Approaches to Information Curation in Engineering

Alex Ball Manjula Patel

24 June 2008

Presented by Alex Ball at the seminar, 'Knowledge and Information Management Through-Life', at the Institution of Mechanical Engineers, 1 Birdcage Walk, London.

Contents

1 Introduction

The kinds of information that have to be dealt with in engineering are many and various: product geometry, dimensions and tolerances; finite element analysis models; design process and rationale models; manufacturing process models; and so on. Engineering organizations have to be able to communicate this information internally and with suppliers, subcontractors, customers and regulators. They also have to ensure that this information is available and understandable over time, both for legal reasons and – in long-lived product-service offerings – to maintain and upgrade the product over its life. Information curation is a collective term for all these activities: keeping information fit for contemporaneous use, now and in the future. In the course of this presentation I'll be giving some general advice on how your information curation practices can be improved straight away, and presenting some current research which hints at practices you might be able to use in the future.

But first, to illustrate the kinds of issues involved, I'm going to take an example from one of the more subtle and inexact branches of manufacturing: cookery. Here's a recipe.

Lenticulam

Aliter lenticulam:

1. Coquis.
2. Cum despumaverit porrum et coriandrum viride supermittis.
3. Teres coriandri semen, puleium, laseris radicem, semen mentae et rutae.
4. Suffundis acetum, adicies mel, liquamine, aceto, defrito temperabis, adicies oleum, agitabis, si quid opus fuerit, mittis.
5. Amulo obligas, insuper oleum viride mittis, piper aspargis et inferes.

This recipe is taken from Cælius Apicius' *De Re Coquinaria*, from the 4th Century. I don't know how good your Latin is, but mine's terrible. Regardless of whether we're talking about contemporary exchange or long-term preservation, we have to at least get to point of being able to understand the recipe. So here's a translation of the recipe.

Lentils with coriander

Another lentil recipe.

1. Boil them.
2. When they have foamed, add leeks and green coriander.
3. Crush coriander seed, pennyroyal, laser root, mint seed and rue seed.
4. Moisten with vinegar, add honey, *liquamen*, vinegar, mix in a little *defrutum*, add oil and stir. Add extra as required.
5. Bind with *amulum*, drizzle with green oil and sprinkle with pepper. Serve.

A contemporary cook would probably be able to work it out from this, provided they were familiar with the ingredients. But for a modern cook, this is still no good, and not just because some of the ingredients are obscure: for a start, modern cooks require a bill of materials so they can plan their procurement strategy, and for another, the instructions are rather more vague than we're used to. So, having done some research and discovered laser root is an extinct type of giant fennel, *liquamen* is a kind of salty fish sauce, *defrutum* is a fig or grape based syrup, and *amulum* is a starch solution then one can do a series of tests and produce the following.

Lentils with coriander

250 g	lentils
2 l	water
1	leek, trimmed, washed and finely chopped
75 g	fresh coriander
5 g	coriander seed
75 g	fresh pennyroyal (or mint)
2 cloves	garlic, crushed
3 g	mint seed
3 g	rue seed (or celery seed)
5 ml	honey

10 ml	anchovy paste
10 ml	vinegar
5 ml	thick fig syrup
	plain flour
	freshly ground pepper
	olive oil

1. Wash the lentils and put them into a saucepan with 2 litres of cold water. Bring to the boil, and skim off the scum.
2. When the water has cleared, add the leek and half of the fresh coriander.
3. Grind the seeds, garlic and pennyroyal (mint), and add them to the pan.
4. Stir in the honey, anchovy paste, vinegar and fig syrup. Let the lentils simmer until they are almost cooked. Check the pan every now and then to ensure that the water has not evaporated.
5. Thicken with flour if required. At the last minute add the olive oil and the freshly ground pepper. Garnish with the remaining coriander.

Adapted from: Patrick Faas, tr. Shaun Whiteside. Around the Roman Table: Food and Feasting in Ancient Rome. *ISBN: 0-226-23347-2.*

Our ingredients now have exact quantities and preparation instructions, obscure ingredients have been replaced with modern equivalents, and the method now has additional hints to aid the less experienced cook.

