Balisage: ***The Markup Conference***

Metadata for Long Term Preservation of Product Data

Joshua Lubell
Computer Scientist
National Institute of Standards and Technology
<lubell@nist.gov>

Abstract

Product data can be usefully defined as structured information about objects that are produced by industrial and business processes. In terms of information types, data formats, usage, and lifespan, product data is both complex and diverse, encompassing 3D image modeling information, dimensions, tolerances, and other model annotations, supplementary material such as test analysis, videos, datasets, and human-readable documentation. Although the metadata issues in this problem space present some unique challenges, there are valuable lessons to be learned from the library metadata and packaging standards and how they relate to product metadata. Extending the library standards to represent subsets of information from emerging product lifecycle management standards could help tame the complexity of long-term archival of product data.

Table of Contents

Digital Product Data Archiving Challenges

ISO 10303 — also known as STEP (the Standard for the Exchange of Product Model Data) — defines a *product* as "a thing or substance produced by a natural or artificial process" and *product data* as "a representation of information about a product in a formal[1] manner suitable for communication, interpretation, or processing by human beings or by computers." [ISO10303-1] Despite the breadth of what STEP considers to be a product, product data is overwhelmingly used in practice to mean structured information about things produced by industrial or business processes.

Prior to the digital information age, product geometry was specified using technical drawings on paper. These drawings were copied using blueprint and other reproduction methods. [Wikipedia1] The drawings were often archived on microfiche, a medium which can last as long as 100 or more years. [Wikipedia2] Nowadays many Computer Aided Design (CAD) systems provide a means for engineers to specify product data through annotated 3D digital models. Thus the de facto definition of a product is no longer drawing-based, but rather is model-based.

Product data is complex and diverse, both in terms of information types and in terms of formats. Therefore, it is not surprising that useful long-term archiving of product data is difficult. Some reasons why are as follows:

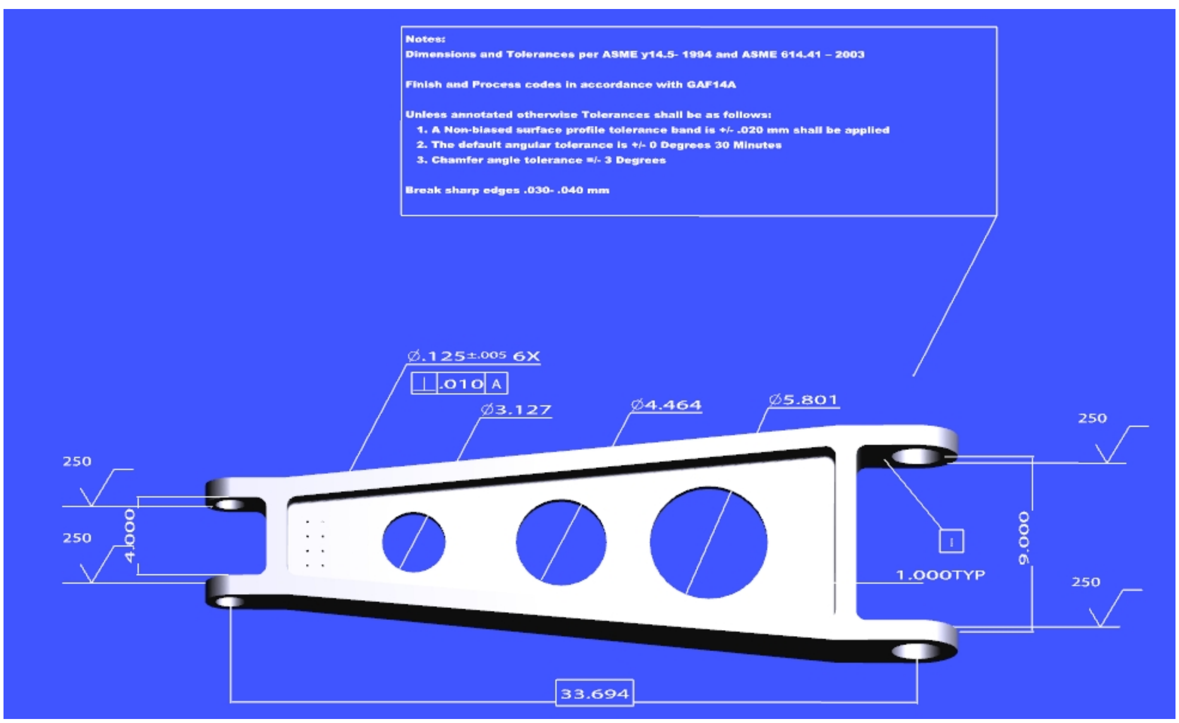
- Engineers often want the digital models and systems they build today to be extensible and reusable by subsequent generations of technologists. This requirement presents a challenge as a digital product model may have a longer lifespan than the data formats, application software and computing platforms used to create the model.
- Archived digital product models should be semantically rich enough to address long-term socio-technical concerns such as forensics (accident and incident investigation) and environmental issues (carbon footprint, disposal). For example, the former might require representation of the rationale behind design decisions, while the latter might

require a product model highlighting any hazardous substances needed to build a part.

