Meeting the Preservation Demand Responsibly = Lowering the Ingest Bar?

Andrea Goethals; Harvard University Library; Cambridge, MA/USA

Abstract

This paper describes (1) a lessening of restrictions on the ingest acceptance criteria, and (2) a growing reliance on repository-supplied Submission Information Package (SIP) creation tools planned for the Harvard University Library's digital preservation repository (the DRS). Reasons for these changes include producer demand for the repository to accept new formats and genres, growing quantities of digital content producers want to preserve but can not adequately process, and the increase in the amount of content that is acquired by instead of created by DRS producers. In response to these needs the DRS will greatly change its SIP requirements and ingest process. Most notably files in any format will be accepted by the DRS. This paper describes steps taken to mitigate the effect of these changes so that the DRS can still meet its preservation commitments while broadening the range of acceptable content. One of these mitigating steps is the development of a new tool, FITS, which will be used to automate format identification and metadata extraction during SIP creation.

Background

Ingest and the OAIS Reference Model

It is widely accepted by the digital preservation community that the concepts and terminology of the OAIS Reference Model [4] are useful for understanding and discussing digital preservation repositories. While the OAIS Model does not prescribe a design or implementation for repository ingest, it does define roles and processes related to the ingest process, specifically the Producer role, the Submission Agreement and the Submission Information Package (SIP).

The Producer role is defined broadly to include systems as well as persons which create and provide the information to be preserved in the OAIS. The Producer and OAIS management negotiate a virtual or explicit Submission Agreement which specifies the formats, contents and logistics for delivering the content to the OAIS. The Producer creates a SIP in accordance with the Submission Agreement and submits it to the OIAS for ingest.

In order for this ingest model to work as described by the OAIS model some assumptions must be made:

- The OAIS management is always able to negotiate with Producers the formats and technical properties of the content that the OAIS will accept.
- Producers have the ability to meet the requirements for the formats, content and SIP packaging required by the OAIS.

As will be discussed later in this paper, in practice these assumptions do not always work with the new realities facing digital preservation repositories.

Repository Requirements for SIPs

Most if not all preservation repositories specify a set of requirements for SIPs that Producers must meet in order to ingest into the repository. An informal survey by the author of this paper of publicly available documentation was used to compile a list of preservation repository ingest requirements. These requirements can be grouped into three general classes: (1) the administrative and technical requirements a producer must meet to be eligible to deposit anything into a particular repository, (2) the requirements for pre-processing, legal clearances and content description a producer must meet prior to SIP creation, and (3) the requirements of the repository for the SIP itself. Only the third group is considered for the purpose of this paper. These requirements can include:

Content requirements

- Format requirements - files must be in particular formats or format profiles
- Format validity requirements - files must conform to their format's specifications
- File technical property restrictions - e.g. files must have particular text character encodings or image resolutions
- Data modeling/Content model restrictions - aggregations of SIP files must conform to repository specifications for classes of objects

Metadata requirements

- Schema restrictions – use of required administrative, descriptive, preservation, rights and other schemas
- Metadata element and value requirements - controlled vocabularies
- Existence, required structure, elements and attributes of metadata container files (e.g. METS XML files)

Packaging requirements

- Existence and structure of auxiliary files required for ingest
- Directory and file naming restrictions
- Composition requirements - e.g. enclosing in an archive format such as ZIP

Content and metadata requirements are imposed for both preservation and non-preservation reasons. The preservation reasons include the desire to influence deposits so that only well-described content in "preservable", valid formats are deposited. Non-preservation reasons include lack of staff knowledge and expertise beyond particular formats and schemas; and limitations and constraints of repository software, database and tools.

In terms of long-term preservation for the content in the repository, SIP restrictions have a positive effect. If they are rigorously defined, they result in producing predictable SIPs that can be ingested by the repository in an automated fashion. Content restrictions can reduce the number and variation of formats that repository management need to monitor and preserve[1]. Limiting the number of formats the repository manages also has the effect of limiting the number of delivery and rendering technologies that need to be managed and/or monitored.

A negative effect of ingest requirements is that they can become a barrier to ingest. If there are reasons that Producers can not meet the requirements, either because they have content that is ineligible for deposit, or because they don't have the technical means to adequately describe or package it, the content can not be submitted in a SIP that meets the requirements of the repository.

