Conference Presentation Proposal for Open Repositories 2009

Crossing the Curatorial Chasm - Lessons from the FACADE Project

MIT Libraries

FACADE ("Future-proofing Architectural Computer-Aided DEsign") - a two-year IMLS grant funded research project undertaken by MIT Libraries, along with the School of Architecture and Planning, was charged with investigating how best to archive the highly proprietary, internally complex, and potentially short-lived digital artifacts of contemporary 3D CAD modeling tools. The methodology was strenuously empirical, rather than theoretical: real, unfiltered, building project datasets from prominent architectural firms were used to define the scope and direction of the work. This lead to challenges both anticipated and unanticipated. This presentation will share analytical results, demonstrate tools developed in the course of the project, and discuss both lessons learned and future directions.

Anticipated Challenges

1. 3D CAD Preservation

In the field of digital preservation, 2D drawings produced by products like AutoCAD present few new problems since they are industry standards (i.e. the DWG and DXF formats) and they can easily be converted into other standard 2D formats such as JPG or PDF. 2D drawings are also not significantly interactive in their native systems, so requirements for their long-term preservation are similar to other static formats like visual images. For situations in which 2D CAD is the best representation of the object, standard formats and migration preservation strategies can reasonably be applied.

However 3D CAD is a different story. 3D models are created in proprietary software using non-standard native formats, and each product uses different techniques for capturing a model's shape information via designer-specified parameters, storing geometry and other properties attached to the geometry, and rendering the model on the computer screen. This results in several challenges:

1.1 3D Format Identification

FACADE utilized and extended several digital preservation tools involved in fine-grained format identification (viz. DROID for identification, and PRONOM as the external format registry) and published newly discovered formats – both proprietary and open standard – for others to leverage. This publication of new formats will also further facilitate the use of third-party registries to help monitor the status of 3D formats and identify migration tools as necessary.

1.2 3D Format Migration

Data acquired by the project was predominately in proprietary native formats produced by the CAD software tools in use. Given the difficulty of working with those formats over time, FACADE evaluated the available standard 3D formats, and made recommendations for a format migration strategy to minimize the risk of permanent data loss. For each 3D file ingested, a standards-based version in either IFC or STEP is created by exporting the file from the original software in the standard format. This is a manual process requiring expertise in both the native software and its underlying data model (e.g the CAD model tree) to create useful standard versions. Automation of this export process is

an obvious area of future research. In addition to the standard STEP or IFC format, another, simpler standard format – IGES – was created to capture the surface geometry and parametrics of the original model. This produces a less functional version than STEP or IFC, but is less prone to translation errors during export. Finally, a presentation version was created that is readily accessible on the Web – 3D PDF – with the understanding that this version will probably need to be replaced fairly often as Web formats for 3D evolve. This set of artifacts – the original, a 'standardized' derivative, a 'dessicated' (preservation-friendly) derivative, and a display copy formed the foundation of the preservation strategy.

1.3 3D Format Emulation

The project also explored the implementation of a software emulation framework to support native CAD files requiring original software to interpret. Current virtualization and paravirtualization platforms (e.g. VMWare, QEMU, XEN) were considered. Software licensing and other considerations made this emulation approach less immediately viable, although discussions with CAD software vendors held some promise.

2. Project Information Modeling

One important FACADE finding was that, at least for large architecture projects, a 3D model is of most value to a designated community (e.g. future researchers, historians, design professionals) if it is available in some context that helps to explain the design intent it implements, and any problems that arose from the design during construction or use of the physical artifact. Achieving this involves creating a “building project” collection, that relates the 3D model(s) to other data from the project. This contextual structure became known as the “Project Information Model” (PIM), and was realized in an RDF ontology, which also proved to be pivotal in the design of both end-user discovery interfaces into the collection, and curators' tools. The PIM also was informed by a nascent building industry approach called the 'BIM' (Building Information Model), which proved suggestive but lacked sufficient maturity and uptake in industry to materially benefit this effort.

Unanticipated Challenges

1. Data Scale and (Dis-)Organization

The building construction datasets received from architecture firms contained a surprisingly large amount of project information in addition to the centerpiece 3D CAD models. There were many varieties of related work artifacts including presentations, communications, reports, project management, schedules, indices, master lists, images, multimedia, etc. Cumulatively, these additional artifacts numbered in the many tens of thousands for commercial projects. Domain experts have deemed virtually all of these artifacts as meriting preservation. Ideally, surfacing the relationships between artifacts and models, and the relationships between artifacts, would form a part of the repository functionality. However, the file system used by the architectural firms as the organizing mechanism for this content was found to be inadequate. For example, metadata, tags, and links were virtually never provided. Efforts to structure directories, and to provide conventions for file and directory naming were attempted, but with almost always inconsistent or at least unreliable results. More traction was obtained by attaching PIM metadata to the constituent files (e.g building phase, document type, discipline) and the project adopted this strategy.

2. Automated and Partial Metadata Generation

A direct consequence of the sheer number of artifacts involved was the prohibitively high cost of creating metadata manually, as it is traditionally practiced in cataloging and repository ingest. FACADE devised several workflows, methods and tools that allowed for automated tagging of artifacts, using contextual, heuristic, or other techniques. These were supplemented with curatorial tools e.g. the “curator's workbench”, a web based application that enables domain experts or library staff to enrich the metadata in efficient, semi-automated, and collaborative ways. Thus artifacts were allowed differing degrees and types of metadata, from minimal to highly described and linked. This process was enabled by the open nature of the PIM data model (an RDF-graph): a rigid, schema-driven data model would have impeded such an effort.

3. Discovery and Exploration (Visualization)

Standard models of visualization proved to be inadequate for collections of this scale. FACADE developed and prototyped several novel approaches to facilitating access. Certain 'high-value' objects were identified, assigned richer metadata, and linked through the 'Project Information Model' to yield browsable collection subsets that shine a spotlight on desirable resources. These interfaces also included the ability to link back to the deeper well of the entire collection, so that more specialized navigation through the corpus would still be possible. These web-based visualizations were constructed using MIT SIMILE project tools for RDF-backed faceted browsing.