- Although content information standards such as STEP are a critical ingredient for long term archiving, not all relevant content information can be made available using standards. Standards' representational capabilities sometimes lag behind those of information models implemented in commercial software, as vendors are constantly adding new enhancements to distinguish their offerings from the competition. Thus content information may use a mix of standard and proprietary formats.
- Data accuracy is required for product quality and manufacturability. A small anomaly in an engineering design can have great economic and social consequences throughout a product’s lifecycle. The verification and validation of engineering data requires complex computations.
- There are many different data types such as product models, geometry, and simulation data such as finite element models. In addition, there may be associated documentation, multimedia, and other information – leading to a variety of different digital formats. Not only do all these data types and formats need to be transparent, but the relationships between the information units that are part of these data types must also be made explicit. Digital formats can include neutral exchange standards such as STEP, 3D visualization formats, office document formats, and proprietary modeling formats.

Product Manufacturing Information (PMI) is a particularly critical constituent of product data. Loosely defined, PMI can include any sort of information defining a product’s components for manufacturing, inspection and sustainment. PMI includes geometric dimensions and tolerances (GD&T) specified in a formal language whose syntax and semantics are defined in several American Society of Mechanical Engineers (ASME) and ISO standards. PMI is essential to product data. Manufacturing cannot take place without it. Today's CAD systems must support not only the ability to edit and display PMI, but also must be able to feed GD&T semantics to computer-aided manufacturing (CAM) systems and other downstream applications. [Srinivasan]

Figure 1



3D visualization of a bracket part with PMI, displayed in Adobe Reader. [2]

As an example of product data consider Figure 1, a screen shot of a 3D visualization of a bracket part created from a digital 3D model developed with a CAD system. There are several points worth mentioning about this example.

- The visualization can be rotated and otherwise manipulated to maximize understanding by humans. However, it is not intended for exchange between CAD systems. It complements rather than replaces the original CAD model.
- The image has PMI callouts, using the standard syntax to specify GD&T annotations. Even though the visualization lacks the full semantics of the original CAD model, it is helpful in communicating how to manufacture the part — arguably more so than an annotated 2D drawing would be.
- The callout at the top of the figure is unstructured text. This is also part of the product data. In fact, product data can include arbitrary attachments such as spreadsheets containing analysis results, video data, and other associated information relevant to the product. These attachments are important and need to be managed and monitored, even if CAD and CAM systems do not "understand" their data formats.

Manufacturers recognize the need for long-term archiving of digital product data and the challenges associated with doing so. As a result, industry-led standardization efforts are beginning to emerge. The most prominent and farthest along of these is LOTAR (LOng Term ARchiving), a suite of standards being developed by an international consortium of aerospace companies. [LOTAR] LOTAR's approach is based upon the Open Archival Information System (OAIS) reference model [ISO14721], a framework widely adopted by archivists and preservationists. LOTAR favors the use of STEP in the archival process while recognizing the reality that a mix of product data archival formats must be supported.

Of course the need for archiving information is not unique to product manufacturers, but in fact applies to many disciplines. For instance, those of us working in archival of digital products can learn much from librarians. They were among the first to embrace OAIS and have implemented numerous tools and standards for digital archives and repositories. In particular, they have long recognized the importance of metadata. Metadata is an information retrieval system's lifeblood. The choice and quality of metadata has a major impact not only on delivery of data, but also on long-term preservation of that data. [Gartner] Librarians understand this well and, in fact, have standardized a multitude of metadata vocabularies, many of them based on the Extensible Markup Language (XML). [XML]

Similarly, this paper takes a metadata-centric approach to product data archiving. The section "Engineering-specific Metadata Requirements for Long-term Archiving" discusses metadata requirements unique to long-term product data access and how these requirements fit into the framework of OAIS. The section "Metadata Vocabularies in the Library and Product Data Worlds" gives an overview of digital library metadata vocabularies and contrasts them with two metadata standards from the world of product data. The section "Summary and Moving Forward" provides some closing remarks and suggestions for future work.

Engineering-specific Metadata Requirements for Long-term Archiving

The linchpin of the OAIS reference model is the "information package," an object containing content (e.g., product model data) along with the metadata needed to adequately interpret, preserve, maintain, and access the content. OAIS provides a taxonomy of metadata types used with information packages. In this paper I consider the three OAIS metadata types most widely implemented in XML:

- Preservation Description Information (PDI): enables adequate preservation of content (includes chain of custody, identification, authentication, and contextual metadata)
- Packaging Information: aggregates and identifies the constituents of an Information Package
- Descriptive Information (DI): supports "the search and retrieval of archived information"

Among these metadata types, DI (also called "descriptive metadata") is most closely connected with enabling and facilitating access to archived data. DI is more domain-sensitive than PDI or packaging information in that a DI schema for product data is likely to look very different from a DI schema for library holdings or archived emails. Because DI is domain-sensitive, and enabling efficient access to archived product data is so important, this section focuses on product model-specific DI requirements. The section "Metadata Vocabularies in the Library and Product Data Worlds" discusses existing metadata vocabularies for all three types.

