NISTIR 7386

Long Term Knowledge Retention Workshop Summary

Joshua Lubell
Sudarsan Rachuri
Eswaran Subrahmanian
William Regli

National Institute of Standards and Technology
Technology Administration, U.S. Department of Commerce

NISTIR 7386

Long Term Knowledge Retention Workshop Summary

Joshua Lubell, Sudarsan Rachuri, Eswaran Subrahmanian
Manufacturing Systems Integration Division
Manufacturing Engineering Laboratory
National Institute of Standards and Technology
Gaithersburg, MD 20899-8263
{lubell,sudarsan,eswaran}@nist.gov

William Regli
Department of Computer Science
Drexel University
Philadelphia, PA 19104
regli@drexel.edu

December 2006

U.S. Department of Commerce
Carlos M. Gutierrez, Secretary

Technology Administration
Robert Cresanti, Under Secretary of Commerce for Technology

National Institute of Standards and Technology
William Jeffrey, Director

Abstract

This report summarizes the presentations, discussions and recommendations of a workshop held at the National Institute of Standards and Technology on March 15-16, 2006. The purpose of the workshop was to identify challenges, research, and implementation issues in digital preservation of information with an emphasis on design and manufacturing. An appendix provides the original call for participation, workshop agenda, and guidelines for breakout sessions, as well as a list of the participants.

1 Introduction

This report summarizes the presentations, discussions, and recommendations of a Long Term Knowledge Retention (LTKR) workshop held at the National Institute of Standards and Technology (NIST) on March 15-16, 2006. The purpose of the workshop was to identify challenges, research, and implementation issues in digital preservation of information, with an emphasis on design and manufacturing. This goal was realized by a diverse group of 35 participants representing industry, government, and academia and bringing together researchers from disciplines including manufacturing engineering, library sciences, knowledge representation, and space science.

Archiving of engineering and manufacturing information has been practiced for a long time in the paper based world. Technologies such as microfiche, while preserving the fidelity of paper, also provided a means to reduce the space required for storing information. Manufacturing organizations created departments to maintain company archives adapting to the technologies available. The advent of computing brought about new means for creating and storing the information, and generated new demands for archiving. For example, when the Incline cable cars in Pittsburgh were refurbished by the then-Westinghouse Corporation in the 1990s, an old TRS80 computer was needed to run an analysis program used a decade earlier to perform some repairs. This example shows that the need for archiving in the digital world is multifaceted. Not only is archiving data (format and content) required, but sometimes it is also necessary to archive the computer (machine) and computer program (software) used to create the data. The ubiquity of computing and digital records requires creation of new and innovative ways to archive information that are not a direct adaptation of the paper based world.

There are many stories which, like the Pittsburgh cable car example, illustrate the consequences of not adopting a proactive approach to digital preservation of engineering design and manufacturing information. These stories are often anecdotal and not widely publicized, perhaps because of fear of embarrassment and possible liability. As a result, a recurring discussion topic throughout the workshop was how best to develop a strong business case for long term archiving.

Section 2 summarizes each keynote and panel presentation. Although there were two panels – one panel emphasizing manufacturing engineering and the other focusing on more generic LTKR concerns, the actual talks more naturally fell into three categories. Therefore we group the summaries into (1) *Broad Perspectives,* (2) *Representation and Quality of Engineering Information: Product Data Representation,* and (3) *Information Standards and Archiving.*

Section 3 relates the issues discussed in the breakout sessions.

Section 4 discusses conclusions and takeaways, and spells out ideas for future work.

Sections 2, 3, and 4 report on what the workshop participants presented and discussed. No endorsement of the ideas expressed in these sections is implied by the authors or by NIST.

We are preparing a follow-up document, *Long Term Knowledge Retention for Engineering Enterprises*, which will provide our own ideas regarding challenges, research, and implementation issues in digital preservation of information with an emphasis on design and manufacturing.

2 Summaries of Workshop Presentations

The workshop agenda included two keynote presentations, two panel discussions, and a breakout session where the participants split into two groups. One breakout group concentrated on manufacturing informatics; the other focused on archiving standards, languages, and representations. The workshop concluded with a group discussion of the breakout results. An appendix contains the call for participation and the agenda distributed to the workshop participants.

2.1 Broad Perspectives

This group of presentations, which includes the two keynotes, discussed general issues in digital archiving and preservation, irrespective of any particular standards or industries.

2.1.1 Research Perspectives on Digital Archiving and Call to Action (keynote)

Dr. Robert Chadduck, US National Archives

The Networking and Information Technology Research and Development (NITRD) Supplement to the President's Budget for FY2007 [1] highlights "maintenance of and access to long-lived science and engineering data collections and Federal records" as a research priority. NIST should take a leadership role in this work. The solution to this problem only partially exists, and different communities working on digital preservation technologies need to coordinate their efforts.

Shipbuilding, aerospace, and civil engineering projects all need to track modifications of original designs over time. This is important both for maintenance and for contractual responsibilities. ISO 10303 (informally known as STEP, the Standard for the Exchange of Product Model Data) [2] should be a centerpiece of coordinated research in this area.

Digital data used, received, or created in the course of federal activities also may need to be preserved. Examples of these activities include major federal acquisitions, purchase and maintenance of airframes or weapons systems, and creation of regulatory data.

Preservation must take place in a user-centered context. The beneficiary is the user, not the repository builder. The objective is to archive the data in the context that can best ensure its future usability. The future usage scenario might be something unanticipated by the repository builder or archivist. Also, there is no "silver bullet" single solution. The solution must evolve as the problem evolves and must fit the requirements.

