Interoperability Issues in Scientific Data Management

Version 1.0

Simon Cox
University of Southampton

Rachel Jones
Instrata Ltd

Bryan Lawrence
British Atmospheric Data Centre

Natasa Milic-Frayling
MSR Cambridge

Luc Moreau
University of Southampton

March 10, 2006

Contents

This paper is a result of a focused discussion and brainstorming that took place at the *Scientific Data Chain Workshop* on January 26, 2006 in Windsor, UK.

The workshop was sponsored by Tony Hey, Senior Vice President of Technical Computing Initiative (TCI), at the Microsoft Corporation. It is the first in the series of initiatives that TCI will carry out to increase a collective awareness of challenges and requirements for interoperability across scientific data efforts.

This document is intended for a community wide discussion about the key drivers, requirements, and issues in the data lifecycle from data collection, curation, storage, and publishing to maintenance and preservation. It is produced for circulation among the representative stakeholders to receive feedback and expand the analysis with proposals for the community wide initiatives.

Authors

Simon J Cox
Southampton Regional e-Science Centre
School of Engineering Sciences, University of Southampton
Southampton, SO17 1BJ, United Kingdom
Email: sjc@soton.ac.uk
Tel: +44 (0)2380 593 116

Rachel Jones
Instrata Limited
12 Warkworth Street
Cambridge, CB1 1EG, United Kingdom
Email: racheljones@instrata.co.uk
Tel: +44 (0)1223 301 101

Bryan Lawrence
British Atmospheric Data Centre
Space Science and Technology Department
CCLRC Rutherford Appleton Laboratory Fermi Avenue
Chilton, Oxfordshire, OX11 0QX, United Kingdom
E-mail: B.N.Lawrence@rl.ac.uk
Tel: +44 (0)1235 445 012

Natasa Milic-Frayling
Microsoft Research
Roger Needham Building, 7 J J Thomson Avenue
Cambridge, CB3 0FB, United Kingdom
E-mail: natasamf@microsoft.com
Tel: +44 (0)1223 479 772

Luc Moreau
School of Electronics & Computer Science
University of Southampton
Southampton, SO17 1BJ, United Kingdom
Email: L.Moreau@ecs.soton.ac.uk
Tel: +44 (0)2380 594 487

Meeting facilitator:

Steve Bather
Specialists in Leadership and Collaboration
Realise (Europe) Limited, www.realisegroup.com
Email: stevebather@realise.ws
Tel: +44(0)1903 778 545
Mob: +44(0)7885 175396

Executive Summary

Scientific data management involves complex and multifunctional systems that support the full lifecycle of data from the production, ingestion, storage, and curation to the publishing and consumption of data. These distinct phases shape the life of a digital object within the system. Moreover, the management and interaction with a data object triggers the creation and lifecycles of new data objects.

Indeed, the design of a reliable and effective data management system requires a ‘self-documenting’ function which gives rise to digital records about the processes applied to data objects. Such are *process documentation* records and *ancillary data* records created by curators. They are digital objects themselves, subject to the same data processes. The system thus perpetually re-enters a succession of steps from data creation to data consumption, here referred to as the *data loop*. Consequently, data management systems are technically intricate and increase in size and complexity with their use.

In order to implement and run effective systems, the organizations and communities that are engaged in data management require appropriate tools, training, and technical support. Interoperability of tools has been flagged as one of the important issues that the community is facing. However, our analysis reveals a number of other related issues that require full consideration.

For example, these communities have not sufficiently defined the roles and skill sets that are needed to perform data management tasks. In fact, there seems to be a lack of community wide recognition and appreciation for the scientific data curation function. As a result, there is no established professional path for individuals who wish to engage in data curation and the teams involved in data preservation have varied skill sets.

Furthermore, the data management and preservation efforts cost while, at the moment, there are no clear strategies for moving them from purely altruistic or government funded efforts to self sustainable operations.

Finally, with the wider availability of effective data management systems and services that potentially include sensitive information, it is important to establish an appropriate legal framework to protect both the organizations that store, preserve, and enable access to the data and the consumers of data services. In order to establish and ensure compliance with legal requirements we need to provide adequate support, tools and best practices for use by the service providers and consumers.

Having these broad issues in mind, we suggest focusing on several problem areas and make some concrete recommendations for the community wide initiatives. However, we particularly stress the importance of establishing a framework within which scientists from different disciplines can effectively communicate, exchange experience and expertise, and create a global awareness about the importance of scientific data management. We expect that everything else will follow from the solid community base, from the development of technologies and best practices to the self-sustainable models for preservation initiatives. We also wish to point out two important aspects that affect the current situation and are important for defining further directions.

First, ‘change’ is the major trait of both the data and the communities dealing with the data. Each data management system is a self-perpetuating cycle of data transformations and processes, affected by changes in technologies and practices. Thus, it requires adequate migration and adaptation strategies to maintain its relevance. Similarly, the teams and communities change and re-organize over time. Any successful approach in addressing current issues will therefore have to keep a long perspective in mind and propose solutions that can evolve and adapt.

Second, the scientific community is diverse while the collaboration is based on common interest and background. We expect that a successful framework for supporting the scientific community will take into account these two important social aspects.

RECOMMENDATIONS

In our analysis of current efforts and practices, we identified several types of initiatives that the scientific community could engage in to ensure further progress in the scientific data management. We group them into three main areas.

FRAMEWORK FOR CONNECTING THE COMMUNITY

- Define the strategy and the framework for unifying the scientific community across disciplines, organizations, and interest groups.

  The framework should exploit the strengths of economic, social, and scientific interests and build a sustainable model for collaboration. Establishing such a framework is crucial for making any significant progress on addressing the scientific data management and related issues.

PROTOCOLS, INTERFACES, AND BEST PRACTICES

- Define open protocols, interfaces, and data models that allow the community of users to develop interoperable tools and services to support all the steps of the data lifecycle.
- Define reference implementations for the open protocols and interfaces, which are configurable, extensible, and deployable in varied environments that end users make use of on a day to day basis. This, in particular, requires effective binding to scripting/programming languages.

TOOLS

- Provide a toolkit stack, i.e., development layers for handling software and hardware that can be easily brought into the user environment.

  This applies to a wide range of users, from administrators to toolkit developers and data consumers. Most of the people involved in the curation are specialists in the data area rather than technology.
- Enable a community of developers to create a corpus of reusable components.

  Provide a framework for developers to connect, share and develop reusable software components. That may include a range of initiatives from informal developers' gatherings and forums to joint investments by interest groups into selected problem areas.

  Develop tools and practices for curation of protocols, interfaces, data models, libraries, and tools. Carefully design the metadata, i.e., descriptions that make them easy to discover, share, deploy, and reuse.
- Invest in tools to help generate, maintain, and index information about the data and processes.

  Process documentation and ancillary information are the basis for further consumption of data and ultimately determine the impact that a data management system has. The community requires tools to help create, define, and process domain specific metadata. We need to enable an appropriate level of automation in order to capturing and manage process documentation and ancillary information.

POTENTIAL FOR A LEADERSHIP ROLE

Problems faced by the scientific community could be addressed through a strong leadership from the industry and funding agencies that combine several strategies to establish a self sustainable and stable ecosystem. These organizations may:

- Fund initiatives in research and technology development to eliminate barriers, such as high initial costs, and help increase the overall standard of data handling

- Actively participate and contribute to the community effort through sharing of the technology and expertise, thus establishing trust and providing the leadership
- Build or enable the building of platforms and components to be licensed to the community and facilitate creation of tools and management of data and processes.
- Build complete applications for general and specific use in specific domains, based on market demand and business opportunities.

Introduction

Over the next five years, science and engineering projects alone will produce more scientific data than has been created over the whole of the human history. The success of many of these projects will depend critically on the ability to collate, distribute, process, and analyze large and diverse data sets.

Data is being generated and collected from various sources, including scientific computations, physical experiments, and sensor networks. Because of the collaborative nature of scientific initiatives, this data often needs to be shared across teams and organizations. Providing reliable access and also ensuring that data is preserved and usable by the community is one of the important factors for further scientific progress. However, that presents challenges from both the technical and the economic point of view. The inherent problem of storing and handling large data sets and the associated costs equally affects both individual researchers and well-funded teams in the scientific community.

