NISTIR 7496

Long Term Sustainment Workshop Report

Joshua Lubell
Mahesh Mani
Eswaran Subrahmanian
Sudarsan Rachuri

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

NISTIR 7496

Long Term Sustainment Workshop Report

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

March 2008

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
James M. Turner, Acting Director

Abstract

This report summarizes the presentations, discussions and recommendations of a workshop held at the National Institute of Standards and Technology (NIST) on April 24-25, 2007. The purpose of the workshop was to identify policies of digital preservation and promising technologies to solve preservation problems in product design, engineering, and manufacturing in particular; with possible extensions to include chemistry, biology, and other scientific disciplines. An appendix provides the guidelines used for breakout sessions.

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.

1 Introduction

This report summarizes the presentations, discussions, and recommendations of *the Long Term Sustainment of Digital Information for Science and Engineering: Putting the Pieces Together* workshop held at the National Institute of Standards and Technology (NIST) on April 24-25, 2007[1]. The workshop focused on policies of digital preservation and on applying promising technologies to solve preservation problems in product design, engineering, and manufacturing in particular; with possible extensions to include chemistry, biology, and other disciplines where critical information must be "future-proofed." A diverse group of more than thirty participated, representing industry, government, and academia and bringing together researchers from disciplines including manufacturing engineering, library sciences, knowledge representation, and physical science.

The workshop was a sequel to March 2006's Long Term Knowledge Retention (LTKR) workshop [1] held at NIST and the February 2007 LTKR workshop [2] held at the University of Bath. The workshop differed from its predecessors in the following ways:

- There was a different mix of participants. The previous workshops had participation from communities developing archiving-related standards such as ISO 10303 – known informally as the Standard for the Exchange of Product Model Data (STEP) [3; 4], a standard for representing product model data; the Open Archival Information System (OAIS) Reference Model [5], a standard for archival information systems; and LOTAR (Long Term Archiving and Retrieval) [6; 7], a standard based on STEP and OAIS for the long-term archiving of 3D geometry and product data management information. This workshop had less representation from these groups, but more representation from implementers of both research-oriented and production-quality archival systems (see 3.4 and 3.5).
- More agenda time was allotted for breakout discussions, and the breakout discussions were conducted in a structured manner designed to facilitate brainstorming, the sharing of ideas, and the setting and future completion of goals.

[1] Presentations are available online at http://www.mel.nist.gov/div826/msid/sima/interopweek/meetings.htm#LTKR

- This workshop emphasized *sustainment*, an idea initially motivated by the Library of Congress (LC) analysis on the sustainability of digital formats [8]. A NIST goal is to extend LC's sustainability criteria to produce metrics specific to digital information used in scientific and engineering applications and, ultimately, to measure the quality of an archival strategy or policy. [9]

The workshop began with several invited presentations, followed by a preparation session for the breakout discussions. Next the participants divided into two breakout groups to discuss issues according to guidelines specified during the preparation session. Each group then presented its conclusions and recommendations to the other group. Section 2 discusses conclusions and takeaways, and spells out ideas for future work. The sections following Section 2 are of interest to readers of this report looking for more than just an "executive summary" of the workshop. Section 3 summarizes the invited presentations. Section 4 describes the process used to facilitate the breakout sessions. Section 5 shows the results of the breakout discussions and enumerates the issues discussed in the breakout groups.

2 Observations, Conclusions, and Next Steps

We begin this section by presenting some overall themes and conclusions that emerged from the discussions. We then enumerate some future research goals.

2.1 Common Themes

Workshop discussions attempted to balance the desire for good metadata from those accessing the archive against requirements for ease of use from those populating the archive. The need for both human effort and automation as well as the need for both technology and policies in achieving success should also be addressed. The underlying objective should be to understand the needs of the designated community and, using the classification of engineering standards that exist for modeling of information for a product life cycle, to arrive at a framework for achieving sustainability of scientific and engineering information.

It was argued that archival models should be such that facilities for archiving are available at the source of information creation. The argument for this approach is that archiving after the fact is often time-consuming and seldom undertaken. This observation and need has important implications for the design of a framework for archiving.

Another theme that emerged from the discussions was the need for an easy-to-use publicly available source containing information regarding digital formats, case studies, and software tools for archiving. A single "one-stop-shop" would make it easier to build an archive.

Yet another wish list item, perhaps at odds with the desire for registry consolidation and interoperability, was a need for more domain-specific standards and tools. The OAIS reference model, content packaging standards, and most software tools are generic in that they are not tailored to any particular application area or user community. In particular, it would be useful for reference architectures and tools to be aware of the underlying model of the information being represented. For example, an archive of digital product designs should support queries based on product structure or geometry.

2.2 Next Steps

Archiving is a socio-technical system design problem that requires cognizance not only of the social needs and mechanisms of archiving, but also of the technical possibilities of achieving the archival goals. Ensuring the long-term usability of science and engineering informatics artifacts is a challenge, particularly for engineered products with longer life cycles than the computing hardware and software used for their design and manufacture. Addressing this challenge requires characterizing the nature of the representation information needed, defining sustainability metrics, and developing methods for long-term preservation.

To address these challenges, it is necessary to characterize the nature of specific domains, developing methods for sustaining long-term usability of their artifacts and defining metrics for long-term digital information management. We consider the following ingredients to be necessary to achieve long-term sustainability:

- Representation methods for both product and process information
- A strategy based on anticipation of future access requirements for managing archived digital objects
- Domain-specific sustainability criteria and metrics
- A registry representing and classifying engineering and scientific digital objects
- Domain-specific extensions to the Open Archival Information System (OAIS) reference model.

