# KIM, ERIM and the Silo of Doom

Lessons from two long-lived data projects

Alex Ball

16th July 2010

My name's Alex Ball. I work for UKOLN at the University of Bath, but I'm here representing no fewer than three projects: the Digital Curation Centre, the KIM Project and the ERIM Project.

I imagine most people here may have heard of the Digital Curation Centre, but for the benefit of those who haven't. . .

Who are we?

- UK-based centre of expertise in digital curation.
- Partnership between Universities of Bath, Edinburgh and Glasgow.
- Primary (but not exclusive) focus on research data.

What do we do?

- Develop curation tools, resources and learning materials, either ourselves or in partnership with others: DRAMBORA, DAF, Introduction to Curation, Technology Watch Papers, Curation Reference Manual, IJDC.
- Provide training and other events such as the annual International Digital Curation Conference.
- Build communities of data curators and foster good practice, e.g. RDMF in collaboration with RIN.
- Collaborate in projects demanding digital curation expertise.

It was because of this collaboration element that I became involved in the KIM Project.

## 1 KIM Project

KIM stands for Knowledge and Information Management; actually, the original title was 'Immortal Information and Through Life Knowledge Management: Strategies and Tools for the Emerging Product Service Paradigm', but that didn't fit very easily on a slide title.

- £5.5 million Grand Challenge project that ran over 3.5 years.
- Funded by EPSRC and ESRC.
- 80 industrial collaborators from aerospace, defence and construction.
- 13 partners across 11 universities.

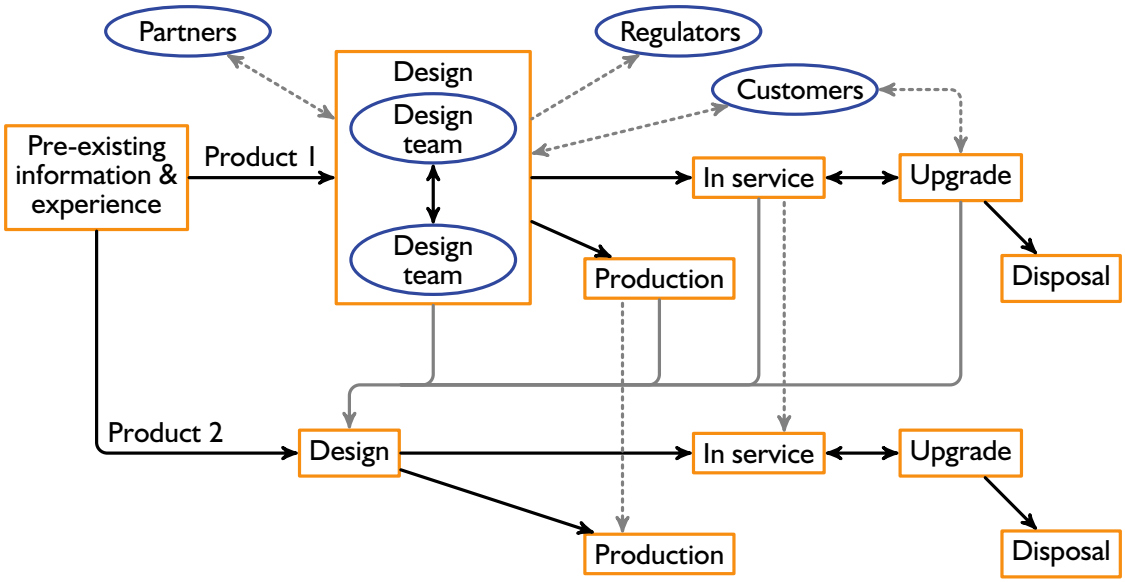


Figure 1: Engineering information flows

- Strategies and tools for the emerging product service paradigm:
  - Advanced product representation: Product/Process/Rationale (PPR) models, Information organisation, Automated capture.
  - Learning throughout the lifecycle: Information capture, Information re-use, Information value.
  - Managing the lifecycle: Incentivisation, HR, Decision Support.
  - Environment, Groups, Individuals, Practices, Tools.

My involvement with the project centred on product models. Product models are of course constructed in Computer-Aided Design systems, which is what we're all here to think about today. CAD models, and indeed the other models and documentation produced in the design phase, are key to entire lifecycle of a product.

