Exploring Variety in Digital Collections and the Implications for Digital Preservation

MacKenzie Smith

Abstract

The amount of digital content produced at academic research institutions is large, and libraries and archives at these institutions have a responsibility to bring this digital material under curatorial control in order to manage and preserve it over time. But this is a daunting task with few proven models, requiring new technology, policies, procedures, core staff competencies, and cost models. The MIT Libraries are working with the DSpace™ open-source digital repository platform to explore the problem of capturing research and teaching material in any digital format and preserving it over time. By collaborating on this problem with other research institutions using the DSpace platform in the United States, the United Kingdom, Europe, and other parts of the world, as well as with other important efforts in the digital preservation arena, we are beginning to see ways of managing arbitrary digital content that might make digital preservation an achievable goal.

Introduction

The modes of scholarship—research, teaching, and communication—continue to evolve toward online digital content that supports critical innovations. Creating digital content and making it available on the Web as a part of the research process is not only getting easier, it is becoming routine. As an example, the U.S. National Institutes of Health (NIH) research proposals now require a description of how data produced with their funds will be made available to share with other researchers. And electronic journals with the ability to provide full-text searching and hyperlinks continue to become the accepted, and expected, norm. The amount of digital information

LIBRARY TRENDS, Vol. 54, No. 1, Summer 2005 ("Digital Preservation: Finding Balance," edited by Deborah Woodyard-Robinson), pp. 6–15

produced annually is, by some estimates, more than thirty-five thousand times the complete contents of the Library of Congress and growing fast (Varian & Lyman, 2003).

But who will ensure that this digital scholarly record continues to exist in an era when the lifespan of digital content is normally measured by a few years? Libraries and archives are just beginning to grapple with the problem of capturing, managing, distributing, and preserving the digital material that their constituents are producing, and to effectively deal with this content requires not only new technological infrastructure but new policies and procedures, new core competencies of staff, and new business lines and cost models—in other words, significant transformation of the current models of institutional scholarly content management.

Preserving this digital material is one of the most challenging components. The digital formats of the content are various and are dictated by the short-term needs of faculty and researchers who have innovation as their driving force; thus, their motivation to use only "good," standard digital formats is very low. Libraries and archives will have to deal with this material whether or not there are well-understood ways to keep it usable over time. Thinking about how to establish the infrastructure and business practices to accomplish this, and keep the costs manageable, is a formidable task.

The MIT Libraries are working with the open-source digital repository platform called DSpace™ to explore the problem of capturing digital research, education material, and publications in any format and preserve them over time; to conduct research and experimentation to learn more about the issues; and to identify what larger, community-based infrastructure is needed by research institutions in order to make digital preservation a practical reality. Working together with researchers from Hewlett Packard and from other research institutions that use the DSpace platform in the United States, the United Kingdom, Europe, and elsewhere, we are beginning to see new ways of managing arbitrary digital content that might make digital preservation an achievable goal. And the emergence of a digital preservation community is helping to educate digital content software producers and authors and the governments that fund research to be more aware of the consequences of current policies and decisions.

The Content Environment

MIT is experiencing what many academic research institutions have noted in the past decade: a rapidly escalating rate of digital data production in every aspect of every activity—research, teaching, and publication on the academic side, and financial and business records on the administrative side. In many cases this digital content is effectively unmanaged, or is managed as current data by the institution's computer center—in other words, it is not under archival control or easily found for reuse by those other than the original creator. And increasingly it is disappearing altogether,

for example, when a graduate student leaves, or a computer is stolen, or a backup regime fails, or a computer platform or piece of software becomes obsolete. It is all too easy for this material to vanish or to lack even the basic metadata needed to manage it over time. Some examples of the digital content in question from the MIT environment include the following:

- Electronic publications in PDF format (articles and preprints, e-theses, technical and working papers, conference proceedings, lab notebooks, etc.)
- Course material, including simulations and visualizations provided as Java applets (that is, program code)
- Images from a range of domains, including digitized fine art photographs and digital images from scientific instruments (for example, MRI scans or confocal microscope images)
- Multimedia archives and publications used for teaching or research in the humanities and social sciences
- Scientific datasets and databases from subjects such as bioinformatics or plasma physics
- Social science statistical and geospatial datasets
- Digitized print collections such as manuscripts and special collections

