

BUILDING A DIGITAL PRESERVATION ARCHIVE: TALES FROM THE FRONT

Priscilla Caplan

First published in *VINE: The Journal of Information and Knowledge Management Systems* Volume 34 Number 1, 2004, pp 38-42. Emerald Group Publishing Limited, ISSN 0305-5728, DOI 10.1108/03055720410530988.

Author information: Priscilla Caplan (pcaplan@ufl.edu) is Assistant Director for Digital Library Services at the Florida Center for Library Automation.

Keywords: Digital Preservation; Preservation Repositories; Florida Center for Library Automation

Abstract: This article describes the evolution of the design of the FCLA Digital Archive, a preservation repository under development for the libraries of the public universities of Florida. The starting assumptions of the designers were challenged as they moved from theory towards implementation. The logic leading to changes in policy and in preservation strategies is described.

Word count approx 3350 without abstract.

BUILDING A DIGITAL PRESERVATION ARCHIVE: TALES FROM THE FRONT

Priscilla Caplan

The Florida Center for Library Automation was established in 1985 to provide technology support to the libraries of the ten public universities of the State of Florida. Initially our main role was to run a central installation of an integrated library system used by all ten university libraries. By the 1990s, however, FCLA services expanded to support the delivery of electronic resources by supporting consortial purchasing and local loading of some indexes and full text journals. Most recently, FCLA has been helping the libraries create and manage their own collections of digital text, images and other media. FCLA systems provide access to electronic theses and dissertations (ETDs) and other born-digital content. FCLA also initiated the PALMM (Publication of Archival, Library and Museum Materials) program to encourage the local digitization of source materials and the collaborative building of web-accessible collections. There are now more than a dozen PALMM collections, ranging in focus from Floridiana to herbarium specimens.

It is a natural step to go from thinking about building digital collections to thinking about preserving them. Building a preservation archive seemed a logical extension of FCLA's mission as a central support organization devoted to building the technical infrastructure and services needed by a set of distributed, independent libraries. An archive developed and maintained by FCLA could give the libraries a trusted and (hopefully) cost-effective alternative to building their own archiving facilities or contracting with commercial services. Because the library directors would also serve as the advisory board of the

archive, this would also give the libraries more control over the archive's policies and services than they were likely to achieve with other options.

FCLA applied for and received a grant from the U.S. Institute of Museum and Library Services (IMLS) to develop a facility called the FCLA Digital Archive (FDA). Our original plan was relatively simple. Our preservation strategy would be based primarily on forward migration. When a file format was in danger of becoming obsolete, a new version of every object in that format would be created in some more current successor format. The original source file would be discarded and the new version would be retained and (if necessary) migrated to the next successor format. In some cases we might also normalize a source file, defined as making a version of it in a format considered more stable. In that case both the original source version and the normalized version would be migrated forward for as long as possible.

A key feature of the plan was for FDA staff to develop an "action plan" for every digital file format we expected to receive. The action plan would specify whether and how we would normalize the format, when we anticipated forward migration, and when the plan should be reviewed. The FDA would only accept formats for which an action plan was in place. Libraries within the state university system would send the FDA materials in accepted formats, and the FDA would ingest them, store them securely on tape, reformat them if necessary, and give them back on request. We proposed that use of the archive would be free for at least three years while we measured costs and developed algorithms for charging, after which we would institute cost-recovery pricing.

The most controversial feature of the project as planned was the intent to separate archiving from access and build a "dark archive" on tape. Current wisdom did and still does call for integrating preservation and access, on the theory that only use itself can ensure that archived copies remain usable. Most of the PALMM projects, however, created archival TIFF masters and derived from these services copies in other formats. We assumed the libraries would only want to send us, and pay for, storage of the archival TIFFs, and we had no system in place designed to serve TIFF images up to end users. Moreover, different campuses might use different applications for display and navigation of complex objects like page image books, and we had no desire to try to run all of the necessary software centrally. We saw no choice but to develop a dark archive and plan to use sampling and other quality control mechanisms to ensure the usability of content. Whatever our reluctance to make this decision, we found a number of immediate benefits, from being able to use cheap tape storage to not having to worry about whole classes of access control and intellectual property questions.

