Institutional Data Management Blueprint

1. Introduction

Research data management has become an international topic of concern to researchers and their funders. Data integrity, validation and security, access, curation and preservation are now all recognised as integral elements in ensuring the quality and value of research in almost every discipline. Defining the responsibilities for managing research data from inception to preservation is also clearly recognised as a complex process shared between individual researchers and research groups, institutions, funders and national agencies.

The demand for effective research data management is driven by many agendas, including those of different funding agencies and programmes, national policies, technology trendsetters and the researchers themselves. A constant factor is the institution - a centre for cohesion, curation and cooperation - which is responsible for its own research data for some, or maybe all, of its lifetime, within a fragmented and volatile world. In order to acknowledge and manage these responsibilities, institutions require an overall framework within which to plan and develop their data management strategy.

As both the total amount of research data produced expands and the complexities of the research landscape increase, data management strategy requires a multifunctional team approach which can bring together the knowledge and expertise of both researchers and professionals within an institutional policy and technical framework. The University of Southampton has a proven track record in creating a team approach to managing research outputs evidenced from the extensive work with research and learning repositories. This Blueprint extends this model to an institutional approach to research data management.

The focus is on making the research process, and thereby the effectiveness of our researchers, easier, and is intended to be researcher-led. To meet the needs of a multi-disciplinary, research-intensive University, it is designed to be both practical and iterative. It will reference policies, guidelines and examples of good practice, and will aim to exploit open standards and service-oriented approaches. It will form the strategic context for developing research data management structure based on business planning.

The Blueprint will be jointly owned by the Research and Enterprise Advisory Group, responsible for University research strategy and the University Systems Board responsible for systems strategy.

2. A Researcher-Led Approach

The University already has guidelines for good practice for the research process:

Throughout their work, it is good practice for researchers to keep full, clear, and secure records, whether in paper or electronic form, of their procedures and results, including interim findings where applicable. They should include accurate and contemporaneous records of primary experimental data and results, in a form that

will provide clear and unambiguous answers to questions concerning the validity of data later. This is necessary both to demonstrate good research practice and to answer subsequent questions.[1]

The researcher survey and the AIDA audit undertaken as part of the IDBM project highlighted that researchers were not clear as to legal, policy and budgetary responsibilities, and that, while best practice was shared at local level through local support networks, the relationship with the institution in terms of rights and ownership was not understood by all. For the individual researcher there was little central information or advice perceived to be available, particularly for their higher-level questions. There was also a disconnect between the support made available by the institution, and what researchers think is available.

Getting this balance between the responsibilities of the researcher and the institution is important at a time when research funders are increasingly requiring evidence of effective research data management, and the pressures on institutions to both manage data and make it available externally are continuing to grow.[2] Experience also shows that in the area of industrial research collaboration, managing and sharing research data appropriately increases the attractiveness of proposals, but there are still significant barriers to making this effective at both a technical and cultural level.

It is clear from the audits that the infrastructure to deliver this is more than what is offered by technology, the equipment, software, hardware, and skills to maintain the digital data management environment and to respond to the changing opportunities offered by technology. It includes the organisational infrastructure, the policies, procedures, practices, and culture fostering the framework for successful data management, and the necessary resources, funding and human resources to deliver the programme.[3]
The audits reiterated that researchers are best incentivised to take up new practices and processes by having a clear policy and service framework for managing their data.

This Blueprint therefore combines two approaches. A bottom up approach based on researchers' needs, designed to enable them to adopt good practice, and a top-down approach, designed to provide the institutional policies and infrastructure to be effective.

---

[1] http://www.soton.ac.uk/ris/policies/integrity.html

[2] The Data Curation Centre has identified two core requirements for funders:

1. Research outputs are created in an appropriate manner to ensure that they can be made widely accessible.

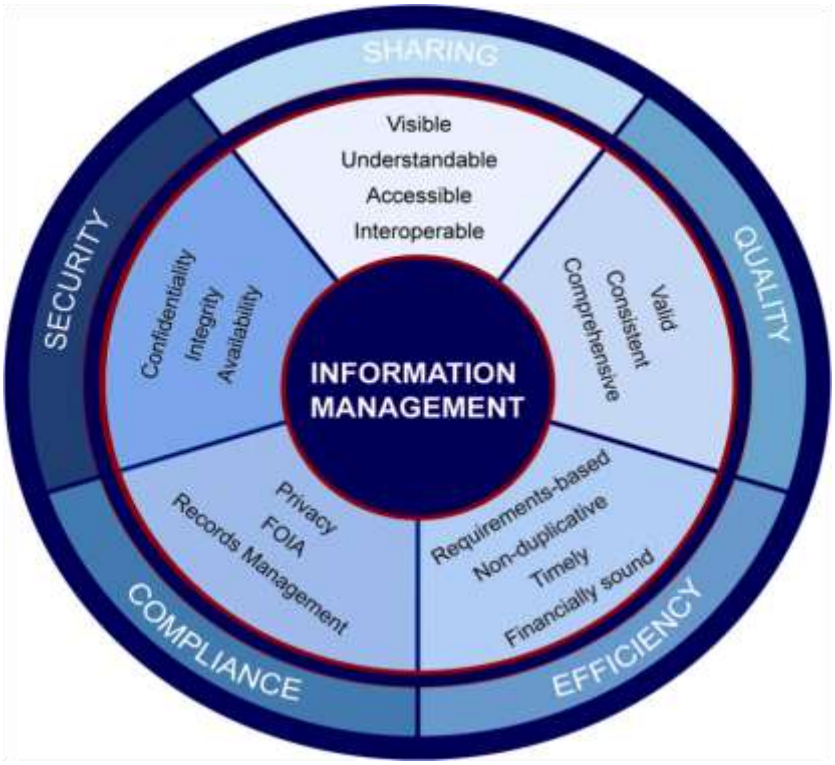
2. They are maintained in the long-term to facilitate future access, either under the auspices of the institution in which the funded researcher is based or by means of deposit in a special repository or data centre.

