KNOWLEDGE & INFORMATION MANAGEMENT THROUGH LIFE

Proceedings of the Atlantic Workshop on Long Term Knowledge Retention 2007

Alex Ball and Lian Ding

kim40mee002ab10.pdf

Access Level: 1

Issue Date: 13 April 2007

Approved by: Chris McMahon

Date Approved: 13 April 2007

UKOLN

University of Bath

DCC

Digital Preservation

CONTENTS

1 INTRODUCTION

This report summarizes the presentations and discussions of the Atlantic Workshop on Long Term Knowledge Retention held at the University of Bath on the 12th–13th February 2007. Twenty-five participants from academia and industry on both sides of the Atlantic contributed to the workshop, with the purpose of sharing knowledge and experience in the fields of engineering-related representations, data trustworthiness, digital repository interoperability and product engineering informatics.

In almost every area of human endeavour, the rate at which digital information generated is far exceeding the rate of consumption. Engineering is no exception to this trend where megabytes of design, manufacturing and production data are generated and modified everyday. The importance of having effective systems for archiving and maintaining digital information has been recognised on both sides of the Atlantic.

In the US, for example, the National Digital Information Infrastructure and Preservation Program (http://www.digitalpreservation.gov/), led by the Library of Congress, has brought together a wide range of researchers from many different disciplines. Similar efforts are underway in the UK with the establishment of the Digital Curation Centre (http://www.dcc.ac.uk/), which aims to support and promote continuing improvement in the quality of data curation and of associated digital preservation.

The specific problems of maintaining industrial data has been brought into sharp focus by the paradigm shift many engineering companies (both in the UK and US) are undergoing from product delivery to through-life service support. Firms are increasingly required to supply products and to provide support services throughout the product lifetime. This requires new business, operational and information system models that extend thirty years or more into the future. In the UK this need has been recognised by KIM (http://www.kimproject.org/), a £5 million research project that brings together researchers from eleven universities.

This report presents the proceedings of the workshop in roughly chronological order. Section 2 contains summaries of the UK keynote presentations that began Day 1 of the workshop. Section 3 contains summaries of the presentations and discussions from the first panel session, which dealt with challenges in engineering informatics. Section 4 contains summaries of the presentations and discussion from the second panel session, which dealt with digital archiving models, representation languages and standards. Section 5 contains summaries of the two breakout sessions, each specializing in the topic of one the panel sessions; some of the plenary discussions from the feedback sessions on Day 2 are also presented here. Section 6 contains a summary of the US keynote presentation that began Day 2 of the workshop. Sections 7 and 8 contain the concluding discussions of the workshop.

2 UK KEYNOTE PRESENTATIONS

2.1 Long Term Knowledge Retention

Chris McMahon, University of Bath, KIM Project Director

Chris started his talk with two quotations. The first, from T. D. Wilson [1], suggested that knowledge management as it is currently practised is essentially information management coupled with good communication policies and practices. The second, from a WHO Bulletin [2], suggested that medical practice today is not sufficiently firmly grounded in evidence. Chris suggested that this applied also in engineering, and that what was needed is the systematic collection and management of information.

The representations that need to be managed in engineering are diverse — models, drawings, finite element analyses — and they are mostly produced with proprietary

software and data, making for a complex curatorial and intellectual property rights (IPR) landscape. Process, rationale, requirements, etc. are mainly recorded in text documents, limiting the possibilities for their retrieval, analysis and reuse.

The work that has been done with computers so far has often been lost due to data corruption, media degradation, and/or hardware and software obsolescence. Current practice is based on good information management and costly format migrations. There is some hope, though. The number of proprietary formats of significance for engineering is reducing as the market consolidates on a few key vendors. The sophistication of translators is improving, and there are various standard exchange formats and lightweight representations, though some work better than others.

There is a persistent problem of too much information coupled with poor organization. Of particular concern is the huge amount of uncontrolled emails. A good search result ranking solution is needed for organizing information: the Google approach does not lend itself to small document sets. More interoperability is also needed: as things stand, aggregation of information is time consuming while discovery is impossible. Perhaps an ideal to aim for would be Peter Murray-Rust's concept of the 'datument': eXtensible Markup Language (XML) documents that integrate well-structured text and the data on which it is based [3].

2.2 Funes the Memorious

Chris Rusbridge, University of Edinburgh, DCC Director

The 2006 Long Term Knowledge Retention workshop held at NIST highlighted some important problems. Critical standards (sc. ISO 10303 [4]) exist but are not sufficient. Engineering format registries could alleviate some of the difficulties.

Funes the Memorious is the subject of a short story by Jorge Luis Borges [5]. Funes is a boy who develops such clear perception and memory that he has trouble ignoring the differences in things caused by different perspectives and the passage of time, to the extent that he can no longer make abstractions and generalizations. This robs him of vital analytical skills. The application of this parable to engineering is obvious.

The Open Archival Information System (OAIS) Reference Model [6] provides a language for talking about repositories and archives. In the engineering domain, what functions or systems have the equivalent scope of OAIS? Information Lifecycle Management (ILM) seems to be merely third party storage management. Knowledge Management (KM) is mainly concerned with person-to-person interactions and just-in-time solutions, and is orientated towards problems. Product Lifecycle Management (PLM) is another third party solution that only works for as long as the current software version lasts.