3 Summaries of Invited Speaker Presentations

The invited presentations included not only general discussions of archiving issues (3.2 and 3.3), but also archival system implementation case studies (3.4 and 3.5). Although Kenneth Thibodeau's talk on the National Archives' Electronic Records Archives program (3.1) was part of the general plenary session on interoperability distinct from the workshop, we include it in our summary because of the subject matter's relevance.

3.1 Archival Interoperability: At the Intersection of Stability and Change

Dr. Kenneth Thibodeau, Electronic Records Archives Program Manager, US National Archives Records Administration (NARA)

The term *archives* implies sustained access. Thus the Electronic Records Archives (ERA) will be a dynamic, living system as opposed to an inactive data store.

ERA has a supply-side interoperability challenge in that during ingest it must cope with a diverse collection of data types, a mix of old and newer technologies, and information items containing multiple data formats and spanning multiple files (e.g. email messages with attachments). ERA also has an internal interoperability challenge in that the meaning of a record is determined not only by its content, but also by provenance and relationships with other records. For example, the 1999 accidental NATO bombing of the Chinese embassy in Belgrade did not result from a lack of detailed maps, but rather occurred because the particular map used by the field commander lacked information that a particular bombing target coincided with the current Chinese embassy location. Finally ERA has a demand-side interoperability challenge in that its most important customers

are those not yet born. ERA must be able to deliver authentic records to unknown systems at any time in the future.

ERA has three basic components: a research effort being conducted in partnership with universities, other government agencies, and other national archives; a business process modernization activity, and a system acquisition activity. ERA will not provide an end-to-end archiving solution, but instead will provide an archiving framework and search mechanisms. Although the cost of a complete archiving solution is estimated at $500 million, ERA is funded at only $300 million. NARA is collaborating with and leveraging efforts of other groups, such as the UK-based Digital Curation Centre.

NARA has gotten conflicting recommendations regarding whether to use XML to solve the persistent document problem.

3.2 Preservation of Engineering Artifacts

Dr. William C. Regli, Drexel University

At the 2006 LTKR workshop, Henry Gladney defined *digital preservation* as "the mitigation of the deleterious effects of technology, obsolescence, media degradation, and human memory." A central question from that workshop was what, if anything makes preservation of digital engineering data different and/or more difficult than for other domains.

To answer this question, let us suppose motion picture film preservation were like engineering. One would have to preserve:

- The film
- Who made it
- How the film was made, including camera position and lighting for each shot and an editing history
- Daily production logs recording who operated each piece of equipment, the weather, etcetera
- Moods and states of the actors.

From this thought experiment, we see that engineering preservation domain has the following special challenges:

- Complexity and diversity of data types whose semantics must be preserved, including design, modeling, simulation, and requirements information.
- Size of the data elements – representing a single machined part can require as much as several gigabytes of data
- The need to preserve changes to an object over time
- Business process workflows within and between engineering organizations must be captured
- Engineering information is both descriptive and prescriptive
- Lack of a well-defined stakeholder. Is it the engineered product's maintainer, the consumer, or is it the legal department?

The following are immediate requirements in order to meet the challenges of preserving engineering artifacts:

- Registries for computer aided design (CAD) and related engineering formats

- A collection of use cases to answer the questions of who will be using archived engineering data and for what purpose
- Representations for important data not covered by existing standards such as STEP
- Software tools to aid with ingest and access
- Open testbeds accessible to industry and academia
- Processes and best practices.

The Drexel University Geometric and Intelligent Computing Laboratory (GICL) has developed a Digital Archiving and Retrieval Tool (DART), available on SourceForge at http://sourceforge.net/projects/archivetool/. DART helps engineers archive, manage, and access information.

GICL is also collaborating with other universities to develop materials for teaching engineering informatics to undergraduates. These include analysis tools and a repository of biology-inspired designs.

3.3 From Knowledge Management to Digital Curation: A UK report and perspective

Chris McMahon, University of Bath

The University of Bath hosted the *Atlantic LTKR Workshop* February 12-13, 2007 as a follow-on to the March 2006 LTKR workshop at NIST. Like the 2006 LTKR workshop, this meeting concentrated on the two thrusts of manufacturing informatics and digital archiving models, representation languages and standards. The agenda and proceedings are online at http://www.ukoln.ac.uk/events/ltkr-2007/. Also like the 2006 LTKR workshop, participants consisted of a mix of universities, government agencies, and product and engineering data management companies, but with more European and less US representation.

Two large UK projects participated: *Knowledge and Information Management through Life* (KIM) and the *Digital Curation Centre* (DCC).

KIM addresses a number of information management issues:

- Product life cycles for complex, expensive items such as aircraft tend to be long
- Industry is moving toward a product support business model where the customer pays a fee for ongoing maintenance in addition to, or in some cases instead of, buying the product
- Organizations are becoming increasingly dynamically distributed, both geographically and through supply chains
- Customers are expecting more personalization, i.e., products customized to their specific requirements rather than one-size-fits-all.

The KIM experience has shown that companies are better at documenting their products than they are at documenting their processes. How a process should be documented depends partly on whether it is synchronous or asynchronous. KIM also introduced the notion of *datuments*, – computer-interpretable data objects embedded in documents. Typically items such as tables and figures in a document cannot easily be extracted in a computer-interpretable form. If these embedded items were represented in a fashion facilitating easy manipulation, a lot of value would be added to the documentation.

