Semi-Automated Preservation and Archival of Scientific Data using Semantic Grid Services

Jane Hunter, Sharmin Choudhury

DSTC,
The University of Queensland,
Brisbane, Australia
jane@dstc.edu.au
sharminc@dstc.edu.au

Abstract

Addressing the long term preservation issues associated with scientific data is a complex challenge compounded by: the scale and multidisciplinary nature of the problem; the wide range of formats and data types involved; and the often proprietary and transitory nature of the hardware and software used to generate the data. In this paper we present the PANIC system - an integrated, extensible architecture based on preservation metadata, automatic notification services, software and format registries and semantic grid services - that we believe, offers a sustainable, dynamic approach to the long term preservation of large collections of heterogeneous scientific data.

Keywords: Preservation, Semantic Grid Services, Scientific Data

1 Introduction

Addressing the preservation and long-term access issues associated with digital scientific data is one of the key challenges facing research, government and scientific organizations today. High performance grid computing, space and earth observation sciences and large scientific experiments produce enormous quantities of data that require effective and efficient management. Digital data files and information objects require constant and expensive maintenance because they depend on hardware, software, data, models and standards which are upgraded or replaced every few years. Accelerating rates of data collection and content creation and the increasing complexity of digital resources means that many organizations can no longer keep pace with the preservation needs of all of the data entrusted to them.

In the field of science the problem is compounded due to heterogeneous nature of the large quantities of data being generated from experiments and observation. For example, astronomers are producing Flexible Image Transport System (FITS) files and VOTable compliant XML files [1]; geoscientists are producing Arc/Info SHP files and GeoTIFF images; bioscientists are generating huge genomic databases and associated EST (expressed sequence tags), GSS (genome survey sequence) and HTGS (high throughput genomic sequence) files; medical scientists are generating vast sets of DICOM images; weather researchers are generating HDF5 datasets.

The task of ensuring long term access to scientific data collections is so overwhelming that scientists spend much of their time managing the data through special purpose handcrafted solutions, rather than using their time effectively for scientific investigation and discovery. This has been partly due to the lack of available tools and services but largely due to the heterogeneous nature of scientific data, not only between different science domains but also within science domains. Organizations such as CODATA [2] have been active in promoting improved management of digital scientific data and the importance of preserving contextual information with the archived scientific datasets. In particular CODATA recommend the use of the OAIS model [3], a high-level conceptual model developed by a consortium of space agencies, to facilitate scientific data management. However to date there has been limited availability of practical, available preservation tools and services. This is

beginning to change primarily as a result of activities in the digital library domain. For example, Cornell's Virtual Remote Control (VRC) project [4] and OCLC's INFORM [5] project are developing risk measurement and notification services. The Global Digital Format Registry (GDFR) [6] initiative, the UK National Archive's PRONOM project [7] and VersionTracker [8] are developing format and software registries that can be used to determine required preservation actions. The UK Digital Curation Centre is developing a Representation Information Repository using ebXML [9]. Projects such as the Typed Object Model (TOM) [10] and IBM's UVC Emulation project [11] are generating migration and emulation services. Many scientific communities are developing their own sets of migration services (e.g., converter programs). Currently each of these components is being developed independently but altogether they can be leveraged to help build a complete preservation solution.

Moreover, it is generally recognized that there is no single best solution to digital preservation. *Differences in the needs and practices of various scientific disciplines, make it difficult if not impossible to define a 'one size fits all' approach to selecting, appraising and retaining scientific data.* [12] The most appropriate strategy depends on the particular requirements of the custodial organization, the producers and users of its collection and the nature of the objects in the collection. Hence within the PANIC project [13] we combine the efforts of the different domain-specific preservation initiatives by integrating the range of tools and services being developed into a single encompassing Grid framework. More specifically PANIC uses a flexible, dynamic, semi-automated approach which provides access to a range of metadata tools and risk assessment, notification, emulation and migration services through a Semantic Web/Grid services architecture.

The remainder of the paper is structured as follows. The next section describes the system objectives through a motivational example. Section 3 describes the overall system architecture. Section 4 concludes with an evaluation of the results and a discussion of problem issues and future work.

