DIGITAL ARCHIVAL OF CONSTRUCTION PROJECT INFORMATION

By

DeWitt Latimer IV[1], Chris Hendrickson[2]

ABSTRACT

The Digital Library Initiative (DLI) concerns itself with the automatic creation, organization, and indexing of complex collections of data. With its own wealth of data types, the construction industry pushes the boundaries of the current technologies, especially with regards to automatic archival of construction project information. Many issues unique to the construction projects, such as data provenance and multi-media searching, require the integration of many technologies being developed at the various DLI funded research universities. This paper seeks to define those issues, the research being done, and provide some direction for future work in this area.

KEYWORDS

Digital library, document management, project document archival

1. INTRODUCTION

Construction projects generate volumes of information on a regular basis. However, access to this information can be difficult or awkward. Typically a person must go to a document area and rifle through papers, copies, or other documents to locate a specific document. Even worse, if a specific item is to be investigated, then the researcher must know where to search for information. If searching for information about a defect in the built entity, the researcher's information may be in the construction plans, meeting minutes between the contractor and the owner, in a bill of materials for parts relevant to the defect, inspector's reports, or even in a pre-construction environmental study. This problem is exasperated by the many media types common in the construction industry, application data files (such as CAD models, project schedule databases, or analysis calculations), images, structured formal documents, analysis results, and even audio or video. Although current web search technologies are becoming more sophisticated, these methods do not necessarily extend well to multiple distinct types of media. In a sense all these various media are the "documents" of the project.

This search problem is even an issue for construction document management. Current efforts in document management systems focus on 4 main issues: central storage of documents, routing of documents to interested parties, document version control, and security of documents. On the other hand, archiving is not only concerned with the saving of all versions of all documents over time but also indexing and providing an interface that allows for search across all the documents. Needless to say, the additional effort, and therefore cost, needed to properly archive documents is outside the budgets of construction projects. What is needed is a means of automatically generating archive documents quickly without much effort.

The digital library initiative [4] is sponsored by many federal agencies, including the National Science Foundation, DARPA, NASA, the National Library of Medicine, and the Library of Congress, among other federal agencies and private sector interests. Currently in its second phase, the goal of this program is to develop digital library technologies on a broad base to support the growing need for access to information. Many university research efforts are sponsored through this initiative. These universities are developing the base technologies and sciences to support the generation of these archives. Much attention is being paid to multi-media archives and archives with little, user-provided structure.

The remainder of this paper is organized in 4 sections. Section 2 provides the goals for archiving the documentation of a construction project as well as a taxonomy of the types of documents for typical construction projects; addressing these goals are the requirements for successful implementation of a construction digital library. Section 3 briefly outlines some commercial systems to indicate where these systems fail to meet the goals. Section 4 outlines some of the current research being performed that may be applicable to solving the issues in archival of construction projects. Section 5 closes the document

[1] dl4s@andrew.cmu.edu
[2] cth@andrew.cmu.edu
Department of Civil and Environmental Engineering
Carnegie Mellon University
5000 Forbes Ave
Pittsburgh PA 15213

by stating some conclusions from the material presented and suggests directions for future work.

2. ARCHIVING A CONSTRUCTION PROJECT

2.1 Goal

The goal of a construction project archive is to provide a searchable data repository that represents both the final constructed artifact and the process that was used to construct the artifact. This archive must be constructed from the materials provided and yet must minimize the user's involvement in the creation of indices or other organizational tools (including keywords) needed to operate the archive. Care must be taken to ensure that the archive can handle the wide scope of items to be covered as well as able to manipulate the digital representation of those items (i.e. the various media formats).

2.2 Construction Information

Table 1 indicates a list of items that can be expected on a typical construction site. This table was inspired by the work of de la Garza and Howitt [18].

