The Implications of Non-Functional Requirements on Cost Models for Data Sustainment

Context

What follows is a short submission to the 4C project on Cost Models for Digital Curation - *curation* being the term used by the libraries sector for what, in engineering, is called *data retention* or *data sustainment* (the term "archiving" is avoided because of the connotation of "file and forget"). The problem is not simply one of storing data, but of ensuring that the data is usable. When the Egyptians carved their hieroglyphics in stone, they created a very reliable data store, but one which was unreadable for a couple of thousand years. At least the hieroglyphics had pictures - digital data is just ones and zeros, and it takes more care (curation) to keep it readable. The 4C project has been developing cost models for that care.

OAIS - Open Archival Information Systems - was developed by NASA as a standard for comparing data sustainment systems and is now developed as ISO 14721
LOTAR is a joint project that started with the US and European Aerospace manufacturers, applying OAIS principles to the sustainment of aircraft design data over the aircraft life - typically 40 to 70 years. LOTAR results are published in Europe as the EN 9300 series of standards.
4C was a European project from the libraries sector developing cost models for data sustainment.
RASSC was a UK Technology Strategy Board funded project developing economic models for data sustainment. Its models are based on a service stack approach - more cloud computing than traditional vertically integrated archiving.

Introduction

The hypothesis proposed is that non-functional requirements are primary cost drivers for data sustainment, and consequently there can be no single or standard cost model applicable to all situations. These non-functional drivers include availability and evidential weight, and meeting them may lead to additional costs such as replicated data stores or detailed audit trails of repository processes. An information sustainment system (a repository in the OAIS sense) will be a trade-off between the costs of meeting those non-functional requirements and the risks arising from not meeting them. A consequence is that users may be better served with a tool that starts by guiding them through the trade-offs rather than one focuses immediately on cost estimation. The definition of such a trade-offs tool is beyond the limited scope of the present note.

Observations on the 4C Deliverables

In conventional Systems Engineering terminology, a functional requirement is one which defines how the system behaves in normal operation, while a non-functional requirement describes aspects of the system not directly defining the operational behaviour. These includes properties such as reliability (the probability of failure over any given period) or, in the case of information systems, "evidential weight", that is, the requirement that the information can be used as evidence in legal cases. In the 4C project, these non-functional requirements have been referred to as "Indirect economic determinates" (deliverable 4.1), however this note will follow more conventional systems engineering terminology.

The motivation for this approach starts with comments in the 4C Deliverable 3.1, that institutions find it easier to develop their own cost models than reuse existing cost models. While it might be supposed that some of the difficulty may be down to different accounting conventions, the deliverable goes on to note that there is some uncertainty in the activities required for the data sustainment system, even when the starting point is the OAIS model.

However, a brief review of both the functions and non-functional factors considered in the 4C project shows that none of the cost models discussed would support the LOTAR requirements for sustainment aircraft design data. In particular, the LOTAR project aims to sustain electrical and mechanical Computer Aided Design (CAD) models for reuse in design and for providing evidence for certification and for product liability (EN9300 Part 2, Requirements).

One of the major costs in CAD implementations is that of training users on the CAD tools, and the aerospace industry has taken the view that when accessing old data, the users will use the current generation of tools that they are familiar with rather than reactivating and learning old and disused tools. Consequently, the LOTAR standard requires additional effort during ingest to the archival system to quality check the CAD model, to convert the CAD model into a common data standard (ISO 10303 aka STEP), and then to validate that conversion. Data sustainment (preservation planning) will require firstly monitoring changes to the STEP standard to see if the format needs updating and secondly validating new versions of the CAD software used to access the models. This validation involves the generation of validation properties which check the correctness of the model *as interpreted* by the CAD tool. The LOTAR standard also requires continual monitoring of repository to demonstrate both that the registered data is still in the archive and also that it is accessible to the user. This approach follows from aircraft safety regulation, the high level of integrity needed for aircraft data and historic failures in IT service suppliers. Note that the LOTAR project builds from the premise that the repository will ensure that the data is maintained unaltered but that this is not sufficient to ensure that (following software updates) the model retains the design intent for which it is certified.