2.1.2 Principles for Digital Preservation (keynote)

Dr. Henry Gladney, HMG Consulting

Digital preservation is the mitigation of the deleterious effects of technology obsolescence, media degradation, and fading human memory. Although the need for digital preservation is widespread, most of the research to date on the topic has come from librarians and archivists. The computer science community has done little so far, and little software has been implemented to aid in digital preservation. The National Archives and Records Administration and Library of Congress are unusual cases with narrow areas of focus. They are huge and relatively well funded. Their problems are driven by the amounts of data they must handle.

LOCKSS (Lots of Copies Keeps Stuff Safe) [3], a system originally developed by Stanford University, is a good idea, but needs to be generalized.

No set of information is truly bounded; no nontrivial bounded set exists. Tools for creating metadata are clunky or nonexistent. To design a good metadata creation tool, one should ask, "What questions do consumers of archived data want to answer?"

The hardest technical problem is the development of durable digital encodings. The key question is how to handle proprietary formats and formats that become defunct over time. Standards and open source software are preferable. Any particular document type may have a minimal set of attributes worth saving. If one can render information in multiple formats to reduce ambiguity, then format migration is less of a problem.

An approach to the durable encoding of digital data is to use software emulation. First develop a virtual machine and an interpreter application for each computer platform. Then develop a data interpreter for each file format. This approach differs from many other preservation methods in that it focuses on the documents to be preserved rather than on repositories, archival, and access methods.

2.1.3 Digital Formats Factors for Sustainability, Functionality, and Quality

Caroline R. Arms, Library of Congress, Office of Strategic Initiatives

There are two types of evaluation factors for digital formats:

1. Sustainability factors for all formats. These influence feasibility and cost of preserving content in the face of future change.
2. Quality and functionality factors that vary by content category. These reflect considerations that will be expected by future users.

Selection of acceptable formats should also take into account the digital content's origin and circumstances concerning its creation. Two examples are as follows:

1. For copyright registration, Joint Photographic Experts Group (JPEG) images [4] must be acceptable, because many digital cameras do not produce uncompressed images.
2. For some content of high cultural value, such as the working files of a composer of electronic music, particular functionality may outweigh sustainability factors.

The Library of Congress (LC) plans to exploit synergy between projects building automated systems for managing information about formats and projects developing tools that can validate, characterize, and transform content in those formats. Some of these projects are funded through LC's National Digital Information Infrastructure and Preservation Program (NDIIPP) [5]. It is suggested that somebody (possibly NIST) maintain a format registry for engineering informatics along the lines of, or in conjunction with, Harvard University's Global Digital Format Registry (http://hul.harvard.edu/gdfr/) and LC's digital formats repository (http://www.digitalpreservation.gov/formats).

2.2 Representation and Quality of Engineering Information: Product Data Representation

These presentations all examine issues strongly related to manufacturing and engineering. Although *The Role of ISO 10303 (STEP) in Long Term Data Retention* focuses specifically on a standard, because that standard is for representing product data, we include this presentation here rather than in Section 2.3.

2.2.1 Archiving Manufacturing Knowledge

Dr. Frank Brown, University of Kansas

A project titled "Knowledge-based Archiving of Manufacturing Part Shape" and conducted by the National Nuclear Security Administration (NNSA) [6] and the National Archives and Records Administration (NARA) [7] started in 2003. The goal of this project is to archive a manufacturing part shape with authenticating features in the NARA Electronic Records Archives (ERA) [8] prototype, and then to retrieve and authenticate it.

The approach taken in this project is to start with ISO 10303 STEP AP-203 [9], transform it into a knowledge-based form, and then deduce authenticating aspects of part shape (features) by reasoning over the part geometry and topology. The part shape is represented in the World Wide Web Consortium's Web Ontology Language (OWL) [10] format for archiving. The project demonstrated how to archive the OWL part, with its authenticating features, in the ERA research prototype, retrieve the OWL part from the ERA, and authenticate it by again deducing its shape features and comparing them to the shape features previously archived.

ISO 10303 AP-203 representations of vertices, edges, loops, faces, points, curves, and surfaces were used to represent parts. The representation of knowledge about parts was then defined using axioms describing different kinds of part features, for example, bosses, open pockets, cutouts, through holes, and face chains. Logistica [11], an automatic deduction system for inference, was used. Logistica is a meta-programming system that allows efficient deduction and translation processes to be quickly specified and refined.

The Kansas City Plant (KCP), an NNSA facility managed and operated by Honeywell Federal Manufacturing & Technologies, was able to represent the action semantics in a LISP-like neutral form (not a standard format). The action semantics were then dynamically read into the Logistica reasoner. The reasoner then applied the action semantics to the geometry and topology of the part to deduce the features of form. Additional rules can be written outside of the reasoner, and they can then be applied

without changing any reasoner code. The KCP neutral format could be replaced by a standardized format in the future.

2.2.2 CAD Model Verification, Validation, and Comparison

Doug Cheney, ITI Transcendata

When analyzing a model, how does one identify "bad CAD" (computer aided design)? And how does one identify which digital artifact represents the master model? Additional information regarding model quality, or fidelity, is needed in any archive. CAD data quality can be tracked during the lifecycle. Quality can be assessed during model ingest into an archive.

"Bad CAD" includes the following:

- Invalid geometry
- Unrealistic features
- Unacceptable changes caused by translation, migration, or re-mastering
- Undocumented changes resulting from design revisions or engineering change orders
- Unintentional changes resulting from revisions or change orders, or caused by parametric relationships

There are two forms of underlying CAD model representation within CAD software systems: procedural definitions that store the recipe on how to create the model, and explicit representations that store only the underlying mathematical forms. Feature-based approaches would be considered as procedural definitions. STEP is geometry based (i.e., most of the existing standard would be considered explicit); features, such as blends, are examples of procedural definitions. Testing with current CAD systems has shown that the risk of significant geometric change is much higher when exchanging a parametric representation of a model. Therefore an explicit STEP representation is geometrically more stable for long-term preservation than a native CAD model.