2 Motivational Example

Russel Coight is an astronomer at the Australian Telescope National Facility (ATNF). He has a massive collection of legacy FITS (Flexible Image Transport System) files. The International Virtual Observatory maintains online registries of the latest recommended formats, format versions and software tools and services, (for authoring, rendering, viewing, and converting files) for the astronomy community. The PANIC system periodically compares the metadata associated with files and information objects in the ATNF collection with the IVO's registries. PANIC determines that IVO now recommends that FITS format files should be replaced with VOTable 1.1 format because the latest version of the Xanadu editing and analysis software will no longer support FITS files. The system sends an email to Russel Coight notifying him that certain datasets that he owns are in danger of becoming obsolete. The email message includes a list of the endangered FITS files. The message also recommends that the FITS format be replaced by VOTable 1.1 which has now become the defacto astronomical data standard, recommended by the IVO. Russel Coight must now find a FITS-to- VOTable conversion service that meets all his service quality parameters. He uses the PANIC system to specify the parameters he requires in the conversion service. For example, he specifies that he requires a free service that converts from FITS to VOTable 1.1. He prefers a distributed converter that will process multiple files concurrently using parallel processors on the Grid. He also specifies that the distributed converter must be highly reliable, high-speed and not result in any loss in data quality. His request is handled by a Discovery Agent which searches a Grid Service registry for the appropriate service description. The Discovery Agent cannot find any exact matches to Russel's request but it can find one near match and two conversion services which can be chained to approximate the required service:

1. A service which converts from FITS to VOTable 1.1 but which is lossy - developed by NASA'a HEASARC
2. One service which converts from FITS to VOTable 1.0 and another which converts from VOTable 1.0 to VOTable 1.1– both are lossless, reliable and high speed and can be distributed across GRID processors.

The Discovery Agent presents these alternatives to Russel, ranked according to how well they match his request. Russel is able to choose his preferred option and the selected Provider Agent then executes this service, (ideally processing multiple files in parallel across Grid procesors e.g., [14]) and returns the VOTable 1.1 files. After migration is complete, the associated provenance and events metadata (which records a history of preservation actions associated with each digital object in the collection) is also automatically updated.

The next two sections of this paper describe the architecture, components and implementation details of the PANIC system that we have developed in order to turn the scenario outlined above into reality.

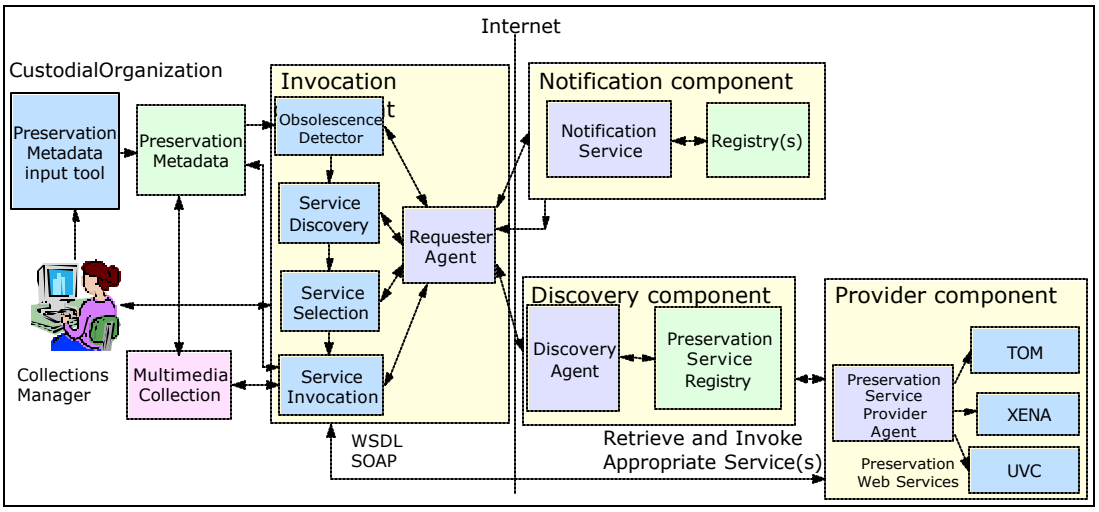

Figure 1: PANIC System Architecture

3 System Architecture

The PANIC system comprises three main components:

- Preservation metadata generation tools;
- Obsolescence Detection and Notification services;
- Preservation Service Description, Discovery and Invocation.

These three components are described in detail in the following subsections. Figure 1 illustrates the overall system architecture.

