PANIC – An Integrated Approach to the Preservation of Composite Digital Objects using Semantic Web Services

Jane Hunter[1], Sharmin Choudhury[1]

[1]DSTC, Brisbane Australia
jane@dstc.edu.au
sharminc@dstc.edu.au

Abstract. To date, long term preservation approaches have comprised: emulation, migration, normalization and metadata - or some combination of these. Most existing work has focussed on applying these approaches to digital objects of a singular media type: text, HTML, images, video or audio. In this paper we consider the preservation of composite, mixed-media digital objects, a rapidly growing class of resources. We describe an integrated, flexible system that we have developed, which leverages existing tools and services and assists organizations to dynamically discover the optimum preservation strategy as it is required. The system captures and periodically compares preservation metadata with software and format registries to determine those objects (or sub-objects) at risk. By making preservation software modules available as Web services and describing them semantically using a machine-processable ontology (OWL-S), the most appropriate preservation service(s) for each object (or sub-object) can then be dynamically discovered, composed and invoked by software agents (with optional human input at critical decision-making steps). The PANIC system successfully illustrates how the growing array of available preservation tools and services can be integrated to provide a sustainable, collaborative solution to the long term preservation of large-scale collections of complex digital objects.

Keywords: *Preservation, Semantic Web Services, METS, MPEG-21, OWL-S*

1 Introduction

Addressing the preservation and long-term access issues of digital resources is one of the key challenges facing informational organizations such as libraries, archives, cultural institutions, scientific organizations and government agencies today. Digital 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.

A number of major initiatives have been established to tackle the problem of preservation of digital content. In 2000, the US Congress appropriated almost $100 million to the Library of Congress to establish a National Digital Information Infrastructure and Preservation Program (NDIIPP) [1]. Other initiatives such as the

CEDARS project [2], CAMiLEON [3], the National Library of Australia's PANDORA project [4], Networked European Deposits Library (NEDLIB) [5] and the PREMIS Working Group [6] have all been investigating strategies for the preservation of digital content. These initiatives have primarily been focusing on three main strategies: emulation, migration or some amalgam of these which relies on the encapsulation of the digital object with detailed preservation metadata. Recently a form of migration, known as *normalisation* (a process by which a digital object is converted to a software- and hardware- independent XML format) has gained popularity. The National Archives of Australia's XML Electronic Normalising of Archives (XENA) [7] project is an example of the normalization approach.

In addition most preservation projects have focused on preserving digital objects of a single media type e.g., web sites (HTML), electronic journals, electronic books, digitally recorded sound, digital moving images, or multimedia objects. Only recently have a couple of mixed-media preservation initiatives emerged from arts organizations and museums, wanting to preserve multimedia, new media or variable media artworks in their collections. *Archiving the Avant Garde* [8] is a collaborative preservation project between the Berkeley Art Museum and Pacific Film Archive (BAM/PFA), the Solomon R. Guggenheim Museum, the Walker Art Center, Rhizome.org, the Franklin Furnace Archive, and the Cleveland Performance Art Festival and Archive. The Guggenheim Museum has also established the Variable Media Initiative [9], now known as the Variable Media Network, which has invited media artists, curators, and museum specialists to a series of meetings to brainstorm strategies for preserving specific case study works in the Guggenheim collection.

The limited availability of practical, available preservation tools and services is beginning to change. For example, Cornell's Virtual Remote Control (VRC) project [10] and OCLC's INFORM [11] project are developing risk measurement and notification services. The Global Digital Format Registry (GDFR) [12] initiative, the UK National Archive's PRONOM project [13] and VersionTracker [14] are developing format and software registries that can be used to determine required preservation actions. Projects such as the Typed Object Model (TOM) [15] and IBM's UVC Emulation project [16] are generating migration and emulation services. Each of these components is being developed independently and individually offer only one piece in the complete preservation puzzle.

The National Library of the Netherlands recently implemented a Preservation Manager component within its e-Depot system to semi-automate the preservation of digital objects stored in its Digital Information Archiving System (DIAS) [36]. Although this approach is similar to the PANIC three-phase approach, it is confined to one institutional archive and is not designed to enable sharing, discovery, composition and invocation of preservation services via registries over the Web.

It is generally recognized that there is no single best solution to digital preservation. The most appropriate strategy depends on the particular requirements of the custodial organization, the producers and consumers of its collection and the nature of the objects in the collection. Hence within the PANIC project [17], our aim is to leverage the efforts of the different preservation initiatives by integrating the growing range of tools and services being developed into a flexible, extensible Web-based framework.

2 Objectives

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

The modular, distributed nature of the Semantic Web services architecture appears to make it perfectly suited to the dynamic, large-scale, heterogeneous nature of the digital preservation problem. Hence 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. We believe that such an architecture will enable access to a suite of independent preservation service components which can be discovered, linked, and used in arbitrary combinations to fulfil the specific preservation tasks and requirements of different archival organizations.

As shown in Figure 1, PANIC acts as the facilitator that assists organisations to share, discover, combine and invoke the optimum preservation services through its Semantic Web services architecture.