Issues arise as to the development of best practices for capture and preservation of CAD artifacts. Are metadata different across different CAD domains? Also, representational transformations may result in losing information or in the introduction of errors. Hence, why not save all models? If one chooses to save all models, than one needs a way of addressing product quality data as it moves through the lifecycle. At the point of archiving, this may be the best time to map quality issues across all objects.

A new STEP resource, *ISO/CD 10303-59 Quality of Product Shape Data* [12], is being developed to standardize the representation of geometry quality measurements. (It was released for Committee Draft ballot in September 2006). Also, an extension to the CAx (computer-aided design, engineering, and manufacturing) Implementers Forum *Recommended Practices for Geometric Validation Properties* [13] was published in March 2006. It defines a methodology for adding mass properties and geometric datum points to a STEP model. These can be used, respectively, to improve the quality of a STEP product model archive by documenting its geometric quality and to enable validation of its geometric shape after import into a future CAD system.

2.2.3 Defense Archiving Issues and Initiatives

James L. Mays, Naval Surface Warfare Center, Carderock Division
Design has evolved from a drafting-based process to a three-dimensional (3D) product-oriented information database used for computer-aided design, analysis, and production. One of the most significant enhancements has been the incorporation or integration of non-graphic attribute information with traditional graphics data.

This expanded database has enabled the maritime industry to share CAD data with engineering analysis, production planning/control, and logistics support tools. These systems allow the industry to reduce errors associated with manual regeneration of data from paper, resulting in an improved product. The use of portable, nonproprietary standard technical data has tremendous potential to facilitate electronic commerce as ship construction, subcontractor, and marine vendor integration proceeds.

In attempting to access a legacy 2D drawing, the Navy encountered several problems including:

- Inability to query scanned, 2D images stored as raster files
- Missing information. Data sometimes cannot be read, or data is lost in interoperability exchange.
- Difficulty interpreting 3D parts from 2D drawings
- Intellectual property rights (IPR) issues. Some parts cannot be stored and used in an open procurement process because they may be designated as proprietary by the vendor.
- Data quality errors, for example if information is entered incorrectly

Product data standards such as ISO 10303 and Department of Defense (DoD) integration standards solve some problems, but they do not address data quality or IPR issues, nor do they address management of libraries/catalogs of parts.

And product data standards create new challenges:

- Storage of redundant data (3D product model data, 2D drawings, etc.)
- Bad CAD data created during conversion
- Standards lagging behind CAD tools in ability to represent industry's data exchange requirements
- Need for translators and viewers

2.2.4 The Role of ISO 10303 (STEP) in Long Term Data Retention

Dr. Burton Gischner, Electric Boat Corporation
Archiving engineering information is a challenge. Long Term Data Retention (LTDR) is a critical problem that needs immediate attention. There are several reasons for this:

- Products in many industries (e.g., aerospace, automobiles, and shipbuilding) have life spans that far exceed the span of the CAD, CAE (computer-aided engineering), or PDM (product data management) systems that create the product models.
- Accurate product model data is needed throughout the lifecycle of the product for repairs, overhauls, in-service modifications, etc.
- Data must be retained for the life of the product, which can be up to 50 years.

- On the other hand, the life of a CAD, CAE, or PDM system used to design the product is often less than 10 years.

LTDR requirements specific to Electric Boat (EB) and other shipbuilders are as follows:

- Products (i.e., submarines) generally remain in service for at least 30 years.
- Data specifying the product must be maintained throughout its lifecycle, which could last up to 50 years.
- CAD/CAE/PDM systems used to generate the product models will not last that long.
- The U.S. Navy requires that ship manufacturers maintain hard copy design drawings for the life of the ship because there is no guarantee the drawings could be reproduced from a digital product model data in the future.

EB is hoping to preserve digital product models rather than two-dimensional (2D) drawings for long term retention. Design is captured as a three-dimensional (3D) product model. The revisions, enhancements, overhauls, and repairs all require modification to be made to the 3D product model. Thus, it only makes sense for EB to maintain the design as a 3D product model, rather than as a set of hardcopy drawings. However, this can only be done if their customer is convinced that the part can be accurately reproduced from the captured 3D product model.

To archive the 3D product model for long-term retention EB is exploring STEP as a possible solution, but has run into some shortcomings. STEP was primarily developed as a tool to capture the finished 3D product model as saved during design. Although extensive efforts at information modeling have gone into development of every STEP specification, the focus of these efforts was on representing the completed model, not capturing the process and reasoning that went into creating that model.

Enhancements to STEP that would improve support for LTDR include the following:

- Representation of design intent, construction history, geometric dimensioning and tolerancing, and analysis results
- An exchange format encoded in the Extensible Markup Language (XML) [14] rather than the non-XML text encoding currently used [15]. Because XML is so widely used and enjoys strong software support, XML-encoded STEP data will be of more value than non-XML STEP data should software vendors stop supporting STEP.
- Assurances that data conforming to the current standard will also conform to future versions of the standard
- Commitments from CAD and PDM vendors to implement future versions of STEP

2.2.5 Archiving Engineering Case Files for Future Reference

Crispin Hales, PhD, CEng, Hales & Gooch Ltd.

Information in archived engineering informatics files can yield useful forensic analysis results. The following table shows which information types from an archive could be helpful in answering particular questions regarding an accident or incident:

Question	Information
What happened?	Reports, statements, photographs, depositions
Why did it happen?	Review, analysis, interviews, timelines, logic
Who was responsible?	Contracts, documents, communications
Who should have done what?	Codes, standards, procedures

The outputs of a forensic analysis include both technical and legal documentation. Examples of technical outputs are opinions, interrogatories, depositions, and reports. Examples of legal documents are mediation or arbitration results, settlements, and (if the case cannot be settled out of court) trial transcripts.