3.1. Preservation Metadata Schema and Capture Tool

There is strong agreement that preservation metadata is crucial to the long term preservation of digital objects. CODATA [12] recommend that scientific organizations adopt the OAIS model [3]. METS [15], a Digital Library Federation initiative, builds on OAIS and provides an XML document format for encoding metadata necessary for both management and exchange of digital objects within and between repositories. Consequently the first phase of PANIC involved developing:

- a preservation metadata schema based on METS but with extensions for audiovisual and discipline-specific needs to support a wide variety of atomic and composite digital objects;
- a preservation metadata capture tool (PREMINT) based on this schema.

Although METS is the more widely used preservation metadata schema, MPEG-21 [16] has also been applied successfully as a preservation metadata format [17]. Hence we decided, for purposes, to develop an MPEG-21 compliant schema to support our preservation metadata requirements.

Based on these metadata schemas, we developed both a stand-alone Java application and a JSP (Java Server Pages) version of the PREservation Metadata INput Tool (PREMINT) metadata input tools. The application consists of a set of metadata input forms, constrained by the underlying XML Schema. PREMINT collects metadata by dynamically presenting the user with a series of forms that collect: Descriptive Metadata, Technical Metadata, Instrumentation Metadata. In order to streamline the preservation metadata capture process we plan to integrate services such as JHOVE, the JSTOR/Harvard Object Validation Environment [18], to automatically extract the format-specific technical metadata and services that automatically record scientific instrument settings (e.g., OME Open Microscopy Environment) to capture precise provenance data. Figure 2 shows a screenshot of the PREMINT tool, illustrating the technical metadata input form for a video file. Users have the option of saving the metadata output to either METS+Extensions or MPEG-21.

3.2. Obsolescence Detection and Notification

The Obsolescence Detection module periodically compares the preservation (formatting) metadata for each object in the collection with information stored in the following three registries:

- Software Version Registry – this contains information about the latest versions of authoring, rendering, viewing, editing or analysis software required to access and use objects in the collections.

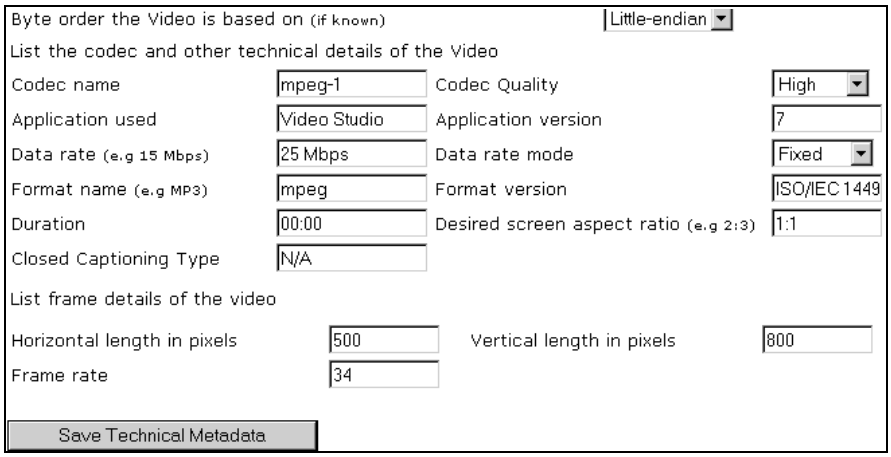

Figure 2: Screenshot of the PREMINT Metadata Input Form

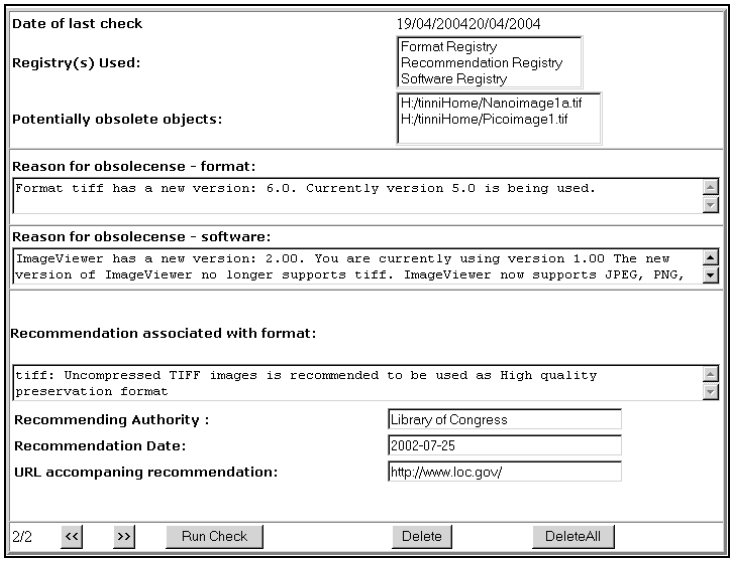

Figure 3: PANIC Notification Screen

For each software tool, the registry stores: Title, Description, Creator, CurrentVersion, ReleaseDate, DeveloperPage, License, Requirements, DownloadSite, DownloadSize, Rating, EaseOfUse, FormatsSupported, Features, Stability, Price.