At the time of this writing, the FDA has been under development for about a year, and the "darkness" of the archive is one of the few aspects of the original plan that is still in place. Action plans, ingest assumptions, and preservation strategies have all been reconsidered and modified in some respects. This article describes some of the areas where we deviated from the original plan and the reasons why. We hope our description of some of the logic going into the functional design of the FDA will invite comparison

with other preservation archiving projects and help increase overall our understanding of different models for digital archives.

One of the significant features of the FDA is the assumption that libraries will decide what they want to archive based on their own evaluation of the costs of archiving and the value of the material. This has at least two important implications. First, the FDA has to be able to deal with a wide range of heterogeneous materials. Some facilities may be limited to archiving ejournal articles formatted according to a finite number of DTDs, or to relatively homogenous publications or web pages. The FDA, however, must be able to anticipate what the libraries are likely to want to archive and accept these materials. Second, the FDA has to assume that the library staff submitting materials to the archive will not be computer professionals or specialists in digital preservation. We needed to accept the fact that there would be errors in submissions and make the ingest system as forgiving as possible. For most of FCLA's projects, FCLA staff and library staff work as partners to deliver services. In respect to the FDA, however, the libraries are essentially our customers, and customer service has to be an important consideration in the design of the archive.

The IMLS grant allowed us to hire a file formats specialist who is writing action plans, format by format, for the archive. However, it became clear early on that we could not simply reject materials for which no action plan existed. Most of the materials we expect to receive from the libraries are compound objects, materials that are not complete in a single digital file. A book, for example, may be comprised of several page images and a file of structural metadata showing how they should be assembled together. A report could be represented as an XML document with embedded links to image format illustrations. An electronic thesis (ETD) might have a main PDF file and one or more associated supporting files in other formats. Although preservation treatment is determined at the level of the digital file, the FDA's customers are going to think in terms of intellectual entities like books, reports and dissertations. It would not be acceptable for us to archive only some parts of a submission simply because we hadn't gotten around to writing action plans for all of its files yet. By the same token, we did not want to reject the entire submission. The only solution that was friendly to the submitting libraries was to accept all file formats they submitted.

Once we decided to accept all formats, however, we realized we could not guarantee that we would have an action plan for each archived format ready in time to do forward migration on it when necessary. This led us to offer two levels of preservation treatment, bit-level and full. Bit-level preservation means we will keep a file safe through secure storage and backup, media refreshment, media migration, and data security measures. Full preservation includes bit-level preservation and a promise to do file migration and/or normalization when called for. Full preservation treatment will be available only for formats with completed action plans.

Having been led to develop a bit-level level of service, it made sense to offer this as an option to the libraries, even for formats with existing action plans. Although charging is not in place yet, it is reasonable to assume bit-level treatment will cost less than full

preservation, and so may be desirable for certain materials. Therefore, for every format identified, we decided to allow a library to specify whether it wants bit-level or full preservation treatment for its submitted files of that format. This is not without complications however. If, for example, a library requests bit-level preservation for a particular format and later changes its mind, should full preservation be applied retroactively to all of the files already accumulated in that format? If so, billing algorithms will have to be designed to discourage libraries from changing to full preservation level after many years of discount bit-level treatment. If not, the archive will have to keep track of and implement preservation policies by date ranges.

The heart of our full preservation treatment, forward migration, is also not as straightforward as it seemed. The process as initially envisioned had five major steps: writing a program to convert from outdated file format A to successor format B (say, from JPEG to JPEG2000), identifying all files of format A in the archive, running the program on the identified files creating a new set of files in format B, examining or sampling the output for quality assurance, and deleting files of format A.

