A Digital Curation Centre
'working level' guide

How to Discover Requirements for Research Data Management Services

Angus Whyte (DCC) and Suzie Allard (DataONE)

With contributions from:
D. Scott Brandt (Purdue University)
Susan Wells Parham (Georgia Institute of Technology)
Sherry Lake (University of Virginia)

Please cite as: Whyte, A and Allard, S. (Eds) 2014. 'How to Discover Research Data Management Service Requirements'. *DCC How-to Guides.* Edinburgh: Digital Curation Centre. Available online: http://www.dcc.ac.uk/how-guide

Digital Curation Centre, March 2014
This work is licensed under Creative Commons Attribution BY 2.5 Scotland

How to Discover Requirements for Research Data Management Services

1. Introduction

This guide is meant for people whose role involves developing services or tools to support research data management (RDM) and digital curation, whether in a Higher Education Institution or a project working across institutions. Your RDM development role might be embedded with the research groups concerned, or at a more centralised level, such as a library or computing service. You will need a methodical approach to plan, elicit, analyse, document and prioritise a range of users' requirements. The term 'requirements discovery' covers these activities, and this guide relates them to the process of developing RDM services. A DCC Guide *How to Develop Research Data Management Services* describes these services in more detail. Further support is available from the DCC[1] and DataONE[2] websites, and from the resources listed at the end of this guide.

In section 2 of this guide we summarise data management roles and responsibilities, and why these make a methodological approach to requirements worthwhile. Effective research data management involves many actors, supporting technologies and organisation, including coordination of human and financial resources. Collectively this research data management infrastructure will call on different support services at different stages of the research data lifecycle. Support services will need to integrate their information systems to some degree. Projects that create research data will also need systems to help manage that data within the project lifetime. And, once the data creators' needs have been fulfilled, other actors and systems may be involved in data archiving and preservation to fulfil the needs of subsequent users and re-users.

Non-academic stakeholders will have requirements, for example any external partners, or research participants who use or help to produce the data. Service offerings from other institutions and commercial providers will also have an important role. For example there will be a need to match individual researchers' needs for long-term storage with the repositories or archives that best serve their discipline, and ensure these also meet the expectations of the funding body or institution.

In section 3 we look at the contexts for RDM requirements. We take as the starting point a high-level model of the services that are typically needed. Then we consider why the research context is different from other areas of business process change, and what challenges that presents for identifying requirements. To help work around the diversity and complexity of research, we suggest three key areas likely to shape requirements;

- Ideas, artefacts and facilities: you will need to identify what kinds of things researchers consider 'data', the implied relationships to other records or information, and what form these take. Consider how far researchers depend on others to get their work done, and whether they use centralised facilities or standardised protocols. Any drivers for researchers to work at a broader scale in their research domain will likely add to the needs for RDM, while those aspects of their process that have the most technical uncertainty will also be the more challenging for RDM.

- Research stakeholders: producers, users and policymakers; developing any RDM service to work at institutional or cross-institutional scale demands a stakeholder analysis, to reduce the risk that some group or service provider's needs are not properly assessed. This should involve people whose roles typically span organisational boundaries, such as academic liaison librarians or research computing staff.

- Institutional rules and research norms: guidelines that articulate the benefits and risks of change will need to be established if stakeholders are to buy in to it. This especially applies to decisions on keeping and sharing data. Key areas of benefit will be around providing technical support during projects, helping researchers get acknowledged for sharing, help to decide what to keep, support for storage and for using data catalogues.

The RDM development cycle can be viewed as a recurring sequence of six phases: envision, initiate, discover, design, implement and evaluate (3). Figure 1 illustrates the first five core activities in context. The sixth phase, 'evaluation', is shown separately as it often involves different actors and may apply to the whole development process as well as its outcomes.

---

[1] *JDCC website. Available at: www.dcc.ac.uk*
[2] *DataONE website. Available at: http://www.dataone.org/*
[3] *These phases are informed by service design principles and stages in business process re-engineering (BPR) e.g. Kettinger, W. J., Teng, J. T. C., & Guha, S. (1997). Business Process Change: A Study of Methodologies, Techniques, and Tools. MIS Quarterly, 21(1), 55–80. doi:10.2307/249742*

In section 4 of this guide we outline key phases of RDM development and the activities these involve. In section 5 we describe tools to support the phases of initiating change and discovering requirements. The guide focuses on these activities, starting from the need for a broad strategy through to description of services to be implemented. It takes a step back from the techniques typically used to gather requirements information, such as surveys, interviews and group discussion. We consider how these techniques fit into service development, and what aspects of the research context need to be considered.

