

PRESERVING OUR DIGITAL HERITAGE

The digital age poses a dilemma to libraries, museums, and archives. On the one hand, electronic information provides unprecedented new opportunities for such institutions to disseminate knowledge and to serve their communities in innovative ways. Digitized versions of physical works, such as paintings or videotapes, can bring these materials to a worldwide audience via the Internet. And exciting new works that exist only in digital form (such as videos and multimedia works) are increasingly finding their way into museum and archive collections.

On the other hand, preserving and maintaining these digital assets for posterity is a monumental problem. The custodians of digital information need to guard against not only hard drive crashes and natural disasters, but also the ongoing obsolescence of technologies, which can render file formats and storage media unreadable in just a decade or two. What good is a 5.25-inch floppy disk containing WordStar documents, for example, if you no longer have a computer that can read the file format, let alone a drive that can read the disk?

“Digitization and digital life are growing exponentially,” says RLG program officer Robin Dale. “Digital archiving is important now, but will be unavoidable for many institutions within two to three years.”

RLG took important steps in 2002 to help its members meet the challenges of digital preservation. Much of that work has focused on defining the characteristics of a “trusted digital repository” where significant digital assets can be archived for the long term.

Defining the digital repository

The starting point for this work has been the Open Archival Information System (OAIS) Reference Model, a comprehensive logical model describing all of the functions required in a digital repository. The OAIS model, which emerged out of an initiative spearheaded by NASA’s Consultative Committee for Space Data Systems, outlines how digital objects can be prepared, submitted to an archive, stored for long periods, maintained, and retrieved as needed.

RLG, OCLC, many members of the two organizations, and several groundbreaking international projects have played key roles in shaping the OAIS model and in adapting it for use in libraries, archives and research repositories. But the OAIS guidelines are just a starting point, which is why RLG continues to help its members understand, implement, and build upon the OAIS model.



RLG program manager Robin Dale and program manager Anne Van Camp meet with Digital Library Foundation president David Seaman, at RLG’s offices, to discuss digital preservation and trusted digital repositories.

In February 2002 RLG launched a new Web page for people working with the OAIS model (www.rlg.org/longterm/oais.html). This page is designed to facilitate communication and collaboration among institutions, and it includes links to resources, publications, and recent workshops. RLG also started a new discussion list, called *oais-implementers*, which is available from this page.

“OAIS is a conceptual framework for the preservation of digital information,” says Dale. “It is not a blueprint to build a digital archive. Its strength is in establishing common terms and concepts for describing repository architectures and comparing implementations, without specifying an implementation an organization should use.”

To help members get a handle on these aspects of digital preservation, RLG and OCLC in May jointly published the final version of a white paper, “Trusted Digital Repositories: Attributes and Responsibilities.” This paper aims to define the characteristics of reliable archiving services for heterogeneous research collections.

In June a joint working group from OCLC and RLG issued another white paper, “A Metadata Framework to Support the Preservation of Digital Objects.” Preservation metadata is the information infrastructure necessary to support digital preservation—such as the descriptive information specified in the OAIS model. This report is a comprehensive guide to preservation metadata.

Hands-on experience

Not content merely to research the issues and publish white papers, RLG also got hands-on experience with digital archiving through a digital repository testbed project, conducted with a third-party provider of archiving services, JPMorgan’s *i-VAULT!SM*.

How the OAIS Reference Model Works

i-VAULT! has provided digital image archiving services for over 15 years. The company's first clients were banks, who needed to store digital images of canceled checks, although the digital objects stored by customers have since expanded to include wills, deeds, titles, and other legal documents. i-VAULT! maintains two separate, redundant archive sites and is currently storing the equivalent of over 700 million document pages. (See www.jpmorgan.com/ivault/ for more information on this service from JPMorgan.)

After extensive discussions with i-VAULT!, RLG decided to test whether the company's image archiving system would fit the bill as a third-party trusted digital archive for libraries, museums, and archives. "i-VAULT! was very attractive," says Dale, "because it could provide the archival storage function of OAIS," potentially saving RLG or its members the trouble and expense of building one of the significant components of an OAIS-compliant archive.

In the test project, RLG took a ten-gigabyte subset of data from RLG Cultural Materials, stored these objects in a test repository provided by i-VAULT!, then retrieved data from the repository to verify that the objects had been successfully archived. In the process, says RLG software development manager Judith Bush, RLG learned a lot about the technical processes required to implement the OAIS model. For instance, RLG had to create submission information packages for transmission to i-VAULT!. The test team also needed to consider the access requirements and the Web interface that would be needed for future retrieval of those objects.