- Format Registry - this contains detailed information about digital formats including: Identifier, Description, Version, Author, Owner, RelationshipsOtherFormats, ApplicationsUsingThisFormat; FormatSpecification; ProvenanceEvents etc [19]
- Recommended Format Registry – this tracks the latest recommended preservation formats and the authority making the recommendation e.g., "the International Virtual Observatory (IVO) recommends VOTable as the preferred preservation format for astronomical data".

For the purposes of demonstrating the PANIC prototype, we have developed three MySQL databases containing sample data, to represent the registries. However there are existing initiatives focusing on developing and maintaining the corresponding real-world registries. For example, Global Digital Format Registry (GDFR) and PRONOM are both developing digital format registries for the purposes of long term preservation. These mainly focus on digital library applications and would have to be extended to include information relevant to scientific data formats. VersionTracker [8] also maintains a website with a human searchable registry of software versions that enables users to determine whether they should update, upgrade or patch their existing applications. Again, this mainly focuses on commercial software. We envisage that each specific scientific community (e.g., astronomers) would have an authoritative body (e.g.,

IVO) responsible for maintaining registries of recommended formats and available software versions.

When there is an incompatibility between a digital object's current preservation/formatting metadata and the latest version recommended in the registry, then a message is sent to the owner of the data or some nominated person(s) or software agent, notifying them of a potential risk. Figure 3 is an example of a notification window.

3.3. Preservation Service Description, Discovery and Invocation

Our aim is to build a system which dynamically incorporates the expanding range of preservation services available and also provides decision-support tools or recommender services which can assist the scientific data manager to select the best single service or combination of services for a particular digital object or a particular set of circumstances.

The modular, distributed nature of the Semantic Grid/Web services architecture makes it perfectly suited to the dynamic, large-scale, heterogeneous nature of the digital preservation problem. A key objective of the PANIC project is to test this hypothesis by developing and evaluating a semi-automated preservation system based on the Semantic Web services architecture which provides access to a suite of independent preservation service components which can be discovered, linked, and invoked in arbitrary combinations across the Grid to fulfil the specific preservation tasks and requirements of different scientific organizations.

3.3.1 Semantic Web/Grid Services

Web services are enabling networked computer programs to process and consume information. Based on open standards such as XML, SOAP and WSDL, Web services provide a standardized way of enabling Web-based application-to-application interoperability. More recently the Semantic Web services initiative has developed OWL-S/DAML-S, an OWL ontology which enables Web services to be described semantically and their descriptions to be processed and understood by software agents. A number of projects are using OWL-S to describe their domain-specific services and enable software agents to automatically discover, compose, invoke and monitor the most appropriate Web services [20, 21]. As far as we are aware, no one is currently applying or extending OWL-S/DAML-S to generate semantic descriptions of digital preservation services so that they can be discovered, invoked and composed by software agents in order to automate the preservation tasks of large archival organizations.

3.3.2 OWL-S Ontology for Preservation Services

The purpose of OWL-S is to provide computer-interpretable descriptions of services so that they can be located, selected, employed, composed and monitored automatically over the Internet. Multiple web services can be matched and chained - interoperating to perform complex tasks and transactions for users dynamically and on-demand.

The OWL-S ontology has a top-level Service ontology with three main subontologies:

- ServiceProfile – provides a description of what the service does, enabling advertising and discovery;
- ServiceModel – provides a detailed description of a service's operation or how it works;
- ServiceGrounding – provides details of how to interoperate with or access a service using messages.

The advantage of OWL-S is that it provides generic upper-level classes that can be refined to describe any Web service. We have extended the OWL-S classes to create more preservation-specific subclasses. Figure 4 illustrates how we have extended the generic Service class by defining a *PreservationService* subclass. *PreservationService* has two subclasses – *emulation* and *migration*. These new types of service are defined in the PreservationService ontology, which extends the Service ontology provided by the OWL-Service Coalition [22]. The *normalization* service is defined as a further subclass of *migration*.