Besides model integrity, the LOTAR project focuses on meeting evidential standards needed for aircraft certification and for legal liability. The processes involved must meet the standards of evidential weight (e.g. the BSI 0008 series), which include generating extensive audit trails and demonstrating to an appropriate certifying authority that the processes and audit trails meet the criteria set down in the evidential weight standards.

Following the OAIS standard, there is a further requirement to maintain the contextual information about the CAD creation and usage processes. This includes both the reference data that is stored as part of the repository indexing system and also the background data, such as the design standards and the ISO 10303 definitions. That is, the repository must maintain data additional to the target of sustainment, a concept familiar to the memory institutions such as libraries, but not to product manufacturers or commercial research institutions.

It was the identification of a complex mixture of sustainment requirements that led the RASSC project to use a layered service model, in which the physical repository provides low level sustainment of uninterpreted data, the information level services (e.g. format translation) are at the second level and the third level covers the knowledge management services such as preservation planning and the identification of the required context information. Moreover, RASSC separated out the repository service stack (which sustains the target data) from the repository management service stack (which provides the security and accounting services needed to run the repository). The service approach is capable of using separate cost models for specific populations of data (e.g. CAD files) and separating them from that for generic repository services such as media migration and cost collection. As a result, users requiring only common data formats such as PDF do not end up cross subsidising those using complex formats where specialist data converters may need to be developed.

Risk and Repository Non-Functional Characteristics

The naive cost model is that the cost of sustaining data is proportional to the volume of data and to the length of time it is stored. However, as the data lifetime increases, the risk that the data becomes unusable increases through factors ranging from the storage media becoming unreadable, through the accession software being incompatible with the current computers and on to the users no longer understanding what the data is telling them.

The term risk here is used in the specific sense of the probability of an event occurring multiplied by the cost incurred when that event actually occurs. That cost could range from the cost of recreating the data to (in the aerospace sector) the consequential cost of having an aircraft fleet grounded until the safety case can be revalidated – a potentially enormous cost for major manufactures such as Airbus or Boeing. In the case of an aircraft crash, criminal liability may hinge on proving that the designers were not negligent, which in turn requires providing the original design documentation with appropriate evidential weight (here recreating the data after the fact is not an option).

In general, users weigh a risk against the cost of mitigating that risk, and may accept a higher level of risk in order to contain costs. For example, usual IT operating practice is to keep a second back-up copy of data off site for disaster recovery. For long term sustainment, the risk of both copies being lost becomes significant, and three or more copies may be maintained at separate sites. For example, computer rooms typically run on a single type of hardware to minimise operating costs, but long term sustainment may require the use of diverse storage hardware to mitigate the risk that there may be no migration path from one storage system to its successor.

The risk mitigation process for a single factor is illustrated in figure 1. On the left is the requirement category – in figure 1 "availability" - from which different risk criteria branch out, e.g. 99.99%, 99.9%, 98% availability. The first risk bar shows the initial level of risk using a conventional IT back-up solution, with the different risk criteria crossing the risk bar at different levels of risk. For example, if the risk criterion is that *99.999% of data must be available within an hour over the working life of the data*, then the probability of that level of service not being met is high, leading to a high risk rating.