An Engineering OAIS would preserve information for its designated community. This community could potentially include all future engineers, various regulators and accident investigators. Most likely the concepts used by the designated community in the future will have drifted away from the concepts used today, rather than simply expanding in number. We can expect polysemantic words like 'understand', 'knowledge', 'information', 'data' and 'files' to be used quite differently, which makes it important to be explicit about what we mean by them now.

Files are never naked data: they are always packaged in some way. This packaging often adds extra information, and it is non-trivial to decide what is content and what is extraneous; for example, with a textual document is it just the sequence of letters that is important, or do the fonts and colours matter as well (or instead)? Information could be described as data that is well-formed and meaningful. In the OAIS model, the information required to interpret data as something meaningful is called representation information. While formats are current, much of the interpretive work is done by software, but for obsolete formats there are various options: old software on virtual machines, newer software that can interpret the older formats, custom software manually built from representation information, generic tools that interpret files using computer-readable representation information, and migration tools that can convert obsolete formats to newer ones. In any case, it is easier to interpret the obsolete format if it is a standard.

OAIS, as a reference model, requires the presence of representation information, but does not require it be collected together or labelled as such. In the engineering case, the questions that need answering are: what information is needed to interpret the underlying model; what information is needed to interpret information encoded using ISO 10303 (informally known as the STandard for the Exchange of Product model data, or STEP); what cannot be encoded in STEP; and what information would be needed by an engineer who had never seen the current tool set, and was unfamiliar with STEP?

Essentially, the problem of curating engineering documentation can be characterized by what makes the information fragile, and what would make it robust. It is also tied up with the question: what other tools, technologies and training do today's engineers need to do a better job?

3 PANEL 1: CHALLENGES AND ISSUES IN ENGINEERING INFORMATICS

3.1 Long Term Archiving and Retrieval in the Aerospace Industry (LOTAR)

Sean Barker, BAE Systems

Computer Aided Design (CAD) systems tend to be obsolete within ten years and forgotten within twenty years. The products they model, though, may be in service for sixty years or more. Worse, the backwards compatibility of CAD software is simply not reliable. LOTAR (`http://www.prostep.org/en/projektgruppen/lotar/`) is a European collaboration focusing on the long term archiving and retrieval of digital technical product documentation, such as CAD models and Product Data Management (PDM) data. Of particular concern are issues of reuse of documentation, and the place of digital documentation in certification and cases of legal liability.

There are four levels at which long term archiving can operate. Keeping hardware is simply not practical. Keeping data is the province of the OAIS Reference Model. Keeping knowledge is the area investigated by the KIM project. LOTAR is at the remaining level: keeping the model. Ideally, an OAIS should operate as a black box from the perspective of data producers and consumers; it succeeds if the data that comes out is the same as the data that went in. It is important to be able to prove that this is the case, thus validation and verification are needed; point clouds are proving a useful technique — if points generated

on the retrieved model match those created on the original model that is good support for the data out being good. Future work by LOTAR will include more representations, validation properties and a virtual schema database.

3.2 ITI TranscenData Activities in the Development of Standards and Technology for the LTKR of 3D Product Data

Andy Chinn, TranscenData Europe Ltd

TranscenData Europe Ltd is an international company that works on five major practice areas: automation, PLM integration, quality testing and repair, STEP translation and development, and the Initial Graphics Exchange Specification (IGES). One basic Long Term Archival (LTA) approach for dealing with the complexity of three-dimensional (3D) CAD is to use the STEP standard. However, there are potential data losses during STEP Export for LTA, such as shape change, position change, and part resize. The objectives of 3D CAD LTA is to enable comprehensive and precise validation of part models and to avoid false migration on both import and export. There are several different techniques that can be used: quality-based metrics (as used in STEP part 59 [7]), sampling point methods, mass property validation (although this is prone to false positives and false negatives) and geometric analysis (as used in STEP part 42 [8]). TranscenData provides CADIQ STEP translation validation to analyse and identify problems during the STEP translation for the LTA process.

3.3 The MIMER project — new solutions to archive PLM and 3D data

Nadia Rincon Turpin, Jotne EPM Technology

There are four basic approaches to archiving 3D data. Reducing to 2D and storing in hard copy loses a lot of information and utility. Archiving the hardware and software requires heavy investment in IT support. Continuous conversion between proprietary formats means exposing the data to continuous risk of corruption. Thus the optimum solution is to use a neutral exchange standard.

The MIMER project aims to deploy the next generation of long-term archiving tools based on STEP (ISO 10303), OAIS (ISO 14721) and LOTAR. The archive process execution includes three major steps at ingest, following creation of the STEP file(s) and validation properties: checking for well-formed STEP syntax, checking for adherence to AP214 rules [9], and checking geometry for faithfulness to the original. At this third step, a quality report is generated, providing error details that can be visualized using the tools.

The tool set includes a search and visualization tool for locating STEP and Tagged Image File Format (TIFF) files related to a particular part. At the moment, search is limited to fields of document names or numbers, and part names or numbers.

The application of the MIMER tools was illustrated through the example of a Scandinavian military project.

