TASK 1
INITIAL SURVEY REPORT

PREPARED UNDER:

TASK ORDER GS-10F-0185P, NAMA-08-F-0016
DIGITAL PRESERVATION BUSINESS CONSULTING SERVICES

FOR:

NATIONAL ARCHIVES AND RECORDS ADMINISTRATION

BY:

SYSTEMS INTEGRATION GROUP, INC.

APRIL 18, 2008

Table of Contents

Initial Market Survey Report of Digital Preservation Projects

The National Archives and Records Administration awarded a Task Order contract to Systems Integration Group, Inc. to provide Digital Preservation Business Consulting Services related to the identification and assessment of emergent information system technologies on behalf of the Electronic Records Archives System (ERA). This market survey is the first of three deliverables, listed as Initial Survey Report. The contract requires that SIG conduct a census and market research to identify digital preservation projects that intend to produce technology, procedures, standards, or knowledge that could be of use in NARA's preservation of electronic records in the ERA system; to identify the types of products, noting assumptions and restraints; and to assess the potential usefulness of each product for NARA.

The information in this report is based on publicly available information. No additional research was conducted, based on the contract stipulation that NARA would indicate what projects should be further researched. A teleconference held at the beginning of the contract with NARA and SIG staff provided further clarification as to the direction this report should follow. Dr. Kenneth Thibodeau, Program Director for ERA, indicated that the initial survey should provide a list of activities that have some promise. Further research will lead to the identification of 3 to 6 projects or organizations that NARA might contact for possible collaboration. The purpose of this Task is to identify procedures, analytic tools, or standards that NARA might apply in implementing the ERA system, and software tools that might be used in the system, for example in the digital preservation framework or search and access framework.

With additional information supplied by ERA staff members, we began our research. In addition, based on Dr. Thibodeau's recommendation to meet with Dr. Mariella Guercio, we solicited recommendations on possible sources of collaboration from Dr. Guercio. We have included the information that Mr. Robert Chadduck and Dr. Guercio provided after our review of publicly available information. Based on comments from NARA on the first draft, we have incorporated additional information from staff of the ERA Research Division, Robert Chadduck and Mark Conrad. In reviewing these projects, a number clearly have been built upon earlier ERA research activities. It is to be anticipated that the projects identified in this paper would welcome continuing collaboration with ERA, which would provide benefits to both communities. It should also be noted that the projects listed are at different stages of development and maturity, and this will influence the timing of possible collaboration.

Much of the current literature about digital preservation focuses on digitization and metadata. Unless we could find specific mention of tools relating directly to preservation of digital materials, we have not included these projects in this survey. There are a number of projects currently underway in the U.S., U.K., Europe, Australia and New Zealand that appear to offer products that might be useful to ERA. Many of these projects have representatives from a variety of countries and professions.

At the end of the report is a table with a summary of ratings for possible collaboration with the projects discussed in this report. A glossary is attached to provide titles for the projects under discussion.

Projects or Tools of the National Archives of the United Kingdom

DROID http://www.nationalarchives.gov.uk/about apps/PRONOM/Tools.htm

DROID (Digital Record Object Identification) is a software tool that the National Archives of the United Kingdom ("TNA") developed to perform automated batch identification of file formats, as part of its broader digital preservation activities. DROID is designed to be able to identify the precise format of all stored digital objects, and to link that identification to a central registry of technical information about that format and its dependencies. DROID uses internal and external signatures to identify and report the specific file format versions of digital files. These signatures are stored in an XML signature file, generated from information recorded in the PRONOM technical registry. New and updated signatures are regularly added to PRONOM, and DROID can be configured to automatically download updated signature files from the PRONOM website via web services. DROID is a platform-independent Java application, and includes a documented, public API for ease of integration with other systems. It can be invoked from two interfaces: Java Swing GUI or a command line interface.

DROID allows files and folders to be selected from a file system for identification. This file list can be saved at any point. DROID can also be used to identify Uniform Resource Identifiers (URIs) and streams (command line interface only). After the identification process has been run, the results can be output in XML, CSV or printer-friendly formats. TNA issued Version 1.1 in August 2006 under an open source license. DROID 1.1 incorporates a number of enhancements, including enhanced signature syntax in PRONOM, enhanced identification algorithm, support for identifying URIs and streams, configuration of proxy server settings from the GUI, and improved XML schemas.

PRONOM: http://www.nationalarchives.gov.uk/preservation/digital.htm

PRONOM is an on-line information system that contains information about data file formats and their supporting software products which can process each format. Information related to these file formats, such as documentation about them, their compression types, character encoding schemes and intellectual property rights, is also included in this system.