In the following subsections I discuss in depth some factors that determine descriptive product model metadata. I first present a classification of types of engineering data archive access I call the "3Rs." [Lubell] I then discuss the multitude, complexity, and evolution of digital formats for product model data. Finally, I discuss the mismatch between standards and implementations.

Product Models and the "3Rs": Reference, Reuse, and Rationale

A key challenge in long-term archiving is ensuring that information be available in the best form for whatever access scenarios the future might bring. This is particularly true for product data. When validating engineering knowledge, it is crucial to be able to replicate the intended behavior of a designed artifact. Successful validation requires that the information be available in forms sufficiently supporting retrieval and reuse. The need to know a designer's intent becomes important in the context of redesign and reuse of existing parts. Another important aspect of engineering archiving is the ability to store the digital objects at different levels of granularity and abstractions as required by design decision making tasks. Without the ability to compose different digital objects for archiving, it would not be possible to maintain the ability to encode rationale or reuse-based access needs.

I therefore consider end-user needs from the point of view of reference, reuse, and rationale – the "3Rs" – to better understand the level of granularity and abstractions required in the definition of digital objects. By "end user," I mean entities who are consumers of OAIS information packages.

The 3Rs – reference, reuse, and rationale – define a set of access capabilities for a designated community. *Reference*, the most elementary type of access, is the ability to read the digital object and produce the digital object for proper reproduction in a given display medium (computer display, paper, etc.). *Reuse* is the ability to refer to and modify the digital object in an appropriate system environment (software and hardware). *Rationale* is the highest level of access in which the end user should be able to refer, reuse, and explain the decisions about the content of the digital object. The need to know a designer's intent becomes important in the context of redesign and reuse of existing parts.

The primary driver for the 3Rs is the specific retrieval needs of each of these scenarios. For example, a product data repository intended primarily for reference may need to be organized differently than one intended for reuse, where not only the geometric aspects of the product are sought but also additional information regarding manufacturing, performance requirements, assembly, and other aspects. In a similar vein, rationale may require content being organized in a different manner to support the inclusion of requirements information along with other performance data on the part or the assembly. Given the range of uses and perspectives of the end users, their needs will have a large impact on the process of archiving and retrieval.

Variety of Formats

As discussed in the section "Digital Product Data Archiving Challenges", many formats are used to represent product model data. Some are native to a specific software application. Others are neutral with respect to software applications. Some are proprietary. Others are open or standards-based. The selection of the product model data format depends upon factors such as the type of data defined in the product model, design stage, and the availability of translators. Because an archive must capture all of the data required to completely define the product and associated processes, the archive may need a variety of data formats, possibly causing redundancies. Good metadata is essential for helping consumers evaluate the quality of data accessed from the archive.

Native formats are typically binary, proprietary, and have a specification not available to the general public. In spite of long-term access concerns, archiving the native data is done as a matter of course because it has such a small impact on resources and because it is universally accepted as a good system management practice. [Kassel] STEP provides an open alternative to native formats for long-term retention of product information. Because STEP is an international standard developed by consensus, it is less subject to change than a proprietary format.

STEP application protocols (APs) specify information models for a specific engineering domain. STEP physical files (informally known as Part 21 files or STEP files) use an ASCII format defined in ISO 10303-21. [ISO10303-21] A STEP processor can be any software application capable of interpreting and/or generating STEP physical files, for example a computer-aided design (CAD) tool capable of importing and exporting STEP files, or a visualization tool that can import STEP data. The objects represented and exchanged using STEP, as well as the associations between these objects, are defined in schemas written in EXPRESS (ISO 10303-11) [ISO10303-11], a product data information modeling language. Recognizing the popularity of XML as an implementation method, the developers of STEP later standardized Part 28 [ISO10303-28], specifying an alternative representation of STEP schemas and data using XML.

Mismatch Between Standards and Implementations

Although open non-proprietary standardized formats are desirable for long-term data retention, even information represented using standards-based methods is subject to format change. For example, consider a long-term archival scenario from the domain of shipbuilding [Kassel] attempting to maximize the use of STEP to represent detail design product model data.

Although there are STEP APs that define information models well-tailored to the ship domain, these APs have not yet been implemented in commercial off-the-shelf software. Other APs such as AP203 (configuration controlled 3D design) and AP214 (core data for automotive mechanical design processes) are supported by today's CAD software applications, but the information models of these APs cannot represent some concepts specific to shipbuilding. So how might ship data be archived, given this imperfect state of affairs? Initially, the data is created in a native format, and the neutral file format is selected as a function of the quality of the available translators. If the desired translators are not available, a compromise is made in order to allow the data to be accessible to the applications used during a specific design phase where access to this neutrally represented data is required. The evolution from this point forward could be as follows.