DCC (http://www.dcc.ac.uk) focuses on the active management and appraisal of data throughout the life cycle of scholarly and scientific materials. The next DCC conference will be held in Washington, DC on December 12-13, 2007.

A good allegory for curation is *Funes the Memorious*, a tale of a boy whose memory and perception are so clear that he has trouble distinguishing between different perspectives and between differences caused by time passages. As a result, he cannot make generalizations or reason using abstractions. Similarly, curation of engineering information requires knowledge of which information is most valuable to engineers and which information is unnecessary, and to keep the useful items while discarding the useless items.

3.4 ADAPT: An Approach to Digital Archiving and Preservation Technology

Mike Smorul, Institute for Advanced Computer Studies, University of Maryland at College Park

ADAPT includes four sub-projects: FOCUS (Format Curation Service), ACE (Auditing Control Environment), PAWN (Producer-Archive Workflow Network), and SRB (Storage Resource Broker) Replication Monitor.

FOCUS is a registry maintaining persistent information on digital formats and applications to validate and manipulate them. FOCUS is accessed either though the Lightweight Directory Access Protocol (LDAP) or through SOAP web services. Digital format processing is implemented using JHOVE. Because of overlap with the Harvard University Library's Global Digital Format Registry project and similar work being done by the UK National Archives, further development of FOCUS is on hold.

ACE supports a modular design, with the auditing function decoupled from the archiving system. Time-stamped certificates verifying an object's integrity throughout its life cycle provide an audit trail. Auditing may either be performed actively or it may be triggered by user events. ACE distinguishes itself from other integrity services based on cryptographic hashing in that it is designed to allow auditing of any component by an independent third party.

PAWN packages ingest workflows using a flexible mechanism for creating custom workflows. This is because experience has shown that a standard interaction model is impossible to achieve. PAWN handles distributed ingestion and includes its own scheduler tailored to the information being archived. It was not possible to use an "off-the-shelf" scheduler. PAWN is an extensible platform and includes both an application programmer interface (API) for creating client applications and pluggable modules for communicating with other components of an archival system.

When asked who will be responsible for operating PAWN, i.e., whether PAWN is intended to be used by producers of information to be archived or by the archivists themselves, the answer was that this depends on where responsibility of producer ends and responsibility of the archivist begins, and that the line of responsibility is unclear.

The Replication Monitor verifies availability of data in the archive using object IDs from the certificate managing system. Mirror sites are automatically synchronized with the

master site. The Replication Monitor was demonstrated in pilots involving the University of Maryland, NARA, the San Diego Supercomputing Center (SDSC), and the National Center for Atmospheric Research (NCAR).

3.5 FDsys Update

Kate Zward, US Government Printing Office

The FDsys (Future Digital System) program is developing a system to automate the collection and dissemination of electronic information from all three branches of the federal government. FDsys has a modular, standards-based architecture independent of specific hardware or software. Standards used include the Open Archival Information System (OAIS) reference model and the Metadata Encoding Transmission Standard (METS) for information packaging. FDsys extends the OAIS information package life cycle such that, in addition to there being an Archival Information Package (AIP), there is also an Access Content Package (ACP). This was needed because the developers of FDsys had found that packages are not fully formed at ingest; they "grow up" over time.

FDsys uses the PREMIS (PREservation Metadata: Implementation Strategies) dictionary for representing provenance and chain of custody information. Other FDsys administration metadata includes schemas for technical information.

The FDsys Dissemination Information Package (DIP) differs depending on a user's access privileges. As an example, the agency responsible for authoring a publication might get a DIP containing a document in a word processing format, whereas the general public would receive a DIP containing a Portable Document Format (PDF) file.

Piloting of Release 1 of FDsys is now underway, with a public launch planned for later this year. Release 1 will support permanent public access of archived documents and will include authentication of information ingested and disseminated as well as version tracking of publications. Future releases will support more advanced authentication and version control and will include tools to harvest information from government agency websites.

4 Process

This section contains the instructions used for the breakout process. They were adapted from the approach used for NIST's Smart Assembly Workshop [10], which in turn was adapted from a collaborative problem-solving method used in industry. [11] The instruction text is identical to that provided on handouts to the workshop participants; hence the use of present and future tense.

4.1 Breakout Session Process

The plan and agenda for the breakout sessions are based on a structured brainstorming and team-oriented problem solving process that has been applied successfully in many organizations.

The general process consists of:

- Problem/Opportunity Identification
- Analysis & Planning

- Root cause analysis
- Recommendation Generation
- Action Planning

A Session Facilitator will help to guide/mentor/facilitate the sessions --- but the content, ideas, analysis, and recommendations are the group's responsibility!!

Each session will also have a scribe to take notes and to prepare the material that will be presented in the group report.

There will be two breakout groups. All workshop participants will contribute to identify/ group the 'Top attributes' pertaining to the workshop, as to lead the discussions in the break out sessions. Depending on the preferences of the participants, the groups will be later assigned. Each group will spend most of the time in the session dealing with the identified Top Attributes, followed by recommendations. The identified breakout groups are:

1. **Archival Implementation and Technology Perspective:**
   The Top attributes for discussion in this group will be identified/ captured from the participants.
   (As a guide this group may explore what needs to be done and why from the applications view. The scope will include all aspects of digital archival system, programming and implementation, automation, searching and retrieval), and will include the functions from preparation through operation. The input from this group is expected to define the functionality needed.)