The first variation we considered was whether to do migration on request instead of mass migration. Migration on request is a technique pioneered by the CAMiLEON project of the Universities of Michigan and Leeds. Under migration on request, the original bytestream is always preserved, along with a tool for migrating the object at the time of use. Translated to our environment, a program to convert from A to B would be written, but it would be run only when a particular file in format A was requested from the archive. Then the transformation would be performed and the file disseminated in format B. One advantage of this approach is that the archive is spared the work of converting large numbers of files “just in case.” It may turn out that most files are not requested until there is successor format C, in which case the majority can be spared the conversion to B and transformed directly from their original format to C. A second advantage is that, since transformations are always performed on the original, there is no opportunity for errors introduced in earlier migrations to propagate through later migrations.

Migration on request moves some of the preservation effort from conversion of files (bytestreams) to program maintenance, because instead of being run once and thrown away, the conversion tool must be maintained over time. Imagine a very simple case, where digital file format A is superseded by formats B, C and D at times T1, T2 and T3. The archive receives files in formats A, B and C at various points in time. With mass migration, program P1 converts all files of type A to B at T1; program P2 converts all files of type B to C at T2; and program P3 converts all files of type C to D at T3. Three programs must be written, and programs P1, P2 and P3 must be retained only as long as the archive continues to receive new files in these obsolete formats. Assuming all new files archived are in format D, none of the programs must be retained.

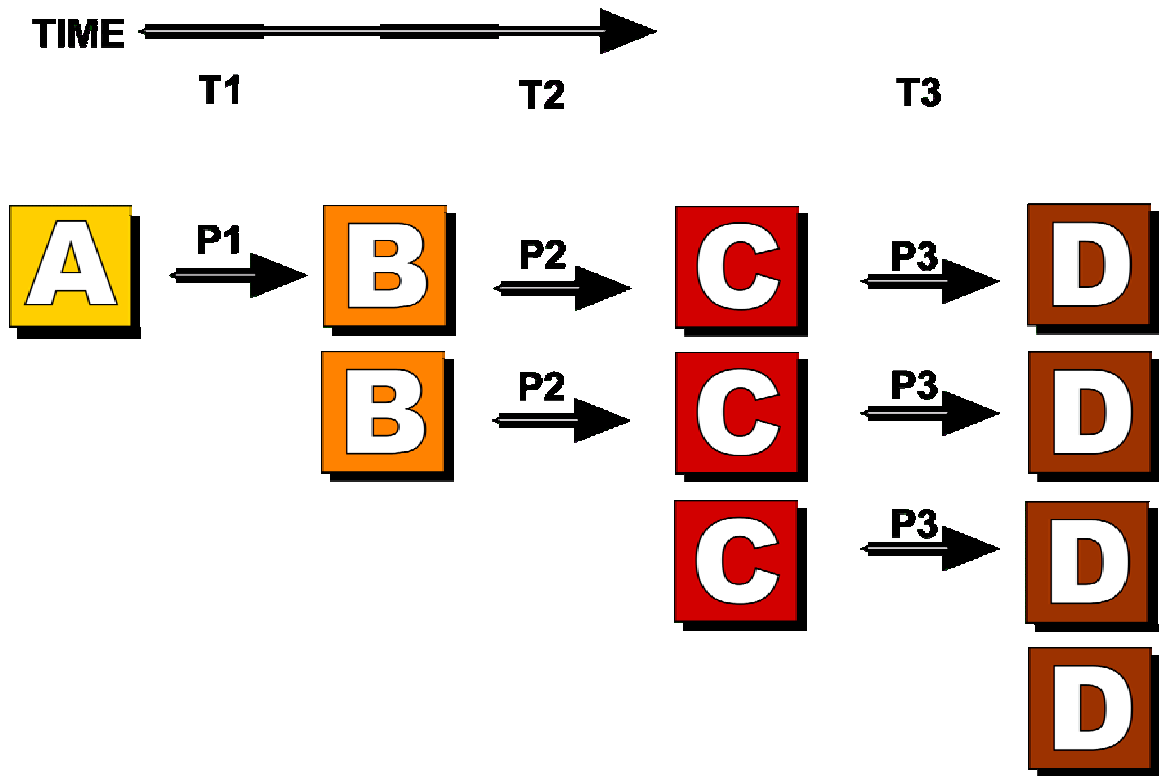


Figure 1.