Whilst some solutions to data handling exist, they are often specific to a particular data set, community, or domain. Coverage of the problems faced and solved also varies significantly across the scientific community. In some domains, tools exist and are often reinvented while in others the data processing and management issues have barely been tackled. The real and perceived cost of migration to proper data handling and providing adequate tools is commensurately high, despite the benefits that users recognize would be achieved. Furthermore, while having a comprehensive and broadly usable toolkit seems highly desirable, there has been little motivation, enthusiasm, and mechanism for the various stakeholders to embrace such a challenge together.

In this paper, we discuss selected data management scenarios and analyze them from several perspectives in order to identify common issues, challenges, and barriers for improvement. We present our summary analysis and key recommendations in the Executive Summary. Subsequent sections provide more detailed background, discussion and suggestions for how to achieve a fluid and efficient way of collecting, curating, storing, publishing, and preserving scientific data.

Overview of the Data Lifecycle

In this section, we analyze a data lifecycle, from creation to consumption, including curation in a repository, storage, and publishing for the reuse of data. We abstract and outline a common underlying model. In the subsequent section we describe several scenarios in more detail, showing how they map onto our abstract model.

Concepts and Data Lifecycle Model

For the sake of clarity and scope we here restrict our discussion to data in digital form. Once created, data is ingested into *digital data repositories* from where it is extracted before use. Data use can result in new data objects which themselves are ingested into the system and then moved through the similar loop from the ingestion, to storage, curation, publishing and usage. Thus, the *data transformation* is one of the key underlying concepts of the data lifecycle. The succession of steps involved in the data lifecycle is referred to as the *data loop*.

DATA LOOP

The loop of data production, storage, curation, publishing, and consumption is continuously triggered throughout the life of a digital object. Indeed, the data lifecycle begins with the data production. However, even the very first transformation of data connects back into the data store and forms a “data loop” since new data items are generated from usage and they need to be preserved and curated. For example, new information created by both curators and service consumers can itself be ingested as additional information, i.e., *annotations* on the existing data objects.

These fluid transitions and emerging cycles in the digital object life are reflected in the nature of the associated roles and processes. In some instances, a single person or a system may assume multiple roles, from a producer to a consumer, while in others multiple people or systems may focus on any particular role, e.g., ingestion of the data objects.

DIGITAL DATA REPOSITORY

Digital data repositories range from simple storage, involving nothing more than storage devices, to complex managed archives that conform to international standards (such as OAIS) for reliability and service provision. Preserving digital content involves a wide range of issues which go beyond the mechanics of preserving information bits and bytes, as discussed below.

METADATA, PROCESS DOCUMENTATION, AND ANCILLARY DATA

The key to reusability of data is to provide information that describes, in sufficient detail, all the processes and data characteristics that are relevant for further data usage. Different communities adopted different terminology to refer to such information.

It is typical to use the term *metadata* to refer to any information about a given piece of data. However, a community such as the Semantic Web reserves this term specifically for commonly accepted domain specific vocabulary or officially standardized tags (Dublin Core) which, once applied, facilitate document retrieval and classification. On the other hand, a generic use of the term metadata does not provide differentiation that is of practical use since any piece of data may be considered metadata relative to some other data.

For that reason we here explicitly recognize two types of information: the automatically generated information about processes, referred to as *process documentation*, and human generated or facilitated information which is a byproduct of data management through the lifecycle, referred to as *ancillary information*. Both process documentation and ancillary information are data objects that require storage, preservation, and access, and incur new data cycles when created.

We recognize that this may not be an ideal use of terminology in general but we adopt it for the purpose of this document.

Phases in the Data Lifecycle

Figures 1 and 2 provide a schematic view of the data loop and roles and tasks involved in data handling throughout the data lifecycle.

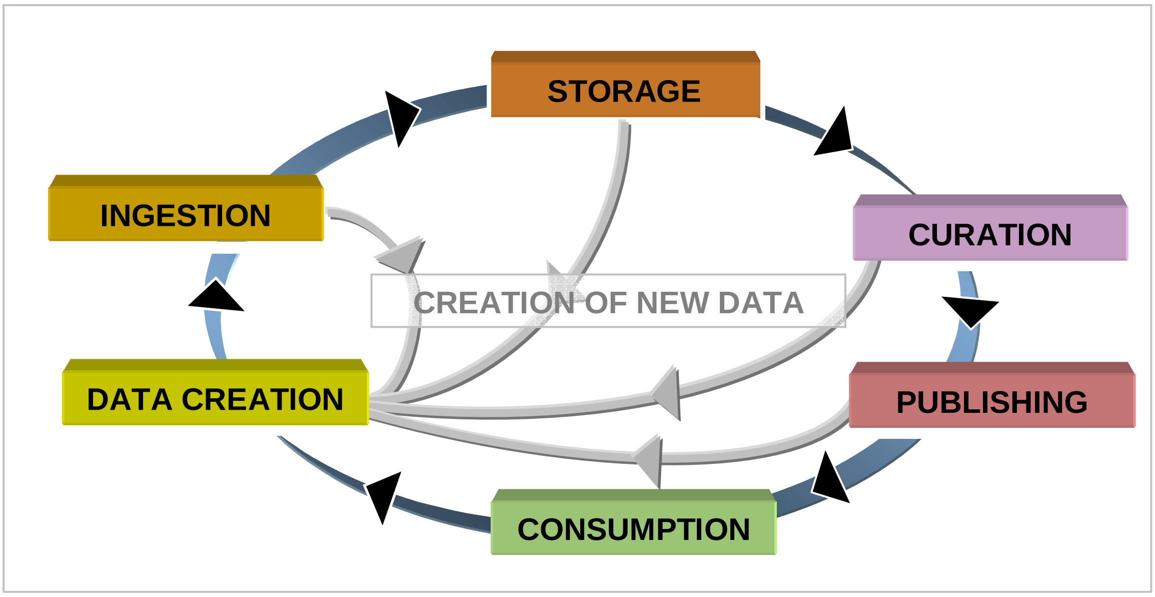

FIGURE 1: Data loop, from data creation to data consumption is continuously instantiated from individual processes in the data management cycle.

DATA CREATION

Any object entering the digital world has properties which are primarily determined by two factors: what has been captured about the object and what initial assumptions are about its storage requirements and future use.

We refer to the person or the process that creates the data artifact as the *producer.* Ideally, when a digital object is created, the producer records information about both the data and the process that created the data. Indeed, the producer's function generally subsumes multiple roles including generation of the data itself and the additional information which captures properties of the process that created the data and the properties of the data itself.

Once the digital object exists, any of its transformations or modifications need to be recorded in the form of *ancillary information,* facilitated by the producer, and the *process documentation* which captures the structure and properties of the of the process executions. From that information, the curator or producer can describe what has been done, how, for what reason, and, if appropriate, by whom.

DATA INGESTION

The digital object is further shaped by the *ingestion process* that moves it into a storage system. The ingestion process covers several activities: acquisition of the digital object, validation against storage requirements, and possibly some data transformation such as compression (see Figure 2). Each of these steps may result in additional property information that ought to be recorded in order to provide information about the ingestion.

