VSMM 2008
Digital Heritage – Proceedings of the 14th International Conference on Virtual Systems and Multimedia
M. Ioannides, A. Addison, A. Georgopoulos, L. Kalisperis *Full Papers*

CURATION AND PRESERVATION OF CAD ENGINEERING MODELS IN PRODUCT LIFECYCLE MANAGEMENT

M. Patel [a], A. Ball[a], L. Ding[b],

[a]UKOLN, University of Bath, Bath, BA2 7AY UK -(m.patel, a.ball)@ukoln.ac.uk
[b]IMRC, Dept. Mechanical Engineering, University of Bath, Bath, BA2 7AY UK -ld218@bath.ac.uk

KEY WORDS: Curation, Preservation, Engineering, PLM, CAD Models, Multilayer Annotation, Representation Information

ABSTRACT:

During the last decade the management of a product's data over its entire life has been gaining prominence. This is largely due to two factors: firstly, an increasingly collaborative environment in which product development and maintenance takes place in a geographically distributed and networked environment. Secondly, there is an emerging economic paradigm shift in which companies that design and build products are increasingly entering into contracts to provide through-life support for them - that is, products are being purchased as services rather than artefacts. For engineering companies, this shift entails a commitment to supporting products over a much longer timeframe than previously expected. At the same time, there is an increasingly greater reliance on CAD models which are now being used as the method for recording definitive product data. However, the CAD software industry is characterised by ephemeral, backwardly incompatible, proprietary applications and file formats which readily become obsolete, making the long-term retention and accessibility of digital product models and data a challenge.
We examine the curation and preservation requirements in Product Lifecycle Management (PLM) and suggest ways of alleviating the problems associated with the sustained representation of CAD engineering models through the use of lightweight formats, layered annotation and the collection of *Representation Information* as defined in the Open Archival Information System (OAIS) Reference Model.

1. INTRODUCTION

The emergence of Product Lifecycle Management (PLM) as a business model over the last decade or so can be attributed to the disruptive effects of ICT which has resulted in a global and networked market place and consequent international collaboration and business practices. This business model is applicable in the engineering, manufacturing, contracting and service sectors amongst others. PLM requires the efficient capture, representation, organisation, retrieval and reuse of product data over its entire life [McMahon et al., 2005].

The importance of managing a product's data over its entire lifecycle is gaining importance, mainly due to two factors. Firstly, many companies now operate in an increasingly collaborative environment, one in which product development and maintenance occur in a geographically distributed and networked environment, with the result that much of the data relating to a particular product or artefact is dispersed over a number of organisations and locations. Secondly, there is an emerging economic paradigm shift such that companies that design and manufacture products are increasingly entering into contracts to provide through-life support for them – that is products are now being sold and purchased as services rather than artefacts. For example, within the aerospace industry, Rolls-Royce has introduced the concept of "power by the hour". For products such as cruise ships, aircraft, rolling stock for railways, hospitals and schools, this could mean a commitment to providing support for as long as the product is in service, extending to 30-50 years or in some cases even longer.

At the same time, there is a much greater reliance on CAD models which are now being used as the main carriers for recording definitive product data as opposed to paper based technical drawings and documentation. Within the last five years or so, the engineering industry has moved over to using CAD models directly for communicating designs, not only to manufacturers and builders, but also to regulating authorities and maintenance crews. However, the switch to recording information digitally presents its own problems, not only in terms of long-term maintenance and accessibility, but also as a potential threat to the recording of the evolution of design, artefacts and products in terms of our industrial, automotive and avionic history and heritage.

The remainder of this paper examines curation and preservation issues in the context of PLM and suggests ways of alleviating the problems associated with the sustained representation of CAD engineering models through the use of lightweight formats, layered annotation and the collection of *Representation Information* as defined in the Open Archival Information System (OAIS) Reference Model [CCSDS, 2003].

2. DIGITAL CURATION AND PLM

Digital Curation

The term *digital curation* is now generally accepted as including the active management of digital data over their useful lifetime, both for contemporary and future use, as well as incorporating archiving and digital preservation. The term also encompasses the notion of adding value to a trusted body of digital information as well as its reuse in the derivation of new information and the validation and reproducibility of results [Beagrie, 2006; DCC, 2007].
The urgency for assuming widespread digital curation activities stem from several issues all of which relate to the proliferation of digital information and the heavy risks associated with its potential loss. A study, conducted by UC Berkeley estimates that the world produces between one and two Exabyte of unique information every year [University of California, Berkeley, 2000]. A recent follow on report forecasts that the "digital universe" will explode to an incredible 988 Exabyte by the year 2010 [IDC White Paper, 2008]. Additionally, legislative and regulatory requirements imposed on certain industries such as

pharmaceuticals and engineering mean that they are required to maintain data and records for considerable periods of time.