The PRONOM Persistent Unique Identifier (PUID) is an extensible scheme for providing persistent, unique and unambiguous identifiers for records in the PRONOM registry. Such identifiers are fundamental to the exchange and management of digital objects, allowing human or automated user agents to unambiguously identify, and share that identification of, the representation information required to support access to an object. This is a virtue both of the inherent uniqueness of the identifier and of its binding to a definitive description of the representation information in a registry such as PRONOM. PRONOM is seen as a precursor and model for the proposed Global Digital Format Registry (GDFR). While PRONOM had the advantage of being available now, if the GDFR initiative bears fruit, it could provide coverage of a much greater range of formats. NARA should study the trade off between waiting for the GDFR and possibly implementing PRONOM before GDFR is fully developed.

DROID and PRONOM offer functionality that potentially could be useful to ERA with their ability to identify files automatically, thus improving processing time for ingest. However, DROID's functionality replicates that of the PERPOS format recognition tool and JHOVE.

Lockheed Martin is including both PRONOM and DROID in the initial ERA system. NARA should consider working with TNA to improve format recognition capabilities, possibly building a combination of DROID and the PERPOS tool. Since both NARA and TNA are still developing these tools, shared knowledge will increase the effectiveness of both sets of tools over time. There are issues that will need to be addressed with the TNA tools. One consideration could be scalability of both DROID and PRONOM. Additionally, NARA needs to determine if the metadata collected by PRONOM meets the requirements demanded for U.S. federal records. Again these issues can be resolved over time by collaboration.

INSPECT: http://www.significantproperties.org.uk/

TNA is also involved with **INSPECT** (**Investigating the Significant Properties of Electronic Content Over Time).** The purpose of INSPECT is to develop and expound the concept of "significant properties" (or characteristics) of digital objects and to analyze a range of digital objects in order to develop a generalized methodology for determining the significant properties of digital object types. According to the project, significant properties are those aspects of the digital object which must be preserved over time in order for the digital object to remain accessible and meaningful. The various properties of electronic records may be categorized as arrangement, content, context, structure, presentation (e.g., layout, color), and behavior (e.g., interaction, functionality). Deciding which aspects of each of these categories must be preserved over time is essential to proper preservation planning. This is a necessary investigation which has not yet been undertaken. There is a critical need to initiate this work in order to establish best practice approaches to preserving digital objects. The project is led by the Arts and Humanities Data Service (AHDS) in association with TNA.

Preserving "essential characteristics" is a key component of the ERA system. NARA staff have spent considerable effort on developing the concept of record templates as structures for capturing information about them and applying this information in preservation processes. In addition NARA staff are devoting many hours to defining essential characteristics for specific types of records that are expected to be accessioned by ERA. Accordingly, it would be useful for NARA to compare the findings of INSPECT with the work completed by ERA.

Other European Projects of Interest

PLANETS: http://www.planets-project.eu/

Planets is a European project that is developing tools associated with digital preservation. The Planets project brings together European National Libraries and Archives, leading research institutions, and technology companies to address the challenge of preserving access to digital cultural and scientific knowledge. Co-financed by the European Commission, Planets seeks to develop services for preservation planning, enabling organizations to define, evaluate, and execute preservation plans. It is developing methodologies, tools and services for the characterization of digital objects and for the support of preservation actions. In addition, it seeks to establish a preservation test bed to provide a consistent and coherent evidence base for the objective evaluation of different preservation protocols, tools and services and for the validation of preservation plans. The implementation of an interoperability framework seamlessly integrates the subprojects within a distributed service network.

One of the first projects undertaken by this project was the comparison of current digital preservation activities. The authors found that few of the current conversion tools have been designed for digital preservation. The authors who analyzed this issue also pointed out the common problem of a dearth of tools for automated Quality Assurance. These findings led to the development of characterization tools.

The characterization of files might be of value to ERA. The work thus far has been to:

1) Develop a "language" that defines file format: XCEL-eXtensible Characterisation Extraction Language. XCEL is designed to be able to allow the expression of all existing file formats.

2) Produce an "extractor" program, which uses a specification to extract the data described by the format, expressed in XCEL, from a file.

3) Provide a generalized model of information contained within files, by providing a language which expresses the content of a file: XCDL-eXtensible Characterisation Definition Language.

4) Develop a software "comparator" able to make a meaningful numerical estimate whether two files contain the same information.

By September 2007, the initial specifications for the eXtensible Characterisation Description and Extraction Language had been completed. A prototype property extraction tool has been developed and is currently being tested. The first iteration of the characterization registry has been completed, based on the PRONOM registry. A characterization framework, which allows the automated deployment of characterization tools through the registry, has been completed, and several of the existing tools, including DROID and JHOVE, have been integrated with this framework.

ERA Research has actively participated in the Open Grid Forum which is seeking is to define an XML-based language, the Data Format Description Language (DFDL), for describing the structure of binary and character encoded (ASCII/Unicode) files and data streams so that their format, structure, and metadata can be exposed. One of the research partners for ERA Research, the National Center for Supercomputing Applications, has been working closely with the University of Cologne. The characterization software tools planned for use by Planets were developed by the University of Cologne. Furthermore, the ERA Research program has funded PERPOS development, which automatically recognizes document types. NARA should seek continued collaboration with Planets to build on the work already begun by the research partners.