With migration on request, program P1 converts A to B at T1, program P2 converts A to C at T2, program P3 converts B to C at T2, and so on. By T3, a total of 6 programs have been written and three of them must be retained (A to D, B to D, and C to D). If the elapsed time from T1 to T3 is long, it is possible these programs may have to be rewritten for different computing platforms.

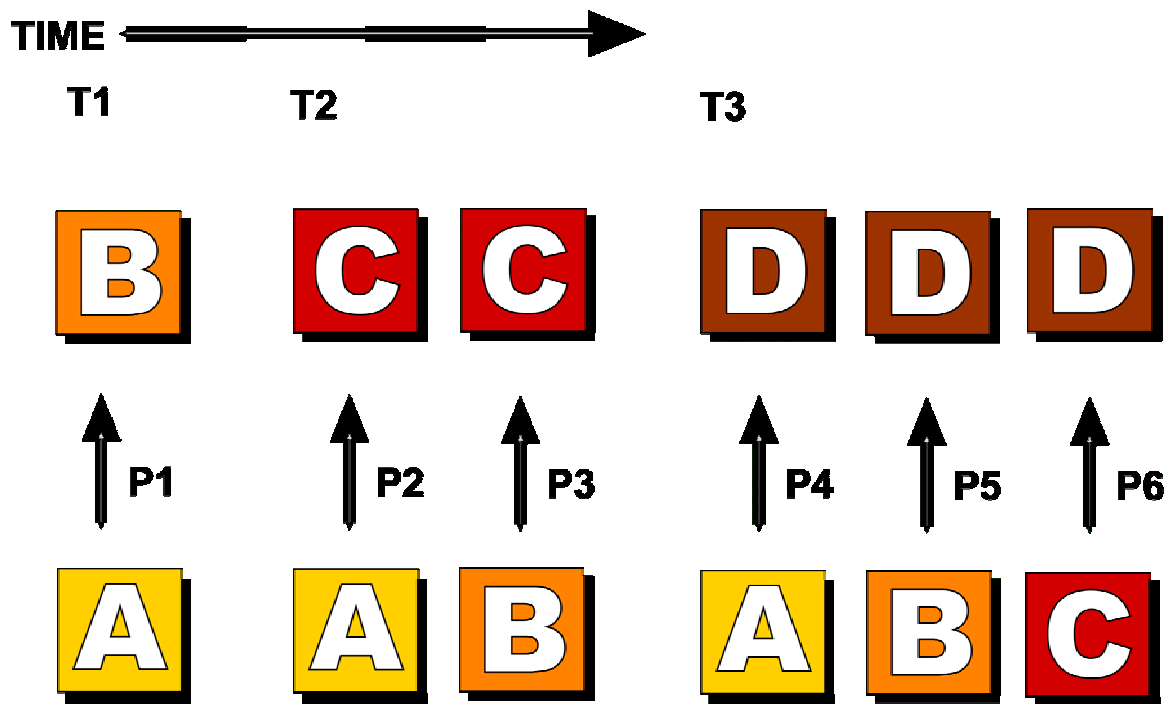


Figure 2

While this discussion was in process, we decided for independent reasons to do bit-level preservation of the original object regardless of preservation actions taken in relation to it. The original plan called for a source object to be deleted after forward migration: if we migrated object 1 in format A to object 2 in format B, we would delete object 1, and if we migrated object 2 in format B to object 3 in format C, we would delete object 2. Under the new plan, if object 1 was the original object as submitted to us for archiving, we would retain object 1 along with object 3 in the above scenario, although we might still delete object 2. The reason for this was conservatism. If a problem was found after the fact in any migration, we would still have the source object with which to repeat the transformation. Also, there is a real possibility that transformation techniques will improve over time; preserving the original will let digital archeologists use any special techniques they can muster.

The decision to retain the original changed the dynamics of the comparison between mass migration and migration on request. Avoiding the propagation of error by doing all transformations from the original was a major rational for migration on request. The same effect, however, can be achieved with mass migration if the original is retained. The scenario for mass migration now looks very much like that for migration on request, with 6 programs being written and three of them retained.