Balancing SIP Requirements and Producer Needs

The SIP requirements made by a repository and the needs of Producers for depositing their content for preservation into that repository may or may not conflict with each other. Factors that can contribute to whether or not these conflict are listed below:

Factors that contribute to a balance (Producers can meet SIP requirements and Repository can meet Producer needs)

- Producers have little and/or homogeneous digital content to preserve.
- The repository requires very little in the way of metadata and/or packaging of Producers.
- The repository has a large and knowledgeable staff. It is able to keep up with Producer demand to support more formats and genres.
- Producers have a large and technically savvy staff. They are able to produce repository-compliant SIPs in a timely manner.
- Producers want to preserve content for which they are also the Creators.
- Producers want to preserve content that is contemporary (is close in time to the creation date).

Factors that contribute to conflict (Producers can not meet SIP requirements and/or Repository can not meet Producer needs)

- Producers have a large amount and/or heterogeneous digital content to preserve.
- The repository staff is few and/or inexpert. They can't keep up with Producer demand to support more formats and genres.
- Producers do not have the resources (time, expertise, technical support) to be able to meet SIP requirements.
- Producers do not have the resources to fully process the content (e.g. fully appraise, resolve relevant legal, privacy and intellectual property rights information) in a timely way.
- Producers want to preserve complex content (email, web, etc.) that is difficult to process and preserve because of the relationships and dependencies among the content.
- Producers want to preserve content that is created by individuals using a variety of standards, tools and technologies, over which the Producer and repository have little or no influence (e.g. personal archives).
- Producers want to preserve content that is distant in time from the point of creation (e.g. content that was donated on a hard drive five years ago).

The Digital Repository Service (DRS)

The DRS is Harvard University Library's digital preservation repository. It can be used by any Harvard organizational unit for digital material that:

- supports research, scholarship and teaching;
- is intended for long-term preservation;
- is described in a publicly-accessible catalog or website; and
- makes a version of the content available to the Harvard community now or in the future.

The DRS is not a records management system, nor is it intended to be an institutional repository capturing all the university output. Harvard has a separate repository, Digital Access to Scholarship at Harvard (DASH), which is intended to serve as an institutional repository.

The DRS is developed and managed by Harvard University Library's Office for Information Systems (OIS). The design of the DRS began in 1999. The software was written in Java by OIS developers and was put into production in September, 2000. Over the years incremental enhancements and bug fixes have been added to the DRS software and storage system.

DRS Content

As of February 2009 the DRS manages a little over 11,500,000 unique files, which sum to 90.3 TB of content[2]. The DRS content conforms to a relatively small set of formats, as shown in Figure 1.

[1] This is only true if the repository's ingest process adequately validates SIPs to check that the SIP actually meets the repository's content requirements.

[2] The DRS stores from three to four copies of each file, depending on whether the file is categorized as high-use or low-use. These additional copies are not included in the 90.3 TB.

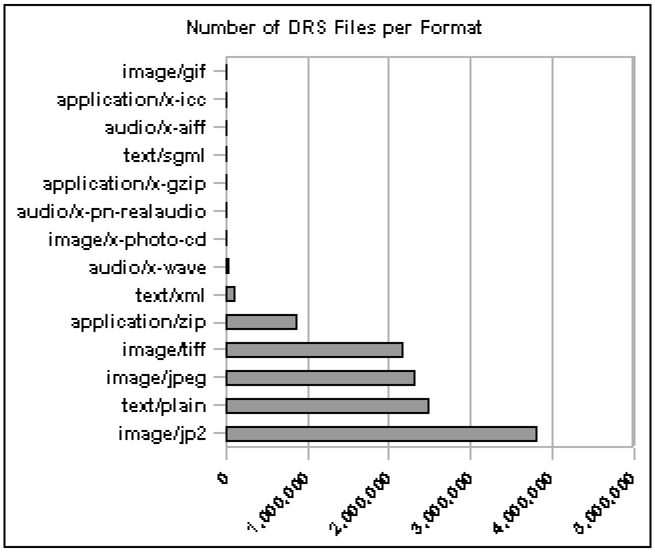

Figure1. DRS files currently conform to a small number of formats.