DATA STORAGE

The result of the ingestion process is a *data object* that conforms to the requirements of the storage environment such as the specifications of the storage media and storage schema. These requirements vary in complexity. In some instances the data object simply has to be in one of the

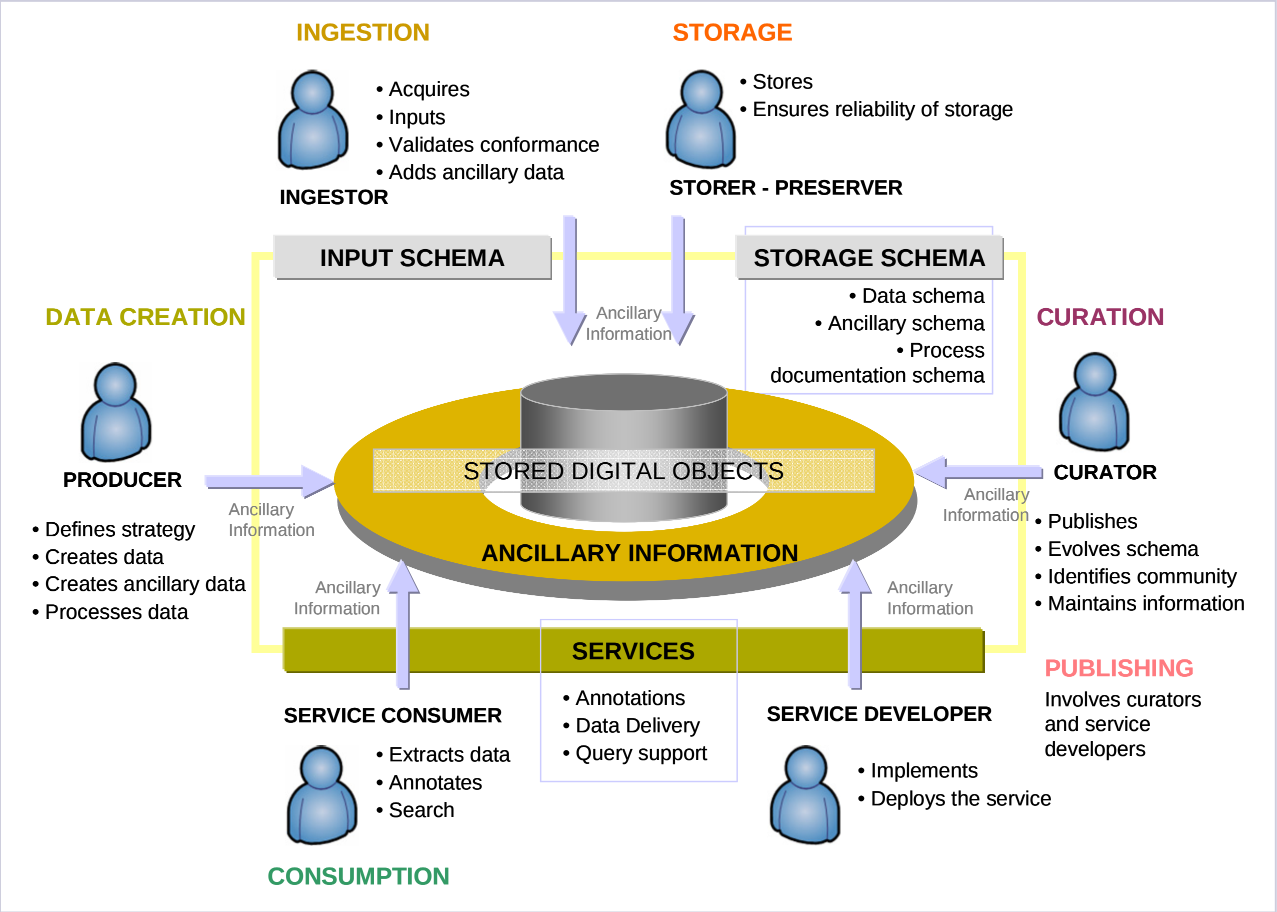

FIGURE 2: Roles and tasks in the data lifecycle. We note that at each stage ancillary data is created and added to the repository of stored digital objects.

supported file formats and its content described in a text file. In others, the data object may have to be stored in a particular binary format and its content description entered using a pre-defined relational database schema. Consequently, the digital *data repositories* range from a simple storage, involving nothing more than storage devices, to complex managed archives that conform to international standards (such as OAIS) for reliability and service provision.

The process of data storing is concerned with the digital object preservation. Preservation involves maintaining confidence that the digital object has not been inadvertently changed, that the storage is reliable, and that the object is migrated to a new storage as necessary.

For the sake of concreteness we shall here primarily discuss repositories that are aimed at preserving content and information about digital data. While this narrows the scope, it still involves a wide range of issues which go beyond the mechanics of preserving information bits and bytes, as discussed below.

DATA CURATION

The storage and preservation need to be distinguished from the digital object *curation*. Curation comprises:

1. Identifying who can access objects and how, i.e., what services are available for operating on the digital object;
2. Cataloguing and/or indexing the object;
3. Evolving the storage schema and expanding information about the data objects to meet the community needs;

4. Providing notification on the data object modifications, i.e., updates and replacements, to the data users who required such a service (see Figure 2).

In some instances, the curator needs to migrate the format of the data object. Thus, the curator assumes the role of a service consumer through accessing and extracting the data, and then acts as a data ingestor, transforming the data object into a new format, re-ingesting it into the storage system, and if necessary creating ancillary data.

DATA PUBLISHING AND CONSUMPTION

Digital objects are made available for consumption via *services* that are controlled by the curator, implemented by the system designers and developers, and consumed by end users (see Figure 2). Such services range from simple publishing of information in support of data discovery, to more sophisticated data search, analysis, visualization, and delivery. The types and the range of possible services typically depend on the resources and tools available to the system developer, the ownership and legal terms of the data access, and the information about the data and its characteristics that is stored in the system.

Connecting a stored digital object to services is the act of publishing and is undertaken by the curator. The curator and the system developer use the available information about the data to provide services. The quality of the content and information about the data object thus determine the scope of the services.

Scenarios

BIO-MOLECULAR SIMULATIONS

BioSimGrid (www.biosimgrid.org) is developing a global repository for bio-molecular simulation data. Data is stored using a pre-specified schema which is continuously maintained. The data repository consists of files with the simulation data and the metadata describing key simulation parameters, provenance information, and summary information computed from the data files.

In the computational bio-molecular research, large amounts of simulation data are generated to capture the motion of proteins. These massive simulation data sets can be analysed in a number of ways in order to identify biochemical properties of proteins. However, the common way of storing these data, typically in the laboratory where the simulations have been run, often hinders data sharing and cross-comparison of simulation results. The data is usually encoded in the format specific to the simulation package that produced the data and, thus, it can only be analysed with tools developed specifically for that simulation package.

The BioSimGrid platform aims to provide a solution to these challenges by exploiting the potential of the Grid to facilitate data sharing. By using BioSimGrid, either in a scripting or the Web environment, users can deposit their data and reuse it for analysis. BioSimGrid tools manage the multi-location storage transparently to the users and provide a set of retrieval and analysis functions for convenient and efficient processing of data. The users of BioSimGrid can store their simulation data, using one of the multiple supported formats, together with the associated metadata. The data

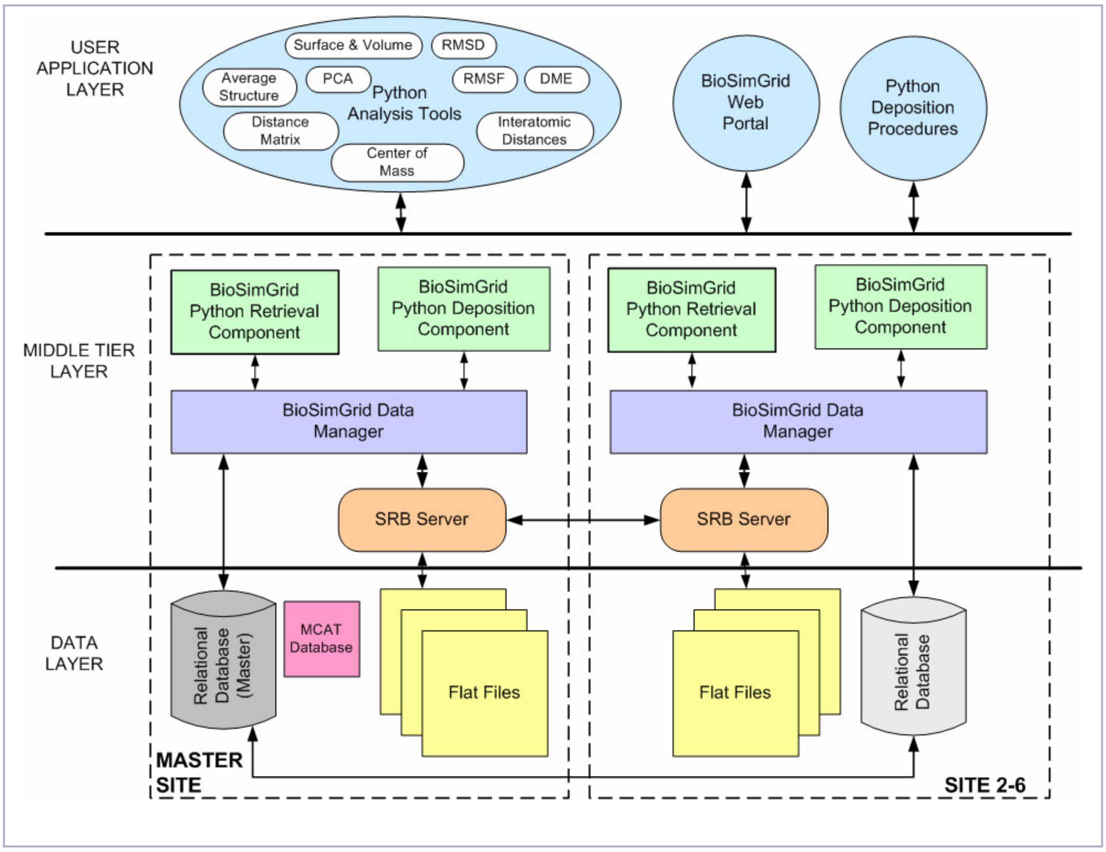

FIGURE 3: Architecture diagram for the BioSimGrid project.