In most of the forensic analyses of engineering files, the investigations are carried out at great cost, and the engineering history is compiled in detail. The evidence is carefully gathered in quantity, and the technical issues are researched carefully. Also the case claims are debated and resolved. After this carefully-executed analysis, in most cases the files are then discarded and forgotten, and unfortunately most of the materials are never used again.

It would clearly be of great public benefit to archive materials from case files. However, there are some difficulties in doing so, such as:

- Determining what is important
- Getting approval for using it
- Determining what format to use
- Integrating different types of material
- Establishing a collective repository
- Creating awareness of what exists
- Dealing with the residual materials

2.3 Information Standards and Archiving

These presentations discuss standards specific to digital archiving, but not specific to a particular industry or application.

2.3.1 OAIS Reference Model Standard: Motivation, Applicability, Follow-on Efforts

Don Sawyer, National Aeronautics and Space Administration (NASA)/ National Space Science Data Center (NSSDC) http://nssdc.gsfc.nasa.gov/

The OAIS (Open Archival Information System) [16] reference model is a standard developed by the Consultative Committee for Space Data Systems (CCSDS) and later ratified by ISO as ISO 14721 [16]. OAIS is applicable to all long-term archives, not just space science applications. OAIS does not specify an implementation; rather it defines a common vocabulary for describing archiving architectures.

Information objects, the key building blocks of OAIS, consist of both *data* expressing the information and *representation information* specifying the data's interpretation. Archives *ingest* and provide *access* to information packages. An information package includes an information object representing its content, as well as *preservation description information* (PDI). The PDI describes how the content originated, the content's chain of custody, its relationships to other information, and how the content can be authenticated. These PDI components in turn each consist of information objects (containing data and representation information).

OAIS has been widely adopted by digital librarians, archivists, scientific data centers, and industries such as aerospace. The OAIS reference model has spawned follow-on efforts to standardize interfaces between producers and archives and to standardize digital repository certification.

2.3.2 XML information Packaging Standards for Archives

Lou Reich, CSC (Computer Science Corporation)
There is a growing interest in XML-based representation of information objects in digital library architectures. These XML formats include ISO/IEC 21000-2MPEG-21 Digital Item Declaration (DID) [17] and Digital Item Declaration Language (DIDL), Metadata Encoding and Transmission Standard (METS) [18], Instructional Management Systems/Content Packaging (IMS/CP) [19], and XFDU (XML Formatted Data Unit) [20]. XFDU is being developed by CCSDS [21]. Benefits of these information objects include:

- Platform-independence
- Industry support
- Longevity and potential migration paths
- Availability of processing tools and validation capabilities

The aim of the METS format is to create a single document format for encoding digital library objects which can fulfill roles of Submission Information Package (SIP), Archival Information Package (AIP), and Dissemination Information Package (DIP) within the OAIS reference model. The initial scope is limited to objects comprised of text, image, audio, and video files. METS intends to promote interoperability of descriptive, administrative, and technical metadata while supporting flexibility in local practice.

Authors' note: METS and XFDU are similar in that they are both XML-based packaging standards intended for use in OAIS information packages. However, XFDU's tag set more closely mimics the structure of the OAIS information model [22].

3 Breakout Sessions

During the second day of the workshop, participants separated into two groups. The "Manufacturing Informatics" group was led by Dr. SK Gupta of the University of Maryland, College Park. The "Archiving Standards, Languages, and Representations" group was led by Dr. William Regli of Drexel University. Each group was assigned a list of discussion topics, detailed in the Appendix. The discussions, as reported by the group leaders, follow.

3.1 Manufacturing Informatics

Manufacturing informatics is sufficiently diverse that no single archiving technical architecture is likely to meet all requirements for all engineering-related applications. However, the breakout group determined that the OAIS reference model seems widely applicable to this domain, even though OAIS was not designed to be engineering or manufacturing-specific.

The breakout group assessed the current state of archival processes and methods and, given the current state of affairs, identified the type of artifacts (intermediate and final) produced during the product lifecycle that need to be archived, gaps in the archival processes, uses of archival information, and the cost of inadequate archival policies and processes. The following paragraphs address these issues.

3.1.1 Current practice

The current state of digital archiving practice suffers from the following:

- Difficulty adapting best practices from the paper-based world
- Failure to keep up with rapid changes in the technologies of creating, codifying (syntax and semantics for representing structure and behavior), exchanging (interactions and sharing), processing (decision making), storing (archiving), and retrieving (accessing) the digital objects that characterize the cross-disciplinary domains of engineering discourse

Companies are disinclined to devote resources to reforming the archiving process because it is expensive and the benefits are long-term rather than short-term. Adding to the confusion, Product Lifecycle Management (PLM) software vendors claim that backup of data in their proprietary format is all that is needed. The lack of a general strategy for archiving has led to a proliferation of ad-hoc approaches, with incompatible local practices hampering interoperability across organizations and causing fragmentation within organizations. For example, use of localized formats for documents and data within a division – without any explicitly decided indexing method for the information – renders the information inaccessible to others.

Other common practices include in-house delivery of data using viewers for product geometry modeling, archiving of scanned printed files or drawings, and restricting access to original data. Although these practices are not necessarily bad when considered individually, taken together they can result in fragmentation within an organization and lead to problems with searching and tracking of information. Improving interoperability and integration is critical to better archiving processes, however there will exist institutional resistance to change. Improving the archiving process is likely to require a combination of policy requirements, support systems and possibly even incentives.