While it has been DRS policy for years to accept files in any format, in practice this policy was not supported. For every format supported by the DRS, OIS staff takes steps to ensure that (1) good preservation decisions are being made and that (2) the repository infrastructure can handle the format. This work includes:

- Acquiring deep knowledge of the format through specifications, workshops and references
- Determining any community- or domain-accepted best practices related to the preservation of this format
- Determining the appropriate technical metadata to use for the format, developing new schemas or extending existing schemas where needed
- Determining any special delivery or rendering needs for the format, researching third party solutions if appropriate
- Adding support for the format to the repository's pre-ingest, ingest, management and delivery systems and tools
- Adding support for the format and its metadata to the DRS database
- Updating internal repository and external depositor documentation

This is not a trivial amount of work for repository staff. In the face of competing priorities and projects it has led to a back log of requests for the repository to support new formats and genres.

DRS Producers

As shown in Figure 2, there are forty-nine libraries, museums, archives and other units at Harvard depositing content to the DRS. The units vary greatly in the amount of staff resources available for the deposit work flow. Some of these units use sophisticated imaging labs to serve as their depositing agents. Other units make use of staff who have many responsibilities in addition to acting as depositors for their units.

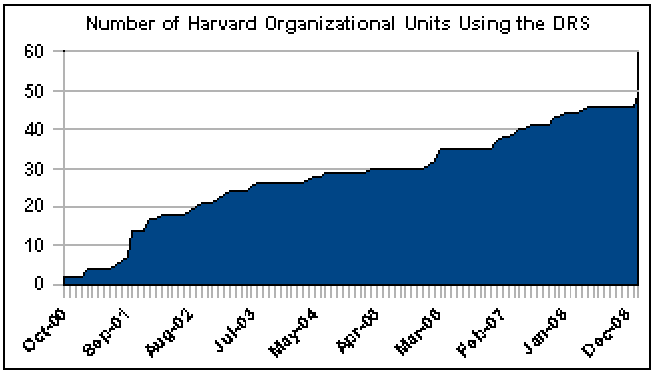

Figure 2. The number of Harvard organizations using the DRS has grown to forty-nine units.

Conflicts in DRS SIP Requirements vs. DRS Producer Needs

While it has been the case for many years that DRS staff have been able to specify fairly strict SIP requirements, and in general DRS producers (collection managers, curators and depositors) have been able to meet these requirements, in the last few years these abilities have started to degrade. There are three main reasons for this degradation:

1. DRS producers are increasingly overwhelmed by the amount of digital content they should be preserving. This reduces their ability to fully process the material they want to deposit into the DRS and reduces their ability to go through any deposit process that is time-consuming.

Some DRS Producers are stewards of archives and special collections at Harvard. Traditionally these units have cared for analog materials but are increasingly becoming stewards of digital material. With analog material these units could "buy time" by storing the material at Harvard's archival media storage facility until they had the resources to fully process it at the item-level. With digital material they know that the processing of this material can not be delayed in the same way that it can for analog material. The media that the digital content is stored on could fail or become obsolete. The files could become corrupt. The formats could become obsolete. For these reasons there is a growing demand on the DRS to assist these units in accepting all this diverse digital content, even if they have only minimally processed it.

2. Increasingly DRS producers are depositing content that they did not create, for example content acquired by web crawlers. This content was created by persons with little or no relationship to DRS staff; therefore DRS staff can not influence the formats or technical properties of this content during creation.

For a number of years DRS Producers deposited primarily digital content converted from analog content. These include images, scanned books and other paged-turned material, and digital audio. Over the last few years, in response to growing demand, the DRS has begun a series of projects to support born-digital genres and formats. In the spring of 2008 the DRS began to accept web-harvested content acquired from Harvard University Library's new web archiving service, WAX. In the fall of 2008 an email archiving project began at Harvard that will result in the DRS accepting email collections by 2010. The typical case for web archiving and email archiving is that the repository has little or no influence over the creation of this content.

3. DRS producers increasingly need to preserve formats and genres that aren't currently supported by the DRS. This preservation demand is straining the DRS' ability to restrict the formats and genres that the DRS will accept and still meet the needs of its producers.

DRS depositing units were interviewed in 2008 to find out what digital material they currently have that they would like to deposit into the DRS if it were supported. The formats and genres they would like to see supported include:

- CAD formats
- 3D Visualization formats
- Additional Audio formats
- Video formats
- Digital fonts
- Word Processing formats
- Spreadsheet formats
- Presentation formats
- Databases
- Locally archived websites
- Raw survey data
- Raw camera files
- Executable files

It is clear from the interview results, and the born-digital projects, that the DRS is on a trend towards supporting a broader range of formats and genres.