In 2008, Planets expects to develop preservation planning tools (PLATO), including decision support and risk assessment modules; integrated preservation planning services, including an automated collection profiling service, a technology watch service, and an advice service; a description language for preservation actions tools; Planets-compliant migration tools for digital objects; emulation tools for specific environments; and final specifications of a characterization description and extraction language. NARA has developed similar tools with the ERA Lifecycle Management Plans. By collaborating with Planets the tools can be compared and improved for both projects.

Finally, the automated profiling service, migration tools and emulation tools might also be useful. NARA has worked closely with two active members: Dr. Reagan Moore and Hans Hofman, enabling this collaboration to build on past activities.

CASPAR: http://www.casparpreserves.eu/caspar-project

The CASPAR project (**Cultural, Artistic, and Scientific Knowledge for Preservation, Access and Retrieval**) is co-financed by the European Commission. The project has four goals: to build a pioneering preservation environment, based on the full use of Open Archival Information Standard (OAIS), to demonstrate its ability to handle the preservation of digital resources of multiple user communities, to advance the current state of the art in digital preservation, and to develop technological solutions supporting the emergence of an offer of systems and services for preservation of digital resources.

The project intends to implement, extend, and validate the OAIS reference model, to enhance techniques for capturing representation information and other preservation-related information for content objects, and to design virtualization services for abstracting from underlying computing and storage environment. Furthermore, CASPAR seeks to integrate services for digital rights management, authentication, and accreditation, to conduct research into more sophisticated access to and use of preserved digital resources such as intuitive query and browsing mechanisms, and to develop case studies for validating its approach across different user communities.

In reviewing the documents posted on their website, we found their methods for development, such as Use Case Scenarios, very similar to ERA development. One document explicitly noted that the questionnaires used to solicit information about possible test beds were drawn from InterPARES and ERPANET. Since the emphasis is building digital preservation services based on OAIS, it would appear that ERA could gain much from this parallel work. Of particular interest might be the packaging and authenticity tools that CASPAR is to develop. We recommend that NARA look at the Preservation Orchestration Manager Use Case, the Representation Information Tool Use Case, and the Digital Rights Manager Use Case to determine if some of the ideas embodied therein might be useful in ERA. Finally, grid technology and open source software are part of the architecture for CASPAR. Both CASPAR and Planets are using the Storage Resource Broker and iRODS(***integrated*** **R**ule**O**riented **D**ata **S**ystems) developed by the San Diego Supercomputer Center with ERA Research funds. There are many shared issues that could benefit from collaboration.

CRIB: http://crib.dsi.uminho.pt/

CRiB (Conversion and Recommendation of Digital Object Formats) is a Service Oriented Architecture (SOA) designed to assist cultural heritage institutions in the implementation of migration-based preservation interventions. The CRiB system assesses the quality of distinct conversion applications or services to produce recommendations of optimal migration strategies. The recommendations produced by the system take into account the specific preservation requirements of each institution. The CRiB is currently being supported by Web services technology and is capable of carrying out the following activities: recommending optimal migration alternatives that take into consideration the preservation requirements of the institution; converting digital objects to up-to-date encodings that most users will be capable of interpreting; evaluating migration's outcome by comparing the original digital object with its converted counterparts and identifying the significant properties that have not been correctly preserved; and generating migration reports to include in the preservation metadata of migrated objects

Since ERA must deal with a variety of formats in preserving the electronic records transferred to NARA, this project appears to offer a superior method for preservation actions. ERA should seek to establish a close relationship with this group as they continue to develop the services described on their website. Although CRiB is a PhD project which is near completion, Planets has adopted CRiB’s migration services. There is also possible synergy between other aspects of CRiB and projects, such as PRONOM, DROID, GDFR and Planets.

DELOS

DELOS is the digital libraries network of excellence. While DELOS is conducting a joint program of activities aimed at integrating and coordinating the ongoing research efforts of the major European teams working in Digital Library-related areas, the emphasis is not on digital preservation. Its main objective and goal is to develop the next generation of Digital Library technologies, based on sound comprehensive theories and frameworks for the life-cycle of Digital Library information. DELOS is funded only until the end of 2008. The DELOS team is currently working on identifying sustainability measures, but there is no question that the long-term availability of the expertise and resources contained within this project is in jeopardy.

DPE: http://www.digitalpreservationeurope.eu

DPE (Digital Preservation Europe) fosters collaboration and synergies between the many existing national initiatives within Europe. DPE addresses the need to improve coordination, cooperation and consistency in current activities to ensure effective preservation of digital materials. Although DPE is not sponsoring specific research and development, it monitors the various activities and is a good resource on explaining current activities. This project provides access to the current European trends. NARA should monitor this project as a means of keeping abreast of European developments and possibly identify new candidates for knowledge or technology transfer.