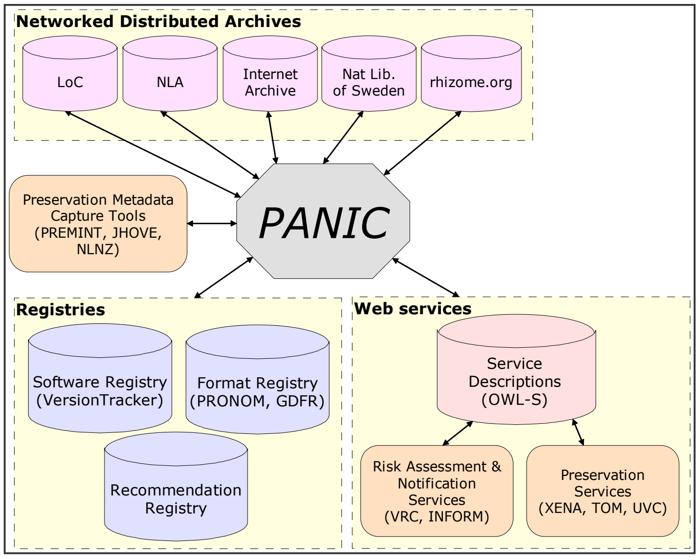

Networked Distributed Archives
LoC
NLA
Internet Archive
Nat Lib. of Sweden
rhizome.org
Preservation Metadata Capture Tools (PREMINT, JHOVE, NLNZ)
PANIC
Registries
Software Registry (VersionTracker)
Format Registry (PRONOM, GDFR)
Recommendation Registry
Web services
Service Descriptions (OWL-S)
Risk Assessment & Notification Services (VRC, INFORM)
Preservation Services (XENA, TOM, UVC)

Figure 1: PANIC integrates complementary preservation registries and services via a Semantic Web Services Architecture

The remainder of the paper is structured as follows. The next section describes the overall system architecture. Section 4 describes our general approach to the preservation of complex mixed-media digital objects. Section 5 describes the preservation metadata schema and capture tools that we developed. Section 6 describes the obsolescence detection and notification components. Section 7 describes Semantic Web Services, our extensions to OWL-S and the preservation service discovery and invocation component. Section 8 concludes with an evaluation of the system to date and a discussion on problem issues and future work.

3 System Architecture

The PANIC system comprises three main software components that were developed to support the following three steps in the overall preservation process:

1. **Preservation Metadata Capture**. This comprises tools which enable the generation of preservation metadata for either atomic or composite mixed-media digital objects. Details of the metadata schema and input tool are provided in the next section. The preservation metadata can be saved in either an extended METS schema [18] or an MPEG-21 Digital Item Declaration Language (DIDL) schema [19].

2. **Obsolescence Detection and Notification**. This software component periodically compares each object's/sub-object's preservation metadata with software and format registries which store information about the latest available authoring, rendering or viewing software and recommended formats. When there is incompatibility between an object's/sub-object's format and the latest available software or format recommendation, a notification is sent to the relevant agent (human or software). Quantitative risk assessment methodologies such as VRC or INFORM, could easily be incorporated to trigger this notification.

3. **Preservation Service Discovery and Invocation**. When preservation action is required, the system allows the collections manager to specify the attributes of the required preservation service. A Discovery Agent then dynamically discovers the most appropriate preservation service by matching the specified attributes against descriptions of available preservation services. This is implemented by making preservation software modules available as Web services and describing them semantically using a machine-processable ontology (OWL-S) [20]. Collections managers then have the option to choose from the ranked list of atomic or composite services retrieved by the Discovery Agent. Service Selection and Invocation Agents then select (and possibly compose) and invoke the most appropriate preservation services for that sub-object and update the provenance metadata.

Figure 2 illustrates the overall system architecture. In the next section we describe how this three-step process can be applied to both atomic digital objects of single media types as well as composite mixed-media objects. An online demo of the Preservation Metadata Capture system (PREMINT) is available at [35]. An online demo of the PANIC system (comprising of components 2 and 3 above) is available at [37].

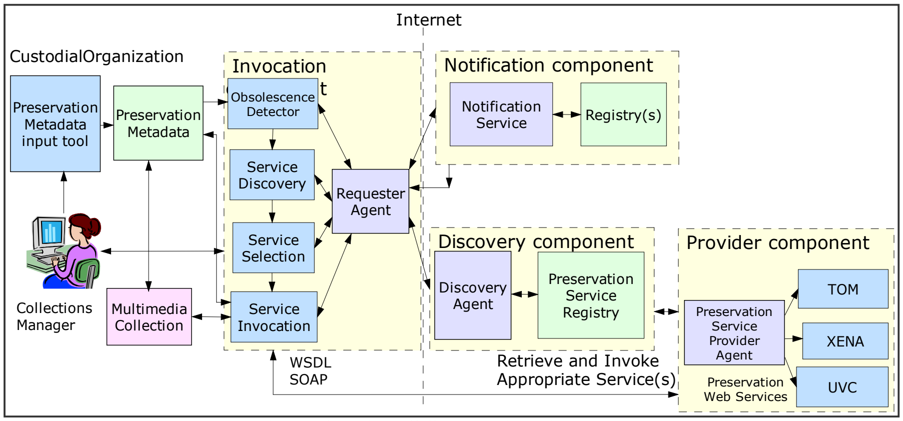


Figure 2: PANIC System Architecture

4 Preservation Process for Composite Mixed-Media Objects

Earlier research [21] carried out by the authors, investigated preservation strategies for a number of case studies involving new media artworks that were composed of a heterogeneous mix of different media types: images, video, audio, text, hyperlinks, software programs and physical objects.