[3] This is adapted from the Cornell 'three legged stool' model for digital preservation, http://www.icpsr.umich.edu/dpm/dpm-eng/tutorialprint.pdf

3. Organisational Policy Framework

Institutional policies and procedures form the context within which this Blueprint can be implemented. The draft Data Management Policy [4] outlines the assumptions behind institutional and researchers' roles, and the commitment by the institution to provide the resources to deliver the policy. It provides researchers with guidance on what is expected and how to manage their data, and it helps the institution to define what is required to manage institutional assets and comply with funders' requirements. It also provides a governance and decision-making framework.

The Data Management Policy will form part of the policy framework for information management at Southampton. It references IPR policy, and will in due course reference Information Security Policy and the Information Management Principles which provide a total framework for the management of Information at Southampton:



Providing a model for implementing these policies is one of the aims of the Blueprint. The AIDA audit revealed that in some areas research practice is embedded and unified whereas capabilities in others varied widely with management being carried out on an ad-hoc basis. Policy and governance at institutional level is not communicated to researchers in the most accessible way.

The Blueprint will provide an enabling framework and more specific guidance will spring from its principles to help build on local good practice. It will be jointly owned by the Research and Enterprise Advisory Group, responsible for University research strategy and the University Systems Board responsible for systems strategy. The key link at senior management level will be the Pro-Vice Chancellor for Research and Enterprise.

[4] See Appendix A

4. Organisational Investment in Storage and Technological Capacity

The AIDA audit indicated that most data management capability tends to be localised, and that capabilities vary locally, with pockets of both best and limited practice throughout. Knowledge of available capability and resources were limited, and researchers do resort to their own best efforts using USB hard drives and other time-limited devices in default of any expectation of institutional provision.

The first steps to address this are already being taken. iSolutions is currently carrying out a review of data storage, and the first tranche of investment, £3m over the next two years with a further £1.5M every 2 years, has been agreed by University Systems Board. The commissioning of a major new datacentre in 2013 will dramatically increase the amount of potential data storage available, allowing each researcher, including PhD students, an automatic right to a set level of resource. This technological investment is in part a response to RCUK requirements, and is focused on data storage both during the life of the project and subsequently to meet RCUK access policies.[5] The long-term storage of data is a significant issue for the University, not only because of the cost of managing effective curation but because most researchers perceive that they need to keep their data forever.

The University will explore four approaches to containing the costs of long-term storage:

1. Balancing the amount of data to be stored with expected reductions in the overall costs of storage;
2. Investigating economies through the use of shared services, including access to cloud storage expected;
3. Incorporating cost charging into bids for new projects requiring data storage over a certain level;
4. Incentivising academic groups to review their data through a recharging model for data costs over a certain level.

Ideally the first two options will be able to meet increased demand, but the second two options require to be scoped to provide balance in investment.[6]

5. Organisational Investment in Metadata Standards and Researcher Support

Data storage is only one element in building the infrastructure. Researchers are interested in front end applications to ease the process of deposit, updating and access, and in the availability and easy adoption of metadata standards. The AIDA audit confirmed that researchers would value help in organising their research data, that they require effective access, both access to other researchers' data and access for others to share their data. Most researchers share, or would like to share, their data, although some would like effective authentication and access control.

Researchers therefore require at institutional level an effective way of creating metadata and ingesting their data; they need a registry of that metadata to facilitate

[5] An outline of funders' data policies is available from the DCC website at http://www.dcc.ac.uk/resources/policy-and-legal/funders-data-policies

[6] See Appendix B for the draft business model.

access and retrieval set by default to open data standards, and they need advice and support in managing the process. All these elements are also prerequisites for the institution to be able to assess data storage requirements, and for those data which will ultimately pass to national research data centres to be identified, and transferred. In defining an institutional framework it is important to build upon existing good practice, enhancing and enlarging the present level of activity to encourage all researchers to engage.

The IDBM project defined a core metadata structure which can be used to provide outline data for research projects and to provide the basis for the registry. This schema is based on Dublin Core as an appropriate standard for an institution-wide metadata framework in line with the approach already used by the National Crystallography Centre at Southampton through eCrystals, and tested in the IDBM project for use in archaeology.

The concept behind this simplified metadata scheme is to encourage people to tag their data by employing a system which is not onerous supported by usable tools for metadata assignment and import and provenance tracking. The first step is to embed good metadata practice across the disciplines and to make it easy for researchers to submit core data. To this end the existing institutional Eprints service will be extended to enable researchers to tag and submit their data, creating Eprints4data. This builds on the existing requirement for researchers to submit their research outputs to the Eprints service, and provides the basis for potentially cross-linking research output with data as is already offered by the eCrystals service at Southampton. [7]

This additional demand on the Eprints service will require investment, but we will also need to be mindful of the long-term capacity of the software to match demand. In parallel, therefore, the University will continue the current pilot work with Sharepoint 2010 to assess how SharePoint 2010 can be used, and extended, to manage data and workflow in a seamless way, and enable users to publish their data with full access control.[8]

This is a complex process. To support researchers in their own understanding and development, there should be a partnership approach between all stakeholders, senior management, researchers, iSolutions, the Library, Research and Innovation Services, Finance and Legal Services. We need a support structure which reaches across the whole research lifecycle and across the spectrum of research careers, from the PhD student and the early career researcher to the mature research group within large scale national and international networks. The experience from the

[7] http://ecrystals.chem.soton.ac.uk/

[8] http://eprints.ecs.soton.ac.uk/21233/

archaeology pilot shows that this can only be effective when delivered through a multi-functional team which is integrated into the institutional data management model.

