SANDIA REPORT

SAND2010-2477

Unlimited Release

Printed November 2010

Long-term Archival and Retrieval of Engineering Data: Implications for the DART Workbench

Vincent De Sapio

Prepared by

Sandia National Laboratories

Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Approved for public release; further dissemination unlimited.

Sandia National Laboratories

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

NOTICE: This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government, nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, make any warranty, express or implied, or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represent that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government, any agency thereof, or any of their contractors or subcontractors. The views and opinions expressed herein do not necessarily state or reflect those of the United States Government, any agency thereof, or any of their contractors.

Printed in the United States of America. This report has been reproduced directly from the best available copy.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865) 576-8401
Facsimile: (865) 576-5728
E-Mail: reports@adonis.osti.gov
Online ordering: <http://www.osti.gov/bridge>

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd
Springfield, VA 22161

Telephone: (800) 553-6847
Facsimile: (703) 605-6900
E-Mail: orders@ntis.fedworld.gov
Online ordering: <http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online>

Long-term Archival and Retrieval of Engineering Data: Implications for the DART Workbench

Vincent De Sapio
Sandia National Laboratories
MS 9159
P.O. Box 969, Livermore, CA 94551 U.S.A.
vdesap@sandia.gov

Abstract

Long term archival and retrieval of engineering data is a key component of product life cycle management. Standards have emerged over time to facilitate archival and retrieval, particularly with respect to CAD data. In this report we assess the state of the art in industry, academia, and government, with respect to archival management and preservation planning. Additionally, we present different scenarios for the integration of the DART Workbench tool with an archival and retrieval system. These scenarios are based on existing CAD/CAE industry standards as well as Semantic Web technologies. The goal is not to design and implement a long-term archival system, but rather, to anticipate potential future systems and consider how the DART Workbench may integrate with them. A secondary goal is to provide some reasoned input for eventual lab-wide implementation of a particular approach, thus steering the design of such a system to some degree.

Contents

1	Introduction	7
2	STEP as a Neutral Archival Format	8
3	Scenario for Workbench Integration with STEP Archival	10
4	Elements of an Open Archival System (OAIS)	11
4.1	Participants, Components, and Information Packages	11
5	Scenario for Workbench Integration with OAIS Model	13
6	Conclusion	15
	References	16

Appendix

A	Summary of the LOTAR project	17
---	------------------------------------	----

Figures

1	Scenario for workbench integration with STEP.	10
2	General process flow of the OAIS model.	11
3	Scenario for Workbench integration with the OAIS model.	13

1 Introduction

A central component of product life cycle (PLC) management is the long term archival and retrieval of engineering data. Standards have been developed over a number of years to facilitate robust archival and retrieval, particularly with respect to CAD data. This report seeks to assess the state of the art in industry, academia, and government, with respect to archival management and preservation planning. Subsequently, different scenarios will be investigated for the integration of the DART Workbench tool [1] with an archival and retrieval system. The actual implementation of this is heavily contingent upon Sandia's lab-wide posture toward product life cycle management and long term archival. Since a lab-wide system is not currently in place, the objective of this report is to consider possible directions and generate high-level strategies for the integration with a yet to be implemented protocol for long term archiving and retrieval. These possible directions are based on existing CAD/CAE industry standards as well other technologies which have emerged from the industrial and academic communities [3] [5] [10] [11].

2 STEP as a Neutral Archival Format

The **ST**andard for the **E**xchange of **P**roduct model data (STEP) is a set of standards (parts) developed under ISO 10303 [4]. It provides a neutral machine interpretable representation of product data throughout the life cycle, independent of any proprietary component. The set of standards comprised by STEP fall into one of the following categories: description methods, integrated resources, application protocols, abstract test suites, implementation methods, and conformance testing. More specifically, the parts include,

- Part 1 - Overview
- Parts 1x - Description methods (the EXPRESS family of information modeling languages)
- Parts 2x - Implementation methods (data exchange mechanisms)
- Parts 3x - Conformance testing methodology and framework
- Parts 4x - Integrated generic information models
- Parts 1xx - Integrated application resource models
- Parts 2xx - Application Protocols (specific models targeted for product data exchange)
- Parts 3xx - Abstract Test Suites (corresponding to the Application Protocol series)
- Parts 4xx - Implementation Modules (corresponding to the Application Protocol series)
- Parts 5xx - Application Integrated Constructs (corresponding to the Application Protocol series)
- Parts 1xxx - Application Modules

STEP uses a modeling language for product data, EXPRESS ISO 10303-11, to specify the product information to be represented. STEP also employs application protocols (AP's) to specify the representation of product information for one or more applications. AP 203 is the application protocol for configuration controlled 3D designs of mechanical parts and assemblies. As with AP 214 (Core data for automotive mechanical design processes) it is supported by a number of CAD vendors, including PTC [7] (Windchill), for file import and export. A number of companies, including Boeing, Lockheed Martin, General Electric, and Northrop Grumman, use STEP AP 203 as a production standard for product model data.