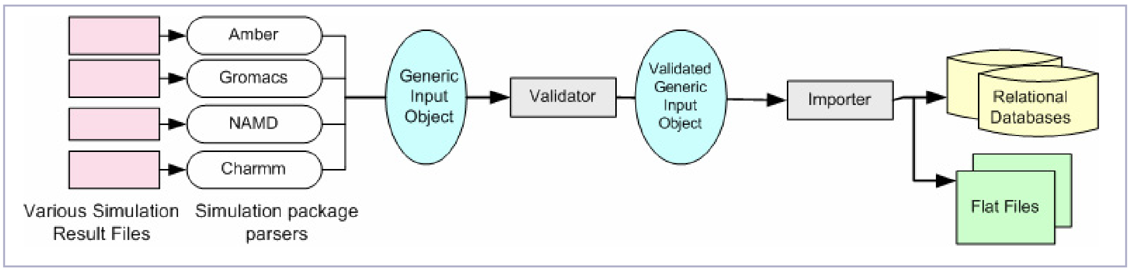

FIGURE 4: Deposition of objects into the BioSimGrid system.

can then be made available to the whole community (see Figure 3 for details of the BioSimGrid architecture and Figure 4 for details of the data deposition process).

The BioSimGrid data retrieval component enables the user to retrieve data transparently, without knowing about the underlying database mechanisms. The flexibility of retrieval tools enables the users to access specific files or groups of files or different slices of a bio-simulation trajectory. The BioSimGrid also provides a set of custom-built analysis tools which can be used to study the functional dynamics of a simulation, e.g., the root mean square deviations, volume and average structure, inter-atomic distance, and surface area. Alternatively, users can create their own analysis tools from the retrieval components and access particular slices of a bio-simulation trajectory that are required by their analysis. They may also invoke post-processing tools provided as a service within the database, to generate various file views (data, picture, and video). The returned files may be used to seed new calculations from which new data is generated and submitted to the repository. In such instances it is of utmost importance to keep a good provenance record.

The deposition, retrieval, and analysis components are implemented in Python, enabling the users to use BioSimGrid in a Python scripting environment. Alternatively, they can use a Web based interface with limited analysis capabilities and without support for data deposition. The rationale behind choosing Python is pragmatic since several analysis dependent post-processing tools, such as PyMOL (http://www.pymol.org, 2002) and the Molecular Modeling Toolkit "MMTK" (http://starhip.python.net/crew/hinsen/) were written in Python and the simulation community is moving towards Python as the preferred environment for post-simulation analysis.

In summary, BioSimGrid seeks to

- Provide a transparency of data location to the users, where the knowledge of the physical location of the data is not essential to the process of data retrieval,
- Maximise the data transfer rate in terms of the speed of delivering data to the computation element, e.g., the analysis toolkit,
- Provide an abstraction of the data layer so that the scientists can focus on their scientific research and avoid dealing with the complexities of the data querying languages and the data storage structure,
- Provide a general purpose analysis toolkit for operating on the BioSimGrid data storage structure.

On deposition, various forms of kite marking occur. The lowest level is a simple deposition with the data validation and verification of the conformance with the given schema. The highest level of kite marking involves additional data post-processing, for example to generate summary information and validate links to publications, as referenced in the data.

The system allows the users to specify various levels of access control for each file and its corresponding metadata items. That enables the system to scale from a personal private repository to a public record that includes, for example, precise details of processes and data files relating to a peer-reviewed publication.

ENGINEERING DESIGN

Engineering Design Search and Optimisation (EDSO) aims to achieve improved designs by exploiting engineering modelling and analysis. Variables in a design are systematically modified to increase or reduce a quality measure called *the objective function*, whilst ensuring that the variables satisfy certain constraints. It often involves computationally intensive processes, producing large amounts of data in a variety of formats throughout the workflows that the engineers are executing. This often happens in a distributed manner, thus processing and producing data at different locations.

The emergence of Grid computing and Web service technologies provides new opportunities for the engineering community to access an extended range of computing resources and manage more effectively the sizable data created by distributed applications. The Geodise Database Toolbox (www.geodise.org) uses Web services, XML, databases, and Grid technologies. It has been developed to support management of data created locally or on the Grid by engineering applications, and to bring these technologies into an environment familiar to engineers. It has been integrated into the Matlab and Jython scripting environments for the ease of use. It interacts with other applications via its Java API.

The Toolbox supports centralised and personal data repositories. The former are accessed via secure Web services from platform independent client applications. Metadata about files, data structures, collections of related data, and workflows can be easily defined. The Toolbox' distinctive feature is the support for user-defined application specific metadata that can be queried to locate required data efficiently. It is used in areas such as Computational Fluid Dynamics (CFD), Structural Mechanics, and Computational Electromagnetic (CEM) engineering problems, and it has been deployed in other domains such as environmental science (see below) and integrative biology. The Toolkit has been further extended for a more convenient use in monitoring application processes, enabling the engineers to intervene by halting or changing long running optimisations if necessary.

Traditionally, data created using engineering applications is stored in files on file systems, with little information to describe them. When the data volume is large, this makes it difficult to search, share and reuse the data. The limitation of this approach becomes more obvious when a group of people affiliated with different institutions, i.e., forming a Virtual Organisation (VO), wish to collaborate to solve a common problem by making use of Grid technology. These issues can be overcome by attaching additional descriptive information (metadata) to the data, so that it can be located by querying its characteristics without having to know its storage location.

In order to encourage the use of metadata within the engineering environment it must be straightforward to specify and include any terms and nested data structures that describe the data. The Storage Resource Broker (SRB) (http://www.sdsc.edu/srb/) provides a uniform interface for connecting to heterogeneous resources over a network and, with the Metadata Catalog (MCAT), provides dataset access based on characteristics rather than names or physical locations. However, MCAT has less support for application specific metadata, particularly for complex, nested data structures and data types that are often essential to locate the problem specific data.

The Geodise Database Toolbox enables engineers to use the Matlab and Jython environment to define metadata conveniently as Matlab structures. Standard metadata (e.g., archive date, file size) are generated automatically by the Toolbox so that users only need to concentrate on defining custom metadata, specific to their applications, and granting access permissions to other users in the VO if they want to share the data.

Metadata is stored in an XML enabled relational database (Oracle 9i) using tables for standard metadata to ensure efficient access and native XML for user defined metadata to achieve required flexibility. These representations can be transparently converted to and from user defined Matlab structures using the XML Toolbox for Matlab (www.geodise.org).

The database is queried using a simple syntax to locate files, variables, or groups of data based on their characteristics. Users specify queries as a combination of named metadata variables and comparison operators, with options for exact or wildcard matches and case insensitivity. A user query is converted to a combination of SQL and XPath and restricted to return results that the user is entitled to access. The returned array of structures may contain the complete metadata for each result or just a specified subset. For example, the user may only want to see unique data identifiers which can then be used to retrieve files from the archive, regardless of where they are stored. Users can incorporate the query function into their Matlab scripts directly or interact with a query GUI which supports hyperlinks for downloading and browsing related data.

A concept of *the datagroup* is used to create logical groups of related data, such as that used in a process monitoring task. The metadata can be added at the group level so that the entire collection is described. A datagroup may contain sub-datagroups. Furthermore, datagroups may share data. This gives users the ability to describe and exploit relationships and hierarchies.