This slide (Figure 1) shows the lifecycle of a single product. A set of requirements from a customer is matched against the set of pre-existing information and experience the firm has to offer and is used to work up first a concept and then a detailed design. Commonly this will be done within a single design team, but sometimes you may have several teams across the world working on it and having to synchronise their work. The finished design is then passed to the manufacturing or construction team to turn into a product. The design also comes in handy for the engineers performing maintenance on the product, and is absolutely essential for people investigating incidents and unexpected behaviour. If there are parts that are wearing more rapidly than expected, or if the customer's requirements change, respective changes will need to be made to the product; so at the upgrade stage the design information needs to be reloaded and those changes implemented. Depending on the longevity of the product this can happen over and over again. Once the product reaches the end of its life, the design information comes in useful yet again for determining how much of the product can be recycled, and whether there are any hazardous materials that should be disposed of in a special way.

The design information has a role to play outside the firm as well *(reveal)*. For the kinds of products our KIM collaborators work on, regulators require a copy of the design

information. It may be necessary to share aspects of the design, though not the full detail, with partners up or down the supply chain and with customers, in order to verify that the product meets the requirements.

But that's not all *(reveal)*, because these designs are costly to produce and when you're making many products that do essentially the same thing it doesn't make economic sense to start from scratch each time. It is common for small design elements from previous products *(reveal)*, and in some cases whole assemblies, to be re-used when designing a new product. There are other information flows from one product to the next that help to ensure continuous improvement: emerging best practice, lessons learned, and so on.

That is the theory. The practice is a little harder to achieve, because each of these boxes has the potential to act as a silo. This is why the title of this talk mentions the Silo of Doom.

No, I don't mean this kind of Silo of Doom (Figure 2), though it can sometimes feel like it.

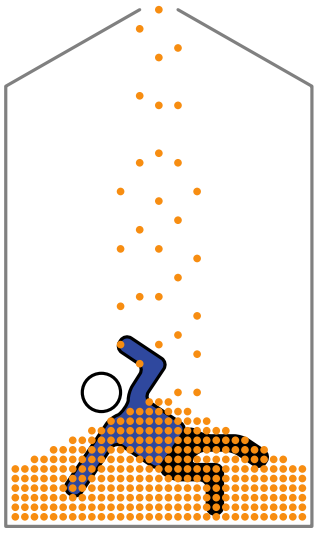

Figure 2: The Silo of Doom

I'm talking about silos of information (Figure 3), where none of the systems talk to one another and the information cannot get from one to the other without considerable manual intervention. NB. CNC = computer numeric control (of manufacturing robots); FEA = finite element analysis (a form of simulation). This problem with silos is nothing new, of course. When CAD was first introduced, it was merely an aid for producing the drawings that would be handed to the production engineers and consulted when upgrade time came along. You could say that since then, the industry has been trying to find ways to break down the silos and make the information in them more useful and usable across different systems.

For those of us in the DCC, this resonates strongly with our idea of curation, which is, broadly speaking, 'maintaining and adding value to a trusted body of information for current and future use.'

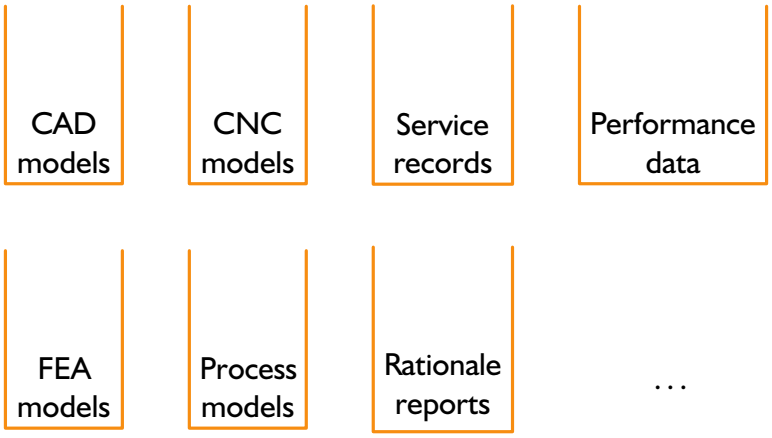


Figure 3: Information silos

- Integrating product information with *current* lifecycle systems.

I think it would be fair to say that industry has been concentrating on contemporaneous integration, producing Computer-Aided Manufacture systems that allow CNC models to be directly generated from the CAD models, and Computer-Aided Engineering systems that integrate CAD models with test and simulation data. Larger firms now make use of PLM – Product Lifecycle Management – systems that do all this, as well as handling design reviewing, providing information resources for service engineers, for sales and for marketing, version control and document management, and so on and so forth.

