Proceedings of IDETC/CIE 2006
ASME 2006 International Design Engineering Technical Conferences &
Computers and Information in Engineering Conference
September 10–13, 2006, Philadelphia, USA

DETC2006-99675

CAD ARCHIVES BASED ON OAIS

Joseph B. Kopena **Joshua Shaffer** **William C. Regli**
Geometric and Intelligent Computing Laboratory
Department of Computer Science
College of Engineering
Drexel University
Philadelphia, PA 19104

ABSTRACT

Within the past few years, there has been a steady, substantial growth of interest in "long-term" archiving of digital data. This problem is particularly acute in many branches of engineering design, where cycles of technological obsolescence in supporting tools happen much more rapidly than those of designed products. Capturing and preserving design knowledge through these cycles is a major challenge that has come to be recognized by many government, industry, and research organizations. The ability to do so has important operational, efficiency, and legal ramifications for the manufacturing industry and its customers.

This paper describes this problem, presenting examples of both why it must be addressed and why it is a challenge. In particular it relates preservation of engineering data to digital archiving efforts in other domains as well as ongoing work within the engineering research community on design repositories. As is shown, long term archiving of digital design knowledge draws upon both but possesses its own unique issues. Much of this discussion is couched within the language of the ISO Open Archival Information Systems (OAIS) Reference Model, including a mapping from an existing significant design repository into the OAIS model. In this way, it is hoped that this paper will widen the discussion on digital archiving within the community of this conference as well as help connect to research in other areas.

INTRODUCTION

Responding to a US Army Air Force call for a new, long-range aircraft for strategic bombing, the Boeing Corporation began studies in 1945 for what would become the B-52 Stratofortress [1]. First fielded in 1955, the B-52 has gone on to serve in combat duty through to the present day. Further, recent studies have shown that the majority of the existing airframes will be usable until 2045. In light of this and the design's continued effectiveness and cost efficiency, current US defense plans and funding provide for its continued deployment until 2040 [2, 3]. At that point the design's lifespan from conception to decommissioning and disassembly will have spanned one hundred years.

Figure 1 charts the first half of this lifecycle against major developments in Computer Aided Design (CAD). The B-52 not only predates CAD, but has existed through every major CAD development to date. In that time it has continually undergone substantial development and redesign. For example, significant modifications were necessitated by a change in mission profiles from high-altitude cruising to structurally stressful nap of the earth flight and back again. Massive, periodic improvements in secondary systems have also warranted continuing retrofits and analysis, e.g. to power extensive new electronics or determine the effect of updated weapons pylons on range. As maintenance and upgrading of the B-52 continues for another 40 years, it stands to reason that it will outlast many more CAD milestones.

This then is the context for archiving in engineering: Supporting the lifecycle and preserving design knowledge of extremely complex artifacts through not just iterations of partic-