When optimizing devices, the optimization process involves a schema which allows querying of the archive of previous design studies. That enables exploiting information about the provenance of the data, the workflows, and the patterns and practices that have been employed. Performing this analysis prior to commencing a new design exercise can provide new insights and inform further steps.

ENVIRONMENTAL SCIENCE

Here we present two classes of environmental science scenarios. In the first instance we consider observations of the real atmosphere and ocean system and in the second, we consider simulation of the atmosphere and ocean system. These examples are chosen from the NERC DataGrid (NDG, http://ndg.nerc.ac.uk/) and the Grid ENabled Integrated Earth system model (GENIE www.genie.ac.uk).

In the case of observations, the observer decides what to observe and develops an observation strategy. This action defines the input schema. In the case of simulations, the input schema is defined based on the set of simulation variables that should be exposed in the output and the time and space resolutions that are required.

As time evolves, both observational and simulation strategies change and so does the input schema. The act of ingestion involves storing numerical objects that result from observations or simulations into repositories and it needs to take into account the changes in the input schema. In the case of the NDG project this may result, for example, in advising the curator to modify the storage schema and take advantage of increasing resolution of the simulations as the computer processing power and speed increase. In GENIE, the data is deposited into the storage environment directly from the producers' run time environment (Matlab/Python) and thus the ingestion role is fully automated.

In general, the metadata accompanying the observations and simulations also changes with time and is often inadequate. On occasion a repository owner may reject the data as lacking sufficient metadata for reliable information preservation over a long term. Even where the metadata is adequate for ingestion, the ingestor may find it necessary to add further information. For example, additional metadata may be required to distinguish between two simulations or describe more precisely the tool that was used to carry out an observation.

Once a storable digital object exists, the storer/preserver assigns a unique identifier and places it in the repository for preservation over time. Preservation involves establishing a backup strategy and migrating from one storage system to another as required. The NDG project aims at preserving the data for posterity and, thus, it is concerned with both the migration and the backup strategies. In the case of the GENIE project, it is recognised that simulations are relatively easy to reproduce and the lifetime of the data is limited. Therefore, the backup strategy is "not to have one" and, consequently, the migration plans are postponed.

In the NDG case, the curators are the professional data management staff at the British Atmospheric and Oceanographic Data Centres. The repositories are devised to be OAIS compliant. Furthermore, a considerable effort is put into ensuring that the requirements of the community are well understood and that the storage schemas continually evolve. Datasets are published via

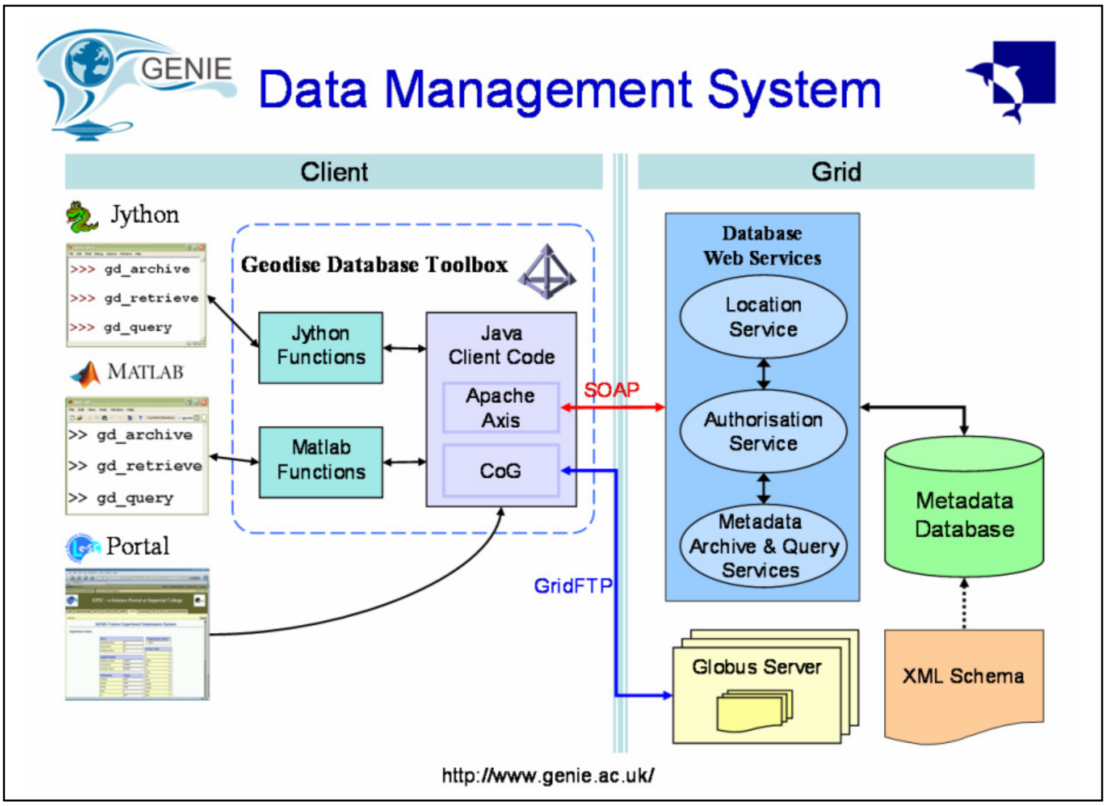

FIGURE 5: Architecture diagram of the GENIE project

catalogues and a variety of services are built and deployed by both the data centre staff and the research community.

The GENIE project is creating a Grid enabled component framework for the composition, execution, and management of the Earth system models. The GENIE codebase consists of mature models of Earth system components (i.e., ocean, atmosphere, land surface, sea-ice, ice-sheets, biogeochemistry, etc.) which can be flexibly coupled and run over multi-millennial timescales, primarily for glacial-interglacial simulations. An important part of such simulations is the parameterisation of many physical processes of the Earth System that occur on relatively small timescales. In order to make meaningful predictions it is vital that these parameters are tuned to appropriate values and that the effects of uncertainties in these parameters are quantified.

An augmented version of the Geodise Database Toolkit (www.geodise.org) is used to provide a generic data management solution for the GENIE project. The Geodise system exploits database technology to enable metadata to be associated with any file submitted to the repository for archiving. The database interface is exposed as Web services and files are archived in the system through a two step process: a) the file is transferred to a user specified file server using the GridFTP protocol, and b) file information is recorded in the database, including its location, its unique system generated identifier, access rights, and any user defined metadata.

Client tools are provided in Matlab and Jython to allow the user to upload, query, and retrieve data from the repository. An XML Toolbox (from Geodise) is used to convert Matlab data structures into XML for communication with the Web service interface to the database. Access to the system is controlled by a user authentication through an X.509 certificate. The system therefore provides an open and transparent facility through which members of the project's VO can share data.

The Geodise system has been designed to provide a flexible management solution for the engineering design process and to handle user-defined metadata that engineers may wish to record. However, the data generated by the GENIE framework is produced by well-defined component codes and, thus, the metadata is more tightly constrained. Therefore, we can significantly improve the efficiency of the system by mapping an XML schema to the underlying Oracle9i database: the metadata is handled by a relational database while the flexibility of the system is ensured by the use of XML schema.

The data management system thus provides a resource for storing metadata, files, and data structures. It is flexible enough to support the evolving needs of the framework while maintaining the efficiency of the database. In both cases, the GENIE and the NDG project, Web services provide interfaces to the data that are consumed widely by the scientific community. They enable data visualisation as well as retrieval of data for external post-processing and analysis. These activities may, in turn, result in further file and metadata deposition into the system.

HEALTH CARE MANAGEMENT

Organ Transplant Management (OTM)
http://twiki.gridprovenance.org/bin/view/Provenance/OrganTransplantManagement

Treatment of patients through the transplantation of organs or tissue is one of the most complex medical processes currently carried out. This complexity arises not only from the difficulty of the surgery itself but also from a wide range of associated processes, rules, and decisions that accompany any such surgery.

Depending on the country where a particular transplant is being carried out, the procedures and the level of electronic automation of information and decision making may vary significantly. Information systems designed to support such medical processes strive to make it possible to:

- Share information immediately between all the actors in a transplant case, from donors to recipients, decision makers, medical teams, and families.
- Provide decision support through sophisticated case data, protocol enforcement, and highlighting of possible matches.