2. **Standards, Specific Domains, and Research perspective**
   The Top attributes for discussion in this group will be identified/ captured from the participants.
   (As a guide from the standpoint of R&D, where are the missing links? What technology voids hinder realization of archival system? What emerging technologies need to be matured? These questions will not explicitly be addressed in the breakout session, but will provide the background thought process for answering the question, "what needs to be done and why" – from the research perspective. This group will also discuss infrastructure and Standards issues like:
   a. *What foundational tools are required to realize an archival system? What computing power is needed and where does it need to be applied? What hardware and software architectures would accelerate the development and implementation of archival systems? What standards are needed to enable interoperability and cost effectiveness?*
   b. *Digital Archival Implementation and Best Practices*
      i. *Tools and Architecture*
      ii. *Methods, procedures, and rules of long-term preservation*
      iii. *Workflow Processes*
      iv. *Metadata Policies*
      v. *Archival Policy*
      vi. *Metrics to evaluate archival methods*

c. *Hardware Issues*
    i. *Storage Media*
    ii. *Archival Networks*
d. *Community of Interest for Digital Preservation*
    i. *Investigate the digital preservation needs of universities and research labs.*
    ii. *Address the issue of decrease of financial resources available for libraries and archives)*

Apart from these discussions, the identified groups could also discuss some common issues connecting both the groups. As a Guide:

- *What are we preserving? What are digital objects?*
  - *Physical, Logical, Conceptual (see Thibodeau paper) and its relationships*
- *Risk of Data Loss in Preservation Environments*
  - *Increasing loss of digital information*
- *Strategy to manage a rising tide of electronic records*
  - *Business case for digital archival*
  - *Is it possible or feasible to archive all digital data automatically and in a cost effective way?*
  - *Information explosion*
- *Plan for continuing technology change and rising user expectations*
  - *Archival Design Plan for Change*
  - *Proliferation of digital formats with hardware and software dependencies.*
- *Social, legal, and ethical issues*
  - *Legal Deposit of Digital Publications*
- *Develop an understanding of the archival usage context.*
  - *Digital preservation requirements*
  - *How much functionality can or must be preserved?*
  - *Maintaining digital information intact, while accessing this information in a dynamic use context – Digital Preservation Paradox?*
- *Digital record integrity and authenticity*
  - *Procedural/technical methods of authentication for preserved electronic records*

4.2 Identification

Top Attributes for Sustaining Long Term Digital Information
Involvement: All participants

Tuesday 24 April
This is the structured brainstorming part – we call it “Gallery of Ideas”

3:30 pm **Introduction to the Process**
Gallery of Ideas
Consider the following questions:
(1) What are the key characteristics and attributes of archival systems and
(2) From the perspective of the identified groups, "what needs to be done and why" to bring to realization archival systems with the characteristics and attributes that you define from Tuesday's presentations and your exercise from step 1.

3:35 pm **Idea generation**
Participants work alone. They will use post-it notes to answer the question, "what needs to be done and why?" They will write their answer in very clear terms so it can be read when posted. They will also record any additional information that will be helpful in understanding and documenting the ideas. Information like: broader statements of ideas, issues that should be addressed, definition of risks, barriers to be overcome, value of success (quantification and metrics). These notes will be posted and recorded. The facilitator will guide the exercise. Please include your name or session sign in number on the note so that we will be able to call you for clarification if needed following the workshop.

3:50 pm **Posting of ideas**
Each participant explains his/her most important ideas to the group and posts notes on board. Explanations/discussion should be brief and for clarification. (Note: The Facilitator will stop you if you over run your time). If someone has already posted your idea, move on to your next idea. Watch the time!

Analysis & Planning

This is where the team-oriented problem solving enters the picture.

4:30pm **Organize the ideas and prioritize categories**

Following the posting of ideas, the facilitator will seek consensus on key theme areas. The identified themes will be written on a flip chart, and the ideas will be grouped with those key themes. Any idea that does not have a place to be recorded will be posted in the parking lot. These parked ideas may find a home under the key themes, or they may be recorded as additional information. Select (through a voting process) the highest priority category(ies) for the team to work on in the next step. As a rule of thumb, five to six Top Attributes may be identified under each Theme.

These will be recorded in the Top attributes List (below) to lead the discussions in the break out sessions

Team ID: Group 1

"Top Attributes" List

Key Characteristics and Attributes of Sustaining Long Term Digital Information	

Remember that ALL of the ideas will be captured and included in the workshop report.

Top Attributes for Sustaining Long Term Digital Information Involvement: Individual Groups

Next Day
Wednesday 25 April

9:15am **Analyze and Generate Recommendations**

For the Top attributes identified for your group, generate recommendations or conclusions by selecting the best ideas or combining ideas. Complete one form (supplied) for each recommendation. Try to cover as much of the top attributes (output from the previous afternoon's exercise, see the figure above) as possible. Recommendations will be presented in the afternoon. Move on to as many priorities and repeat this process. A word of caution is offered: this workshop is a strategic event. The intent is to cover the broad topic of digital archival at an overview level. Therefore, the groups are cautioned that detailed analysis and long discussion will negate the realization of the goal of a relatively comprehensive overview. The facilitators will drive the group to move forward.

Select a spokesperson for the group

2:00pm **Reconvene for group reports.**

Submit completed electronic "Recommendation" forms to the Facilitator
Your team spokesperson will present for the group.
All of the ideas (post-it notes) and summary of the key discussions in the groups will be included in the workshop report to capture thoughts that were not analyzed by the group during the session.