As you see, this digital content is extremely varied in subject and purpose; exists in a wide range of technical formats, with and without software dependencies; arrives at different distances from the time of creation; and requires very different metadata to describe it both functionally and technically. And these examples do not begin to reflect some of the complexity that can be seen in real life, for example, files with other files (in other formats) embedded in them, groups of files that have been glued together (for example, with the UNIX "tar" program), XML files without documentation for what the markup signifies, and similar conundrums with datasets and databases.

Adding to the complexity of this technical variation in digital content is the current legal and regulatory landscape we live in. All of these works are copyrighted, patented, or otherwise "owned" by some legal entity. In some cases it is the institution that employs the creator, in others it is the creator himself, and in many cases it is some third party (for example, a publisher) to whom the creator has turned over ownership or licensed the patent rights, whether or not that was allowed by local policy. And in many cases the creator does not actually know who controls the material in question. Institutions wishing to archive digital material have the interesting choice of either accepting it with associated risk of copyright or patent violation or requiring material to be cleared of restrictions first with a warrant from the depositor—a surefire way of limiting what is deposited.

It is our observation at MIT that in *some* cases the institution can suggest what file formats are desirable for archiving purposes (for example, use

PDF rather than Microsoft Word), and in *some* cases the archive could convert the file on ingest from a less to a more desirable format (for example, from Microsoft Word or Adobe PDF into PDF/A, the archival profile of the PDF standard). In many other cases, however, neither option is possible. Many faculty authors make format choices based on the practical realities of their own domain—accepted formats in their field or the current state of Web browsers for rendering—and do not have the time, patience, or expertise to convert or re-create their materials in formats more conducive to preservation. At that point it becomes a matter of institutional policy whether the material will be accepted in the archive or not and at what level of preservation support. If a university's president turns up with a laptop filled with her professional and administrative records, do you say no? If a Nobel Laureate scientist appears with a key scientific dataset (the Human Genome database, say) do you refuse it because it is not in ASCII or XML? There are numerous anecdotes from every archival organization along these lines, and clearly we cannot refuse everything that does not meet high standards of creation, especially when there are no absolute guarantees that we can preserve even the best-quality digital material (it could prove to be too expensive, for example). Therefore, at MIT we have decided that it is important to tackle the problem from the perspective of an archive that has to deal with everything and still drive costs down to the point where digital preservation is a practical possibility.

DSpace

This section describes how the MIT Libraries and Archives are using the DSpace platform to begin to tackle the problem of institutional archiving and preservation of digital material produced by the local community of scholars. DSpace[1] is free, open-source software that has the functionality to capture, store, manage, and support preservation of digital objects in any format.

There are two important caveats to everything said here about DSpace. First, what is described here represents MIT's current thinking, policies, and procedures, and not those of other institutions using the DSpace platform. It is our hope that the community of DSpace implementers (known as the DSpace Federation) will collaborate on the problem of digital preservation to develop collective best practices and possibly even policy, but it is early days in the use of DSpace. For now there is wide variety in the ways that different institutions are beginning to think about the technological approaches, costs, risks, and benefits associated with digital preservation and what they are willing to do.

Second, DSpace is software, and as such, it does not *do* digital preservation per se. Preservation is a collection management, or digital life cycle management, activity with a technology component but also associated poli-

cies and procedures. Especially now, when so little is known with certainty about how we will accomplish digital preservation and at what cost, human involvement is critical, and DSpace can merely support our ambitions in the aspects that lend themselves to automation. But I will attempt to describe how the software can help and where it cannot.

Best practice in software development today, especially in areas that are poorly understood like digital archiving and preservation, defines a process by which the system evolves rapidly as our understanding of the problem increases. This is known as "spiral development" (Boehm, 2000), and in practice it means that systems should be designed with modularity in mind and the assumption that the code will all be thrown away and re-created often as understanding evolves. Prototypes are created to try new things, and experimentation is encouraged. The assumption is that any attempt to define a "perfect architecture" for the system that solves the entire problem once and for all is naïve and creates too much risk for the organization that depends on it.