The current situation is such that while the cost and investment in digital information creation is huge, the benefits are likely to be short-lived and the threat of the "Digital Dark Ages" [Kuny, 2007] will remain omnipresent unless digital information and data is curated and preserved adequately. Due to technological obsolescence (hardware, software and file formats), which constitutes one of the major threats to digital information, data can become inaccessible within a very short time. Moreover, much digital information requires software applications in order to make is accessible to humans.

A variety of techniques have been proposed and explored to combat the effects of rapidly changing technologies and media degradation: bit-stream copying, refreshing, the use of durable media, digital archaeology and replication. Strategies aimed at preserving access to the information content and providing functional preservation include: technology preservation, analogue backups, migration, normalization, emulation and encapsulation. A particular strategy concerned with mitigating the effects of technology evolution is based on the use of *Representation Information* (RI) – a concept used in the Reference Model for an Open Archival Information System (OAIS) [CCSDS, 2003]. RI is all-encompassing; it is essentially any information that is required to render, process, visualize and interpret data and includes: file formats, software, algorithms and standards as well as semantic information.

The Open Archival Information System (OAIS)

The OAIS Reference Model establishes a common framework of terms and concepts for use in the preservation of information. An archival information system consists of an organisation of people and systems, which has accepted the responsibility to preserve information and make it available for a *Designated Community*. The latter being an identified group of potential stakeholders and users. The Model is set in the context of producers (who generate information to be archived), consumers (who retrieve that information) and management (the wider organisation responsible for maintaining the OAIS).

The model has achieved widespread adoption, influencing: the development of preservation planning [Strodl et al, 2007]; preservation metadata [PREMIS, 2008]; architectures and systems of repositories [Giaretta, 2007]; and conformance and certification criteria for archives [TRAC, 2007]. Of particular note is OAIS PDI or *Preservation Description Information*, comprising several types of metadata to help ensure the quality of the data and its fitness for purpose:

- *Reference*: One or more mechanisms used to provide identifiers for unambiguous access to content e.g. object identifier or a persistent identifier.
- *Provenance*: Documents the history of the content information, to provide some assurance as to its likely reliability.
- *Context*: Documents the relationships of the content information to its environment and other content information e.g. calibration history; relationship to other data; or pointers to related documents.
- *Fixity*: Provides data integrity checks including validation and verification keys used to ensure authenticity e.g. encoding and error detection schemes such as checksums.

The OAIS is in effect a comprehensive reference model for the development and operation of a preservation and curation environment for all types of data.

A Digital Curation Lifecycle Model: Digital curation is a multi-faceted and complex process involving social, political, organisational and financial as well as technical issues. In order to clarify these numerous aspects and the relationships between them, the Digital Curation Centre (DCC) has developed a Curation Lifecycle Model [DCC Curation Lifecycle Model, 2008], see Figure 1.

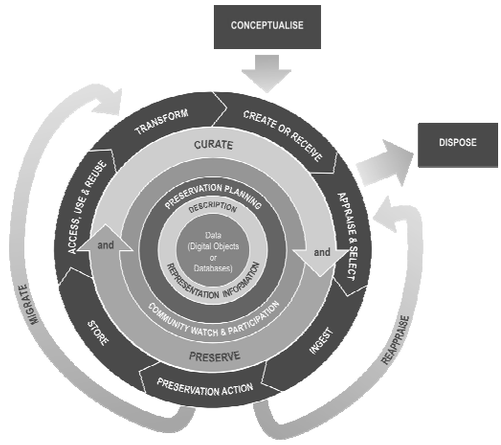

Figure 1: DCC Curation Lifecycle Model © DCC 2008

The Model provides a graphical high level overview of the stages required for successful curation and preservation of data from initial conceptualisation or receipt. It can be used to plan activities within an organisation or consortium and enables granular functionality to be mapped against itself and particular information workflows. The DCC Curation Lifecycle Model defines digital objects, as well as databases and splits the processes into those that are:

- Full lifecycle stages (Description and Representation Information; Preservation Planning; Community Watch and Participation; Curate and Preserve)
- Sequential actions (Conceptualise; Create or Receive; Appraise and Select; Ingest; Preservation Action; Store; Access, Use and Reuse; Transform)
- Occasional actions (Dispose; Reappraise; Migrate)

Product Lifecycle Management

The scope of PLM is extensive and includes a number of phases: Conceptualisation (innovation, requirements); Design Organisation (people, infrastructure, knowledge); Design (product, process); Evaluation (analysis, simulation, performance, quality); Manufacture and Delivery (production, supply, delivery); Sales and Distribution (advertising, marketing); Service and Support (maintenance, upgrades, warranties); Decommissioning (retirement, recycle, disposal). The Knowledge and Information Management through Life Project (KIM) is currently investigating the implications of PLM and the paradigm shift to a product-service approach [Ball et al., 2006]. A scenario has been devised by the Project to illustrate ideal information flows in a product's lifecycle, Figure 2 shows a summary. In all cases the flow of information must be managed to ensure that appropriate information is

transmitted and understood at the different stages. A major challenge is that data needs to be shared and exchanged between multiple organisations involved in the lifecycle of the product which can extend to considerable lengths of time.