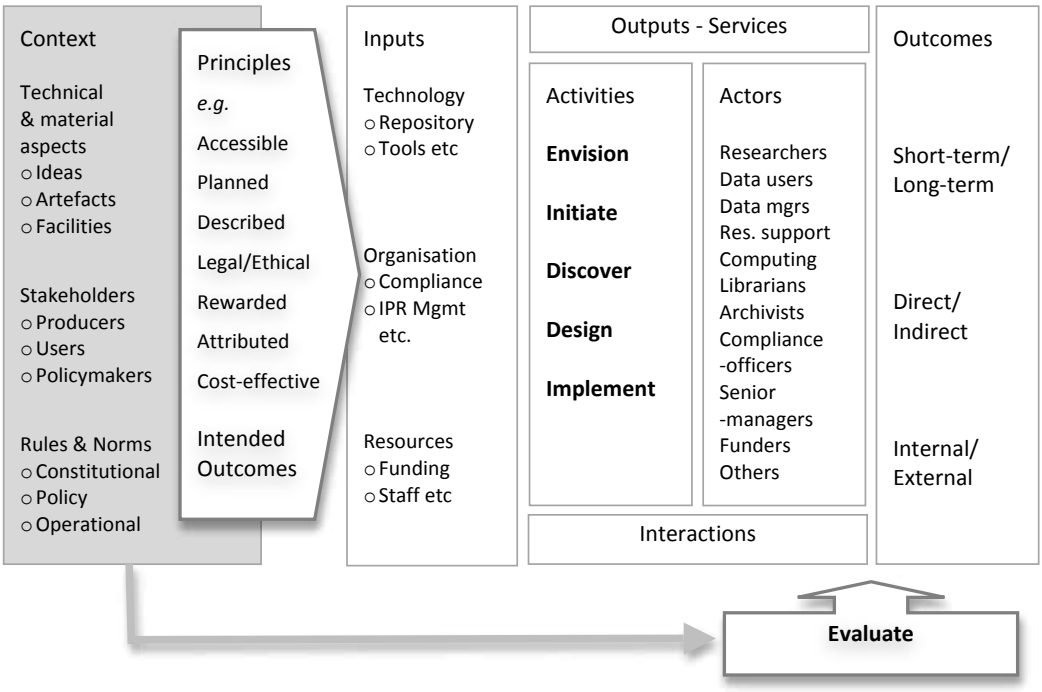

Figure 1 The process of developing research data management services

In section 5 we also identify the main elements of some common RDM requirements approaches. These include:

- Performing in-depth case studies
- Surveying data practices and benchmarking service capabilities, using the Data Asset Framework (DAF), Collaborative Assessment of Research Data Infrastructure and Objectives (CARDIO), or DMVitals.
- Documenting data lifecycles with Data Curation Profiles
- Stakeholder profiles, personas and scenarios
- Development workshop events, e.g. hackdays and mashups

Finally in Section 6 we consider the next step of managing requirements once they have been scoped, and some future challenges for scoping RDM requirements.

DCC and DataONE

The Digital Curation Centre is funded by the UK organization Jisc to build the capability for good data curation practice across the UK higher education sector. The DCC provides coordinated training and support to institutions, and enables the transfer of knowledge and best practice. This guide draws on the DCC programme of engagement with universities in the UK, and from working alongside related projects developing RDM services (4).

DataONE is a cyber-infrastructure funded by the US National Science Foundation (NSF). It aims to facilitate biological and environmental data discovery and access across distributed repositories, and to provide scientists with an integrated set of tools that support all phases of the data lifecycle including data management and curation (www.DataONE.org).