This may seem like a frivolous example, but it illustrates many of the problems for information curation in engineering:

- Exchanging the information requires a common language
- Well documented languages are more likely to be readable in the future
- Basic components and methods change over time: understanding of the design intent, tolerances, etc. is required in order to make suitable substitutions later on
- Some simple but often overlooked points: things that are common sense now may not be common sense in future, and
- things that are common sense to those working on the design may not be common sense to engineers elsewhere. Both of which suggest a need for conventions and practices to be codified to some extent.
- One last point I haven't brought out so far is the secret ingredient. When cooks give away their star dishes, they will sometimes leave out a small detail to ensure no-one else can make the dish quite as well. The complication of managing both complete and incomplete versions of a design side by side also exists for engineers. I'll be coming back to this later.

But next, I'm going to introduce you to some the information curation work coming from the scientific community.

2 OAIS Reference Model

The Open Archival Information System (OAIS) Reference Model is an ISO standard developed by the space science community to provide a common terminology for talking about data repositories, their functions and contents. Illustrated here (Figure 1) is the functional model.

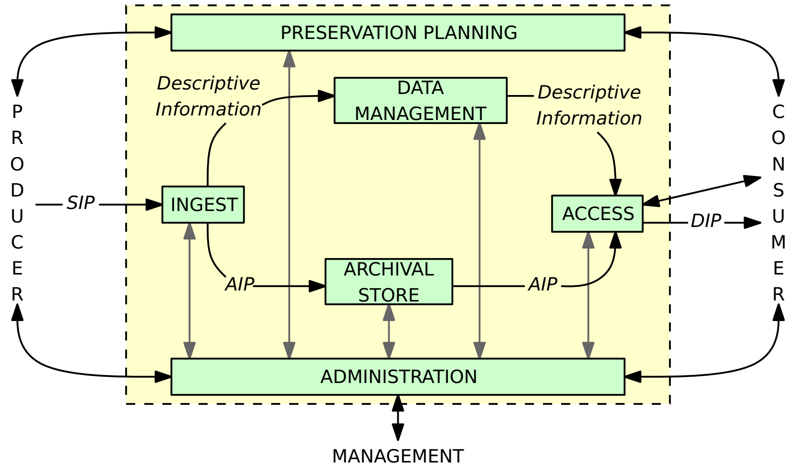


Figure 1: OAIS Functional Model

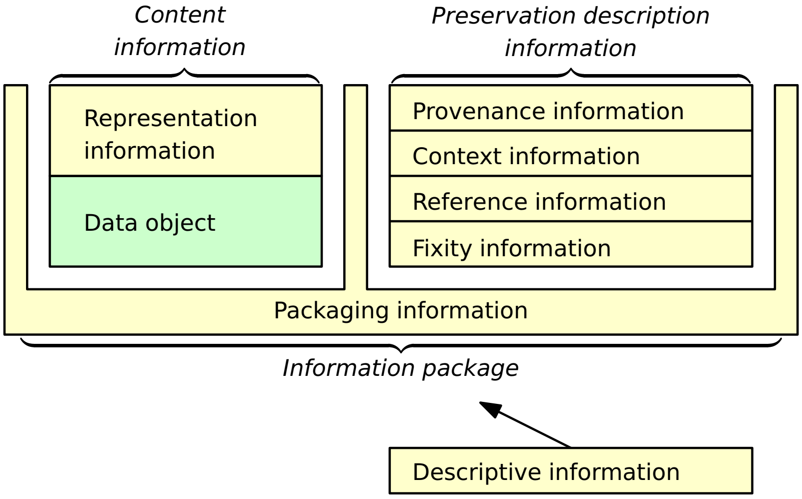
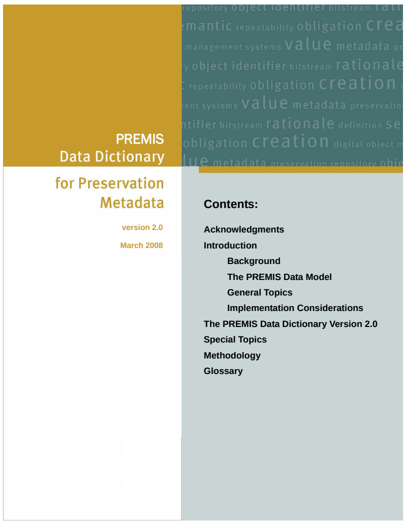