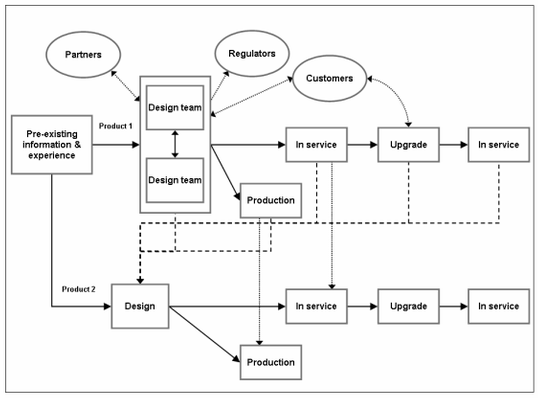

Figure 2: Information flows in the lifecycle of a product

Up until the turn of the millennium, engineering software was used to support a paper-based workflow. CAD packages were used to create virtual models of designs, from which drawings and other design documentation could be produced. The manufacture or construction process was based on the resulting documentation. However, current digital environments necessitate an electronic flow of information between heterogeneous systems for Computer Aided Design (CAD), Computer Aided Engineering (CAE) and Computer Aided Manufacture (CAM) as well as Enterprise Resource Planning (ERP), Customer Relationship Management (CRM) and Supply Chain Management (SCM). Additionally, users in the differing stages of a product's lifecycle require different information and representations - i.e. there are multiple viewpoints on the product models [Ding et al, 2006]. For example, machining features are useful for manufacturing engineers, but not for marketing staff, for whom a visualisation of a product, unencumbered with production and manufacturing information, is of far greater use.

Curatorial issues in PLM

The active management of all product data is vital to PLM as data is created, added to, modified and extracted over the course of the lifecycle of a product. It is apparent that elements of the DCC Curation Lifecycle Model (Figure 1) could be incorporated into the information flows at various stages in PLM (Figure 2). However, although current PLM systems such as Agile, IBM/Dassault, MatrixOne, PTC and UGS PLM, cater for some aspects of digital curation, they do not place an emphasis on issues relating to preservation. Additionally, the scope of PLM is wide-ranging, involving a large volume and variety of data, information and knowledge, all of which needs to be managed and maintained. This can range from highly structured data (such as geometric models and databases) to unstructured textual documents (e.g. email) to the tacit knowledge held by employees (e.g. design rationale and lessons learned from experience).

Whilst digital product information may share many issues in common with other types of data, such as text documents and scientific datasets (e.g. issues relating to bit-preservation such as information security, integrity, authenticity and longevity of digital storage media) there are several aspects which are specific to PLM and that make the digital curation of engineering product data particularly rife with problems.

Within PLM there is also a requirement to support global and distributed collaboration. However, product data tends to be amongst the most valuable intellectual property of a company, which will therefore only be prepared to share selective information depending on the role of the collaborating partner.

In addition, not all stages of PLM require all product data and it is useful to extract a simplified view of the product for use in later stages. Here the notion of *Significant Properties* comes to the fore. Significant properties are those aspects of the digital object which must be preserved over time in order for it to remain accessible and meaningful [Significant Properties Workshop, 2008]. In PLM the significant properties of a product may vary depending on the view and the stage in the lifecycle.

Due to the complexity of many contemporary products, the volume of data generated during development tends to be huge and distributed, so that it becomes difficult to make decisions regarding which data and how much should be kept and maintained.

Issues relating to technological obsolescence are exacerbated in PLM, mainly due to the state of the CAD software industry and the move to the product-service paradigm. Complex dependencies and relationships exist between file formats, software and hardware. In addition, the CAD software market is competitive and characterised by a proliferation of CAD formats which are proprietary, closed and subject to frequent change. Interoperability between such systems is virtually non-existent – indeed many CAD tools do not even maintain reliable backwards compatibility with their own versions.

Solutions such as emulation and migration pose problems in the engineering domain. Emulating old software incurs difficulties with integrating it into complex and more modern workflows and systems. A major issue with migrating old designs to newer formats is that there is always the risk of data loss and subtle design corruption. Also, the cost of re-checking and re-validating a design after migration can be substantial. Even the use of open and neutral standards, such as IGES (Initial Graphics Exchange Standard) [IGES, 1996] and STEP (Standard for Exchange Product Data) [STEP, 2005] is not without issues. The rigours and long timescales of developing a comprehensive exchange standard for CAD models, means that it is difficult to keep up to date with the latest capabilities of CAD tools. Furthermore, the level of support for such standards can be variable between tools. As a consequence, data created using a particular application is in danger of becoming inaccessible once that software is retired or replaced. Furthermore, in such a complex and dynamic environment, it becomes extremely difficult to reliably retrieve and trace provenance information to check the veracity and reliability of data. To facilitate the development of new generations of a product, especially in the face of greater awareness of environmental impact and efficiency, it is necessary to cater for long-term retention and preservation so that older designs can be reused and adapted or customised.