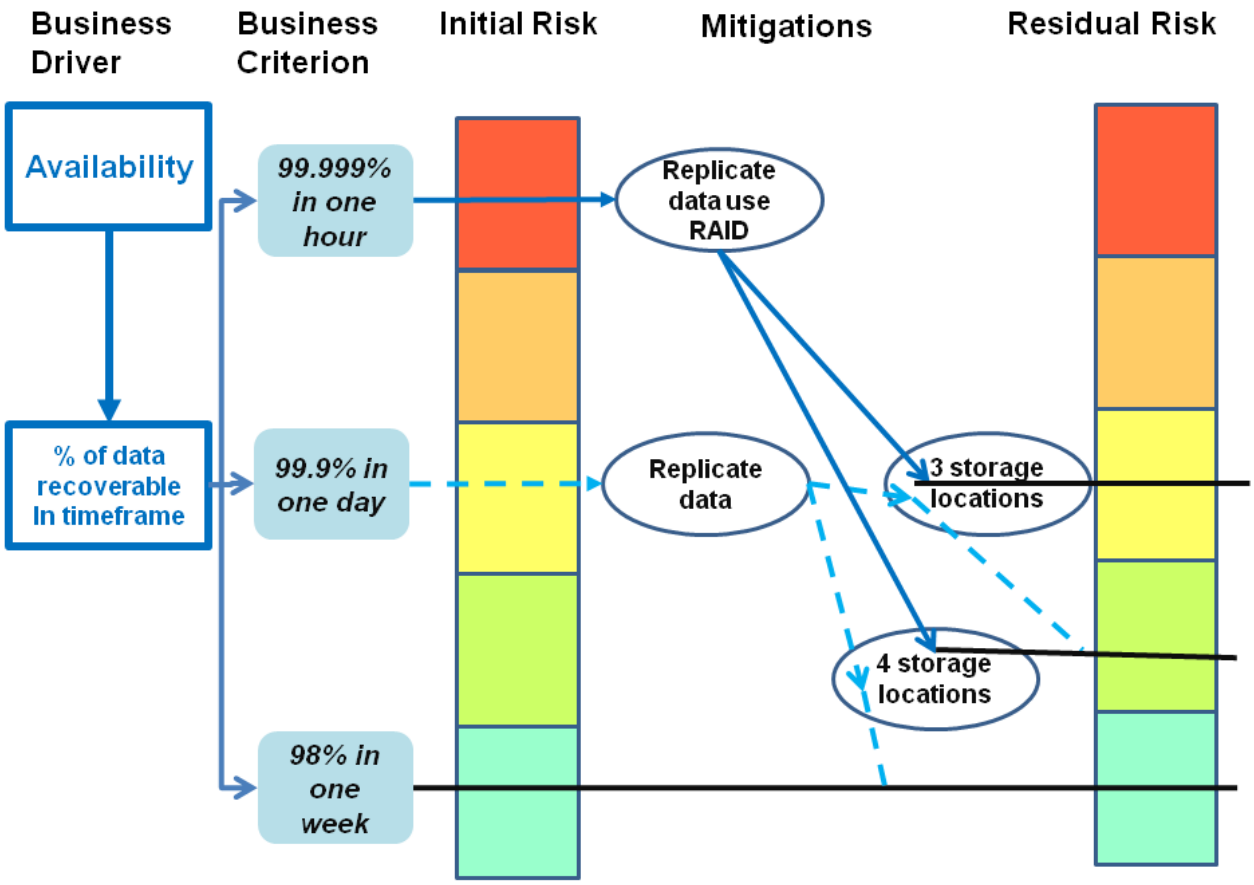

Figure 1: Non-functional Requirements and Risk Mitigation – Availability

Between the two risk bars are the mitigation actions; the rightmost risk bar shows the risk level after mitigation. For example, mitigation options for availability may be to replicate the data to three or four data stores. Following the 99.999% line, replicating to three stores reduces the risk to medium, and to four reduces it to medium-low.

If the organization were to reduce the requirement to *99.9% of data available in a day*, they might meet the medium-low requirement with only three stores, and may choose to re-evaluate their original requirement and accept a lower performance level in order to meet both risk and cost criteria.