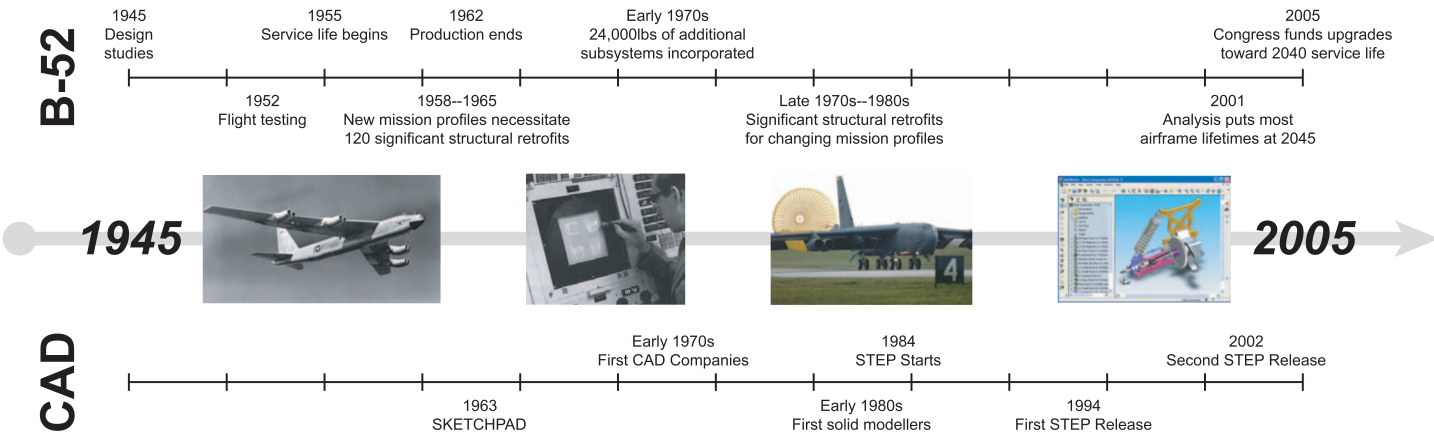

Figure 1. MAJOR POINTS IN THE HISTORY OF THE BOEING B-52 AND COMPUTER AIDED DESIGN.

ular software packages, but major revolutions in CAD itself. Although the B-52 is perhaps an outlier, it is not hard to come up with many examples from aerospace, rail, naval, or powerplant engineering of designs with lifecycles measured in the decades.

Design repositories [4] and product lifecycle management systems both aim to support design and operational activities throughout products' lifetimes. However, both are based on the implicit assumption that stored data will always be accessible, irrespective of changes in the applications themselves, supporting platforms, or even the underlying hardware. Over a long enough timespan, e.g. decades, this is simply not realistic. Without specific procedures and technology for preserving digital data through fundamental system changes, design knowledge stands little chance of surviving. To see this one merely needs to consult the literature and track the long history of research and development motivated by the difficulty of geometry exchange.

Current industrial and legal policies for addressing this problem and archiving engineered designs remain largely based on paper drawings as the gold archival standard. Paper has the advantage that its long term preservation is well understood. However, paper drawings are unsuited to preserving design data for any number of reasons, including ambiguity, difficulty of reuse, and limited scalability of management and retrieval. Some archives have begun to incorporate TIFF encoded, digital equivalents of paper drawings. Although perhaps easier to manage, among other problems they are just as unsuitable for reuse.

Recently, however, many groups have come to realize the value of and need for effective design archiving. This has taken the form of several workshops, research and pilot projects, and some legislative action, e.g. mandates for the handoff in digital form of 3D models for several substantial US government procurements. The aim of this paper then is to widen the discussion in this community on the preservation of digital design knowledge over long periods of time and through major changes in the underlying tools. This is a significant problem currently unresolved in either engineering research or practice. Only by addressing this challenge may it be hoped to:

— Archive the copious and varied range of data involved in any project, e.g. geometry, simulations, and control code.

— Ease redesign, continued analysis, disassembly, and other tasks that occur throughout a long-lived product's lifetime.

— Support beyond-lifecycle tasks such as archaeology and historical research or management of potentially hazardous "rediscovered" artifacts such as ship wrecks or waste sites.

Toward this goal, the following section briefly describes relevant work in general digital archiving and efforts specific to the engineering domain. A brief overview of the ISO Open Archival Information Systems Reference Model (OAIS) is then given. This provides the framework for the next two sections, which compare the National Design Repository [1] with the OAIS model and use this as to discuss the relationship of design repositories and product lifecycle management systems to engineering archives. The subsequent section closes with a few conclusions.

RELATED WORK

Other domains of significant digital archiving efforts include document and science data preservation. In comparison to document archiving, e.g. [5, 6], the structure of both the data itself and the relationships between the data is much more complex in the engineering domain. General document and image formats also have a much larger user base, substantially improving the chances that data formats will be interpretable far into the future. Metadata extraction and effective seach functionality is much easier to provide. Science data, e.g. astronomy observations or protein sequences, typically possesses a simpler and more ho-

[1] http://designrepository.org/

mogenous structure than engineering data, most often taking the form of voluminous databases of measurements or other records.