3. CHALLENGES FOR CAD ENGINEERING MODEL REPRESENTATIONS

CAD models have traditionally been used in the design, evaluation and manufacturing phases of PLM and product information is still largely stored in CAD models based on conventional product representations, including boundary representation (B-rep), freeform surface modelling, feature-based models or parametric models. If CAD models are to become the main carriers of extra product lifecycle information, it is clear that they need to be extended, augmented and supported in additional ways.

PLM makes several demands that need to be taken into account:

- Protection of commercially sensitive information (Intellectual Property)
- Generation of view-point specific representations to support differing processes
- Rapid sharing of information between geographically distributed applications and users (interoperability, platform and application independence, use of standards, reduced file sizes etc.)
- Support for recording feedback from downstream processes
- Long-term preservation (recording of metadata, design rationale, open formats, RI, use of standards etc.)

4. A STRATEGY TO SUPPORT CURATION IN PLM

To extend a CAD model from purely the design stage into the whole product lifecycle, a framework of lightweight representations is proposed together with a method for annotation and the use of a Registry/Repository of Representation Information (RRoRI) to support decision making.

In the proposed strategy, all users throughout a product's lifecycle can annotate the CAD model according to their specific requirements and experiences. The information is stored in a series of separate XML-based files, each of which is linked to the CAD model through a specific element (e.g. a face) using a mechanism of references. With the support of these markup files, the CAD model can be compressed into various lightweight representations according to different levels of security and viewpoints. Additionally, RI relevant to the CAD model and lightweight formats, as well as the XML schemata for markup documents, is stored in a RRoRI to aid the interpretation of the accumulated data in the longer term. Issues relating to the collection and use of engineering RI are explored in Patel & Ball [Patel & Ball, 2007].

Lightweight CAD Model Representations

Full CAD formats tend to be large, complex and proprietary, and are rarely backwards compatible. This makes them unsuitable both for long-term archiving, reference and reuse, and for distributed collaborative design work. Lightweight formats provide a potential solution. They comprise simple formats that are easier to preserve but which do not try to retain all the richness of the full CAD model. By producing files in these formats at the time of the original design, they can be validated at the same time as the full model. Their simplicity makes them easier to read back into newer software. They have smaller file sizes, simpler and more open specifications, and more affordable software support. Also, they need only contain as much information as a particular recipient needs (this is analogous to the notion of *desiccated formats* [Kunze, 2005] which retain only essential information).

There are a number of different lightweight representations in current use, each with properties and characteristics better suited to some purposes than others. In this section we introduce a number of these formats, with a particular regard to their capabilities with respect to: fidelity to the full model, metadata storage, data security, file size reduction, support for the format by software and openness.

3D XML: 3D XML [Versprille, 2005; Dassault Systèmes, 2007] is an XML-based format for describing a model's geometry, structure and visualization, and is optimized for interactivity and compactness. It can represent geometry using compact NURBS-like surface descriptions, XML polygon meshes and compact syntax polygon meshes, but does not have any additional security features. File sizes are kept down by a reference-instance mechanism (allowing the same data to be re-used several times within a model), a modification mechanism (allowing an instance or reference object to build on the properties of another reference object) and raster graphic compression. Models may be expressed by a single file or split across several` files. Native support for the format is largely restricted to Dassault Systèmes products, although free plug-ins are available for Lotus Notes and Microsoft Word, PowerPoint and Internet Explorer, as well as a free standalone viewer. The format is owned and controlled by Dassault Systèmes; the specification for the format is available cost-free to those who register.

JT Format: JT Format [UGS, 2006] is a binary format for encoding product geometry using boundary representations and wireframes, and supports additional product manufacturing information and other metadata. It does not have any built-in security features other than approximating data using tessellating polygons. File sizes are kept down using a reference-instance mechanism, zlib compression of various data elements and datatype-specific compression using algorithms such as uniform data quantization, bit length codec, Huffman codec, arithmetic codec, and Deering Normal codec. Models may be expressed by a single file or split across several files. Native support for the format is largely restricted to UGS products, although free plug-ins are available for Microsoft Word, Excel and PowerPoint, as well as a free standalone viewer. The format is owned by UGS, but the specification is freely accessible on the Web and blanket permission is given to implement it.