3.4 Product Life Cycle Support — from standard to deployment

Rob Bodington, Eurostep Limited

Product Life Cycle Support (PLCS) is a STEP application protocol (AP239) that aims to support the exchange and sharing of product information throughout the product lifecycle [10]. The drivers for developing this new application protocol include the increasing complexity of products and the enterprises that produce them, long product lifecycles, increasing in-service commitments placed on producers, and poor existing integration across the product lifecycle. PLCS includes product definition information, maintenance schedules, product operation information, tools and test equipment support facilities.

PLCS is already being applied in industry; notable implementers include Motorola (requirements management), BAE Systems (design for manufacturing), Volvo (manufacturing for sales), the Swedish Defence Material Administration (through-life configuration management), the Norwegian Defence Logistics Organisation (product acquisition), and the US Department of Defense and the UK RAF (in-service feedback). PLCS still has challenges for long term knowledge retention: the through-life information model is more than just CAD, involving systems engineering, training and transactional standards. Further work is needed, such as integrating PLCS with non-STEP standards, providing unique identification of information and resources through life, ensuring persistent through-life semantics and keeping justification or intent well-known through life.

3.5 Long Term Sustainment of Digital Information for Engineering Design: Model Driven Approach

Sudarsan Rachuri, NIST

The challenges that faced digital preservationists ten years ago are still with us today: proliferating information and formats with IPR-related restrictions abounding, falling investment in libraries, the privatization of archives, and standards that don't address fundamental issues such as licensing. Key questions need to be answered; for example, what granularity of data needs to be preserved?

OAIS is an ISO standard reference model for long term digital preservation. Could any protocols come out of OAIS, as they did from the Foundation for Intelligent Physical Agents (FIPA) standards? For enhanced preservation of product models, the current work being conducted — the NIST Core Product Model (CPM) and Open Assembly Model (OAM) — is focusing beyond geometry [11, 12]. The CPM is based on the form, function and behaviour of a product. It is able to capture and share the full engineering context in product development; the objectives of CPM are generic, extensible and independent of any one product development process. OAM tackles the problem of attaching assembly and system-level tolerance information to archival product models. The approach uses annotations on the product model, such as the semantic mark-up method.

Standards for semantic interoperability, and sustenance of digital information for the whole product life cycle, are two important issues for the development of public standards such as STEP, and the development of theories on product representations, languages and domains.

3.6 UKCeB Challenges in Long Term Knowledge Retention

Andy Voysey, UKCeB

The UK Council for Electronic Business (UKCeB) is a trade organization working on secure information sharing for the Aerospace and Defence industries. One of its seven units is the Information Management Working Group, which commissioned research into the downstream retrieval of information. It turned out that the research landscape was fragmented, not least because the issues involved people and culture as much as engineering. For example, traditionally engineers worked in the same field, if not the same company, all their life, so their tacit knowledge was always available. Now, engineers move around a lot more, meaning that companies have to adopt strategies for retaining tacit knowledge: exit interviews or consultancies, for example. Further strategies are needed to promote information flow: automated capture of design processes is just one example.

The UKCeB was heavily involved with the University of Bath in the development of the KIM Grand Challenge project researching through-life information and knowledge management for long-lived product-service projects.

3.7 Discussion

When discussing engineering information, it is always worth bearing in mind that CAD data and the information covered by STEP only account for a small proportion of all the information that needs curating. A vast amount is tacit, or else recorded as unstructured text. There are standards outside STEP emerging to cover some of this other information, but they need tying together.

The reason why CAD has been the focus of attention is partly political and partly practical. CAD modelling is mathematically defined, and so in principle should be one of the easier questions to solve. Having said that, one of the major markets for CAD software, the automotive industry, is much more concerned with supply chain integration and high level business information than it is about CAD.

While there is an obvious business case for archiving issues — quality assurance, compliance, etc. — there are also disincentives. Does legal liability increase if more information is available? Will designers be penalized for reusing existing work, or encouraged to do so? Work is needed on judging what needs to be archived. Sometimes, quirks can be as important as recognised effects, as shown in the pharmaceutical industry.

The problem with archiving is that the immediate beneficiaries are often unknown. The term ‘archiving’ itself can also be a barrier, as it seems to imply data frozen for future historians rather than data kept available for reuse. Managers are therefore loathe to make the necessary investment. Engineers must be presented with incentives or mandates if they are to change their behaviour; ideally, good preservation practice should clearly reduce their workload, rather than increase it. The perfect tool — one that is as simple and transparent to use as a web browser, yet powerful and intelligent enough to perform automatic metadata capture, versioning, indexing and format migration on new and existing file stores — has yet to be invented, and so the preservation problem remains abstract and intangible.

4 PANEL 2: DIGITAL ARCHIVING MODELS, REPRESENTATION LANGUAGES AND STANDARDS

4.1 STEP approaches to LTKR

Mike Pratt, LMR Systems

STEP was fundamentally conceived as a standard to enable sharing and exchange, although it was also recognized as a basis for archiving. AP203 was one of the original parts of STEP, covering such issues as configuration management, product structure and geometry. The forthcoming second edition will add more capabilities, such as geometric dimensioning and tolerancing (GD&T).