Another area of general digital archiving research is that of bit level, format-agnostic preservation. Although a cricital archiving component, this is not enough. Unless the syntax and semantics of those bits are also preserved, there can be little hope of interpreting the data in a meaningful fashion. This is particularly true in engineering design, which features intricately structured data with tight margins of error in interpretation.

One important archiving effort specific to engineering is the studies of the LOTAR Group, a consortium of aerospace manufacturers. Much of the group's work focuses on processes for preserving geometry [7], largely based on STEP. This is an important element, particularly as any successful engineering archive must ultimately be incorporated into business and design workflows. However, as discussed in this paper, there remains much work to be done in constructing engineering archives.

OVERVIEW OF OAIS

Much of the work in digital archiving has come to be framed by ISO Standard 14721:2003—Reference Model for an Open Archival Information System (OAIS) [8]. The OAIS reference model began as an effort to meet the need for standardizing and constructing archives of datasets created in the space sciences. However, it has rapidly been adopted throughout many other domains. It provides a common language with which to discuss archiving, in the form of an abstract breakdown of the information flows and functional components present in any archive. This section quickly overviews the OAIS model. In addition to the quite readable, publicly available reference model document itself, several other longer guides also exist, such as [9].

Most basically, OAIS defines archiving as "Long Term" preservation of information over cycles of technological obsolescence, rather than any specific time frame:

> *Long Term* is long enough to be concerned with the impacts of changing technologies, including support for new media and data formats, or with a changing user community. Long Term may extend indefinitely.

This definition matches with the challenge outlined in the introduction of preserving engineering knowledge through major changes in the underlying toolset. The problem is therefore equivalent to that of archiving as defined by OAIS.

Figure 2 presents the basic, overall OAIS model, under which there are three main abstract actors:

— *Producers*, who generate and provide information which is input to the archive for long term maintenance.

— The *Archive*, which provides interfaces and functionality for receiving information, storing it over the long term, and retrieving it for later use.

— *Consumers*, who request and receive information kept in the archive via direct addressing or queries.

The central data handling functionality of the archive is broken up into four top-level components:

— *Ingest* provides an automated or manual interface through which Producers submit content to be archived.

— *Archival Storage* implements functionality for permanent storage of the data, including tasks such as compression, media migration, error checking, and disaster recovery.

— *Data Management* implements indexing and search functionality which is used to identify, navigate, and query the information stored in the archive.

— *Access* provides an interface through which Consumers query for and retrieve archived content. The Data Management module is used to identify content to be retrieved, which is then requested of Archival Storage.

Each of these functions operate on and produce *Information Packages*, each of which bundle:

— The *Data Object*, the package's actual bitstream content.

— *Representation Information* which defines the syntax and semantics necessary for interpretation of that bitstream.

— *Preservation Description Information* used to identify, verify, and certify the package.

As shown in Figure 2, there are several generic types of Information Package transmitted throughout an archive:

— The *Submission Information Package (SIP)*, in which content to be preserved is submitted to the archive.

— The *Archive Information Package (AIP)*, which contains SIPs and complete Representation Information; it is the basic unit kept in permanent storage by the system.

— *Descriptive Information* which contains metadata, indexes, features, and other content summarizations used to identify and search over the archive contents.

— The *Dissemination Information Package (DIP)*, which bundles content requested by and sent to Consumers.

The top-level OAIS model also includes several entities related to general oversight and control of the archive. The *Preservation Planning* component monitors the archive and makes recommendations to ensure the safety of its contents, for example to migrate data or upgrade systems. *Administration* functionality conducts general operations, such as negotiating SIP formats, auditing submissions, and maintaining the archive's systems. A fourth abstract actor, *Management*, provides larger control of the archive, e.g. procuring submissions and setting access policies.

