

**The Art Institute of Chicago
Department of Architecture**

Collecting, Archiving and Exhibiting Digital Design Data

Section 3: Implementation

Starting a Digital Collection

The *Archiving Digital Design Data: Practices and Technology* section documents the steps required to create and maintain a digital collection:

- Preparing
- Collecting and processing
- Cataloging
- Storing
- Preserving
- Providing access.

The next question is: How to start such a digital collection?

The major areas of consideration when adding digital objects to an institution's collection are:

- Institutional commitment
- Hardware/software/communications infrastructure
- Transition and training.

Institutional Commitment

There is a necessary level of institutional commitment:

- The administration and funders must understand that, in addition to the initial system cost, they are making a long-term commitment to maintaining, expanding and periodically updating the hardware and software infrastructure supporting the digital archives.
- The institution's information services department must provide the skills and staff level to set up, test and manage adequately the hardware/software infrastructure and also to interface the data repository with any existing collection management and public access systems.
- The institution must provide appropriate training and tools to registrars, curators and archivists to allow them to perform their roles appropriately in relation to the digital content.
- The institution's counsel, registrar, curators, archivists and information services personnel must establish and implement policies on intellectual property, security and privacy specific to the digital collection.
- The institution must create a mechanism for monitoring technology that will impact the characteristics of the digital material submitted to the archive, attitudes about appropriate content, the tools available for accessing the data and obsolescence of data in the archive. The recommendation is to establish a Preservation Policy Committee, with representatives from the registrar's office, the curatorial and/or archive department and information services.

Hardware/Software/Communications Infrastructure

The systems required for the digital collection must support the following functions:

- Receive, review and process digital submissions
- Catalog and search the digital collection
- Store and preserve the digital collection
- Make the digital collection available online to various audiences.

Each of these functions is described in more detail below.

Receiving, Reviewing and Processing

The design office will prepare the submission of digital data, possibly with the assistance of a curator or archivist from the institution. The submission will be transferred to the institution via either an Internet connection or suitable media such as CD or DVD. It should be noted that email is a poor choice for most transfers, due the quantity of data to be transferred. Rather, the institution should make available some type of FTP (file transfer protocol) site for these transfers, if Internet submissions are to be accommodated.

Starting a Digital Collection

As discussed previously in this report, the registrar, the curator and the archivist may all be participating in the evaluation of the technical suitability of digital design data: their file formats, naming and organization. Someone will be creating the thumbnails and low-resolution derivative images that will be made available to the public. Metadata for the submission and its component objects will be entered into the collection management system. The intellectual property provisions of the Deed of Gift will be recorded. The electronic provenance of the files—format, software version, operating system and so forth, must also be recorded so that appropriate preservation strategies can be applied.

Those participating in this process will need a system where incoming submissions can be staged, validated, viewed, manipulated and cataloged. A high-end desktop system with large memory and storage capacity will suffice. This system will need drives to read the media types, such as CD or DVD, on which submissions will be received, as well as high-speed connections to the FTP site for Internet submissions and to the collection management system, the repository and the format registry. Adobe Photoshop and Acrobat software will be needed to create derivative images and compilations for electronic delivery to the public. This ingest function will also need a multiformat viewer to access the native data. There is a broad range of viewing products described in the Accessing Digital Design Data chapter of the Archiving Digital Design Data: Practices and Technology section. For certain collections, the institution may want to solicit a workstation equipped with the software in which the data were created. This should not be seen as a solution for long-term access, rather as a tool for the initial creation of the Archival Information Package (AIP).

Cataloging and Searching

It is highly likely that an institution planning to implement a digital archive will already have in place a collection management system used for cataloging, searching and retrieving information about the physical collection. This system should be extended to the digital collection if at all possible to provide a unified and familiar interface for searching the institution's entire collection.

This front-end collection management system provides the following functions:

- Stores and manages a database of descriptive and administrative metadata
- Provides a search function
- Interfaces with the repository system.

In order to ensure that an existing collection management system is capable of cataloging digital architectural works, its metadata schema must comply with the following requirements:

- Extensible to include descriptive metadata fields necessary to catalog digital architectural works
- Customizable administrative metadata fields to track preservation actions taken on the digital collection
- Capable of a metadata hierarchy that allows information to be recorded at both a master level and at a component level (or at an architectural job level and individual document level).

The metadata schema recommended for the collection management system is the Categories for the Description of Works of Art (CDWA), which complies with the above requirements.

Storing and Preserving

The back-end data repository provides the following capabilities:

- Store each item of digital content and return a unique ID to the collection management system
- Extract and maintain technical metadata such as checksum, format and version information
- Ensure bitstream preservation of the digital items
- Provide tools to support functional preservation of the data
- Maintain or link to a format registry to track file formats, versions and associated preservation strategies.