First, geometry is exchanged using any means possible, but most likely using AP214 or AP203. The non-graphical data can be extracted separately and saved in a project-specific XML format. The XML often contains only basic product structure, enabling minimal exchange of geometric data, graphics, and basic properties. It also has the greatest potential for minimizing the dependency on a particular CAD vendor.

A more desirable long-term solution would be for the product model software supplier to implement AP239 (the Product Life Cycle Support (PLCS) standard) [PLCS] to define product structure, the relationships between objects, and reference data libraries to define an extensible set of properties. The section "Product Data and Lifecycle Management" discusses PLCS in further detail.

Assuming future vendor support, the ideal long-term solution would be to employ implementations of AP214 for general purpose geometry and PLCS for configuration management and product structure. This approach would be further enhanced by defining application-specific product model data using APs with a more specific scope, such as the shipbuilding APs.

In summary, even with a policy favoring STEP for representing product information, the detail design data may manifest itself in many different formats, each conforming to a different AP, or in some cases not conforming to STEP at all. Over the life of the product, format choices will evolve as a function of the product model complexity, the quality of the translators, and the evolution of the standard itself. Therefore, the goal of a single standard to represent product model data may not be realistic, and we should be prepared to encounter multiple formats over the lifecycle of a product.

Metadata Vocabularies in the Library and Product Data Worlds

So far I have discussed product data's diversity and complexity, the importance of metadata in archival systems in general, and why Descriptive Information (DI) in particular is key to ensuring the long-term usability of product data. Now I attempt to answer two questions:

- How should we go about defining metadata vocabularies needed for product data archival information systems?
- Can the metadata vocabularies developed for digital libraries help?

Digital Library Metadata Standards

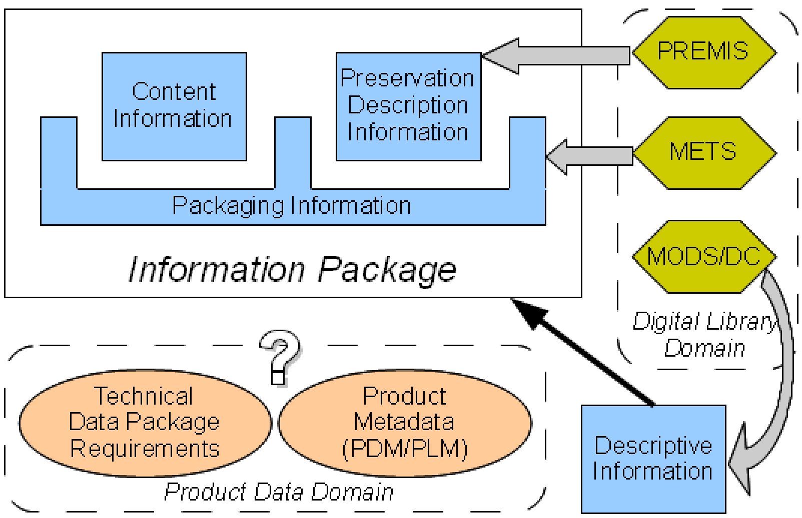
I begin by addressing the second question, surveying some of the more widely adopted XML-based metadata vocabularies for digital libraries. The Metadata Encoding and Transmission Standard (METS) [METS] specifies an XML schema [XSchema] for encoding packaging information and is extensible such that other vocabularies can be incorporated for encoding PDI and DI. METS Profiles, detailed and unambiguous descriptions of classes of METS instances, are used to document METS schema extensions and customizations. The PREMIS (**Pre**servation **M**etadata: **I**mplementation **S**trategies) data dictionary and XML schema [PREMIS] define a vocabulary suitable for representing PDI. METS and PREMIS are often combined to create a more expressive schema, though there are some inconsistencies and overlaps. [Guenther] Dublin Core (DC) [DCMI] and the Metadata Object Description Schema (MODS) [MODS] are both vocabularies for descriptive metadata and are easy to integrate with METS. MODS offers more precision than DC, but DC is more widely used. [Gartner]

An additional advantage of DC is that it is encoded using the Resource Description Framework (RDF). [RDF] RDF's simple, modular data model facilitates extensions to DC. The wide variety of software tools available for RDF can be used to edit, process, and query DC metadata. Most important, it is easy to merge and harmonize RDF metadata from heterogeneous sources. The importance of this last benefit will become clearer later on in this paper.

Given the variety and widespread adoption of digital library metadata vocabularies, it would seem that these vocabularies could be helpful for encoding the packaging information, PDI, and DI needed for a product data information package. But there is one big caveat, and this is that the DI requirements for product data archives differ significantly from those of digital libraries, as previously discussed in the section "Engineering-specific Metadata Requirements for Long-term Archiving". Not surprisingly, product data-specific distinctions such as the 3Rs, product model evolution, and changing data formats are outside the scope of the digital library metadata standards.

