Journal of Zhejiang University-SCIENCE C (Computers & Electronics)
ISSN 1869-1951 (Print); ISSN 1869-196X (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Report:

Requirements and characteristics of a preservation quality information management system

Gabrielle V. MICHALEK
Carnegie Mellon University Libraries, Pittsburgh, Pennsylvania, USA
E-mail: gabrielle@cmu.edu

doi:10.1631/jzus.C1001013

1 Introduction

The proliferation of digital materials has changed not only how information is presented but also how people expect information to be available. People want access to all forms of information, from simple text to complex multimedia. Whether or not the items in question were created digitally, they can be made to behave digitally through scanning and conversion. This improves access, but makes preservation more difficult because of the rapid rate of obsolescence of formats, hardware and software systems.

In the early days of the digital age, the gap between librarians and the people working in information technology was vast. The past 20 years, however, have seen thousands of digitization projects that include scanning paper analog resources and making them available in digital format. Such work has coupled the worlds of tech and libraries. This cross pollination has resulted in rich, robust online resources worthy of preservation. Libraries have a role in preservation of such resources, much as they have had a role in preservation of the hard-copy word.

The Carnegie Mellon University Libraries (CMULs) have been at the forefront where digitization of rare and primary sources is concerned. Our pioneering efforts in digitization have also involved a creation of preservation strategies for digital content.

2 Strategies used by Carnegie Mellon University Libraries in system development

This paper focuses on the requirements of an information management system that is designed to ensure the long-term preservation of the information managed and delivered by that system. It is not an all-inclusive list or description, but an overview of the approaches and methodologies that have been used by CMULs as we have built, developed, and now, acquired, information management systems that have the onerous task of supporting collections we have promised to keep in perpetuity.

This paper will focus on those approaches, activities, and philosophies maintained at CMULs to ensure long-term preservation is part of the design and management of our systems. These activities break down into four categories: adherence to standards for image and metadata creation, management of data, system architecture, and security and auditing.

2.1 Use of standards for image and metadata creation

Use of standards is basic library practice and one that has been accepted by those tech specialists who partner with libraries. By now, most viable digital library programs have gotten onboard with the standards that pertain to image creation, derivative format creation, and some metadata creation. Most programs have embraced the standards that support image creation simply because they lead to the production of better image quality, better reproduction quality and longevity and interoperability of the data. CMULs participated in the work by the Digital Library Federation for the creation of their 2002 publication, “Benchmark for Faithful Digital Reproduction of Monographs and Serials” (http://www.diglib.org/standards/bmarkfin.htm). The CMULs have been

following these recommendations along with the standards set forth by the Library of Congress (http://memory.loc.gov/ammem/about/techStandards.pdf).

In addition, standards are beginning to evolve around derivative formats. JPEG2000 and PDF-A are two of these formats. The PDF-A format is good for archiving since it is 100% self-contained and even allows for the embedding of metadata into the file. Most files that began life as textual documents can be converted to PDF-A format, making it an attractive choice for professionals concerned about long-term preservation. CMULs maintain PDF-A files for most of our digital collections where the PDF format makes sense. Textual documents, and JPEG2000 format for image collections, are examples.

One of the most important methods of supporting long-term preservation for digital collections is through the creation and/or acquisition of reliable metadata. There are several different types of metadata, each representing a different purpose such as descriptive or technical. If each of the metadata types is used correctly and if they are used in concert with one another, chances of long-term preservation increase significantly.

The descriptive metadata describes the collection. The standards for descriptive metadata that we use most frequently at CMULs include, machine-readable cataloging (MARC) for library materials, Dublin Core for library materials and historic image collections, encoded archival description (EAD) for archival collections, and visual resources association data core (VRA) for art image collections. These metadata schemas are in extensible mark-up language (XML) format and work as part of the XML platform that makes up the system architecture.