SIP Requirements and Tools

The DRS SIP is the package that is submitted to the DRS loader for ingest, and is known as a "batch" to depositors. The batch consists of:

- The content files in DRS supported formats, and
- a "batch loading file". This is an XML file containing administrative and technical metadata for the content files.

The batch is built by depositors on their computer systems and then transferred by the depositor via sftp to a directory monitored by the DRS ingest loader. When the loader sees that a batch is ready for import, it imports the metadata found in the batch loading file into the DRS database and the content files into the DRS file system.

DRS depositors, especially those from smaller Harvard units, have always had difficulty creating the XML batch loading file. Many batches have been rejected during ingest because of mistakes in the loading file. In the first few years of the DRS, depositors developed their own scripts and programs to help with the creation of the batch loading file. This model is very OAIS-like where the Producers (DRS depositors) have the full responsibility for creating the SIP required by the repository. Over the years it became apparent that the deposit process would work much better in terms of ease to the depositors, validity of the batch loading file and quality of the purported metadata if the creation of the SIP were assisted by repository tools. Essentially this makes the SIP creation a shared responsibility of the repository and Producers.

Beginning in 2005, the DRS started to create SIP-creation tools for DRS depositors. A prototype, DSIP, was developed by OIS in 2005. It was a command-line script built as an extension to JHOVE [8]. It was tested within OIS and used for in-house projects but never released to DRS depositors. The second iteration in 2006 produced Batch Builder, a tool that is still use by DRS depositors today. Batch Builder has both a desktop GUI and a command-line interface. It combines metadata manually input by the Batch Builder user, metadata stored in configuration files and JHOVE output to automatically generate the batch loading file and populate it with administrative and technical metadata.

DRS 2

Beginning around 2006 the OIS starting planning for a multi-year enhancement of the DRS - named "DRS 2". The current implementation of the DRS was retroactively dubbed "DRS 1". The largest changes moving from DRS 1 to DRS 2 are:

- A new data model – In DRS 1 the content is modeled very simply as files. The content in DRS 2 will be modeled explicitly as objects, files and bitstreams.
- All DRS 2 objects will conform to one of a newly designed set of "content models", or classes of objects. Each content model definition will specify the acceptable formats, file roles, relationships, metadata requirements, delivery and known rendering applications for the object. One of these content models will be for "opaque objects". There are no format restrictions on an opaque object's files.
- In DRS 1 the metadata schemas were primarily custom schemas developed in-house. Now that there are community-standard metadata schemas such as PREMIS, MIX, and textMD, they will be used for DRS 2.
- In DRS 1 all of the metadata was stored in the DRS database. The metadata in DRS 2 will be stored on the file system in METS XML object descriptor files stored with the content files, as well as in the database. The requirement of this descriptor file per object would be an impassable barrier for DRS depositors if tools were not provided for them by the repository to generate the descriptor files. For this reason the Batch Builder tool will be enhanced to use new tools, FITS and OTS, to generate the descriptor files.

As a response to the needs of DRS Producers for the DRS to support minimally processed content in any file format, one of the first DRS 2 enhancements is support for opaque objects.

Opaque Objects Enhancement

Opaque objects are represented by a new content model in DRS 2. They permit files in any format, but they are "opaque" to the repository – it doesn't know what the object represents, it does not record any relationships between its content files, and it does not permit it to be delivered except to its Harvard owning unit.

The primary goals of the enhancement are:

- to accept files in any format for ingest into the DRS, which is consistent with DRS policy
- to provide minimal preservation services to content that otherwise would not receive any preservation services, including secure, redundant storage; file integrity monitoring; and technical characterization of files
- to make accessible to DRS staff, content that needs to be represented by new or modified DRS content models

While these changes are being made to handle the preservation demand from DRS depositors, it is also recognized that the DRS is taking on a large preservation responsibility for accepting this content. To mitigate this risk the DRS is creating new tools and processes to:

- Automatically identify file formats
- Automatically extract metadata from files
- Automatically validate files according to their purported formats
- Flag DRS files with ambiguously identified formats in order to improve file identification tools
- Flag DRS files found to be invalid so that they can be analyzed at a later date
- Clearly label DRS opaque objects so that they can be later analyzed, re-characterized and/or migrated into new or modified DRS content models.