The data repository system recommended is DSpace. To make the digital collection searchable across multiple institutions, DSpace can be configured in a federated model with other institutions and can be used as a second point of access to the digital collection. To enable researchers to search using DSpace, some fields of metadata associated with the digital content must be passed from the collection management system to DSpace and stored in the format of its metadata schema: Dublin Core. The Dublin Core was designed to be the "least common denominator" of metadata schemas with its 15 basic cataloging fields and to allow for discovery of content across digital repositories worldwide through what is called the Open Archives Initiative (OAI). The

appropriate fields of CDWA metadata should be copied to Dublin Core format by an automated programmed routine.

Making the Collection Available Online

Most institutions already have robust, publicly accessible Web sites that provide access to a variety of educational resources. Hopefully, the institution has made parts of its collection management system Web-accessible so that the public can search the descriptive metadata for items, physical or digital, in the collection. An OAI-compliant repository, such as DSpace, will make a subset of that metadata accessible to cross-institutional Web searches.

If the institution does not yet have a public Web presence, it will require a Web server with software capable of searching the digital collection and presenting the graphic material identified through such a search. Searching and presenting are really two separate functions, unless digital images are actually embedded in the collection management database. In step one, the user interface permits the entry of the search criteria and returns descriptive metadata for each item retrieved through the search. In the second step, the user interface permits the selection of one of the items described, retrieves appropriate content from the repository system and presents it to the user.

There are a number of policy and business issues to be considered when providing public access. Most institutions with a database of scanned images of the collection allow low-resolution images to be accessed from the Web site but continue to require payment for use of high-resolution images. The Art Institute of Chicago's Department of Architecture receives requests for copies of working drawings in their archives from developers and architects when historic structures are being renovated, and they charge for the copies. The Web server could actually include e-commerce capabilities that would allow the public to order digital data or printed copies and pay by credit card. Protecting intellectual property is a major consideration if digital data are delivered to the public. Particularly with native data, there is really no way to control how they are used once delivered.

The two-tier digital archive approach described throughout this report makes Web viewing of the output data housed in the repository extremely simple. Although uncompressed TIFF images are too large for Web viewing, low-resolution derivative images for quick browser viewing are created during ingest in the JPG format. No additional software is required and these images would have too low resolution for publication or commercial use. The PDF content, which includes CAD drawings, PowerPoint presentations, animations and any non-graphic materials, can be viewed with the free Adobe Reader that is pre-installed on most new desktop computers or can be downloaded from the Internet. The PDF format can be published as non-editable, so that the content is protected.

Files in native format, however, will not be so easily viewable. The institution may decide to not make native data Web-available or to install Web server-based multi-format viewing software to provide access to a portion of the native data. However, the need to do this may be obviated in the near future by emerging output types that provide navigable 3D models in formats that can be read by free viewers.

Transition and Training

As most businesses have discovered, moving operations into the electronic environment requires not only new tools, but new roles, responsibilities and skills. Many work processes must change and the time and effort devoted to each step in a process may change as well. In the course of this study, The Art Institute of Chicago Department of Architecture's current archiving process was compared to that proposed for the digital archive. Although the digital archiving process required more initial effort, the downstream activities of retrieving and repurposing the digital collection for exhibition, publication or use by the public were almost effortless.

Preparing the Institution

It will be up to each institution to determine who should shoulder which responsibilities specific to the digital collection and make sure that the individuals tasked with these new responsibilities receive appropriate tools and adequate training to fulfill them.

Receiving, processing and cataloging new digital submissions will require a conceptual understanding of the format, color, metadata, intellectual property and preservation issues, as well as training in a number of

Starting a Digital Collection

software programs. These programs include, initially, the institution's collection management system, DSpace, Photoshop and Acrobat. Although the registrars, curators and archivists will soon become expert with these tools, the institution should consider hiring a skilled technician for the transition phase. The institution must also understand that, as design practice evolves over time, its changing tools will demand new methods and tools for the ingest function. Ongoing training will be a requirement.

Educating Design Firms

The primary responsibility for creating archivable design data lies with the design firm. It will be necessary to communicate to potential donors the best practices for formatting, naming and maintaining their data. These are described in the *Preparing Digital Design Data* chapter in the *Archiving Digital Design Data: Practices and Technology* section. Firms must invest in informing and training staff to follow these best practices. Staff should be educated about image resolution, format and color profile and know how to manipulate these in a program such as Adobe Photoshop.

In order to adhere to the recommended practices, design firms need little additional hardware or software. Maintaining checkpoint data sets will increase storage requirements, but online storage is inexpensive and continues to drop in price. According to the survey conducted as part of this study, most firms already use Photoshop. Some firms may need to acquire software for creating PDF documents from CAD drawings, PowerPoint presentations, animations, and so forth. Options include Adobe Acrobat and a number of third party offerings.

Resource Requirements

This chapter describes the hardware, software and personnel requirements for initial implementation of a long-term digital archive.

Programming Requirements

The first step in implementing a digital archive is a programming effort to create or customize the collection management system to accommodate the digital collection.

Maintaining a digital collection will require preservation actions be taken, so preservation should be tracked as metadata including the type of preservation action taken (such as an upgrade to a newer software version or conversion from one format to another) as well as the date and name of the person carrying out the action.

To accommodate the hierarchical nature of design documentation, the data structure of the collection management system may need to be reconfigured. It should include a Master/Components data structure to accommodate a *Project* → *Document Group* → *Individual Document* cataloging hierarchy. The best implementation of such a metadata hierarchy is for information to be “inherited” from the master level to the component level to minimize redundancy in metadata entry. The Art Institute of Chicago’s CITI collection management system has a planned inheritance tool for this purpose.

The second step in implementing a digital archive is to integrate the back-end data repository with the front-end collection management system. This involves, at a minimum:

- Creating a procedure to link the front-end and back-end systems so that the unique ID assigned to the digital document by the back-end DSpace is returned to the front-end database and stored, allowing digital documents to be retrieved
- Creating a procedure to map the metadata of a digital document from CDWA (or other front-end metadata schema) to Dublin Core to be stored in DSpace.

Software Requirements

DSpace software can be freely downloaded from SourceForge.net, subject to terms of the BSD distribution license. DSpace is built on free, open-source tools such as Apache Web server, the Tomcat Servlet engine and the PostgreSQL relational database system. It is possible to modify DSpace to work with other Web servers and database software. DSpace packages JDBC (Java Database Connectivity) and other drivers and libraries with the DSpace download.

Server and Data Storage Requirements for Repository

DSpace requires a UNIX operating system and runs on everything from a PC to high-end server configurations. Possible high-end server options are listed below (taken from the DSpace Web site). By the time this report is issued, these specifications will be out of date. However, they provide an indication of the range of costs and vendors.

- HP Server rx2600, powered by dual 64-bit Intel Itanium 2 processors (900MHz), 2GB RAM, 26 GB internal disk storage. HP StorageWorks Modular SAN Array 1000 (msa1000) with a single high-performance controller. Options include a second controller and, with the addition of two more drive enclosures, controls up to 42 Ultra2, Ultra3, or Ultra320 SCSI drives. Total capacity can be six terabytes. Cost starts around \$40K and goes up to around \$1.8M
- Sun Fire 280R Server, two 900MHz UltraSPARC-III Cu processors, 8MB E-cache, 2GB memory, two 36GB 10,000rpm HH internal FCAL disk drives, DVD, 436-GB, or 12 x 26.4 Gbyte 10K RPM disks, Sun StorEdge A1000 rackmountable w/ 1 HW RAID controller, 24MB std cache. Around \$30K.
- Dell PowerEdge 2650 with dual Xeon processors (2.4GHz), 2GB RAM, 2x73GB SCSI disks. One 2.5TB Apple XServe. A DLT tape library for back up, etc. Around \$10K.

Because of the types of data being archived, submissions will be very large. For example, the PowerPoint presentations collected for the case studies documented in this report were as large as 50 megabytes each.

Resource Requirements

Full archival documentation of a project, including images suitable for print publication, could easily reach one gigabyte of data. In addition, as preservation policies are applied, data will be duplicated: the original files will be preserved and new, updated versions created for functional preservation. The institution therefore should select servers understanding that they may need to store terabytes (1 terabyte = 2^{10} gigabytes, or approximately 1,000,000 megabytes) of data, although storage space may be added incrementally as needed.

Data Maintenance Requirements

- Adequate backup and failover systems
- Disaster recovery plan to house a copy of the data at an offsite location. One possibility is to store duplicate data with another museum or institution to avoid costly commercial disaster recovery options.

Personnel Requirements

- Computer programmer with experience in Java, HTML/Java Server Pages and SQL databases, to customize the DSpace software, as well as other relevant experience to customize the existing collection management system
- Someone from the information services department to setup and configure hardware, data storage and data maintenance systems, and to maintain the system on a day-to-day basis.

Web Server Requirements

The Web server for public access to the digital collection should be a different server from the repository server, and it would probably be providing access to larger collection and institutional resources than just the digital data. Sizing of Web servers is driven by traffic: the number of visitors to the site, the duration of their visits, the content accessed and downloaded and the number of simultaneous visitors at peak times. This aspect is beyond the scope of this study.