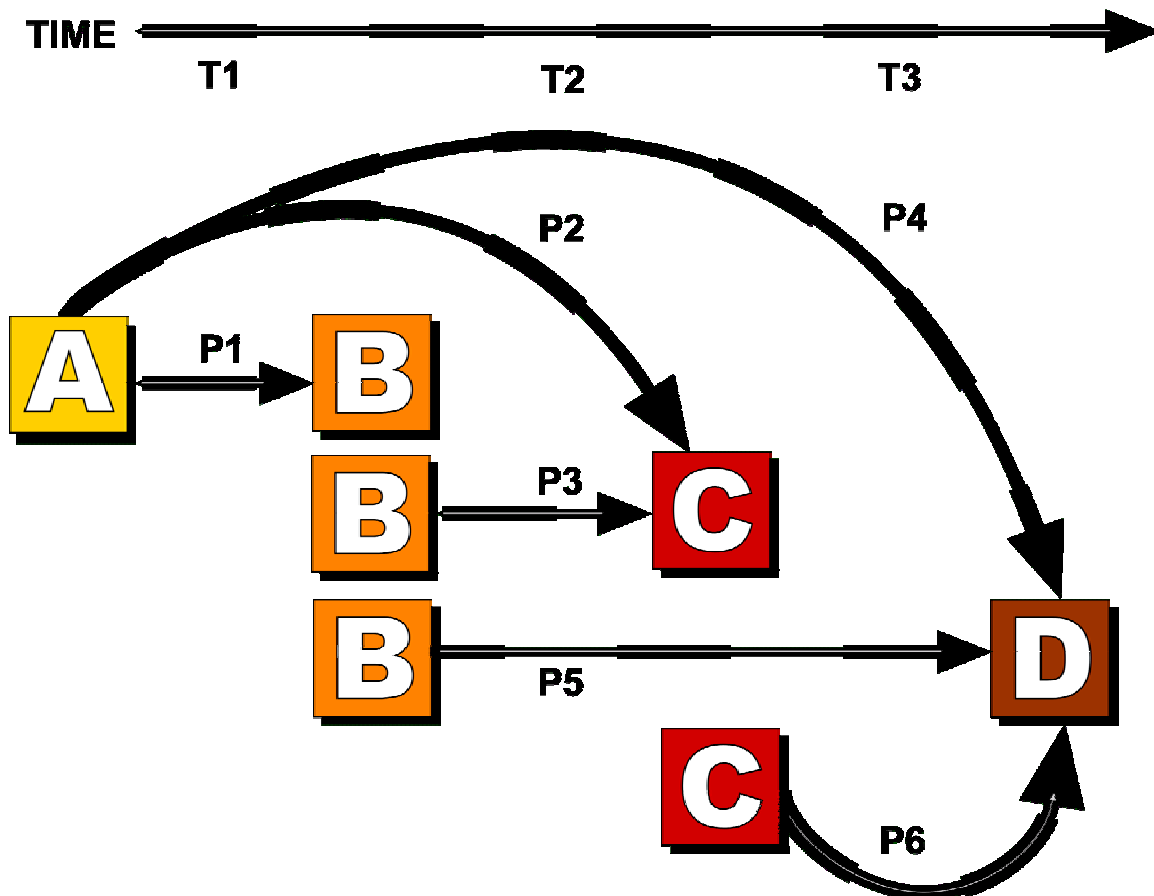


Figure 3.

Having arrived at this point, we concluded that the choice between migration on request and mass migration is more of a tactical decision than a strategic one. Initially, we may prefer mass migration to ensure we have the capacity, and to feel confident that our migration programs are written to handle all objects in the archive. For some objects and formats, we may prefer migration on request. Similarly, the decision whether to retain intermediate versions beyond the source will be made on a format by format basis, as part of the action plan for that format.