Technical metadata for each digital collection is also created. This metadata describes the attributes of the file, such as which software was used to create the file, the date the file was created, the file size, which software was used to create the image, etc. Much of this metadata is created by the software used to create the digital image. This information is captured and stored as an XML file and is used to manage the collection over time. It is especially useful when the collection is being migrated to a new platform.

2.2 Management of data

Libraries and Archives have long been involved in management of data sources. The digital world spends its first few decades not taking advantage of the data management expertise so easily found in the local library. Much as digitization projects have brought information technology (IT) methods into the library world, library sciences such as data management have gradually become incorporated into the IT mindset. To help ensure both long-term preservation of data and persistent integrity of that data, CMULs have put into place a program, mostly administered by system managers, to develop tools and techniques to manage data effectively over time. One technique is the daily backup of scanned images to a remote server. This automated process includes email alerts that notify the system mangers of the progress and of transfer problems. The data is then periodically backed up to tape. According to industry standards and best practices, these tapes have a life expectancy of up to 25 years. However, because tape capacities continue to increase, each time we change to a new type of tape with a larger capacity format, all of the tapes are read back and re-written to the new backup tapes. At a minimum, tapes are spot checked annually. In essence, what we have here is a disk-to-disk tape backup strategy which parallels not only the method of data preservation of Carnegie Mellon's Computing Services Department, but also mirrors best practices of the IT industry at large.

The utility used for disk-to-disk copying is an open-source program called 'rsync'. rsync first compares the files (via an rsync algorithm) on the source and the destination, and copies only files that are newer or updated. rsync effectively transfers only the changes across the network and merges them with the last full backup on the remote machine, thus providing full backups of the local server with the added efficiency of incremental backups. There are other ways to reliably copy and check large volumes of data; however, this is the method selected by CMULs. When the data is moved to backup tape, the Libraries use a reliable utility and format called 'tar', tape archival.

Post-processed 'fixed content' is stored on networked volumes that are part of our 10 TB storage area network (SAN) which is located in a climate controlled server room in Hunt Library. This SAN resides on a private 'storage network' and is accessible only by the hosts that are also on this private

storage network. This is a security precaution. The data is further protected by RAID 5 (distributed parity) technology. If there is any hardware failure or pending hardware failure, the system managers are notified via email alerts from the RAID array. In such a case, the spare disk is written to and the bad disk is immediately replaced thus restoring active RAID protection. All data has now been copied to our new SAN which has RAID protection.

To help ensure long-term preservation, managers must verify the integrity of the data within the repository. The method we use at CMULs is by adding checksum capability at those locations where data is transferred from one place to another. The managers compute checksums as early as possible and store the checksum alongside the data. Later on the manager runs the checksum again and compares the checksum with the results. If the sums do not match, the file has been corrupted in transmission or storage. This capability helps verify the integrity of the data and that the data has not been corrupted over time. This not only has the advantage of catching network problems, but also tapes problems when data is brought back from tape.

For over a decade CMULs have been using persistent uniform resource locators (URLs) as a way of ensuring that links to our digital objects do not break over time. This is especially important when scholars cite digital resources in their research. Broken links make it impossible for future scholars to retrace or continue the work. The persistent URL is actually a tool in the shape of a database that directs a request to the current URL for a resource. That link will always be associated with the resource even when data gets moved to different servers over time. This functionality is critical to preservation activities and it is essential that it be included as part of a digital library program.

2.3 System architecture

XML is an open standard, and a precise way of storing and communicating data. XML is easy to work with and aids preservation because it is human readable. It is especially good at representing complex data structures and hierarchies such as EAD, MARC, and Dublin Core as described in Section 2.1. Several application programmable interfaces (APIs) have already been developed to explore and extract data stored in XML format, making it robust and giving the specification long-term viability. XML is platform independent, so it allows a project to create data that can be read by applications on other platforms. This improves the odds of preservation by eliminating the concern that applications and platforms will become obsolete.