Request for information	Materials Management	Equipment Management	Cost Management	Submittals	Safety	QC/QA	Interpersonal	Schedule means and methods	Jobsite record keeping	Future Trends
Design intent and clarification	Access to material management	Equipment location	Budget	Test results	Accident reporting	Initiate inspections	Emails	Schedule	Recording timesheets	Positioning data
Subcontractor information	Material location	Fuel monitoring	Material cost accounting	Revisions to submittals	Reporting violations	Report QC/QA problems	Voice Mails	Schedule updates	Progress reporting	Sensory data
Contract specifications	Material order status		Equipment cost accounting	Physical Artifacts	Safety Plans	Reporting inspections results	Meeting Minutes	Delay recordings	Exception reporting	
Contract drawings	Request materials to site		Personnel cost accounting	CAD Models		Test Artifacts		As-built records	Visitor's log	
Work package information	Place material orders		Purchase Orders			Test Plans		Productivity information	Daily Constructio n Reports	
Means and methods	Material Specifications					QC/QA Plans			Images / Photos	
Implementation problems										

Table 1 – Typical documents of a construction project, by type

From the construction documents listed in Table 1, a list of digital data/file types was generated [Table 2]. File types are the digital representation of documents, as they would be stored in a digital archive. It is essential for any archival system to cover the breadth of these file types, given that each may contain information that is only referenced in that specific type of file.

Text	Digital Images	Application Data Files	Voice	Video	Formal Structured Documents
ASCII	BMP	MS-Project Files	WAV	MPEG	XML
MS-Word	JPG	VRML	MP3	AVI	CIM-Steel
PDF	TIFF	Pro-E	Real Audio	Real Video	spread sheets
Word Perfect	GIF	AutoCAD	AU	MOV	CIS/2 documents
…	…	…	…	…	…

Table 2 – Digital file types by application

A system must eliminate or mitigate all the fallacies, handle all the formats and document types would be a necessary requirement/criteria for successful implementation of a construction digital library.

3. COMMERCIAL SOLUTIONS

Several commercial products exist that provide document management services, but do not pass muster as an automated archival tool. Yet these products are likely to evolve over time, as the research described in section 4 matures.

JobDocs.com [8] provides an image management system for clients. While this service provides an engine for storing digital imagery about a job site, it does not automatically link the images with other project information, such as CAD models, product information, or meeting minutes. Additional storage of image versions is apparently handled by the user and not provided by the system.

Autodesk [1] provides a project collaboration environment for the management of print materials. While the environment does provide for managing document versions, the service does not provide for searching the contents of the documents themselves.

Meridian project systems [9] provides products for project management. The ProjectTalk [13] suite offers document management [14] functions. However, like Autodesk, this system is limited in its approach to searching for information. ProjectTalk behaves more like a file repository than a searchable archive.

In addition to the above systems, Amor/Faraj [17] provides a look into some of the reasons that integrated

project databases (document management systems) fail. After some revision and simplification, we can identify 10 fallacies that may lead to a failure of integrated project databases:

1. An object-oriented system provides the complete solution;
2. A coordinated model solves representation problems;
3. A complete model of reality required;
4. User views are reconcilable;
5. The Internet solves the communication problem;
6. Printed documents will disappear;
7. CAD is the center of the archive;
8. No data ownership problems exist;
9. A database guarantees coordinated, consistent information;
10. The industry is ready to adopt such databases.

We will address these points in turn.

An object-oriented system provides the complete solution - From a computer science perspective, object-oriented system development is a method to develop software systems, and in itself does not solve any problems.

A coordinated model solves representation problems - Coordinated model(s) for very complex systems are difficult to create as can be evidenced by the slow development of industry foundation classes (IFC) [20] and the standard for the exchange of product data (STEP) [19] product descriptions; plus these representations, and systems based on these representations, will need constant updating as building technologies change.

A complete model of reality require - Completely modeling the reality of the site is not likely to be possible, however, any candidate system should be able to operate in a partial-information world.

User views are reconcilable - User views may not be reconcilable (certainly unless there is an information model).

The Internet solves the communication problem - The Internet does not, by itself, solve data communication issues, but is a tool that can be used to build systems that solve the problem.

Printed documents will disappear - Paper documents will likely still exist even with an easy access digital system.

CAD is the center of the archive - While useful for laying out spatial problems, CAD also cannot serve as a central data model for items such as equal opportunity reports, administrative budget items, and other items may not be linkable directly to structures.

No data ownership problems exist - In the business world, not all data may be viewable to all users, as the data may be proprietary, sensitive, confidential, or otherwise restricted for legal or business reasons.

A database guarantees coordinated, consistent information - Databases do not inherently solve the data consistency issue; by themselves databases do not guarantee the consistency and coordination of data, but respond to the data models with which they are designed.