Kopal: http://kopal.langzeitarchivierung.de/

Kopal, a three-year project, has as its objective the practical testing and implementation of a long-term preservation system for digital information, developed and operated cooperatively. The project started in July 2004 and is funded by the German Federal Ministry of Education and Research (BMBF). The project partners, Die Deutsche Bibliothek (DDB), the Göttingen State and University Library (SUB), IBM Deutschland GmbH and the Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG) will implement a cooperatively run and reusable solution for the long-term archiving of digital data, hosted at the scientific computer center GWDG in Göttingen. The technical realization will be based on DIAS (Digital Information and Archival System), jointly devised by IBM and the Koninklijke Bibliothek in The Hague, the National Library of the Netherlands. The customization and further development of DIAS into a cooperatively run system will be administered by IBM. The software developed by the project partners DDB and the SUB Göttingen for the import and export of data will be released under an open-source license. The koLibRI software will be a part of the Shaman testbed (see Shaman).

ERA might want to examine any free and open source software developed in this project to see if it has potential use, particularly in the preservation of scientific records.

LIFE http://www.life.ac.uk/

LIFE (Lifecycle Information for E-Literature) is a British Library project that models the digital lifecycle and calculates the costs of preserving digital materials, particularly publications, allowing organizations to understand costs and focus resources where most needed. Like the **DRAMBORA (Digital Repository Audit Method Based on Risk Assessment)** tool, that characterizes digital curation as a risk management activity and provides a metric and a methodology for repository self-assessments, it is an interesting project that does not have direct value for the ERA Project.

LOTAR: http://www.prostep.org/en/projektgruppen/lotar/

LOTAR (Long Term Archiving and Retrieval in the Aerospace Industry) is a project that is focused on preserving current digital information for long periods of time. It comes out of the aeronautic industry. This project was suggested to us by Mr. Robert Chadduck. LOTAR is focusing on digital technical product documentation, such as 3D-CAD and PDM data. This is an ASD-STAN and ProSTEP IVIP association supported development and standardization project driven by Airbus, EADS-Military Aircraft, MTU Aeroengines, Alenia Aeronautics and BAE Systems. The project began in December 2001 and will run through December 2009. The objective of the project group is to establish methods, processes and a data model for archiving 3D geometry data and product structure information.

ERA Research is a member of PDES, which is working closely with LOTAR in reviewing and offering comments about engineering records, based on ERA Research work with the University of West Virginia, the Department of Energy and the Naval Sea Systems Command, for the preservation of ship records. LOTAR seeks to establish the results as a standard for the European aerospace industry within the EN9300 series. Since PDES is playing an active role, it is to be anticipated that an international standard will develop from this effort. We recommend that ERA follow this work closely, since these types of records offer significant challenges, and the project could take advantage of the expertise examining these issues

Nestor http://www.langzeitarchivierung.de/index.php?newlang=eng

Nestor, the German Network of Expertise in long term storage of digital information, began in June 2003 as a cooperative effort between six partners led by the German National Library (Die Deutsche Bibliothek). The goal is a permanent distributed infrastructure primarily for long term accessibility of digital resources and less so for pure preservation aspects, comparable to the Digital Preservation Coalition in the United Kingdom. Nestor will provide a broad communication platform for all interested institutions as well as various guidelines for standardized digital preservation policies and coordinated activities to be utilized mostly within Germany and possibly to be extended through international cooperation

ERA Research is participating in a working group sponsored by the Consultative Committee for Space Science Data Systems, which developed the OAIS standard, to look at the Trusted Repository Audit and Checklist (TRAC) to see if this can be used for engineering drawings. Some of this work will be coordinated with the NESTOR and DRAMBORA projects with the goal of taking the requirements of the TRAC document and making it an ISO standard. Mark Conrad, a member of the ERA Research staff is the NARA representative in this working group. Clearly, ERA will want to monitor the developments of this project, since it could lead to an establishment of a standard.

PARADIGM: http://www.paradigm.ac.uk/index.html

PARADIGM (The Personal Archives Accessible in Digital Media) was a project that involved Bodleian Library, University of Oxford, and John Rylands University Library, University of Manchester. The project focused on personal records, working with politicians, archivists and researchers to investigate the challenges of preserving personal archives of born digital materials. This project does not appear to offer information that could be of value to ERA.

PRESTO: http://presto.joanneum.ac.at/index.asp

The **PRESTO (Preservation Technology for Broadcast Archives)** project was one of the European Union Information Society Technology (IST) sponsored initiatives that looked at a range of preservation issues and prospective outcomes for digital and analog film, audio and video archival materials. The initial partners were several major European public broadcast archives and commercial research partners. The aims of the project were to evaluate the preservation status of collected audio-visual materials, to establish what resources and technologies were utilized or were required for cost-effective preservation, to develop a preservation process chain of automated copying and playback with the necessary quality control and management of metadata, and to evaluate and test the new methodologies. The project ran from late 2000 until mid 2002. The information gathered from this project has hopefully been absorbed by either PLANETS or CASPAR.