The problem with these systems is that they tend to work with only a limited number of tools. Say, one for each function. So this is very little help for partners with different systems working on the same contract, and very little help the next time a new version of a tool comes out. We haven't escaped the silos, they've just grown bigger. At this point ISO comes to the rescue *(reveal)*.


- Integrating product information with *future* lifecycle systems.

ISO 10303, also known as the Standard for the Exchange of Product model data (STEP), began development in 1984 and has evolved constantly since then. It is by far the largest ISO standard, with literally hundreds of parts building up a near-complete system for encoding product data. Its most notable success has been AP203, the tightly-defined, vendor-neutral CAD format, but there's all sorts of good stuff coming out, such as AP239, Product Life Cycle Support (PLCS) for integrating systems across the lifecycle. As the name suggests, STEP is again focussed on contemporary interoperability – it's an *exchange* standard after all – but a happy by-product of using a carefully controlled standard like this one is that conforming data and interfaces should remain valid long into the future.

There's got to be a catch, though, otherwise we wouldn't be here. And the catch is the delay inherent in the system. It takes a lot of time and effort to get the STEP parts right, and once they have been published they tend to be of such extent and intricacy that it takes another age before vendors become convinced of the economic benefit of implementing them, and while STEP has more stringent compliance requirements than the standards it superseded, there's still no guarantee that vendors will implement them properly. In the meantime, the state of the art has moved on, and people have data that the standards can't help them with.

Clearly there's no magic wand here. Interim solutions have to be found for individual problems on a case-by-case basis, and the problem we were concerned with in KIM was to get the CAD data working as hard as possible throughout the lifecycle in as cheap yet robust a fashion as we could think of.

I should explain that the key limitations we hoped to address were as follows (Figure 4).

We wanted it to be easy for designers to load up the design for a particular assembly and see what engineers throughout the lifecycle – from manufacturing, from in service – had discovered about it.

We wanted engineers to be able to view the CAD data in a way that makes sense to them: if you think of a edge of a cog, are they teeth that come out or gaps that go in?

We didn't want engineers to be locked into using a specific version of a specific piece software until the computer that runs it breaks, especially given how frighteningly quickly some of the respective data formats become obsolete.

We didn't want CAD models to be trapped in one site because they're too big to move around the network, but at the same time we wanted firms to be sure that they