PLM XML: PLM XML [UGS, 2005] is a set of XML schemata for describing a model's geometry, structure, features, ownership, and visualization. It is designed to be interoperable between a number of different tools from across the lifecycle of a product. The native schemata for representing geometry can support 2D and 3D vector graphics, NURBS surfaces and features, although non-native representations can also be used or referenced in a PLM XML document. It also allows for a single logical product model to have several different geometric representations, tailored to different purposes. Metadata of several different types – mass, material, texture, product manufacturing information, dimensions and tolerances, user markup, application-specific data – can be attached to logical parts of the model or specific geometric representations. File sizes can be reduced using a reference-instance mechanism and by splitting out various sections of data into separate files (so that data not needed for a particular purpose need not be transmitted). As well as approximating and sub-setting data,

PLM XML also supports mechanisms for restricting access to parts of the model data on the basis of person, organization or place. The format is used extensively by UGS products but is not widely supported otherwise. The format is owned and controlled by UGS; the XML schemata are freely accessible on the Web, but the software development kit must be purchased.

PRC: PRC [Adobe Systems, 2007b] is a binary format that promises to encode the full range of CAD geometry, along with model trees, history trees and various forms of markup. Alternative geometries (e.g. exact and tessellated) can be provided for each part; markup can be associated with entire parts or tessellations but not with items of exact geometry. Arbitrary non-PRC data can be included at various points, notably at the end of entity code. Summary data sections enable files to be accessed without being fully parsed, but the format is not suitable for streaming. File sizes are reduced through a number of mechanisms: compact mathematical encoding of geometry, a reference-instance mechanism, and gzip encoding of data sections (header sections remain uncompressed). The precision of the geometry may be reduced to provide lossy compression. Proprietary converters are available for a wide range of CAD formats, and PRC is supported as a native 3D model format within the Portable Document Format (PDF) specification from version 1.7 (corresponding to Adobe Acrobat 8.1), which adds some conservative security measures on top of the otherwise unprotected format [Adobe Systems 2007a]. The format was initially proprietary but is expected to form part of ISO 32000.

Universal 3D (U3D): Universal 3D [ECMA-363, 2007] is a binary format for encoding product geometry using sets of tessellating triangles and (from the 4th edition) NURBS surfaces. A mesh update mechanism allows meshes to be rendered progressively, providing basic streaming support. Metadata, stored as key/value pairs, may be attached to any node in the model tree. It does not have any in-built security features other than approximating the geometry. File sizes are kept down using a reference-instance mechanism and a bit compression algorithm on numeric data fields. The format is most notably supported as a native 3D model format within the PDF specification, with the 1st edition of U3D supported from version 1.6 (Adobe Acrobat 7) and the 3rd edition supported from version 1.7 (Adobe Acrobat 8.1). PDF also adds some conservative security mechanisms of its own. It was developed by the 3D Industry Forum and is published and maintained as ECMA standard 363; the specification is freely available on the Web.

X3D: X3D (ISO/IEC 19775 2004; ISO/IEC 19776 2005; ISO/IEC 19777 2006) is an improved version of Virtual Reality Markup Language (VRML); it is an XML format optimized for animation and interaction. It can represent 2D and 3D vector graphics, 3D tessellating polygon meshes, and NURBS surfaces as well as identifying bones and joints for human animation. Any node in the model tree may have metadata attached, in a format specifying a value (string or number), a metadata schema and a key. It does not have any in-built security features other than data approximation. X3D has a reference-instance mechanism and a relatively compact XML syntax, with coordinates expressed as space/comma delimited lists within attributes, rather than through a hierarchy of tags; a binary syntax is available that compresses field values according to Fast InfoSet principles, using zlib compression, quantization of floating point number arrays, integer range reduction and conversion of absolute values to relative values. Open source libraries and viewers are available for processing and rendering X3D files. X3D was developed by the Web 3D Consortium, and is published and maintained as ISO standards 19775, 19776 and 19777; these standards are freely available on the Web.

XGL/ZGL: XGL [XGL Working Group, 2006] is an XML-based encoding of the Open Graphics Library (OpenGL) application programming interface for rendering 2D and 3D computer graphics. When compressed it is known as ZGL. It uses tessellating triangles to encode geometry, and is optimized for display. It does not have any capabilities for storing metadata, nor does it have any in-built security features other than approximating the geometry. File sizes are kept small using a reference-instance mechanism and a relatively compact XML syntax, with vector coordinates expressed as comma delimited lists rather than through a hierarchy of tags. XGL is supported by Autodesk and a few smaller CAD vendors. It was developed by the XGL Working Group but no longer appears to be maintained; the specification of the format was once freely available on the Web, but now only appears in 'unofficial' locations.