STEP is starting to address issues of the reliability of data. To date, most of the effort has gone into addressing data that describes a finished model/product; now, data describing the process by which a model/product was designed deserves attention. ISO 8000 [13] is a general standard for data quality and includes issues of provenance and currency; this is being drawn up by the same subcommittee as STEP (TC184/SC4). A specialized version of the standard with respect to CAD models will be released as part 59 of STEP [7]; a draft is expected by the end of 2007. The guidelines for the quality of shape data in models are being drafted by the Strategic Automotive product data Standards Industry Group (SASIG) and will be formalized and quantified by ISO. Efforts are being made to ensure the LOTAR standards will be compatible with part 59.

The various other new areas being addressed by STEP include:

- Part 55: Models represented by their procedural or construction history.
- Part 108: Parameter based models (this is being tested by NIST).
- Part 109: Assembly models.
- Part 111: CAD 'feature' definitions for use in part 55.
- Part 112: Procedurally represented 2D CAD models.

Final models and construction history models each have their advantages — final models are more compact, whereas construction history models are easier to reuse — so STEP will recommend keeping models in both forms.

4.2 Navigation and browsing of large archives of 3D CAD models

Andrew Sherlock, ShapeSpace Ltd

ShapeSpace specializes in software to search, navigate and browse through databases of 3D CAD models. The problem that ShapeSpace aims to avoid with its software is that of searches that are too vague (returning too many results) or too specific (returning no results). The tools therefore work on the principle of ranking shapes by similarity to a chosen shape. The part browser, for example, displays a 3D grid of 3D thumbnails of part models. When a part is selected, the remaining parts are re-ranked by similarity to the chosen part and the resulting grid displayed, allowing iterative searching until the right part is retrieved.

The principle behind the software is that each shape is characterized by a signature string. The algorithm for generating this string is tailored to the kinds of searches likely to be performed by the target user group for the software.

4.3 Definition and integration of product-service data

Alison McKay, University of Leeds

There are at least five levels of LTKR. The bottom three characterize the IT infrastructure: hardware, software and applications. The fourth level is the information viewpoint. The fifth level is the enterprise viewpoint, where matters of provenance, organizational context and socio-political context are of prime importance [19].

The purpose of LTKR is so we can retrieve information and use it during throughout the period of contractual obligation (e.g. for maintenance). The pertinent information ranges from general information to information on specific projects, and from open to proprietary data. This is not just a technical problem but also one at the enterprise level.

Researchers into Product Lifecycle Management, and indeed all proponents of LTKR, have the challenge of building a strong case for these issues to be tackled. There is a whole range of beneficiaries from LTKR activities, but are the benefits always higher than the cost? There may be information that it is not worth keeping, and that information should be weeded out. The metric for deciding should be linked to the strategic intents of the organization, and not just to protection from litigation.

It is impossible to predict the uses to which our current product model data will be put in future, nor who the users will be or what their contexts and information needs will be. This has an impact on our knowledge management systems, which build activity models, assess user requirements then devise ways of fulfilling these requirements. Information systems need to be low-risk and flexible.

The representations needed include CAD models, data history (including more than CAD), and the enterprise level rationale, processes and history. The quality of CAD models can be judged on whether the shapes are well defined, since solid modelling has a good underlying mathematical model.

When implementing meta-models, there is no reason to feel restricted to just one. Multiple layers of meta-models can be stacked together to cater for multiple viewpoints.

4.4 Representing Policies for Long-Term Data Retention

MacKenzie Smith, Associate Director for Technology, MIT Libraries

One of the obstacles facing a repository wishing to share its content with others is communicating the repository's policies. The MIT Libraries are involved in the PLEDGE project, which is looking at extracting policies and making them discoverable, configurable and sharable in an automated way.

A six stage model is used. The repository may well have a standard by which it is judged, such as the Trusted Repositories Audit Checklist (TRAC) being developed by RLG and NARA [20] (A). The repository takes this checklist and develops a set of local policies to bring it into compliance with the standard (B). These policies are then mapped to functions and capabilities in the preservation environment (C). These functions are then

translated into rules for the repository software to follow (D). The rules are enforced and monitored using microservices (E) which produce state information logs and reports (F), which in turn are used as the data on which assessments against the checklists are based.

The envisioned policy framework has four categories: organization, environment and legal policies; community and usability policies; process and procedural policies; and technical and infrastructure policies. The latter of these categories is where many of the issues about CAD would be decided.

Various approaches were considered for encoding policies in machine-readable form. Many schemata were either too specific or orphaned now that the originating project had finished. The most suitable languages turned out to be RuleML and Rei (N3) RDF. The next step will be developing a suitable protocol for rule exchange.

4.5 Collaborating to compile information about formats: the vision, the current state, and the challenges for format registries

Caroline Arms, Library of Congress

In the task of keeping digital information interpretable, dealing with formats is not enough, but it is a good starting point. In order to deal with formats, information is needed for both humans and systems, along with services that can be invoked as needed. Humans need information that will help them develop policies, for example, which formats to prefer and what to do with the other formats. Systems need information for identifying, validating, characterising, and transforming formats, as well as for converting them to delivery formats for current use, and for assessing the risks associated with each format.

The Library of Congress is charged with collecting all sorts of material. Under legal deposit they are entitled to the best quality format of a resource, but there is no actual quality threshold. For digital materials, the question of format is not just a preservation issue: it has implications for research on accumulations of data, commercial repurposing, compliance and forensic purposes and for maintaining long-lived facilities.