Figure 2: OAIS Information Model

- Ingest takes submission information packages, performs the necessary transformations to turn them into archival information packages, and creates descriptive information for them.
- Archival Storage looks after and retrieves the archival information packages.
- Data Management looks after and searches through the metadata associated with the archival information packages.
- Preservation Planning monitors the environment and target user group (‘designated community’) to make sure the archival information packages remain understandable.
- Access receives queries and orders from consumers, provides metadata and converts archival information packages to dissemination information packages ready for consumers to use.

These information packages are all modelled with the same structure (Figure 2). The heart of the package is the data object itself and *representation information*, the information required to render the data object understandable to the person looking at it. This is supported by the preservation description information, which tells you the processing history of the data object, its relation to other data, identifiers with which to cite it, and checksum or digest information



PREMIS
Data Dictionary
for Preservation
Metadata

version 2.0
March 2008

Contents:

Acknowledgments
Introduction
Background
The PREMIS Data Model
General Topics
Implementation Considerations
The PREMIS Data Dictionary Version 2.0
Special Topics
Methodology
Glossary

Figure 3: The PREMIS Data Dictionary version 2.0

so you can check if it's changed. Packaging information links it all together, and descriptive information helps you to find it again later.

The main purpose of the model, as I've mentioned, is to provide terminology, but it is also useful as a checklist of the minimum functions a repository needs to perform, and types of information it needs to store. What it doesn't do is recommend a particular way to go about it, as this varies with the type of material being stored.

3 Preservation Metadata

Getting more practical is PREMIS – Preservation Metadata: Implementation Strategies – which is a project to determine the metadata that most working preservation repositories are likely to need to know to support digital preservation. It combines the theoretical work of the OAIS model and the CEDARS [CURL Exemplars in Digital Archives] Project with practical metadata schemata from the National Library of Australia, the National Library of New Zealand, the Networked European Deposit Library and the OCLC Digital Archive Service, and as a result of this activity the project has produced the data dictionary you see here (Figure 3). While it is mainly concerned with representation information, it also covers the provenance and rights information needed by repositories.

To give you a broad idea of the scope of PREMIS, here are the top level metadata categories (Table 1).

Under 'Objects' we have metadata relating to data objects within the repository: everything from checksums and file sizes through structure and format to software and hardware environments where the object is known to be readable. 'Events' refer to the processing stages that lead to the creation of the data object within the repository. 'Agents' are any people or organizations connected to the data object, and under 'Rights' we have the permissions that allow the repository to carry out its preservation functions.

Again, PREMIS isn't particularly prescriptive about how one stores this information, or even *that* one should store it, as long as one knows it. For the purposes of exchanging this information between repositories, however, the PREMIS Maintenance Activity has produced PREMIS XML schemata; these can either be used to store information directly or to act as a Rosetta stone for translating between the metadata profiles of different repositories.

Table 1: Sample semantic units from the PREMIS Data Dictionary v2

Objects

- objectIdentifier
- objectCategory
- preservationLevel
- significantProperties
- compositionLevel
- fixity
- size
- format
- creatingApplication
- inhibitors
- objectCharacteristicsExtension
- originalName
- storage
- environmentCharacteristic
- environmentPurpose
- environmentNote
- dependency
- software
- hardware
- environmentExtension
- signatureInformation
- relationship
- linkingEventIdentifier
- linkingIntellectualEntityIdentifier
- linkingRightsStatementIdentifier