6. Roadmap

This Blueprint is intended to be part of an iterative, dynamic model for supporting data management at Southampton. As part of the IDBM Project, the Project Team identified three phases of what was in effect a roadmap for delivering effective data management infrastructure.

Short-term (1-3 years)

Crucial to supporting researchers is a higher profile role for the institution in developing the core infrastructure. This requires an integrated approach to policy, technical infrastructure and support which can meet the demands of the growth in the level and complexity of research data, the requirements by funders and the need for the institution to manage its digital assets effectively.

In this phase the infrastructure must have sufficient capacity and be affordable/free to attract users, rather than forcing them to develop/procure local solutions. This should be piloted, and then grown over the short-to-medium term.

The core components for this phase are:

- A robust institutional policy framework which is agreed and implemented by the institution.
- An agreed scalable and sustainable business model for storage based on the three components of active data, descriptive metadata and archive storage.
- A working institutional data repository at institutional level which can satisfy the majority of researchers" data management requirements for ingest, metadata creation and retrieval, and be extensible. It must have sufficient capacity and be affordable/free to attract users, rather than forcing them to develop/procure local solutions. This should be piloted, and then grown over the short-to-medium term.
- A one-stop shop for data management advice and guidance to provide information on policy, legal issues and guidance, so that they can rapidly create data management plans, access advice on technical capability, understand funder requirements and the benefits of managing data to exploit and share, and developing their skills and own practice.

During this phase it is assumed that there will be close engagement with disciplines acting as early adopters, but sufficient institutional profiling to take principles forward at institutional level.

Medium Term (3-6 years)

During this phase it is assumed that the demands for the management of very large amounts of data of increased sophistication and complexity will increase, and that some disciplines will require potential high levels of data management input than can be managed within one institution. Although the cost of storage is likely to continue to decline, the management process itself will increase demand on staff skills. There will also be a higher profile for open and shared data, and the value of pooling and sharing between institutions will be explored through specific exemplars.

The core components for this phase are:

- An extensible research information management framework to respond the variations in discipline needs.

- A comprehensive and affordable backup service for all, but one based on clear organisational thinking about the cost-benefits of backing up different classes of data.
- An effective data management repository model able to manage the potential full range of data deposit.
- Building an infrastructure to respond to a commitment to open research data creating a model for data publication.
- Based on the cost-benefit analysis for backing up different classes of data providing comprehensive solutions for managing research data across its whole lifecycle.
- Embedding data management training and support across the disciplines through partnership working between services and researchers.
- Pilots with consortia to manage data collectively using standard infrastructure applications including cloud computing, and shared staff knowledge and expertise.

Longer term (6-10 years)

Long-term aspirations will focus on providing significant benefits realisation across the whole University and a stable foundation for the future. The institution would have policies and infrastructure in place to make strategic judgements on how to manage its digital assets, and would have moved to a mixed-mode of data management within consortia or national framework. There would be a higher level of partnership between funders, organisations, local consortia and national facilities. Data management processes would be embedded throughout the research data lifecycle, and the infrastructure would fully support researchers with supply meeting demand via an easy-to-use data management service. This would significantly improve research productivity, allowing them to concentrate on their research, rather than worrying about data management logistics.

The core components for this phase are:

- Coherent and flexible data management support across all disciplines across the whole data management lifecycle.
- Agile business plans for continual improvement in response to changing requirements and technology changes and new business models evolve.
- Strong commitment to innovation in open data publication and the infrastructure to support this across the institution.
- Active participation in consortia and national framework agreements, contributing capacity and skills to building overall capability.

Authors: Mark Brown, Oz Parchment and Wendy White
Date: 31 August 2011

Appendix A Draft Data Management Policy

Introduction

The purpose of this policy is to: (i) create model research data management practices for all staff and researchers within the University of Southampton (the University); (ii) foster responsibility for such research data management through the development of research data management plans; and (iii) ensure that research data is stored, retained, accessed and disposed of securely in accordance with all legal, statutory, ethical, contractual and funding requirements.

A robust research data management policy is required to demonstrate and ensure:

- good research practice and procedures
- protection of intellectual property rights (IPR)
- proper recording, maintenance, storage and security of research evidence and results
- compliance with relevant legislation and regulations
- appropriate access to research data.

For the purposes of this policy 'research data' is considered to be all data generated in the course of the research process in both raw and analysed form.

The Policy

The University is committed to achieving the highest standards for secure research data management and recognises that this is a shared responsibility between the researchers, Faculties and data owners. This policy addresses the following areas relating to data and research materials: (i) ownership and intellectual property rights (IPR) (ii) storage and management (iii) retention (iv) access (v) disposal and destruction (vi) exceptions.

Ownership and IPR

The researcher should clarify issues surrounding ownership of and rights to research data at the outset of a project. However all research data created at the University will be subject to the University's Intellectual Property Regulations – http://www.calendar.soton.ac.uk/sectionIV/ipr.html

Where research involves external funding and/or collaboration with other institutions or external parties, IPR ownership and rights should be explicitly identified and documented in the relevant funding agreement or contractual documentation prior to commencement of the project.

Where a research project involves usage of data owned or controlled by a third party, regard must be had to any relevant laws, regulations and/or restrictions on use - e.g. copyright law or conditions of licensing agreements.

Further useful guidance with regard to the University's IPR and research related policies can be found at http://www.soton.ac.uk/ris.

Storage and Management

The University recognises the importance of providing safe and secure storage of research data to protect against damage, loss, misuse or theft during and beyond the research project. However, it is the researcher's responsibility to ensure that all research data is stored securely in a durable format appropriate for the type of data and with adequate metadata and/or documentation to facilitate identification. The researcher should also ensure that the data is backed-up regularly and logged in the Eprints Soton Research Repository (the Repository) -http://www.soton.ac.uk/library/research/eprints/policies/oapolicy.htmlt. Non-digital data unsuitable for digitisation should be identified and documented.