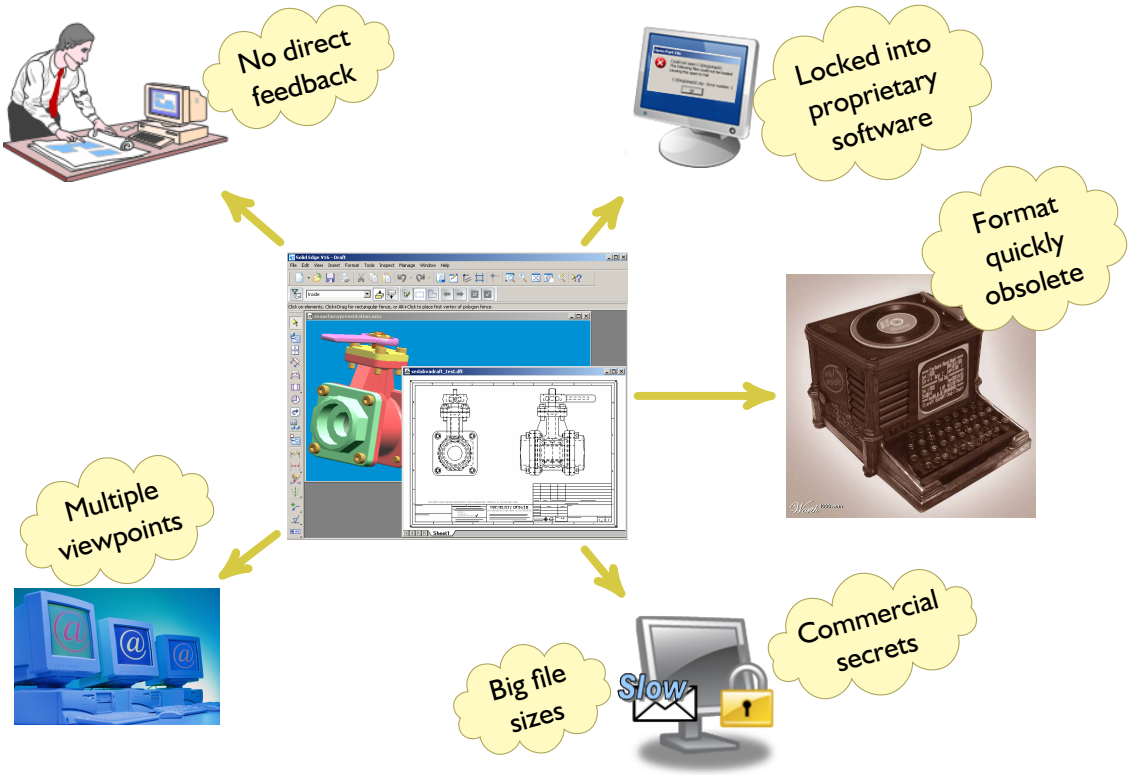


Figure 4: Limitations of CAD models

when they did move CAD models around, there was no danger of sensitive data falling into competitors' hands.

So the idea we came up with was a system of lightweight models with multilayer annotations (Figure 5). The idea is that the master CAD design stays where it is, and various surrogates are made for distribution along the lifecycle and to other stakeholders. The format for these surrogates, and how much of the CAD information that goes into them, depends on what each recipient needs, what they are allowed to see, and what tools and systems are available for them to use them on. A production engineer needs to know exact geometry, dimensions and tolerances, whereas service engineers are more interested in having the information to hand when inspecting the product. On top of these surrogates we have a series of XML files that each represent a layer of annotation;

Different annotation layers for different *viewpoints* (design, manufacture, service) and for different *security levels* (internal, public)

Geometry layer

Figure 5: Lightweight Models with Multilayer Annotations


Figure 6: Registry/Repository of Representation Information for Engineering

each layer is specific to a particular viewpoint and a particular security level. Tying all this together is a hybrid referencing system that uses named entities and geometric point mapping to tie the annotations to shapes and surfaces in the geometry. This means that the CAD model or any of its surrogates can act as the geometry layer, and any combination of corresponding annotation files can be layered on top. So a service engineer can mark annotations on a bare-bones model on a PDA, send those annotations back to a central store, where a designer can load them up on top of the master CAD model. Since all of this is well documented, it should also be possible to store a STEP surrogate in the archive, with any information lost by that conversion stored in annotation files. Many decades later, the STEP file and annotation files can be used to reconstruct the full product model. That's the hope, anyway.

To aid in this process (Figure 6), we put together a tool for determining the ideal surrogate format for the geometry, and the most suitable tools for generating it, determined by comparing user requirements for things like exact 3D surfaces or file streaming, and the support given by the various formats and converters.

Time is pressing though, so I need to say a few words about ERIM.

## 2 ERIM Project

- Engineering Research Information Management.
- Funded by JISC.
- Research Data Management Programme, Research Data Management Planning for Research Funders' Projects strand, the funder in question being the EPSRC.
- University of Bath: IdMRC and UKOLN/DCC.
- Managing data produced by the KIM Project and other IdMRC research.