Project StORe

Project StORe was conceived as an initiative to apply digital library technologies in the creation of new value for published research. Its primary objective was the design of middleware to enable bi-directional links between source repositories containing research data and output repositories containing research publications. The one research paper discussing the project points out the difficulties in working with the wide range of designated communities.

Sherpa DP2: http://www.sherpadp.org.uk/

Sherpa DP2 will extend the collaborative, shared preservation environment developed by the Sherpa DP project. The project will build on that work by extending the implementation model to interact with repositories holding different and varied types of digital content and by using a more diverse range of content management systems. A tool for creating Metadata Encoding and Transmission Standard (METS) packages will be developed, and automation of digital object preservation will be investigated. The focus is on digital library content, not electronic records. The automation of digital object preservation might be of some interest, so this project should be monitored.

Shaman http://www.d-nb.de/eng/wir/projekte/shaman.htm

Shaman is a multiple partner/multinational source preservation project that received EU funding this past summer. Its goal is to develop a prototype for the grid-based system for archival preservation by investigating and developing a long-term next generation digital preservation framework. The project builds on earlier work in the development of Multivalent Document software, which offers potential for on-demand preservation. The project will create a framework and application development environment, and will develop a test-bed which will employ real-life use scenarios. The koLibRI software, developed by KOPAL, will play a crucial role for the creation of the testbed. In addition, the SHAMAN project will deliver infrastructure and services which will improve access and reuse of this content.

ERA Research has worked with Dr. Paul Watry at the University of Liverpool Library, who is the overall project coordinator, in the work of building scalable grid systems by funding NSF research at Berkeley and Liverpool. This project offers great promise, building upon previous research. ERA should seek close collaboration with this group.

In a meeting with Dr. Guercio, she mentioned several possible sources for collaboration. Her first recommendation was that we contact Mr. Luigi Briguglio of Engineering, one of the largest national IT companies in Italy, who is a member of the CASPAR project. *3D Informatics* is focused on transforming software into open source. Dr. Guercio is working with them in developing a recordkeeping system for the University of Urbino, which she will use as a test bed for InterPARES 3 in managing email at a university. Mr. Franco Bazzigotti is using diplomatics to develop typologies of documents. The region Emilia Romagna is seeking software to manage the various electronic documents created by the region. The software company *UNIMATICA* (Ms. Rosella Bonora) is hoping to develop this software. Another company responsible for much of the financial documentation is *InfoCamere*, and they are interested in working with open source software. Finally, *Consorzio Cilea* in Milan has been involved with scientific records, has experience both with Open Archives and GRID, and is well versed in XML.

Projects of the National Archives of Australia

Although most of our focus has been on European projects, the survey would be inadequate if activities at the National Archives of Australia, the National Library of Australia, and the Provincial Records Office of Victoria were not included in this survey.

XENA: http://www.naa.gov.au/records-management/secure-and-store/e-preservation/at-naa/software.aspx#section1

XENA (XML Electronic Normalising of Archives) is a project that has been under development at the National Archives of Australia (NAA) for at least four years. In many respects the goals of the NAA mirror what ERA seeks to accomplish. XENA is evolving to keep abreast of changes in information technology and the dynamic nature of digital recordkeeping by undertaking ongoing research and development. The focus of the NAA research is in the following areas:

- Process. The digital preservation process must evolve in line with changes in the Archives' own business processes and developments in digital records creation, management and preservation.

- Infrastructure. Regardless of what the Archives does to minimize changes required in infrastructure, the prototype itself will, over time, become obsolete and require change. In part this will be driven by the Archives' need to scale its operations in line with the storage capacity required to maintain digital records created by government agencies.

- Software. The preservation software must not only evolve to meet changes in the process and infrastructure but also keep up with changes in the digital recordkeeping environment. Currently, the Archives converts (normalizes) office documents, emails, images and some other files into open file formats, but there are many more digital formats, and more will evolve in the future, that will have to be normalized and preserved. Xena's plug-in architecture enables the software to be readily enhanced to meet this challenge. While its digital preservation approach is already in operation as a prototype, the National Archives continues to plan and prepare for a future when the volume of records will require a more sophisticated system to handle significantly larger operational requirements.

XENA currently normalizes the following file types to the specified open format. If asked to normalize a file format not yet supported, the process will fall back to binary normalization:

> **Archives and Compressed Files.** GZIP, JAR, MAC Binary, TAR, TAR.GZ, WAR, ZIP; Audio: AIFF,FLAC, MP3 and WAV. All are converted to FLAC, which is an open format.
>
> **Databases:** SQL files are processed as plain text wrapped in XML.