Effective data management should include exit planning. Each Faculty should establish procedures to deal with circumstances where researchers leave the University or withdraw from a collaborative project. Such procedures should ensure that all exit arrangements are fully documented and take proper account of any relevant requirements – e.g. legal, funding or ethical.

Retention

All research data should be held for a minimum recommended period of 10 years, subject to relevant regulatory/legislative requirements and/or guidance. However, the requisite period of retention can vary according to the discipline and type of research, as well as any contractual/funding/sponsorship/internal policy requirements.

Permanent or long term retention may be advisable where research has, for example, a public interest or heritage value – or where outcomes may be contentious or subject to challenge. In such cases data should be retained pending review and not destroyed/disposed of until the matter is fully resolved.

Disposal and Destruction

Disposal of research data should be undertaken in accordance with the University's recommended practices to ensure secure and safe destruction. The agreed processes for the timing, manner and recording of research data disposal should be included in data planning and stored with other project information and documentation.

Prior to any scheduled disposal, the research data records should be reviewed and authorised for destruction by the appropriate University Faculty or the data owner (where the University is not the owner). Disposal shall be managed in line with any regulatory and contractual obligations, and in accordance with the sensitivity of the data in question.

Research data must not be disposed of without written authorisation and the destruction process must be fully documented.

A record of the deletion of data should be logged in the Repository and include the reason and authority for deletion.

Access

The University recognises the benefits of making its research data accessible to the wider academic community. However, before sharing data during or after a project it is essential to consider the implications of doing so – e.g. in terms of IPR ownership, any ethical/privacy/confidentiality requirements or any legal/regulatory/funding restrictions.

Exceptions

If a researcher and/or Faculty considers that any research data needs to be dealt with in a manner outside the scope of this research data management policy, the matter must be referred to either the Dean or Associate Dean (Research) for the appropriate Faculty, or any individual or body (e.g. Ethics Committee) authorised to make exceptions on their behalf. The only exceptions permitted will be those approved by these authorities.

Policy ratification and implementation

This draft policy for the University of Southampton has been produced as part of the JISC funded Institutional Data Management Blueprint project with advice from Legal Services. We gratefully acknowledge the model provided by Monash University http://policy.monash.edu.au/policy-bank/academic/research/research-data-management-policy.html .
This policy will now go to the Research and Enterprise Advisory Group, chaired by the Pro-Vice Chancellor for Research, for final consultation with the Faculties and will be tabled for approval at Senate by end 2011. It will be made available with a series of one page guidance sheets on specific topic which are expected to grow over time in response to academic requirements. We expect to launch four key guidance sheets with the policy.
Data management roles and responsibilities
Advice on disposal and destruction
Storage options
Managing licenced data sets
Promoting the benefits of the policy, developing further guidance and embedding good data management practice will form the next phase of the institutional data management roadmap.

Appendix B Business Model

Crucial to the successful deployment of an institutional data repository is a reasonably clear understanding of the business and cost models for the associated infrastructure, whether that is people or technology. In providing this model we need to arrive at an understanding of a number of areas or issues,

- The elements included in the model
- The current and future University research data landscape
- Cost modelling

The elements included in the model

The business model is applied only to the IT services delivery of an infrastructure (technology & people) to deliver and sustain an institutional repository for the University's digital assets.

The model will be based on a putative high level architecture with indicative order of magnitude costs only. Further refinement will require rigorous market testing.

The basic assumption will be that "**the University wishes to provide a secure and sustainable repository capable of hosting the University's entire digital assets"**.

Current and future needs for the University's research data

Whilst the IMDB survey goes someway to providing insight, the University in common with many, if not all, of its peer institutions has no detailed knowledge of

- **The quantity of research data that it holds, or**
- **The growth in research data in any specific timeframe**.

Nevertheless some indicative data is available from the researcher survey and from the University's current mid-scale research storage platform to allow estimates for both of these quantities.

Estimating the University's digital research assets:

The University today offers ~200TB of secure storage for research data. This service, while not universally subscribed to, by our research community currently hosts ~120TB of research data.

The IMDB researcher survey suggests that the University centrally hosts some 10-15% of the research data of the surveyed researchers. If representative of the University as a whole, we can estimate that the University digital research data assets are of the order **0.8 – 1.2PB**.

Predicting the growth in digital research assets:

The rate at which new research data is being generated is more difficult to estimate. Various market research firms (see for example IDC 2008) estimate that file-based or unstructured data have a compound annual growth rate (CAGR) of 62% compared to 22% for structured or transactional data.

However, it is not clear if research data, can be simply aligned to either of these definitions. A significant and growing percentage of research data is generated through, for example scientific instruments or computers (super or otherwise).

It might be more appropriate to consider research data in a similar manner to "content depots" [IDC 2008] (e.g. Google, Flickr and YouTube) wherein the combined effect of millions of subscribers is similar to the instrumentally rich nature of modern day research environments. It has been suggested that for this category, the CAGR's is as high as 121%[9] ["Storage is cool", Henk Wubbolt, IDG Storage World, 2009, based on data from IDC 2007]. However, more recent data[10] [IDC's 2010 Enterprise Disk Storage Consumption Model] suggests the CAGR is 76%.

Early indications from Southampton's mid-scale research data platform suggest that the CAGR for research data is in the region of 170% but as noted this service is not yet fully subscribed by the University research community and the user base is yet to reach a steady state.

While we have to be careful in our treatment of these numbers they do provide some guidance as to possible bounds to the future requirement.