Events

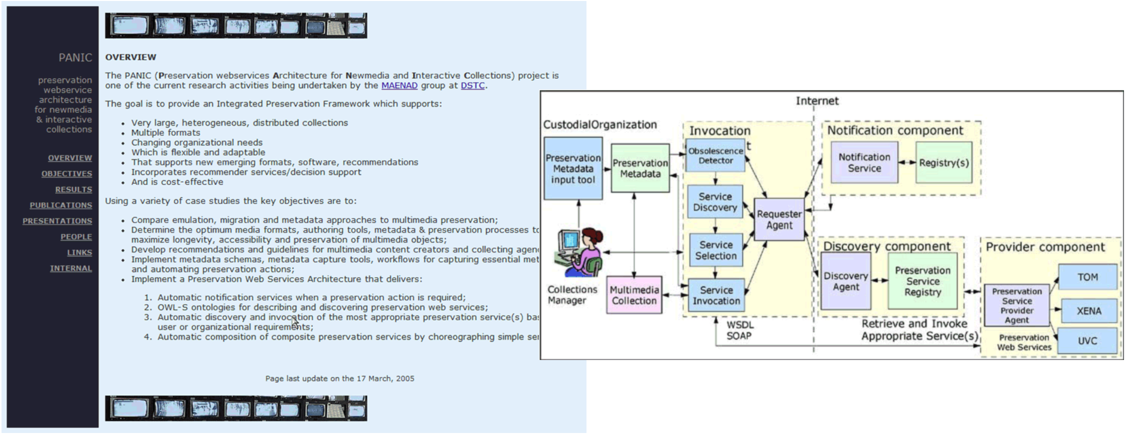
- eventIdentifier
- eventType
- eventDateTime
- eventDetail
- eventOutcomeInformation
- linkingAgentIdentifier
- linkingObjectIdentifier

Agents

- agentIdentifier
- agentName
- agentType

Rights

- rightsStatementIdentifier
- rightsBasis
- copyrightInformation
- licenseInformation
- statuteInformation
- rightsGranted
- linkingObjectIdentifier
- linkingAgentIdentifier
- rightsExtension



PANIC

preservation webservice architecture for newmedia & interactive collections

OVERVIEW
OBJECTIVES
RESULTS
PUBLICATIONS
PRESENTATIONS
PEOPLE
LINKS
INTERNAL

OVERVIEW

The PANIC (**P**reservation webservices **A**rchitecture for **N**ewmedia and **I**nteractive **C**ollections) project is one of the current research activities being undertaken by the MAENAD group at DSTC.

The goal is to provide an Integrated Preservation Framework which supports:

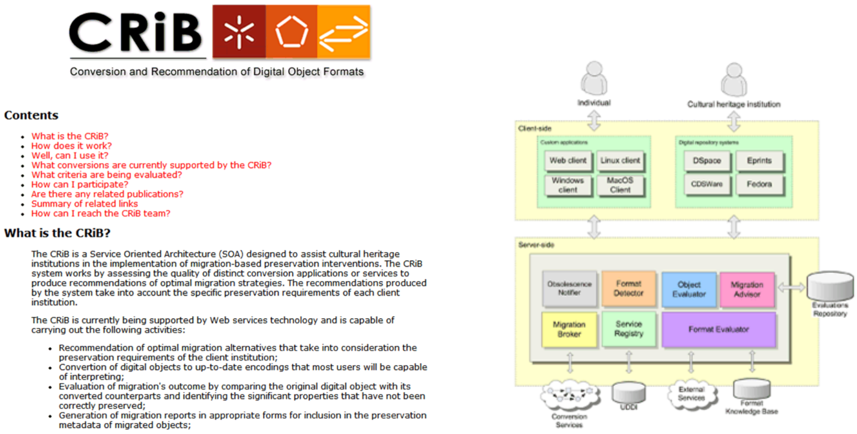
- Very large, heterogeneous, distributed collections
- Multiple formats
- Changing organizational needs
- Which is flexible and adaptable
- That supports new emerging formats, software, recommendations
- Incorporates recommender services/decision support
- And is cost-effective

Using a variety of case studies the key objectives are to:

- Compare emulation, migration and metadata approaches to multimedia preservation;
- Determine the optimum media formats, authoring tools, metadata & preservation processes to maximize longevity, accessibility and preservation of multimedia objects;
- Develop recommendations and guidelines for multimedia content creators and collecting agen
- Implement metadata schemas, metadata capture tools, workflows for capturing essential met and automating preservation actions;
- Implement a Preservation Web Services Architecture that delivers:
  1. Automatic notification services when a preservation action is required;
  2. OWL-S ontologies for describing and discovering preservation web services;
  3. Automatic discovery and invocation of the most appropriate preservation service(s) bas user or organizational requirements;
  4. Automatic composition of composite preservation services by choreographing simple ser

Page last update on the 17 March, 2005

Figure 4: The PANIC environment



CRiB

Conversion and Recommendation of Digital Object Formats

Contents

- What is the CRiB?
- How does it work?
- Well, can I use it?
- What conversions are currently supported by the CRiB?
- What criteria are being evaluated?
- How can I participate?
- Are there any related publications?
- Summary of related links
- How can I reach the CRiB team?

What is the CRiB?

The CRiB is a Service Oriented Architecture (SOA) designed to assist cultural heritage institutions in the implementation of migration-based preservation interventions. The CRiB system works by assessing the quality of distinct conversion applications or services to produce recommendations of optimal migration strategies. The recommendations produced by the system take into account the specific preservation requirements of each client institution.

The CRiB is currently being supported by Web services technology and is capable of carrying out the following activities:

- Recommendation of optimal migration alternatives that take into consideration the preservation requirements of the client institution;
- Convertion of digital objects to up-to-date encodings that most users will be capable of interpreting;
- Evaluation of migration's outcome by comparing the original digital object with its converted counterparts and identifying the significant properties that have not been correctly preserved;
- Generation of migration reports in appropriate forms for inclusion in the preservation metadata of migrated objects;

Figure 5: The CRiB environment

4 Preservation Environments

While these standards provide a lot of guidance about what one needs to do in order to stand a chance of preserving one's digital content, they don't give much away about *how* one does it, so I'm going to introduce you to a few projects which give much more practical guidance.

4.1 PANIC

PANIC [Preservation Web services Architecture for New media and Interactive Collections] is a semi-automated preservation environment developed by the University of Queensland (Figure 4). PANIC uses representation information about existing resources, possible target formats, web services, and other software tools in conjunction with local policies, to generate obsolescence alerts, recommend preservation actions and co-ordinate complex operations involving multiple tools.

4.2 CRiB

CRiB is a similar environment being developed by the University of Minho in Portugal (Figure 5). It incorporates a number of internal feedback mechanisms so that the efficiency and effectiveness of previous preservation activities are automatically taken into account when deciding on future activity.

4.3 PLANETS

PLANETS [Preservation and Long-term Access through Networked Services] is a four-year project co-funded by the European Union under the Sixth Framework Programme to address core digital preservation challenges. Partners include the national libraries of Britain, the Netherlands, Austria, and Denmark, the national archives of the UK, the Netherlands and Switzerland, as well as some major research libraries and IT companies. It is looking specifically at:

- Defining, evaluating and executing preservation activities
- Characterizing digital objects
- Tools for transforming or emulating obsolete digital assets
- Integrating tools and services in a distributed network
- Testbeds for checking the effectiveness of preservation plans
- Delivering a comprehensive Dissemination and Take-up program to ensure vendor adoption and effective user training.

Deliverables produced so far include:

- Software
  - Plato, a preservation planning tool
  - Modular emulator and Universal Virtual Computer
  - Characterization registry (based on The National Archives' PRONOM)
  - Significant property extraction tool
  - Software environment for testing preservation techniques
  - Integration framework for PLANETS tools
- Preservation planning approaches
  - S. Strodl, C. Becker, R. Neumayer, A. Rauber. 2007. *How to choose a digital preservation strategy: evaluating a preservation planning procedure.* Proceedings of the 7th ACM/IEEE Joint Conference on Digital Libraries.
- Case studies
  - Electronic theses and dissertations, interactive multimedia art

5 Standards

These preservation environments are largely tailored toward digital libraries, so you could pick them up and run with them for reports, spreadsheets and the like, but for more specialist information the capabilities aren't quite there yet. In the KIM Project we've been looking at ways in which Engineering information could be brought under a similar kind of system, and as a case in point I'd like to share with you the ideas we've had for curating CAD data.