The Open Archives Initiative, OAI-PMH, is a method used to expose structured metadata such as EAD and Dublin Core to Web crawlers. The metadata can then be discovered by Internet search engines used by harvesters and aggregators such as OCLC's OAIster, or the National Science Digital Library (NSDL) in the United States, making it accessible to the world. In addition, the metadata from one set of collections can then be added to a database with metadata from other collections making the collections interoperable. Interoperability allows content from a diverse group of collections from multiple organizations to be organized, managed, and disseminated in ways that promote new learning to a wide range of communities. Since the EAD metadata schema is XML encoded, there are many ways in which to use the OAI standard to communicate the data between repositories and harvesters. Making data interoperable and sharing it with other repositories help preserve information by distributing it across locations and platforms.

2.4 Security and auditing

It is important that systems include software tools that allow for content versioning and the ability to conduct a digital audit. These two pieces of functionality will indicate if data has been modified, when the changes occurred, and who made the changes. This helps ensure accuracy of the information, making researchers more confident in the integrity of the information. It also helps system managers track down any problem with the data and locate where in the data stream the problem may have occurred. While DIVA, the current home-grown information system used by CMULs, has rudimentary content versioning to produce an audit trail, a more robust version of this functionality is required as the

Libraries move to a new commercial grade information management system.

To aid in the creation of an audit trail, managers must include computer security that provides administrative control. This will also protect the data from theft, corruption, and disaster. Users may have read permission and can access documents, but they may not alter the document or metadata in any way. Personnel working with data may have limited permissions to work within the system. They must be required to log in to add, delete, or modify the data. System managers and administrators may have access to all of the data within the system to be able to change the data, and also to modify the system itself. However, they must be forced to log in as well. This, too, is functionality required of the new system CMULs are purchasing.

Good security controls authorization right down to the object level. For the end user, this security determines whether a user can access a folder, or a file, and whether the end user can move from file to file, or folder to folder. Security control also determines whether the end user can see all of the metadata associated with a file or just a limited set of metadata. For personnel, this determines whether the user can access, create, modify, or delete a folder or file. For a system manager or administrator, it determines whether the user can do all of the above as well as change permissions down to the file level. This functionality is especially important as 'born digital' materials are ingested to the system. As materials are digitized we have the luxury of reviewing each document to determine whether it should be scanned and added to the system. This may not be the case with born digital materials such as email. It may be that managers would like to make some of the emails available to end users because of reasons of confidentiality. This functionality would support these objectives.

3 Conclusions

For the past 20 years the Carnegie Mellon University Libraries have been a leader in the field of library digitization. Whenever the Libraries commit to creating and maintaining a digital collection, it is making a promise to preserve that digital collection in perpetuity, or for as long as it is technically feasible. In the process of making these unique library resources available online we have gained experience and knowledge concerning the preservation of these digital objects. This paper has described some of the tools and approaches our managers have used to aid in the long-term preservation of the digital library. While this paper does not purport to be a comprehensive list of all of the technical approaches necessary to ensure long-term preservation, our techniques are easily replicated and represent some of the best practices in the field.

References

Caplan, P., 2009. Understanding PREMIS. Library of Congress, Washington DC, USA. Available from http://www.loc.gov/standards/premis/understanding-premis.pdf

Federal Agencies Digitization Guidelines Initiative, 2009. Digital Conversion—Documents and Guidelines: a Bibliographic Reference. Available from http://www.digitizationguidelines.gov/stillimages/documents/Guidelines_Bibliography-2009rev.pdf [Updated on Aug. 28, 2009].

Peterson, Z.N.J., Burns, R., Ateniese, G., Bono, S., 2007. Design and Implementation of Verifiable Audit Trails for a Versioning File System. Proc. Conf. on File and Storage Technologies, p.93-106.

Rose, K.L., 2009. Preserving Our Digital Collections. Presentation for Cultural Memory Class, Carnegie Mellon University, Pittsburgh, PA.

Rumsey, A.S., (Ed.), 2010. Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information. Technical Report, Blue Ribbon Task Force on Sustainable Digital Preservation and Access. Available from http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf