Final Report for the MIT FACADE Project: October 2006 – August 2009

Introduction

Since the introduction of Computer-Aided Design (CAD) software in the 1960s, industries that design and develop our built environment have been trading pencil and paper for computers and digital files. The earliest adopters of the new technology were industries like aerospace and automotive, but since then the fields of architecture and design have been enthusiastic adopters. CAD has allowed architects to take previously unimaginable risks in their designs, and to experiment with new forms and materials without the need of building prototypes or performing expensive structural analyses until much later in the process. U.S. architects like Frank Gehry led the way, and institutions like MIT have provided the educational programs and technical expertise to marry architecture, engineering, software design.

That this has led to a new generation of architects leveraging the technology, and to a number of buildings that they have designed, is well-known. The impact of this on the record of architectural innovation and practice – in architecture libraries, archives, museums, among others – is only beginning to be appreciated. No longer can libraries acquire blueprints or drawings, a few images, and a scale model or two, to represent a major work of architecture in their collections. Now they must acquire the 3-D CAD models and 2-D drawing files, Building Information Models (BIM), digital images, videos and documents, all delivered on a computer hard drive with no annotation whatsoever. No library or archive is currently prepared for this new reality, but they are increasingly under pressure to figure out how to acquire these 21st century collections, to support the next generation of architectural students and historians.

CAD is particularly problematic for libraries, museums, and archives since it is highly volatile, relying on proprietary mathematical algorithms to represent shapes and structures, and packaged in complex, proprietary, and rapidly evolving software products which are expensive, digitally encrypted and obsolete within years… a digital preservationist's worst nightmare. Fortunately the standards for CAD are catching up, so that options are emerging to represent CAD drawings and models in ways that achieve a degree of "interoperability" across systems and time. These standards are complex and offer many trade-offs among them. Different software programs support different standards, and each standard supports different aspects of the represented design. In addition to simply capturing these digital design files, archiving this material raises a host of questions about what purposes the digital designs should serve, their authenticity, and how to technically manage such assets in the digital future.

Recognizing this growing problem, in 2006 the MIT Libraries applied for, and was awarded, a grant from the U.S. Institute of Museum and Library Services for a research project to develop a practical strategy for processing and preserving the output of modern architectural project involving 3-D CAD and other digital material. The project is called FACADE: Future-proofing Architectural Computer-Aided Design, and this document represents our final report to the IMLS on our findings. The project was originally funded for the two years beginning October 1, 2006, but was granted a one year no-cost extension in August of 2008, allowing us to continue work up to the present time.

For the past two and a half years the MIT FACADE Project has investigated the complexities of acquiring, processing, archiving, preserving, and disseminating digital materials produced during major architectural projects. The primary focus of our work was on 3-D CAD models of buildings, but also included the myriad other files that are generated during such projects (documents, images, videos, spreadsheets, websites, and so on). Over the course of the project we acquired the digital files of several recent buildings designed by noted architects who rely on 3-D CAD modeling, and we have processed these collections into a prototype digital archive based on the DSpace institutional repository (or digital archive) software. This report encapsulates all of our findings and recommendations on the range of research topics we originally proposed, and a few more that we did not anticipate at the outset but proved necessary to achieve our goals.

The proposed research questions for the FACADE Project were:

- What techniques can and should be applied to preserve the native CAD architectural models over archival time frames? Given that CAD models require particular versions of specific software programs to interpret them, is it necessary and sufficient to archive the software as well, or is an "emulation" framework needed for the digital archive platforms that host the material?
- What additional process information is needed to capture the entire building life cycle, and how can that information best be stored in digital archives? Is a new standard necessary for encoding that information, or is a linked document sufficient?
- What other annotations need to be supported to capture the architect's intentions and instructions to the contractors and subcontractors who do the construction (i.e. the Building Information Model) and where and how should that information be kept?
- How can we archive this type of data into institutional digital repository systems like DSpace, which are designed to cover the entire range of digital data formats that libraries, archives and museums need to manage and preserve?

Building on these research questions, the FACADE Project defined five major deliverables:

- Analysis, identification and description of native digital formats produced by top CAD software used by architects, primarily CATIA and AutoCAD formats. Registration of these formats into the Global Digital Formats Registry and similar registries for general access.
- Analysis, design and implementation of native CAD file ingestion, management, preservation and dissemination practices, and development of necessary modules for the DSpace digital archive system. These may include archiving of relevant CAD software packages for future processing, or development of emulation tools and frameworks for rendering these files in the DSpace platform at a minimum.
- Analysis and recommendation related to process documentation (relationships between various CAD files and versions, and between CAD files and other project communication and documentation).
- Analysis and recommendations related to annotation of CAD files for important related information, such as non-graphical files related to materials used.
- Documentation, training, outreach and dissemination of results to the digital library, digital preservation, and DSpace user communities.

These objectives have now been met or modified, as explained in our interim reports, to adapt to the changing landscape over the two-and-a-half year period of work. The status and outcomes of each deliverable are described in detail below.

Project Results

The FACADE Project began actual work in January of 2007 and recently completed its final tasks as we reach the two and a half year mark. The following report is organized into seven sections, covering the structure and process of the project and the five deliverables listed above.

1. FACADE Project Team and Process

The team for the FACADE Project was composed of a Principal Investigator and Project Manager from the MIT Libraries and, at various points in time, software developers, metadata specialists, architecture specialists, and staff from the MIT Rotch Library of Architecture and Planning. We worked especially closely with the School of Architecture, including project Research Assistants who were graduate students in the Master of Science in Architecture Studies (SMArchS) program with extensive experience working in architectural firms as project architects and technology experts. We also, at various times, consulted staff of the MIT Institute Archives, the MIT Museum, the MIT Facilities Department, and faculty from other departments such as Computer Science and Mechanical Engineering. MIT is fortunate, and nearly unique, in its ability to leverage this nexus of expertise for this purpose.

We also benefited from an excellent project Advisory Board, chaired by Professor William Mitchell from the MIT School of Architecture, and including:

- Stephen Abrams (Senior Manager for Digital Preservation Technology, California Digital Library, University of California),
- Alonzo Addison (Special Advisor to the Director of the UNESCO World Heritage Centre),
- Howard Burns (Architectural historian, Scuola Normale Superiore, Pisa),
- Kristine Fallon (CEO, Kristine Fallon Associate, Inc.),
- William Regli (Professor, Department of Computer Science, Drexel University),
- Dennis Shelden (CTO, Gehry Technologies).

The Project was structured around the annual Advisory Board meetings, which served as focal points for delivering prototypes of the archive and its user interface for reactions and direction from the Board. By delivering consecutive working prototypes of the archive we were able to explore our assumptions about priorities and use of the material with a representative group of stakeholders, helping to insure that the final product would be useful to them. Each prototype was an end-to-end solution, covering the entire process from identifying a new building collection to acquiring, processing, ingesting, and publishing that collection. In other words, the desired User Experience drove our ultimate requirements for the material so that throughout the project we were simultaneously working on each aspect of the workflow. This meant, for example, that the user requirements helped us choose appropriate standards for digital preservation of the 3-D CAD models based on requirements for future access and use of the models.

At the Project's initiation, we undertook an Outcomes-Based Planning and Evaluation process to help us identify and prioritize the project's target audiences, and how to assess our success with each. The possible target audiences included:

- Librarians, archivists, and museum staff who work with architectural collections

- Instructors and students in architecture programs
- Architectural historians
- Architecture and design practitioners
- Members of the public

Over the course of the project our thinking changed about which audience to target first. While our ultimate audience is clearly the library, archives and museum community to which we belong, we initially we thought that practitioners were the most important near-term target audience, since they control the digital archives we want to acquire and would be motivated to contribute those collections if they could use the archiving system themselves (something they typically struggle with now). However we discovered that the technical environment in the majority of architectural firms is not as advanced as it is in modern research libraries, so these firms would be unlikely to adopt our system for internal use. With that in mind, we changed our primary focus to architecture instructors and students, with a secondary focus on historians. Instructors and students have an immediate need for this type of data, while architectural historians are not yet studying contemporary architects who rely on digital design software so we could not assess their needs quite as directly. The public we considered to be covered by the roles of students and historians. With this target audience in mind, we proceeded to design the information model and the archiving system, while realizing that all of these audiences are important to support.

Fortunately, our Advisory Board and the various focus groups that we were able to convene adequately represent all the target audiences, and were very clear in their recommendations for how the material should be organized and what functionality the system should support.

2. The FACADE Research Collection

Over the course of the project we created a collection of digital material from several major contemporary architects that we could use as a research test bed. Our goal for this collection was to identify major architects using different 3-D CAD modeling tools in their normal work practices, and acquire examples from them of actual building project data to experiment with. We are fortunate to be collaborating with the MIT School of Architecture and Planning, and in particular with Professor William Mitchell, so that we have access to the world's great architects in order to both collect data and test ideas (e.g. licensing terms for using or publishing the data). Given the very large size of these project archives, comprising tens of thousands of computer files for building projects that lasted several years and cost millions of dollars, we have so far limited ourselves to a small number of projects. But the data we have now exercised all of the issues described in our research agenda, and we will be able to collect further data as needed. The current collection is described below:

Moshe Safdie and Associates[1]

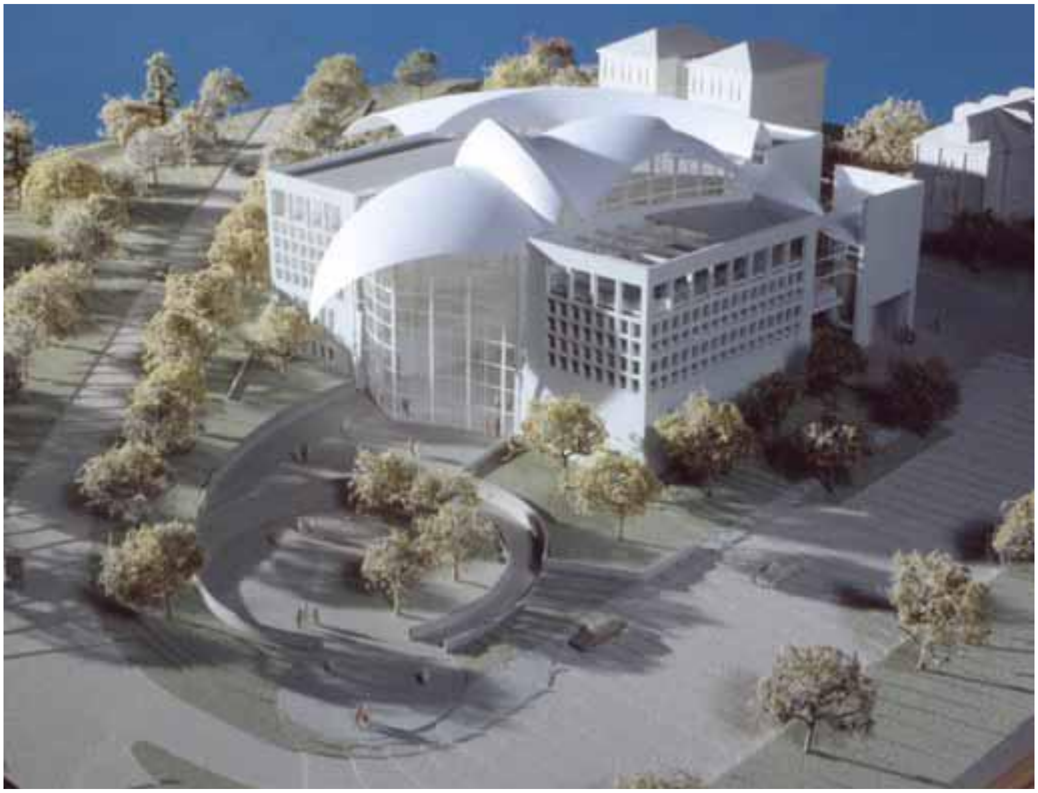


Figure 1. Model of the U.S. Institute of Peace designed by Moshe Safdie Associates

The test data from Moshe Safdie's architecture firm is for the new United States Institute of Peace, currently under construction on the National Mall in Washington, D.C. It is scheduled for completion in the fall of 2010, and has already completed major phases of design and construction. MSA used CATIA for early phases of the project (particularly the complex roof design) and then switched to Autodesk's Revit for the remainder of the project, including migration of the roof model. Interestingly, Moshe Safdie's physical archives (drawings, sketchbooks, models, and printed project files) are being archived by the Canadian Architecture Collection at McGill University in Montreal, but they are unable to accept his digital archives because of the preservation problem described in our research agenda. MSA has offered to provide additional building projects to our research collection if we need them.

[1] Moshe Safdie and Associates is based in Boston, Massachusetts and is described in detail on their website at http://www.msafdie.com/.

Frank O. Gehry & Associates[2]

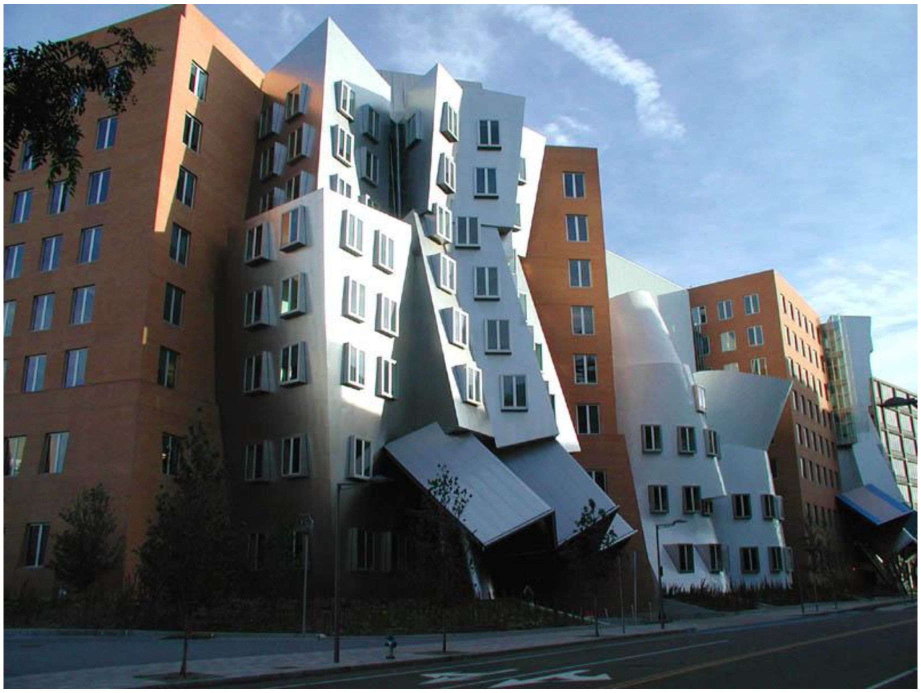


Figure 2. the Ray and Maria Stata Center at MIT designed by Frank O. Gehry & Associates

Frank Gehry is regarded as one of the pioneers in using sophisticated CAD modeling software in his architectural practice. He was the first architect to employ the CATIA software for architectural design (it was originally created for the aerospace industry to design aircraft and other complex engineered products). Gehry has said:

"This technology provides a way for me to get closer to the craft. In the past, there were many layers between my rough sketch and the final building, and the feeling of the design could get lost before it reached the craftsman. It feels like I've been speaking a foreign language, and now, all of a sudden, the craftsman understands me. In this case, the computer is not dehumanizing; it's an interpreter."[3]

Gehry completed the MIT Stata Center[4] in 2004, and MIT retained a full set of his 3-D CAD designs and related material which are now part of our research collection. Gehry has used CATIA exclusively over the past two decades, and started a technology company – Gehry Technologies, Inc. – to provide technology and services to leading owners, developers, architects, engineers, general contractors, fabricators, and other building industry professionals worldwide.

Morphosis Architects[5]

[2] Gehry’s company is based in Los Angeles, California and is described on their website at http://www.foga.com/

[3] See the Case Study of CATIA at Frank O. Gehry & Associates, Inc. http://www.cenitdesktop.co.uk/html/case_frank_gehry.htm

[4] See http://en.wikipedia.org/wiki/Stata_Center for general information, and the article in Wired Magazine 12(05) May, 2004 “Frank Gehry’s Geek Palace”, http://www.wired.com/wired/archive/12.05/mit.html for more detail

[5] Morphosis is based in Santa Monica, California, and is described on their website at http://www.morphosis.com/

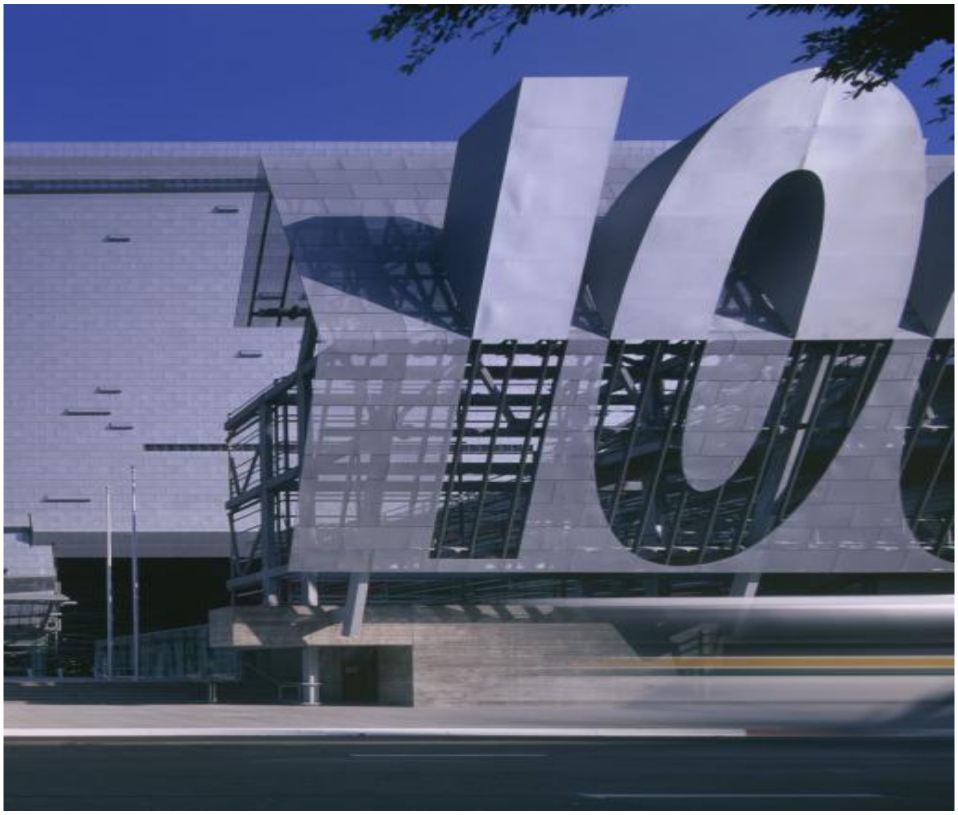


Figure 3. the Caltrans District 7 Headquarters building designed by Morphosis

Morphosis principal Thom Mayne provided a third building for our research collection, the designs and project files for the Caltrans District 7 Headquarters building in Los Angeles, California, completed in 2004. The building was designed using the Bentley Microstation CAD modeling software, and we have received the complete project archive for the building.

Faculty of the MIT Department of Architecture

Finally, we are also working with members of the MIT Faculty of Architecture, including architect Larry Sass, who provided design data for his *Digitally Fabricated Housing for New Orleans* building that was included in the New York Museum of Modern Art's 2008 exhibition *Home Delivery: Fabricating the Modern Dwelling*[6]. Sass used 3-D modeling tools including Rhino and Maya in his design process, and computer programs to create the component pieces of the building that were assembled on site in New York. This construction process is growing in popularity and has major implications for the future of architecture design and construction.

[6] The exhibition is detailed at http://www.momahomedelivery.org/

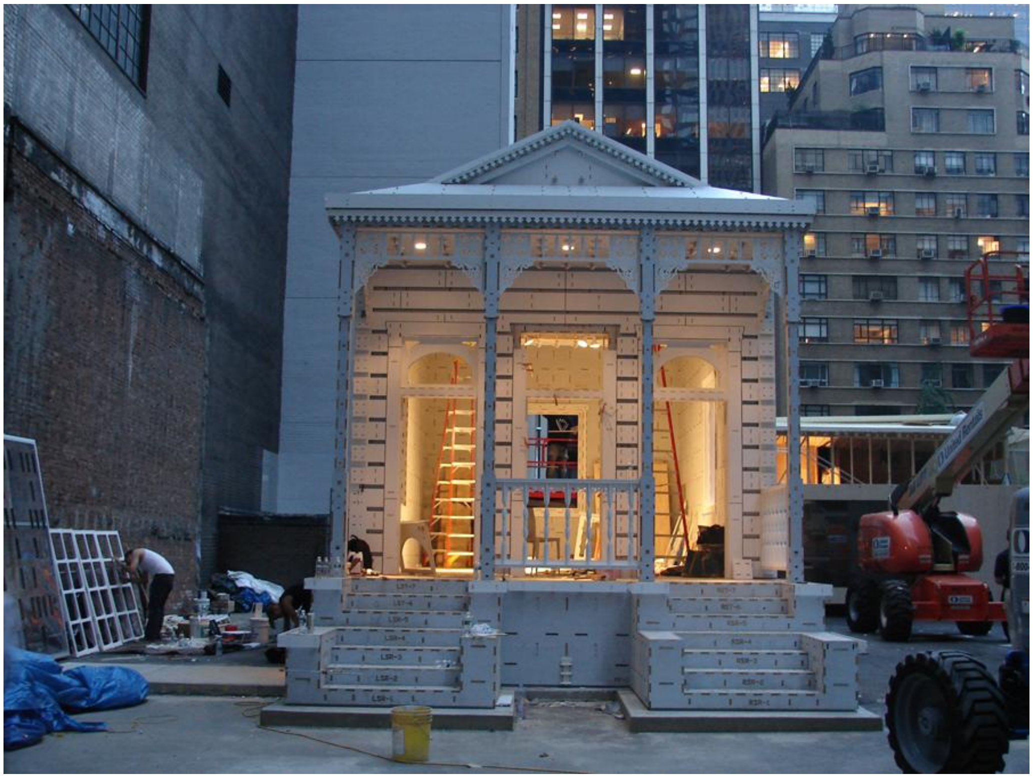


Figure 4. Digitally Fabricated Housing for New Orleans designed by Larry Sass

Summary

What characterized each of these building projects was that we acquired the material on a hard drive or set of DVDs in whatever file system was in use by the firm, and without annotation to help us determine what was included. There has been much speculation that the solution to the archiving problem is for architectural firms to provide their data in file formats and organized in the manner of our choosing. In discussing that scenario with firms, it became clear that this is an unrealistic expectation for the foreseeable future, and that our best option is to work with software companies that support architectural firms (e.g. Newforma[7]). Beyond that, we will continue to get ad hoc file collections and need to annotate and organize them as part of the acquisition and processing workflow.

Of the test collections acquired, the size ranged from just under 20,000 files (10Gb) to almost 100,000 files (50Gb) for a building-in-progress. The 3-D CAD models in particular are each very large (comprised of one or more separate files) but are usually few in number. The 2-D CAD drawings and other files are smaller, but extremely numerous. If the firm has culled the project files for their own archives then we acquire a smaller set consisting of what the firm considered important to keep, but ideally we would acquire complete sets of data so that they include more than just the designs and client presentations (since the other material is often of high historical value). In future we will develop guidelines for architectural firms of what material we recommend they keep, to help insure that the handover to the long-term archive includes everything we want to acquire.

[7] Newforma http://www.newforma.com/ is a popular Project Information Management product used by architectural firms to organize their project data. It is based in New England and its leaders are open to discussions with us about the long-term archiving problem and how they can help.

For each building, we asked for material from all stages of the project, including: concept design, schematic design, design development, construction documents, and construction administration. While 3-D models were the focus of our research on digital preservation, we found that the context provided by the other materials in the collection were key to understanding the models (e.g. client presentations, correspondence with clients and contractors, and digital images). Since architects cannot currently indicate their design intent directly on a 3-D model, having the complete collection gives students and historians a means of understanding what the architect was trying to achieve.

A final note on the test collection is about the intellectual property concerns of the architectural firms. Each of the firms we worked with was willing to provide research data under a very liberal license, but was unwilling to allow open public access initially. Now that we have a working prototype to show them, and a proposed archive license, we are beginning to discuss a more permanent arrangement with them. These firms understand the seriousness of the situation for archives and for the historical record of their work, and to that extent they are very open to discussing an archive license with us. However they have legitimate concerns about their legal exposure and client confidentiality if we acquire the complete records of projects and make them publically available. So the discussions we are having now relate to which parts of the project data can be made public immediately and which later (via an embargo), and what records they cannot share with us at all. We expect those negotiations to be ongoing for the next few years as we become more sophisticated about this and gain experience with the license.

3. CAD File Format Representation Information (i.e. Format Metadata)

> *Analysis, identification and description of native digital formats produced by top CAD software used by architects, primarily CATIA and AutoCAD formats. Registration of these formats into the Global Digital Formats Registry and similar registries for general access.*

Given the highly proprietary nature of CAD software and the internal data formats they each use, acquiring detailed information about those internal formats proved predictably difficult to obtain. Some software vendors have made this information publically available (e.g. Autodesk's AutoCAD formats. Others have proven willing to discuss the issue with us, and to offer alternative solutions (e.g. a license- and DRM-free copy of the software to permanently archive for future use). We have collected information where we could, and plan to continue negotiating with software vendors to acquire their format representation information. Representatives of both Autodesk (i.e. Revit) and Bentley (i.e. Microstation) and in discussions with us, and we anticipate that if the top two or three vendors of CAD software for architects supply this information to us then others will follow suit. However, it is clear that these vendors do not want their format information made publically accessible (for obvious reasons) so we will probably be required to escrow the information ourselves or work with the public format registries to make this information inaccessible for some contractual period of time.

CAD Format Information for PRONOM

Per the original deliverable, the FACADE Project provided to the PRONOM digital format registry all information we were able to determine about representation information for 3-D CAD software formats and other formats found in the test collection. Thirty-five format additions or modifications were submitted to PRONOM for inclusion in the registry at http://www.nationalarchives.gov.uk/pronom/

- AutoCAD Drawing 2004-2005
- X-Windows Dump File
- Java Compiled Object Code
- Apple QuickTime
- Tab-Delimited Text File
- RealAudio Metafile
- Extensible Markup Language 1.0
- TeX Binary File
- Windows shortcut file
- Portable Document Format 1.0
- 3DM 4 openNURBS, Rhino
- DWG (2007-2008) AutoCAD
- CATIA Model 4
- CATIA Project 4
- CATIA Material Description 5
- CATIA Model (Part Description) 5
- CATIA Product Description 5
- AutoCAD Database File Locking Information
- form*Z Project File
- Adobe InDesign Document
- Revit Family File
- Revit Family Template
- Revit Template
- Revit External Group
- Revit Project
- Revit Workspace
- Steel Detailing Neutral Format
- SketchUp Document Backup
- SketchUp Document
- TrueType Font
- Internet Shortcut
- JPX (JPEG 2000 Extended)
- Initial Graphics Exchange Standard
- Windows Bitmap V3
- MPEG-1 Video Format

Additionally, because DSpace requires a MIME-type (a.k.a. Internet media type) to apply to each file for Web browsers to use, every file format in our test collection required a MIME-type to be included with its PRONOM entry. Since many PRONOM entries were lacking the field or had incorrect data, we provided some 155 corrections and additions to MIME types for the PRONOM registry.

Update on the Global Digital Formation Registry

Our original deliverable called for registering the CAD format information into the Global Digital Format Registry, which was, in 2006, a funded development project based at Harvard University and backed by the U.S. National Archives (among other prominent institutions both in the U.S. and internationally). In the intervening years, the GDFR lost momentum and ultimately failed to reach an operational state. In its place there is a new project called the Unified Digital Formats Registry has emerged which will unify the GDFR and PRONOM communities towards a single, common registry with a defined governance model and plan for sustainability. Unfortunately, that effort has just begun and its organizers anticipate that it will be at least sixteen months before there it becomes operational. Fortunately, the plan includes importing existing data from the PRONOM registry, so the FACADE Project's contributions to PRONOM should be carried forward to whatever emerges from this new initiative.

4. Building Information Models

> *Analysis and recommendations related to annotation of CAD files for important related information, such as non-graphical files related to materials used.*

In our initial proposal we discussed a new development in digital architecture: Building Information Models or BIMs[8]. The BIM concept is a next generation 3-D CAD model that adds annotations and other data to support the entire lifecycle of a building, from design through its years of future use – data not just for designers, but for building owners too, evolving as the building does over time. Building information modeling includes geometry, spatial relationships (e.g. parametrics), geographic information, quantities and properties of building components (e.g. manufacturers' specs). As 2-D drawings gave way to 3-D models, now 3-D models are giving way to BIM databases that bring together the range of building-related information into one place.

The AEC industry is energized by this new concept and discusses it often. FACADE Project team members spent considerable time learning about BIM (e.g. from Professor Chuck Eastman at the Georgia Institute of Technology, often credited with coining the name "BIM") and its potential for building data communication and archiving. But as a practical matter, BIM adoption by architects and the software they currently rely on proved immature. The concept is beginning to emerge in real software products, but its adoption by design practitioners has not yet reached the mainstream. Certain BIM software products (e.g. Digital Project from Gehry Technologies[9], newer versions of Autodesk's Revit and Bentley Architecture) are available to architects now, but we now believe it will be several more years before a BIM model is the dominant data communication tools for architecture.

In the meantime, it is still necessary to relate the 3-D CAD models to their corresponding 2-D drawings, specifications, material lists, and so on. Since there was no provided solution to this need, we took the approach of annotating models and relating them to other data via the Project Information Model (PIM) described in the next section. BIM will never include every file related to a building project, but will gradually come to incorporate more and more of the building data needed for ongoing maintenance. Our approach of using a PIM to relate building data is very flexible and can easily adapt to emerging BIM use in the future.

[8] Details on the standard are available on the buildingSMARTalliance website at http://www.buildingsmartalliance.org/nbims/

[9] http://www.gehrytechnologies.com/index.php?option=com_content&task=view&id=97&Itemid=211

5. Organizing Architectural Project Files: the Project Information Model

> *Analysis and recommendation related to process documentation (relationships between various CAD files and versions, and between CAD files and other project communication and documentation).*

A major deliverable of this project is the information model (or relationship map) that reflects the relationships among the materials received from the architectural firm. We currently keep all the files received except for system "junk" files that are not actual content, and some files that are duplicated, so the information "ontology" that we developed covers every type of file we can receive and places each file into a context that allows target audiences to locate and retrieve it. Our initial version of the ontology had exhaustive relationships between each files so that, for example, letters that referred to drawings were explicitly linked together, drawings were linked to their corresponding models, and letters were linked to each other in a series – all generated by hand.

While these linkages were seen as very useful, this approach was clearly not going to scale to tens of thousands of files for each building. Our Advisory Board suggested a different approach that would separate the collection into two parts: first, a small number of key "selected" items from the collection (e.g. 3-D CAD models; client presentations; important images) that are of high value and should be annotated more carefully and showcased in the user interface, and second, the remainder of the collection, to be given basic descriptive tags and made available more generically in the user interface. The rationale for this suggestion was that the "selected objects" in the curated set would meet 80% of the users' needs, but the other 20% should be kept for the student or historian who is motivated to browse through that materially manually, given a starting point. These "selected objects" are similar in concept to the "outputs"[10] recommended for archiving in a 2004 report on archiving digital data prepared by Kristine Fallon Associates for the Art Institute of Chicago.

From that point, we developed our final Project Information Model (PIM) ontology, and designed a workflow to allow the staff to organize the collection this way: identify the "selected objects" of note, annotate the files appropriately, and expose them via a user interface in the two categories.

The ontology (see Figure 5) is organized around the central concept of a "file" (or set of files). Every file is then assigned five properties:

- Building project phase (when), e.g. concept, design, construction
- Architectural discipline (why), architectural, electrical, mechanical
- Building zone (where), e.g. Stata Center, Gates Tower, 4th floor
- Document type (what), e.g. presentation, drawing, communication
- File format (what), e.g. CATIA 3-D CAD model, JPG image, Word document

File formats are further categories by purpose (e.g. original, standardized for preservation, display), and are linked to a record for the corresponding software that created them. We also noted where in the original file system from the architectural firm the file was located, since file co-location could be exposed via the user interface as a possible clue to historians of design intent or other interesting aspects of a project. Finally, we note the access policy for each file, in case there is an embargo or other limitation required by the architect. Selected objects are further assigned properties of design

[10] See the report "Collecting, Archiving and Exhibiting Digital Design Data" published in 2004 at http://www.artic.edu/aic/depts/architecture/ddd.html

type, e.g. 3-D model, 2-D drawing or drawing set. Since so much of the data were digital designs in which many files are linked together to create a model or a drawing set, we provided a way to relate files together in a particular sequence.

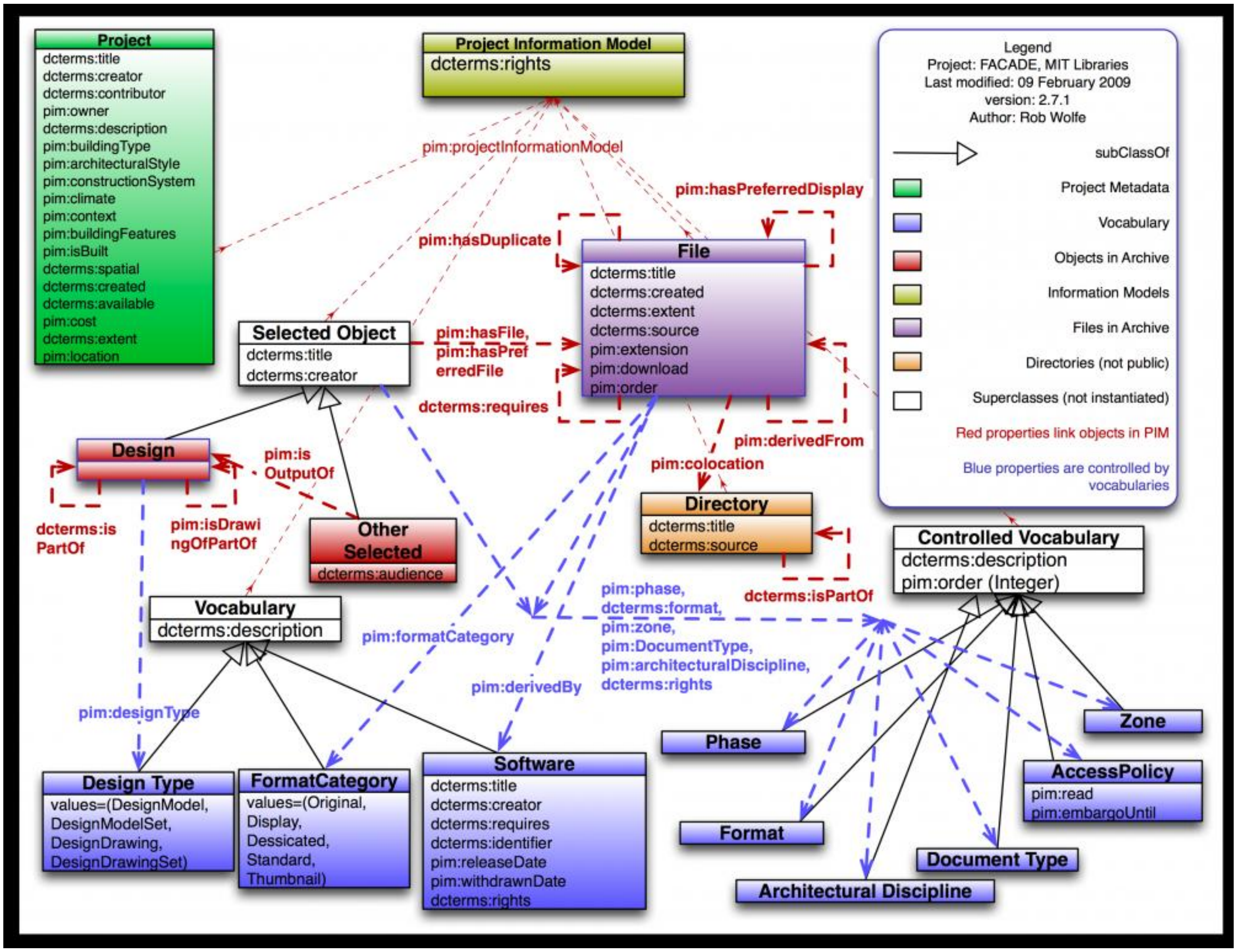


Figure 5. FACADE Ontology diagram, for "Project Information Model"

File formats are further categories by purpose (e.g. original, standardized for preservation, display), and are linked to a record for the corresponding software that created them. We also noted where in the original file system from the architectural firm the file was located, since file co-location could be exposed via the user interface as a possible clue to historians of design intent or other interesting aspects of a project. Finally, we note the access policy for each file, in case there is an embargo or other limitation required by the architect. Selected objects are further assigned properties of design type, e.g. 3-D model, 2-D drawing or drawing set. Since so much of the data were digital designs in which many files are linked together to create a model or a drawing set, we provided a way to relate files together in a particular sequence.

Finally, the ontology includes a "project" entity to represent the building as a whole, and provides a place to add cataloging metadata for the building. The building properties supported were suggested by the Advisory Board and will be harmonized with standards for metadata from the art and architecture library community (e.g. the CCO and CDWA schemas).

As will be discussed later, the metadata prescribed by the PIM ontology is created by automated tools or by library staff as part of the processing workflow for a new building collection. It can be done all at once or iteratively over time. The metadata is encoded in a Web technical standard called "RDF"[11] and is stored as a file alongside the other data files for use by the system and in its user interface. This approach allows us to change the ontology very easily as we gain experience with this type of collection, without the corresponding need to change either the workflow software or the user interface to the material.

While the PIM changed significantly over the course of the project, it was reviewed in key states by several experts from the digital library and archives field[12] and has been widely disseminated to interested parties. There is speculation in the community that while it was designed for architecture projects, with very minor modification is might be made to work for any type of collection that is acquired in this way – i.e. on a hard drive consisting of an ad hoc file system containing tens of thousands of un-annotated files. We hope to test that theory in the future.

6. Archiving Architectural CAD and Related Digital Files

> *Analysis, design and implementation of native CAD file ingestion, management, preservation and dissemination practices, and development of necessary modules for the DSpace digital archive system. These may include archiving of relevant CAD software packages for future processing, or development of emulation tools and frameworks for rendering these files in the DSpace platform at a minimum.*

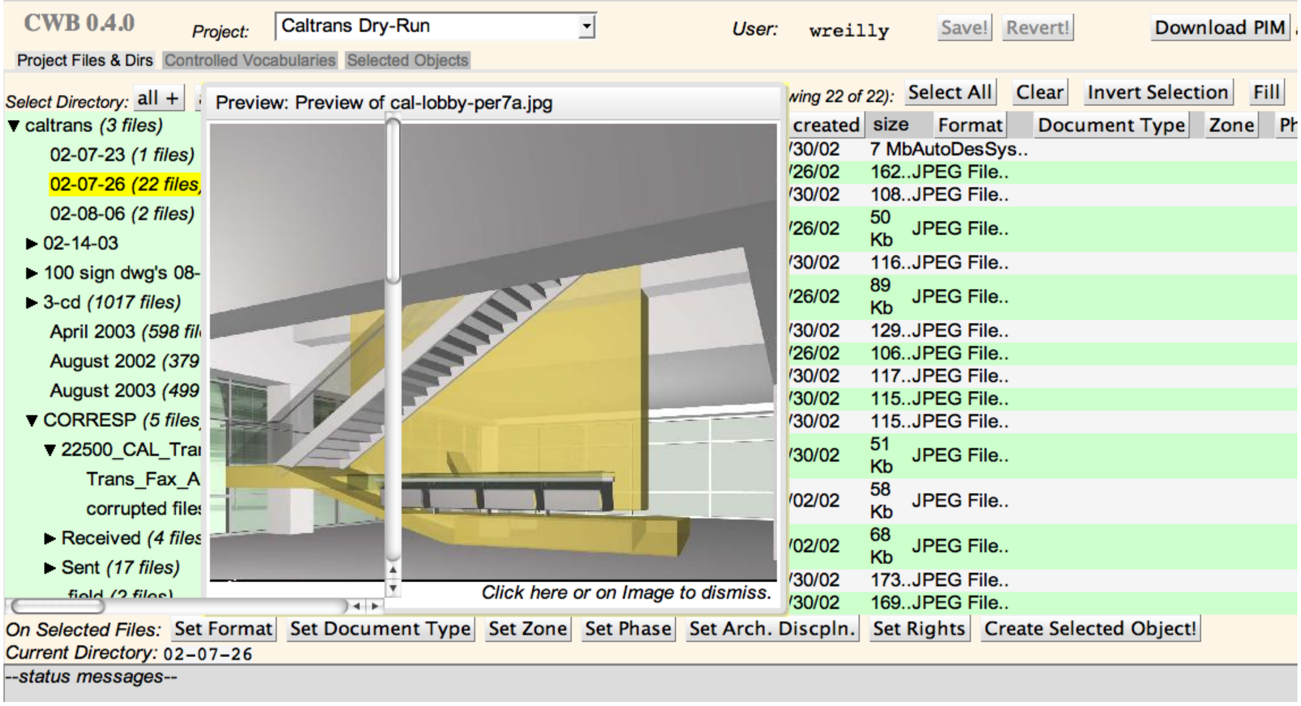
This deliverable covers the vast majority of the development effort for the FACADE Project. It encapsulates all of our work in specifying a workflow for library curators to acquire and process these digital collections, building the archive itself (based on the DSpace open source software platform), implementing a user interface to the collections, and a curation/preservation strategy for digital design data. Each of these aspects of the work will be described here, with more detail available on the project website. Specifically, we

- Developed a DSpace data model for architectural project materials and generated metadata;
- Programmed DSpace ingest of metadata and all materials; and export of metadata for new user interface software;
- Developed several iterations of a prototype public user interface to the complete data collections, using Semantic Web software originally developed for another project (i.e. the Simile Project's Longwell and Exhibit software);
- Completed a prototype "Curators' Workbench" (i.e. a tool to assist library and archives staff with scalable processing of digital acquisitions)
- Documented end-to-end workflow to address needs of architectural curators, digital preservation specialists, and technical operations staff, including archive workflow, preview workflow, post-publish workflow, and license workflow;
- Concluded architectural domain expertise activity in CAD tools export to derived standard formats for archiving and display;
- Established preservation strategy policy recommendations for received architectural materials;

[11] The Resource Description Framework, http://www.w3.org/RDF/ is standard developed by the W3C as part of the Semantic Web initiative. It is encoded in XML, and constrained by OWL ontologies such as our PIM ontology.
[12] For example, Professor Jerry McDonough at the University of Illinois Urbana-Champagne i-School and by Robin Wendler of the Harvard University Library Office of Information Systems.

Software Development for Metadata: Curators' Workbench

Beyond the scope of the original proposed deliverables, but necessary to process the size of collection received from architectural firms to the specifications of our target audience, was the creation of a "workbench"-like software system for use by curators. The tool, dubbed the “Curator's Workbench”, allows staff to apply metadata tags and other designations (e.g. selected objects) for the tens of thousands of files received from a firm. This is done with a Web-based computer application that exposes the original file system as received, in an environment where staff can bulk tag the files, including entire directories, in minutes (see Figure 6). Many of the metadata properties we provide (as described in section 5 on the PIM) are common to large numbers of files that are co-located in the source collection, e.g. a directory of mechanical system drawings, all of which will get an architectural discipline tag of “mechanical”, were done in the same “construction” phase, and are stored in “DWG” file format (see Figure 7). The goal is for a curator to be able to process an entire new collection of tens of thousands of files in a few weeks, or even days. The curator also identifies the key “selected objects”, ideally in communication with the architectural firm, but not necessarily. Finally, the curator archives the data and newly created metadata into DSpace.



Project: Caltrans Dry-Run

User: wreilly

Select Directory: all +

▼ caltrans *(3 files)*
02-07-23 *(1 files)*
02-07-26 *(22 files,*
02-08-06 *(2 files)*
▶ 02-14-03
▶ 100 sign dwg's 08-
▶ 3-cd *(1017 files)*
April 2003 *(598 fil*
August 2002 *(379*
August 2003 *(499*
▼ CORRESP *(5 files*
▼ 22500_CAL_Trai
Trans_Fax_A
corrupted files
▶ Received *(4 files*
▶ Sent *(17 files)*
field *(2 files)*

Preview: Preview of cal-lobby-per7a.jpg

wing 22 of 22):

created	size	Format	Document Type	Zone	Ph
/30/02	7 Mb	AutoDesSys..			
/26/02	162..	JPEG File..			
/30/02	108..	JPEG File..			
/26/02	50 Kb	JPEG File..			
/30/02	116..	JPEG File..			
/26/02	89 Kb	JPEG File..			
/30/02	129..	JPEG File..			
/26/02	106..	JPEG File..			
/30/02	117..	JPEG File..			
/30/02	115..	JPEG File..			
/30/02	115..	JPEG File..			
/30/02	51 Kb	JPEG File..			
/02/02	58 Kb	JPEG File..			
/02/02	68 Kb	JPEG File..			
/30/02	173..	JPEG File..			
/30/02	169..	JPEG File..			

On Selected Files:

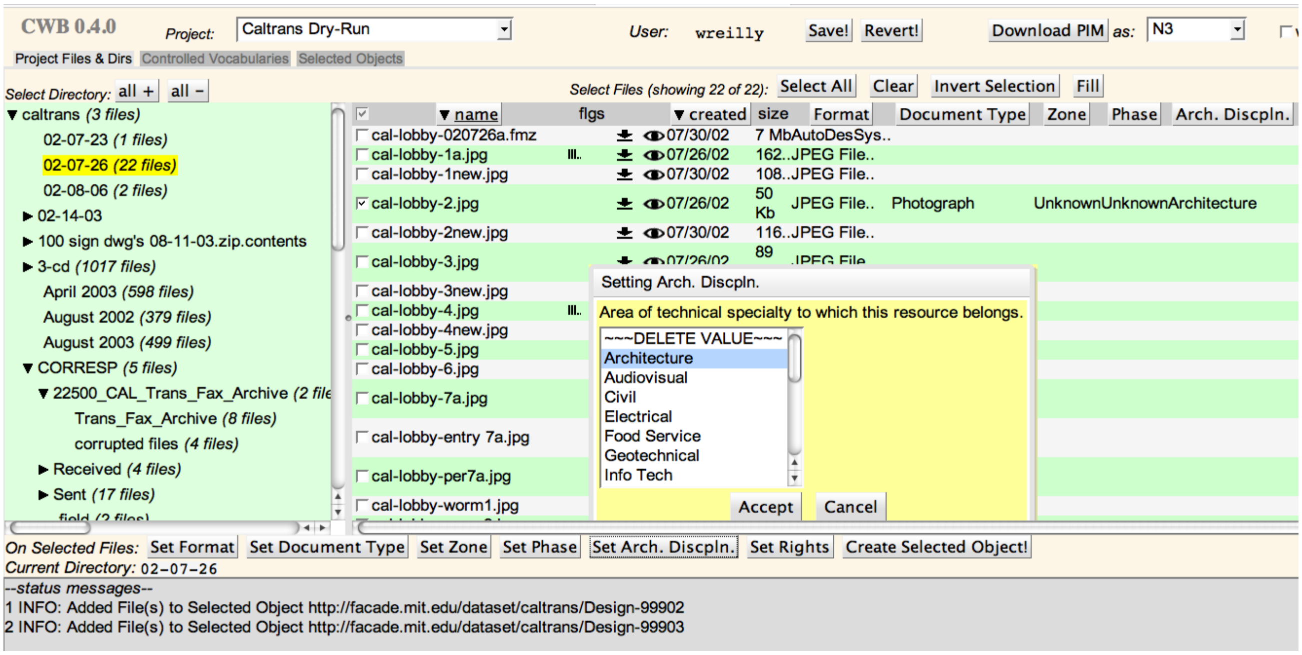
Current Directory: 02-07-26

--status messages--

Figure 6. Curators' Workbench, preview function (JPG thumbnail)

The workbench interface provides:

- a "download" feature to get a copy of a file for local study;
- a "preview" feature to show a version of the a file in the Web browser;
- an icon flagging presence of duplicate files (commonly found in these collections);
- a feature to set intellectual property rights statements on files;
- functionality supporting the complex task of assigning multiple files to the "Selected Object" status.



Project: Caltrans Dry-Run *User:* wreilly *as:* N3

Select Directory:

- ▼ caltrans *(3 files)*
  - 02-07-23 *(1 files)*
  - 02-07-26 *(22 files)*
  - 02-08-06 *(2 files)*
  - ▶ 02-14-03
  - ▶ 100 sign dwg's 08-11-03.zip.contents
  - ▶ 3-cd *(1017 files)*
  - April 2003 *(598 files)*
  - August 2002 *(379 files)*
  - August 2003 *(499 files)*
  - ▼ CORRESP *(5 files)*
    - ▼ 22500_CAL_Trans_Fax_Archive *(2 file*
      - Trans_Fax_Archive *(8 files)*
      - corrupted files *(4 files)*
    - ▶ Received *(4 files)*
    - ▶ Sent *(17 files)*
    - field *(2 files)*

Select Files (showing 22 of 22):

☑	▼ name	flgs	▼ created	size	Format	Document Type	Zone	Phase	Arch. Discpln.
□	cal-lobby-020726a.fmz		07/30/02	7 Mb	AutoDesSys..				
□	cal-lobby-1a.jpg	lll..	07/26/02	162..	JPEG File..				
□	cal-lobby-1new.jpg		07/30/02	108..	JPEG File..				
☑	cal-lobby-2.jpg		07/26/02	50 Kb	JPEG File..	Photograph	Unknown	Unknown	Architecture
□	cal-lobby-2new.jpg		07/30/02	116..	JPEG File..				
□	cal-lobby-3.jpg		07/26/02	89	JPEG File				
□	cal-lobby-3new.jpg								
□	cal-lobby-4.jpg	lll..							
□	cal-lobby-4new.jpg								
□	cal-lobby-5.jpg								
□	cal-lobby-6.jpg								
□	cal-lobby-7a.jpg								
□	cal-lobby-entry 7a.jpg								
□	cal-lobby-per7a.jpg								
□	cal-lobby-worm1.jpg								

Setting Arch. Discpln.

Area of technical specialty to which this resource belongs.

- ~~~DELETE VALUE~~~
- Architecture
- Audiovisual
- Civil
- Electrical
- Food Service
- Geotechnical
- Info Tech

On Selected Files:

Current Directory: 02-07-26

```
--status messages--
1 INFO: Added File(s) to Selected Object http://facade.mit.edu/dataset/caltrans/Design-99902
2 INFO: Added File(s) to Selected Object http://facade.mit.edu/dataset/caltrans/Design-99903
```

Figure 7. Curators' Workbench, popup to "Set Architectural Discipline" metadata facet on selected file(s)

A goal of the design of this tool and the workflow it supports is ensuring that curation staff do not need to process metadata outside of this system, e.g. using a text editor on the raw RDF metadata. This lowers the risk of inadvertent errors introduced to the metadata due to staff inexperience with complex metadata formats and ontologies.

Related to the Curator's Workbench, we also created a set of utility programs to supplement the Workbench metadata for large-scale processing by more technical staff. For example, if the dates of project phases are generally known, a software tool can process each file to find its technical "create date" and assign it to the corresponding phase. While this would never be 100% accurate, it provides default values for properties that would otherwise take time to assign by hand in the Workbench. These software tools also validate processing, e.g. for consistency and adherence to the PIM ontology rules. These tools, combined with the Curators' Workbench, provide a prototype for a complete solution to the acquisition workflow of large-scale, complex digital archival collections that are acquired in contemporary computer file systems.

The Curator's Workbench is designed as a stand-alone system, so that it is only run when needed (e.g. when a new collection is acquired). It allows the curator, together with technology staff, to process a collection, archive it to DSpace (or another system), shut it down, and restart it in the future to do more processing, or if a missing item is discovered, etc.

Since it was not a proposed deliverable, we were limited in how much time we could apply to the Curator's Workbench. We consider it a working *prototype* and not a production-quality tool as yet. But its utility and obvious applicability to other archival problems is so clear that we intend to seek further funding to finish this work and provide it for widespread use beyond MIT.

Software Development for Data: DSpace

DSpace Data Model

The basic data model of DSpace is very simple: there are one or more digital files that constitute a useful "item" of content (e.g. a book, article, image, dataset, etc.). These items are grouped into logical "collections" (and items can be in multiple collections), and collections are grouped into "communities" like departments or centers. This model has worked well for the normal uses of Institutional Repositories in research libraries, but does not immediate suggest how to organize a collection of hundreds of thousands of files representing a building project. We analyzed this problem and developed a data model for architectural collections in DSpace that meets the goals of: a) supporting long-term preservation of received materials; b) maintaining established relationships between files (e.g. drawing sets); c) capturing supporting metadata for the PIM and each file; and d) delivering content files to the Web on request from the end user interface (described below).

The model treats each original digital file as its own DSpace "item" to achieve maximum flexibility for future changes to the metadata, content files (e.g. for preservation migrations), and access methods. It also provides for straightforward migration of the collections from DSpace to other archiving platforms in the future. Each DSpace item includes the original file, as received and all of its derivatives for preservation and display, along with simple metadata needed for ongoing curation of the collection (but not for end user discovery, which is handled in the external user interface). The PIM file is also treated as a DSpace item and can be maintained and retrieved separately from the data files themselves.

Ingest

The FACADE project has created a custom DSpace ingest tool for these large collections, relying on special "packages" of content to load large numbers of files simultaneously. The import processing of FACADE materials represents one of the largest scale operations using DSpace, with scaling up to tens of thousands of files successfully ingested.

For FACADE, two specialized DSpace add-ons were developed to improve processing of received image PDFs (2-D and 3-D) into thumbnail images, and for performing enhanced text extraction from text PDFs. These have been shared with the larger DSpace community.

We also developed sophisticated validation checks against the system logs, to insure that all expected files were correctly ingested (and none that weren't expected). We believe this will have general applicability well beyond this project.

Export

Additional development was needed to permit DSpace to export collection metadata (i.e. the PIM) to the external user interface, and to allow the users to request digital items from the archive for viewing online. We developer specialized tools to export the building metadata in formats expected by our user interface (e.g. N3 and JSON), and these can be adapted to other technical formats as needed in the future for newer, more sophisticated user interfaces.

Technical File Format Support

FACADE building collections include hundreds of thousands of digital files in a miscellany of formats, some labeled and other not (or even better, mislabeled). For long-term preservation it is essential to identify and validate the format of each and every file received so that it can be placed into a curation program with a defined preservation strategy (see, for example, the discussion of 3-D CAD model files below). In order to provide that higher quality file format identification and validation, we redesigned and prototyped an implementation of a revised data ingest subsystem for DSpace.[13] This work supported automated identification and validation of digital file formats, using standard open source software tools like JHOVE and DROID[14], and provided standard metadata and identifiers for identified formats. We also did work to integrate DSpace with the international file format registries that are emerging as part of our digital curation infrastructure. Since the Global Digital Format Registry (GDFR)[15] is not yet open for business, we are working initially with the PRONOM registry[16] developed by the UK National Archives, but we designed the subsystem to work with any format registry (including a local registry), and also worked with GDFR project staff to provide feedback and test results. All of this work was done publicly, with input and feedback from the DSpace community and other experts (e.g. members of our advisory board and collaborators from other repository platforms). We have communicated with the DSpace developer group about this work, and have made the software available for integration into the standard DSpace open source software release.

Open Source Software Releases

Software created for the FACADE Project falls in two categories:

- of general interest to the DSpace community
- of potential interest to a subset of the DSpace community, for handling digital collections similar to the FACADE test collection.

Software in the first category has been submitted to the DSpace developer ("committer") group for inclusion in the current version of the open source DSpace software. This includes things like new processing tools for generating thumbnail images and full-text from PDF files, and improvements to the way DSpace handles file formats for long-term preservation (e.g. integration with external format registries like PRONOM, described above).

Software in the second category has been carefully archived in the MIT Libraries' software repository, and made available under a standard open source software license[17] on request. Our FACADE website will post information about how to request the software, as well as links to the software repository for downloads. We will make our best effort to answer questions and support any external software adopters, and will make any future software improvements available this way even if they are internally funded.

Software Development for the Archive User Interface

The user interface (UI) to the FACADE archive was developed over time via a series of prototypes. Each iteration of the UI was presented to the project's Advisory Board and one or more focus groups for feedback, allowing us to refine the UI as the amount of data increased, and as the requirements for relating and preserving large test collections emerged. The final prototype was finished in the

[13] Details are available from the DSpace project wiki http://wiki.dspace.org/index.php/BitstreamFormat_Renovation

[14] JHOVE and DROID are freely available software tools to read files and establish their encoding format, and verify them as a valid instance of that format. See http://droid.sourceforge.net/wiki/index.php/Introduction

[15] http://hul.harvard.edu/gdfr/

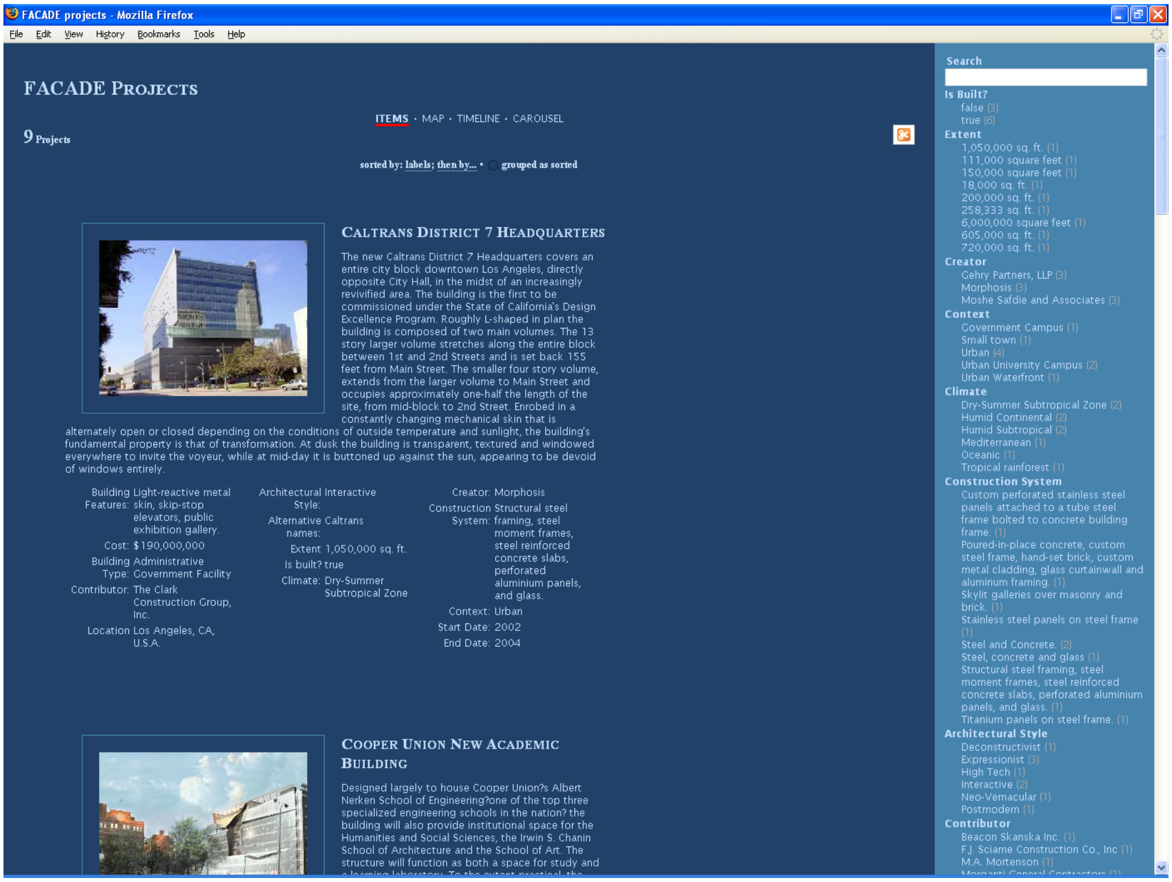
[16] http://www.nationalarchives.gov.uk/pronom/

[17] The MIT Libraries normally use the Berkeley Software Distribution (BSD) license for our open source software releases, and both the DSpace and Simile software products used for FACADE are under that license. So it makes sense to release the new code under that same license for consistency and clarity to new adopters of the software. The BSD license is described as "commercial-friendly" since it allows for-profit companies to use the software and redistribute it commercially with only attribution to the original software author. The license is described in detail on the Open Source Initiative web page here: http://www.opensource.org/licenses/bsd-license.php

spring of 2009, and includes data from the USIP and Caltrans Headquarters buildings. Due to our limited programming resources, the UI was built with technology developed for a different project at the MIT Libraries called Simile. These tools use Semantic Web standards for data representation to allow integration of heterogeneous and rapidly changing data together into a common framework. Because we did not have to develop the UI tools for FACADE, we were able to focus our efforts on developing a useful "User Experience" of the data and on improving our Project Information Model to make that User Experience more rewarding.

Building on the recommendations of our Advisory Board (among them: architects, architectural historians, technology and digital preservation experts, cultural heritage experts, and librarians) we designed the UI to have three major components: 1) a catalog of all archived buildings (including rich metadata for each building); 2) a curated "exhibit" for each building of selected items from the collection that are of high value and deserving of richer description and easier access; and 3) a more traditional, complete archive for each building collection, with poorer description and more constrained search and visualization options.

1) The building catalog is a simple Exhibit (i.e. search and faceted browsing of metadata for each of the buildings). Metadata for the catalog includes: building extent, creator, context, climate, construction system, architectural style, contributor, building type, cost, building feature, location, date, and whether or not it was built. This metadata is displayed, but can also be browsed to quickly find the buildings of interest. Metadata can be plotted as separate items (see Figure 8) or on a timeline, a map, or by thumbnail images.


FACADE projects - Mozilla Firefox
File Edit View History Bookmarks Tools Help
ITEMS · MAP · TIMELINE · CAROUSEL
sorted by: labels; then by... · grouped as sorted
Search

FACADE PROJECTS

9 Projects

CALTRANS DISTRICT 7 HEADQUARTERS

The new Caltrans District 7 Headquarters covers an entire city block downtown Los Angeles, directly opposite City Hall, in the midst of an increasingly revivified area. The building is the first to be commissioned under the State of California's Design Excellence Program. Roughly L-shaped in plan the building is composed of two main volumes. The 13 story larger volume stretches along the entire block between 1st and 2nd Streets and is set back 155 feet from Main Street. The smaller four story volume, extends from the larger volume to Main Street and occupies approximately one-half the length of the site, from mid-block to 2nd Street. Enrobed in a constantly changing mechanical skin that is alternately open or closed depending on the conditions of outside temperature and sunlight, the building's fundamental property is that of transformation. At dusk the building is transparent, textured and windowed everywhere to invite the voyeur, while at mid-day it is buttoned up against the sun, appearing to be devoid of windows entirely.

Building Features: Light-reactive metal skin, skip-stop elevators, public exhibition gallery.
Cost: $190,000,000
Building Type: Administrative Government Facility
Contributor: The Clark Construction Group, Inc.
Location Los Angeles, CA, U.S.A.

Architectural Style: Interactive
Alternative names: Caltrans
Extent 1,050,000 sq. ft.
Is built? true
Climate: Dry-Summer Subtropical Zone

Creator: Morphosis
Construction System: Structural steel framing, steel moment frames, steel reinforced concrete slabs, perforated aluminium panels, and glass.
Context: Urban
Start Date: 2002
End Date: 2004

COOPER UNION NEW ACADEMIC BUILDING

Designed largely to house Cooper Union?s Albert Nerken School of Engineering?one of the top three specialized engineering schools in the nation? the building will also provide institutional space for the Humanities and Social Sciences, the Irwin S. Chanin School of Architecture and the School of Art. The structure will function as both a space for study and a learning laboratory. To the extent practical, the

Is Built?
- false (3)
- true (6)

Extent
- 1,050,000 sq. ft. (1)
- 111,000 square feet (1)
- 150,000 square feet (1)
- 18,000 sq. ft. (1)
- 200,000 sq. ft. (1)
- 258,333 sq. ft. (1)
- 6,000,000 square feet (1)
- 605,000 sq. ft. (1)
- 720,000 sq. ft. (1)

Creator
- Gehry Partners, LLP (3)
- Morphosis (3)
- Moshe Safdie and Associates (3)

Context
- Government Campus (1)
- Small town (1)
- Urban (4)
- Urban University Campus (2)
- Urban Waterfront (1)

Climate
- Dry-Summer Subtropical Zone (2)
- Humid Continental (2)
- Humid Subtropical (2)
- Mediterranean (1)
- Oceanic (1)
- Tropical rainforest (1)

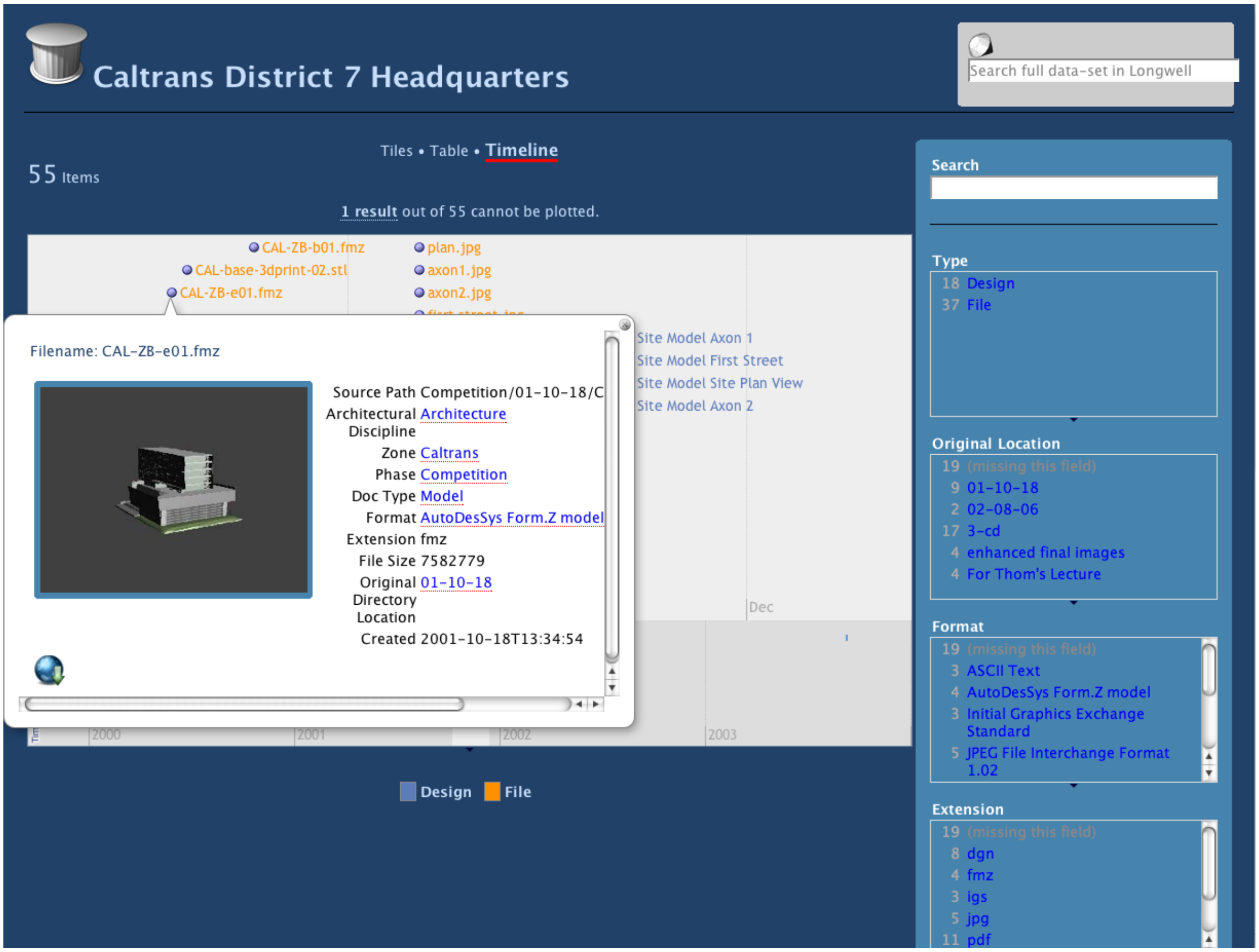
Construction System
- Custom perforated stainless steel panels attached to a tube steel frame bolted to concrete building frame. (1)
- Poured-in-place concrete, custom steel frame, hand-set brick, custom metal cladding, glass curtainwall and aluminum framing. (1)
- Skylit galleries over masonry and brick. (1)
- Stainless steel panels on steel frame (1)
- Steel and Concrete. (2)
- Steel, concrete and glass (1)
- Structural steel framing, steel moment frames, steel reinforced concrete slabs, perforated aluminium panels, and glass. (1)
- Titanium panels on steel frame. (1)

Architectural Style
- Deconstructivist (1)
- Expressionist (3)
- High Tech (1)
- Interactive (2)
- Neo-Vernacular (1)
- Postmodern (1)

Contributor
- Beacon Skanska Inc. (1)
- F.J. Sciame Construction Co., Inc (1)
- M.A. Mortenson (1)
- Morganti General Contractors (1)

Figure 8. The FACADE building catalog

2) The building “Exhibit” displays the set of "Selected Objects" that the curators identified from the complete records of the building project. As explained earlier, these are either “Design” objects (typically 3-D models or 2-D drawings or drawing sets), or else they are “Other Selected” objects such as interesting images, client presentations, key documents, and so on. Each of these objects has the usual five properties associated with it, a thumbnail graphic of some sort. The User can perform keyword searches against the text of the data and metadata, or faceted browsing of the metadata elements, to quickly find items of interest (see Figure 9).



Filename: CAL-ZB-e01.fmz

Source Path Competition/01-10-18/C
Architectural Discipline Architecture
Zone Caltrans
Phase Competition
Doc Type Model
Format AutoDesSys Form.Z model
Extension fmz
File Size 7582779
Original Directory Location 01-10-18
Created 2001-10-18T13:34:54

Type
18 Design
37 File

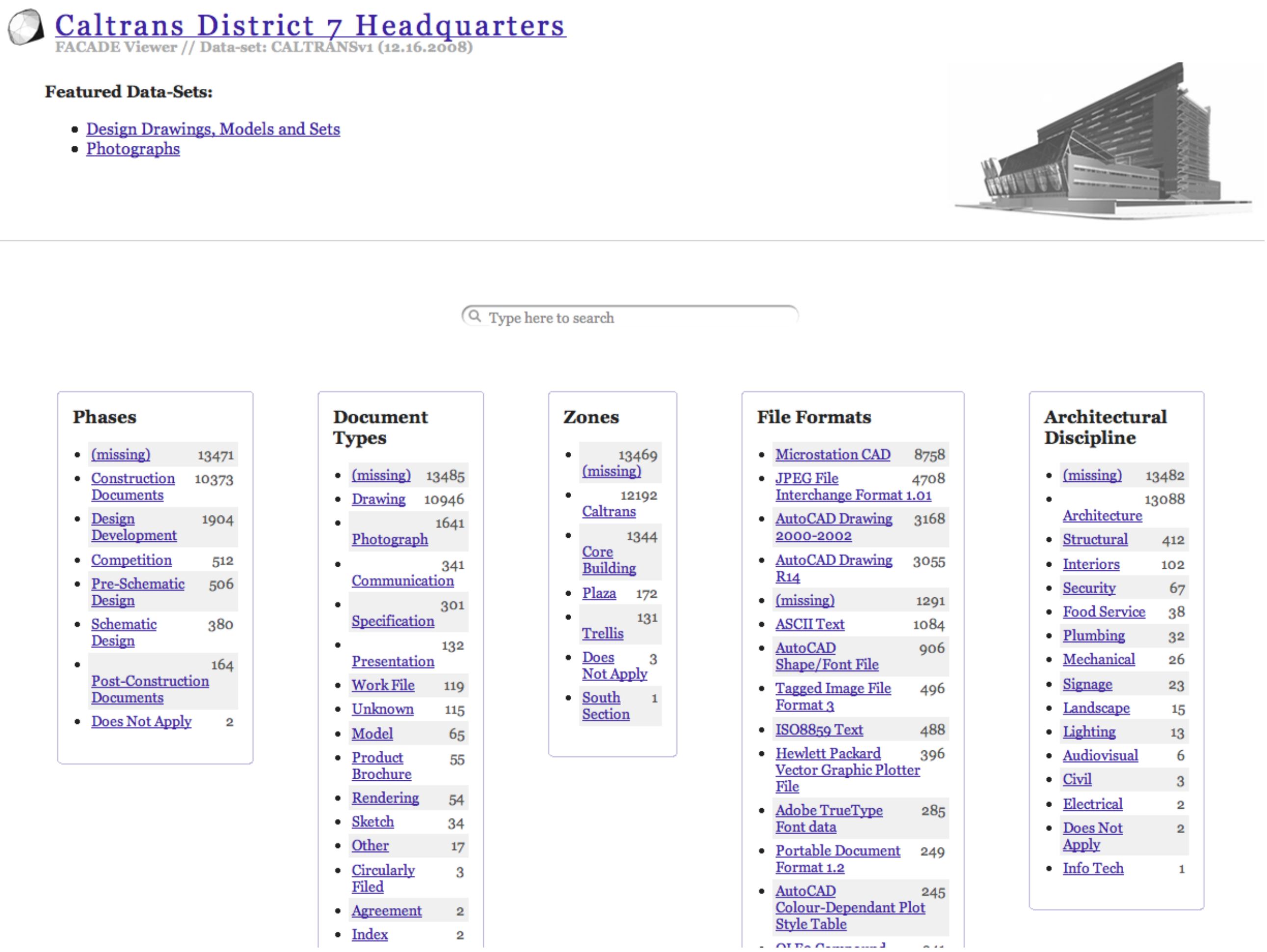
Original Location
19 (missing this field)
9 01-10-18
2 02-08-06
17 3-cd
4 enhanced final images
4 For Thom's Lecture

Format
19 (missing this field)
3 ASCII Text
4 AutoDesSys Form.Z model
3 Initial Graphics Exchange Standard
5 JPEG File Interchange Format 1.02

Extension
19 (missing this field)
8 dgn
4 fmz
3 igs
5 jpg
11 pdf

Figure 9. Exhibit of Caltrans project on a timeline; popup shows 3D PDF of original FormZ 3-D CAD model

3) The entire building collection is available via another of Simile’s faceted browsing tools, called Longwell, which scales to much larger sized collections than the Exhibit. Longwell can handle the tens of thousands of files included in each building collection and still provide the sort of display, search, faceted browsing, and data navigation provided in Exhibit. For the complete building collection we provide a “jump off” page that lists the standard five browsable properties (see Figure 10). In Figure 11 you can see the result of filtered the collection by data type of “Design”, and selecting a single record displays the available metadata for that item (i.e. the five properties described earlier).



Caltrans District 7 Headquarters

FACADE Viewer // Data-set: CALTRANSv1 (12.16.2008)

Featured Data-Sets:

- Design Drawings, Models and Sets
- Photographs

Type here to search

Phases

(missing)	13471
Construction Documents	10373
Design Development	1904
Competition	512
Pre-Schematic Design	506
Schematic Design	380
Post-Construction Documents	164
Does Not Apply	2

Document Types

(missing)	13485
Drawing	10946
Photograph	1641
Communication	341
Specification	301
Presentation	132
Work File	119
Unknown	115
Model	65
Product Brochure	55
Rendering	54
Sketch	34
Other	17
Circularly Filed	3
Agreement	2
Index	2

Zones

(missing)	13469
Caltrans	12192
Core Building	1344
Plaza	172
Trellis	131
Does Not Apply	3
South Section	1

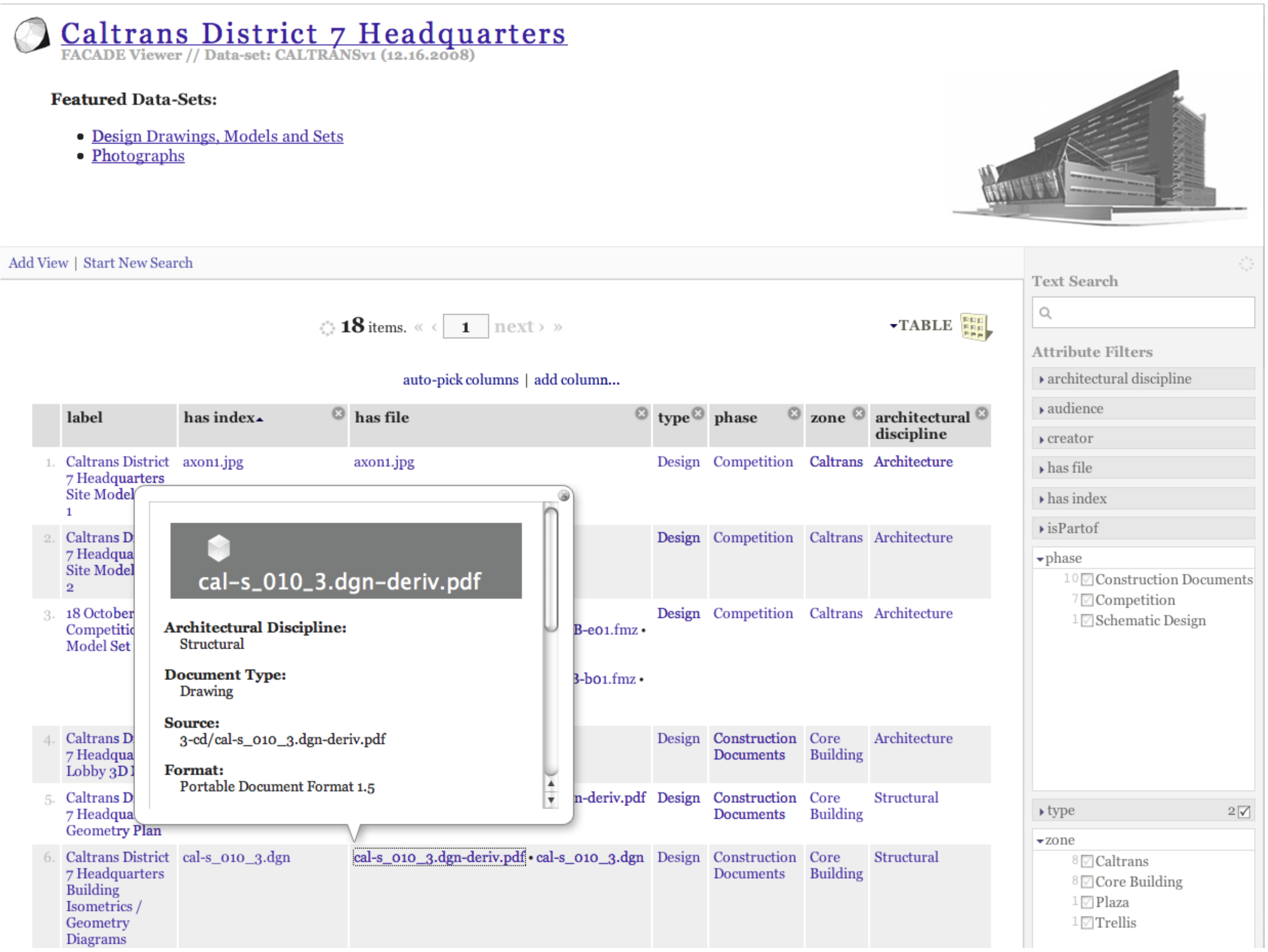
File Formats

Microstation CAD	8758
JPEG File Interchange Format 1.01	4708
AutoCAD Drawing 2000-2002	3168
AutoCAD Drawing R14	3055
(missing)	1291
ASCII Text	1084
AutoCAD Shape/Font File	906
Tagged Image File Format 3	496
ISO8859 Text	488
Hewlett Packard Vector Graphic Plotter File	396
Adobe TrueType Font data	285
Portable Document Format 1.2	249
AutoCAD Colour-Dependant Plot Style Table	245
OLE2 Compound	241

Architectural Discipline

(missing)	13482
Architecture	13088
Structural	412
Interiors	102
Security	67
Food Service	38
Plumbing	32
Mechanical	26
Signage	23
Landscape	15
Lighting	13
Audiovisual	6
Civil	3
Electrical	2
Does Not Apply	2
Info Tech	1

Figure 10. "Starting Points" page for Caltrans Headquarters building in complete collection UI



Caltrans District 7 Headquarters
FACADE Viewer // Data-set: CALTRANSv1 (12.16.2008)

Featured Data-Sets:

- Design Drawings, Models and Sets
- Photographs

	label	has index▴	has file	type	phase	zone	architectural discipline
1.	Caltrans District 7 Headquarters Site Model 1	axon1.jpg	axon1.jpg	Design	Competition	Caltrans	Architecture
2.	Caltrans D 7 Headqua Site Model 2			Design	Competition	Caltrans	Architecture
3.	18 October Competitic Model Set		B-e01.fmz • 3-b01.fmz •	Design	Competition	Caltrans	Architecture
4.	Caltrans D 7 Headqua Lobby 3D I			Design	Construction Documents	Core Building	Architecture
5.	Caltrans D 7 Headqua Geometry Plan		n-deriv.pdf	Design	Construction Documents	Core Building	Structural
6.	Caltrans District 7 Headquarters Building Isometrics / Geometry Diagrams	cal-s_010_3.dgn	cal-s_010_3.dgn-deriv.pdf • cal-s_010_3.dgn	Design	Construction Documents	Core Building	Structural

cal-s_010_3.dgn-deriv.pdf

Architectural Discipline:
Structural

Document Type:
Drawing

Source:
3-cd/cal-s_010_3.dgn-deriv.pdf

Format:
Portable Document Format 1.5

Figure 11. Entire collection, filtered by type "Design"; Popup record for chosen item.

The two interfaces (Exhibit and entire collection) are very similar in functionality. The key difference is scale. When a user is casually seeking a popular item for a building – its final 3-D model, or the initial client presentation – then searching through a collection of a hundred thousands files can make that difficult. We wanted to present a simple, quick UI for the 1% of the collection that is wanted 90% of the time. However historians and other researchers will need access to everything, and will have the patience to look through hundreds of files for what they need. The UI to the entire collection is provided for them.

To connect the two interfaces we provide a full-text search box in the Exhibit UI that searches the complete collection and moves the user to the fuller UI to see the search results. The usage scenario for this design was a researcher who looks through Exhibit, sees the models and key correspondence, and decides to explore the rest of the collection around a particularly controversial feature of the building (a particularly interesting roof design, for example). By doing a keyword search on the feature of interest (either the word "roof", or an RFI number for the correspondence on that feature, or any other term related to the feature) the researcher gets a (typically large) set of results in the complete collection UI and can begin to browse and refine their results there.

This UI design is very flexible in that any component of it can be changed or disposed of easily, with any other UI software tool that supports the RDF data model. If the curated Exhibit view proves too time-consuming to create, we can simply drop that component of the UI and keep the complete collection view.

This UI has undergone three rounds of focus group review, but we still plan to do more usability testing with target audience members in future, since there are outstanding questions about the value of some of the properties we identified and other metadata we kept. For example, the location of the file original file system as received from the architectural firm may provide clues about relationships in the collection that would not be evident from the data itself. But since such collections have never been available to researchers before, this and much else should be tested with them as the archive grows.

The FACADE Archiving Workflow

The workflow developed for FACADE we refer to as the “archiving workflow”, but it might also be called the "publishing" workflow as it includes the end-to-end processing of received data through archiving to DSpace and on to the end user interface. Figure 12 shows the major workflow steps, involving diverse staff to process the new collection into the FACADE system and perform the annotation and selection described above.

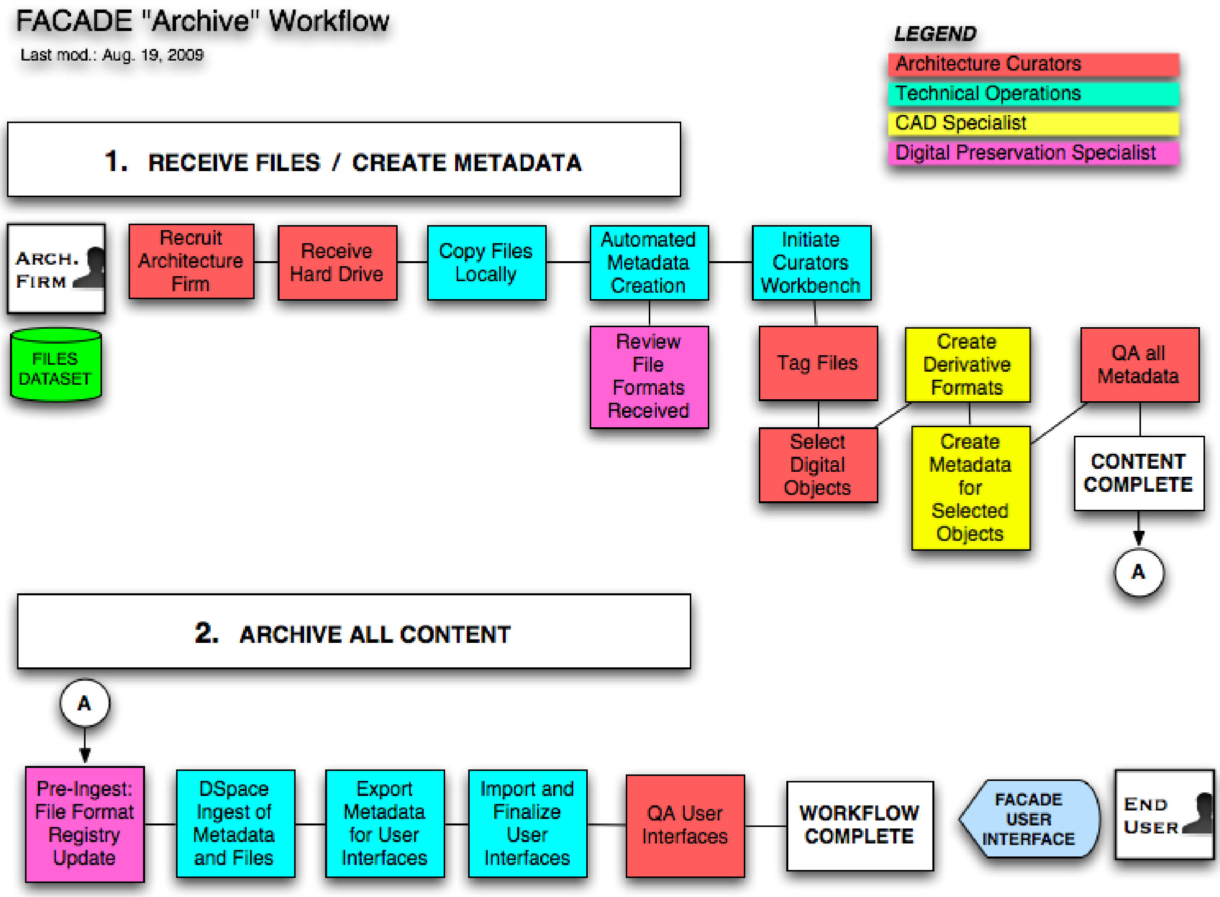

FACADE "Archive" Workflow
Last mod.: Aug. 19, 2009
LEGEND
Architecture Curators
Technical Operations
CAD Specialist
Digital Preservation Specialist
1. RECEIVE FILES / CREATE METADATA
ARCH. FIRM
FILES DATASET
Recruit Architecture Firm
Receive Hard Drive
Copy Files Locally
Automated Metadata Creation
Review File Formats Received
Initiate Curators Workbench
Tag Files
Select Digital Objects
Create Derivative Formats
Create Metadata for Selected Objects
QA all Metadata
CONTENT COMPLETE
A
2. ARCHIVE ALL CONTENT
A
Pre-Ingest: File Format Registry Update
DSpace Ingest of Metadata and Files
Export Metadata for User Interfaces
Import and Finalize User Interfaces
QA User Interfaces
WORKFLOW COMPLETE
FACADE USER INTERFACE
END USER

Figure 12. Workflow for FACADE Processing

During the process of designing the archiving workflow described above, we identified a few additional workflows that would usefully serve various curator-driven use cases in future. These include:

- A "preview" workflow to allow curators working on metadata to be able to see results in the user interface prior to publication;
- A "post-publishing" workflow to support revisions to an existing collection, for example to add newly received data, fix mistakes, or remove files that triggered to intellectual property issues;
- A "license" workflow to track the end-to-end activities, technical and other, to secure licensing and related intellectual property considerations for a project and to document the process.

Each of these additional workflows can be done now, but very clumsily, and we hope to streamline them in a future phase of the project.

During the project, the archiving workflow was tested via several "dry runs" with staff of the MIT Libraries. The purpose of the dry runs was to test the reliability of the workflow and to estimate time and effort requirements for processing new collections. This proved to be an essential part of the project since it helped us identify remaining gaps in the process and things that could be done more efficiently, as well as documenting that it was, in fact, possible to process a new collection of large size in a matter of weeks. This was particularly valuable for the curation staff, since the idea of bulk processing digital collections via the Curators' Workbench was novel and foreign. The process we designed will take some time to become familiar and comfortable to library staff, but we have proven to our satisfaction that they can and will be able to use this process in a fully operational setting.

CAD Preservation

As demonstrated by the project's bibliography provided below, a major focus of FACADE was on strategies for preserving 3-D CAD models into the far future. Descriptions of why that is challenging have already been provided. Here we focus on what can be done today, and on our general recommendations.

CAD Preservation Strategy Recommendations

In light of the scale and wide variety of materials received, we arrived at a set of recommended best practices for a reasonable preservation strategy for all the materials received in a building collection. This includes:

- Special processing of 3-D CAD models to generate derivative versions with greater long-term archiving potential than the native software format (see below);
- Semi-automated conversion processing of other key design file formats (e.g. 2-D drawings into PDF);
- Automated conversion processing of common digital file formats (e.g. Microsoft Office documents and JPEG images) as part of archive ingest;
- No processing for remaining classes of file formats; although these will come under more generalized digital repository preservation strategies outside the scope of FACADE's focused concerns.

For 3-D CAD models we identified the need for four versions with distinct formats to insure long-term preservation. These are:

1. Original (the originally submitted version of the CAD model)
2. Display (an easily viewable format to present to users, normally 3D PDF)
3. Standard (full representation in preservable standard format, normally IFC or STEP)
4. Dessicated (simple geometry in a preservable standard format, normally IGES)

Appendix 1 is a report we created to provide instructions for creating derivative versions of 3-D CAD models for some of the more popular CAD software we encountered in our test collection. Given the high rate of change in the CAD industry, these instructions are necessarily an ongoing work-in-progress that we will add to and edit as CAD software and standards evolve. In the document we provide a brief rationale for recommending the IFC and STEP ISO standards for the "Standard" version, and the IGES standard for the "Dessicated" version. 3D PDF was chosen for the "Display" version since it is not needed for long-term preservation and is natively supported in modern Web browsers for 3-D display. We believe that it is important to keep the original 3-D model as well, both for authenticity purposes and because most native software is still improving on export capabilities so that it may be possible to create even better standard export versions from the originals in future.

More discussion of these preservation strategies and their rationale can be found in the articles listed in the project bibliography below, and particularly "Curating Architectural 3D CAD Models" (2008).

CAD Software Emulation Recommendations

Contemporary CAD software systems are usually designed to run on PCs with the Windows operating system. There is often to UNIX or Mac version available, and certainly the 3-D CAD models we received in the test collection were produced on PCs. These software systems are commercially sold, and often require a "license key" to enable that are provided by the vendor to the customer, and are time sensitive (e.g. a key might be good for one year, after which time it "expires" and the software becomes unusable to that customer).

For the FACADE Project we were able to acquire all the CAD software products that were used by the architects who contributed to the research test collection, and we had valid access to those products throughout the project. Should an archive need to keep CAD software in perpetuity to view older CAD models, that archive would need to continue to buy license keys for the software forever, and hope that those CAD companies don't go out of business. This is obviously not a realistic strategy for long-term preservation, yet ideally we need access to that software for many decades. We have briefly discussed this issue with several of the leading CAD software companies (e.g. Autodesk and Bentley) and they are open to the idea of escrowing unrestricted copies of the software with appropriate libraries and archives, so we feel that is the best avenue to pursue.

As for software emulation, we performed a detailed case study of that strategy for the AccuDraw software on the Apple II platform (long since obsolete) and were able to view AccuDraw models by running the software in a virtual machine[18] environment. We documented the process and lessons