Note that OAIS is a reference model, not an architecture. Any particular architecture or implementation will contain its

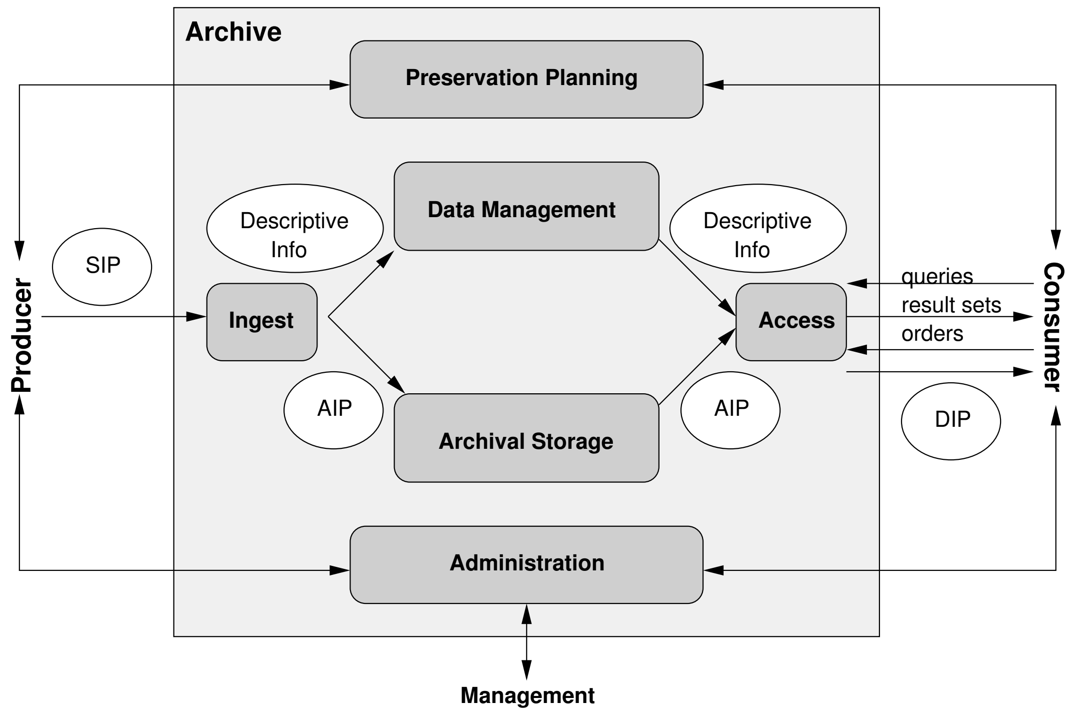

Figure 2. OVERVIEW OF FUNCTIONAL COMPONENTS AND INFORMATION FLOWS WITHIN THE OAIS MODEL.

own set of components and connections. However, if an architecture is OAIS compliant, it will be mappable back to this model. In addition, any implementation of the OAIS model will include a mix of manual and automated processes implemented by humans as well as machines. For example, Archival Storage may be essentially a database and Administration human staff.

THE NATIONAL DESIGN REPOSITORY AS AN OAIS-BASED ENGINEERING ARCHIVE

The National Design Repository [10] is a unique digital library of engineering designs from a variety of domains, currently consisting of over 55,000 CAD models and assemblies. Over time, the repository has evolved from a simple file depot to a web portal incorporating several sophisticated search and browsing techniques [11–14]. Recently the repository has been reconfigured into a Web Services architecture, exposing several internal capabilities such as file retrieval, search, and translation [15].

Many of the features and goals an archive would possess are also held by the the National Design Repository. The repository is a place for participants from research, industry, and education to post and retrieve design data from mechanical, architectural, electrical, and other domains. A primary aim is to enable widespread dissemination of challenge problems and results in CAD research. It has had a substantial influence on progress in geometry-based search, translation, and other areas by providing large, consistent datasets for experimentation.

In addition to the goal of preserving and providing access to significant artifacts, the repository shares many functional similarities with digital archives. Figure 3 presents some of these via a mapping from the top-level OAIS functional model into the repository's major software and procedural elements.