5.1 STEP

We can't get very far in this vein without talking about standards, and in particular about STEP, ISO 10303. As I'm sure you're all aware, STEP is a vast standard for dealing with Engineering data, and has been particularly successful in creating vendor-neutral CAD representations [AP203]. In recent years it has extended its capabilities in this area to include parametric, procedural and construction history modelling [IGR55, IAR108, IAR111], along with greater support for the full product lifecycle [AP239].

Is STEP the answer? Yes, except:

- there is a long lead time between a modelling technique or information requirement arriving at the bleeding edge and it being standardized in STEP
- there is a long lead time between something being standardized in STEP and it receiving widespread (and reliable) vendor support

What can one do in the meantime?

5.2 Little standards and quasi-standards

In preservation circles there is a school of thought that one should store information in the simplest possible format that captures all the significant characteristics of the information. If you think all the important information in a report lies in the words, rather than logos or page margins, you're better off storing it for the long term in a plain text format than leaving it in a binary word processor format. The idea is that these simpler formats are a lot easier to recover information from later.

Lightweight formats are:

- simpler formats dedicated to a single purpose
- usually well documented to encourage wide support
    - real standards
    - quasi-standards (freely available specifications)
- supported by free software
- supported on many platforms
- small in file size

The same sort of principle is cautiously being applied in the Engineering domain. Many CAD vendors and consortia now offer lightweight visualization formats that provide compact representations of 3D geometry.

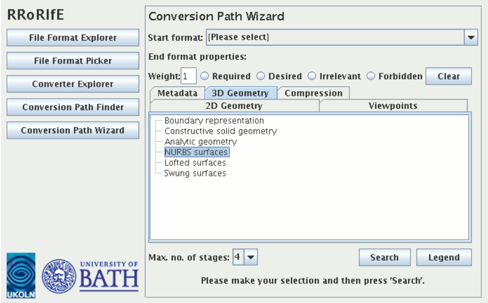
Examples:	3D XML	OpenHSF	JT	PLM XML
	PRC	U3D	X3D	XGL/ZGL

Benefits:

- Quicker file transmission: smaller sizes, streaming capabilities
- Greater freedom of platform: handheld devices, extended enterprise devices, etc.
- Cost: free viewers vs. full CAD package licences
- Resilience: published specifications, lightweight software

But:

- Design protection only by omitting detail or embedding in more secure format
- Ability to reuse geometry unclear
- Loss of useful information



RRoRIfE

Conversion Path Wizard

Start format: [Please select]

End format properties:

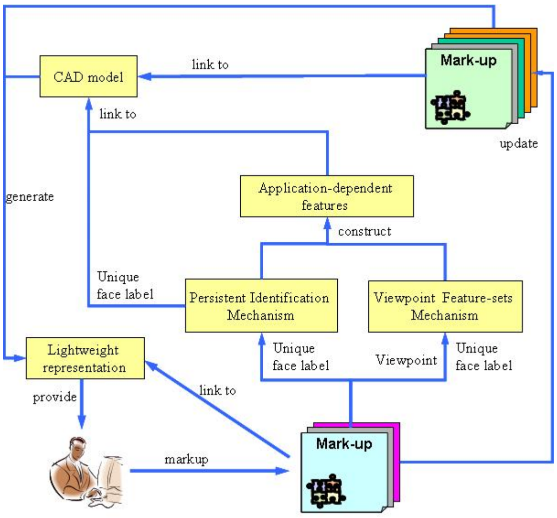
Weight: 1 ○ Required ○ Desired ○ Irrelevant ○ Forbidden

Boundary representation
Constructive solid geometry
Analytic geometry
NURBS surfaces
Lofted surfaces
Swung surfaces

Max. no. of stages: 4

Please make your selection and then press 'Search'.