In the next subsections, I will try to answer the first question by looking at two potential metadata sources unique to the world of product data: technical data packages (TDPs) and product data and lifecycle management (PDM/PLM) information. A TDP defines and aggregates the information objects needed to support a specific activity in a product's lifecycle, and is analogous to OAIS packaging information. PDM/PLM information is highly detailed metadata used to "integrate people, data, processes, and business systems and provide a product information backbone for companies and their extended enterprise." [PLM]

Figure 2



Metadata sources for OAIS Information Packages in the digital library world versus the product data world.

But first let us review our discussion so far regarding OAIS and metadata. As shown in Figure 2, an OAIS information package is a container bound via packaging information, inside of which is content as well as PDI providing the administrative metadata needed for long-term preservation of the content. DI is domain-sensitive metadata used to facilitate information retrieval from the archive. XML metadata standards for digital libraries (shown as hexagons) provide vocabularies for representing packaging information, PDI, and DI. However, the digital library vocabularies for DI do not address product data-specific retrieval concerns.

Technical Data Packages

The United States Department of Defense defines a Technical Data Package (TDP) as follows in MID-STD-31000 [MID-STD-31000]:

> A technical description of an item adequate for supporting an acquisition strategy, production, and engineering and logistics support. The description defines the required design configuration or performance requirements, and procedures required to ensure adequacy of item performance. It consists of applicable technical data such as models, drawings, associated lists, specifications, standards, performance requirements, quality assurance provisions, software documentation and packaging details.

As the MIL-STD-31000 options selection worksheet in Figure 3 shows, a TDP provides packaging and administrative information as well as some DI such as whether the product model data is conceptual, a more detailed design, or sufficient for manufacturing (e.g., box 1.A.).

Figure 3

TDP OPTION SELECTION WORKSHEET

SYSTEM: DATE PREPARED:

A. CONTRACT NO.	B. EXHIBIT/ATTACHMENT NO.	C. CLIN	D. CDRL DATA ITEM NO(s).

1. TDP Level (X and complete as applicable .)

A. ☐ CONCEPTUAL LEVEL
☐ DEVELOPMENTAL LEVEL
☐ PRODUCTION LEVEL

B. REMARKS:

2. TYPE AND FORMAT (X all that apply and complete as applicable .)

A.
☐ TYPE 2D: 2D DRAWINGS
☐ TYPE 3D: 3D MODELS ONLY
☐ TYPE 3D: 3D MODELS WITH ASSOCIATED 2D DRAWINGS

B.
☐ NATIVE CAD (SPECIFY TYPE) ____________
☐ ISO 10303 STEP FORMAT (Specify STEP PROTOCOL AP203, AP 214 etc.) ____________
☐ ISO 32000 PORTABLE DOCUMENT FORMAT ____________
☐ OTHER ELECTRONIC FORMAT (SPECIFY TYPE) ____________
☐ HARDCOPY ____________
REMARKS: ____________

3. CAGE Code AND DOCUMENT NUMBERS

A. ☐ CONTRACTOR CAGE AND DOCUMENT NUMBERS
☐ GOVERNMENT CAGE (COMPLETE 3B & 3C OR 3D)

B. USE CAGE CODE:

C. USE DOCUMENT NUMBERS:

D. To Be Assigned By:

4. DRAWING FORMATS (X one and complete as applicable)

☐ CONTRACTOR FORMAT.
☐ GOVERNMENT FORMAT.
REMARKS: ____________

5. TDP ELEMENTS REQUIRED (X all that apply)

☐ ELEMENTS REQUIRED TO BE DETERMINED BY CONTRACTOR - OR THE FOLLOWING ARE REQUIRED:

☐ CONCEPTUAL DRAWINGS/MODELS AND ASSOCIATED LISTS
☐ DEVELOPMENTAL DESIGN DRAWINGS/MODELS AND ASSOCIATED LISTS
☐ PRODUCT DRAWINGS/MODELS AND ASSOCIATED LISTS
☐ COMMERCIAL DRAWINGS/MODELS AND ASSOCIATED LISTS
☐ QUALITY ASSURANCE PROVISIONS
☐ SPECIAL INSPECTION EQUIPMENT (SIE) DRAWINGS/MODELS AND ASSOCIATED LISTS
☐ SPECIAL TOOLING (ST) DRAWINGS/MODELS AND ASSOCIATED LISTS
☐ SPECIFICATIONS
☐ SOFTWARE DOCUMENTATION
☐ SPECIAL PACKAGING INSTRUCTIONS (SPI) DRAWINGS/MODELS AND ASSOCIATED LISTS