Multilayered Annotation (LiMMA)

Although lightweight representations alleviate many of the challenges outlined in section 3 (collaborative exchange, customised views, security, application independence, preservation, reduced file sizes etc.) there is still an out-standing requirement – that of being able to augment the geometric model of the product with additional information from different phases of PLM. For this purpose we propose the use of annotation which allows the incorporation of varied information:

- Design rationale, context, provenance, RI etc.
- Extra information needed for a certain point of view
- Embedding of commercial security levels to restrict access to certain partners or users

Use of a machine processible language, such as the Extensible Markup Language (XML) [XML 2006], is particularly useful for collaborative ventures over the Web.

There are two methods for applying annotation, 'inline' and 'stand-off'. Inline annotation involves adding information directly into the text of a document or model, whereas the stand-off (external or reference) method allows markup information to be stored separately, being linked back to the document or model using references or pointers [TEI Standoff Markup Working Group, 2003; Thompson & McKelvie, 1997]. The latter is the more appropriate for use with CAD models for several reasons:

- It allows a 3D geometric representation of a product to be progressively expanded to include additional metadata without changing the representation method used for the geometry of the product.
- The CAD model itself need not contain all the information required for every user and purpose: context-specific information can be extracted into a number of separate files to provide multi-layered annotation which can be passed around as required, allowing the CAD model to remain smaller in size.
- It allows the same annotation to be applied to different representations of the same model, granting the annotation information some independence of the CAD format used.
- It enables downstream processes (e.g. finite element analysis and manufacturing processes) to be independent of the CAD model through the reuse of annotation.

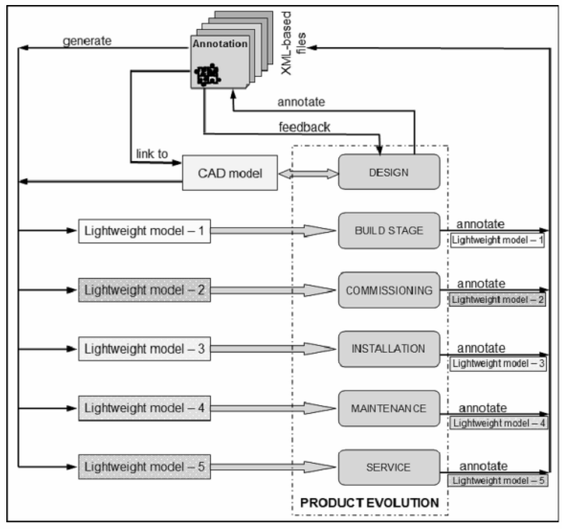

Figure 3: A Framework for the annotation of CAD models

LiMMA (Lightweight Models with Multilayered Annotations), is a framework for representing CAD models using lightweight geometric models with additional layers of XML-encoded information, as shown in Figure 3. To date, LiMMA plug-ins have been written in C/C++ and NX Open for UGS's NX CAD package, as well as in JavaScript for the 3D PDF viewer Adobe Acrobat Reader. In addition, a standalone X3D viewer has been written in Java as another component in LiMMA. Annotations are currently linked to the geometric models by means of unique identifiers attached to the entities that comprise the model. However, an alternative system of referencing making use of co-ordinate sets is also under development. Figure 4 shows the annotation environment which has been developed within the NX package.

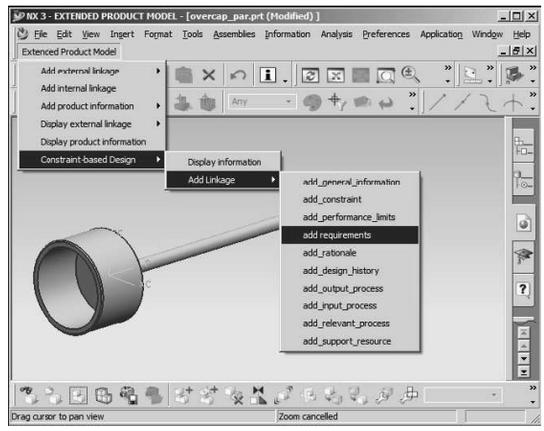

Figure 4: Interface of the internal NX markup environment

Representation Information (RRoRIfE)

Although the use of lightweight formats and multi-layer annotation seemingly address all of the challenges outlined in section 3, there is a further issue with the proposed strategy in that a wide range of lightweight formats with differing characteristics are available. The problem therefore lies in selecting a format which is most appropriate not only for a particular use and view of the product, but also for long-term retention.

For this purpose we have developed the Registry/Repository of Representation Information for Engineering (RRoRIfE), a simple preservation planning tool based on RI relating to various characteristics of file formats and their associated conversion software. The premise behind the tool is that an intellectual object can only be faithfully reproduced in a new format or environment if the latter supports properties or characteristics equivalent to those used by the intellectual object in its native format or environment. Furthermore, different tools may be better or worse at re-expressing the constructs of the old format or environment in the constructs of the new.