Figure 6: The RRoRIfE interface



Lian Ding, University of Bath

Figure 7: The implementation of LiMMA

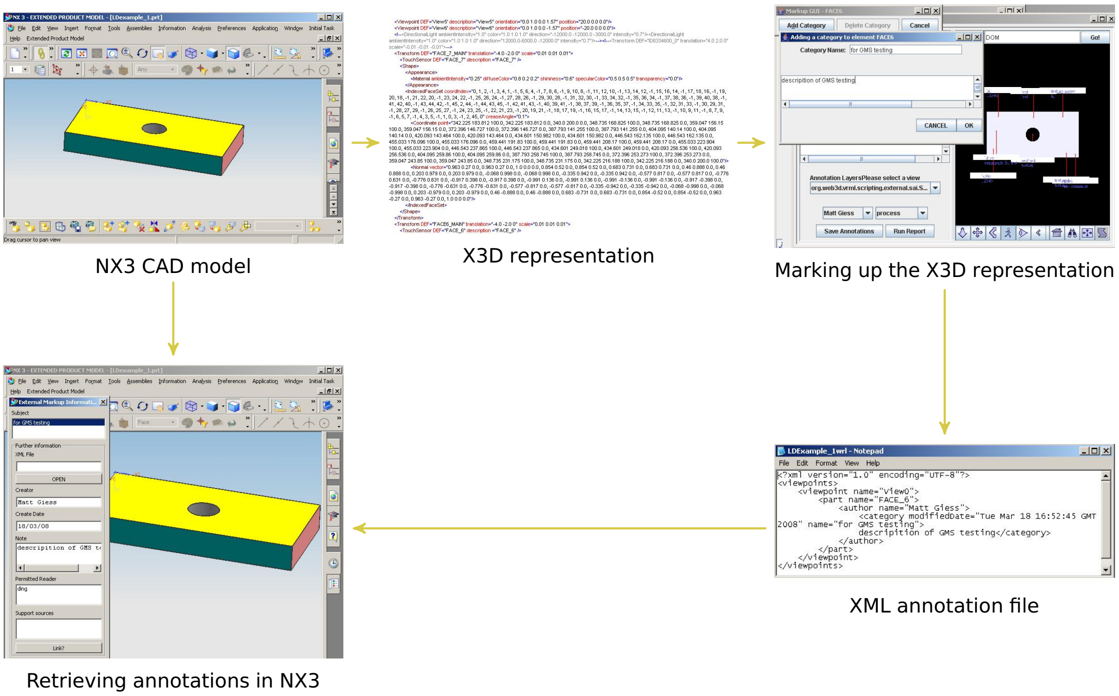
6 Preservation planning

6.1 RRoRIfE

Therefore one of the things we've been working on is a very simple preservation planning tool that enables you to check what information is at risk if you go from one format to another, and whether that's inherent in the format or a limitation of the tool you're using (Figure 6). Conversely you can also say what information you don't want to risk, and how long a tool chain you're prepared to use, and it will give you some possible migration paths to choose from. It's called the Registry/Repository of Representation Information for Engineering, and it works by storing and analyzing XML files that record how well formats and software cope with certain significant properties of files. A full-blown version of the tool would also have information about the availability of specifications, APIs, licences and perhaps costs.

6.2 LiMMA

Lightweight formats may seem like replacing one solution with a worse one, but the point is that you're not restricted to one lightweight format. You can split the information between several simple formats and link them together to form a whole. We've been looking at how one can layer text and data on top of geometric models and we've come up with a system we call LiMMA, standing for Lightweight Models with Multilayered Annotations. This diagram, by my colleague


NX3 CAD model
X3D representation
Marking up the X3D representation
XML annotation file
Retrieving annotations in NX3
NX 3 - EXTENDED PRODUCT MODEL - [LDexample_1.prt]
Adding a category to element FACE6
Add Category
Delete Category
Cancel
CANCEL
OK
Save Annotations
Run Report
LDExample_1wrl - Notepad
External Markup Informati...
OPEN
Link?