6. ASSOCIATED LIST (X and complete as applicable)

☐ A. PARTS LIST (X ONE)	☐ (1) INTEGRAL	☐ (2) SEPARATE
☐ B. DATA LISTS (X ONE)	☐ (1) NOT REQUIRED	☐ (2) REQUIRED (SPECIFY LEVELS OF ASSEMBLY)
☐ C. INDEX LISTS (X ONE)	☐ (1) NOT REQUIRED	☐ (2) REQUIRED (SPECIFY LEVELS OF ASSEMBLY)
☐ D. WIRING LISTS (X ONE)	☐ (1) NOT REQUIRED	☐ (2) REQUIRED (SPECIFY LEVELS OF ASSEMBLY)
☐ E. INDENTURED DATA LISTS (X ONE)	☐ (1) NOT REQUIRED	☐ (2) REQUIRED (SPECIFY LEVELS OF ASSEMBLY)
☐ F. APPLICATION LISTS (X ONE)	☐ (1) NOT REQUIRED	☐ (2) REQUIRED (SPECIFY LEVELS OF ASSEMBLY)

7. APPLICABILITY OF STANDARDS. The following Standards apply: (X as applicable)

☐ ASME Y14.100 ENGINEERING DRAWING PRACTICES WITH APPENDICES: ☐ B ☐ C ☐ D ☐ E	☐ ASME Y14.24 TYPES AND APPLICATIONS OF ENGINEERING DRAWINGS ☐ ASME Y14.34 ASSOCIATED LIST ☐ ASME Y14.35 REVISION OF ENGINEERING DRAWINGS AND ASSOCIATED LIST ☐ ASME Y14.41 DIGITAL PRODUCT DEFINITION DATA PRACTICES ☐ ASME Y14.5 DIMENSIONING AND TOLERANCING	☐ OTHER STANDARDS APPLY AS DESCRIBED: COMPANY STANDARDS PERMITTED ☐YES ☐NO

Technical Data Package option selection worksheet from MIL-STD-31000.

The options selection worksheet and accompanying instructions in MIL-STD-31000 imply a schema for TDP metadata. Instances of the implicit schema can be thought of as contracts between various parties in a product's lifecycle specifying the TDP contents needed for the lifecycle activity. The TDP metadata and contents vary depending on the activity, actors, and lifecycle stage. For example, consider a TDP supporting procurement by an aircraft manufacturer of a jet engine from a tier 1 supplier versus a TDP supporting maintenance of the aircraft. The procurement TDP might require a 3D model (or visualization) specifying PMI needed for the supplier to manufacture a jet engine compatible with the aircraft design. The

TDP supporting maintenance, on the other hand, might require technical manuals and troubleshooting documentation.

From an OAIS viewpoint, a TDP is an information package that may contain any content and metadata necessary for a product to be designed, manufactured, purchased and/or maintained. Recalling Figure 2, the library metadata standards collectively can represent much of an information package's metadata. But can these standards also represent TDP metadata? To answer this question, consider the following four top-level elements of the METS schema. `<dmdSec>` defines an information package's DI (represented with embedded Dublin Core or MODS). `<admSec>` defines the information package's PDI (with embedded PREMIS markup). `<fileSec>` specifies the locations and grouping of files comprising the information package. `<structMap>` defines the information package's lexical structure and relates it to the files in `<fileSec>`.

Now consider the TDP option selection worksheet in Figure 3. Table I shows how the numbered sections of the worksheet could map into the top-level METS elements. The table is a first step toward demonstrating the feasibility of using library standards to encode TDP metadata. As mentioned in the section “Engineering-specific Metadata Requirements for Long-term Archiving”, DI is the most domain-sensitive component of information package metadata. Looking at the TDP metadata, it indeed seems to be true that the DI (e.g., TDP level, applicability of PMI standards) is more unique to product data and more challenging to encode using library standards than the non-DI metadata (e.g., CAGE codes, document numbers).

Table I

Possible mapping of TDP metadata to top-level METS elements.

METS section	TDP metadata
Descriptive Metadata (`<dmdSec>`)	TDP Level, Drawing Formats, Applicability of Standards
Administrative Metadata (`<admSec>`)	Type and Format, CAGE Code[3] and Document Numbers
File Section (`<fileSec>`)	[locations of required TDP elements and associated lists]
Structural Map (`<structMap>`)	TDP Elements Required, Associated List

Given the preceding discussion, it stands to reason that success using the library standards approach to represent TDP metadata hinges on being able to extend library DI standards (MODS or Dublin Core) such that they can be used within a METS instance to specify TDP descriptive metadata. Because Dublin Core is encoded in RDF and has provisions for defining application profiles (context-specific extensions to the basic DC metadata terms) [DCAP], it appears to be the strongest candidate for representing TDP DI. Thus the discussion in this section suggests that using METS and Dublin Core is a promising approach for encoding a TDP in a manner interoperable with digital repository frameworks and other software applications supporting library metadata standards. By creating TDP-specific application profiles for METS and Dublin Core, these standards can represent the TDP metadata elements specified in MIL-STD-31000.