As a result of this investigation we concluded that no single approach is the right one. Each object or class of objects needs to be evaluated on a case-by-case basis. However we did determine that generally a combination of preservation metadata plus migration to high quality, platform-independent, standardized formats is the optimum approach for composite multimedia objects. More specifically we proposed the following preservation approach:

1. Atomic single media components of the composite digital object should be stored in formats which are as high quality, standardized and platform-independent as possible (e.g., MPEG-2, MP3, TIFF);
2. Composite mixed-media objects should be stored in an XML-based structural markup format such as HTML, SMIL [22], XHTML+SMIL [23] or HTML+TIME [24], rather than proprietary formats such as Shockwave, Macromedia Director or Flash;
3. Sufficient descriptive, structural, administrative and technical metadata should be stored in order to interpret and display the canonicalized object(s);
4. Either METS [18] (with extensions) or the MPEG-21 Digital Item Declaration Language (DIDL) [19] should be used to encapsulate the platform-independent composite digital object with its descriptive, administrative, technical and structural metadata, into a complete preservation package;
5. The system should periodically consider the accessibility and preservation of each of the atomic sub-objects, prior to monitoring and processing the composite object.

By only considering those composite objects constructed using the XML-based languages, HTML, SMIL [22], XHTML+SMIL [23] and HTML+TIME [24], the preservation process is greatly simplified. Because these XML-based composite formats only specify temporal, spatial and presentation layouts for the atomic sub-objects, it means that the composite objects can be considered independently of the sub-objects and their formats.

Hence our approach is to capture the preservation metadata for each of the sub-objects first, prior to defining the structure and capturing the preservation metadata for the composite object. Similarly with the obsolescence detection and notification and the preservation service discovery and invocation steps. We consider the atomic sub-objects first and only after each of the sub-objects has been dealt with, is the preservation of the overall composite object considered. For example, all atomic JPEG images may first need to be migrated to JPEG-2000, after which the composite objects which contain these images may need to be migrated from SMIL 1.0 to SMIL 2.0.

In the next three sections we describe in detail, the three software components that were developed to support: preservation metadata capture; obsolescence detection and notification; and preservation service discovery and invocation.

5 Preservation Metadata Schema and Capture Tool

The Library of Congress's Metadata Encoding and Transmission Standard (METS) schema [18] provides a flexible mechanism for encoding descriptive, administrative, and structural metadata for a digital object, and for expressing the complex links between these various forms of metadata. The METS schema provides a standardized XML syntax for identifying the digital components of a composite digital object, encoding the different types of metadata describing the components and for expressing the relationships between the components. A METS document consists of seven major sections, including a METS Header, Descriptive Metadata, Administrative Metadata, File Section, Structural Map, Structural Links, and Behavior.

For the composite mixed-media digital objects that we are considering, METS provides an ideal base schema. However in order to support the preservation of mixed-media objects that contained images, video, audio and text, we determined [21] that the following refinements to the basic METS schema were required:

- technical audiovisual format information for each component object using extensions developed through the AV Prototype Project [25].
- The use of SMIL 2.0 or HTML+TIME in the Structural Map section to specify the overall structure i.e., spatio-temporal relationships between the component objects.

In addition, the new media artworks that we were considering [21], required three additional metadata profiles to capture information specifying:

- the artist’s intention;
- the artists’ attitude towards preservation;
- installation and presentation information.

Figure 3 illustrates the structure of the METS-based schema that we developed for capturing the preservation metadata required for a collection of new media artworks .

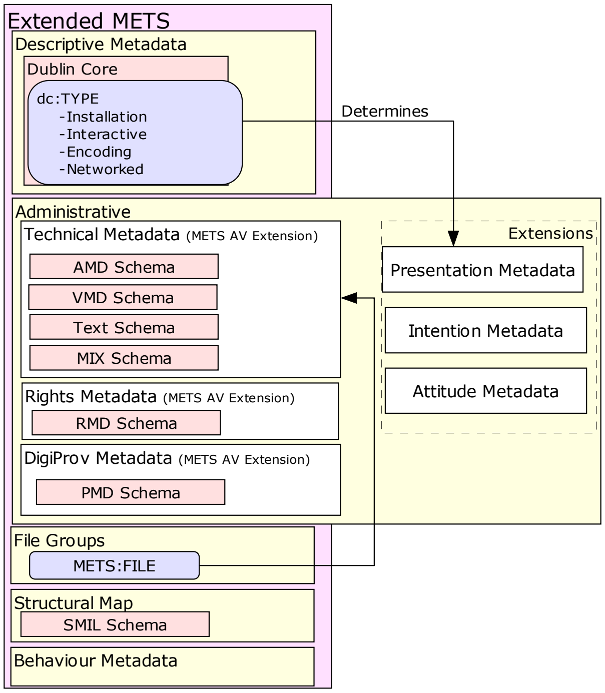
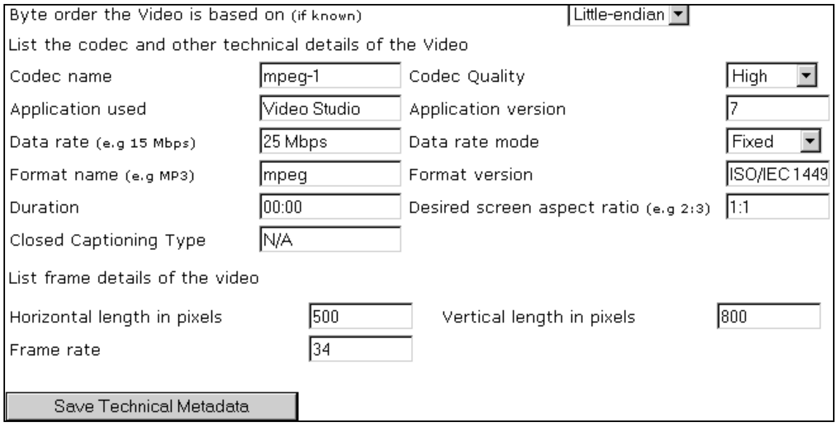