Documents: CSV/TSV; DCO/PPS/PPT/XLS-Microsoft Office documents are converted to the Open Document Format (ODF); HTML; ODS/ODP/ODT-Open Document Files are preserved as they are; RTF is converted to Open Document Format; SYLK-this spreadsheet format is converted to ODF; SXC/SXI/SXW-StarOffice formats are converted to the newer ODF; TXT is stored in plain text wrapped in XML; WPD is converted to ODF; WRI-Microsoft Write files are converted to ODF; and XHTL is stored in native XHTML format.

Email: MBOX Mailboxes are converted to individual XML files, and a XENA index file is created which will display the files in a table when opened with the XENA viewer; PST files are converted to XML files.

Mr. Mark Conrad, a member of the Research Division of ERA, performed an assessment of XENA v. 4.0.1 in the Research Lab on 24 October 2007. In his analysis, he noted many improvements from earlier versions. However, he noted that as it is currently, it would not work at scale because it does not maintain hierarchies, and essential characteristics might be lost. Yet, there are some functionalities that could be of possible use to NARA. Since ERA is building the infrastructure but does not yet have tools, and NAA is seeking to move from prototype to a more robust system, this is an opportunity for collaboration.

The National Library of Australia: http://www.nla.gov.au/preserve/digipres/tools.html

The National Library of Australia has undertaken a pragmatic but principles-based approach to managing and preserving its digital collections. The Library has shown a readiness to find and, if necessary, develop tools and practices to address the challenges suggested by the principles involved in keeping its digital collections accessible.

The tools and procedures produced by such an approach are often speculative, but because they are based on an intelligent analysis of principles, and because they are tested in managing actual collection management problems, they often work quite well. In turn, they provide a platform of experience that the Library reflects on and uses in developing further tools and practices.

The Digital Object Storage System (DOSS) is the primary arrangement for storing and refreshing data, while the Digital Collection Manager (DCM) is being developed as the metadata repository and management system that can manage a range of automated preservation processes, as well as other needs. Other important components of the Library’s growing suite of tools, procedures and infrastructure supporting digital preservation address the issues of creation, preservation, initial processing, naming and description, repository and management frameworks to support preservation, data security and authenticity, and ongoing access.

Since the emphasis is on digital library materials, ERA should look for commonalities, but focus more on the archival system developed by NAA.

Victorian Electronic Records Strategy(VERS): http://www.prov.vic.gov.au/vers/vers/

VERS has been developed by the Public Record Office Victoria (PROV) to provide leadership and direction in the management of digital records. VERS has created a framework that deals with the problem of capturing, managing and preserving electronic records. This framework includes standards, guidance, training, consultancy and implementation projects, which is centered around the goal of reliably and authentically archiving electronic records. The VERS project began in 1995 with the report "Keeping Electronic Records Forever," which was the starting point to an ongoing collaborative effort between the Victorian State Government, industry and academia to find a practical way to deal with digital records.

A key software tool is the VERS Encapsulated Object (VEO) format, which provides the ability to capture, manage and discover digital records. The Public Record Office Victoria has created the Centre of Excellence, which is responsible for overseeing the rollout of VERS across the Victorian government. The Centre provides resources, advice, and guidance to Victorian government agencies as well as conducting further research into the long term preservation of electronic records and overseeing the construction of an electronic records repository at PROV (Digital Archive).

ERA could benefit from the Lessons Learned from this project, since it is now in an implementation stage. The software could also be of interest.

Projects Supported by ERA

PERPOS

ERA Research has funded projects with Georgia Tech Research Institute to work with the Bush Presidential Library's electronic records to analyze and design software tools that support the accessioning, preservation, arrangement, review, and description of electronic records. The work is evaluating experimental and commercial natural language processing (NLP) search and retrieval tools for use in reviewing Freedom of Information Act (FOIA) exceptions, reviewing Privacy Act and Presidential Records Act restrictions, and responding to routine reference and FOIA records requests. Dr. William Underwood has developed software to recognize document types created by office applications found in the records from the Bush Presidential Library.

One of the key ERA funded projects is Integrated Rule-Oriented Data System (iRODS), which grew out of earlier work with Storage Resource Broker (SRB) technology developed by San Diego Supercomputer Center. iRODS enables users to handle the full range of distributed data management needs, which includes extracting descriptive metadata and managing data to moving it efficiently, sharing data securely with collaborators, publishing it in digital libraries, and ensuring long-term preservation. The most powerful new feature is the “rule engine” that lets users easily accomplish complex data management tasks. Users can automate enforcement, or “virtualize” data management policies by applying rules that control the execution of all data access and manipulation operations. Rather than having to hard code these actions or workflows into the software, the user-friendly rules let any group easily customize the iRODS system for their specific data management needs.

Another ERA Research partner, The National Center for Supercomputing Applications (NCSA) has been gathering and analyzing information about decision processes using geospatial electronic records. The goal is to understand the cost of information preservation and the value of the preserved information. A team from NCSA has been working with NARA to provide software tools for simulating complicated high-confidence decision scenarios using geospatial electronic records and preserving the gathered information in temporally sustainable data containers and reconstructing high-assurance decision making processes.