4.3 Recommendations

The recommendation template looks like this:

NIST Workshop: Sustaining Long Term Digital Information
Team Report/ Recommendation Template

Breakout Team ID	
Identify Problem or Issue	
Analyze Root Cause	
Recommendation	
Benefit	

Plan: Action(s) to Implement	**Owner/ Time Frame**

- The problem is a statement of what's wrong.
- The root cause is a statement of why the problem exists. Every "why" is basically another problem statement. Sometimes you have to ask "why" a number of times to try to get to the root cause. The "root cause" is a re-statement of the observed problem in a way that lends itself more readily to "corrective action planning". This is like trying to get from the SYMPTOM to the DISEASE.

- The recommendation is a high-level plan or strategy to address the root cause of the problem. You will fill out ONE FORM for each recommendation. You may have more than one recommendation for the same problem, or your team may have time to consider more than one problem.
- The benefit is a statement of the business value of a solution. If your recommendation were implemented, what would that be worth to you? The more you can quantify this, the better!
- The action plan is a specific set of tasks, with assigned "owners" and (if possible) due dates, to *implement* the recommendation. The owner is responsible to see that the task gets done. It is usually a "rule" that the owner must be someone on the team who has agreed to accept the responsibility – not necessarily to DO the work, but to see that the work gets done. If you can't find an owner on your team, then suggest a natural owner (preferably someone at the workshop), or identify someone who will contact the owner to try to get them to agree to the task.
- Action plans should be as specific as possible, and correspond to things to be done AFTER the workshop to implement your recommendation. By itself, identifying a problem and making a recommendation is great – but if we can also document steps to take after the workshop, that is even better!

5 Breakout Sessions

There were two breakout groups: Group 1 focused on archival technology and generic issues, while Group 2 concentrated on standards, digital formats, and specific domains. Breakout sessions were conducted using an approach described in detail in the Appendix. Prior to forming the two groups, workshop participants collectively classified issues pertaining to long-term digital information sustainability to focus discussions in the breakout sessions. Participants began by brainstorming, submitting issues written on post-it notes to a "Gallery of Ideas" whiteboard, shown in Figure 1. Next the post-it ideas were divided between the two groups, the result photographed in Figure 2. Participants then extracted a set of key issues from the Gallery of Ideas (Figure 3). The key issues were then apportioned among Groups 1 and 2 as shown in Figure 4.

Workshop participants were then divided into the two breakout groups. Group 1, *Archival Information and Technology*, explored what needs to be done and why from an applications viewpoint. Their scope covered all aspects of digital archival system implementation and deployment, including the functions from preparation through operation. Group 2, *Standards and Specific Domains*, concentrated on the role of standards in solving fundamental issues with respect to digital information, with an emphasis on the product engineering domain. Each group used the methodology described in the Appendix to arrive at a set of problems and root causes, with a list of recommendations and an action plan for resolution.

Figure 1. Gallery of Ideas

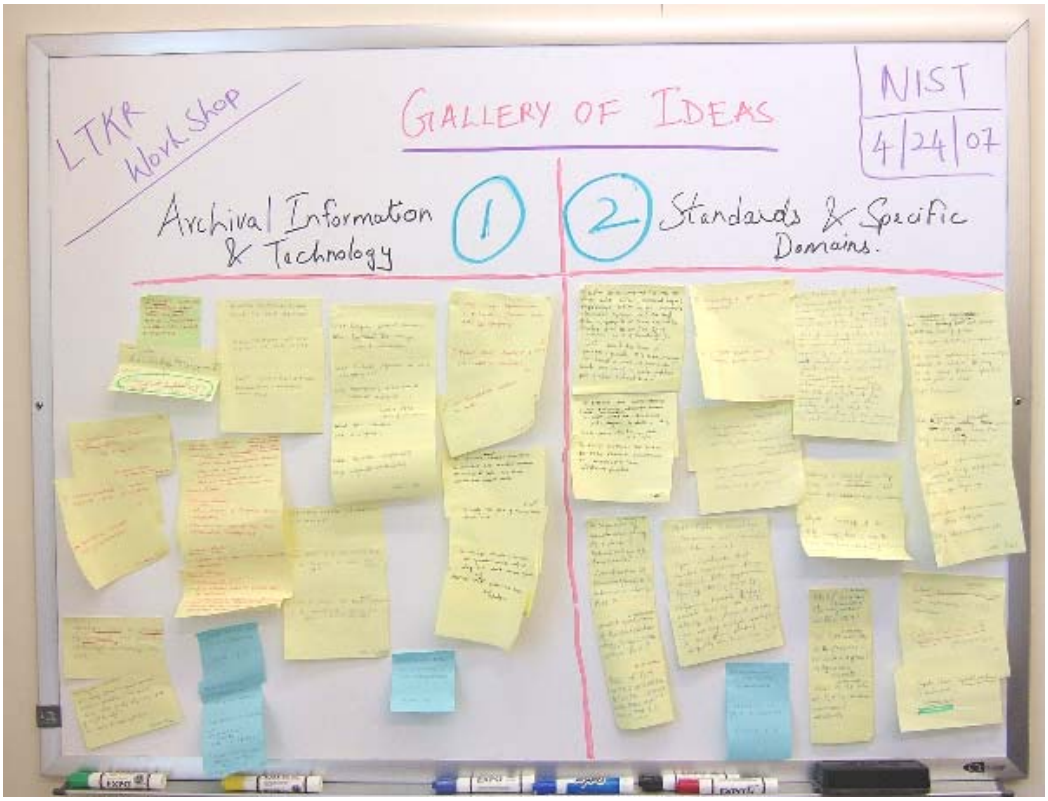

Figure 2. Post-its divided between the two breakout groups.