Within the ServiceProfile ontology, the *Profile* class provides three types of information:

- Service name, description and contact (person or organization);
- Functional description in terms of inputs, outputs, pre-conditions and effects;
- An extensible set of properties used to describe features of the service e.g., service category, quality rating, etc.

We have also extended the ServiceProfile ontology to create a PreservationServiceProfile.

Semantic Matchmaker [23] is used as the Discovery Agent in PANIC. When the collections manager (see Figure 1) specifies the parameters required in the preservation web service, a query is created and submitted to *Semantic Matchmaker*. ServiceProfiles are used to match service requesters to service providers. Requests from the service requesters, are converted to ServiceProfile documents

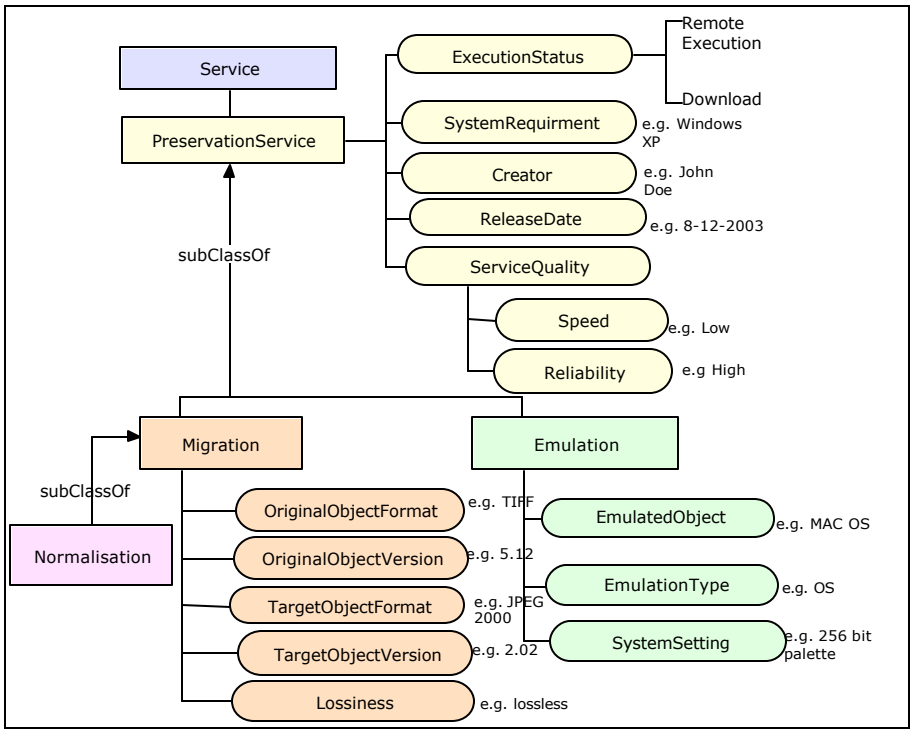

Figure 4: OWL-S Service and ServiceProfile Extensions

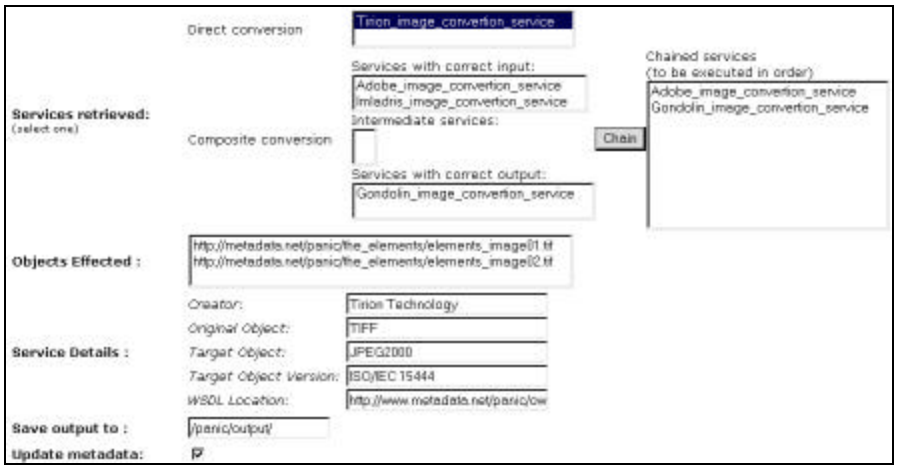

Figure 5: Service Selection Screen

and compared against the stored ServiceProfiles for available services. A ranked list of matching services is retrieved and displayed. Matching services can be either atomic or chained, composite services.

For example, a user might specify that he requires a service that converts from FITS to VOTable 1.1. He also prefers a distributed converter that must be highly reliable, high-speed and not result in any loss in data quality. His request is forwarded by the Requester Agent to a Discovery Agent which searches a Web Service registry for a service description matching this specification. In the past UDDI registries [24] were used to advertise available Web services but dynamic discovery was difficult due to the lack of semantics. Using OWL-S and *Semantic Matchmaker*, enables more precise and dynamic discovery of appropriate services. In this case, the *Semantic Matchmaker* determines that there are two matches to the service request: a simple process (which converts directly from FITS to VOTable 1.1) and a composite, chained process that first converts FITS to VOTable 1.0 and then converts VOTable 1.0 to VOTable 1.1. The matching service descriptions (the WSDL, ServiceGrounding,

ServiceProfile and ServiceProcess documents) are sent back to the *Requester Agent*. These are then used to present the search results to the user as shown in Figure 5.

3.3.3 Service Selection and Invocation

Given the results of the search and the recommendations of the Discovery Agent, the collections manager can (through the system configuration interface) choose to allow the system to automatically invoke the best matching service or interactively select a particular preservation action and invoke it manually. The system configuration interface enables the collections manager to set certain runtime parameters prior to service execution e.g., whether to process multiple files in parallel, where to save the output files, whether to update preservation metadata, where to email the logfile etc.

After service selection, the Requester Agent sends the inputs (e.g., FITS files) to the Provider Agent which schedules and executes the conversion services (ideally across multiple parallel procesors using a Grid DataFarm such as [14]) and returns the outputs (e.g., VOTable 1.1 files) to the Requester Agent. The Requester Agent saves the output files locally to the specified location and updates the preservation action metadata – recording what files were converted, when, authorized by whom, and the service that was used. Finally an email is sent to the user, notifying him that the migration of FITS files has been completed.

4 Conclusions

In this paper we have briefly described how PANIC, a prototype preservation system which we have developed based on: preservation metadata; software and format registries; and Semantic Grid Services, can streamline the long term preservation of scientific data.

The distributed nature of the proposed Web/Grid services architecture offers many advantages. It leverages existing work on preservation metadata and preservation software tools (e.g., emulation and migration services) by integrating them and making them available through a single interface. It enables institutions to coordinate and share their digital preservation activities whilst retaining the flexibility to meet local requirements. The proposed system is scalable and extensible. It has the potential to provide preservation services for a very wide variety of data and media formats. Because the system is based on standards including: METS, MPEG-21, XML, SOAP, WSDL, UDDI, OWL, OWL-S, interoperability between services and information is optimized. The design offers maximum flexibility - as an organization's preservation needs change, the system adapts accordingly. As new preservation services, tools, standards and recommendations evolve, they can automatically be incorporated into the system by adding their semantic descriptions to the relevant registries. As well as providing unified access to the wide range of preservation services available, the system also provides decision-support and recommender services to assist the scientific data collections manager to select the best single service or combination of services for a particular set of objects. The user interface allows easy customization of the system and human intervention where required – offering the best combination of human and software agents.

To conclude, we believe that the semantic web services approach, as employed within the PANIC system, provides the optimum architecture for viable long term preservation of large scale collections of scientific data. By enabling the automatic detection of potentially obsolescent data objects and the dynamic discovery and execution of the most appropriate preservation service – the system can potentially save research, government and scientific organizations vast amounts of time and effort, as well as prevent the loss of valuable data.

Acknowledgements

The work reported in this paper has been funded in part by the Co-operative Research Centre for Enterprise Distributed Systems Technology (DSTC) through the Australian Federal Government's CRC Programme (Department of Education, Science, and Training).