Our plans for normalization also changed. Originally, we thought we would create normalized versions of files only for certain file formats considered to be poor candidates for forward migration. This might be because the format was proprietary, had a history of changing frequently, was written for a lesser-used computing platform, or included executable code. An Excel spreadsheet, for example, might be normalized on ingest into some more generic data format that might require less frequent forward migrations, even as we attempted to move the original Excel file forward.

Over time we came to regard normalization less as exceptional treatment and more as a routine way of “hedging our bets,” increasing the chances that the intellectual content of an object will survive. The more we investigated format migration, the more we realized how little we know about it, and how little experience there is in the community as a whole. Under these circumstances, we feel that the more versions of an object we preserve, the greater the odds are for the objects’ long-term usability. Our current strategy is to make a normalized version of any format that can be reasonably (even if not losslessly) represented in another format that also has an action plan. PDF, for example, is considered by some to be a good archival format because there are so many PDF files in use, the development of a robust forward migration path is probable. Nonetheless, we will normalize any PDF file received into page-image TIFFs, thus providing two paths for the content’s survival. We also routinely normalize XML, HTML and any other format that can contain embedded links, replacing the address in the link with the identifier of the linked-to object in the archive.

Our overall preservation strategy means that there may in some cases be many versions of submitted materials maintained in the archive: the original submitted version, one or more migrated versions, a normalized version, and one or more migrated versions of the normalized version. It is probable that our cost recovery pricing when instituted will have some component for storage, and that the “invisible” maintenance of multiple versions will increase costs to our customers. As we gain more experience we may become less conservative and opt for retaining fewer versions. The library directors, functioning as the advisory board to the FDA, will doubtless have some influence in determining the proper balance between minimizing costs and minimizing risk. In the absence of experience, however, starting off with this conservative strategy seems reasonable.

Another design area that required considerable discussion was whether to alter incoming “packages” and if so to what extent. In the Open Archival Information System (OAIS) framework, the content, metadata and packaging submitted to an archive is called the Submission Information Package, or SIP. Initially, we imagined a rather simple model, in which libraries would send us SIPs and we would archive the contents. Each SIP would include an XML descriptor in the Metadata Encoding and Transmission Standard (METS) format listing each file included along with whatever metadata was available to describe the file and its relation to other files. Ingest routines would examine the descriptor and the other files in the SIP in order to store metadata about them in the archive management system, but we would not alter the content of either the descriptor or the files.

This plan proved insufficient when we set up our ingest routines to validate XML documents against their schema or DTDs. We found that in most cases libraries were not providing all of the referenced schema in their SIPs. Even if, for example, they were providing the METS schema, they were unlikely to submit all the extension schema it referenced, or the schema that these schema referenced. Rather than expecting the libraries to have the expertise to identify and copy all of the schemas in use, or alternatively rejecting quantities of submitted SIPs, we wrote our ingest routines to find and download any referenced DTD or schema files that are not included in the SIP before performing validation. At that point, we realized we should also include the downloaded files in the package being archived. Because this changes the contents of the SIP, a new descriptor is created that references the files added by the FDA as well as the original files.

This led to a second problem, that as the archive grew, schema used by METS (and possibly other common XML documents) would be downloaded and stored hundreds of thousands of times. To obviate this, we created a small set of “global” files that are stored only once by the archive but can be referenced by any number of packages. If a SIP contains a schema on our list of global files, we compare its checksum to the checksum of the stored global schema to ensure they are identical. If so, we don’t archive the submitted schema and change the file reference in the descriptor to address the global file.

The discussion above is only a sampling of areas where our original assumptions were challenged by the reality of actually designing and implementing a working preservation archive. We fully expect that when the FDA goes into full production sometime in 2004 we will discover a whole new set of circumstances that will require rethinking and redesign of our policies, procedures, and technical implementations. Building the FCLA Digital Archive sometimes seems less like a development project than a journey of discovery. The path twists and turns, and it is not always easy to tell if we are going in the right direction. The goal remains to provide trusted preservation services that the libraries of the public universities of Florida can rely upon to keep their valuable digital materials usable in some form for as long as possible.

(The author thanks Christopher Vicary and Andrea Goethals for their contributions to this paper.)