However, despite these similarities, the repository as currently implemented is not truly suitable as an archive for providing long term preservation. Particular concerns include:

— There is no systematic procedure in place for monitoring degradation in the reliability of the underlying media.

— There is no formal "technology watch" monitoring for imminent obsolescence of technology or data formats.

— There are only limited automated procedures for migrating the dataset in response to the above concerns or others.

— There is no systematic version control over (re-)submitted content, nor explicitly recorded artifact relations.

— There is no concrete, formal AIP, only conventions for structuring data in the Archival Storage filesystem.

— There is little Representation Information included in the Information Packages, at most URIs for XML schemas.

— There is little evidence that data will be interpretable in the future, and a lack of formal syntax and semantics.

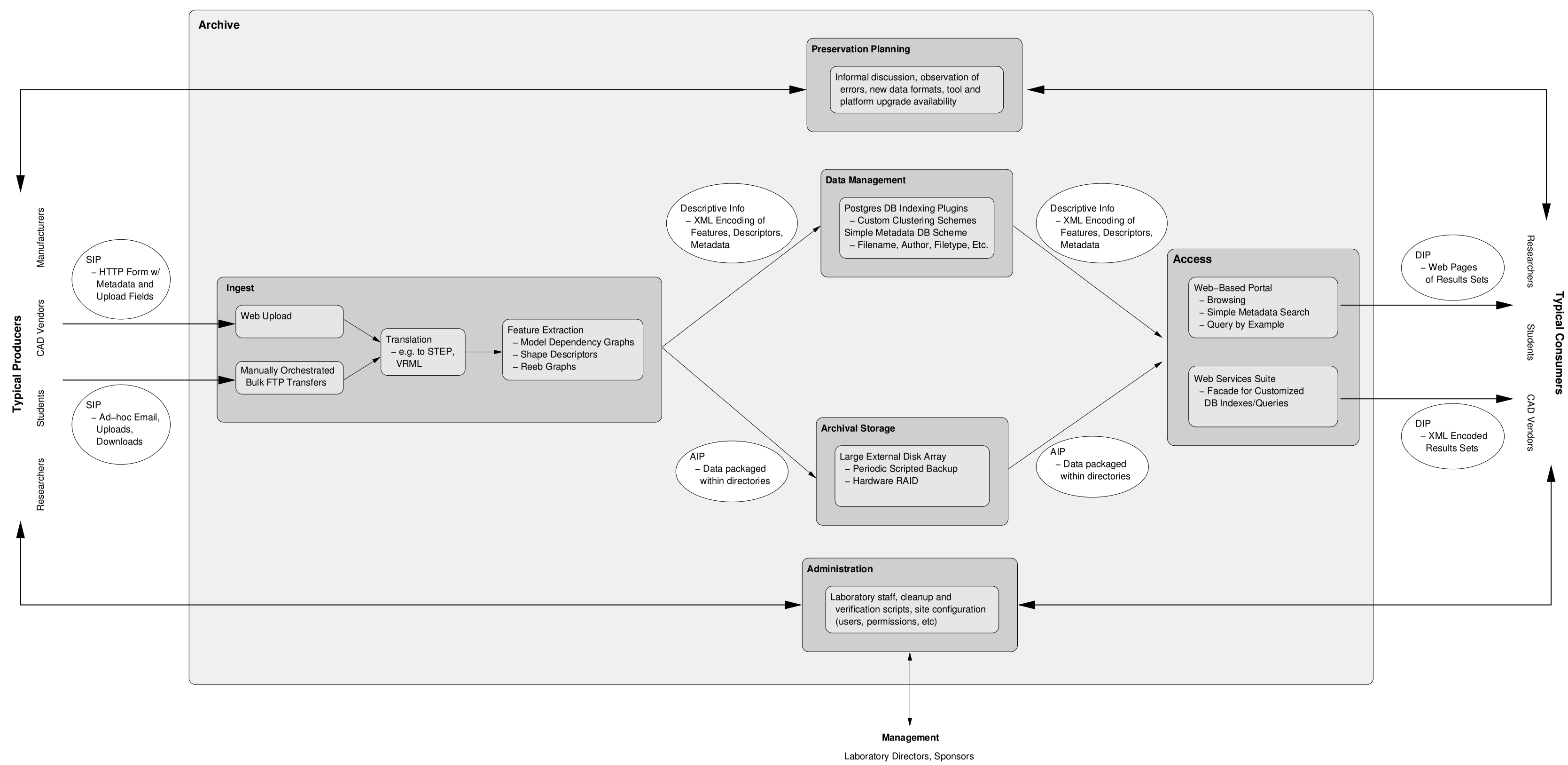

Figure 3. THE NATIONAL DESIGN REPOSITORY MAPPED INTO THE OAIS TOP-LEVEL FUNCTIONAL MODEL.

— There is little design knowledge currently captured for most stored artifacts beyond final, detailed geometry.

These problems are significant, but are true of nearly all design repository and product lifecycle management systems.

CHALLENGES IN DIGITAL ENGINEERING ARCHIVES

The mapping of the current National Design Repository into the OAIS functional model in Figure 3 makes several points. Most immediate is that there is a strong similarity between design databases, design repositories, product lifecycle management, and digital engineering archives. However, they are not identical. Long term preservation of design data introduces many new elements. Some of these are procedural concerns to address in practice, such as monitoring for media degradation and technological obsolescence. Others require application and continued development of current research, such as search techniques for CAD data. A few, however, are research and development topics unique or particularly stressed by the archiving task.

As an opener, renewed focus must be placed on capturing all aspects of a design project. Geometry, analysis meshes, simulation code, requirements, functional models, etc. must all be archived as any data successfully archived may rapidly become meaningless outside its original context. Although a prominent goal of design repositories since their conception, results remain limited. Long term archiving also increases the pool of data which must be captured. For example, in addition to any generated data, the tools used to create that data, and any operating context or parameters, must also be preserved. Archiving of software raises its own concerns at least as complex as those of CAD, but is an important component of design context.

