# DataPool: Engaging with our Research Data Management Policy

Wendy White, Head of Scholarly Communication, University of Southampton; Mark Brown, University Librarian, University of Southampton

## Rationale

There has been a clear steer from many funders that data are part of the core outputs of research activity and should be more readily sharable, with a move towards “intelligent openness”[[1]](#endnote-1). The Royal Society report that emphasised the desirability of this cultural shift, was published towards the start of the DataPool project, providing a useful context for discussion. Initial policy proposals had been developed as part of the preceding Institutional Data Management Blueprint Project, dovetailing with the production of a 10 year institutional roadmap[[2]](#endnote-2). The roadmap was deliberately visionary and has provided a useful framework for benchmarking institutional progress and maturity of research data management practice. Clearly the timing of the policy and roadmap were influenced by the requirements of the EPSRC and consequent sector-wide debate about implementation, responsibilities and cost. However it became clear as we debated the issues relating to the policy with Faculties, members of the University Executive Group and the DataPool Steering Group, that there was a strong view that the motivation for the policy was as much about supporting our own [[3]](#endnote-3)academic practice as meeting external requirements. Improved services and support for researchers are seen as crucial to underpin change, some of it discipline-specific and/or influenced by new technical possibilities such as wide-spread availability of a variety of imaging facilities.

It was felt that the policy[[4]](#endnote-4) needed to provide a clear statement of principles, but that it should also provide a framework for roles, responsibilities and decision-making. It also needed to take account of the legal and ethical context of research data management activity and provide a top-level commitment to core services offered by the institution in order to support compliance.

## Approach

The approach was therefore one which placed emphasis on enabling activity as well as outlining requirements. The engagement of researchers was seen as key for proper exploration of the complex issues through the drafting of the policy and right from the off the process was framed as iterative; not just for “finalising” a policy, but recognising that this is a developing issue over the longer term. There was a commitment to continue to refine the policy based on experience of implementation and return to Senate with updates and amendments as required. One of the challenges was a balance between helpful support for decision-making and procedures, but avoiding specificity that would require too frequent formal amendment. One solution to this was to cross-refer to other University Policies as appropriate e.g. Intellectual Property Regulations and to formally refer to some of the key guidance that was written to support the policy e.g. “Guidance on Retention Periods”. This gives weight to the importance of key guidance, whilst allowing for this underpinning material to be frequently updated and presented flexibly. It is currently available as web-based information but we plan to enhance this with additional audio and visual content.[[5]](#endnote-5) A detailed account of the development of the website is available in the project report on data management planning and guidance.

Creation of the policy was led by the Library with substantive input from other Professional Services and drafting by Legal Services. The academic iteration was targeted at as many formal and informal opportunities as possible. The University’s Research and Enterprise Advisory Group is chaired by the Pro-Vice Chancellor Research with Associate Deans contributing from all Faculties. Key policy issues were identified and discussed over several of these meetings. These included the definition of Research Data and how to decide which data need to be kept and for how long. Specific Faculty contacts were also nominated for more in-depth individual consultations. They provided feedback on each draft, explored problematic issues from a disciplinary perspective and acted as an on-going point of contact for the development of guidance and services. This role was extremely valuable as they could act as an embedded interface with their colleagues. As we consolidate the implementation of the policy local understanding and reach is essential to promote engagement. There is recognition that policy is only effective where it is seen as relevant, however the level of discussion of the policy in all fora right the way through to Senate gives encouragement that the issues covered in the policy are seen at the heart of research activity.

## Roles and responsibilities

Many of the issues that elicited the most vigorous discussion related to judging scope of requirements and decision-making at all stages of the research life-cycle. Therefore clarification of roles and responsibilities was both a point of debate, but also part of the resolution to problems. One of the main challenges was how to account for disciplinary difference in an institution-wide policy, particularly for retention requirements. Our survey of researchers for the IDMB project showed us that the majority want to keep their data forever[[6]](#endnote-6). The EPSRC now require 10 years after the last request for access. However there is disciplinary difference in custom and practice e.g. where it is cheaper and it is possible to re-run experiments data will not be kept. It was decided that our policy should require “significant” data to be stored and described appropriately, with researchers responsible for applying this “significance test” in relation to disciplinary norms and in the context of the specific project. Linked to this researchers have a requirement to be aware of any data management plans that relate to their research. We are still working with researchers with individual examples to assess the impact of this key component of the policy. At our recent Steering Group meeting we discussed the relative expense of making good quality retention judgements in some disciplines given the human effort required to assess the data compared to lowering storage costs in the longer term.