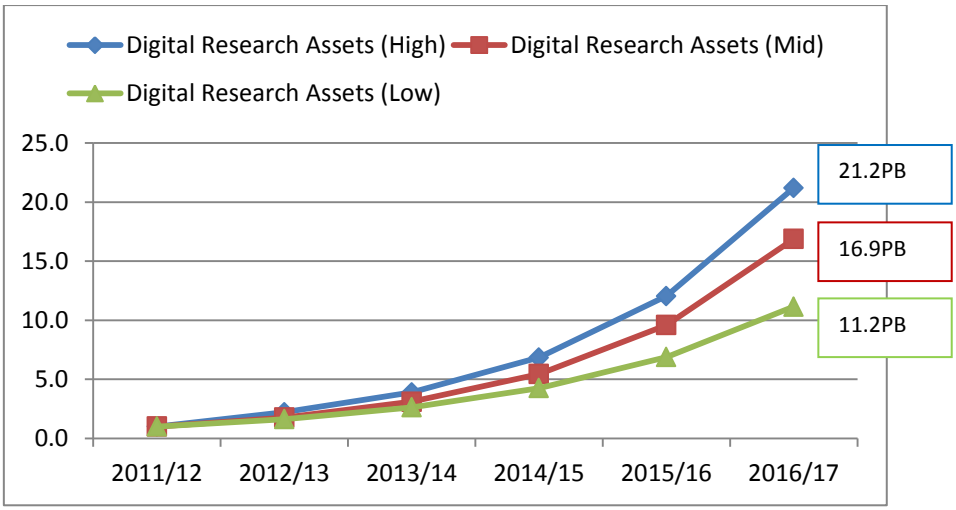
Table A: Estimated Growth in Research Data Assets

	%CAGR
"Content Depot" (**High**)	121
"Content Depot" (**Mid**)	76
"Unstructured Data" (**Low**)	62

[9] "Storage is cool", Henk Wubbolt, IDG Storage World, 2009, based on data from IDC 2007
[10] IDC's 2010 Enterprise Disk Storage Consumption Model

Using this data we can estimate the possible bounds for the storage and archiving requirement for the University of Southampton into the future. The estimated growth is only displayed over a 5 year time frame for clarity.

Figure A: 5yr Predicted Storage Requirements (PB)



While the total digital research assets of the University are, today (**year 0**), estimated at ~1PB, HIGH case (121% CAGR) estimates the University requirement at **12PB** by 2016/17 (**year 5**) rising to **~0.36EB** in 2021/22

We will further consider, only the "mid-case" scenario for cost modelling which estimates the University requirement for storage and archiving of to be ~**10PB** in 2016/17 rising to ~**0.29EB** in 2021/22.

Cost Modelling

In modelling the costs associated with an institutional repository we have already defined the purpose of the repository which effectively translates into "keep everything forever". Although there is clearly a downstream opportunity for the University to define what is a "digital asset", and to refine this in terms of the Data Management Policy referencing to disposal and destruction of data.

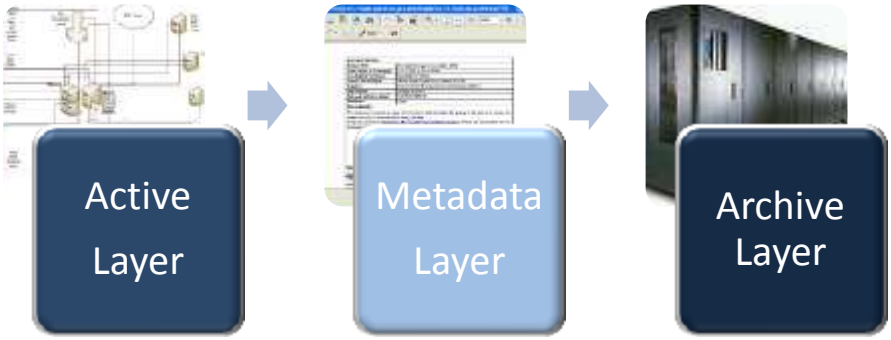
We have already estimated the envelope of infrastructure required to deliver the repository based on the mid-case scenario. However, three further elements are needed,

- Conceptual architecture,
- Staffing and
- Facilities.

High Level Architecture

It is possible to create a simplified model of the institutional requirement and thus to provide an initial IT infrastructure in order to obtain a high level architectural model with which to inform a business and cost model.

Figure B: Components used in the high level architecture



At its simplest, the proposed architecture provides several components

- **Active Storage.**
  - Data is created onto or received into the active data layer from instruments, manually or from downloaded content.
  - The active layer contains secure, resilient storage for working data sets, highly available and accessible via various network protocols, e.g. NFS, CIFS/SMB, HTTP(S), SFTP etc.
- **Metadata layer**
  - This contains descriptive information on archived data sets.
- **Archive Storage**
  - Here the data is preserved and secured for the long term.

It can be seen that the active storage layer enables the creation or receiving of and transformation of primary or secondary data whether that be from surveys, instruments, computers or other data sources or archives. Movement of data from the active storage layer to the archive layer is via the metadata layer.