One consideration unique to archiving is that in the absence of perfect solutions to capture all aspects of design, it may be better to "simply grab everything." This may require substantial "digital archaeology" efforts to utilize such unstructured, unspecified data. However, it is certainly better than preserving no data and such efforts may be worth it in the future for significant artifacts. This also addresses the consideration that it is hard to predict precisely how data will be used over the long term.

Ideally, however, that wide range of data would be captured and packaged in some manageable form. To do this, schemas must be developed for Information Packages to encapsulate the wide variety of data generated in the CAD domain. Well-defined, formal package schemas would also promote a distributed, open, services-based approach to developing archive functionality. Information Package schemas are therefore not just a concern of particular projects, but a topic for possible standardization.

Importantly, these schemas need to incorporate Representation Information such that every discrete piece of transmitted and archived information is assured to be interpretable in the future. Standards such as STEP will clearly play a large role in this. However, there are many types of engineering data for which no standards exist. Long term preservation only heightens the need for complete interoperability, as there will be no original source to question or examine when questions of interpretation arise. Suitable Representation Information must therefore leave no data open to interpretation. In addition to the formality required to do so, scalable, automated conduct of archive tasks such as collection migration may require machine interpretable semantics.

The above are significant issues which must be addressed in developing digital engineering archives. However, they are also not without precedent in related areas.

CONCLUSION

This paper has aimed to widen the discussion of engineering archives within the design research community. As shown by the example of the National Design Repository, this is not a wholly new problem. Instead it draws upon design repository, lifecycle management, and other research, but introduces several new dimensions and nuances to the problem. By presenting the context and need for digital archiving, as well as challenges and approaches to doing so, it is hoped this work will motivate further research and development in this area.