Depositors are encouraged to submit license, original order and other documentation with opaque objects. A file naming scheme for these documentation types of files is specified so that they can be easily identifiable for future uses. To support the automatic file identification and metadata extraction, a new tool called FITS is being developed by OIS.

FITS (File Information Tool Set)

The File Information Tool Set (FITS) identifies, validates and extracts technical metadata for a wide range of file formats. It acts as a wrapper, invoking and managing the output from several other open source tools. The native output from these tools is converted into a common format (FITS XML), compared to one another and consolidated into a single XML output file. FITS is written in Java and is compatible with Java 1.5 or higher. FITS can be invoked by its command-line interface or through its Java API. The external tools wrapped by FITS currently are:

- JHOVE [8]
- Exiftool [11]
- National Library of New Zealand Metadata Extractor [10]
- DROID [13]
- Ffident [5]
- File Utility [6]

The philosophy behind the design of FITS is that there is a need among domains and communities external to the digital preservation community for tools that identify formats and extract metadata from files. These communities include the open-source community, computer scientists, artificial intelligence researchers, imaging professionals, and audio and video applications. The digital preservation community can leverage the tools created by these communities and create or enhance tools where tools don’t exist for particular formats, or where desired metadata is not extracted.

FITS produces a “status” value for each format identification it makes. When the status is SINGLE_RESULT, all tools that were able to identify the format agree on the file’s format. When the status is CONFLICT, there is more than one purported format identified for the file.

Because FITS combines the output of multiple tools it has to be able to handle conflicts among the tool’s output when they don’t agree. It handles this conflict in many ways:

- Tool output is normalized before it is compared for conflicts. For example, one tool might report for a file format that it is “PNG”, while another tool may output it as “Portable Network Graphics”. In another example, one tool might output the resolution unit as “2”; another tool might output it as “inches”. These values are normalized in the XSLT file that converts the tool’s native output to FITS XML before the FITS XML for each tool is compared to each other. This reduces the number of false positive conflicts.
- Users configure a tool ordering preference. In cases of format identification conflicts, the format identified by the preferred tools will determine the format FITS reports.
- Tools can be excluded from reporting on particular formats and/or on particular metadata elements if its output is found in testing to be incorrect or buggy. This is very useful for incorporating a tool into FITS because it is good at some things without having to accept known unreliable information from the tool.
- FITS consults a configurable “format tree” to know when two reported formats for a file are not really conflicts because one of the formats is a more specific form of the other format. For example the format tree documents that the OpenDocument Text format is a more specific form of the Zip format. If a file is identified as being in both of these formats by FITS tools it is not reported as a conflict because technically they are both correct. Instead the more specific format, OpenDocument Text, is reported as the format.
- Users configure whether or not to report format identification conflicts. The default is to report conflicts (i.e. report all formats that the tools identified for a file).

Discussion and Future Work

The opaque object enhancement and the development of FITS and other SIP-creation tools is the first step toward meeting the expanded needs of DRS depositors. It is recognized that this first effort does not solve the entire problem and that there is a lot of work to be done in this area in the future.

Opaque Objects in the DRS

The opaque object content will at least receive some preservation services. Instead of the content living on various computer systems, external hard drives, or various removable media outside of the repository it is receiving professional care in a well-monitored secure system. It is technically characterized using a variety of file identification and metadata extraction tools.

Files for which the format could not be identified, or for which there were conflicts in the identified format, are flagged in the file metadata for subsequent preservation services.

From the depositor's perspective the most obvious loss in service for opaque objects as compared to other objects is the lack of a delivery service for opaque objects. At least initially there will be no delivery services for opaque objects. These objects will only be available through an administrative interface to the Harvard units who were responsible for depositing the objects into the DRS. One of the reasons for restricting access to these objects is that the legal clearances for public redistribution may not have been obtained by collection managers before depositing the content into the DRS. One of the use cases for depositing opaque objects is depositing the contents of a hard drive donated to a Harvard archive. In these cases the content won't have been processed at an item-level yet.