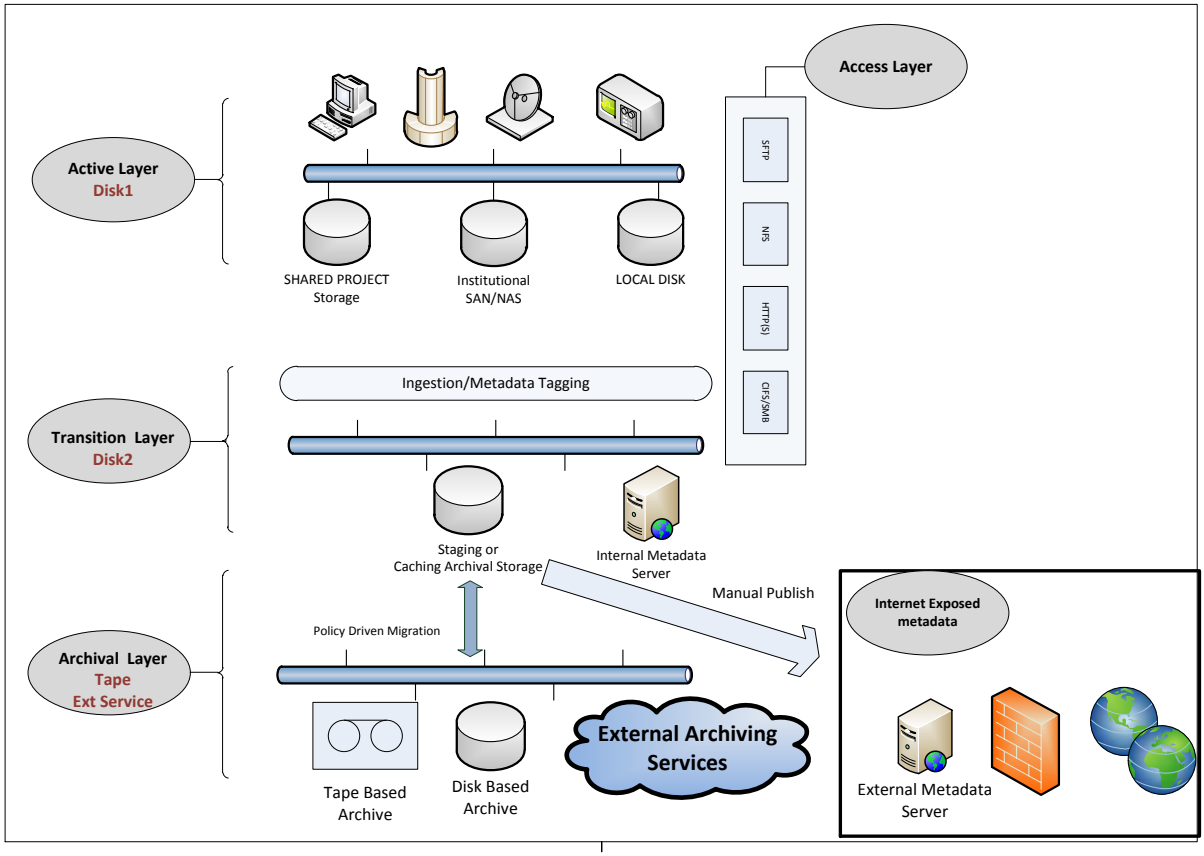
In considering the high level architecture, a small number of initial questions need to be addressed, with regard to availability, accessibility and the required performance. These questions would normally be addressed by understanding the research workflow that the architecture is created to enable. On an individual or group research level, this is a straightforward task, on an institutional level many more broad assumptions need to be made in order to provide guidance on an initial solution.

Assumptions

- The active storage layer will consist of both localised and institutional resources,
- Localised resources will be used where it is impractical to use institutional resources.
- Institutional resources will require a high level of availability and accessibility.
- Access to the archive layer will require the deposited metadata to be appropriately tagged.
- Separate metadata servers will be required, for externally published metadata and for internal use.
- Access will be required through numerous network protocols to both the active and transition storage layers but only for internal viewing.

- Movement between the active layer and the archive layer should be available in both directions.
- Use of external archive and storage services are directly available from the transition layer.

Figure C: Architecture for Institutional Archive for Research Data



Staffing

The cost model will consider only those additional staff deemed necessary to manage the additional IT infrastructure associated with the University repository. A small IT team is envisaged of 4FTE, consisting of storage and server administrators and service management. Clearly there will be staffing needs in other areas e.g. data management, and potentially software development, however neither of these are included. In addition we need to consider staffing requirements in the start-up costs, i.e. business case development, project management, training and consultancy. With the exception of training which is an operating expense the remaining areas are treated as capital expenditure associated with the initial roll-out or major upgrade of the repository.

Facilities

Storage is quickly becoming a major consumer of space, power and cooling in enterprise data centres, this is likely to accelerate as with increasing data requirements and server virtualisation. The space requirement of an institutional repository capable of hosting the University's digital assets will have a significant footprint within a data centre.

Operating Costs

The University's own figures[11] [University of Southampton Data Centre Options Review 2008] for its current data centre, estimates annual operating costs of ~£0.5M per annum for ~300 M^2 of data centre space or £1667 per M^2 per annum including electricity costs.

The Uptime Institute[12] [A Simple Model for Determining True Total Cost of Ownership for Data Centers, Johnathan Koomey, 2008] suggests for a Tier III[13] data centre the cost is ~£9000 per M^2 per annum.

The University has a facility which was originally built in 1975, the operating costs are therefore not necessarily appropriate today. A Tier I or Tier II facility typically seen in Higher Education institutions would attract a premium of ~£3500 per M^2 per annum.

The bulk of these costs are attributed to electricity costs, both IT equipment and facilities, and contribute ~60% of the operating costs of a data centre. The estimated power consumption of the various IT equipment is available, and the facilities use of electricity is modelled using a conservative estimate of the data centre power utilisation efficiency (PUE) of 2.0. As the electricity costs associated with the IT infrastructure will be included explicitly a figure of £1400 per M^2 per annum will be used to cover the miscellaneous operating costs of the data centre.

Capital Costs

As with operating costs, the capital costs are heavily dependent on the level of availability or Tier rating of the data centre and on the size and location of the facility and indeed on the state of the construction industry at the time of the build. While some variables are relatively straight forward the latter components are much more difficult to estimate. Therefore we will not consider this component further, but it should be easy to add to the model if known.

Cost Modelling

For the overall architecture shown in in **Figure C**, the cost model has both capital and operational elements and a number of assumptions associated with each element of the cost model, these are explained under the relevant headings in the calculator spread sheet.