American Projects Other Than ERA

Digital Library Projects

There are efforts to develop mechanisms for preservation such as LOCKSS, Fedora, and DSpace, which focus more on digital libraries rather than archival records. LOCKKS (Lots of Copies Keeps Stuff Safe) provides tools which use local, library controlled computers to safeguard readers' long-term access to web based journals. The system ensures that hyperlinks to material it is safeguarding continue to resolve and deliver the appropriate content to their readers even if in the Internet at large the links no longer resolve and the content is no longer available. Fedora uses open source software to give organizations a flexible service-oriented architecture for managing and delivering their digital content. DSpace has as its focus capturing data for reuse in any format – in text, video, audio, and data. It distributes it over the web. It indexes the work, so users can search and retrieve specific items. The goal is to preserves digital work over the long term. DSpace provides a way to manage research materials and publications in a professionally maintained repository to give them greater visibility and accessibility over time.

NDIIPP

The Library of Congress with funding from Congress established a collaborative project, called the National Digital Information Infrastructure and Preservation Program (NDIIPP). The goal is to develop a national strategy to collect, archive and preserve the increasing amount of digital content, especially those materials created only in digital formats, for current and future generations. NDIIPP is working with partners from universities, libraries, archives, federal agencies and commercial content and technology organizations. Much of the funding is towards developing possible tools to preserve and provide access to born digital materials.

PeDALS (Persistent Digital Archives on Library System is one of the projects funded by NDIIPP. Led by Arizona State Library and Archives, this project seeks to develop a curatorial rationale to support an automated, integrated workflow process and to implement “digital stacks” using an inexpensive, storage network that can preserve the authenticity and integrity of the collections. Minnesota Historical Society, “A Model Technological and Social Architecture for the Preservation of State Government Digital Information,” is another NDIIPP-funded project that will work with legislatures in several states to explore enhanced access to legislative digital records. This will involve implementing a trustworthy information management system and testing the capacity of different states to adopt the system for their own use. Content will include bills, committee reports, floor proceedings and other legislative materials. States working in this project are Minnesota, California, Kansas, Tennessee, Mississippi, Illinois and Vermont. A third project led by North Carolina Center for Geographic Information and Analysis, “Multistate Geospatial Content Transfer and Archival Demonstration,” will focus on replicating large volumes of geospatial data among several states to promote preservation and access. The project will work closely with federal, state and local governments to implement a geographically dispersed content exchange network. Content will include state and local geospatial data. States working in this project are North Carolina, Utah and Kentucky.

Chronopolis

Chronopolis is a collaboration among the San Diego Supercomputing Center (SDSC), NCAR/CISL, the University of California Library System, and the University of Maryland Institute for Advanced Computer Study (UMIACS). It is aimed at developing national-scale digital preservation infrastructure that has the potential to serve the broad science and engineering community. This new effort encompasses studying viable models and effective systems that facilitate establishing standard reference datasets, preserving collections that evolve over time, and establishing preservation resources "of last resort" for digital assets that might become lost. Digital collections that must persist for 100 or more years are one important focus of this activity.

This project builds upon work developed by ERA Research, the Transcontinental Persistent Archives Prototype (T-PAP) with SDSC and UMIACS. NARA should monitor this project as it moves forward with its research.

DAITSS: http://www.fcla.edu/digitalArchive/daInfo.htm

DAITSS (Dark Archive in the Sunshine State) began in 2000 when the libraries of the state university system of Florida began looking for a long-term preservation solution for digital dissertations and master files from digitization projects. A survey of available content management, “digital library,” and institutional repository software indicated that no existing product actually performed active preservation strategies such as format migration. The decision was made to write repository software locally, and planning for DAITSS development began.

The Florida Center for Library Automation (FCLA) was awarded a three-year grant in 2002 from the Institute of Museum and Library Services that allowed FCLA to add a full time formats specialist to the DAITSS development team. In November 2005, an early version of the DAITSS application that lacked dissemination and withdrawal functions went into production for the Florida Digital Archive. In December 2006, the first version of DAITSS with all major planned functionality was completed. The software was released as open source under the GPL license in 2007.

Plans for DAITSS 2.0 include moving to web service architecture; making the transformation logic rule (table) driven, not hard coded in data file class; using transformation service that reformats based on rules (agnostic to migration, normalization, etc.); and recording more detailed provenance information by describing software as PREMIS agents. We recommend that NARA monitor this project closely to determine if the software under development could be incorporated into ERA functionality.

ERA Research has funded work that has demonstrated the capability to perform the functions that DAITSS is currently developing. It would make sense to form a collaboration so that improvements can be encouraged and thus incorporated into the ERA system.