3.1.2 Cost – To archive or not to archive

One important aspect of changing the culture of archiving is to make a clear cut business case with respect to both the cost of archiving and that of not archiving. There are several reasons for the cost of archiving to be perceived as too high given the pressures to shorten the product cycles. However, the cost of not archiving often leads to second order costs related to re-creating information, cost of repetitions of errors and retesting, cost of legal

challenges to the product viability, cost of training and education of personnel in the product design and manufacturing processes, and decisions and cost of re-engineering a product at a later date.

In the manufacturing and design world the following real-world scenarios exemplify the need for a robust archiving process from a cost point of view:

- Fixing aircraft stranded in remote places. This requires well archived data, given the life span of aircrafts. E.g., Boeing's "airplane on the ground" team had to rebuild the bottom half of an Air France 747 that belly-landed in Delhi on site. The cost of transporting aircraft to be fixed elsewhere is often prohibitive
- Reconstruction of the events leading to an Airbus A330 accident in Long Island in 2001. Accident reconstruction is expensive, but very critical for liability and legal purposes.
- Emergency spares. The Department of Defense, which consumes products with long operational lifetimes, creates a demand for spares for these products in times of emergency. Sometimes emergencies require onsite reconstruction of the product. Military planners need to consider maintenance and replacement costs for products of different and overlapping life times.

The problem of performing a realistic cost-benefit analysis will require collection of data to justify the cost of not archiving. The cost of archiving, on the other hand, will require estimating the cost of changes in technology for storage, format, computing machinery and the ease with which the archival process can be supported and made useful. This is not an easy task. These costs must be weighed against the cost of not archiving in order to make an informed decision on the level of archival needs to be supported.

3.1.3 Requirements

The fundamental question of what should be archived is important in the digital world. Gathering all information without systematic indexing and organization will not lead to good archives, just as collecting and boxing all the paper documents in a warehouse does not lead to useful archives either. In the digital world, the ability to archive more than what we do in the paper world creates new possibilities, but it also creates new problems such as information overload as well as information loss due to changes in technology. With these observations in mind, the breakout group identified the following as important categories of digital documents to be stored:

- Design rationale
- Minutes from meetings and design reviews
- CAD Models
- Engineering drawings
- Test data used for validation
    - Photographs/Images
    - Loading diagrams
- Manufacturing process plans
- Assembly plans
- Inspection, maintenance, and service
- Documents delivered to customers

- Manufacturing logs
- Product operation
- Failure results
- Materials used to manufacture the product
- Engineering changes
- Relevant algorithms and assumptions
- Relevant software
- Design/production logs
- Physical artifacts
- Certifications/Authorizations of materials and processes

While the list is a long one, the justification for storing each of these can be made by looking at the tasks that need to be supported during the lifecycle of a product and its variants. The initial categories of tasks where the above information may be used are as below:

a) Legal
    a. accident investigation, failure analysis
    b. customer delivery requirements
    c. Merger and acquisition
    d. Patent infringements
b) Operational and support
    a. Historical data to provide lifecycle support (maintenance, spares, recycling and disposal)
c) Product development management
    a. Effectivity; tracing design rationale in cases of failure, etc.
    b. Design re-use (important for parts used in multiple products or models)
    c. Engineering change proposals/analysis
    d. Reverse engineering
d) Comparison with new work, test beds, validation suites

To support these tasks, several issues need to be considered. These include how to store the information to be archived, which standards to use for addressing the archived documents, and which standards to use for indexing, archiving, maintaining, and tracing the provenance of the archived information.

3.2 Archiving Standards, Languages, and Representations

From the standpoint of LTKR, archiving requirements for engineering are unique in that (1) data formats tend to be more complex, and (2) representation of discrete processes – both manufacturing and business processes – is particularly important. Keeping these differences in mind, one can apply generic archiving methods and technologies to engineering applications, adapting them as needed.

There are five major ingredients to successful archiving. The first is keeping the people in an organization committed and engaged. This requires sustaining a focus on data preservation and quality. The commitment to archiving must always be present.

Next is a thorough analysis of use cases and user communities likely to access the archive in the future. Archiving policies (e.g., determining what to save and what to discard) must take these requirements into account.

The third ingredient is metadata, i.e., what the OAIS reference model calls *representation information* and *preservation description information* (see Section 2.3.1). Key questions are "How much metadata is enough?" and "How can the metadata be captured most easily?" Ideally, metadata should be obtainable without an undue burden on the part of data providers. The Google™ search engine is a good example of painless metadata capture. Metadata is obtained through usage patterns rather than relying on web page authors to add metadata tags to their pages.

Fourth is a strong business case. This can be a challenge because the likely beneficiaries of good archiving practices are often not the same people as the ones responsible for creating the archive. Saving everything with minimal metadata, and leaving it up to those accessing the archive in the future to determine what is relevant and what is not, can reduce archiving costs. This might actually be a reasonable strategy if archive creators are weak on the second ingredient. Economic analyses of the cost and benefits of archiving for different application domains would be beneficial for determining a particular business case.

Finally, effective archiving is more likely to occur in a climate where potential users expect digital data to be preserved. This is already happening to some extent as mass-market consumer technologies such as digital photography create a widespread societal need for preserving large amounts of data. Articles published in the popular press or in non-information technology venues such as law review journals can also raise expectations and awareness of the need for digital preservation.

Some other observations:

- Libraries and governments are developing a number of digital format registries. It would be useful to the engineering community if CAD/CAE formats were added to these registries.
- Currently available technologies can implement some useful capabilities.
- The "going forward" problem is different from the "looking backward" problem. In other words, archiving data available today and developing an archiving system for data yet to be created require different solutions.