Product Data and Lifecycle Management

Product Data Management systems have evolved over the years from simple file managers for engineering drawings into complex software suites satisfying a wide variety of functions. Currently supported capabilities include version control of product models, management of design modifications, and management of engineering and business processes. A thorough discussion of PDM systems would require a paper all its own, and those who wish to learn more are encouraged to read Srinivasan's chapter in “Advanced Design and Manufacturing Based on STEP” [Srinivasan] for a detailed introduction to PDM and PDM standards. Suffice it to say that the information used in PDM systems is essentially product metadata. Because of the complexities of product data, PDM systems must provide very rich vocabularies for describing parts, people, resources, documents, activities, and other entities involved in the design, manufacturing, and support of products.

Since PDM is such an integral part of product development, STEP not only supports the exchange of 3D product geometry and PMI but also supports the exchange of PDM information as well. Ongoing growth of the capabilities of PDM systems has resulted in the emergence of newer standards that build upon the STEP information model for PDM. The newest of these standards is the Product Lifecycle Support (PLCS) suite of data exchange specifications [PLCS] from the Organization for the Advancement of Structured Information Standards (OASIS) and based on STEP AP239.

PLCS has a number of appealing characteristics as a source of DI for product data:

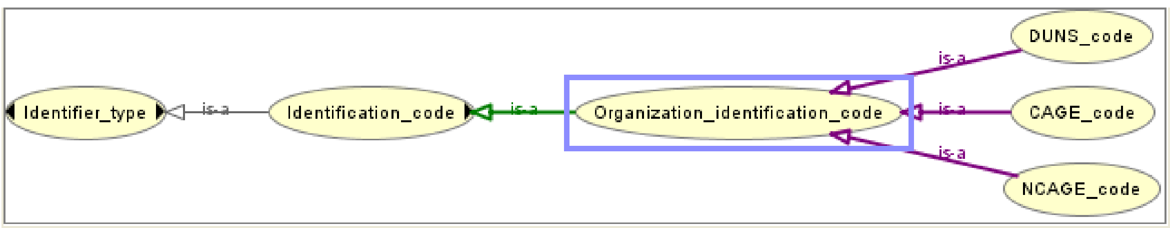
- Its scope covers the entire product lifecycle, including later stages such as maintenance and disposal. Therefore, PLCS can represent product metadata for a wide variety of access scenarios spanning the 3Rs (discussed in section “Product

Models and the “3Rs”: Reference, Reuse, and Rationale”).

- PLCS embraces implementation technologies popular with software developers such as XML and semantic technologies based on RDF.
- PLCS has a Reference Data Library (RDL), a managed collection of classes and individuals defining an extensible controlled vocabulary for representing business-specific concepts. The RDL is encoded in the the Web Ontology Language (OWL) [OWL]. Because the RDL is an OWL ontology, it can be browsed, queried, and processed using the many software tools developed for the semantic web.

Figure 4 shows the reference data class representing a CAGE code, its ancestors, and its siblings. PLCS models `CAGE_code` and its siblings (`DUNS_code` and `NCAGE_code`) as subclasses of `Organization_identification_code`, which is a subclass of `Identification_code`, which in turn is a subclass of `Identifier_type`. To keep the size of the example manageable, Figure 4 hides the numerous siblings of `Identification_code` and of `Identifier_type`. This example illustrates only a tiny subset of PLCS. The standard RDL contains hundreds of classes and covers a wide variety of PDM/PLM concepts such as assemblies, bills of material, work breakdown structures, maintenance procedures and more. And the RDL continues to grow as new PLCS data exchange specifications get standardized.

Figure 4



Portion of the PLCS RDL showing `CAGE_code`, its ancestors, and its siblings.

The PLCS RDL provides a comprehensive and fine-grained source of product metadata. But DI need only be sufficient for facilitating access to an archival information system. Therefore DI should be coarser-grained and less encompassing than PLCS. The good news is that, thanks to the RDL being encoded in OWL — a language built upon RDF — reference data subsets can easily be extracted and combined with DI represented using Dublin Core. The challenge is determining the subsets to extract. Analysis of access use cases similar to those discussed in section “Engineering-specific Metadata Requirements for Long-term Archiving” can guide the selection of the RDL subsets.

Summary and Moving Forward

In this paper, I explained why it is difficult to archive product data, discussing requirements and challenges unique to engineering. Next, I surveyed some of the more widely used metadata standards from the digital library community and argued that, although they could be useful in a product data archival system, they need augmentation in order to represent product data Descriptive Information. I then discussed two sources of product metadata — the options selection worksheet for Technical Data Packages and the PLCS Reference Data Library. While TDP metadata is of a top-down nature and maps well to METS, the PLCS RDL is extremely detailed and is more bottom-up. However, useful DI should be extractable from the RDL by considering potential long-term access scenarios. This RDL subset can then serve to enrich the more generic Dublin Core metadata terms standardized by Dublin Core Metadata Initiative. [DCMI]

The next step is to test the effectiveness of these ideas by developing a METS profile for TDPs and implementing a product data ingest and access prototype. The prototype should support one or more access scenarios, with the scenario's data requirements determining the subset to be extracted from the RDL. The METS descriptive metadata (`<dmdSec>`) section should contain RDF-encoded Dublin Core combined with the RDL subset, using semantic technologies to merge and harmonize the information. If successful, the METS profile and prototype implementation could benefit developers and implementers of product data archival standards.