Each of these objectives has a potential to significantly improve the decision making, speed up the reaction times and, ultimately, improve the outcome of the patient care. However, as with any medical information system, a great care must be taken to ensure that the system procedures, protocols, and data storage follow strict guidelines laid out by the law and medical practice.

These efforts involve distributed applications that are run by multiple institutions, each maintaining their own institutional medical data repositories and ingesting data that originates from multiple sources and in various formats. Such are patient records, laboratory analyses, etc. The flow of information in such systems is well understood and defined by workflows that are activated by external events such as availability of a donor, of an organ, etc. The course of action followed within such workflows is dictated by complex regulations that impose constraints on the exposure of the patient data to ensure confidentiality.

In order to facilitate auditing, every step in the execution of the workflow is documented by a well-structured process documentation. Such process documentation is also archived but typically kept in a separate repository with its specific access control policy. Regulations dictate the nature of the documentation, the kind of patient data it can or cannot contain, and the users who are entitled to access it. Auditing takes place by issuing queries to the repository that contains process documentation and to the repositories of medical data, provided that the access is granted. Such comprehensive queries over the patient's medical history may ultimately lead to insights into the patient's current conditions.

PERSONAL DATA SCENARIO

While the above scenarios outline the main functions and issues in data management by the scientific community, we claim that the same concepts and practices apply to each individual among us who needs to manage data over a period of time. For example, most computer literate users are now amassing a corpus of personal digital data that range from digital photographs and music to electronic versions of important documents, such as personal financial records, etc. A very few of us have engaged in the curation of personal data beyond possibly ensuring a backup to a physical media or onto a remote server and, perhaps, setting the access control on individual files and folders. Yet, we shall all soon face the issues of format migration and storage media migration as the technology evolves and new applications become available for content generation and information management.

We expect that with an increased awareness of the importance of data management we shall have on our disposal tools and services to help us with that process. These would be applications to help us maintain data through the digital lifecycle, from annotation of existing digital objects to creating information for provenance analysis and migration of important digital objects into a new generation of archival solutions, as required. We anticipate that the standard schooling will provide basic and refresher courses on this important aspect of our future life.

Issues, Concerns, and Barriers

Examples in the previous section illustrate the efforts that have been undertaken by the scientific community to manage and preserve data of great importance for our existence and our posterity. Here we discuss the main issues and concerns, and the barriers that stay in the way of further progress. Devising successful strategies for removing these barriers requires a good understanding of the complexities that result from intricate interaction of social, legal, economic, and technological issues. Here we provide a brief account of main issues from these four perspectives.

As a general remark, we note that many of the roles described in the digital object lifecycle are non-existent or poorly executed in practice, including the situations where professionals with relevant experience and knowledge are involved. Similarly, while the outlined repeated loops in the digital object lifecycle require that all actions which modify a digital object are documented through ancillary information or process documentation, only a few existing digital pipelines respect that requirement. This leads to problems in the exploitation of digital objects, sometime so serious that the data is effectively lost; in effect, although the content is preserved, its meaning is not captured.

Social Issues

In this section we describe social issues that have been identified either in the academic literature or from our own experience and that need to be considered when developing an infrastructure to support e-Science.

In the Computer Supported Cooperative Work (CSCW) and Human Computer Interaction (HCI) literature, we only found one reported study on the impact of technology on collaboration in scientific communities (Star & Ruhleder, 1994). Two other CSCW/ HCI papers that are particularly relevant are Dourish et al. (1999) who discuss the area of customisation and mutual intelligibility in shared category management, and Bannon and Bodker (1997) who discuss constructing common information spaces. In the following, we collate and summarise the issues we feel need particular consideration.

COLLABORATION

- Barriers to usage arise as a result of defining information to fit an individual's need and by the tools and data sources with which the individual is familiar with.
- Multiple meanings and interpretations can occur at all levels.
- Resolving multiple meanings requires an effort on the part of the contributor and the consumer of the data. It results in the creation of what are called "boundary objects" in the field of CSCW.

Star & Griesemer (1989) introduce the concept of *boundary objects* that characterize common intellectual tools, which play the role of containers and carriers, "both plastic enough to adapt to local needs and constraints of several parties employing them, yet robust enough to maintain a common identity across sites. They are weakly structured in common use, and become strongly structured in individual site-use. Like a blackboard, a boundary object 'sits in the middle' of a group of actors with divergent viewpoints".

The ordering and registration of animals in the museum is one example of a boundary object, a map is another. They are both there to be used by all the users of a museum, though these users use the boundary objects in very different ways.

CUSTOMIZATION

- Collaboration involves a continual adaptation, appropriation, and reconfiguration of technologies, artefacts, and environment. These are performed as a result of periodic organisational changes, local group customisations, and individual patterns of use. As a result, the infrastructure needs to be built for the expectation that change will occur.
- The manipulation of structures that results from continual adaptation, appropriation and configuration of technology can interfere with the shared understanding and intelligibility.

REWARD

- The maintenance and upkeep of data sets requires community commitment and appropriate incentives and rewards.
- There are no clear rewards for tool building.

PRIVACY

- Human mediators, even a directory service, can incur privacy issues.
- Query trails can incur privacy issues.

TRUST

- Trust and reliability of information is a concern.

PUBLISHING

- Choices exist around what and when to publish. This touches upon issues of the intellectual property (IP) and ownership.
- Front stage publication of data can indicate closure, e.g., a completion of data analysis, and hide the ambiguities and other issues that they have been encountered during the analysis process.
- Published data may be incomplete in that it does not contain information about how it was produced and what informal data was available at its conception, or it is in stored a redundant format that is no longer readable.
- The data may require linkage across multiple content files and format. When metadata about the link is not maintained properly the organization of the data is impaired and the meaning may be lost.

SUPPORT

- Some users revel in “playing around” with computing whilst others need technical support.
- “Ownership” of a computing problem can be locally determined.
- Baseline computing expertise is unevenly distributed.
- There is a clash of cultures between computer scientists’ tools and domain experts’ knowledge and training.

SYSTEM

- The development and use of the infrastructure is defined by complex social and organizational relationships.
- There is incompatibility across potential collaborators at the system and a tool level.
- The granularity of the metadata and data identifiers determines the granularity at which the data is accessed and can be exploited in further analysis. The usage of the data may introduce a requirement for a finer control over data which could not have been anticipated from the outset.
- Legislative and regulatory frameworks determine data attributes that need storing.

Legal Issues

COMPLIANCE AND AUDITING

- Legislative and regulatory frameworks may dictate the access and schema requirements for the data documentation.

Some legislative frameworks may require service providers to be auditable. For instance, the process that produces the data may have to be compliant with existing regulations. To this end, the documentation about the process should be accessible in order to be verified for compliance with regulations. Such function is performed by a regulator who acts as a specific 'consumer' of the data services, given the right to retrieve the data process documentation and analyse it.

- Legal requirements may determine the curation practices and documentation characteristics.

In order to be useful, the process documentation has to satisfy several important properties. It must be immutable and non-repudiable. It should be factual and at appropriate level of detail. Overall, its properties must ensure that one cannot deny that a given process execution took place as described. Since it often needs to be stored for a long time and published, its usage has implications for the storage and service planning.

- Auditing requirements may span all the processes in the data lifecycle.

Indeed, all the steps in the data loop including curation, storage, and publishing may be subject to audits. This is often the case when commercial services claim to offer a good quality of curation and preservation, third party auditors may wish to examine the process that curators and preservation personnel apply. Furthermore, there are instance when regulatory authorities agree with particular application owners on the third party repositories that are trusted to enforce the expected properties of process documentation. These repositories are then subject to detail audit.

ACCESS CONTROL

- Access to data content may need to be regulated separately from the access to data documentation.

Access control over process documentation is crucial and as important as access control over data. The two need to be handled separately and with care. Auditors may not be able to access actual data, e.g., for privacy reasons, but should gain access to information about the processes that resulted into the data production. Thus, the process documentation may be stored in repositories that are different from those used for storing data.

DATA LICENSES

- Storage plans and enforcement of legal access rights are frequently not synchronized.