Another testing topic was deciding what we meant by data for the purposes of the policy. What are researchers responsible for? After due consideration, we decided that human tissue samples should be specifically excluded from this policy which led to a definition focussing on electronic and print data. This will be revisited, but for now this also means that other physical objects are out of scope and we will look at the impact of this, particularly with disciplines like art and archaeology. There has been a lot of interest in exploring improvements to the archiving of print data, both in terms of workflow processes and for ethics agreements taking account of the potential for format shifting, with consequent additional issues for data security.

To help us address the decision making across the life-cycle of a project and also over a time period that is likely to transcend individuals in posts, we are working with Arkivum on mapping some use-cases for ingest of data through the repository, making those data available as part of a specific project and then sending to long term archival storage, looking at decision points in the workflow and the potential of built-in triggers to aid timely decisions. In policy terms Deans have Faculty responsibility for ensuring that data are reviewed appropriately prior to any destruction. This emphasises the long-term institutional responsibility for its data, whilst working with a framework which supports PIs with their data management planning and decision making.

## Implementation

The Research Data Management Policy has now been in place for a year and we are working to increase awareness of the policy and facilitate engagement. The DataPool project had three clear strands of activity. The development of one-stop-shop guidance to provide thematic information to aid implementation and development of support services. Enhancements to relevant repositories to provide a robust data catalogue with appropriate storage. Roll out of a training model to explore research data management principles and practice with target groups of data creators. This inter-linked approach, like the approach taken to the development of the policy, is seen as essential to support long term cultural change and engagement with the policy. Success needs to be evidenced, but we anticipate progress will be incremental.

These three strands are at different stages of take-up. In the current financial climate storage requirements and investment options have been the subject of sector-wide scrutiny and strategic soul-searching. The Russell Universities Group of IT Directors has had this issue at the forefront. There has been on-going financial assessment at Southampton, taking the IDMB cost-modelling as a base, and a new investment proposal is shortly going to senior staff, providing a suite of targeted options for stepped investment. The ability to store data appropriately is a key part of the policy so improving storage options is a priority, but it is fair to flag that this is a sector-wide challenge. Case studies produced as part of DataPool have helped provide evidence of particular gaps in provision e.g. 3D imaging where third party providers are not seen to provide appropriate facilities for necessary curation. The aim is for the policy to support transition from reliance on external hard drives to risk-assessed and curation-enhancing solutions.

 The policy requires that data are effectively described, through both metadata and supplementary documentation, to facilitate future identification and reuse. The institutional repository already had the facility to support ingest of data through the JISC funded DataShare project with Edinburgh and Oxford. Through IDMB we developed a three tiered approach to metadata, which Essex has referenced[[7]](#endnote-7) to develop a new app to provide an enhanced data catalogue for EPrints. This has been tested with researchers at Essex and Southampton with input from Glasgow and Leeds. This is designed to support inter-operability through compliance with key schema such as DataCite and INSPIRE. This enhancement has just been integrated with the Southampton repository with an additional 20TB of storage. This underpins engagement with this strand of the policy and uptake of this service will be monitored over the next period. There is on-going work on the relationship between an institutional catalogue and data stored in disciplinary and/or funder data centres and this will be a key issue as this service matures. The ethos of service development to support policy implementation is of enablers, with the burden for researchers kept to a minimum. Where there may additional work, for example data cleaning and guidance for reuse where there can be a significant and costly time commitment, this should be relevant activity of benefit to researchers. As content grows in the enhanced data catalogue, the issue of the cost of added-value for reuse can be assessed at scale.