Two architectural models are shown for comparison. **Scenario 1** is primarily tape based using a ratio of 3:1 to assign data either to tape or disk. **Scenario 2** is wholly disk based, a ratio of 9:1 used to assign data either higher performing disk (disk1) or lower performing disk (disk2). In both scenarios an allowance is made for an offsite copy of the data to be held. This would be similar for both scenarios assuming a tape based clone was assumed for both scenarios.

One of the key assumptions in the model is that storage costs will reduce significantly over time associated with the rapid increases in storage density e.g. Kryders Law[14]. Estimates vary as to the rate of reduction in storage costs over time, figures from 15% to 40% per

[11] University of Southampton Data Centre Options Review 2008
[12] A Simple Model for Determining True Total Cost of Ownership for Data Centers, Johnathan Koomey, 2008
[13] Data Centre Infrastructure …….
[14] Walter, Chip (July 2005). "Kryder's Law". *Scientific American*.

annum are routinely discussed[15] [Storage Economics, Four Principles for reducing Total Cost of Ownership, David R. Merrill, June 2011], which would lead to a £1M difference in the 5 year costs of scenario 1. We have simply used a figure of 25% per annum in the model and applied it to both tape and disk costs.

The expected lifetime of disk against a tape library is an additional factor, the model assumes a 4 year lifetime for disk, and a >5 year life time for tape. This leads in year 3 to a new purchase of disk and in year 4 execution of a migration exercise for the disk based data in year 4.

Tables B and C below, provide further details on the elements for the two scenarios. Capital equipment costs are estimated and clearly could be substantially different subject to more rigorous market testing.

Table B: Cost modelling for Tape Based Repository.(Scenario 1)

Cost Models (Scenario 1)	2011/12	2012/13	2013/14	2014/15	2015/16
Equipment (Capital)					
Disk1	£125,000	£71,250	£505,913	£124,146	£163,873
Disk2					
Tape	£75,000	£42,750	£56,430	£74,488	£98,324
Offsite Copy	£100,000	£57,000	£75,240	£99,317	£131,098
Servers	£40,000			£40,000	
Start Up and Migration Costs	£120,000			£50,000	
Equipment (Operational)					
Maintenance	£11,250	£17,663	£26,127	£37,300	£52,049
Offsiting Service (5% uplift p.a.)	£15,000	£15,750	£16,538	£17,364	£18,233
Power Costs	£5,913	£11,448	£22,162	£42,907	£83,067
Facilities (Non Power)	£400	£705	£1,240	£2,182	£3,841
Staffing	£175,000	£180,250	£185,658	£191,227	£196,964
Training	£10,000		£10,000		£10,000
Totals	**£677,563**	**£396,815**	**£899,307**	**£678,931**	**£757,448**

Table C: Cost modelling for Disk Based Repository (Scenario 2)

Cost Models (Scenario 2)	2011/12	2012/13	2013/14	2014/15	2015/16
Equipment (Capital)					
Disk1	£50,000	£28,500	£202,365	£49,658	£65,549
Disk2	£450,000	£256,500	£1,821,285	£446,926	£589,942
Tape	£0	£0	£0	£0	£0
Offsite Copy	£100,000	£57,000	£75,240	£99,317	£131,098
Servers	£40,000			£40,000	
Start Up and Migration Costs	£120,000			£50,000	

[15] Storage Economics, Four Principles for reducing Total Cost of Ownership, David R. Merrill, June 2011

Equipment (Operational)					
Maintenance	£0	£0	£0	£0	£0
Offsite Service (5% uplift p.a.)	£15,000	£15,750	£16,538	£17,364	£18,233
Power Costs	£21,024	£40,702	£78,800	£152,557	£295,350
Facilities (Non Power)	£824	£1,449	£2,551	£4,490	£7,902
Staffing	£175,000	£180,250	£185,658	£191,227	£196,964
Training	£10,000		£10,000		£10,000
Totals (2)	**£981,848**	**£580,152**	**£2,392,436**	**£1,051,539**	**£1,315,037**

Cost Modelling Comments

Overall the cost difference in the two scenarios over 5 years is estimated to be £2.9M or £600K per annum in favour of a tape based model. Clearly a tape based model provides a lower overall total cost of ownership, this is not unexpected.

The high densities available recently for disk arrays and tape libraries see for example, IBM[16], Xyratex[17], DDN[18] and Spectralogic[19] reduce considerably the footprint required in enterprise data centres. Although, in the case of disk, at the expense of increasing the power density per square metre.

The operating expense is a significantly higher (40%) percentage of the overall 5 year cost for a tape and disk, than for the disk only scenario (26%).

The power requirement for scenario 2, is a significant cost factor rising to 22% of the annual cost in year 5. It should be noted that the model does not account for any probable reductions in power consumption per petabyte.

The estimated price for scenario 1, averaged over a 5 year period, is £355 per TB, while the wholly disk based solution scenario 2, over the same period is £659 per TB.

Table C: Cost per TB

	2011/12	2012/13	2013/14	2014/15	2015/16
Price Per TB (Scenario 1)	£678	£225	£290	£125	£79
Price Per TB (Scenario 2)	£982	£330	£772	£193	£137

The per TB price reduces significantly over the 5 year period with the exception of year 3 where the upcoming (year 4) end of life of the disk requires a major procurement. Smoothing out this cost hike will be necessary to provide a sustainable model.

[16] www.ibm.com
[17] www.xyratex.com
[18] www.ddn.com
[19] www.spectralogic.com

Business Modelling

The costs for the IT infrastructure required to provide an institutional repository is a necessary step in informing the business model to deploy and sustain the facility. The crucial question is how to pay for it all, now and into the future? The architecture has two key elements an active storage layer and an archival layer. Many different business models are possible, to provide the necessary services required to manage the research data output of the institution but we will only consider two,

- A default allocation available to all researchers, with any requirements in excess of the default provided as a chargeable service. (Model A)
- A free at point of use service for the entire institution regardless of need, which of course does not require any further analysis.