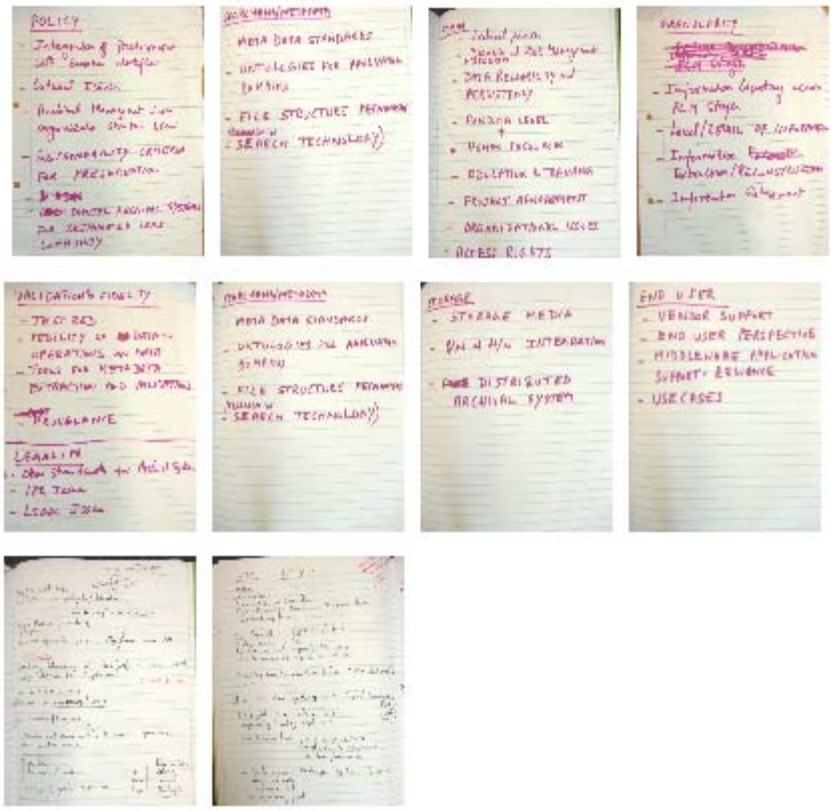

Figure 3. Extracting key issues from the Gallery of Ideas.

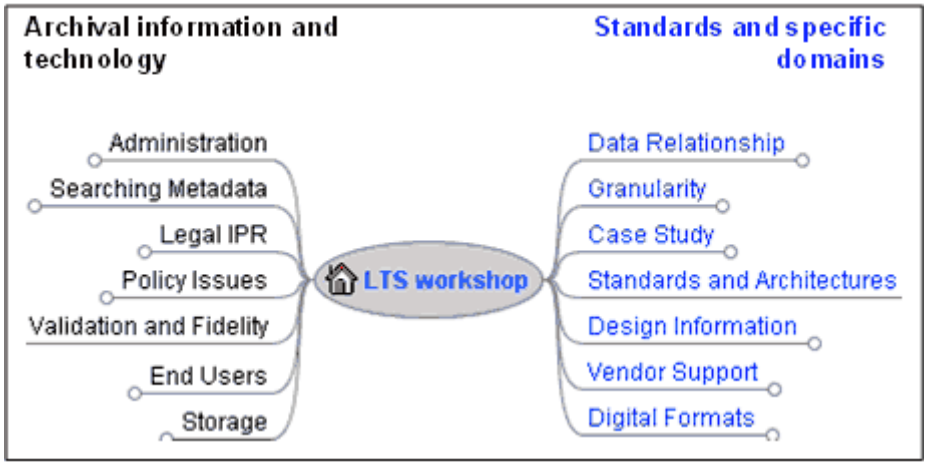

Figure 4. Key issues apportioned between the two breakout groups.

Sections 5.1 and 5.2 list the key issues and the corresponding subtopics for each issue (distilled from the Gallery of Ideas post-its) taken up for further elaboration by the breakout groups, followed by the issues the groups chose to discuss in detail. Owing to the time constraint only a small subset of the key issues were discussed in detail during the breakout sessions. The discussion for each issue was guided by the recommendation template as shown in Section 4.3.

Groups 1 and 2 used different approaches to choosing discussion items. Group 1 chose categories intended to span multiple subtopics whereas Group 2 chose discussion items

corresponding to individual subtopics. Ultimately, this did not matter much since the topics and subtopics are interrelated. The purpose of the breakout method was to serve as a catalyst for useful and productive discussions.