The training model developed through DataPool is designed to equip the next generation of researchers with core knowledge about good practice in managing research data as part of the overall reflective approach to doctoral training and early career mentoring.[[8]](#endnote-8) Cross-disciplinary training in first principles is run through the Researcher Development Graduate Centre (RDGC) generic training programme. This is enriched by workshops targeted at Faculties and Doctoral Training Centres (DTCs) where there is more emphasis on disciplinary case studies and on sessions for early career PIs developing research data management plans. So far training has run across Medicine, Health Sciences, Humanities, Biological Sciences, Web Science DTC and through the general programme. The training refers to the policy, raising awareness and linking the aspects of the policy to practical exemplars and personal interests. The partnership approach taken for the development and delivery of training reflects the integrated approach that we believe is necessary for the successful delivery of many aspects of the policy. PhD students, researchers, Library and IT staff have all been involved in developing the training and we will be working with the RDGC on ensuring PhD researcher involvement over the long term. This partnership approach was also crucial in development of the web-based guidance and in contributing advice as part of the Data Management Planning support service that was launched just after the ratification of the policy. Research and Innovation Services also contributed to the guidance and we are committed to drawing on expertise across the University and further developing skills and knowledge.

## Impact

One of the critical success factors for policy implementation is engagement with the principles of the policy whilst formulating data management plans. Over time this will mean refinement of the plan over the course of the project, but in this first phase of implementation the emphasis has been on supporting new projects at the proposal stage. This is also an important opportunity to improve the impact of the research, as well as the quality of research through improved practice. The Data Management Planning Service is co-ordinated by the Library, but draws in relevant expertise from across disciplines and services. Researchers have used the policy and associated guidance to reflect on possible use of relevant institutional services and we have found that the nature of proposals so far has covered a wide range of policy areas; advice on metadata standards, storage requirements, data visualisation options to enhance usability of open data, retention and data security. Researchers who have engaged with the DMP service have provided positive feedback and have referred other colleagues on or additionally engaged with the project in other ways, e.g. giving feedback on iteration of guidance, contributing to building up FAQs or testing the data catalogue. We anticipate that the DMP service users along with those contributing to training are helping to build a critical mass of change agents to embed the policy into multiple strands of institutional practice.

A number of case studies developed as part of DataPool have explored thematic, multi-disciplinary issues. One of these focusses on presenting migration data to the public in a way that enhances the usability of the data and provides exemplar guidance alongside the data, thus seeking to enhance the reach and impact of the research. We can see from work with DMPs that there are plans to make data publicly available in a variety of ways and in disciplines as diverse as English and Medicine. As one of the objectives of the research data management policy is to make data publicly available where possible and to support effective reuse, we will seek to extend such exemplars as projects progress to emphasise the benefits of the policy to researchers.

To encourage early engagement we have agreed with the Researcher Development Graduate Centre that the formal process for electronic submission of PhD theses be amended to provide the option to deposit data alongside the text of the theses. We will also work with them to extend a pilot where data was made available as part of the internal PhD upgrade process and whilst we want to proceed with this carefully so there is no risk to the students, we would like to gather more feedback on the potential of this approach. This is a good example of the research data management policy influencing other policies and processes and shows that an organic institutional approach provides the flexibility to adapt and change in all areas of activity.

An important part of promoting the impact of data is ensuring that they are easy to cite correctly. To this end the University of Southampton is an early institutional adopter of the British Library’s Data Cite DOI service. This will be used first by Crystallography who is well placed to review any early issues and help develop a policy on when DOIs will be minted. They will develop scripts for automatic generation of DOIs for the National Crystallography Service at Southampton, as well as an app for use in other EPrints repositories, including ePrints Soton. They are also developing a way of assigning DOIs to LabTrove notebooks. Ensuring researchers are appropriately credited for their work, providing possible tools for increased citation of work and engaging with innovative research practice is critical for the success of the research data management policy.