Underlying RRoRIfE is an ontology of properties, characteristics and constructs of engineering information; since RRoRIfE is at present focused on CAD models, the current ontology includes various two dimensional and three dimensional geometric entities, as well as different compression techniques and forms of metadata. This ontology was derived from a superset of the properties supported by a sample of CAD formats. Two XML schemas have been written using the ontology. The first relates to file formats and describes whether or not the format supports a particular property. An intermediate value of 'partial' support is allowed, to indicate that support is limited in some way; for example, NURBS surfaces may be allowed, but only with 256 or fewer control points. In cases of partial support, explanatory text must be provided.

The second XML schema relates to processing software. For each format conversion the software is able to perform, a record is created as to how well the conversion preserves each property. Four levels of preservation are allowed: 'none' indicating that the property has never knowingly survived the conversion intact (perhaps because the destination format does not support the property); 'good' indicates that the conversion has so far preserved examples of the property sufficiently well that it would be possible to reconstruct the original expression of the property from the new expression; 'poor' is used when tests have found it at least as likely for the property to be corrupted or lost as it is to survive; while 'fair' is used otherwise, alongside an explanatory note.

Figure 5: User Interface to RRoRIfE

Where preservation is less than 'good', it is possible to record whether the property survives in a degraded form, and if so, whether this degradation always happens in a fixed way, a configurable way or an unpredictable way. For example, when moving from a format that supports NURBS to one that only supports tessellating triangles, there may be a fixed algorithm

for approximating surfaces, or one may be able to specify how detailed the approximation is.

RRoRIfE reads files in these two XML formats, and uses them to answer simple preservation planning queries. As well as being able to look up the characteristics of individual formats and conversion utilities, RRoRIfE also allows one to select certain characteristics as significant and discover which formats support them. It can generate possible migration pathways between two formats, and given a starting format and set of significant characteristics, it can generate a list of suitable destination formats and conversion pathways. Figure 5 shows the GUI to RRoRIfE.

5. FURTHER WORK

There are several aspects to continuing the work described in this paper, not least that of developing an integral framework for LiMMA and RRoRIfe. One issue currently receiving much attention concerns the persistent identification of geometry between translations from native CAD models into lightweight formats, such that product information can be reliably associated with the same entities in both models. This is also known as the "persistent naming problem" [Marcheix & Pierra, 2002; Mun & Han, 2005], and is not peculiar solely to engineering data. Buneman et al. discuss an analogous problem in the context of curating databases [Buneman et al, 2008].

With respect to RRoRIfe, it is important to accumulate a corpus of RI to enable informed decision making both in terms of the characteristics of conversion software as well as investigating the significant properties of various formats.

Wider consideration of digital curation in PLM includes non-intrusive and automated capture of information as well as issues relating to the selection and appraisal of product lifecycle data.

6. CONCLUSIONS

We have examined the digital curation challenges posed by PLM in engineering and suggested several techniques to improve the robustness of product data to serve the needs of both PLM and long-term accessibility.

Full CAD models tend to have closed, proprietary formats and are difficult to pass around between organisations and the stages in PLM. Lightweight formats provide a more promising approach in that: they have open specifications; they are simpler and have smaller file sizes; they can cater to the need for multiple viewpoints; and restrict access for security purposes. In addition, multilayered annotation allows models to be augmented with much valuable data including that required for preservation. Finally, accumulation of RI facilitates informed decision making with respect to which lightweight formats and conversion software to use both for data exchange within PLM and for accessibility in the longer run.

7. REFERENCES:

Adobe Systems, 2007a. PDF Reference and Related Documentation. Adobe Acrobat SDK version 8.1, http://www.adobe.com/devnet/pdf/pdfs/PDFReference16.pdf (accessed 14th June 2008).

Adobe Systems, 2007b. PRC Format. Version 7094, http://www.adobe.com/devnet/acrobat/pdfs/pdf_reference.pdf (accessed 14th June 2008).

Atlantic Workshop on Long-Term Knowledge Retention (LTKR), 2007. Dept. Mechanical Engineering, University of Bath, UK, 12-13th February 2007.

Ball A., Patel M., McMahon C., Culley S., Green S., 2006. A Grand Challenge: Immortal Information and Through-Life Knowledge Management (KIM), IJDC 2006, Vol. 1, No. 1.

Beagrie N., 2006. Digital Curation for Science, Digital Libraries, and Individuals, IJDC 2006, Vol. 1, No. 1.

Buneman P., Cheney J., Tan W-C., Vansummeren S., 2008. Curated Databases, PODS'08, June 9-12th 2008, Vancouver, BC, Canada

CCSDS 2003. Reference Model for an Open Archival Information System (OAIS). CCSDS 650.0-B-1. Blue Book (ISO 14721:2003).

Dassault Systèmes, 2007. 3D XML User's Guide. 0th ed. Version 4.0.

Digital Curation Centre (DCC), 2007. What is Digital Curation? http://www.dcc.ac.uk/about/what/ (accessed 10th June 2008).