[18] The distinction between emulation and virtualization is important: the former requires the original software to be rewritten to run on a modern computing system (it is "emulated" in the new environment), while the latter requires the original software to run in an emulation of the original computing system (e.g. the Apple II). Both approaches have their advantages and disadvantages, but the latter proved far more practical given the highly proprietary and complex nature of CAD software and the unlikelihood of being able to recreate it for a new platform.

learned in detail, and we feel it is a viable *technical* approach for preserving modern CAD software and data, but the issue of legal access to the software via license keys is a significant barrier.

7. Outreach

> *Documentation, training, outreach and dissemination of results to the digital library, digital preservation, and DSpace user communities.*

Over the course of the project, members of the FACADE team have spent significant time on outreach and education to a range of audiences about the project's goals and findings. These included architects and architectural firms; technology companies supporting the AEC industry; library, archives, museum and higher education staff; digital library and digital preservation specialists; and others as the opportunity arose. A list of presentations and papers follows.

Smith, MacKenzie. "Future-Proofing Architectural Computer-Aided Design: MIT's FACADE Project." *Architecture and Digital Archives: Architecture in the digital age: a question of memory*. Ed. Peyceré, D. and Wierre, F. Paris: Editions InFolio, 2007. 409-423.

Smith, MacKenzie. "Curating Architectural 3D CAD Models." *International Journal of Digital Curation*, 4(1). December 2008. http://www.ijdc.net/index.php/ijdc/article/view/105/80

FACADE: MIT Libraries CAD and BIM Preservation Research Project
Presentation to the National Collegiate Facilities Management Technology Conference (NCFMTC) by William Reilly, Technology Projects Manager, MIT Libraries. Cambridge, MA; August 2007.

Archiving of Digital Design Data: Formats for Long-Term archiving. Presentation at the American Institute of Architects (AIA) Annual Convention by MacKenzie Smith, Associate Director for Technology, MIT Libraries. Boston, MA; May 2008.

FACADE: Future-proofing Architectural Computer-Aided Design. Presentation at the Society of Architectural Archivists Research Forum by MacKenzie Smith, Associate Director for Technology, MIT Libraries. San Francisco, CA; August 2008.

Preserving Brand-new Buildings: Digitally Archiving 3D CAD and Related Architectural Materials. Presentation at the Digital Library Federation Fall Forum by William Reilly, Technology Projects Manager, MIT Libraries. Providence, RI; November 2008

Curating Architectural 3D CAD Models. Presentation to the 4th International Digital Curation Conference by MacKenzie Smith, Associate Director for Technology, MIT Libraries. Edinburgh, Scotland; December 2008

Crossing the Curatorial Chasm - *Lessons from the FACADE project.* Presentation at the 4th International Conference on Open Repositories by William Reilly, Technology Projects Manager, MIT Libraries. Atlanta, Georgia; May 2009
http://hdl.handle.net/1853/28505

FACADE Revisited: Future-proofing Architectural Computer-Aided Design. Presentation to Moshe Safdie and Associates staff by the FACADE Project Team. Cambridge, MA; June 2009

MIT's FACADE Project: Future-proofing Architectural Computer-Aided Design. Presentation to the Society of Architectural Archivists (SAA) by Tom Rosko, the MIT Institute Archivist. Austin, TX; August 2009.

In additional to these formal papers and presentations, members of the project team have had numerous ad hoc meetings with other staff at universities, in their library, archives, or architecture school. As the project concludes, we are developing a final version of the project's Website to capture our findings and recommendations, and to post information about future developments.

8. Other Notable Issues

Two final issues should be noted, since they were not anticipated and have bearing on the future of our project and its larger goals for libraries, archives and museums.

First is the legal landscape that we encountered for these digital building collections. Many types of collections acquired by library, archives and museums have intellectual property and other legal challenges associated with them, and architectural data is among the most difficult type of material in this. The collection data itself raises many concerns for architects: legal liability, potential for functional misuse by others, potential for loss of creative credit, and so on. These concerns are somewhat offset by the equally strong desire to see the designs survive, to leave a legacy and to retain credit even after the physical buildings are long gone. What makes digital collections so different is that the library or archive does not need to acquire the intellectual property rights to the material in order to archive, preserve and publish it… they merely need a license from the architect to get a copy of the collection for that purpose. During the project we drafted such a non-exclusive, royalty free license for review by the architectural firms we worked with, but were not able to complete that negotiation in this timeframe. We believe that the architects will ultimately agree to these terms, probably with exclusions and/or embargo periods for some of the more sensitive data (e.g. contracts with clients). But we anticipate months, if not years, of negotiations between lawyers before agreement is reached, and setting the right precedent for architectural collections in the digital environment is *critical* if we hope to afford such collections in future.

In addition to the legal complexities of the data itself, we have elsewhere in this report described the difficulty of collecting the software products used to create the key designs. The products are expensive, complicated to learn and use, rapidly changing, proprietary and encrypted (via license keys). All of these barriers can be overcome for some few products, but it is difficult to imagine doing so for the hundreds of products potentially needed. From this we anticipate the need to negotiate terms with a few of the leading vendors (e.g. Autodesk) to escrow unencrypted software with a trusted archive, and to use that precedent with other vendors to do likewise. There are organizations like the American Institute of Architects whose internal archives receive copies of most CAD software in use, so working out arrangements with them and the vendors is a useful strategy for this problem. The library, archives and museum community will then need to cooperate on who has what software, since it is unlikely that each archiving organization can maintain copies of all of the software they might ever need.

The second unexpected issue relates to the potential for broad adoption of the FACADE system and workflow by other archiving institutions. MIT is fortunate to possess a nexus of expertise in digital libraries and archives, architecture, and technology. We have faculty on one side of campus inventing the next generation of CAD software, and on the other side of campus applying it to architecture, and a library in the middle working with both, and with a high degree of technical sophistication and resources. Based on earlier projects conducted by the MIT Libraries in this domain (e.g. DSpace, Simile, etc.) we anticipated enthusiastic adoption of our software by other institutions that we know to be facing the same problems with digital design data. What we found is that while the problem is widely shared, the expertise and resources to solve them is very rare. Libraries, archives and museums specializing in architecture have few or no programmers or technology staff who can implement and manage a digital archive (no matter how simple), nor do they have access to architecture experts who can collaborate with them on reformatting CAD models for long-term preservation. We found that while there is great enthusiasm in the community about our work, we and a very small number of other organizations are the only ones who currently have the capacity to run such an archive.

This finding means that while we have made all of the DSpace code customizations and other software available as Open Source Software, and will advertise its availability on the project's website and in the community, we doubt if there will be much use of it outside of MIT and a few other organizations. MIT is planning to create its own production archive of digital architecture, and we will focus our efforts on making it as comprehensive as possible, so that these digital collections are not lost forever while other archival institutions find the resources they need to make the switch to digital archives.

Conclusions

The FACADE Project was very ambitious in its conception, dealing with one of the most complex challenges found in the field of digital libraries and long-term digital preservation. We are therefore pleased with the quality of the results provided here, and believe that the solid base of research produced will lead to a useful ongoing research program as well as real operational digital archives of digital architectural collections. We continue to believe that this project has been unique in its scope and goals: developing strategies for archiving, preserving, and making available the digital products of major 21st century architectural projects. And our work on the common data types of architecture, i.e. complex digital content such as 3-D CAD models, has general applicability well beyond architecture. Also generalizable is our work on processing workflows for large, heterogeneous archival collections (e.g. hard drives from companies loaded with archival data files). Finally, the platform we developed to provide end user access to the digital archives can, of course, be re-purposed to serve many digital archives.

All FACADE project documentation and products are available via the project's website and wiki at http://facade.mit.edu/

MacKenzie Smith, Associate Director for Technology, MIT Libraries
August 14, 2009