STEP focuses on product model data and CAD. It is more limited with regard to CAE data. CAE vendors like MSC (PATRAN) support AP 203 through the import of product model geometry. STEP AP 209 (Composite and Metallic Structural Analysis and Related Design) addresses FEA data, but it is not deployed in the CAE industry to the extent that AP 203 is deployed in the CAD industry. It may be a potential solution for commercial analysis codes (contingent upon industry adoption), but there is still the need to address archival of data from Sandia in-house analysis codes.

While STEP is a well established standard there are newer technologies based on XML standards for e-commerce and the Semantic Web [9]. The Semantic Web is based on ontologies that formally define all concepts within a given domain (in our case, the product life cycle domain). A fundamental component of the Semantic Web is the Resource Description Framework (RDF) [8]. The RDF is a standard for the

conceptual description of information using subject-predicate-object expressions to define relationships between objects. A network of these expressions form a directed graph, similar to Class and Object diagrams in object-oriented design. In this respect RDF can be viewed as an ontology language. In keeping with Semantic Web standards RDF can be formatted as XML.

Some recent developments in STEP are leveraging XML and Semantic Web technologies. There is an application protocol that addresses Product Life Cycle Support (PLCS), AP 239. This is relevant with regard to long-term archival of product data. PLCS, as defined through AP 239, is characterized by the creation and management of an Assured set of Product and Support Information (APSI), over a period of time (life cycle). AP 239 specifies an information model that can be customized by an organization through the use of Reference Data Libraries (RDL's). Similar to an RDF, an RDL specifies the semantics necessary for deployment in a PLCS system. Eurostep is developing a set of Web Ontology Language (OWL) ontologies [6]. A prototype RDL Server has been developed by Eurostep [2] under contract to the UK Ministry of Defense.

In addition to AP 239, AP 28 specifies XML representations of EXPRESS schemas and data, and in this way provides STEP support for XML. One particular benefit of this is the capability for facile data transformation using XSLT/XSL, a declarative, XML-based language for the transformation of XML documents into other XML documents.

Figure 1. Scenario for workbench integration with STEP.

3 Scenario for Workbench Integration with STEP Archival

A minimal scenario for Workbench integration of a STEP archival model is simply to use AP 209 as a way of exchanging commercial CAE data into a neutral STEP format for long-term archival. This is illustrated in *Figure 1*. This scenario is highly contingent upon commercial CAE support for AP 209 and, as mentioned earlier, there is still a need to address archival of data from Sandia in-house analysis codes, as well as the vast amount of additional engineering data that the Workbench needs to manage. *Figure 1* emphasizes these contingencies with dashed lines and cross marks.

Using AP 239 and AP 28 ontologies/RDL's would need to be developed to describe the other data sources. This is certainly not a plug-and-play solution and would require substantial system development. This approach is analogous to the Open Archival System to be addressed in the following section.

Figure 2. General process flow of the OAIS model.

4 Elements of an Open Archival System (OAIS)

The Open Archival Information System (OAIS) is a standard developed under ISO 14721:2003. The OAIS is an archive with affiliated constituents, including an organization of people and systems [3]. The information being archived in this model has a requirement for long term preservation. This requirement introduces numerous challenges associated with evolving software technologies (both proprietary and open source) and file formats, evolving hardware and media technologies, and a changing user base.

4.1 Participants, Components, and Information Packages

The OAIS reference model defines the three types of **participants** in an open archival system to be the *producers*, the *archive*, and the *consumers*. The four types of **components** are, *ingestion*, *archival storage*, *data management*, and *access*.

Information packages involve a combination of the bit-stream file content data (data object), the syntactic and semantic information needed for data interpretation (representation information), and the information needed to identify, verify, and certify the data, (preservation description information). The three types of

information packages are *submission*, *archive*, and *dissemination*. Submission information corresponds to information the producer wishes to archive. Archive information corresponds to information the archive stores. Dissemination information corresponds to information retrieved by the consumer from the archive.