5.1 Group 1: Archival Information and Technology

Group 1 took up the following set of key issues and subtopics for further elaboration, starting with this set and identifying the most common themes. The subsections 5.1.1, 5.1.2, and 5.1.3 are the result of the re-categorization and prioritization

- Administration
    - Archival policies
    - Disaster management, risk management, and security
    - Data reliability and persistence
    - Funding level and human resources
    - Education and training
    - Project management
    - Organizational
    - Access rights
    - Configuration management
    - Revised/deleted archives
- Searching Metadata
    - Metadata standards
    - Ontologies for application domains
    - File structure mechanisms (file organization) – implementation metadata /application profiles
    - Research in search technologies
- Legal and Intellectual Property Rights (IPR)
    - Open (format) standards for archival systems
    - IPR
    - Legal
- Policy
    - Integration of preservation with business workflow
    - Culture
    - Archival management from organization structure view
    - Sustainability for preservation
    - Digital archival systems for designated user community
- Validation and Fidelity
    - Testbed
    - Fidelity of data
    - Tools for metadata extraction and validation
    - Provenance
- End Users
    - Vendor support - end-user tools

    - End user perspective
    - Reliance on middleware application support
    - Use cases
- Storage
    - Storage media
    - Software/hardware integration
    - Distributed archival systems – metadata for the archivist

Group 1 explored the following problem areas.

5.1.1 Data and metadata identification

The premise of this category is that data and metadata requirements differ for the various kinds of usage scenarios. From an end-user's perspective the issues include identification of the required data and metadata, extent, depth, and data formats.

Identified problem or issue: Identification of data and metadata required for ingest, scope, format, and data types for specific users and scenarios
Analyze root cause: Lack of knowledge about who the end users are and how they will use the data.
Recommendation: Collect access scenario use cases
Benefit: Prediction of current and future usage from an end user perspective. Help in identifying what should be saved and what should be discarded.

Plan: Action(s) to implement

- Identify end users and collect use cases and/or survey current end users
- Capture lessons learned from existing (including paper-based) archives (historical point of view)
- Provide mechanism, such as a wiki, for sharing lessons learned and use cases.

5.1.2 Getting the right information

The purpose of this category is to ensure that the creator of the information is able to support the archivist's task of providing relevant archival information packages (AIPs) for future end-users.
Identified problem or issue: Getting the right information from the creator for archiving from an archivist's point of view (which is to serve the end user)
Analyze root cause: Lack of rationale and rich documentation required for archival packaging of data for future use.
Recommendation: Create guidelines/requirements for creators to follow, while recognizing that they won't be universally followed. Potential enforcement can be through contractual mechanisms or peer review. Guidelines are needed to educate the creators of information in the information's preparation for archiving. Support creators through (semi-)automated tools.
Benefit: Facilitation of end-user support for archiving.
Plan: Action(s) to implement

- Survey different communities to determine whether or not adequate guidance exists
- Explore metadata extraction tools and watch for opportunities to influence vendors
- Establish Communities of Practice for archive submission requirements and guidelines. Develop guidelines for creators of information with participation of archivists and end-users.

5.1.3 Searching metadata

This category identifies the various needs and mechanisms for searching metadata. Research on search technologies was not included. The focus was exclusively on how to represent information in an archive.
Identified problem or issue: Encoding of traditional metadata and unstructured metadata (annotations, social tagging). Determination of metadata required for every object in the archive. Identification of format-specific metadata (e.g., binary image versus text, content – image could be x-ray or CT-Scan).
Analyze root cause: Lack of metadata that is context sensitive and format specific.
Recommendation: Follow guidelines and best practices for creation of structured and informal metadata. Classify different kinds of data objects and their metadata requirements. Review work done by NARA ERA research partners.
Benefit: Better searching, characterization of objects, validation of archives, metrics.

Plan: Action(s) to implement

- Review LC sustainability criteria to determine need for metadata for various formats, and do gap analysis for missing formats.
- Develop classification for engineering and science domains for different data formats.
- Articulate preferred set of data elements and encoding for each classification.

5.2 Group 2: Standards and Specific Domains

Group 2 took up the following identified key issues and corresponding subtopics for further elaboration.

- Data Relationship
  - Traceability
  - Information models for archival systems
- Granularity
  - Information capture through product life cycle stages
  - Level/detail of information
  - Information extraction by reconstruction
  - Information refinement
- Case Study
  - Archival engineering body of knowledge
  - Tools repository

    - Analysis of existing archival systems
    - Engineering data corpus
- Standards and Architectures
    - Extensions to OAIS for multimedia documents and environments
    - AIP interoperability
    - Classification of AIP
    - Reference architectures based on OAIS
    - Standards landscape for archival
    - Identification and use of established standards and best practices
- Design Information
    - Design history capture
    - Standard for extraction of design and manufacturing rules from knowledge based support tools
- Vendor Support
    - Incentives for software vendors to implement standards
    - Reconciling multiple data formats
    - Vendor support for post delivery maintenance
- Digital Formats
    - Preservation of format registry
    - Merging and differentiation standards
    - Detection of format obsolescence
    - Format migration
    - Flexible handling of formats
    - Workflows
    - Software version

The following subsections contain the subtopics Group 2 chose to discuss in detail.

5.2.1 Case study

There is a lack of proper case studies demonstrating and validating the usefulness of long term sustainability approaches for engineering information. Case studies are needed to create an archival engineering body of knowledge (experience base), a software tools repository, and analysis of existing archival systems.

Identified Problem or Issue: Lack of repositories of available case studies, archival tools, and methods, leading to difficulty in identifying best practices
Analyze Root Cause: Lack of vendor interest and lack of coordinated efforts across interested parties.
Recommendation: Create consolidated registries to serve as a clearing house for available technologies for archival purposes and to promote educational awareness (e.g., Global Digital Format Registry (GDFR)).
Benefit: Sharing of resources and tools through a single portal.
Plan: Action(s) to implement

- Promote the use of wikis for communities of interest
- Create/ add- to online informational web resources (e.g., www.dcc.ac.uk/tools)

- Contribute to GDFR and other consolidated registries
- Assemble test data for engineering archival systems and organizations

5.2.2 Granularity

Granularity represents the relative size, scale and level of detail of information generated, extracted, and refined throughout the product life cycle stages.