[4] *Jisc (2012) Managing Research Data Programme 2011-13 Retrieved from: http://www.jisc.ac.uk/whatwedo/programmes/di_researchmanagement/managingresearchdata.aspx*

2. Data Management Roles and Responsibilities

Policy makers in funding bodies and institutions are placing more importance on research data management and digital curation, putting new responsibilities on institutions and researchers. Scoping their needs may be complex, and stakeholders may be unfamiliar with data management and digital curation terminology. As their needs and wants will change, a requirements management process should help address challenges such as the following:

- **Data policy:** publishers and funding bodies are mandating that researchers make data openly accessible and reusable. Policy on research data continues to develop as new services and infrastructure become available.

- **Diverse stakeholders in research:** these include data producers, research funders, participants and users, any of whom may influence data management requirements - both during research projects and afterwards in terms of their needs for data produced to be reusable.

- **The 'data intensive' research paradigm:** new digital data sources and modelling technologies provide opportunities to integrate new and archived datasets, and look for patterns across them that can generate new hypotheses and theories. This 'inductive-deductive' approach, newer to science than to arts and humanities, is common in emerging bioscience fields. Effective data management and digital curation is key to data intensive fields, and requirements analysis will need to take into account their need for specialised data management tools and platforms.

Underlying all the above is the need to prioritise requirements according to the expected benefits and projected costs. A large proportion of the costs of developing RDM infrastructure are 'start up' costs to address issues like those listed above: developing policy, capturing requirements, and investigating implementation options – costs that will diminish with scale and experience 5). So while none of these issues should be obstacles to progress they do make the case for a methodical approach to identifying requirements, and managing them as they change.

Funding bodies define responsibilities for data management at a broad level. In the UK for example the Research Councils UK (RCUK) has set out Common Principles on Data Policy 5). Individual Research Councils adopt their data policies, which set out responsibilities that are shared between individual researchers and their institutions.
Most universities and other research institutions will have already developed a repository for publications. To help make research data accessible many now have a need to provide a 'core' repository for data outputs that are considered valuable assets, alongside services to help researchers ensure more digital data is of value to others. ,Another area of active development is in Current Research Information Systems (CRIS) to manage information on research outputs and integrate this with other administrative functions.

A US National Science Board report *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century* (7) gives a broad view of the roles and responsibilities of individuals and institutions to make data accessible. The report identifies roles for funding agencies, e.g. in creating a culture that gives as much consideration to digital data as to journals; encouraging the creation of an accessible 'digital commons', supporting the development of norms and standards, and providing resources and oversight.

The report also identifies responsibilities for data authors and users (see 'Research stakeholders' in section 3) and the main actors:

- Data managers: the organizations and data scientists responsible for database/ repository operation and maintenance

- Data scientists: information and computer scientists, database and software engineers and programmers, disciplinary experts, curators and expert annotators, librarians, archivists and others crucial to the successful management of a digital collection.

Data managers involved may already provide support to researchers through an institutional repository or subject-based archive. They will usually require software development capabilities, or access to these, to analyse the requirements and select the enabling technologies. In 'data intensive' fields this may call for detailed analysis of current practices, involving data scientists from the groups concerned. That analysis might also involve specialists in e-research or cyber-infrastructure.

---

[5]*Wilson JA, Fraser MA, Martinez-Uribe L, Jeffreys P, Patrick M, Akram A, et al. (2010) Developing Infrastructure for Research Data Management at the University of Oxford. Ariadne. 2010 Oct;65.*
[6]*RCUK (2011) Common Principles on Data Policy. Available at: http://www.rcuk.ac.uk/research/Pages/DataPolicy.aspx*
[7]*National Science Board (2005). Available at: http://www.nsf.gov/pubs/2005/nsb0540/*

Projects to develop infrastructure in institutions will typically involve research support units in the library, computing services, and research administration. The information professionals involved may come from a variety of backgrounds, as non-technical issues are likely to be at least as important as technical ones. Any process of organizational change may give rise to new requirements for policy development and training, and research data management is no exception. This may also require input from information governance or records management units, or experts in intellectual property management, whose role may be as important as that of e-research or scientific computing. Academic liaison or subject librarian roles may be expected to fulfil some of the service roles involved.

Beyond the individual research institution, many subject-based repositories exist (8), and some funding bodies mandate deposit to specific data centres. Some research groups may be aligned with disciplinary e-research infrastructures or cyberinfra-structures such as DataONE. Other groups involved in coordinating guidance, standards and workflows include national academies, learned societies, and professional bodies.