4 Observations and Conclusions

The workshop presentations and discussions exhibited two key characteristics: (1) a view of LTKR as an archiving process, and (2) an emphasis on business case development.

Looking at LTKR as an archiving process, the problem becomes one of applying an archiving model such as OAIS to a particular collection of digital artifacts. Although most workshop participants adopted this viewpoint, Henry Gladney's keynote (Section 2.1.2) offered an alternative "document-centric" view of LTKR. In Gladney's view, ensuring a document's preservation and authenticity are most important. The archival process and data representation method are secondary (provided that you have software that can interpret the data). Both views of LTKR are valid; neither is right or wrong.

The workshop's emphasis on business case development was driven primarily by a perceived need to "sell" decision-makers on the need to devote resources to long-term archiving. There was also a feeling that any future research agenda should be heavily driven by industry and government need.

A key lesson learned from the workshop was that the engineering community and the digital preservation community have much to gain by pooling their efforts. Both groups are grappling with many of the same issues. One such issue is lack of support and understanding of LTKR. Companies want to avoid the upfront overhead needed to archive. After all, the people burdened with archiving the data could be long gone before that data needs to be retrieved. An economic model is needed to rationalize archiving. However, the increasing volume of digital information being produced and companies' dependence on that information is making the business case for archiving more compelling.

But in spite of the common issues across archiving disciplines, every archiving scenario has its own unique requirements. It is very difficult for an archivist to determine what and how much information about a product needs to be submitted to a repository for effective retrieval in the future. No good software tools are available to help with this task. Requirements for such software tend to be application-specific.

The major barriers to archiving in the digital world are the lack of formal methods and standards for long term retention of engineering knowledge, uncertainty in the utility of the archived data, lack of good cost-benefit analysis of archiving, and inefficient archival procedures. To address these problems, one has to recognize that one-size-fits-all solutions for archival problems will not work for all engineered products. These products have different life times, from a few months for a cell phone to decades for a nuclear power plant. Policy guidelines and cost-benefit models are needed for choosing the archival model that best suites an industry.

Given the policy guidelines, the development and use of standards in the generation and recording of product information is crucial. An example of such an effort is the LOTAR (LOng Term ARchiving) project [23], which is developing a series of specifications for OAIS-compliant long-term archiving of product models represented using STEP. Here new standards are being created for the archiving based on an already standardized data representation.

The workshop concluded with a group discussion outlining a plan for advocacy and research and development. The following actions were proposed to address the issues and concerns raised in the context of design and manufacturing industry:

- Write a number of articles, with each tailored to a particular audience. The first article should be a report on the workshop. Additional articles should be written for research-oriented publications as well as for funding agencies and laypeople.
- Determine how best to capture workflows of business and manufacturing processes, and develop software tools to help automate the process capture
- Collect and preserve archiving case studies, as recommended in Crispin Hales' presentation (Section 2.2.5). Case studies can provide lessons learned by pointing out examples of poorly organized archived data.

- Collaborate with other groups concerned with long-term access to engineering designs such as the Design Society (http://www.designsociety.org/) and the ASME Design Automation Committee (http://divisions.asme.org/ded/daccomm/). Somebody should investigate these groups' activities and write a survey paper.
- Create a registry of engineering data formats, or contribute information on STEP and other engineering formats to the Harvard Global Digital Formats Registry

Acknowledgments
We are grateful to KC Morris, John Messina, Mark Carlisle, Jim Mays, Caroline Arms, Doug Cheney, and Crispin Hales for their helpful review of earlier drafts of this manuscript. We also wish to thank all of the LTKR workshop participants for their energy, enthusiasm, and good ideas.

Disclaimer
Mention of commercial products or services in this paper does not imply approval or endorsement by NIST, nor does it imply that such products or services are necessarily the best available for the purpose.

Appendix: Call for Participation, Agenda, and List of Participants

Long Term Knowledge Retention (LTKR): Archival and Representation Standards
March 15-16 2006
National Institute of Standards and Technology
Gaithersburg, MD 20899

Goal:

To identify challenges, research, and implementation issues in digital preservation of information with an emphasis on design and manufacturing.

Problem Statement:

In this age of Internet and networked economy, the rate at which the digital information generated is far exceeding the rate of consumption. According to some reports, today it takes about 15 minutes for the world to churn out new digital information equivalent to the entire collection in US Library of Congress. It does so about 100 times every day, for a grand total of five exabytes annually. This phenomenal proliferation of information clearly underscores the ease with which we can produce digital data. But our capacity to make all these digital information accessible in 200 or even 20 years remains a work in progress.

Recognizing the importance of these electronic records for its mission of preserving "essential evidence," the National Archives and Records Administration (NARA) launched a major new initiative, the Electronic Records Archives (ERA) initiative, in 1998. The Consultative Committee for Space Data Systems (CCSDS) recommendation established a common framework of terms and concepts which comprise an Open Archival Information System (OAIS) which was later adapted as ISO 14721:2003. Various other efforts are being explored to address the needs for long term knowledge retention in specific areas like manufacturing, health care and life sciences, legal, and military applications.

In all these efforts, standards play a very crucial role. In the area of engineering informatics, the LOTAR (LOng Term Archiving and Retrieval of digital technical product [23] documentation, such as 3D-CAD and PDM data) project studied the applicability of the international standards such as ISO 14721:2003 and ISO 10303 (STEP). The importance of digital preservation is clearly emphasized by various efforts as mentioned above and more specifically by the Digital Preservation Project of US Library of Congress (www.digitalpreservation.gov). But the long term retention of digital information is a work in progress and there are various issues that need to be addressed. In this workshop we intend to provide a forum for information and archival specialists, domain knowledge experts from manufacturing and product engineering, and other stakeholders, to discuss, among other things, the following set of issues:

1. **Digital Archiving Models, Representation Languages and Standards**
   - What constitutes a canonical representation for archiving?
   - How to compress data and develop data reduction schemes?