Figure 3: Structure of our Extended METS Preservation Metadata Schema

Although METS is the more widely used preservation metadata schema, MPEG-21DIDL [19] has also been successfully applied to packaging of complex digital objects and has potential as a preservation metadata schema [26]. Hence for comparative purposes, we also developed an MPEG-21 DIDL compliant schema to support the preservation metadata requirements for our composite mixed-media objects. Our comparative research (to be published in the near future) indicated that although METS is well structured, clearly defined and quite easy to use, it is relatively fixed, compared to MPEG-21 DIDL. METS *Profiles* now allow individuals or organizations to specify limitations and restrictions on the creation of METS instance documents in order to define new document classes whilst maintaining interoperability. Because MPEG-21 DIDL provides only a high-level metadata framework, it is more flexible than METS and allows multiple, alternative ways of representing a single set of preservation metadata tags. Again community *profiles* will be required to ensure that this flexibility doesn't lead to ambiguity and loss of interoperability.

Because we believe that there are advantages to both METS and MPEG-21 DIDL, we developed a PREservation Metadata INput Tool (PREMINT) to support both schemas. PREMINT is available as both a stand-alone Java application and a JSP (Java Server Pages) version [35]. The application consists of a set of metadata input forms, constrained by the underlying XML Schema. PREMINT captures metadata by dynamically presenting the user with a serious of forms that collect: Descriptive Metadata, Technical Metadata, Intent Metadata, Presentation Metadata, and Attitude-to-Preservation Metadata.

Figure 4 shows a screenshot of the PREMINT tool, illustrating the Technical Metadata input form for a video file. Currently technical metadata must be manually input but in order to streamline the process we plan to integrate services such as JHOVE, the JSTOR/Harvard Object Validation Environment [28], to automatically extract the format-specific technical metadata.



Byte order the Video is based on (if known) Little-endian

List the codec and other technical details of the Video

Codec name	mpeg-1	Codec Quality	High
Application used	Video Studio	Application version	7
Data rate (e.g 15 Mbps)	25 Mbps	Data rate mode	Fixed
Format name (e.g MP3)	mpeg	Format version	ISO/IEC 1449
Duration	00:00	Desired screen aspect ratio (e.g 2:3)	1:1
Closed Captioning Type	N/A		

List frame details of the video

Horizontal length in pixels	500	Vertical length in pixels	800
Frame rate	34		

Figure 4: Screenshot of the PREMINT Video Technical Metadata Input Form

For specifying the Structure Map (structural relationships) of composite mixed-media objects, SMIL 2.0 [22] provides a simple platform-independent XML-approach. Because of its simplicity, human-readability and platform-independence, it is preferable to application-dependent formats such as *Director*, *Acrobat, Shockwave* or *Flash*. Rather than build our own SMIL editing tool, we invoke an existing SMIL authoring tool, *Fluition*, by *Confluent Technologies*, [27] from within our Java application. Users specify the location of the individual digital objects and their temporal and spatial layout relative to each other by defining regions and attaching the digital objects to them.

When the metadata specification is complete, users have a choice of saving the structural metadata to SMIL or HTML+TIME and the overall metadata output to either METS+Extensions or MPEG-21 DIDL.

6 Obsolescence Detection and Notification

The Obsolescence Detection module periodically compares the preservation (formatting) metadata for each object and sub-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 and viewing software required to access objects in the collections. For each software product, the registry stores: Title, Description, Company, CurrentVersion, ReleaseDate, DeveloperPage, License, Requirements, DownloadSite, DownloadSize, Rating, EaseOfUse, FormatsSupported, Features, Stability, Price.