DCC Curation Lifecycle Model, 2008. http://www.dcc.ac.uk/docs/publications/DCCLifecycle.pdf (accessed 10th June 2008).

Ding L., Li W.D., McMahon C.A., Patel M., 2006. Lightweight Representations for Product Lifecycle Management, Virtual Concept 26th November – 1st December 2006, Mexico.

ECMA-363, 2007. Universal 3D File Format. 4th Edition.

Extensible Markup Language (XML) 1.1 (Second Edition), 2006. W3C Recommendation. World Wide Web Consortium.

Giaretta, D., 2007. The CASPAR Approach to Digital Preservation, International Journal of Digital Curation, Vol 2 (1) 2007.

IDC white paper, 2008. The Expanding Digital Universe: A Forecast of Worldwide Information Growth Through 2010. Sponsored by EMC, http://www.emc.com/about/destination/digital_universe/ (accessed 14th August 2008).

Initial Graphics Exchange Specification (IGES) v5.3, 1996. ANSI, from US Product Data Association (USPRO), http://www.uspro.org/documents/IGES5-3_forDownload.pdf (accessed 14th August 2008).

ISO 10303 STEP -Standard for the Exchange of Product Model Data, 2005.

Kuny T., 2007. A Digital Dark Ages? Challenges in the Preservation of Electronic Information, 63rd IFLA Council and General Conference, http://www.ifla.org/IV/ifla63/63kuny1.pdf (accessed 14th June 2008).

Kunze, J., 2005. Future-Proofing the Web: What We Can Do Today, 16th September 2005, International Conference on Preservation of Digital Objects, Göttingen, Germany.

Marcheix D. and Pierra G., 2002. A survey of the persistent naming problem, Proceedings of the Seventh ACM Symposium

on Solid Modeling and Applications. Saarbrücken, Germany. ACM : New York. pp13–22. ISBN: 978-1-58113-506-0. DOI: 10.1145/566282.566288.

McMahon C., Giess M. and Cully S., 2005. Information management for through life product support: the curation of digital engineering data, IJPLM, Vol. 1, No. 1, pp26-42.

Mun D. and Han S., 2005. Identification of topological entities and naming mapping for parametric CAD model exchanges. International Journal of CAD/CAM Vol. 5, No. 1, 2005.

Open HSF Initiative, 2008. The HOOPS 3D Product Suite, http://www.openhsf.org/docs_hsf/index.html (accessed 14th June 2008).

Patel M. & Ball A., 2007. Challenges and issues relating to the use of Representation Information for the digital curation of Crystallography and Engineering Data, 3rd International Digital Curation Conference, 12-13th December 2007, Washington, D.C.

PREMIS Data Dictionary for Preservation Metadata version 2.0, 2008. Preservation Metadata Maintenance Activity, http://www.loc.gov/standards/premis/ (accessed 14th June 2008).

Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC), Version 1.0, 2007. Center for Research Libraries and RLG Programs.

Significant Properties of Digital Objects, 2008. JISC/BL/DPC Workshop, April 2008

Strodl S., Becker C., Neumayer R., Rauber A., 2007. How to Choose a Digital Preservation Strategy: Evaluating a Preservation Planning Procedure, JCDL'07*,* 17–22 June, 2007, Vancouver, British Columbia, Canada.

TEI Standoff Markup Working Group, 2003. Stand-off markup. Working paper SOW 06, Text Encoding Initiative.

Thompson, H. S. and McKelvie, D., 1997. Hyperlink semantics for standoff markup of read only documents: The next decade – pushing the envelope, Proceedings of SGML Europe '97. Barcelona. Graphic Communications Association, pp227–229.

University of California, Berkeley, 2000. How much Information, http://www2.sims.berkeley.edu/research/projects/how-much-info/summary.html (accessed 14th August 2008).

UGS, 2005. Open product lifecycle data sharing using XML, http://www.ugs.com/products/open/plmxml/docs/wp_plm_xml_14.pdf (accessed 14th June 2008).

UGS, 2006. JT File Format Reference. Version 8.1, http://www.jtopen.com/docs/JT_File_Format_Reference.pdf (accessed 14th June 2008).

Versprille K., 2005. Dassault Systèmes' strategic initiative: 3D XML for sharing product information. Technology Trends in PLM. Collaborative Product Development Associates.

XGL Working Group, 2006. XGL File Format Specification, http://web.archive.org/web/20060218/http://www.xglspec.org/ (accessed 14th June 2008).

8. ACKNOWLEDGEMENTS

The Digital Curation Centre is funded by the UK's Joint Information Systems Committee (JISC). The KIM Project is supported by the UK Engineering and Physical Sciences Research Council (EPSRC) and the Economic and Social Research Council (ESRC) under Grant Numbers EP/C534220/1 and RES-331-27-0006.