The Global Digital Format Registry (GDFR) promises to be a organized registry of representation information, including format information, software and codecs to aid in the delivery of data, and relationship and dependency information. It is led by Harvard University, with system development from OCLC. Some of the information will be held in the registry itself, some will be held by the Library of Congress on an escrow-type basis, and some will be held elsewhere, possibly on a commercial basis.

In the meantime, the (US) National Digital Information Infrastructure and Preservation Program (NDIIPP) is working on small parts of the problem, for example, the fine-grained (sub)format characterisation and validation required by specialized archives. The Library of Congress itself has put together a website dealing with factors for sustainability of various formats.

4.6 Knowledge retention infrastructures and archives, and their validation

David Giaretta, CCLRC, CASPAR Project Director

The OAIS Reference Model defines information as knowledge that can be exchanged in the form of data, and long term as ‘a period of time long enough for there to be concern

about the impacts of changing technologies, . . . and of a changing user community, on the information being held in a repository.' We can be reasonably sure of being able to preserve original streams of bits and paper documents into the long term, and we can rely on there being people, computers and remote access in the future. A useful concept is the chain of preservation, as each generation attempts to pass on its information to the next; like all chains it is only as good as its weakest link.

There are several strategies that can be adopted. In decreasing order of preference:

1. Creating a preservation plan at the point of creation.
2. Reliance on those familiar with the information to curate it. (Producing machine-processable descriptions of the information would be ideal but is not a necessity.)
3. Software maintenance or emulation (although this means one is restricted to the functionality of the old software).
4. Digital archaeology (although it is much harder to guess semantics than it is structure/syntax).
5. Just guess!

The OAIS Information Model defines an information object as a data object that is interpreted using representation information (which is itself one or more information objects). Representation information can be information about semantics (using ontologies, folksonomies, etc.) or about structure. This is an important point: XML itself isn't enough to ensure long term interpretability, as the semantics of a schema can be lost.

A shared preservation infrastructure helps to cut the cost of preserving material. Such an infrastructure should provide persistent identifiers (none are guaranteed to last!), access to standards and preservation description information, interfaces to support interoperability, and accreditation to ensure trustworthiness. The Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval (CASPAR) Project is a European project aiming to set up such an infrastructure. It will operate on a basis of virtualization, and will have a modular architecture to allow different software to be swapped in and out as required.

CASPAR is trialling its infrastructure on three diverse testbeds, covering the three target areas of culture, performing art and science.

4.7 Discussion

The CASPAR testbeds have shown that changes in hardware and software is the relatively straightforward problem to solve: interoperability is key. The difficult issue to predict and plan for is the evolution of the designated community.

One of the peculiarities of CAD models is that many of the recent formats do not completely encode the model data. Rather, a set of instructions is recorded to allow the software to recreate the model. Thus, CAD models are useless without the software to interpret it. Even if the code for the software were released to, say, the Library of Congress under escrow, it would take decades of work to construct a useful tool from it.

Vendor lock-in is a phenomenon that is diminishing, thanks to increasing emphasis on producing end-to-end, enterprise-level PLM systems. Such systems rely on interoperabil-

ity between tools, hence vendors are finding it increasingly in their interests to support interoperability. Patents are a significant stumbling block.

Representing geometry as (arbitrarily large) bags of (arbitrarily small) triangles is one technique for preserving information. Tools are being developed to re-interpret bags of triangles as solid objects. The problem with bags of triangles is that a) they have large file sizes, b) problems can arise at the limits of resolution, c) a lot of reinterpretation needs to be done.

Digital storage space at the scale of gigabytes is inexpensive and of no real concern. At the scale of exabytes, the power consumption of the disks becomes a significant cost, even if the cost of disks themselves is disregarded.

What is missing from the debate is a comprehensive set of use cases for the engineering preservation problem.

In terms of indexing archived information, most classification is currently based on definitions which are culturally sensitive. A classification based on objective and rigorous definitions is required, so that semantic drift does not make it unintelligible.

A funding model whereby the cost of preserving information is borne by those retrieving it rather than those depositing it could make the price of replacement parts (say) highly volatile. This would not be a good business model. The cost of preservation, and the risk of having to re-engineer parts, needs to be considered up front and factored into cost models.

5 BREAKOUT SESSIONS

5.1 Engineering Informatics

The session began with those present voicing the questions they would most like to discuss. These were rationalised into the following discussion points.

1. Abstracting from silos of information
    - Bringing in interested parties from outside engineering
    - Usage of data
    - What is the most important information to archive?
2. What needs to be done to make practical progress on these issues?
    - Open community with vendor involvement
    - What are the costs of implementation?
3. Business/risk issues
    - Risk-cost models to aid decision making
    - Dealing with intellectual property rights
4. Representation information
    - What knowledge can be preserved by the designated community?
5. Retrieving information from large repositories

5.1.1 Abstracting from silos of information

At present, organizations work with literally hundreds of different tools. Each of these works with information in its own format, and only rarely can these formats be used by other tools. Thus, in this way each tool can be said to create its own silo of information.

The STEP standard helps to some extent, but it does not cover all the silos. PLCS is making headway in integrating across STEP components and adding additional capabilities, but it is simply not wide-ranging enough. While it may be possible to advance STEP into other areas, this will necessarily take time; there will always be a lag of a few years between current practice and the standards. In general, standards are always tricky to define as specialists can take too long discussing fine detail: generalists are needed to keep development moving.