During the planning of data ingestion and storage, legal issues are generally not even considered. Are the licenses associated with the formats, algorithms, and intellectual property rights likely to last as long as the preservation of the content? If the digital rights management (DRM) mechanism exists, does it have the same potential longevity as the content which it protects?

Economic Issues

INITIAL COST

- Implementing an infrastructure for supporting the data lifecycle is expensive and has a high start up cost.

STAFFING

- Scientific data curation is a new field and lacks clear definition of professional skills that are required and recognized career paths for the key roles involved in the data cycle.

Indeed, in many cases the data curation is an altruistic activity from which the corporation or a team performing the work is unlikely to reap the immediate benefits from. The career prospects and the community respect are relatively low compared to other computing careers. Indeed, data curation activity is often classed as “cataloguing”, i.e., a “library” activity within technology focused research communities for whom such an adjective is pejorative.

REWARD STRUCTURES

- Lack of incentives for persistent and consistent high quality curation.

Metadata creation is hard work, requiring persistence and technical acumen, and yet there are no suitable reward structures. Similarly, planning for process documentation, and ensuring that the production software is equipped with automatic logging is often perceived as an un-necessary overhead. These are seen as slowing down the production of software that would otherwise focus on “useful” functions.

- Lack of informal rewards through the community awareness and appreciation.

In practice the reuse of data, internally and externally, has not been valued and given high priority. Individual group’s interests thus prevail. It is often assumed that data management only entails preserving the bits and bytes.

INVESTORS AND STAKEHOLDERS

- Lack of clarity around possible economic and business models.

It might be that the only sustainable model that insures the quality tools and processes are competitive business models which promote quality services that use best practices in data curation and preservation. However, it is not clear who the shareholders are and what their business incentives would be. We speculate that the slow progress in the production and consolidation of scientific data curation tools is due to the lack of clear business proposition for the players who can invest in the creation of such tools.

It is interesting to recall Ted Nelson (http://www.xanadu.com/) who associates a micro payment mechanism whenever data is used. Nelson promoted Transpayment: “Minuscule payment systems that will allow a user to buy and assemble electronic documents (even web pages) transpublished from various sources–with exact microscopic payment to each source for each piece, pro rata by character”, (http://www.g4tv.com/techtvvault/features/4605/Ted_Nelson_Hypertext_Pioneer.html).

Nelson's team has demonstrating the prototype, the HyperCoinTM system. The transpayment gives an incentive to annotate data properly as the quality of annotations impact the data services and their usage.

Technology and Management Issues

Even when the importance of scientific data management is understood, the technological issues present a barrier for implementing sound practices and services. Furthermore, there is a lack of experience and management processes to support the vision and its implementation in the particular application domain. Here we outline some of the major issues.

TOOLKITS

Probably the single biggest barrier to creating systems that support the digital object lifecycle is the lack of effective toolkits to build and maintain required services and programs.

- Requirements on the documentation of the data loops yields to the complexity of the programs

Scientific data services and programs are complex and made more elaborate by the requirement to manage the data and processes themselves and the information about the data and processes that is dynamically changing with the use of the system.

- Tools are often immature and based on programming paradigms that are not yet established.

Existing tools are immature, based on programming paradigms that are on the bleeding-edge of technology. This leads to systems which cannot interoperate, either because their architectural concepts are incompatible or because the toolkits themselves are not interoperable (or both). In practice, to be useful, such toolkits need to be easily extensible (because they cannot be complete by definition, since any new digital process must be decorated with methods to generate and capture ancillary information). They need to be easily used by the information communities as well as (or perhaps instead of) computer specialists – the curation experts are going to be domain experts, and it is unrealistic to expect them to be using complex computer languages and libraries.

- It is necessary to ensure the compatibility of the toolkit with the higher level languages used in the community

There are no suitable tools that can be easily accessed and deployed in a familiar manner by the scientific and engineering communities. These tools need to support the processing that is performed in the languages with which these communities are familiar with. For example, providing a toolkit in Java for a community that primarily uses Python or Fortran is not a suitable support. Even where toolkits exist in native languages, they do not cover a sufficient range of issues such as support for an effective query structure over ancillary information.

Because the aim of the work is to ensure longevity, the toolkits need to be based on open protocols, standards and data formats, and they need to exploit a coherent architectural framework with common interfaces.

- Toolkits for metadata generation and editing are needed.

These toolkits need to support the development and evolution of specialised metadata editors which are easy to use and specialised to the target communities. They must be easily configurable, and they must exploit templates which minimise the information which needs to be re-used. In today's world it is recognised that the schema generation, evolution and maintenance are necessary but hard the existing tooling is poor.

TRAINING

- Training is sporadic and generally not for the tools in use.
- Query languages can require the use of complex and unfamiliar technologies.
- In view of technological advances, there is a need to maintain a skill set to deal with older versions of software.

  Thus, it is essential to manage the re-education of the staff and deal with the turnover adequately.

- It is necessary to identify ways to exchange knowledge among communities and support them.

WORKFLOW AND THE SCALE OF OPERATIONS

- Workflow support is essential to handle complex and large scale e-Science processes.

In the context of e-Science, complex processes are frequently expressed as workflows describing the composition of simpler processes.

The challenge in this context is that large data sets have typically to be operated over by complex processes. Hence, a large number of processors, large storage space, high-bandwidth communication networks, etc., need to be coordinated in order to undertake such computations efficiently. This is usually referred to as Grid Computing.

A lot of research activity is concerned with the design of workflow languages, user tools to express workflows easily, discover workflows and adapt them to specific purpose and systems to run workflows.

ANCILLARY DATA STANDARDS AND METHODS

- There is a lack of standard methods for creating ancillary information that is associated with data and processes.

In order for data to be understandable and usefully stored, it is necessary to provide documentation, i.e., ancillary information about the process that led to data production. Currently, there exist no standard methods to make such documentation available and no standard model to structure such documentation.

We note that some computing environments may provide specific logs about their execution. Such are an http server log or a computational grid monitoring service. However, they adopt different structures and semantics. It is, thus, the role of the tool developers to agree on the standard documentation schemas and standard ways to record such documentation, and make sure that the toolkit provides the adequate support.

- It is crucial to ensure consistency in data representation and access methods.

Given the iterative nature of the data loop, the process of curation, for instance, may result in specific indexes that are further used to discover and retrieve particular data, which in turn yields a particular final result. All the processing steps that are involved in producing the result have to be documented uniformly since each step that is causally related to the final result may have a significant influence and that may be relevant for further analyses. Conversely, all the data should be represented in a way that ensures consistency in data description, discovery, and access. Otherwise we may introduce a bias into the data search process.

DATA REPRESENTATIONS AND PUBLISHING

For data to be publishable, it needs to be accompanied by suitable descriptive information, known as ancillary data that enables the classification, discovery and reuse of data. This information includes details about the format, structure and semantics, so that the data can be maintained over time.

We already have a good understanding of what ancillary data is required for good ingestion and curation practice. A good portion of ancillary data can be derived by querying the documentation about the processes that are executed and logged automatically. Process data that is already in the human readable form can be used in further stages to complement readable metadata. For example, it could be used for data access, as is the case with inverted indices of textual data.

Similarly, in order to facilitate provenance information requests, it is important for the technology involved in processes to generate logs about the process itself. This is the data that complements the content and metadata in provenance analysis.

- It is important to establish best practices for executing processes using software technologies that include logging facilities.

Most of the home grown tools focus on the immediate need of the process rather than down the stream usage. It is important to enhance the technology to create data records about the particular stage in the process and the data involved. Furthermore, all the logs need to satisfy the common interface for access and search in order to be used for provenance analysis.

- Recording metadata related to data processing requires the definition of flexible and expandable schemas.

Data processing typically involves enhancement with metadata, in various stages. Ensuring that the data moved from one stage to another complies with the corresponding representation schema is a challenge. If the schemas are not designed in advance in a consistent manner then it is important to introduce a validation phase at each stage. Typical approach is to add additional data and count on the redundancy in the metadata as a safeguard towards the miss-placement or loss of data.

- The quality of the ancillary data, process documentation, and query methods determine the scope and depth of the data provenance analysis.

Detailed documentation about processes that led to the data enables the users to derive the data provenance. We note that there is not a unique provenance for a given piece of data. Data provenance depends on the perspective and the interest of the data user. For instance, a person may be interested in different restorations of a painting while others may require information about owners of the painting over time.