- How to manage interoperability among different archival systems?
- How to convert submission information to archived information and how to create disseminated information taking a holistic view of information package? This is essential to avoid fragmentation of creation, storage, and retrieval.
- Authentication and trustworthiness of archived information?
- What is the role of standards in information packages? How to develop standard schemas for submission information package, archival information package, dissemination information package, and Producer-Archive Interface Methodology Standard?
- Domain Taxonomies, Thesauri and Ontologies
- Role of markup languages and achieving Semantics Interoperability

2. **Challenges and Issues in Manufacturing Engineering Informatics**
   - What is to be archived beyond geometry information? How is this information to be represented?
   - Is STEP a starting point for content information?
   - How to scale from part level to system level information?
   - How to incorporate tolerance information?
   - What is the initial requirement (draft) for Preservation Description Information (PDI) for product data?
   - What are the Access points (for retrieval) for product data? Is there a role for generic features and contextual indexing?

Expected outcome:

A detailed roadmap identifying areas of investigation and experimental testbeds for archival of design and manufacturing information.

Organizing committee:

Co-Chairs

Joshua Lubell, NIST

Sudarsan Rachuri, NIST and George Washington University

William Regli, Drexel University

Committee members

Robert Chadduck, National Archives Records Administration

Eswaran Subrahmanian, Carnegie Mellon University and NIST

John Zimmerman, Dept. of Energy, National Nuclear Security Administration

Agenda

Panel 1: Challenges and Issues in Manufacturing Engineering Informatics
Panel 2: Digital Archiving Models, Representation Languages and Standards

March 15 2006	
Time	**Description**
8:30-9:00 AM	**Refreshments and Registration**
9:00-9:15	**Welcome and Introduction**
9:15-10:00	***Call to Action*, Dr. Robert Chadduck, National Archives**
10:00-10:15	**Coffee Break**
10:15-12:15	**Panel 1** **Doug Cheney (ITI Transcendata) – *AMBER Geometry Analysis*** **Crispin Hales (Hales-Gooch) – *Archiving Engineering Case Files for Future Reference*** **Jim Mays (Navy) – *Defense Archiving Issues and Initiatives*** **Frank Brown (Kansas Univ.) – *Design Geometry Inferencing***
12:15-01:15	**Lunch**
1:15-3:15	**Panel 2** **Don Sawyer (NASA) – *The Open Archival Information System (OAIS) Standard*** **Lou Reich (CSC) – *Metadata Standards for Archives*** **Burt Gischner (Electric Boat) – *The Role of ISO 10303 (STEP) in LTKR*** **Caroline Arms (Library of Congress) – *Sustainability of Digital Formats***
3:15-3:30	**Coffee Break**
3:30-5:30	**Parallel Break out sessions (2 -3 groups)**

March 16 2006	
Time	**Description**
8:30-9:00 AM	**Refreshments**
9:00-9:45	***Principles for Digital Preservation*, Henry M. Gladney, HMG Consulting**
9:45-10:00	**Coffee Break**
10:00-12:00	**Report from the break out sessions** **Will be divided among the groups**
12:00-12:15	**Concluding Remarks**

Breakout Group Agenda

Group 1: Manufacturing Informatics (Moderator: SK Gupta, Note taker: Lalit Patil, Jeff Abrahamson)

Group 2: Archiving Languages, Standards, and Representations (Moderator: Bill Regli, Note taker: Joe Kopena)

Outputs from Breakout Groups

1. Slides detailing group discussion (see topics below)
2. Roadmap based on gap analysis (see below)
    a. ASAP (desperately needed yesterday)
    b. Medium term (important but not as urgent)
    c. Long term (would be nice)

Discussion Topics for Both Groups

1. Evaluate state of archiving practice
    a. Current industry practice
    b. Current technologies and tools available
2. Archiving requirements analysis (what we should be doing)
3. Gap analysis
    a. Business processes
    b. Preservation processes
    c. Technologies
4. Develop archiving case studies (good anecdotes)
5. Enumerate ways in which archived information is used
    a. Legal (liability, incident investigation)
    b. Historical
    c. Design Rationale and Analysis (especially in the case of Engineering Informatics)
    d. …
6. Establish application-specific business cases for preservation (Electronic Record Archive project makes a generic business case).
    a. Impact of not preserving
    b. Ideas for providing economic incentives (when people doing the preservation work are likely to be long gone before the archived information is needed)
    c. Impact for taxpayer
    d. Impact for industry
    e. Archiving for fast changing information (e.g. internet digital information)
7. Technical architecture for archiving
    a. Does one size fit all?
    b. Is OAIS suitable for most applications?

Topics for Manufacturing Informatics Breakout Group

1. What would a document-based archiving approach (as opposed to repository-based) look like for engineering data?
2. How should assemblies be represented?
    a. Assembly in the small (geometry, mating, features, constraints)
    b. Assembly in the large (material handling, processes, packaging, shipping)

Topics for Archiving Languages, Standards, Representations Group

1. Intellectual Property issues (how much should we care?)
2. Authenticity and provenance
    a. Should it be part of the representation?
    b. Should it be part of the system (repository architecture)?
3. How do we address the issue of technology (hardware/software) evolution and how to synchronize the archival system with the fast pace of technology evolution