## Governance

To ensure that the next phase of our 10 year roadmap fully embeds the tenets of the policy into practice right across the institution, we are proposing at the next Research and Enterprise Advisory Group that the Steering Group for DataPool evolves with a link to REAG to provide continuity of governance for research data management issues. The PI of the project based in the Library will continue to lead co-ordination of research data management services as this has also been formalised into a longer term role, recognising the strategic importance to the institution. We look forward to a period of consolidation, whilst continuing to encourage an “innovation to service” approach to progress. A year on it still feels early days for the policy. There is a lot of the roadmap to go, but we also feel that a considerable amount of progress has been made in a relatively short period of time.

## Shared services

The initial experience of implementing the policy has helped us refine our understanding of the sector context. The collaborative nature of the Jisc programme with many examples of cross-working with other institutions, research centres and publishers has been instructive. The focus on institutional responsibility for research data was sharpened by the dismantling of the Arts and Humanities Data Service in 2008, followed by the EPSRC requirements. Equally many of our researchers have significant experience of working with National Data Centres, particularly those supported by NERC and the ESRC, where there are models of data flow and shared responsibilities for each part of the data lifecycle. We have been building on this experience as we have considered our longer-term developments, including cost assessment and investment priorities.

There is clear agreement that data supporting published results are an institutional priority. Our “significance” test for other categories of data retention remains a live issue. However as funders debate the cost of managing data which could be replicated and storage trajectories, we can benefit from working with them on these shared problems. The benefits of a collaborative approach have emerged strongly from our Steering Group discussions. The capacity of the researchers to match policy requirements in terms of time and system knowledge is linked to the capacity of service providers to supply expert support in areas like data cleaning /documentation and technical skills. Institutions find it challenging to carry the full range of discipline specific support, at the same degree of depth, provided by the data centres. Our developing assessment of major curation and preservation issues raises queries about institutional capability as well as affordability. There is appetite for shared services approaches, but a tension between the low barrier use of drop-box style third party services to facilitate sharing and the complexity of developing formal cross-provider shared service offers with consequent IPR and data security issues. We seek to build on our work with the data centres, other institutions and possible commercial providers to provide robust and achievable solutions to support our research data management policy. There are difficult problems but we are committed to work collaboratively on innovative, realistic and sustainable solutions.

## Acknowledgements

The development and implementation of the policy and guidance would not have been possible without the contribution of many people across services and Faculties. Particular thanks go to the Associate Deans Research, Faculty Champions, Oz Parchment and Brevan Miles (iSolutions), Mylene Ployaert (Research and Innovation Services), the project managers Dorothy Byatt and Steve Hitchcock and colleagues in Legal Services for their significant contribution to the drafting of the policy. We are grateful to JISC for funding this project and the Programme Manager Simon Hodson for his expertise and support.

1. Royal Society (2012), Science as an Open Enterprise, <http://royalsociety.org/policy/projects/science-public-enterprise/report/> [↑](#endnote-ref-1)
2. Brown, M., Parchment, O., White, W. (2012) Institutional Data Management Blueprint, <http://eprints.soton.ac.uk/196241/> [↑](#endnote-ref-2)
3. Byatt, Dorothy, Scott, Mark, Beale, Gareth, Cox, Simon J. and White, Wendy (2013) Developing researcher skills in research data management:: training for the future – a DataPool project report <http://eprints.soton.ac.uk/351026> [↑](#endnote-ref-3)
4. University of Southampton (2012) Research Data Management Policy, <http://www.calendar.soton.ac.uk/sectionIV/research-data-management.html> [↑](#endnote-ref-4)
5. University of Southampton, Research Data Management [website] <http://www.southampton.ac.uk/library/research/researchdata/> [↑](#endnote-ref-5)
6. Takeda, K. (2010), Institutional Data Management Blueprint project: initial findings report, <http://eprints.soton.ac.uk/195155/> [↑](#endnote-ref-6)
7. Ensom, T. (May 28, 2012), Re-using the IDMB toolkit [blog post], <http://researchdataessex.posterous.com/tag/metadata> [↑](#endnote-ref-7)
8. Byatt, Dorothy (2013) Research data management planning, guidance and support: report <http://eprints.soton.ac.uk/351027>

 This work is licenced under a Creative Commons Attribution 3.0 licence.

To view a copy of this license, visit <http://creativecommons.org/licenses/by/3.0/> [↑](#endnote-ref-8)