OIS is providing guidelines to depositors for "designing" their opaque objects. The intention of the guidelines is to ease the expected future migration of opaque objects into supported content models. These guidelines are:

- Depositors should try to minimize the amount of future "object manipulation" (splitting and merging) needed. This object manipulation may come with a price by having to store lengthy and complex process history. For instance two video works that in the future could be migrated to two separate video objects (once the DRS supports video objects), should be deposited as separate opaque objects now.
- Try to include within the same opaque object, files that are related by derivative relationships (e.g., masters and deliverables); or have rendering dependencies (for example style sheets, scripts and images should be grouped together in the same object as web pages using them); or require the same descriptive metadata (are part of the same work with the same bibliographic record)

It is expected that the DRS guidelines for opaque objects will grow and adapt as experience is obtained. As implied by the guidelines, in the future DRS depositors and staff will need to be able to "rearrange" the contents of opaque objects. They will need to be able to split and merge them. DRS depositors have a similar need for other types of objects that are already in the DRS. It's not uncommon for portions of page-turned books to be scanned as separate projects and then later merged together for delivery purposes after the separate pieces are already in the DRS. A generic service to rearrange object content would serve both purposes.

Future of FITS

It is the intention of OIS to release FITS as open source under the LGPL license. Any documentation or code for FITS will be available on the FITS website [7]. There are a number of tools that will be evaluated for incorporation into FITS in the future:

- Apache Tika [2] for document and other formats
- JHOVE 2 [3]
- Aduna Aperture [1] for document, text, email formats
- MediaInfo [9] for audio and video formats
- The tools listed in the Cairo Tools Survey [12]

FITS will be tested against the particular formats expected to be deposited into the DRS. It is a known fact that it will end up being used with formats found in opaque objects for which it was not tested. The potential problems in these cases can include unidentifiable formats, false positive conflicts in format identifications, true conflicts in format identifications, and a failure to extract some technical metadata. In the cases of format conflicts, flags in the metadata will be used to find these problems and improve FITS. The unidentifiable formats and failure to extract some technical metadata could be fixed by adding new tools to FITS that work better for these formats and metadata.

References

[1] Aduna and DFKI. "Aduna – Aperture", <http://www.aduna-software.com/technologies/aperture/overview.view> (22 March 2009).

[2] Apache Software Foundation, "Apache Tika", <http://lucene.apache.org/tika/> (22 March 2009).

[3] California Digital Library, Portico and Stanford University, "JHOVE 2 Home", 28 August 2007, <http://confluence.ucop.edu/display/JHOVE2Info/Home> (22 March 2009)

[4] Consultative Committee for Space Data Systems, "Reference Model for an Open Archival Information System (OAIS)", Blue Book, Issue 1. (2002).

[5] "Ffident", <http://schmidt.devlib.org/ffident/index.html> (February 2009>.

[6] "File for Windows", GnuWin32, 19 February 2009, <http://gnuwin32.sourceforge.net/packages/file.htm> (22 March 2009)

[7] Harvard University Library, "fits – Google Code", <http://code.google.com/p/fits/> (22 March 2009).

[8] Harvard University Library, "JHOVE – JSTOR/Harvard Object Validation Environment", 25 February 2009, <http://hul.harvard.edu/jhove/> (22 March 2009).

[9] "MediaInfo", <http://mediainfo.sourceforge.net/en> (22 March 2009).

[10] National Library of New Zealand, "Metadata Extraction Tool", <http://meta-extractor.sourceforge.net/> (22 March 2009).

[11] Phil Harvey, "Exiftool", 20 March 2009, <http://www.sno.phy.queensu.ca/~phil/exiftool/> (22 March 2009).

[12] Susan Thomas, Fran Baker, Renhart Gittens, Dave Thompson, "Cairo tools survey: a survey of tools applicable to the preparation of digital archives for ingest into a preservation repository", version 1.0, JISC. (2007).

[13] The UK National Archives, "Introduction – DROID", 29 August 2006, <http://droid.sourceforge.net/wiki/index.php/Introduction> (22 March 2009).

Author Biography

Andrea Goethals received an M.Sc. in Computer Science from the University of Florida. Before coming to Harvard she worked at the Florida Center for Library Automation. She is leading the establishment of Harvard's new digital preservation program and manages its preservation repository, the DRS. She is currently involved in a major upgrade of the DRS and an international effort to establish a format registry for digital preservation.