Figure 2 shows a second non-functional requirement, in this case meeting legal obligations. Here the consequences of losing the operating licence or of being sued by a large company create very high risks that need to be mitigated by both data replication and increased security. On the other hand, if the risk is being sued by an individual, the expected risk (cost times probability) maybe low enough not require further mitigation (although the individual whose data is lost may not agree). Figure 3 provides another illustration of the risk graph for data complexity. Here the risk arises from data formats becoming unreadable as the software goes out of date or if the software developer ceases trading.

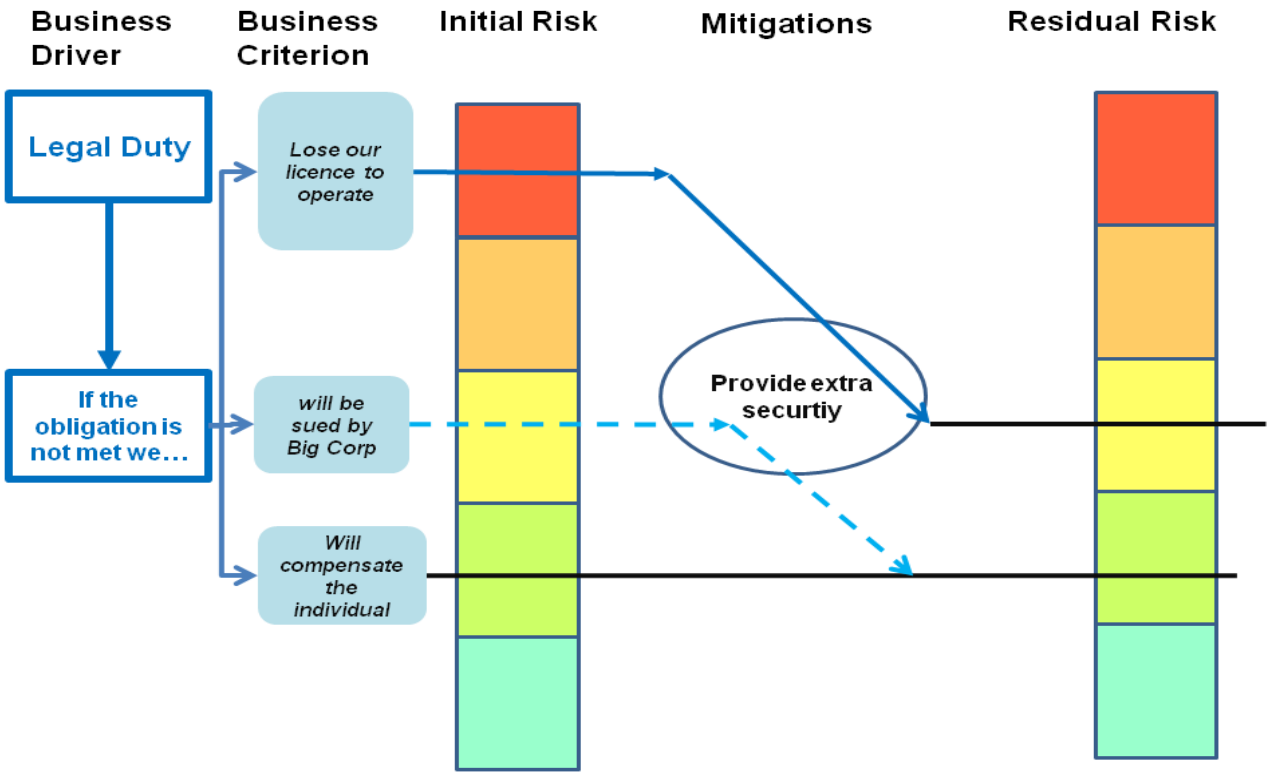

Figure 2: Non-functional Requirements and Risk Mitigation – Legal Duty

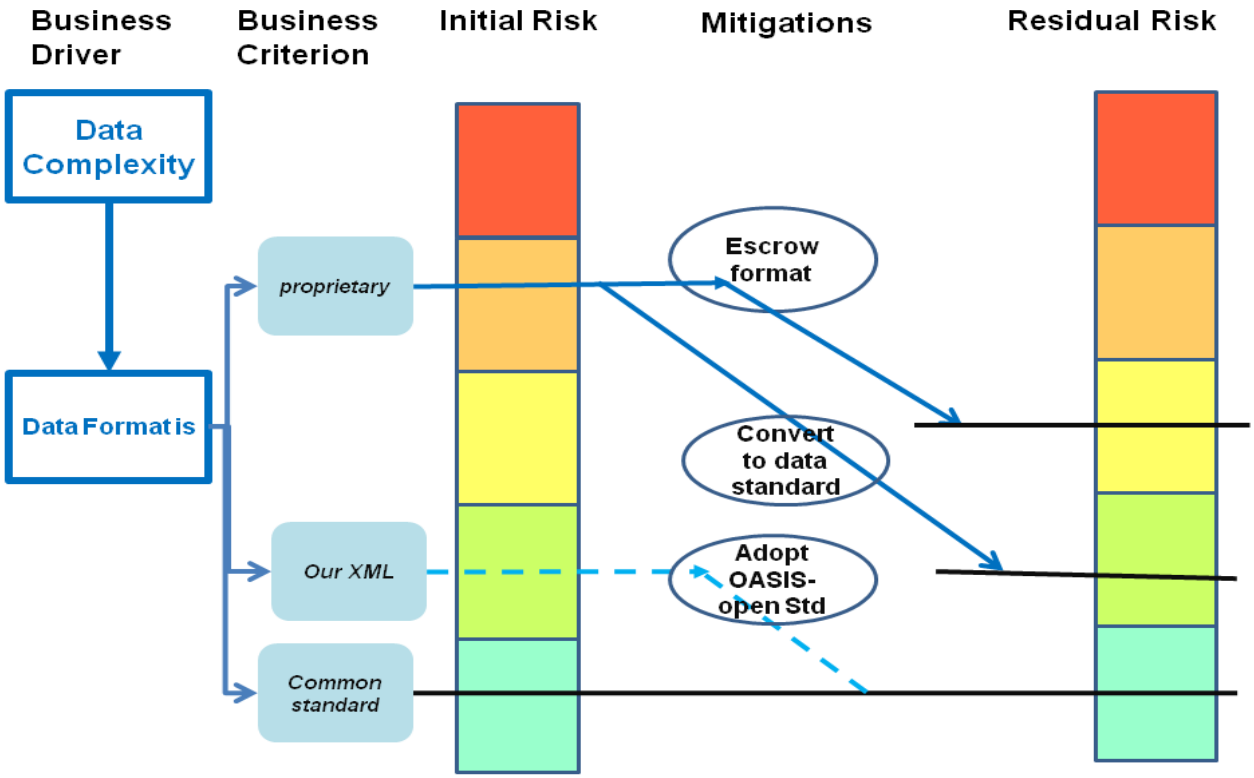

Figure 3: Non-functional Requirements and Risk Mitigation – Data Complexity

Implications for Cost Modelling

The worst possible outcome for the 4C project is that every cost model will be need to be unique in order to tailor it to the various requirements and risk appetite of the user. However, in the three examples shown, there are recurring features, such as the need to replicate the physical repositories or proved enhanced security.

The RASSC model made the assumption that there are common features that can be factored into separate services, and that the repository would be built up by selecting the services needed. The aim behind this approach was not cost modelling as such, but rather looking at the prior problem of how an economic ecosystem for long term sustainment might be created. In particular, one implementation use-case assumed that initially the preservation of the uninterpreted data would be outsourced to a commodity storage supplier such as Ovation (one of the partners in RASSC), noting that there are already a number of suppliers for this service (e.g. Ovation, Archivum, Tessela). Commodity storage services have an additional advantage that the costs of understanding how to store data – understanding media migration, technology watch on storage devices, etc. – and the costs of facilities for multi-site replication would be spread out across multiple users, rather than the user company having to develop and maintain the expertise and facilities in-house. However, the use-case assumed that the user company would need to provide both information and knowledge level services for data such as CAD models because there are very few organizations operating in its business segment in the UK.