Acknowledgments

I am grateful to Paul Witherell and Barbara Guttman for their meticulous and helpful reviews of an earlier draft of this paper. Any remaining mistakes are my sole responsibility. I also wish to thank my colleagues in the NIST Design and Process Group for many thought-provoking discussions about long-term archiving, information modeling, and engineering design.

References

[DCAP] K. Coyle and T. Baker. "Guidelines for Dublin Core Application Profiles." DCMI Recommended Resource. 2009-05-18. http://dublincore.org/documents/profile-guidelines.

[DCMI] DCMI Home: Dublin Core Metadata Initiative (DCMI). http://dublincore.org/.

[DCMI] Dublin Core Metadata Initiative. DCMI Metadata Terms. DCMI Recommendation. 2008-01-14. http://dublincore.org/documents/dcmi-terms.

[Gartner] "Metadata for digital libraries: state of the art and future directions." JISC Technology and Standards Watch. April 2008. http://www.jisc.ac.uk/techwatch.

[Guenther] R. Guenther, et al. "Guidelines for using PREMIS with METS for exchange." Revised September 17, 2008. http://www.loc.gov/standards/premis/guidelines-premismets.pdf.

[ISO10303-1] ISO 10303-1. Industrial automation systems and integration — Product data representation and exchange — Part 1: Overview and fundamental principles. First edition 1994-12-15.

[ISO10303-11] ISO 10303-11:2004: Industrial automation systems and integration — Product data representation and exchange — Part 11: Description methods: The EXPRESS language reference manual.

[ISO10303-21] 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.

[ISO10303-28] ISO 10303-28:2007. Industrial automation systems and integration — Product data representation and exchange — Part 28: Implementation methods: XML representations of EXPRESS schemas and data, using XML schemas.

[ISO14721] ISO 14721:2003. Space data and information transfer systems — Open archival information system — Reference model.

[Kassel]]B. Kassel, P. David. "Long Term Retention of Product Model Data." In Proceedings of the 2006 SNAME Maritime Technology Conference and Ship Production Symposium (Fort Lauderdale, Florida, October 10-13, 2006).

[LOTAR] LOng Term ARchiving - LOTAR International. http://www.lotar-international.org.

[Lubell] J. Lubell, S. Rachuri, M. Mani, E. Subrahmanian. "Sustaining Engineering Informatics: Toward Methods and Metrics in Digital Curation." International Journal of Digital Curation. Vol 3. No 2 (2008). http://www.ijdc.net/index.php/ijdc/article/view/87.

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

[MODS] Metadata Object Description Schema: MODS (Library of Congress). http://www.loc.gov/standards/mods/.

[OWL] World Wide Web Consortium. OWL 2 Web Ontology Language Document Overview. W3C Recommendation 27 October 2009. http://www.w3.org/TR/owl-overview

[PLCS] Product Life Cycle Support (PLCS). http://www.plcs-resources.org.

[PLM] PLM Technology Guide. http://plmtechnologyguide.com.

[PREMIS] PREMIS: Preservation Metadata Maintenance Activity (Library of Congress). http://www.loc.gov/standards/premis/.

[RDF] World Wide Web Consortium. RDF Primer. W3C Recommendation 10 February 2004. http://www.w3.org/TR/rdf-primer

[MID-STD-31000] U.S. Army. MIL-STD-31000. Defense Standard: Technical Data Packages. Draft. 8 September 2009.

[Srinivasan] V. Srinivasan. "STEP in the Context of Product Data Management." In "Advanced Design and Manufacturing Based on STEP." Xun Xu and Andrew Y.C. Nee eds. Springer London. DOI 10.1007/978-1-84882-739-4. 2009. Pages 353-381.

[Wikipedia1] Blueprint - Wikipedia, the free encyclopedia. http://en.wikipedia.org/wiki/Blueprint.

[Wikipedia2] Microform - Wikipedia, the free encyclopedia. http://en.wikipedia.org/wiki/Microform.

[XML] World Wide Web Consortium. Extensible Markup Language (XML) 1.0 (Fifth Edition). W3C Recommendation 26 November 2008. http://www.w3.org/TR/xml/.

[XSchema] World Wide Web Consortium. XML Schema Part 0: Primer Second Edition. W3C Recommendation 28 October 2004. http://www.w3.org/TR/xmlschema-0/

[1] The authors of STEP do not give an explicit definition for “formal.” However, based on their usage of the word throughout the standard, it seems “formal” is intended to mean unambiguous and adequate for specifying an implementation.

[2] 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.

[3] A CAGE (Commercial and Government Entity) code, as defined in MIL-STD-31000, is a five character identifier assigned to vendors doing business with the Government.

Balisage:* *The Markup Conference