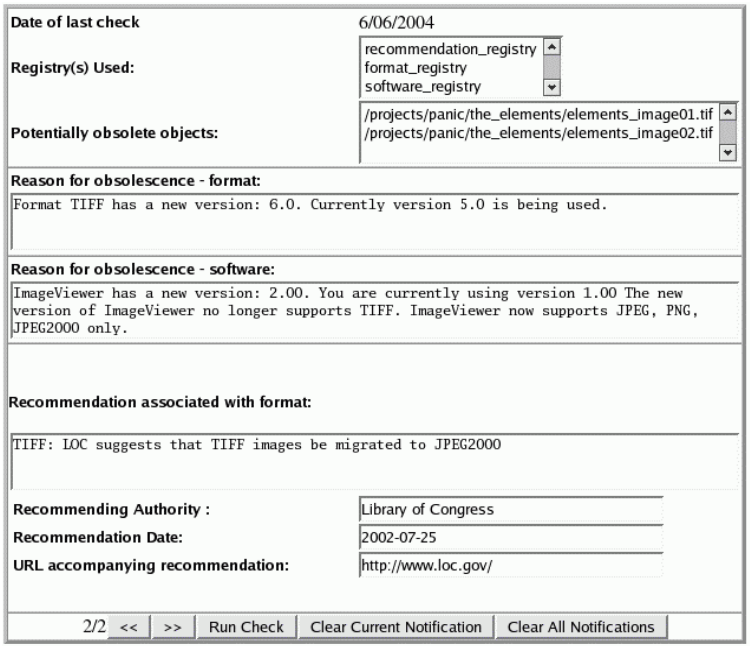
- Format Registry - this is a subset of the GDFR [29]. It contains information about digital formats including: Identifier, Description, Version, Author, Owner, RelationshipsOtherFormats, ApplicationsUsingThisFormat; FormatSpecification; ProvenanceEvents etc.
- Recommended Format Registry – this tracks the latest recommended preservation formats and the authority making the recommendation e.g., "the Research Libraries Group (RLG) recommends JPEG-2000 as the preferred preservation format for images".

For the purposes of demonstrating the PANIC prototype, we have developed three MySQL databases containing sample data, to represent these three registries. However we expect to replace these with corresponding real-world registries under development by existing or evolving initiatives. For example, the Global Digital Format Registry (GDFR) and PRONOM are both developing digital format registries for the purposes of long term preservation. VersionTracker [14] 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.

The system determines when there is an incompatibility between an object's current preservation/formatting metadata and the latest version recorded in the registry, then sends a message to the nominated person(s) or software agent, notifying them of a potential risk. Figure 5 is an example a notification window. In this example, the Obsolescence Detector has determined:

1. From the Software Version Registry, that a number of TIFF images are at risk because the latest version of ImageViewer no longer supports TIFF but does support JPEG-2000.

2. From the Recommended Formats Registry, that the Library of Congress now recommends that TIFF files be migrated to JPEG2000.



Date of last check	6/06/2004
Registry(s) Used:	recommendation_registry format_registry software_registry
Potentially obsolete objects:	/projects/panic/the_elements/elements_image01.tif /projects/panic/the_elements/elements_image02.tif

Reason for obsolescence - format:

Format TIFF has a new version: 6.0. Currently version 5.0 is being used.

Reason for obsolescence - software:

ImageViewer has a new version: 2.00. You are currently using version 1.00 The new version of ImageViewer no longer supports TIFF. ImageViewer now supports JPEG, PNG, JPEG2000 only.

Recommendation associated with format:

TIFF: LOC suggests that TIFF images be migrated to JPEG2000

Recommending Authority :	Library of Congress
Recommendation Date:	2002-07-25
URL accompanying recommendation:	http://www.loc.gov/

Figure 5: Notification Screen listing at-risk objects

7 Preservation Service Description, Discovery and Invocation

After the system has determined that a digital object is at risk and the relevant person has been notified, the next step is to determine what preservation action needs to be taken. In this section we describe the technical components required to enable the dynamic discovery, (composition), selection and invocation of the optimum preservation service for the endangered digital object(s).

7.1 Semantic Web Services

Web services enable networked computer programs to process and consume information. Based on the following open standards, Web services provide a standardized way of enabling Web-based application-to-application interoperability:

- XML (Extensible Markup Language) – is used to define the syntax and structure of the application-to-application messages;
- SOAP (Simple Object Access Protocol) – provides the message format for communicating and invoking Web services;
- Web services Description Language (WSDL) – describes how to access Web services;
- Universal Description, Discovery and Integration (UDDI) - provides a registry that clients can use to discover available services.

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 [30, 31]. 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.

7.2 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. Figure 6 shows the structure of the OWL-S ontology. There are three main subontologies to the top-level Service ontology:

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

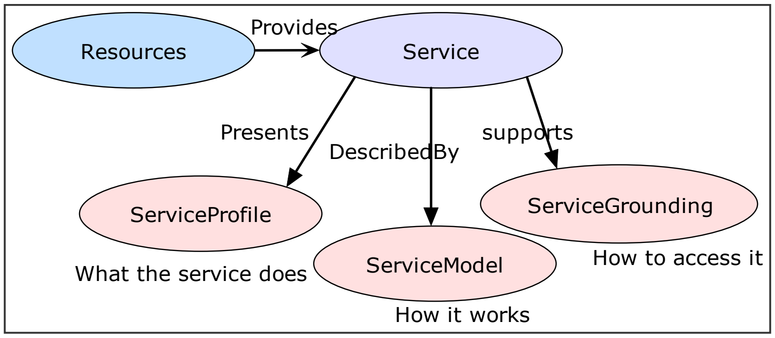


Figure 6: Top level view of the Web Service Ontology

The advantage of OWL-S is that it is very general and can be adapted to describe any Web service. We have extended the OWL-S classes to create more preservation-specific subclasses. Figure 7 illustrates how we have extended the top-level 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 [20]. The *normalization* service is defined as a further subclass of *migration*.