The industry is ready to adopt such databases - Finally, the construction industry is notoriously IT poor; end users may not be receptive or have the inherent computing capability to support and use these IT systems.

4. DIGITAL LIBRARY TECHNOLOGIES

Many universities are participating in the Digital Library Initiative, however research at seven universities (Carnegie Mellon, Columbia, Cornell, Penn, Stanford, Berkeley, and UC-Santa Barbara) is quite relevant to the development of a digital archive of construction projects. These projects were noted because they directly answered one or more of the challenges presented in the previous section.

4.1 Carnegie Mellon University

The Informedia-II Digital Video Library [6] project at Carnegie Mellon University (CMU) focuses on building digital libraries from video sources. The project is automatically indexing and providing searchable interfaces to video news media. These indices are built by segmenting the video stream and using speech recognition to automatically generate transcripts of the sessions. In addition to normal textual querying of the archive, visual interfaces are being developed to query the archive with an image as the key. This visual querying works by having the user selecting a figure of interest in the video stream (typically a face) and then asking for matches. This provides a multi-media querying interface.

In a related project at CMU CCRHE (Capturing, Coordinating, and Remembering Human Experiences) [7], the video stream is augmented with sensory and position data. In this case, querying for video can be done by querying for the geographic position. As intended, this would be for looking for a memory of a trip; however in a construction application this capability could be used to ask for video from the location of a specific building feature.

4.2 Columbia University

PERSIVAL (PErsonalized Retrieval and Summarization of Image, Video And Language resources) is a patient care digital library [10] being developed at Columbia. This library take a pre-existing user model of health care professionals and provides an interface that attempts to predict the user's needs and interests. The information presented is from an online collection of patient records maintained by the hospital.

These records, in addition to information brought in from journal and web pages, are on distributed systems

covering many different media (voice, image, text, etc.). Attention must be given to fusing together potentially conflicting information; methods such as automatically identifying source type, quality, and level of intended audience is needed so that the system can present information appropriate for the type of user performing the query (from the user model).

Other component technology to be developed are multi-modal query input, automatically augmenting queries based on current patient records, search and presentation of multi-media resources, supporting browsing over automatically constructed categories, graphical layout of multi-modal material, presentation and summarization of multiple relevant documents, merging repetitive information, and generation of appropriate level summaries.

The user based model approach may be helpful in the construction context since different participants have very different interests.

4.3 Cornell University

The Prism project [12] at Cornell focuses on digital library research in many areas. Some areas that are most relevant to construction project archiving are digital object architecture, human-centered research, and policy expression and enforcement.

The FEDORA (Flexible and Extensible Digital Object and Repository Architecture) project focuses on reliable and secure means to store and access digital content. This project seeks to address the creation of objects that aggregate heterogeneous types of data from distributed sources. Then the architecture allows objects to have global and/or domain-specific behaviors. The architecture is designed to support multiple viewing modalities, based on client access. The architecture utilizes a rights management model to aid in the creation of a "view of an object" contents.

The human-centered research group has deployed a field access system that utilizes global position sensing (GPS) and laser tracking (for indoor applications) to provide context-aware delivery of resources. On a construction site, this may be used to queue the plans to the room in which the person is currently located. Further, the group has studied the search patterns of users based on resource availability, relative expertise, and user characteristics (such as gender), which can greatly help deliver the right information in the field to individuals.

The policy expression and enforcement group focuses on a suite of tasks to deliver automated policy enforcement for digital libraries. The first task is generating a formal definitions and declarations, which provides a policy specification. Then, automated enforcement agents must be capable of understanding and acting upon the declarations. These behaviors manifest by the system allowing and encouraging permitted actions and denying unauthorized actions. The exact details of this behavior can be complex when considering distributed, mobile objects with various methods available dependent on the security access of the user. This work is moving towards a formal logic for such security specification.

4.4 University of Pennsylvania

The PENN Database Research Group has undertaken the exploration of data provenance [11]. Data provenance is the path the data took to become in the form being viewed. The provenance answers questions such as where and how the information was produced, who has corrected it, and how old it is. This tracking is essential to understand the past history of an item in the database, as well as fully appreciate and rely on the current state of the item.

4.5 Stanford University

The Stanford Digital Libraries Project [15] is looking into issues of interoperability, mobile access to digital libraries, and archival.