Business Model A

It would be fair to say that based on the principle of a "well-founded laboratory" it is not unreasonable to assume that any research intensive institution will provide some basic level of storage dedicated to one of its core business activities. Obviously a number of questions arise from this statement, but only two will be considered,

1. How is the initial capacity scoped?
2. How is it allocated?
    - By researcher, Discipline, Faculty, Project?
3. How is it paid for?
    - Taxed?
    - At point of use?

In attempting to explore these questions, we can consider the current situation at the University of Southampton, where

1. The number of principal investigators is ~1100 (based on RAE returns)
2. The number of Post Graduate Research (PGR) students ~2000
3. The number of Post- Doctoral Research Fellows/Assistants (PDRF) ~400 (Estimated)

Scoping the initial Capacity

The question of an "appropriate" allocation is difficult to estimate on a pan-University basis as the requirements of each discipline will vary considerably, whereas researchers in Engineering, Chemistry or Physics may require multiple terabytes per researcher, other disciplines may require only a few gigabytes.

It is perhaps coincidental that the estimated research data holdings at the University (0.8-1.2PB), suggest that a capacity of 1TB per principal investigator would approximately encompass the starting requirement for the majority of the University's research community. Indeed the IDMB researcher survey suggests that 80% of researchers surveyed had less than 1TB of research data.

Allocating and Accounting for the Capacity

The principal investigator is the heart of the research activity at any University, PGR's and PDRF's are short-term investigators within the organization. It is therefore prudent to assign any allocation to the principal investigator. While simple to understand, this is however, complex to manage, often requiring manual intervention to manage quotas and complicated by increasing collaboration between disciplines. It is not clear that a "per researcher" model is scalable or workable. For similar reasons the discipline based allocation is also challenging, especially as it is difficult to define what a discipline is.

A higher level model is required in order to help ease the management and indeed understanding from the research community. The University of Southampton has recently re-organized into 8 faculties which is a manageable number of academic units to provision in a high level model. Faculties also tend to be coherent long term business units and therefore provide a reasonable basis for accounting.

The model needs to encompass some metric which reflects the varying research intensive natures of each of the faculties. Therefore we propose a model which envisages,

- Faculties as the basic business unit.
- Research intensity defines the faculty allocation
- Research intensity is measured by some ratio of;
    - #Principal Investigators
    - #Post-Graduate Students
    - Research Income.

Table B below provides some starting estimates for a University with 4 faculties. In calculating the allocation per faculty the following ratios will be used.

Research Intensity: Post-Graduates: Principal Investigators = 8: 1:3

Table B: Model Faculty metrics

Faculty	% PI	% PGR's	%Research Income
A	20.30	25.79	18.08
B	33.26	24.67	48.56
C	14.23	11.48	3.69
D	32.22	38.06	29.67

Assuming the capacities predicted for Southampton, 1PB of research data today growing to 9.6PB in 2015/16, the following allocations would be provided to each faculty based on research intensity.

Table C: Model Faculty based allocations (PB) based on research intensity.

Faculty	2011/12	2012/13	2013/14	2014/15	2015/16
A	0.19	0.34	0.60	1.05	1.85
B	0.43	0.75	1.32	2.33	4.10
C	0.07	0.12	0.22	0.38	0.67

D	0.31	0.55	0.96	1.69	2.97
Total	**1.00**	**1.76**	**3.10**	**5.45**	**9.60**

Assigning Costs

Model A assumes a default allocation per faculty based on research intensity, requirements outside of the allocation will need to be funded on an individual project level. The basic requirement of the research data repository is to store the data forever, which would suggest an operational model to cost recovery, i.e. the storage is priced on a monthly or annually. However, research projects overwhelmingly are funded capitally. This disconnect has been and continues to be problematic, but is exacerbated by the need to store research data for decades, when most research grants and contracts are finished in 5 years or less.

We have looked at Princetons "Pay Once Store Forever" (POSF)[20] model [DataSpace: A Funding and Operational Model for Long-Term Preservation and Sharing of Research Data, S.J.Goldstein and M. Ratliff. 27 Aug 2010], while we understand the model has certain weaknesses particularly focusing on the cost of storage hardware, which we have seen is 53% and 70% of the total costs for the two scenarios, nevertheless we believe the model is sufficiently valid to provide a good starting point. A good discussion on the pros and cons of the model can be found at http://blog.dshr.org/2011/02/paying-for-long-term-storage.html.

The model basically proposes a "pay-upfront" model based on the fact that the cost of storage per GB reduces over time. This provides for a surplus to be generated sufficient to fund the next upgrade. The model requires two variables, the estimated reduction in the price of storage (D) and its replacement cycle (R). The storage factor which is multiplicative on the estimated cost per storage unit is given by

Storage Factor = 1/((1-(1-D)^R))

We have applied this to the two scenarios,

Table D: Pay Once Store Forever

POSF Model	2011/12	2012/13	2013/14	2014/15	2015/16
Price Per TB (Scenario 1)	£991	£330	£425	£182	£115
Price Per TB (Scenario 2)	£1,436	£482	£1,130	£282	£200

Storage Factor	1.46
D=0.25 (25%)	
R = 4 years	

The real purpose for any business model is getting your customers to buy-in, with USB and other external drives at £50 per TB, £1000 per TB for a POSF storage model is a challenging sell, and requires a clear cost-benefit analysis.

[20] DataSpace: A Funding and Operational Model for Long-Term Preservation and Sharing of Research Data, S.J.Goldstein and M. Ratliff. 27 Aug 2010.