Identified Problem or Issue: Lack of metrics and methods for determining the minimal granularity requirements based on essential domain-specific characteristics
Analyze Root Cause: Lack of potential use cases and use of service level agreements.
***Recommendation*:** Map granularity to usage requirements.
Benefit: Targeted delivery of information based on cost/risk assessments.
Plan: Action(s) to implement

- Identify domain-specific characteristics of granularity needs based on quality and functionality factors
- Identify service level agreements
- Identify metrics to characterize the granularity of information
- Characterize user requirements with respect to granularity

5.2.3 Standards and Architectures

Discussions on standards and architectures focused on extensions to the OAIS reference model for domain specific multimedia documents and environments. The extensions would require the specification of a domain-specific archival information package (AIP) interoperability and classification, reference architectures, and a repository of established standards and best practices.

Identified problem or issue: Lack of domain-specific reference architectures, related use cases, and knowledge of representation information
Analyze root cause: Lack of clear understanding of usage scenarios and requirements. Lack of policies and business case for archiving.
Recommendation: Understand the needs of the user community and analyze industry-specific requirements.
Benefit: Reference architectures for specific communities
Plan: Action(s) to implement

- Identify and classify usage requirements for specific domains (e.g., engineering design, construction)
- Identify reference architectures based on use cases, requirements and scenarios
- Participate in the LOTAR project
- Analyze domain-specific file formats
- Identify information handles and annotation practices
- Identify representation standards for specific architectures

5.2.4 Design Information

Design information, in addition to geometry, includes design history, design rationale, and rules for design and manufacturing. No processes exist for standardized representation and extraction of non-geometric information for specific analytical tasks.

Identified problem or issue: Insufficient methods and representation to capture design history and design rationale
Analyze root cause: No comprehensive methods to capture design history explicitly and, even if captured, they are disconnected
Recommendation: Create context-sensitive, analysis needs-based definition of digital records.
Benefit: Efficient methods to capture design history and rationale

Plan: Action(s) to implement

- Identify methods and representation for meaningful and efficient extraction of design history and rationale
- Create information representations through analysis of use cases for "as designed, as manufactured and as maintained" stages of the product life cycle.
- Research manufacturing history, supply chain environment, collaborative tools used

5.2.5 Digital Formats

Issues include preservation of format information, merging and differentiation standards, detection of format obsolescence, format migration, flexible handling of formats, workflows, and software versioning.

Identified problem or issue: Proliferation of digital formats and transformation problems among digital formats
Analyze root cause: Vendor business models and independent application development, leading to proliferation of specific formats and slow standards development
Recommendation: Promote standardized digital formats. Create methods for converting cost (current and estimated) of transformation of digital format into the total cost of ownership of the digital object.
Benefit: Interoperability, cost reduction in future data retrieval and use

Plan: Action(s) to implement

- Classify and characterize digital formats using public registries (e.g., GDFR)
- Identify business models for competing needs of the vendor
- If IPR issues prevent vendors from contributing format information to a public registry, develop mechanisms to transform data into a non-proprietary standard format, or provide methods to protect intellectual property while maintaining benefits of public registries

6 References

1. Lubell, J., Rachuri, S., Subrahmanian, E., Mani, M., and Regli, W. C., "Long Term Knowledge Retention Workshop Summary," National Institute of Standards and Technology, NISTIR 7386, http://www.nist.gov/msidlibrary/doc/NISTIR_7386.pdf, Dec. 2006.

2. Ball, A., and Ding, L., "Proceedings of the Atlantic Workshop on Long Term Knowledge Retention 2007," University of Bath, http://www.ukoln.ac.uk/events/ltkr-2007/proceedings/, Apr. 2007.

3. Kemmerer, S., "STEP: The Grand Experience, (Editor)," National Institute of Standards and Technology, NIST Special Publication 939, http://www.nist.gov/msidlibrary/doc/stepbook.pdf, 1999.

4. ISO 10303-1: Industrial automation systems and integration -- Product data representation and exchange -- Part 1: Overview and fundamental principles. 1994. Geneva, Switzerland, ISO.

5. 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.

6. Dutz, A.. LOTAR Standards - The perfect alternative and not just for the aerospace industry. Product Data Journal [2], 15-17. 2007. ProSTEP.

7. LOTAR - LOng Term Archiving and Retrieval of digital technical product documentation, such as 3D-CAD and PDM data. http://www.asd-stan.org/Lotar.html . 11-29-2007. ASD-STAN.

8. Sustainability of Digital Formats. http://www.digitalpreservation.gov/formats/ . 5-21-2007. Library of Congress.

9. Lubell, J., Rachuri, S., Subrahmanian, E., and Mani, M., "Sustaining Engineering Informatics: Toward Methods and Metrics for Digital Curation," 3rd International Digital Curation Conference, http://www.nist.gov/msidlibrary/doc/dcc07.pdf,2007.

10. Smart Assembly: Industry Needs and Technical Challenges. http://smartassembly.wikispaces.com/ . 2008. National Institute of Standards and Technology.

11. Shuit, D. P.. GM Goes Fast. Workforce Management , 36-38. 2004. Crain Communications, Inc.