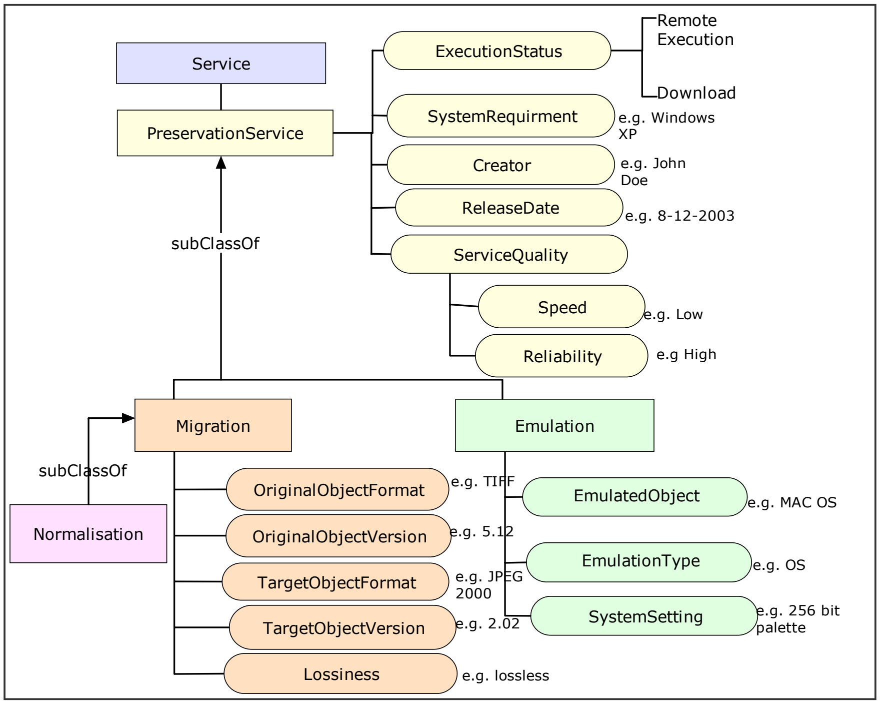
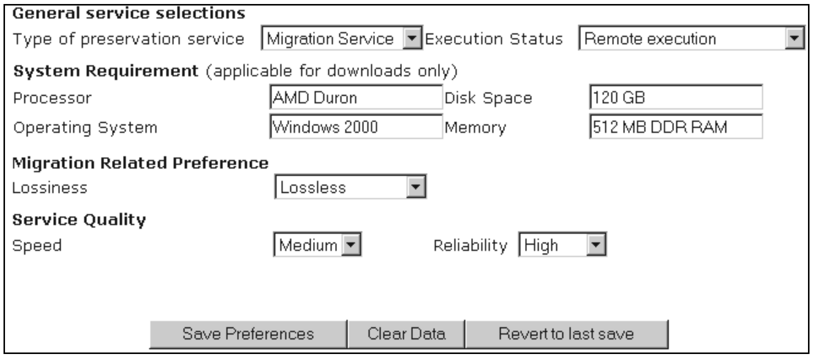


Figure 7: OWL-S Service and ServiceProfile Extensions

Semantic Matchmaker [32] is used as the Discovery Agent in PANIC. When the collections manager specifies the parameters required in the preservation web service, a query is created and submitted to *Semantic Matchmaker*. Requests from the service requesters, are converted to ServiceProfile documents and compared against the stored ServiceProfiles describing available services. A ranked list of matching services is retrieved and displayed. Matching services can be either atomic or chained, composite services. Figure 8 illustrates a Service Specification screen in which the collections manager specifies the details of a required preservation service.



General service selections

Type of preservation service: Migration Service — Execution Status: Remote execution

System Requirement (applicable for downloads only)