If using STEP is Plan A, Plan B is probably getting reliable backwards compatibility between versions of the same CAD system. Plan C is using cut-down, lightweight formats. One peculiarity in the field of lightweight formats is that vendors are keen to develop and push their own (e.g. Dassault's 3D XML, UGS' JT Format). LOTAR is bringing in interest from outside its own area of applicability. While it is mainly concerned with aerospace engineering, there is interest from the automotive industry in applying the results of the project. The pharmaceutical industry is also keeping an eye on LOTAR.

5.1.2 Making practical progress

One of the major problems for long term knowledge retention is the attitude that CAD vendors have towards the issues. It is not so much the case that CAD vendors have a business policy of poor inter-version compatibility or poor import/export facilities, it is rather that compatibility and translation is not given a high enough priority, so that any functionality developed in that area is unreliable. How can this attitude be changed?

The case of Microsoft's Office OpenXML was discussed. Microsoft's decision to open up the specifications for its new version of Office was probably based on three issues: a) the increasing concern that many large companies have about the longevity of their Office-format documents, b) the existence of a credible open-source alternative, and c) the combination of both these issues leading the State of Massachusetts to ban MS Office in favour of OpenOffice.org. The problem with the CAD case is that: a) only a small proportion of CAD customers have long term knowledge retention needs, b) CAD is a relatively small concern for most engineering firms, and c) it is not always immediately obvious when CAD interoperation has gone wrong. It was suggested that even the combined lobbying power of the Western defence industry may not be sufficient to get CAD vendors to take these issues seriously. Perhaps more aggressive lobbying is needed.

It would be useful to enumerate the different silos that firms use: we could expect at least three quarters to be common to all defence contractors. A gap analysis should then be performed, to see which silos are not covered by STEP or other recognised standards; a taxonomy would be needed for these gaps. This work could be used to manage expectations of clients and managers.

5.1.3 Business and risk issues

IPR issues can be obstructive to progress. PLCS may have been driven by the UK's MOD, but who knows how many patents may be involved in implementing it?

Airbus is aligning all its information silos on XML, to allow easier cross-site collaboration and to permit greater integration. Problematic formats such as spreadsheets have been rejected in favour of more easily curated alternatives.

In order to minimise risk, the following tactics were suggested.

- Use STEP, de facto standards (e.g. PDF for word processed documents) and simple formats wherever possible.
- Save copies in several alternative formats.
- Recognize that some elements will have to be reverse engineered, and budget accordingly.
- Pay for gap analysis; identify the art of the possible.

Much of the problem is not with representation information, but with fixity information (proof that the information is unchanged) and provenance information (proof of where the information came from).

5.2 Archiving Standards, Languages and Representations

This discussion focuses on how to integrate the two communities: engineering and digital curation; the problems of archiving engineering data; the relationships between product data and digital formats, archiving standards (e.g. OAIS and STEP), languages and representations for engineering data.

5.2.1 What information needs to be retained?

Should one retain all information from the early stages of a design, or only keep the information on the final stage? The answer is still not clear as there are conflicting views from users and producers of the information. In general, the users want to acquire as much of the information as possible for future generations; but the producers are less enthusiastic about retaining such a volume of information. It may be of value to keep all the information, but the question is, is it possible to keep all the data? For example, is it possible to record the history of a design in different versions, like 'heavy' CAD models. On the other hand, as the future is impossible to predict, should we keep everything throughout the entire life cycle? This might be a sensible strategy, if by archiving the final information the information from earlier stages is kept alive.

If a two tier approach to archiving is taken, where live information and information likely to be needed in the near future is kept online, while other information is kept in offline storage, the question of where to draw the line arises, and what sort of 'air gap' should exist between the two file stores. Given the retrieval times for offline files, the search mechanisms for these files need to be well honed.

5.2.2 Exploration of use cases to demonstrate the usefulness of archiving, e.g. STEP in 50 years.

In the short term, especially for business, we need to demonstrate the usefulness quickly. We need to establish which industries are already archiving data, especially engineering data, and what businesses will benefit from archiving their data. In the engineering industry, it is desired to have the requirement of long term preservation, e.g. aerospace needs data to be retained for periods of 30-50 years. Engineering redesign is a good example use case for archiving. Obviously, such use cases are incredibly valuable, but they may be expensive to produce. However, it is necessary to investigate use cases from both engineering community and library community.

5.2.3 Discussion on STEP

STEP is an international standard addressing the representation and exchange of product data. It is extensible and covers the product design phase and additional life-cycle phases, such as maintenance and repair. Currently, two Applications Protocols are closely related to product lifecycle management (PLM): AP239 and AP233.

- AP239, Product Life Cycle Support (PLCS) supports the exchanging and sharing of product information throughout the full product life, and addresses the complete product support domain.
- AP233, Systems Engineering Data Representation, is more related to system engineering design and mathematical models [21].

Archiving as STEP files (e.g. EXPRESS schema and XML schema) may offer a solution for product model preservation. It is generally good for web services that use open source enabling people to access them freely. Unfortunately there are major factors limiting the potential implementation of STEP:

- STEP may be too 'heavy' for most applications;
- STEP is too big: there is uncertainty on how to use different modules;
- The cost of purchasing the standard can hinder its application.