The RASSC project considered the possibility that the market for repository services would develop into a set of horizontal service sectors, much as the Cloud Computing sector has. The cost models identified in the 4C project appear to have been developed for a vertically integrated organization that keeps all curation costs and activities in-house, much as might be expected given that many of the institutions involved in 4C are specialised in the permanent sustainment of data created by other organizations. Moreover, since the RASSC situation does not yet exist, it would not be practical to build cost models based on it.

The contention is, however, that the existing cost models provide a workable basis for cost estimation only if one can identify which model one wants. This is because the cost models available embed a set of non-functional requirements and risk mitigations deriving from the assumptions of the institution that created that model. It is hypothesised that the costs of active curation of data dominate over IT costs, although this comment is based on anecdotal evidence rather than experiential.

Implications For Cost Model Definition

As an intermediate step to developing service-based cost models, it is suggested that there will be a series of use-case based cost models, based on an indistinct "quality of service", much as the Myles na gCopaleen “book handing” service ran from basic opening a book a few times through *dog-earing and inserting used railway tickets as book marks* to the deluxe service including *underlining and marginal remarks and the inclusion of a signed programme from the Abbey Theatre in Dublin as a lost bookmark*. In data sustainment, the equivalent qualities would include long term guarantees of data integrity, the level of indexing and accessibility provided, and the diversity of the collection in terms of the data formats supported. The aerospace industry requirement would be for a high level of integrity and evidential weight, a wide diversity of specialist formats (including mechanical and electrical CAD, product structures and Finite Element Analysis), but where the data would become obsolete between twenty and a hundred years after its creation.

The concept proposed is therefore of a series of cost models, represented in figure 4 as an example series of curves of cost against planned sustainment duration. Each curve is defined for a quality of service, and a meta-model provides guidance of when one needs to transit from one curve to

another. In figure 4, the bottom curve represents a simple cost-per-unit-of-storage v. time, and the upper curves represent three different levels of service: one providing three separate repositories, one providing four repositories, and one providing additional data format migration services to keep a collection accessible.