To this end, DSpace was originally created as a breadth-first system (that is, having all the necessary functionality to start using it out of the box, so to speak) with very little depth in any particular aspect of its functionality so that the community of implementers, including MIT, could decide what they need most from it and how it should evolve. The original support for digital archiving and preservation in DSpace was limited. It consisted of the following elements:

1. An internal file format registry with entries for each format known to the institution and what level of support is offered *by policy* for that format at that institution. The three levels of support defined now are

    - "supported"—the organization will make every attempt to preserve the current functionality of the file into the future (what we call "functional preservation");
    - "known"—the organization knows about the format but cannot, or will not, insure preservation over time. Typically this would be the case if the organization might be able to provide preservation support but for some reason cannot promise it. For example, if the format is defined by Microsoft so that it is both proprietary and unpublished, then the organization *might* be able to migrate it to another format in the future using third-party tools but cannot guarantee that such tools will be available;
    - "unsupported"—the organization will preserve the deposited bits but not attempt anything further. Typically, this would be the case for formats that are completely unknown or too expensive to preserve, for example, a compiled, binary program in a programing language that was invented by the author and never documented.

2. Minimal technical metadata for each deposited file (or "bitstream" as they are known in DSpace) consisting of about five pieces of information:
   - A unique identifier for the file in a local namespace
   - The file format, as defined in the internal format registry
   - The file size in bytes
   - A date of deposit to DSpace (or the closest approximation to creation date that is programmatically available)
   - A checksum and checksum type (currently MD5) for the file
3. A "History" file, consisting of metadata stored in Resource Description Framework (RDF) format, that is updated with each significant event in the system (for example, when an item is updated or a bitstream changed). This serves as a sort of digital provenance tracking system to record what happens to a particular deposited item over time.

Clearly these parts of the DSpace infrastructure are necessary but not sufficient to perform digital preservation in any meaningful way. What is missing includes, at a minimum, the following:

1. Batch ingest tools to do basic file quality assurance and technical metadata extraction during the deposit process. This would include things like virus checking, verifying that the files are indeed of the format specified and are viable instances of that format, and extracting all available technical metadata directly from the file using tools like JHOVE.[2]
2. Better support of technical metadata that varies with each file format and changes quickly over time. Currently this metadata can be stored as another file (for example, in XML) along with the deposited files, but that makes it difficult to interact with in the performance of collection management activities. Having tools to interact with this technical metadata will be necessary to support preservation, but these tools cannot be too proscriptive or difficult to change since as a community we still have very little idea of exactly what technical metadata will be necessary to support digital preservation or whether it will prove to be affordable to generate.
3. A more modular system architecture that implements the Open Archival Information System (OAIS) framework[3] more closely, separating the digital asset store from the database that manages it more cleanly. Currently DSpace uses a relational database to manage the asset store, which is implemented on the computer's file system. Ideally the metadata would be packaged together with the content in an OAIS Archival Information Package (AIP), which is also stored in the asset store, with the database serving more as a way of optimizing certain system operations such as lookup and reporting. While OAIS does not require this approach, having self-contained AIPs will make preservation services like replication and distribution of content much safer and easier.