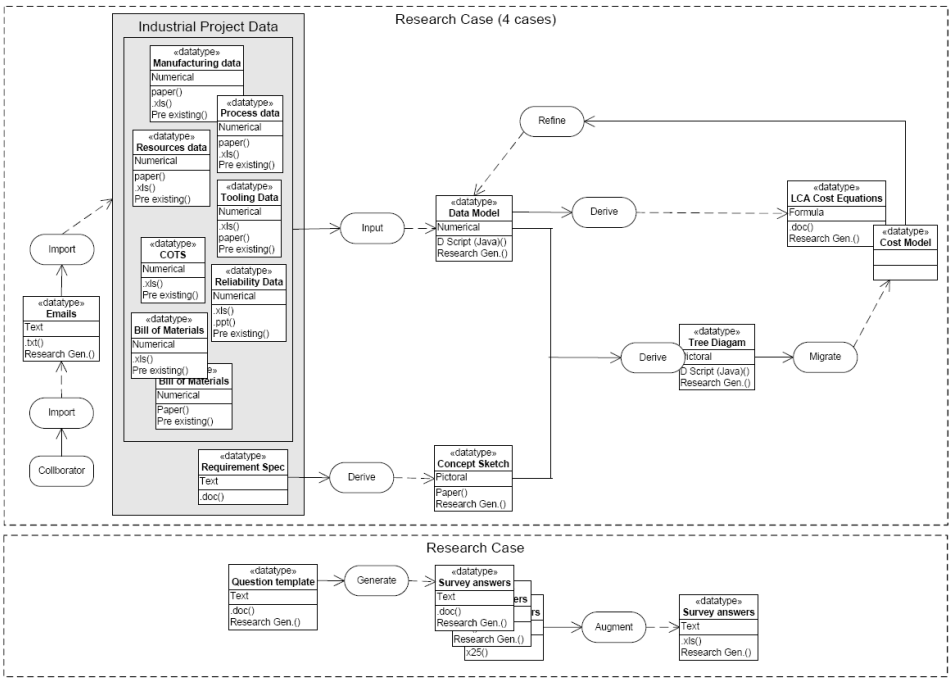


Figure 7: Map of data processing steps

We're only about half way through the project, but it's already clear that silo problems are striking again.


There's no consistency on where the data are being stored: desktop hard drives, portable hard drives, pen drives, CDs, personal network drives, team shared drives, institutional shared drive. Admittedly this is getting better, with greater use of the institutional shared drive for project data. But without keeping track of where all the data are, it's impossible to integrate them into archival systems that would perform fixity checks, format migrations and so on.

Confidentiality *(reveal)* is also a widespread barrier to integration, but it goes with the territory in engineering. We need clear summaries of access restrictions to the data, and we need to be able to find quickly the agreements that set those restrictions. We also need to make sure when negotiating these agreements that we remember to argue for rights to perform archival actions that keep the data usable.

CAD data in engineering research has many of the same contextual *(reveal)* requirements as industrial models: you need to know the process and rationale that lies behind the finished model in order to understand and re-use it. If researchers don't make the effort to document this at the time, it can be lost for ever, thereby rendering the data useless. So we're looking at systems to make that easier.

By way of example, in the course of our investigations, both we and the researchers themselves have found diagrams like these (Figure 7) really helpful in understanding the structure and relationships between their data. We're now considering how researchers

might be able to create diagrams like these as they go along, and whether any of this can be automated.

So, to wrap up...

- STEP where possible.
- Simple solutions elsewhere, as they simpler they are, the less there is to go wrong.
  - Identify the information needed.
  - Identify a simple way of storing that information.
  - Find a way of getting information there that arises from a natural workflow. Example: instead of getting a designer to write a report of why they designed something in a particular way after the fact, use a system that can generate rationale maps from their workings. In general, it's easier to get information at the time rather than leaving it to the last minute.
- Avoid creating new silos.
- Manage the silos you are forced to have carefully.

# 3 Further information

Ding, L. et al. (2009). Annotation of lightweight formats for long-term product representations. *International Journal of Computer Integrated Manufacturing, 22*(11), 1037-1053. DOI: 10.1080/09511920802527616

Ball, A. (2010). Review of the State of the Art of the Digital Curation of Research Data. (ERIM Project Document erim1rep091103ab12). University of Bath. `http://opus.bath.ac.uk/19022`

# 4 Other work

FACADE (Future-proofing Architectural Computer-Aided DEsign)

- Archiving architectural CAD models in DSpace.
- `http://facade.mit.edu/`
- Smith, M. (2009). Curating Architectural 3D CAD Models *International Journal of Digital Curation, 4*(1), 98-106. `http://ijdc.net/ijdc/article/view/105`

SHAMAN (Sustaining Heritage Access through Multivalent ArchiviNg)

- Enabling preservation in PLM systems
- `http://shaman-ip.eu/shaman/`
- Brunsmann, J. & Wilkes W. (2009). Enabling product design reuse by long-term preservation of engineering knowledge. *International Journal of Digital Curation, 4*(3), 17–28. `http://ijdc.net/ijdc/article/view/131`

See also: Lubell, J. et al. (2008). Sustaining Engineering Informatics: Toward Methods and Metrics for Digital Curation. *International Journal of Digital Curation, 3*(2), 59-73. `http://ijdc.net/ijdc/article/view/87`

# 5 Acknowledgements


- Figure 4: Lian Ding.
- Figure 5: Images by Lian Ding.
- Figure 7: Tom Howard.

- KIM Project: Lian Ding, Manjula Patel, Jason Matthews, Chris McMahon, Glen Mullineux, and many others...
- ERIM Project: Mansur Darlington, Tom Howard, Chris McMahon, Steve Culley, Liz Lyon.



*Alex Ball. DCC/UKOLN, University of Bath.* `http://www.ukoln.ac.uk/ukoln/staff/a.ball/`


---






The DCC is funded by JISC.

For more information, please visit `http://www.dcc.ac.uk/`