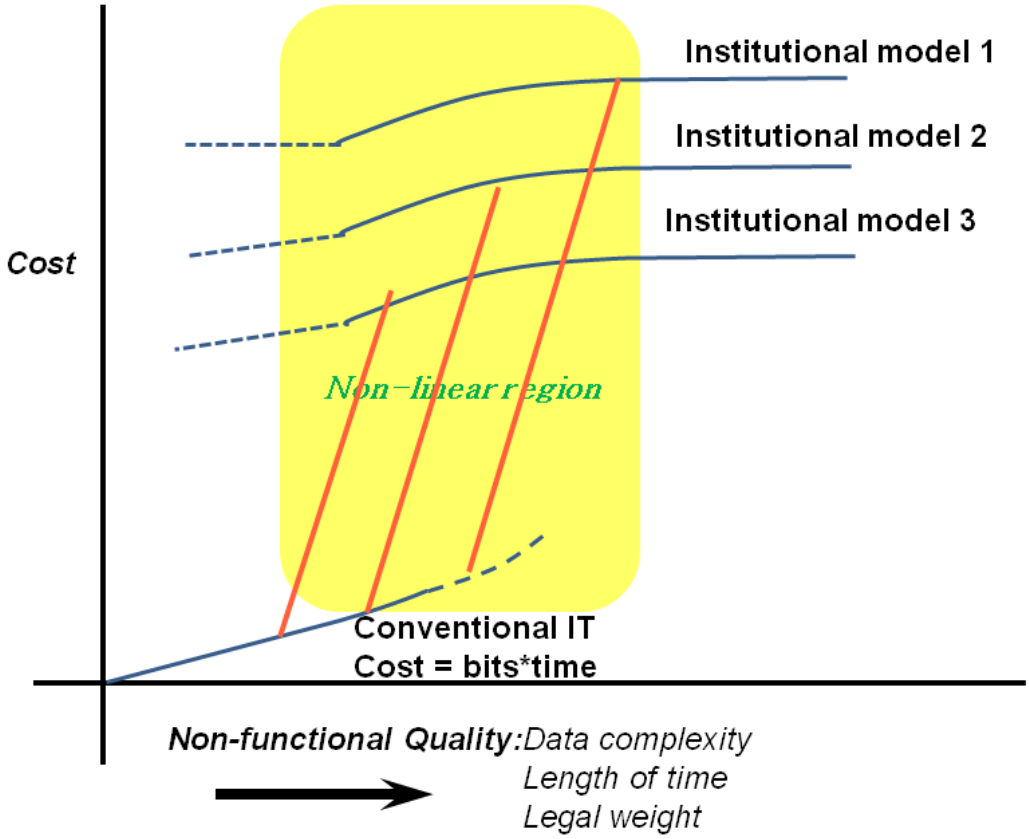

Figure 4: Transition between Cost Models

Cost Model Usability

Cost modelling for data sustainment is likely to be intrinsically complicated, both because of the new technological issues involved and because of the varying institutional approaches to risk. At least in industry, the current problem in data sustainment is less about accurately costing sustainment and more about making the users aware that it is not an IT problem but rather is driven by corporate policy and governance. One should therefore propose a number of risk scenarios to the user and find out which of the cost model available is most like the one that matches their requirements and risk appetite. It is only when the user understands that the cost model is sensitive to corporate requirements that they are likely to put sufficient time and effort into identifying the cost model they need.

The user must also be appreciate that cost modelling for data sustainment is at an early phase of its development, and therefore will require a level of expertise and knowledge to use – there is as yet no Mrs Beaton of cost modelling for data retention. It is probably a fair complaint that cost modelling tools appear complex, but the problem is one of developing user understanding as to why they are complex rather than expecting everything to be done with an intelligent user interface. As

with mathematics text books , the difficulty lies not in the sentence structures used, but in the concepts they are trying to express.

In the longer term, specialist cost estimation services will be needed – that is, people with enough knowledge of the complexities of curation to ask simple enough questions to allow costs to be estimated. For example, in the UK it would be unreasonable to expect that every PhD student should know enough about data curation to keep their experimental data live as long as the research councils mandate. It would make more sense to provide data sustainability advice to ensure that data is stored in a form that is easily sustained, and for which there is a known cost model.

Summary and Conclusion

The hypothesis is that the costs of mitigating perceived risks is the major determinant of long term data sustainment, and that these risks arise mainly from non-functional requirements such as data integrity and evidential weight. Consequently, it will be difficult to read across any cost model developed for a particular vertically integrated organisation to any other such organization unless they have the same approach to risk management.

It is suspected that some of the major risk mitigations, such as storage replication, may be applicable to multiple risks and generally applicable across a range of data types, while others, such as audit trails or validating CAD translators, will have a much narrower range of applicability. It is recommend therefore that work on cost collection should be supported by collecting information about institutional and organizational approaches to risk, and identifying the links between risk and the sustainment solution adopted. With sufficient data it should then be possible to test the hypothesis.

It would also be useful to examine a service-based approach to cost modelling. However, this area is yet to be well exercised and as yet only the first steps have been taken towards service-based sustainment. If such an approach is valid and viable, this should lead to a meta-model of services and costs which can be used to tailor a cost model for a particular institution, company or business sector. The availability of such a cost modelling approach may also encourage the growth of an information sustainment service sector.