Interoperability thrust is focused on the SDLIP (Stanford Digital Library Interoperability protocol). SDLIP describes an interface for clients to request searches be performed over multiple clients by a library service proxy. This proxy would then communicate, using protocols defined elsewhere (such as Cornell's FEDORA[12]) to contact the data sources and perform the actual query to the databases. This protocol is already past its first release and is being refined through involvement with other, non-Stanford libraries.

The mobile access to digital libraries thrust is spearheaded by the development of a power browser for personal data assistants (PDAs). This power browser integrates automatic generation of indices of web pages with the necessarily limited resources of remote PDA computing.

The archiving thrust is concerned with the development of methods for local and wide-area archiving of file systems and replication of archives. Further, this thrust is concerned with the generation of cost/benefit analysis tools, such as a simulator, to help guide the design of archives.

4.6 University of California at Berkeley

The Digital Library Project at Berkeley [16] has three research areas of interest: computer vision in digital libraries [3] and a geodetic information system (GIS) viewer [5].

The use of computer vision in creating a digital library enables the creation of large, automatically annotated data stores from digital images. Performing object recognition on images enables each image to be labeled with its contents. From here, interfaces for statistical based queries on the data are being developed. Further work is being done to be able to query a database for images relevant to a passage of text (such as one from another document).

The GIS Viewer 4.0 is designed as a tool for displaying, manipulating, editing, querying, and otherwise interacting with layered geo-spatial information. In construction, spatial information is normally presented in various layers, floors, electrical, plumbing, etc.

4.6 University of California at Santa Barbara

The Alexandria Digital Earth Prototype [2] project at UC Santa Barbara is seeking to develop a digital library environment and service based on the digital Earth metaphor. The focus of this research is to create and use Iscapes (Information Landscapes), which are a personalized digital information collection. These Iscapes are intended to have layered services which organize, access, and use different types of information. Currently, the project is focusing on creating these Iscapes for learning/educational purposes in undergraduate classes.

5. CONCLUSIONS

From the survey of digital library technologies, there is much support for the argument that a digital archive, or library, can be built for construction projects. Although research is not currently directed at this specific domain, the research of the many universities provide the key technologies to solve most of the 10 issues in developing an integrated project database brought forth by previously [Table 3].

Issues	Technologies	Universities
1. OO does not provide complete solution	Developing paradigm independent models of archives	Cornell
2. Coordinated model does not solve representation	Do not assume documents conform to a single structure	Cornell, Stanford
3. Complete model of reality required	Create reports and views based on partial information	Columbia
4. User views are reconcilable	Architectures to support multiple user views	Columbia, Cornell, Berkeley, Santa Barbara
5. Internet solves communication problem	Specific digital library/archive protocols for communication	Stanford
6. Paper documents will disappear	Multi-media archiving	CMU, Berkeley, Santa Barbara
7. CAD is the center of the database	Loosely distributed document structure	Cornell, Stanford
8. No data ownership problems	Data security, migration, and ownership model	Cornell
9. Database guarantees coordinated, consistent information	Data provenance theories to aid in maintaining data	U Penn
10. Industry is ready to adopt such a system	Not Addressed	NA

Table 3 – Technologies to Address Issues of Integrated Project Databases

Although object oriented systems are heavily promoted to be the principal architecture upon which a library is built, there is no assumption of the structure of the underlying data components. Indeed, the systems that propose object structures for the digital library interaction assume no structure for the underlying objects, which are normally items such as web pages, digital books, or image files.

While all digital library systems propose a model of the library, almost all research into searching the library and structuring the library is by viewing the underlying documents as "unhelpful." In this case, unhelpful means that the document contains little or no meta-information specifically designed to help a search engine, or to help build links between different documents. On a higher level, this means that no one model of underlying documents is assumed.

Some research is being done to address answering queries when incomplete information is available. Specifically the work at Columbia targets generation of reports based on potentially incomplete information. Also, the work at Penn on data provenance gives client applications the ability to determine the sources, and thus the potential gaps, in the knowledge available.

In almost all systems, a single user view is not assumed, and a great deal of attention is being devoted to the creation of personalized user views. This will allow disparate user views to be built from various user profiles to suit the needs of the various users.

The current architecture research being performed in digital libraries understands that the internet is not a panacea that solves all networking problems. Indeed the current architecture projects appear to be very cognizant of the need to develop reasonable protocols for interacting with the digital libraries.