Provenance interests are specified by users and therefore should be expressed as queries over process documentation. Repeated provenance requests, across members of the user population, may be facilitated by pre-computed results. However, in its essence, data provenance is the result of the user’s search over the available information about data and processes and, for the inter-operability purposes, we should aim at standardizing the querying methods.

STORAGE MIGRATION

- The evolution of production tools and storage media require the migration and transformation of data into alternative formats.

Technical challenges around the storage involve the evolution of the storage media and the production tools. If we are trying to preserve, for example, a WordPerfect document in its original form, we need to ensure that it is stored on the media that is readable at the time we need to access it and that we have an appropriate application version available to process it.

In some instances, the transformation of the format is the very remit of the data management effort. For example, digital libraries' role is to preserve the original, unaltered document. We understand that requirement well when applied to analogue documents, i.e., the physical document archives. However, the same may simply not be feasible in the digital media.

It is interesting to note that the preservation of physical objects, such as manuscripts and audio tapes, is now attempted through digitization, introducing the appropriate digital format but, nevertheless, exposing it to even more volatile digital object life.

DATA ACCESS STRATEGIES AND SUPPORT

- Data access involves a variety of strategies and supporting tools that need to evolve with the user information needs.

Main data discovery and access methods comprise (1) search over metadata in order to gain access to the data itself, (2) search and analysis of the data itself when retrieved, and (3) routing or distribution of information to the consumers based on their interests and needs. Challenges around search involve handling of heterogeneous data types, from structured data that is automatically or manually generated and stored in the databases, to the free-text content and annotation about the data. Furthermore, the metadata and the representation of the data may not satisfy new types of data requests and search queries.

Data classification using controlled vocabularies provide alternative ways of searching and browsing the data. However, the tools for creating and maintaining controlled vocabularies are often home grown. Organizations may create their own workbench for data classification to enable editorial support and quality control.

Conclusions and Recommendations

From the several efforts in scientific data management we identified the main phases of the data lifecycle: the production, ingestion, storage, curation, publishing, and the consumption of data. While these are distinct roles that shape the life of a digital object in the system, the system itself accommodates multiple data loops at any give time. This results from the fundamental requirement of the system design to provide a 'self-documenting' function. The system stores and manages digital object but also creates digital records about the processes applied to the object. These records become digital objects themselves that are subject to the same data management process. Consequently, the systems are technically complex.

However, equally important and influential is the human factor. The data management processes are put in place by teams of people with a common interest in preserving and exploiting the data. Thus the characteristics of the teams and their objectives shape the type of the data management system and supported services. With the evolution of technologies and decrease in computing and storage costs we are witnessing a growing interest in tackling scientific data management challenges across a wider community. This creates new opportunities for beneficial services and a significant impact on the progress of science. At the same time it introduces a host of new issues that have not been present in the smaller scale efforts within individual organizations.

Here we summarize our discussion by pointing out several guiding principles and providing suggestions for specific community wide initiatives.

SUPPORT THE CHANGE

'Change' is one major trait of the data lifecycle and thus we ought to plan for it. Advances in technology challenge the stability and coherence of existing architectures, schemas, provenance methods, and storage. We plan and implement migration of data, services, and technologies. Similarly, reorganizations of teams are inevitable and ensuring that the required skill set stays within the organization is a challenge.

BUILD ON COMMONALITY AND DIVERSITY

The second important aspect is the diversity of perspectives and common approaches involved in handling the data. Scientists use the tools that are optimal or most comfortable for the particular task. They collaborate and make formal alliances with teams they are close to and compatible with. Comfort drives many decisions and established practices. These are the principles of collaboration and need to be enforced in providing support for data management systems. They are linked to the community based award system and provide the incentive for altruistic contribution in the community.

Having these in mind, we suggest focusing on several areas. First and utmost, we need to work towards a framework that will enable the community to engage, communicate, and exchange knowledge and experience. We expect that everything else will follow from that solid community base, from the development of technologies to establishing best practices and identifying self-sustainable models for data preservation, whether collaborative or competitive. Here we outline our recommendations in more detail.

FRAMEWORK FOR CONNECTING THE COMMUNITY

- Define the strategy and the framework for unifying the scientific community across disciplines, organizations, and interest groups. The framework should exploit the strengths of economic, social, and scientific interests and build a sustainable model for collaboration. We expect that establishment of such a framework is crucial for making any progress on addressing the scientific data management and related issues.

PROTOCOLS, INTERFACES, AND BEST PRACTICES

- Define open protocols, interfaces, and data models that allow the community of users to develop interoperable tools and services to support all the steps of the data chain.

- Define reference implementations for the open protocols and interfaces that are configurable, extensible, and deployable in varied environments that end users make use of on a day to day basis. This, in particular, requires effective binding to scripting/programming languages.
- Develop tools and practices for curation of protocols, interfaces, data models, libraries, and tools. This requires careful design of metadata, i.e., descriptions that make them easy to discover, share, deploy, and reuse.
- Provide tools to help create, define, and process domain specific metadata. Identify both domain-independent and domain-specific metadata and annotation schemas, and use them to guide the standardization efforts.

TOOLS

- Provide a toolkit stack, i.e., development layers for handling software and hardware, which can be easily brought into the user environment.

This applies to a wide range of users, from administrators to toolkit developers and data consumers. Most of the people involved in curation are specialists in the data area. They identify what services are needed and develop them. While some teams have software engineers, a significant amount of work is undertaken by data specialists, not software developers. For this reason, software toolkits must be robust and exploit high level languages, such as Python, Ruby, Perl, and similar. They should be easily extensible and used by individuals who are not professional software engineers.

- Enable a community of developers to create a corpus of reusable components.

Provide a framework for developers to connect, share and develop reusable software components. That may include a range of initiatives from informal developers' gatherings and forums to joint investments by interest groups into selected problem areas.

Develop tools and practices for curation of protocols, interfaces, data models, libraries, and tools. Carefully design the metadata, i.e., descriptions that make them easy to discover, share, deploy, and reuse.

- Invest in tools to help generate, maintain, and index information about the data and processes.

Process documentation and ancillary information are the basis for further consumption of data and ultimately determine the impact that a data management system has.

Besides the lack of standards, we recognize an acute need for effective tools for metadata generation and management. The tools need to be easy to use across user communities and in different software environments. In particular, they should integrate well with software environments that are native to major science and engineering communities. They need to support the migration of underlying information models, even when scale and complexity become issues.

With regards to the ancillary information and process documentation, we need to enable an appropriate level of automation. In particular, it is important to provide the means for constructing the descriptions of algorithms that are applied by systems' processes and record information about data transformations that result from the deployment of algorithms and processes. This would help promote the current status of algorithm description and information indexing from an obscure academic discipline to a key activity which will underpin our investment in the digital objects.

References

Open Archival Information System (OAIS), http://ssdoo.gsfc.nasa.gov/nost/isoas/. Other useful sites: http://www.rlg.org/en/page.php?Page_ID=3681, http://www.rlg.org/en/page.php?Page_ID=377.

Start, S.L. & Ruhleder, K. (1994). Steps towards an ecology of infrastructure: Complex problems in design and access for large-scale collaborative systems. In Proc. CSCW'94, pp 253-264. ACM.

Dourish, P., Lamping, J. & Rodden, T. (1999). Building bridges: customisation and mutual intelligibility in shared category management. In Proc. Group, pp 11-20. ACM.

Bannon, L. & Bodker, S. (1997). Constructing common information spaces. In Proc. ECSCW. Elsevier.

Star, S.L. & Griesemer, J.R. (1989). Institutional Ecology, Translations and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. Social Studies of Science 19, 387-420.

Theodor Holm Nelson, Project Xanadu http://www.xanadu.com/.

Paul Groth, Simon Miles, Victor Tan, and Luc Moreau. Architecture for Provenance Systems. Technical report, University of Southampton, October 2005. http://eprints.ecs.soton.ac.uk/11310/ .

Provenance in Organ Transplant Management Applications: Application Scenario Summary. http://twiki.gridprovenance.org/bin/view/Provenance/OrganTransplantManagement.

Provenance Aware Service Oriented Architectures (Project PASOA), www.pasoa.org.

EU Provenance Project (Enabling and supporting provenance in Grids for complex problems) www.gridprovenance.org.