List of Participants

Caroline Arms, Library Of Congress, Washington, DC
Abdelaziz Bouras, University of Lyon, France
Lawrence Brandt, National Science Foundation
Frank Brown, Kansas University, Lawrence KS
Tony Brown, Atomic Weapons Establishment, United Kingdom
Robert Chadduck, National Archives and Records Administration, College Park MD
Wo Chang, Information Technology Laboratory, NIST
Douglas Cheney, ITI Transcendata, Milford OH
Mark Conrad, National Archives and Records Administration, College Park MD
Richard Eckenrode, BAE Systems, York PA
Burton Gischner, Electric Boat Corporation, Groton CT
Henry M. Gladney, HMG Consulting, Saratoga CA
Satyandra Gupta, University of Maryland, College Park MD
Crispin Hales, Hales & Gooch Ltd., Winnetka IL
Hyoil Han, Drexel University, Philadelphia PA
Martin Herman, Information Technology Laboratory, NIST
Ben Kassel, Naval Surface Warfare Center, West Bethesda MD
Joseph Kopena, Drexel University, Philadelphia PA
Joshua Lubell, Manufacturing Engineering Laboratory, NIST
James Mays, Naval Surface Warfare Center, West Bethesda MD
John Messina, Electrical and Electronics Engineering Laboratory, NIST
Gilles Neubert, University of Lyon, France
Yacine Ouzrout, University of Lyon, France
Lalit Patil, University of Michigan, Ann Arbor MI
Sudha Ram, University of Arizona, Tucson AZ
William Regli, Drexel University, Philadelphia PA
Louis Reich, Computer Sciences Corporation / NASA, Lanham-Seabrook MD
Donald Sawyer, NASA, Greenbelt MD

Ali Shokoufandeh, Drexel University, Philadelphia PA
Eswaran Subrahmanian, Manufacturing Engineering Laboratory, NIST
Rachuri Sudarsan, Manufacturing Engineering Laboratory, NIST
Jan Vandenbrande, Boeing Company, Seattle WA
David Wilkie, Drexel University, Philadelphia PA
Ronald Wood, Northrop Grumman Ship Systems, Pascagoula MS
John Zimmerman, Dept. of Energy – National Nuclear Security Admin., Kansas City MO

References

1. Supplement to the President's Budget for Fiscal Year 2007. http://www.nitrd.gov/pubs/2007supplement/07SuppEntireBook/07Supp_FINAL-022306.pdf, The Networking and Information Technology Research and Development Program, A report by the Subcommittee on Networking and Information Technology Research and Development, Committee on Technology National Science and Technology Council, February 2006.

2. STEP - NASA Website. http://step.nasa.gov . 2006.

3. Lots Of Copies Keep Stuff Safe. http://www.lockss.org . 2006. LOCKSS Alliance.

4. ISO/IEC 10918-1:1994: Information technology - Digital compression and coding of continuous-tone still images: Requirements and guidelines. 1994. International Organization for Standardization.

5. The National Digital Information Infrastructure and Preservation Program (NDIIP). http://www.digitalpreservation.gov/ . 2006.

6. National Nuclear Security Administration (NNSA). http://www.nnsa.doe.gov/ . 2006.

7. National Archives and Records Administration (NARA). http://www.archives.gov/index.html . 2006.

8. Electronic Records Archives (ERA). http://www.archives.gov/era/ . 2006.

9. ISO 10303-203: 1994, Product Data Representation and exchange - AP 203: Configuration controlled 3D design of mechanical parts and assemblies. International Organization for Standardization (ISO), Geneva, Switzerland.

10. Web Ontology Language (OWL). http://www.w3.org/2004/OWL/ . 2005.

11. David Leasure. A Logistica Deduction System for Solving NonMonotonic Reasoning Problems Using the Modal Logic Z. 1993. University of Kansas.

12. ISO/CD 10303-59. Product data representation and exchange: Integrated generic resource: Quality of product shape data, N4471. 8-14-2006. ISO TC184/SC4/WG12.

13. CAx Implementors Forum. Recommended Practices for Geometric Validation Properties. http://www.cax-if.org/ . 3-24-2006.

14. ISO/DIS 10303-28e2. Implementation methods: XML representations of EXPRESS schemas and data, N258. 1-20-2006. ISO TC184/SC4/WG11.

15. ISO 10303-21:2002. Industrial automation systems and integration - Product data representation and exchange - Part 21: Implementation methods: Clear text encoding of the exchange structure. 2002. International Organization for Standardization.

16. CCSDS 650.0-B-1: Reference Model for an Open Archival Information System (OAIS). Blue Book. Issue 1. ISO 14721:2003. http://public.ccsds.org/publications/archive/650x0b1.pdf . 2005. Consultative Committee for Space Data Systems.

17. ISO/IEC 21000-2:2003. Information technology -- Multimedia framework (MPEG-21) -- Part 2: Digital Item Declaration (1st ed.) . 2003. International Organization for Standardization.

18. Metadata Encoding and Transmission Standard (METS) Official Web Site. http://www.loc.gov/standards/mets/ . 2006.

19. IMS/CP. http://www.imsproject.org/content/packaging/ . 2006.

20. XML Formatted Data Unit (XFDU) Structure and Construction Rules - White Book. Consultative Committee for Space Data Systems (CCSDS) Secretariat, Office of Space Communication (Code M-3), National Aeronautics and Space Administration, Washington, DC 20546, USA, September 2004.

21. The Consultative Committee for Space Data Systems (CCSDS). http://public.ccsds.org/default.aspx . 2006.

22. Alex Ball. Briefing Paper: the OAIS Reference Model. http://homes.ukoln.ac.uk/~ab318/eprints/oaisBriefing.pdf . 2006. UKOLN, University of Bath.

23. Long Term Archiving and Retrieval of Product Data within the Aerospace Industry (LOTAR). http://www.prostep.org/file/17291.WP_LOTAR#search=LOTARproject [08/30/2002]. 2002. ProSTEP iViP Association.