4. A framework to support such useful technical advances as highly distributed storage (to help deal with the petabyte-sized digital objects that are part of the archive) and replication of the content to support redundancy as an alternative backup strategy.[4]
5. Better support for content versioning and proper content life cycle management, with appropriate metadata and access controls to manage which version is for what purpose and who has access to each. Having multiple versions of an item or a bitstream in DSpace is possible now, but it is complex, in part because there are multiple types of "versioning" that are appropriate for digital repositories: different versions in time (for example, new editions), different versions in format (for example, a Microsoft Word file and a PDF), and different versions in quality (for example, an archival master TIFF image and its associated thumbnail). Today, versions in DSpace are managed either by creating separate items and relating them with identifiers stored in the Dublin Core metadata or by adding new files to existing items, which are identified by labels and/or suppressed from public view. And while assigning varying access control for different files attached to an item is possible, no friendly user interface is provided to make it simple to do. Lastly there are issues that arise in records management like destruction of digital content when retention policies call for it and how to certify that as part of the content management process. Certainly for the cases involving preservation masters or other versions of files that are not for public display, better support to make this easy is needed.
6. Better support for emulation in cases where a simple format migration is not practical (for example, for binary formats of things like simulations or video games). In general DSpace assumes that items are either directly viewable in modern Web browsers or should be downloaded to the local computer for manipulation. A third, currently unsupported option would be for DSpace to render the item itself in a manner that allows it to be viewed in a browser. If a software emulator exists for a particular item of content, then it would be desirable for DSpace to run that emulator to show the content rather than forcing the user to download the content and its emulator to the desktop.[5]
7. Automated integrity checking for the digital material on deposit. DSpace stores a checksum for each digital file it manages, and the system should constantly monitor its asset store for file corruption or other problems. Using RAID storage arrays is some help, but problems can go undetected if a file is not read for a long period of time. If a good backup regime has been implemented, then restoring a corrupted file is always possible, so each file in the system should be read and checked against its checksum on a regular schedule. This would also work in conjunction with recommendation 4 above (file replication) if that were implemented.

The plan for how to do all this (and more) in DSpace brings us back to the open-source software development model and the spiral development model. Some of these things could be added quite easily to the current version of the system (for example, the batch tools and the versioning support, both of which are in development now), but some require a major change to the system's architecture. This has led to the idea of a DSpace version 2.0, which will be quite different internally from the version out there today. Hopefully it will not affect the public user interface too drastically, nor the way collection managers interact with the system, but the internal plumbing will be quite different. But creating a new version of DSpace today is quite a different matter than creating the original was, as there are hundreds of institutions relying on the system now and many of them have good ideas to contribute to the 2.0 design and implementation. The open-source software promise is that, by being publicly scrutinized and criticized, the finished product is much, much better than it would be if it were just done by a small number of developers in one place. But the cost is in the complexity and time it takes to do distributed development by many people.

Having said that, there is a move afoot to start work on DSpace 2.0 led by one of the original Hewlett Packard developers and with extensive involvement from other developers at MIT, the University of Cambridge, the Australian National University, and others. They are in the midst of defining a project to work collaboratively to build the 2.0 system, and we hope to see DSpace continue to evolve at this pace over the coming decades.

Other Relevant Initiatives

While the DSpace 2.0 development gets sorted out, we are collaborating actively with two other initiatives that could be of great benefit to DSpace in support of digital preservation (both are described in detail elsewhere in this issue). These include the Storage Resource Broker (SRB)[6] work at the San Diego Supercomputer Center (SDSC) (in collaboration with the University of California, San Diego Libraries), and the Global Digital Format Registry (GDFR)[7] initiative of the Digital Library Federation.

The SDSC SRB is client-server middleware that provides a uniform interface for connecting to heterogeneous data resources over a network and accessing replicated datasets. SRB, in conjunction with the Metadata Catalog (MCAT), provides a way to access datasets and resources based on their attributes and/or logical names rather than their names or physical locations. The project underway will evaluate how DSpace and SRB might be integrated to provide DSpace managers access to managed, distributed, replicated, grid-based storage when desired.

The GDFR initiative is being developed under the auspices of the Digital Library Federation. The initiative has so far developed a data model

for technical format metadata and a set of services and has submitted a grant proposal for further development of the system. A prototype is in development at the University of Pennsylvania Library under the name of FRED (Format REgistry Demonstration) as part of the TOM (Typed Object Model) project.[8]

These are just two of many initiatives, too many to mention individually, that we are paying close attention to as models for where the DSpace community needs to go, what works, what does not work, what the costs are, how to assess risk, how to collaborate, and so on.