References

[1] R. Williams, F. Ochsenbein, C. Davenhall, D. Durand, P. Fernique, D. Giaretta, R. Hanisch, T. McGlynn, A. Szalay, A. Wicenec, "VOTable: A Proposed XML Format for Astronomical Tables", Apr 2002, http://www.us-vo.org/VOTable/VOTable-1-0.pdf

[2] Committee on Data for Science and Technology (CODATA) http://www.codata.org/

[3] B. Lavoie, "Meeting the challenges of digital preservation: The OAIS reference model", Feb 2002, http://www.oclc.org/research/publications/archive/2000/lavoie/

[4] N. McGovern, A. Kenney, R. Entlich, W. Kehoe, E. Buckley, "Virtual Remote Control: Building a Preservation Risk Management Toolbox for Web Resources", D-Lib Magazine, April 2004 http://www.dlib.org/dlib/april04/mcgovern/04mcgovern.html

[5] A. Stanescu, “Assessing the Durability of Formats in a Digital Preservation Environment: The INFORM Methodology” D-Lib Magazine November 2004 Volume 10 Number 11 http://www.dlib.org/dlib/november04/stanescu/11stanescu.html

[6] Global Digital Format Registry (GDFR) http://hul.harvard.edu/formatregistry/

[7] PRONOM –The file format registry http://www.nationalarchives.gov.uk/pronom/

[8] VersionTracker http://www.versiontracker.com/

[9] D. Giaretta, “Draft DCC Approach to Digital Curation”, Jan 2005. http://dev.dcc.rl.ac.uk/twiki/bin/view/Main/DCCApproachToCuration

[10] The Typed Object Model (TOM) http://tom.library.upenn.edu/

[11] J.R. van der Hoeven, “Permanent Access Technology for the virtual heritage”, May 2004 http://jeffrey.famvdhoeven.nl/Researchtask%20IBM%20TU%20Delft%20-%20J.R.%20van%20der%20Hoeven.pdf

[12] ERPANET/CODATA Workshop Biblioteca Nacional, Lisbon, Dec 2003 http://www.erpanet.org/events/2003/lisbon/LisbonReportFinal.pdf

[13] Preservation webservice Architecture for Newmedia & Interactive Collections (PANIC), http://www.metadata.net/panic/

[14] N. Yamamoto, O. Tatebe, S. Sekiguchi, "Parallel and Distributed Astronomical Data Analysis on Grid Datafarm", Proceedings of 5th IEEE/ACM International Workshop on Grid Computing (Grid 2004), pp.461-466, 2004.

[15] Metadata Encoding and Transmission Standard (METS) http://www.loc.gov/standards/mets/

[16] ISO/IEC TR 21000-1:2001(E) (MPEG-21) Part 1: Vision, Technologies and Strategy, MPEG, Document: ISO/IEC JTC1/SC29/WG11 N3939 http://www.cselt.it/mpeg/public/mpeg-21_pdtr.zip

[17] J. Bekaert , P. Hochstenbach and H. Van de Sompel, “Using MPEG-21 DIDL to Represent Complex Digital Objects in the Los Alamos National Laboratory Digital Library”, D-Lib Magazine, November 2003 http://www.dlib.org/dlib/november03/bekaert/11bekaert.html

[18] JSTOR/Harvard Object Validation Environment (JHOVE) http://hul.harvard.edu/jhove/jhove.html

[19] Global Digital Format Registry (GDFR), Data Model v.3 Dec 2003 http://hul.harvard.edu/gdfr/DataModel_v3.doc

[20] D. Wu, B. Parsia, E. Sirin, J. Hendler and D. Nau, “Automating DAML-S Web Services Composition Using SHOP2”, 2nd International Semantic Web Conference, ISWC 2003, Sanibel Island, Florida, USA, October 2003 http://www.mindswap.org/papers/ISWC03-SHOP2.pdf

[21] T. Tsai, H. Yu, H. Shih, P. Liao, R. Yan, S. Chou, “Ontology-Mediated Integration of Intranet Web Services”, Computer No 10, Volume 36, October 2003 http://computer.org

[22] The OWL Services Coalition, “OWL-S: Semantic Markup for Web services” July 2004 http://www.daml.org/services/owl-s/1.1B/owl-s/owl-s.html

[23] Massimo Paolucci, Katia Sycara, Takuya Nishimura, Naveen Srinivasan, “Using DAML-S for P2P Discovery”, International Conference on Web Services, ISWS 2003, Las Vegas, Nevada, USA, June 2003 http://www-2.cs.cmu.edu/~softagents/papers/p2p_icws.pdf

[24] OASIS, “UDDI Spec Technical Committee Draft”, Nov 2004 http://uddi.org/pubs/uddi_v3.htm