Given the highly distributed nature of the digital libraries being considered, no one data item appears to be driving the structuring of the archives. Thus, there is no misconception that CAD, or any other document type will be central to the archive.

The data security work at Cornell is specifically targeting the creation of ownership and access management technologies. With these technologies, a system can be envisioned which will allow contractors, subcontractors, architects, engineers, owners, and other interested parties to be able to connect systems together with each managing their own access rights. And upon commissioning, the mobile library projects at several universities can make sure that information allowed to be mobile is automatically migrated to the owner's systems.

Penn's work on data provenance directly addresses the coordinated and consistent information issue. When the provenance work is combined with the Columbia work, which focuses on ranking various input sources for appropriateness and handles inconsistencies among sources, then there is the ability to generate a system that

can recognize and handle coordinating information to present consistent information to the users.

Two areas that are lacking in the current research are very specific to the construction industry and must be solved in the scope of developing a library specifically for construction projects: the issue of physical documents and the readiness of industry to adopt such a technology. In order to successfully manage documents and ensure they are entered into the system, a series of procedures and field protocols need to be developed. To prepare the industry to receive this research will require economic impact analyses that are outside the scope of the current research.

Overall, there is clearly a critical mass of technology being developed that can support the creation of a digital archive for construction projects. The next step should be the formalization of requirements for such an application and the development of a prototype based on the components already available and/or under development.

6. REFERENCES

[1] "Autodesk BCS: Project Collaboration and reprographic print management for building, design and management for building, design, and management professionals", http://www.buzzsaw.com/, accessed on 2002-03-30

[2] "Award #9817432 – DLI Phase2: Alexandria Digital Earth Prototype", https://www.fastlane.nsf.gov/servlet/showaward?award=9817432, accessed 2002-03-30

[3] "Computer Vision Meets Digital Libraries", http://elib.cs.berkeley.edu/vision.html, accessed 2002-03-30

[4] "Digital Libraries Initiative Phase 2", http://www.dli2.nsf.gov/, accessed 2002-03-30

[5] "GIS Viewer 4.0", http://elib.cs.berkeley.edu/gis/index.html, accessed 2002-03-30

[6] "Informedia Digital Video Library at CMU", http://www.informedia.cs.cmu.edu/, accessed 2002-03-30

[7] "Informedia-II Digital Video Library", sub-title: "CCRHE – Capturing, Coordinating and Remembering Human Experience", http://www.informedia.cs.cmu.edu/ccrhe/index.html, accessed 2002-03-30

[8] "JobDocs.com online construction documentation image management system", http://www.jobdocs.com/, accessed 2002-03-30

[9] "Meridian Project Systems - Comprehensive solutions for organizations that rely on successful projects", http://mps.com/, accessed 2002-03-30

[10] "PERSIVAL – A Patient Care Digital Library", http://www.cs.columbia.edu/diglib/PERSIVAL/, accessed 2002-03-30

[11] "Penn Database Research Group: Data Provenance", http://db.cis.upenn.edu/Research/provenance.html, accessed 2002-03-30

[12] "Prism Home Page", http://www.prism.cornell.edu/, accessed 2002-03-30

[13] "ProjectTalk.com – where the project team meets", http://www.projecttalk.com, accessed 2002-03-30

[14] "ProjectTalk.com | Project Management | Document Management", http://www.projecttalk.com/MK/pm/doc.shtml, accessed 2002-03-30

[15] "Stanford Digital Libraries Project, The", http://www-diglib.stanford.edu/, accessed 2002-03-30

[16] "UC Berkeley Digital Library Project", http://elib.cs.berkeley.edu/, accessed 2002-03-30

[17] Armor, Faraj, "Misconceptions About Integrated Project Databases", Electronic Journal of Information Technology in Construction, Volume 6, 2001, http://www.itcon.org/2001/5/paper.pdf, accessed 2002-03-30

[18] Garza, Howitt, "Wireless communication and computing at the construction jobsite", Automation in Construction, Volume 7, Number 4, pg. 327-347

[19] "Industrial automation systems and integration - Product data representation and exchange", International Organization for Standardization, Standard 10303-1:1994

[20] "IFC Object Model Release 2.0", Industry Alliance for Interoperability, CD-ROM, 1999