3. Research Data Management in Context

3.1. What services will be needed?

Any development project will need to identify the services its users and stakeholders need, according to their and the organisation's goals. There is some consensus on the range of support services that a research institution typically requires, which can help frame the options. Figure 2 below gives a high level view.

Services will follow the typical stages of the research cycle, from planning the data to be used to depositing and publishing it. Each of these will need online support in the form of guidance or software tools, and some degree of human guidance and training, depending on need and affordability.

Developing a research data management service in a university is like Business Process Re-engineering (BPR) in some respects. For example senior management need to support it, customer requirements need to be met, and internal

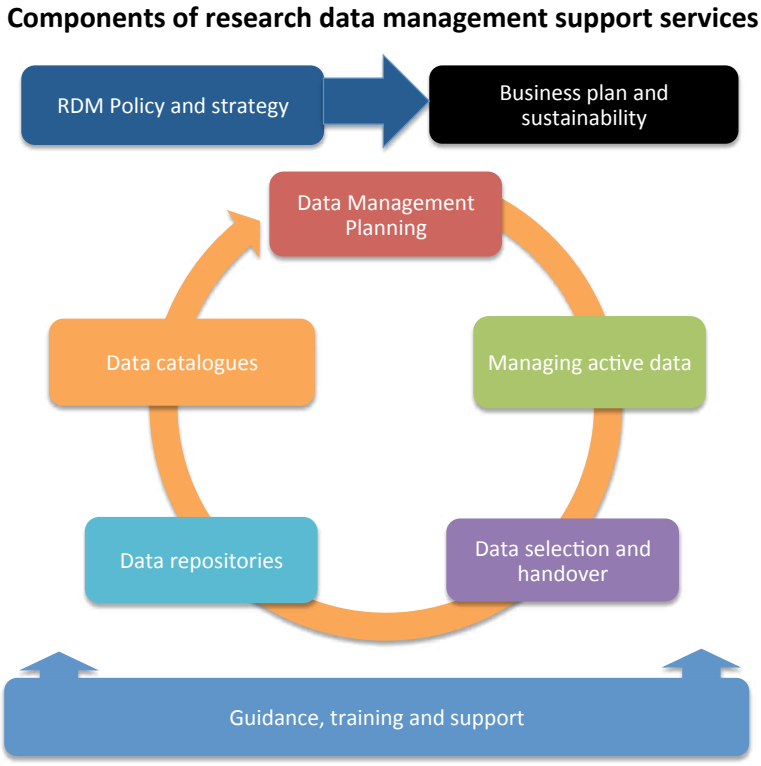

Figure 2. Research data management support services – an overview

[8]Directories of data repositories include Databib.org and Re3data.org

stakeholders' buy-in is essential. RDM is also like BPR in that its outputs typically include tools for transforming digital data and the workflows around data production. Also like BPR, cultural change may be at least as significant to success as the enhanced capabilities that are on offer from changed systems.

We also need to consider the important differences between the RDM context and business process change, as these present challenges for service development. Research contexts are unlike business processes in that they are typically more complex – i.e. the activities involved are difficult, uncertain, and with highly variable levels of interdependence and standardisation. Research domains have fuzzy and changing boundaries, organised through informal networks and according to norms that allow academic independence to pursue high-risk ideas and resist institutional control.

Catering for a research environment limits the scope to introduce standardised 'enterprise' solutions, as these may not meet the needs of research to explore novel ways of working with data.

Designing cross-disciplinary RDM infrastructure is a relatively new challenge, but research on e-infrastructure shows that, as in other forms of institutional design for shared resources, there is a need to appreciate the current norms and the 'patterns of interaction' that relate everyday practice to the principles, capabilities, service components and outcomes that are envisaged by funders and other stakeholders. Norms may be embedded in everyday research practice in a taken-for-granted way. It makes sense to identify them early in design, to avoid costly changes that may result if assumptions only become apparent when a new process threatens to damage the research process it was intended to serve.

Information professionals should ideally co-design with research communities, i.e. work closely with researchers to understand how they see the requirements, and gain their active participation to shape solutions that fit these requirements. Where feasible, RDM should build on the platforms they already use, and minimise the effort to share data – for example by extracting contextual information and metadata automatically from existing workflows (9).