ACKNOWLEDGMENT

This work was supported in part by the Library of Congress National Digital Information Infrastructure Preservation Program (NDIIPP) via National Science Foundation (NSF) DIGARCH Program grant CISE/IIS-0456001. Additional support was provided by the National Science Foundation Cyber-Infrastructure Teams (CI-TEAMs) Program grants CISE/SCI-0537370 and CISE/SCI-0537125.

Opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the other supporting government and corporate organizations.

REFERENCES

[1] US Air Force, 2005. Factsheets: B-52 stratofortress. http://www.af.mil/factsheets/factsheet.asp?fsID=83, October. Accessed May 2006.

[2] US Air Force, 2001. Long-range strike aircraft white paper. http://www.af.mil/library/posture/bmap01.pdf, November. Accessed May 2006.

[3] Defense Science Board, 2004. B-52H re-engining. http://www.acq.osd.mil/dsb/reports/2004-06-b52h_re-engining.pdf, June. Accessed May 2006.

[4] Szykman, S., Bochenek, C., Racz, J. W., Senfaute, J., and Sriram, R. D., 2000. "Design repositories: Engineering design's new knowledge base". *IEEE Intelligent Systems*, **15**(3), May/June, pp. 48–55.

[5] van der Werf-Davelaar, T., 1999. "Long-term preservation of electronic publications: The NEDLIB project". *D-Lib Magazine*, **5**(9). Available at http://webdoc.sub.gwdg.de/edoc/aw/d-lib/dlib/september99/vanderwerf/09v%anderwerf.html.

[6] Tansley, R., Bass, M., and Smith, M., 2003. "The integration of product data and workflow management systems in a large scale engineering database application". In European Conference on Research and Advanced Technology for Digital Libraries, Springer Verlag, pp. 446–460. LNCS 2769.

[7] Project LOTAR, 2002. Long term archiving and retrieval of product data within the areospace industry (LOTAR): Technical aspects of an approach for application. Tech. rep., ProSTEP iViP Association, August.

[8] Consultative Committee for Space Data Systems, 2002. Reference model for an open archival information system (OAIS). Tech. rep., January. Adopned as ISO Standard 14721:2003.

[9] Bull, A., 2006. Briefing paper: The OAIS reference model. Tech. rep., UKOLN, University of Bath, February.

[10] Bespalov, D., Ip, C. Y., Regli, W. C., and Shaffer, J., 2005. "Benchmarking search techniques for CAD". In ACM Symposium on Solid and Physical Modeling, Association for Computing Machinery, ACM Press, pp. 275–286.

[11] Cicirello, V., and Regli, W. C., 2001. "Machining feature-based comparisons of mechanical parts". In International Conference on Shape Modeling and Applications, ACM SIGGRAPH, The Computer Graphics Society, and EUROGRAPHICS, IEEE Computer Society Press, pp. 176–187.

[12] McWherter, D., Peabody, M., Shokoufandeh, A., and Regli, W. C., 2001. "Transformation invariant similarity assessment of solid models". In ASME Design Engineering Technical Conference, American Society of Mechanical Engineers, ASME Press. DETC2001/DFM-21191.

[13] Ip, C. Y., Lapadat, D., Sieger, L., and Regli, W. C., 2002. "Using shape distributions to compare solid models". In Seventh ACM/SIGGRAPH Symposium on Solid Modeling and Applications, Association for Computing Machinery, ACM Press, pp. 273–280.

[14] Bespalov, D., Regli, W. C., and Shokoufandeh, A., 2003. "Reeb graph based shape retrieval for CAD". In ASME Design Engineering Technical Conferences, 23rd Computers and Information in Engineering Conference, American Society of Mechanical Engineers, ASME Press. CIE-2003-48194.

[15] Shaffer, J., Kopena, J. B., and Regli, W. C., 2005. "Web-service interfaces for design repositories". In ASME International Design Engineering Technical Conferences, 17th Computers and Information in Engineering Conference, American Society of Mechanical Engineers, ASME Press. DETC2005-85386.