JHOVE: http://hul.harvard.edu/jhove/index.html

JHOVE (JSTOR/Harvard Object Validation Environment) is a project to automate the identification and validation of file formats. This process is based on the file itself rather than the filename extension. **JHOVE2** seeks to retain existing JHOVE functionality and refactor the existing architecture, support enhancements, increase the range of supported formats, and develop modules for supporting key preservation processes. Since Harvard is a key member of the Global Digital Format Registry, we assume NARA has access to JHOVE materials.

Global Digital Format Registry : http://hul.harvard.edu/gdfr/

The GDFR initiative led by the Harvard University Library (HUL) aims to develop an architecture to support a distributed global registry for file format information. Once deployed, the GDFR will provide services for the centrally-organized collection of format representation information, and the distributed storage, discovery, and delivery of that information. The architecture and software for the GDFR is being developed by Harvard University Library with support from the Andrew W. Mellon Foundation.

Table 1 Summary of Potentially Useful Preservation Products

Function	Product	Status	Sponsor
Data type characterization	GDFR	In development	HUL
	PRONOM	Developed, implemented and being improved; Available as free and open source software (FOSS)	TNA
Data type recognition & validation	DROID	Developed, implemented and being improved; Available as FOSS	TNA
	JHOVE	Developed, implemented and being improved; Available as FOSS	HUL
	PERPOS format recognition tool	Developed at Georgia Tech Research Institute under ERA funding. Implemented with test data	ERA Research
Characterizing electronic records	PERPOS document characterization tools	Implemented in Georgia Tech Research Institute ERA research project	ERA Research
	XCDL	Specified	PLANETS
	Representation Information Tool	In analysis	CASPAR
Defining properties that must be preserved	INSPECT		TNA
	Record Templates	Described	NARA
Preservation Planning	PLATO	In development	PLANETS
	Lifecycle Management Plans	To be implemented partially in initial ERA system	ERA
	CRiB	In development	CRiB
	DAITSS 2.0 preservation format decision tool	Available	FCLA
	LOC preservation format decision tool	Available	LOC
Preservation Management	Lifecycle Management Plans	To be developed in later ERA increments	ERA
	Preservation Framework	Architectural concept	ERA
	CRiB		CRiB
	automation of digital object preservation		Sherpa DP2
	Digital Collection Manager		NLA
	DAITSS 2.0		FCLA

Preservation Services	CASPAR authentication tool	In analysis	CASPAR
	CASPAR Preservation Orchestration Manager	In analysis	CASPAR
	CRiB migration service	In development	CRiB
	LOTAR 3D process and data model for geometry data and product structure information		LOTAR
	tool for creating Metadata Encoding and Transmission Standard (METS) packages	Planned	SHERPA DP2
	MVD tool		Shaman
	XENA normalization tools	Developed, implemented and being improved	NAA
	VERS Encapsulated Object	Available	PROV
	DAITSS 2.0 transformation tool		FCLA
Access Services			CASPAR

Project Rating Table

The table shown below summarizes the findings of this market survey. The rating scale is to be interpreted as follows:

3 The project shows high potential for continuing collaboration should these projects continue

2 The project shows moderate potential for continuing collaboration with ERA

1 The projects shows little potential for continuing collaboration, but should be monitored

None The project has nothing specific to offer ERA.

Name of Project/Software Product	Rating
The National Archives DROID, PRONOM	3
The National Archives, INSPECT	3
Planets	3
CASPAR	3
CRIB	3
DELOS	1
DPE	2
Kopal	1
LIFE	
DRAMBORA	
LOTAR	3
Nestor	2
PARADIGM	
PRESTO	
Project StORe	
Shaman	3
Sherpa DP2	1
National Archives of Australia XENA	3
National Library of Australia	2
VERS	2
Chronopolis	1
FCLA DAITSS	3
JHOVE	3

Glossary of Acronyms

CASPAR	Cultural, Artistic, and Scientific Knowledge for Preservation, Access and Retrieval
DAITSS	Dark Archive in the Sunshine State
DELOS	Network of Excellence in Digital Libraries
DPE	Digital Preservation Europe
DRAMBORA	Digital Repository Audit Method Based on Risk Assessment
DROID	Digital Record Object Identification
INSPECT	Investigating the Significant Properties of Electronic Content Over Time
JHOVE	JSTOR/Harvard Object Validation Environment
LIFE	Lifecycle Information for E-Literature
LOTAR	Long Term Archiving and Retrieval in the Aerospace Industry
METS	Metadata Encoding and Transmission Standard
NARA	National Archives and Records Administration (U.S.)
NDIIPP	National Digital Information Infrastructure and Preservation Program
NESTOR	Network of Expertise in Long-Term Storage of Digital Resources
OAIS	Open Archival Information Standard
PARADIGM	The Personal Archives Accessible in Digital Media
PLANETS	Preservation and Long-term Access through Networked Services
PRESTO	Preservation Technology for Broadcast Archives
PRONOM	Online technical registry developed by TNA
PUID	Persistent Unique Identifier
TNA	The National Archives (Great Britain)
URI	Uniform Resource Identifier
VERS	Victorian Electronic Records Strategy