Conclusion

The success of DSpace as a digital archiving and preservation platform depends on several things: our ability as a community to rapidly evolve the system as we learn more about how to do digital preservation; our ability to educate the library and archive professionals responsible for collection management and preservation today to be able to cope with digital material alongside the print material; and our ability to learn from each other and coordinate the many useful initiatives that are underway in this area. If there is one thing we do know with certainty it is that this problem is far too big and complex to be solved by any one organization, system, or preservation strategy. We need to collaborate and share, to build common infrastructure where it makes sense, and to support alternative models of archiving. Biodiversity is good, monocultures are bad, and none of us really knows what is going to happen in the future. But stay tuned.

Notes

1. DSpace is described in detail at the DSpace Federation Web site, http://dspace.org. Tansley, Bass, and Smith (2003) describe it in the digital archive context.
2. JHOVE was created by staff at the Harvard University Library to ingest various file formats that they are archiving (see http://hul.harvard.edu/jhove/jhove.html), and the New Zealand National Library is working on similar tools for a digital archiving program.
3. The OAIS reference model was developed by the Consultative Committee for Space Data Systems (http://ssdoo.gsfc.nasa.gov/nost/wwwclassic/documents/pdf/CCSDS-650.0-B-1.pdf); it defines the components of a digital archive and the functions each component supports, as well as an abstract data and metadata model for digital content.
4. Currently, the most famous example of this comes from the LOCKSS project at Stanford University (http://lockss.stanford.edu/), which uses a combination of Web harvesting of published e-journals and content replication among LOCKSS sites to help insure preservation of the bits over time, no matter what happens to an individual archive. This is widely considered to be a good approach to bit-preservation since it allows for the possibility of a single archive being abandoned or destroyed.
5. This is the approach taken by the Universal Virtual Computer (UVC) project, originally developed by researchers at IBM. A good article describing the approach is Lorie (2001).
6. The Storage Resource Broker is a system developed by the San Diego Supercomputer Center as part of their research engagement with the National Science Foundation National Partnership for Advanced Computational Infrastructure (http://www.npaci.edu/DICE/SRB/).
7. For more information on GDFR, see the project Web site at http://hul.harvard.edu/gdfr/.
8. The TOM project is described at http://tom.library.upenn.edu/, and http://tom.library.upenn.edu/fred/ hosts the format registry prototype.

References

Boehm, B. (2000). *Spiral development: Experience, principles, and refinements.* Presented at the Spiral Development Workshop, February 9, 2000. Reprinted in W. J. Hansen (Ed.), *Special report CMU/SEI-00-SR-08, ESC-SR-00-08, June, 2000.* Retrieved February 7, 2005, from http://www.sei.cmu.edu/cbs/spiral2000/february2000/BoehmSR.html.

Lorie, R. A. (2001). A project on preservation of digital data. *RLG DigiNews, 5*(3). Retrieved February 7, 2005, from http://www.rlg.org/preserv/diginews/diginews5-3.html#feature2.

Tansley, R., Bass, M., & Smith, M. (2003). DSpace as an Open Archival Information System: Current status and future directions. In T. Koch & I. Solvberg (Eds.), *Research and advanced technology for digital libraries: 7th European Conference on Digital Libraries, ECDL 2003, Trondheim, Norway, August 17–22, 2003* (pp. 446–60). London: Springer-Verlag.

Varian, H., & Lyman, P. (2003). *How much information 2003?* Retrieved February 7, 2005, from http://www.sims.berkeley.edu/research/projects/how-much-info-2003/.

MacKenzie Smith, Associate Director for Technology, MIT Libraries, 77 Massachusetts Avenue, 14S-308, Cambridge, MA 02139, kenzie@mit.edu. MacKenzie Smith is the Associate Director for Technology at the MIT Libraries, where she oversees the Libraries' use of technology and its digital library research program. She is currently acting as the project director for DSpace, MIT's collaboration with Hewlett-Packard Labs to develop an open source digital repository for scholarly research material in digital formats. She was formerly the Digital Library Program Manager in the Harvard University Library's Office for Information Systems, where she managed the design and implementation of the Library Digital Initiative, and she has also held positions in the library IT departments at Harvard and the University of Chicago. Her research interests are in applied technology for libraries and academia, in particular digital libraries and archives.