In addition to the participants, components, and information packages described, another key element of the OAIS reference model is a *file format registry*. This is the source of representation information for all of the digital content. This representation information provides an ontological description for all data streams. This includes a formal description of the syntax and semantics for a file format. In terms of Semantic Web technologies the file format registry can leverage the RDF (as well as OWL) which acts as a formal language (using subject-predicate-object expressions) for representing ontologies. This would allow for automated reasoning about archived data.

There is one additional administrative component that directly supports the file format registry. That is the *preservation planning component*. This component would contain facilities for accessing the archive and file format registry, and reasoning about data preservation based on the state of the archive and the data ontologies represented in the file format registry. After a CAD/CAE file is submitted by a design engineer or analyst the preservation planning component would be able to access the file format registry and, based on the ontological description, convert the file to a long-term sustainable format using a suite of existing translators at its disposal. It could also orchestrate the creation of other files from the source file (2D models and rendered images/movies from the 3D model or CAE results, etc.). These would be archived as deemed appropriate. A process flow for the OAIS reference model is illustrated in *Figure 2*.

Figure 3. Scenario for Workbench integration with the OAIS model.

5 Scenario for Workbench Integration with OAIS Model

A scenario for Workbench integration with an open long-term archival system, based on the OAIS model, is illustrated in *Figure 3*. Of course this is contingent upon a lab-wide commitment to such a system, since the design and implementation would be beyond the funding of the Workbench project.

In this scenario the ingestion, access, and preservation planning phases would be subsumed by the Workbench. It is conceived that the bulk of these phases would be transparent to the users (producers/consumers) and that their interaction with the Workbench would be largely unchanged. During ingestion of any data object (file/data stream) from a producer there would be the generation of a submission information package (SIP). This would include data objects directly submitted by a producer as well as any data objects derived from source data (i.e., analysis results, etc.). The SIP would contain the bit-stream file content, representation information, and preservation description information. During the ingestion phase the format registry would be queried for validation and transformation of the SIP to an archive information package (AIP). The AIP would be submitted to the the archive. During the access phase the format registry would again be queried for validation and transformation of the AIP to a dissemination information package (DIP).

The Workbench would provide facilities for the consumer to receive the data. The preservation planning component would periodically query the archive and format registry and reason about appropriate data preservation. This would entail automatic conversion of data to sustainable formats. Again, this would be ideally transparent to the user.

6 Conclusion

A number of technologies directed at facilitating the long-term archival and retrieval of design and analysis data have been discussed. These include open standards-based technologies like STEP and OAIS, as well as Semantic Web technologies like XML, RDF, and OWL. The goal is not to design and implement a long-term archival system, but rather, to anticipate potential future systems and consider how the DART Workbench may integrate with them. A secondary goal is to provide some reasoned input for eventual lab-wide implementation of a particular approach, thus steering the design of such a system to some degree.

Two scenarios have been detailed. The first is the most limited and involves the use of STEP standards. STEP AP 203 and AP 214 address neutral long-term archival formats for CAD design data. In the case of the Workbench, where CAE data is of particular relevance, AP 209 may be a potential solution for commercial analysis codes. This is contingent upon widespread industry adoption. Additionally, there is the need to address archival of data from Sandia in-house analysis codes. The use of more recent developments in the STEP community, including XML and Semantic Web technologies as well as Product Life Cycle Support (AP 28 and AP 239), offers the capability of transcending the current limitations of STEP. Additionally, the OAIS reference model offers an integrated approach to managing long-term archival. The second scenario builds on this model and lays out a high-level scheme for integration of the OAIS model components with the DART Workbench.

References

- [1] DART Workbench. <https://dta.ran.sandia.gov/dart>.
- [2] Eurostep Global. <http://www.eurostep.com/>.
- [3] M. J. Grauer, I. K. Howley, J. B. Kopena, and W. C. Regli. Towards a format registry for engineering data. In *Proceedings of the 2007 ASME International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, September 2007.
- [4] ISO standards. <http://www.iso.org/iso/catalogue>.
- [5] D. Mcwherter, M. Peabody, A. Shokoufandeh, and W. C. Regli. Database techniques for archival of solid models. In *Proceedings of the 6th ACM symposium on Solid Modeling and Applications*, pages 78–87, 2001.
- [6] OWL Web Ontology Language. <http://www.w3.org/TR/owl-features/>.
- [7] Windchill. <http://www.ptc.com/products/windchill/>.
- [8] Resource Description Framework (RDF). <http://www.w3.org/RDF/>.
- [9] Semantic Web. http://www.w3.org/2001/sw/wiki/Main_Page.
- [10] R. Sharma and J. Gao. STEP PLCS for design and in-service product data management. In *Advances in Design*, volume 5 of *Springer Series in Advanced Manufacturing*, pages 293–301. Springer, 2006.
- [11] X. W. Xu and Q. He. Striving for a total integration of CAD, CAPP, CAM and CNC. *Robotics and Computer-Integrated Manufacturing*, 20(2):101–109, 2004.