Category Name: for GMS testing

description of GMS testing

```
<?xml version="1.0" encoding="UTF-8"?>
<viewpoints>
    <viewpoint name="View0">
        <part name="FACE_6">
            <author name="Matt Giess">
                <category modifiedDate="Tue Mar 18 16:52:45 GMT 
2008" name="for GMS testing">
                description of GMS testing</category>
            </author>
        </part>
    </viewpoint>
</viewpoints>
```

Subject: for GMS testing

Further information

XML File:

Creator: Matt Giess

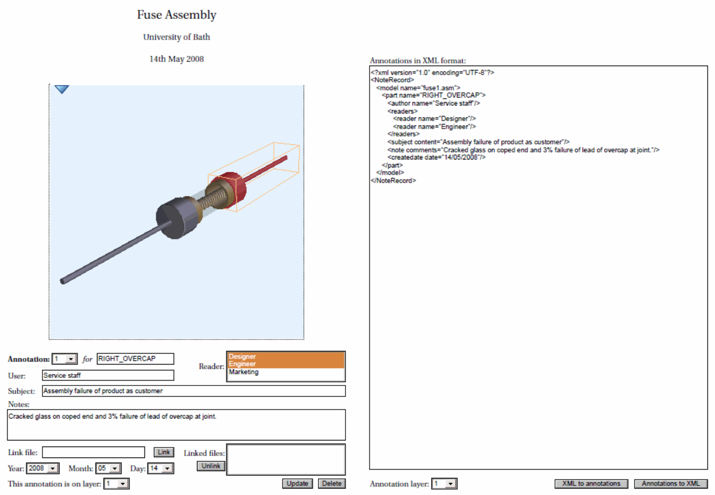
Create Date: 18/03/08

Note: descripition of GMS te

Permitted Reader: ding

Support sources:

Figure 8: LiMMA demonstrated using NX3, the X3D format and a custom Java-based viewer



Annotation: 1 *for* RIGHT_OVERCAP

Reader: Designer, Engineer, Marketing

User: Service staff

Subject: Assembly failure of product as customer

Notes: Cracked glass on coped end and 3% failure of lead of overcap at joint.

Link file:

Linked files:

Year: 2008 Month: 05 Day: 14

This annotation is on layer: 1

Annotations in XML format:

```
<?xml version="1.0" encoding="UTF-8"?>
<NoteRecord>
  <model name="fuse1.asm">
    <part name="RIGHT_OVERCAP">
      <author name="Service staff"/>
      <readers>
        <reader name="Designer"/>
        <reader name="Engineer"/>
      </readers>
      <subject content="Assembly failure of product as customer"/>
      <note comments="Cracked glass on coped end and 3% failure of lead of overcap at joint."/>
      <createdate date="14/05/2008"/>
    </part>
  </model>
</NoteRecord>
```

Annotation layer: 1

Model courtesy of Jason Matthews, University of Bath

Figure 9: LiMMA demonstrated using U3D embedded within PDF

Lian Ding, (Figure 7) shows how it works. We start with the CAD model in the top left; from this is generated the lightweight representation, which is given to someone in the extended enterprise who annotates it. These annotations are saved to an XML file which is sent back to the design team, who can layer the annotations over the original CAD model using a persistent identification mechanism.

Our first implementation of this was using a plugin developed for the NX CAD system and a custom Java-based viewer for reading X3D files. Here (Figure 8) you can see annotations being applied to an X3D file, and those same annotations being retrieved later from the NX model.

We have also managed to get this working using the new 3D capabilities of PDF (Figure 9). In both cases, the system works using persistent names for entities within the model, but we're working on a way to allow annotations to be associated with the model using sets of co-ordinates as well. Next on our to-do list is to demonstrate how this helps with the other aspects of curation, specifically generating annotations automatically from the full CAD model, and the reverse operation of combining lightweight models and annotations into a full CAD model.

7 Conclusions

That's about all I have time for, so I just want to leave you with some take-home messages.

- Curating information benefits both contemporary and future use
- Don't neglect the obvious – it won't always be obvious
- Preservation planning tools and environments make it easier to:
  - preserve and re-use expert curatorial knowledge
  - explore all the available options
- Small and simple solutions can be cheap and effective solutions for the extended enterprise and make good back-ups for the lifecycle

Thank you for your attention

Questions?

This work 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.