The tests showed that i-VAULT! has much promise as a digital repository. "It is a very robust service," says Linda West, RLG's director of member programs and initiatives, "and could form the backbone of a certified digital repository. However, such services don't come cheap." After considering "sunk costs" (those already spent on staff and servers), RLG concluded that using a third-party archiving service would be less economical than building an archive in-house. RLG expects to have over ten terabytes of data in RLG Cultural Materials alone within five years, notes RLG chief information officer Jack Grantham, and for that quantity of data, RLG judged that the cost of i-VAULT! would be too high to justify, since archival storage is only one component of a full-fledged digital repository.

Grantham believes that many institutions will make a similar calculation regarding third-party archiving services, especially when considering labor costs. Universities, for instance, have access to existing IT staff resources—and inexpensive undergraduate labor pools. Because these costs are low (or aren't included in budgets for in-house archiving projects), third-party archiving services look costly by comparison. "One huge issue that any certified digital archive will have

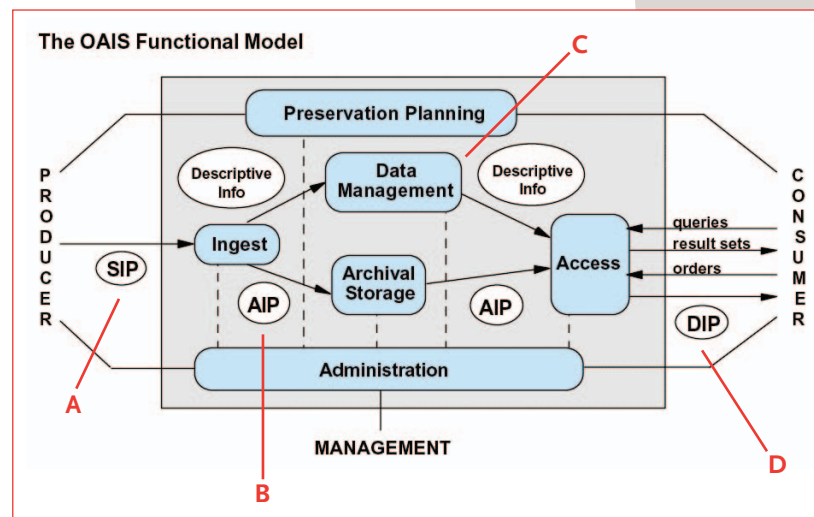
A In the OAIS model, the producer of a digital object combines it with data about the object, which may include descriptive metadata (details about provenance, context, and so forth) and details about how the object is to be stored and accessed. The combination is then submitted to a repository as a submission information package (**SIP**). An SIP may also include information about the object's data format—ensuring that future generations will be able to retrieve the object and will have information about how to open, use, and view it. For example, a PDF file might be submitted to a digital repository along with information indicating where a technical explanation of the PDF format itself can be found.

B The archive's **ingest** function generates an archival information package (**AIP**) from one or more SIPs. The AIP may include metadata associated with the object, as well as information about the object's data format. The **archival storage** function is responsible for storing, maintaining, and retrieving the archive's AIPs.

C Descriptive metadata about archived objects can be stored separately by a **data management** function, which allows users to retrieve details about what's in the archive without accessing the archived objects themselves.

For example, data about an archived document's author, creation date, and subject matter could be stored here, enabling users to search against that information and to view details about a file before requesting that it be retrieved.

D When a consumer retrieves an object from the archive, the **access** process uses the stored metadata and file format information to deliver a dissemination information package (**DIP**) that ensures the user can actually access the stored object. In the case of a PDF file, that package might contain the file itself as well as a copy of Adobe Acrobat® Reader™ for viewing the file.





to face when approaching big universities is that you can't create a pricing model that will be convincing to them," says Grantham.

However, the initial steps taken by RLG and i-VAULT! bode well for the future. "It's encouraging that there's a large, world-renowned vendor out there that's willing to work with our community to create a digital archiving service that might work for us," says Dale. Eventually, the price of third-party solutions such as i-VAULT! may converge with the amount that libraries and other institutions are willing to pay for such services. "At some point in the future, people are going to have to bite the bullet and make the investment in archiving," Dale says. "i-VAULT! might be a reasonable solution for institutions who cannot build and maintain their own archival storage."

More on the horizon

At the close of 2002 a new initiative was just beginning. RLG and the National Archives and Records Administration (NARA) have formed the joint Task Force on Digital Repository Certification to produce certification requirements for establishing and identifying reliable digital information repositories.

For digital archiving services to be effective and trustworthy, all parties involved in the process need to have a shared understanding of what is to be done, and how, by the repository. Institutions are not likely to entrust their valuable digital assets to an archiving service if they don't have assurances that the service provider is reliable, stores and manages data according to accepted standards, and is likely to be in business for a long time. Similarly, institutions building their own repositories may need assurances that their archiving methods conform to best practices (in order to meet funding guidelines, for instance). A certification process for digital repositories is one way to provide that assurance to all interested parties.

Along with developing a certification process, the new task force is charged with identifying a certifying body (or bodies) that can implement the process. For more information about the task force and its membership, see www.rlg.org/longterm/certification.html. ■