Processor: AMD Duron — Disk Space: 120 GB

Operating System: Windows 2000 — Memory: 512 MB DDR RAM

Migration Related Preference

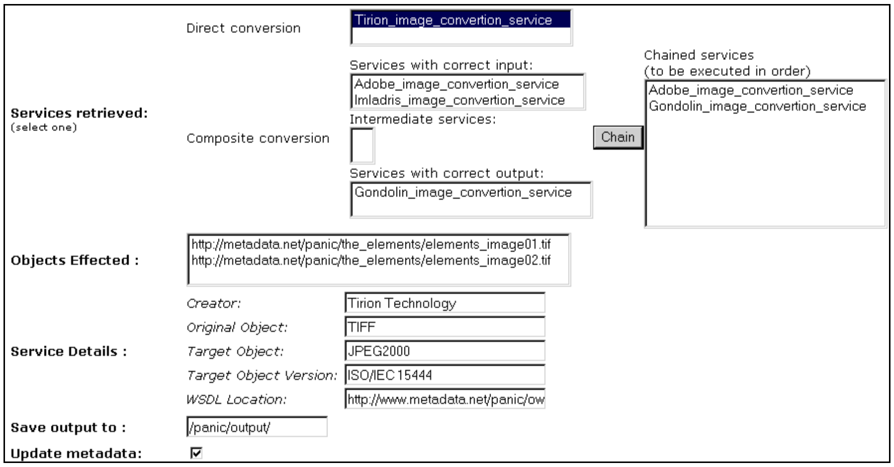
Lossiness: Lossless

Service Quality

Speed: Medium — Reliability: High

Figure 8: Service Specification Screen

For example, a user might specify that she requires a service that converts from TIFF 4.0 to JPEG-2000. She also prefers a distributed converter since she does not want to download a converter to her local machine, which is old, slow and has limited memory. She also specifies that the distributed converter must be highly reliable, high-speed and not result in any loss in image quality. Her 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 [33] were used to advertise available Web services but dynamic discovery was difficult due to the lack of rich semantics. Because we are using OWL-S, we can use *Semantic Matchmaker* which performs more precise and dynamic discovery of appropriate services. In this case, the *Semantic Matchmaker* determines that there are two matches to the service request: one simple process and one composite, chained process (TIFF 4.0 is converted to TIFF 5.0 and then TIFF 5.0 is converted to JPEG-2000). The matching service descriptions are sent back to the *Requester Agent*. These are then used to generate an email and present the search results to the user as shown in Figure 9.



Services retrieved: (select one)

Direct conversion: Tirion_image_convertion_service

Composite conversion:

Services with correct input: Adobe_image_convertion_service
Imladris_image_convertion_service

Intermediate services:

Services with correct output: Gondolin_image_convertion_service

Chained services (to be executed in order): Adobe_image_convertion_service
Gondolin_image_convertion_service

Objects Effected : http://metadata.net/panic/the_elements/elements_image01.tif
http://metadata.net/panic/the_elements/elements_image02.tif

Service Details :

Creator: Tirion Technology

Original Object: TIFF

Target Object: JPEG2000

Target Object Version: ISO/IEC 15444

WSDL Location: http://www.metadata.net/panic/ow

Save output to : /panic/output/

Update metadata: ☑

Figure 9: Service Selection Screen

7.3 Service Selection and Invocation

Given the results of the search and the recommendations of the Discovery Agent, the collections manager can either allow the system to automatically invoke the best matching service or interactively choose a particular preservation action and invoke it manually. The collections manager may need to set certain runtime parameters prior to service execution e.g., where to save the output files, whether to update preservation metadata, where to email the log file etc.

The Requester Agent then sends the inputs (e.g., TIFF files) to the Provider Agent which executes the service and returns the outputs (e.g., JPEG-2000 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 her that the migration of TIFF files has been completed.

8 Evaluation, Future Work and Conclusions

8.1 System Evaluation and Future Work

To date we have only tested the system on a small collection of SMIL objects containing a combination of text, images and video of different formats - but it has proven highly effective in capturing the preservation metadata associated with such objects, in monitoring their accessibility, notifying nominated system users and in facilitating the semi-automatic migration of the contained objects. Based on a comparison of the digital objects' preservation metadata with prototype registries that track the latest software versions and recommended preservation formats, the system can discover and execute the most appropriate migration services. By making the metadata capture tools, the software and format registries and the preservation services available through a Semantic Web Services architecture and software agents, we have significantly reduced the effort required by individual organizations to establish their preservation infrastructure. The manual effort required in monitoring new versions of rendering software or file formats is carried out automatically by the Obsolescence Detection and Notification Agents. The additional effort required in searching the web for the appropriate migration or emulation service is executed by the Discovery and Invocation Agents.

Further testing and evaluation of the system will involve applying it to: larger collections; composite objects containing more diverse formats (e.g., text, Word, web pages, video, audio, multimedia etc.) and requiring a wider variety of preservation services. As more real-world registries and services (such as PRONOM, GDFR, INFORM, TOM and UVC) become available, we plan to incorporate them into the system. The design of the system is such that this will require relatively little effort. We are also planning to install and evaluate the system across a number of distributed digital archives to determine if there are additional requirements or constraints which we need to consider or support. In the next few months we will begin a collaborative project with the National Library of Australia and the UK Digital Curation Centre, that will expand and evaluate components of PANIC by applying it to a number of (FEDORA, DSpace and PANDORA) digital object repositories in Australia.

Finally, to date we have only invoked atomic and simple processes or services. Further work is required to enable the semi-automatic composition of more complex, composite or chained preservation services using a tool such as Web Service Composer [34].

8.2 Conclusions

In this paper we have described the PANIC system - a prototype digital preservation system which we have developed that is based on a combination of preservation metadata capture, software and format registries and Semantic Web Services. We believe that the semi-automated, distributed Web services approach which we describe in this paper is the optimum architecture to provide a viable, cost-effective solution to the long term preservation of large scale collections of complex digital objects. By enabling the automatic detection of potentially obsolescent digital objects and the subsequent discovery and execution of the most appropriate preservation service – the system can potentially save organizations vast amounts of time and effort, as well as prevent the loss of valuable digital assets.