In addition, it is unclear what STEP can do on different levels, like requirements, functions and behaviours.

6 US KEYNOTE PRESENTATION

6.1 Challenges in Preservation of Digital Engineering Artefacts and Processes

William C. Regli, Drexel University

As noted at the 2006 LTKR workshop, digital preservation is the mitigation of the deleterious effects of technology obsolescence, media degradation and fading human memory [22]. There are several factors that distinguish digital preservation in the engineering context. The data types are unusually complex, diverse, and hard to describe. The data elements can have large file sizes, with single parts on the scale of gigabytes. The temporal aspect is important, particularly where something changed and

why. The business process workflows need to be preserved as well as the end products. Engineering information is both descriptive and prescriptive, with no clearly defined set of stakeholders.

We can be confident in making the following assumptions:

- Disk space is nearly free.
- People don't wish to burdened by additional work, so preservation workflows need to be automated.
- Simple formats stand the greatest chance of remaining interpretable.
- It is hard to envision the ultimate use of the data; it may be used in digital archaeology, forensics or historical research.

The (US) Department of Energy's official format for the preservation of product model data is the 2D technical drawing. This omits all sorts of useful data: manufacturing planning data, how a part fits into an assembly and how it is fixed there, the specification of its function, inspection data, fabrication methodology, provenance information, etc. Even STEP neglects some key pieces of information, such as behaviour models and the forces, torques and power consumption needed for robust physics-based models.

There are several immediate needs for digital preservation in engineering.

- *Engineering Format Registries.* Drexel is working on a taxonomy of engineering data formats, captured using Web Ontology Language (OWL) on a globally accessible Wiki. The information captured about each format can include its name, formal specifications, provenance, and example files.
- *Use cases.* More work needs to be done.
- *Representations for important data not covered by STEP.* Simple, self-documenting formats would be most suitable; if necessary, techniques can be developed for inferring richer models from these simple formats. The choice depends on the use cases.
- *Software tools to make ingest transparent, data interrogable, and preservation tangible.* Drexel has developed a Digital Archive Tool (DArT) that allows 'instant archiving'. When a file is dragged and dropped onto the DArT icon, the tool uploads it to the archive server, performs various migrations and conversions, adds metadata and backs up other relevant files, all automatically.
- *Open testbeds.* The two main testbeds used by Drexel are both Cyberinfrastructure TEAM (CI-TEAM) demonstration projects. One, Engineering for Bio-Inspired Robotics, involves the design and manufacture of robots; the other project, CIBER-U, involves teaching design through the dissection and reverse engineering of artefacts such as waterproof cameras. The work from these projects is currently being collected on a Wiki.
- *Best practice policies.* NIST was once asked to investigate a battery additive. The investigation did not reveal a purpose for the additive, but did find that the batteries worked better when the terminals were cleaned — a procedure carried out when the additive was tested; this emphasises the importance of best practice.

7 FUNDING PROGRAMMES

The following programmes were identified as possible sources of funding for further research in this area:

- ESPRC: Collaboration for Success through People. This programme funds travel and subsistence for collaborating working; it could potentially fund exchanges of researchers between the (UK) KIM Project and the (US) Digital Engineering Archives Project, for example. The funding would run for eighteen months starting in October 2007. Chris McMahon volunteered to co-ordinate a bid for this programme.
- EU FP7 ICT Challenge 4: Digital libraries and content. The DCC is applying for funding from this programme to research preservation techniques for databases. Jonathan Corney volunteered to investigate the feasibility of a LTKR-based application, with input from Sudarsan Rachiri on NIST activities.
- Until March 2008, the DCC has funds to support visiting researchers for periods of one to three months. Researchers from the US and from industry would be eligible. The KIM Project also has some funds that could be put towards exchanges, such as sending a KIM researcher to work on the DArT software.

8 CONCLUSIONS

All participants were invited to contribute to NIST's Digital Preservation Wiki at `http://digitalpreservation.wikispaces.com/`, and to attend the workshop *Long Term Sustainment of Digital Information for Science and Engineering: Putting the Pieces Together* , to be held on 24-25 April 2007 at NIST, Gaithersburg, MD 20899, USA. This workshop will be weighted much more heavily towards breakout discussions than panels and presentations.

Sean Barker and Peter Bergstrom (EuroSTEP) are working on the VIVACE Project (`http://www.vivaceproject.com/`), which is an EU FP6 Integrated Project aiming to deliver an aeronautical collaborative design environment. The KIM Project should establish a dialogue with VIVACE to determine whether any LTKR issues come out of the latter project.

Despite occupying the thoughts of specialists in the area, the importance of LTKR for engineering is not appreciated widely enough. A clear statement of the problem is needed; the phrasing used at the previous LTKR workshop would be a good starting point that could be developed by the community using a Wiki: either the NIST Digital Preservation Wiki or the DCC Development Wiki (`http://twiki.dcc.rl.ac.uk/bin/view`). Some sort of hard sell is also required, perhaps based on some proof of concept involving old CAD data or current, artificially aged CAD data. Such a demonstration could showcase what can be done, and highlight what cannot yet be done. The problems associated with distributed working can be effectively illustrated by groups of mutually geographically remote universities working on the same data. Accompanying cost models would also be useful. Experience from the chemistry domain suggets that an effective case can be made by combining worst case scenarios with illustrations of

the immediate benefits of addressing the problem. Tony Brown (AWE) and Alex Ball (UKOLN) will look at creating a set of industrial use cases for LTKR.