A Summary of the LOTAR project

A.1 Introduction

LOTAR is an acronym for LOnG Term Archival & Retrieval. It is a project with the aerospace industry involving four areas of engineering data archival:

- Geometry topology and shape representation
- Geometry dimensioning and tolerancing attributes
- PLM and technical data
- Validation rules, properties and data quality attributes

The driver in the aerospace industry is the FAA's 50 - 70 year retrieval requirements. For Digital Data, the challenge is that the data is often stored in a proprietary, native format and will most likely be uninterpretable over time. Consequently, archiving data in its native form requires periodic migration to the new release (version) and this method quite often leads to data loss and the repair can be costly. A typical technological obsolescence cycle of a CAD generation roll (i.e., CATIA V4 V5) is 3 - 5 years.

The use of a neutral archiving format safeguards the interpretability of the data for a much longer period of time, perhaps its entire retention period. Neutral forms make it easier to migrate the data based on the way that the Application Protocols (AP's) are structured. In addition, their life expectancy (obsolescence cycle) is significantly longer in duration.

To meet these requirements the participants have adopted an open source neutral 3D model format in STEP. Many CAD vendors support the STEP AP 214 format directly or through external translators (<http://www.cadverter.com/>). PTC supports AP 214 Product Lifecycle Management (PLM) services (Windchill).

A.1.1 Life Cycle Information Planning

- What information should we archive?
- What is the configuration of the information?
- What is the information context?
- What is the format of the information and what form does it need to be stored in?
- How long do we need to keep the data?
- How frequently do we need to access the data?

A.1.2 Life Cycle Phases

- Ingest
- Preservation planning
- Access

A.1.3 LOTAR Approach

Develop an architecture that supports:

- Data architecture containing:
 - Semantic representation
 - Open and Neutral forms (STEP)
 - Data Quality and Validation
- Process architecture:
 - Based on Open Archive Information System (OAIS - ISO 14721)
 - ISO 10303 (STEP)

A.1.4 Participants

- Boeing
- Airbus
- General Dynamics
- Rockwell Collins
- Dassault
- PTC
- NIST
- DoD
- DoE (Sandia)
- NASA

and others.

A.1.5 NAS9300-xxx Standards

- 9300-010 Common Process
- 9300-011 Data Preparation
- 9300-012 Ingest
- 9300-013 Archival Storage
- 9300-014 Retrieval
- 9300-015 Removal
- 9300-016 Test Suites
- 9300-017 Audits

A.2 LOTAR and OAIS

This summary has been gleaned from LOTAR documents:

The Open Archive Information System (OAIS) model defines the processes and actors which ingest the data into an archive, and which provide services to consumers of the data, including both query and retrieval. The most subtle area, and possibly the least understood, is the construction of the web of information needed to correctly read the data once it has been retrieved. The LOTAR standard uses the OAIS reference model as a basic framework, providing specific guidance on specialized types of data; initially Mechanical CAD/CAM/CAI and non-geometric meta data. The problem here is not to be sure that the data comes in and out correctly, but that it is being correctly interpreted by the new generation of software. That is, if information is data in context, and the context is the application which interprets the data, then LOTAR looks at information retention. In short, how do we know that the design we look at in twenty years time is the same as the design we look at in our current system? LOTAR makes the assumption that we know what we need to archive. Lifecycle Information Planning asks the question, “how do we retain our product knowledge (i.e. Design Intent) throughout the life of the product?” This is wider than the OAIS question, “what do we need to be able to understand this particular package of data?”, rather asks “what data about a product should we keep?” Although the answer starts with obvious elements such as the design and the configuration, it soon gets into areas such as the preservation of design rationale, the processes by which the product was designed, and the organizational structures that enable those processes to operate.

Figure A.1. From *LOTAR 101 - A Project Overview*.

DISTRIBUTION:

1	MS 9155	Robert Clay, 8953
2	MS 9159	Vincent De Sapio, 8953
2	MS 0933	David White, 9510
2	MS 9018	Central Technical Files, 8945-1
2	MS 0899	Technical Library, 4536