3.2 Research ideas, artefacts and facilities

One definition of research data is '… the primary building block of information, comprising the lowest level of abstraction in any field of knowledge, where it is identifiable as collections of numbers, characters, images or other symbols that when contextualised in a certain way represent facts, figures or ideas as communicable information' (10). As this suggests, what counts as 'data' in a research field depends on accepted ideas of what it does or represents. That in turn may depend on the ideas, artefacts and facilities used in the field to test assertions that a given dataset is indeed what it is or represents.

Funding bodies define research data as a 'public good'. To fulfil that role, research data needs to be packaged with enough contextual information for it to be understood by (at least) the originating researchers' peers, using standards to define generic details and describe the contents in a form relevant to the discipline. This information will allow deposition in a well-governed archive, repository or community database, enabling the data to become useful to others.

Some of the information needed to frame a particular stream of bits as 'research data' could be located in a research proposal, data management plan, or in the instruments and software used to capture those bits. These will be relevant to the extent they help understand and test any assertion that the data is evidence for particular research findings. So digital data becomes 'research data' through its relationships to ideas that are embodied in this contextual description and representation information (11).

In a growing number of domains 'data' is recognised as an output to be shared with a gradually widening range of peers and stakeholders as a research project progresses. In other domains data may be the raw material for analysis and interpretation, but only shared to the extent that it is described in a research article or book. tThese cultural differences affect degrees of public data sharing. The levels of cooperation and competition between research groups affect this and can shift rapidly, especially when technology brings new opportunities to pool data and translate analysis methods across research fields.

Research fields vary in many ways, but two relevant distinctions are how interdependent researchers are, and how much technical uncertainty their tasks involve (12). More technical certainty means standards are more likely to be viable, and data or code more replicable. Greater mutual dependence is accompanied by economies of scale and centralisation of resources around large-scale facilities. It is probably no coincidence therefore that data management and sharing are better established in fields with these characteristics, such as astronomy, earth sciences and genomics. Even in data intensive disciplines there is wide diversity in data types and

[9]*Beitz, A. (2013) Growing an Institution's Research Data Management Capability through Strategic Investments in Infrastructure. International Digital Curation Conference. Amsterdam. 17 Jan 2013 Retrieved from: http://www.dcc.ac.uk/events/idcc13/programme-presentations*

[10]*Pryor, G. (2011). Managing Research Data. Brighton, UK: Facet Publishing.*

[11]*See DCC 'What is representation information?" http://www.dcc.ac.uk/node/9558*

[12]*Fry, J. (2006). Scholarly research and information practices: a domain analytic approach. Information Processing & Management, 42(1), 299–316. doi:10.1016/j.ipm.2004.09.004*

standards, including for metadata. A centralised service may require researchers to walk a fine line - between supplying the standard data and metadata needed for generic catalogues, and 'shoehorning' their research into inappropriate standards.

Research facilities, instruments and analysis platforms shape data production and management. Their location, capacity and capabilities will set boundaries on what more can be achieved by developing new RDM infrastructure that serves the the institution as a whole. Local 'data intensive' research clusters can also offer examples of good practice. These need to be considered in identifying requirements and evaluating results. Considering institutional backup solutions, for example, users' requirements for data retrieval may vary across different instruments that produce different file quantities, sizes, complexity and overall volumes. The more general point is that 'one-size-fits all' approaches will need to be minimal in scope and accommodate research groups' different backgrounds and changing needs.

The *artefacts* that embody data and relate it to its producers' expressed ideas will shape how data may be harvested and pooled. RDM tools or services can facilitate more use of 'digital research objects' that link digital workflows to data. These could include, for example, electronic laboratory notebooks. Dimensions to consider when assessing the options for, and benefits of, digital research objects include the extent to which they need to meet the ideals of being 'reusable, repurposeable, repeatable, reproducible, replayable, referenceable, revealable and respectful' (13). When introducing tools or services that replace physical artefacts such as notebooks with digital ones, and these are used in collaborative ways, it will be important to observe how people interact with them in the context of use to ensure the digital artefacts actually improve users' ability to make sense of data in that context (14).