The distributed nature of the proposed Web services architecture offers many advantages. It leverages existing work on preservation metadata capture and preservation services (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 discover and incorporate new preservation services as they become available. Because the system is based on standards including: METS, 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 grows and changes accordingly. As new preservation services, tools, standards and recommendations evolve, they can automatically be incorporated into the system by advertising them via semantic descriptions in the appropriate 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 librarian, archivist or collections manager to select the best single service or combination of services for a particular digital object or a particular set of circumstances. 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 PANIC system which we've described in this paper represents a significant advance towards the design of a sustainable digital preservation system for libraries and archival organizations responsible for maintaining long term access to large digital collections.

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] National Digital Information Infrastructure and Preservation Program, http://www.digitalpreservation.gov/ndiipp/

[2] CEDARS, CURL Exemplars in Digital Archives http://www.leeds.ac.uk/cedars/

[3] CAMiLEON http://www.si.umich.edu/CAMILEON/

[4] PANDORA http://pandora.nla.gov.au/

[5] Networked European Deposits Library (NEDLIB) http://www.kb.nl/coop/nedlib/

[6] PREMIS (PREservation Metadata: Implementation Strategies) Working Group http://www.oclc.org/research/pmwg/

[7] XENA: XML Electronic Normalising of Archives http://xena.sourceforge.net/

[8] Archiving the Avant Garde http://www.bampfa.berkeley.edu/ciao/avant_garde.html

[9] Guggenheim Museum, Variable Media Initiative http://www.variablemedia.net

[10] Virtual Remote Control (VRC) http://irisresearch.library.cornell.edu/VRC/

[11] Andreas 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

[12] GDFR http://hul.harvard.edu/formatregistry/

[13] PRONOM http://www.nationalarchives.gov.uk/pronom/

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

[15] TOM http://tom.library.upenn.edu/

[16] 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

[17] PANIC, http://www.metadata.net/panic/

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

[19] MPEG-21, Information Technology, Multimedia Framework, "Part 2: Digital Item Declaration," *ISO/IEC 21000-2:2003*, March 2003.

[20] The OWL Services Coalition, OWL-S: Semantic Markup for Web services, DAML, 24th of July 2004 http://www.daml.org/services/owl-s/1.1B/owl-s/owl-s.html

[21] J. Hunter, S. Choudhury, "Implementing Preservation Strategies for Complex Multimedia Objects", The Seventh European Conference on Research and Advanced Technology for Digital Libraries, ECDL 2003, Trondheim, Norway, 17th – 22nd August 2003 http://metadata.net/panic/Papers/ECDL2003_paper.pdf

[22] Synchronized Multimedia Integration Language SMIL 2.0, W3C Recommendation 7 August 2001 http://www.w3.org/TR/smil20/

[23] XHTML+SMIL Profile, W3C Note 31 January 2002 http://www.w3.org/TR/XHTMLplusSMIL/

[24] Timed Interactive Multimedia Extensions for HTML (HTML+TIME) Extending SMIL into the Web Browser http://www.w3.org/TR/NOTE-HTMLplusTIME

[25] AV Prototype Project Working Documents, Extension Schemas for the Metadata Encoding and Transmission Standard, Revised February 2003. http://lcweb.loc.gov/rr/mopic/avprot/metsmenu2.html

[26] Bekaert J., Hochstenbach P. and Van de Sompel H., "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

[27] Fluition, Confluent Technologies http://www.confluenttechnologies.com/fluition.html

[28] JHOVE http://hul.harvard.edu/jhove/jhove.html

[29] GDFR, Data Model v.3 http://hul.harvard.edu/gdfr/DataModel_v3.doc

[30] Dan Wu, Bijan Parsia, Evren Sirin, James Hendler and Dana Nau, "Automating DAML-S Web Services Composition Using SHOP2", 2nd International Semantic Web Conference, ISWC 2003, Sanibel Island, Florida, USA, 20th – 23rd October 2003 http://www.mindswap.org/papers/ISWC03-SHOP2.pdf

[31] Tse-Ming Tsai, Han-Kuan Yu, Hsin-Te Shih, Ping-Yao Liao, Ren-Dar Yan, Seng-cho T. Chou, "Ontology-Mediated Integration of Intranet Web Services", Computer No 10, Volume 36, October 2003 http://computer.org

[32] 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, 23rd to 26th June 2003 http://www-2.cs.cmu.edu/~softagents/papers/p2p_icws.pdf

[33] UDDI, http://www.uddi.org/

[34] Web Service Composer, http://www.mindswap.org/~evren/composer/

[35] PREMINT Preservation Metadata Input Tool http://maenad.dstc.edu.au:8080/premint/index.jsp

[36] Erik Oltmans, Raymond J. van Diessen, Hilde van Wijngaarden, "Preservation Functionality in a Digital Archive", Proceedings of the Fourth ACM/IEEE Joint Conference on Digital Libraries, Tucson, Arizona, June 7-11, 2004

[37] PANIC Preservation Architecture for New Media and Interactive Collections http://maenad.dstc.edu.au:8080/webservice/login.jsp