One of the significant problems with providing a tool to migrate files between the various versions of a format is that of permuting all the possible migration pathways. One efficient solution is to migrate files via a single intermediate format. A related technique is to generate the migration software automatically using pattern ontologies as a basis: this meta-metamodel is used to identify the semantics of the metamodel. BAE Systems did some exploratory work in this area, but the tool is not being actively developed. Sean Barker (BAE Systems) would be willing to open this work up to academic development, perhaps as part of a taught Masters course.

The emulation approach to reusing old CAD data is often overlooked, but William Regli (Drexel University) has some experience of using CAD software on VMware, and David Giaretta (CCLRC) also has an interest in virtual machines. Possible candidate software to test would be open software such as Open Pelorus or OpenCASCADE, or venerable commercial products such as Romulus or early UGS offerings.

Further work in this area will be helped if the key institutional players in information management are identified. The UKCeB and John Lawless (MoD) will put together a list.

REFERENCES

[1] T D. Wilson. The nonsense of 'knowledge management'. *Information Research*, 8(1):144, October 2002. URL http://informationr.net/ir/8-1/paper144.html.

[2] Roberto J. Rodrigues. Information systems: The key to evidence-based health practice. *Bulletin of the World Health Organization*, 78(11):1344–1351, November 2000. URL http://www.scielosp.org/scielo.php?script=sci_arttext&pid=S0042-96862000001100010&lng=en&nrm=iso.

[3] Peter Murray-Rust and Henry S. Rzepa. The next big thing: From hypermedia to datuments. *Journal of Digital Information*, 5(1):248, 2004. URL http://jodi.tamu.edu/Articles/v05/i01/Murray-Rust/.

[4] ISO 10303. Industrial automation systems and integration – Product data representation and exchange. Multipart standard.

[5] Jorge Luis Borges. Funes, the Memorious. In *Ficciones*, pages 107–116. Grove Press, New York, 1962.

[6] CCSDS. Reference model for an Open Archival Information System (OAIS). Blue Book CCSDS 650.0-B-1, Consultative Committee for Space Data Systems, 2002. URL http://public.ccsds.org/publications/archive/650x0b1.pdf. Also published as ISO 14721:2003.

[7] ISO/CD 10303-59. Industrial automation systems and integration – Product data representation and exchange – Part 59: Integrated generic resource – Quality of product shape data.

[8] ISO 10303-42:2003. Industrial automation systems and integration – Product data representation and exchange – Part 42: Integrated generic resource: Geometric and topological representation.

[9] ISO 10303-214:2003. Industrial automation systems and integration – Product data representation and exchange – Part 214: Application protocol: Core data for automotive mechanical design processes.

[10] ISO 10303-239:2005. Industrial automation systems and integration – Product data representation and exchange – Part 239: Application protocol: Product life cycle support.

[11] Steven Fenves. A core product model for representing design information. Interagency Report NISTIR 6736, National Institute of Standards and Technology, April 2001. URL http://www.mel.nist.gov/msidlibrary/doc/ir6736.pdf.

[12] Mehmet M. Baysal, Utpal Roy, Rachuri Sudarsan, Ram D. Sriram, and Kevin Lyons. The Open Assembly Model for the exchange of assembly and tolerance information: Overview and example. In *Proceedings of the 2004 ASME Design Engineering Technical Conferences*, Salt Lake City, UT, 28 September – 2 October . American Society of Mechanical Engineers.

[13] ISO/NP 8000. Catalogue management systems – Requirements.

[14] ISO 10303-55:2005. Industrial automation systems and integration – Product data representation and exchange – Part 55: Integrated generic resource: Procedural and hybrid representation.

[15] ISO 10303-108:2005. Industrial automation systems and integration – Product data representation and exchange – Part 108: Integrated application resource: Parameterization and constraints for explicit geometric product models.

[16] ISO 10303-109:2004. Industrial automation systems and integration – Product data representation and exchange – Part 109: Integrated application resource: Kinematic and geometric constraints for assembly models.

[17] ISO/PRF 10303-111. Industrial automation systems and integration – Product data representation and exchange – Part 111: Integrated application resource : Elements for the procedural modelling of solid shapes.

[18] ISO 10303-112:2006. Industrial automation systems and integration – Product data representation and exchange – Part 112: Integrated application resource: Modelling commands for the exchange of procedurally represented 2D CAD models.

[19] ISO/IEC 10746. Information technology – open distributed processing – reference model. Multipart standard.

[20] OCLC and CRL. *Trustworthy Repositories Audit and Certification: Criteria and Checklist*. Online Computer Library Center, Dublin, OH, 1.0 edition, 2007. URL `http://www.crl.edu/PDF/trac.pdf`.

[21] ISO/WD 10303-233. Industrial automation systems and integration – Product data representation and exchange – Part 233: Systems engineering data representation.

[22] Joshua Lubell, Sudarsan Rachuri, Eswaran Subrahmanian, and William Regli. Long term knowledge retention workshop summary. Interagency Report NISTIR 7386, National Institute of Standards and Technology, February 2006. URL `http://www.nist.gov/msidlibrary/doc/NISTIR_7386.pdf`.