3.3 Research stakekolders – producers, users and policymakers

In most research fields there will be a range of stakeholders involved in data production, as well as academic researchers, data managers and data scientists employed by a research institution. These stakeholders could include for example companies, policy-makers, non-governmental organisations

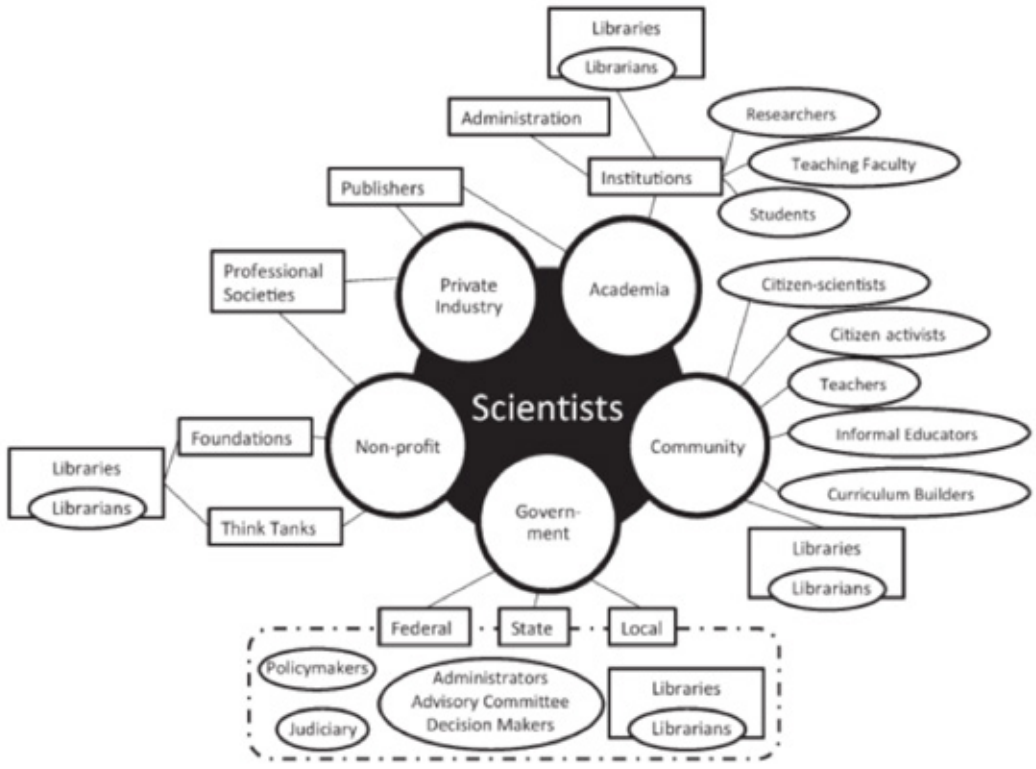

Figure 3. RDM Stakeholders: an example from DataONE. Individual roles are shown as ovals and organisations as boxes (reproduced with permission, from *Michener et al* (13)

[13]*De Roure, D., Bechhofer, S., Goble, C., & Newman, D. (2011, September 5). Scientific Social Objects: The Social Objects and Multidimensional Network of the myExperiment Website. 1st International Workshop on Social Object Networks. Retrieved from http://eprints.soton.ac.uk/272747/*

[14]*See for example the case study in Hartswood, M., Procter, R., Taylor, P., Blot, L., Anderson, S., Rouncefield, M., & Slack, R. (2012). Problems of data mobility and reuse in the provision of computer-based training for screening mammography. In Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems (pp. 909–918). New York, NY, USA: ACM. doi:10.1145/2208516.2208533*

RDM development will need to consider these stakeholders' needs, e.g. if they are likely to interact with the planned services. For example, the developers of the DataONE infrastructure for biological and environmental research identified primary stakeholders as 'scientists', with a further range of secondary stakeholders represented by those who regularly interact with scientists during the research process. As Figure 4.3 shows, librarians were identified as key secondary stakeholders, as they are present in each of the institutional environments in which scientists work (15).

The development project should identify how an RDM service will interoperate with other service providers, within a research institution and externally. These will include cyber-infrastructures such as DataONE in the US, and European research infrastructures catering for specific disciplinary communities (16).

Within the institution, senior management will need to steer the project to ensure the planned outputs are feasible, desirable and sustainable. They will also need to support any policy or guidelines introduced as part of the development, consulting on these through (e.g.) relevant committee structures and staff communication structures, and engage with the institutional business planning process.

