arizona.edu/faculty/coleman/papers/iskoasc.pdf> [20 Dec 2005].

De Roure, D., Jennings, N. R. and Shadbolt, N. R. 2005, The Semantic Grid: Past, Present and Future. *Procedings of the IEEE* 93(3):pp. 669-681. Available from: <http://ieeexplore.ieee.org/iel5/5/30407/01398019.pdf?arnumber=1398019> [20 Dec 2005].

Dspace 2005, *DSpace Federation*. Available from: <http://www.dspace.org/> [20 Dec 2005].

Fedora 2005. *Fedora*. Available from: <http://www.fedora.info/> [20 Dec 2005].

Hill, Linda, et al. 2001, *A Content Standard for Computational Models.* D-Lib Magazine, vol. 7, no. 6. Available from: <http://www.dlib.org/dlib/june01/hill/06hill.html> [20 Dec 2005].

Houghton J., Vickery G. 2005, *Digital Broadband Content: Scientific Publishing*, Working Party on the Information Economy, Committee for Information, Computer and Communications Policy, Directorate for Science, Technology and Industry. Available from: <http://www.oecd.org/dataoecd/42/12/35393145.pdf> [20 Dec 2005].

IBM 2002, *BPWS4J*. Available from: <http://www.alphaworks.ibm.com/tech/bpws4j >[20 Dec 2005].

Lagoze, C., Hunter, J. 2001, *The ABC Ontology and Model*. Journal of Digital Information, vol. 2 issue 2, Article No. 77. Available from: <http://jodi.ecs.soton.ac.uk/Articles/v02/i02/Lagoze/> [20 Dec 2005].

Library of Congress 2005, *METS Metadata Encoding and Transmission Standard*. Available from: <http://www.loc.gov/standards/mets/> [20 Dec 2005].

Nielsen P.F. and Halstead M. D. 2004, *The evolution of CellML*. in Proceedings of the 26th Annual International Conference of the IEEE EMBS San Francisco, CA. Available from: <http://ieeexplore.ieee.org/iel5/9639/30463/01404512.pdf?tp=&arnumber=1404512&isnumber=30463> [20 Dec 2005].

Raspl S. 2004, *An Overview of PMML Version 3.0*. KDD-2004 Workshop on Data Mining Standards, Services and Platforms (DM-SSP 04). Available from: <http://www.ncdm.uic.edu/workshops/dm-ssp04/pmml3_overview.pdf> [20 Dec 2005].

RuleML 2005, *The Rule Markup Initiative*. Available from: <http://www.ruleml.org/> [20 Dec 2005].

ScienceCommons 2005, *ScienceCommons Initiative*. Available from: <http://sciencecommons.org/> [20 Dec 2005].

Taverna 2005, *Taverna 1.3.1*. Available from: <http://taverna.sourceforge.net/> [20 Dec 2005].

W3C, 2001, *Web Services Description Language (WSDL)* W3C Note. Available from: <http://www.w3.org/TR/wsdl> [20 Dec 2005].

W3C 2004a, *RDF/XML Syntax Specification (Revised)*, W3C Recommendation 10 February 2004. Available from: <http://www.w3.org/TR/rdf-syntax-grammar/> [20 Dec 2005].

W3C 2004b, *OWL Web Ontology Language Overview*, W3C Recommendation 10 February, 2004. Available from: <http://www.w3.org/TR/owl-features/> [20 Dec 2005].

W3C 2004c, *SWRL – A Semantic Web Rule Language Combining OWL and RuleML*, W3C Member Submission, 21 May 2004. Available from: <http://www.w3.org/Submission/SWRL/> [20 Dec 2005].

Zhao, J., Wroe, C., Goble, C., Stevens, R., Quan, D., Greenwood, M., 2004a, *Using Semantic Web Technologies for Representing e-Science Provenance* in Proc 3rd International Semantic Web Conference ISWC2004, Hiroshima, Japan, Springer LNCS 3298

Zhao, J., Goble, C., Stevens, R., Bechhofer, S. 2004b, *Semantically Linking and Browsing Provenance Logs for E-science*, Proceedings of ICSNW 2004. pp. 158-176