Research support and administrative services will, as mentioned earlier, typically be drawn from an institution's library, computing, records management and research administration functions. In many cases their collaboration on research data management will involve exploring and negotiating new roles. Commonly this will be through participation in an RDM team to implement actions defined by a steering group. Identifying the needs for skills development, training and policy advocacy are likely to be within the team's remit, and these will be one focus of the discovery and design phases. Research computing staff and information professionals in library roles, e.g. repository managers and subject librarians, are likely to be considered the main actors in identifying the requirements for information systems.

Librarians, especially those in academic liaison or subject librarian roles may already have suitable 'boundary spanning' skills to apply to RDM service development. The need to work across academic, IT and other professional domains will be at least as strong as in digital library development. The requirements around metadata and cataloguing for research data will stretch these skills. As digital library research shows, stakeholders' different networks of practice will shape how they frame questions on what to record about data, and what needs to be shared (17).

The stakeholders will extend across institutional boundaries. For example they include discipline-related data repositories, and the cyberinfrastructures or research infrastructures (such as DataONE) that bind these together, providing the services that enable data management across institutions and disciplines (18). This infrastructure can be thought of in terms of layers of services. The fundamental layers are networked computing and storage services, from commercial providers of cloud services as well as public NRENS (National Research and Education Networks). On top of this, intermediary services providing standards and systems to support interoperability between data repositories. Developers may need to interact with a wide range of service providers; for example for author identification (e.g. ORCID), dataset citation (e.g. DataCite), or research information standardisation (e.g. OpenAIRE, CASRAI, EuroCRIS).

3.4 Institutional rules and research norms – why change?

In recent years 'top-down' policies on data sharing and governance have flowed from public funders and regulatory bodies. These reflect the need to address technology changes in science, and the broader public interest in research transparency and integrity (19). Principles for data governance such as those defined by Research Councils UK or the National Science Foundation help to identify the capabilities an institution needs in order to contribute to a research data commons. It can be useful to distinguish between such 'constitutional rules' and 'collective choice' policies that institutions and research groups need to formulate at an operational level. These are rules that define, for example, who may submit research data to a repository and who may access it.

The take-up of RDM services will also depend on community norms forming around the collective choices that research groups and their stakeholders make on what data is worth keeping and sharing. Institutional policies cannot determine every decision, but can set parameters that allow details to be worked out at a lower level such as the research group or department, or by specific service areas. For example

[15] *Michener, W. K., Allard, S., Budden, A., Cook, R. B., Douglass, K., Frame, M., … Vieglais, D. A. (2012) Participatory design of DataONE—Enabling cyberinfrastructure for the biological and environmental sciences. Ecological Informatics, (0). doi:10.1016/j.ecoinf.2011.08.007*

[16] *European Commission (2014) 'Research Infrastructures'. Available at: http://ec.europa.eu/research/infrastructures/index_en.cfm*

[17] *Khoo, M., & Hall, C. (2013). Managing metadata: Networks of practice, technological frames, and metadata work in a digital library. Information and Organization, 23(2), 81–106. doi:10.1016/j.infoandorg.2013.01.003*

[18] *Whyte, A. (2011) Emerging Infrastructure and Services for Research Data Management and Curation in the UK and Europe. In Pryor, G. (ed.) Managing Research Data. Brighton, UK: Facet Publishing.*

[19] *Royal Society (2012) Science as an open enterprise. Retrieved from http://royalsociety.org/policy/projects/science-public-enterprise/report/*

decisions on what services should be provided by which part of the organisation will take into account how researchers perceive the current offerings (e.g. of the library, computing, research support and training providers), and whether these are offered centrally or at faculty level.

Norms of data sharing are a contentious subject and will strongly affect the take-up of data repository services. There may be wide gulfs between top-down policies and researchers' beliefs and established practice. Many research investigators see it as their moral right to decide the terms of access, and surveys show a preference for sharing data only under conditions that directly benefit their own research, such as collaboration or joint authorship (20).

Policies necessarily set out the expectations of funders and institutions, but that may not be enough to make data sharing happen. Advocacy for RDM alongside Open Access publication may encourage the synergy between these forms of openness, since there is evidence that researchers publishing in OA journals are more likely to share data (21). Experience in developing institutional repositories for published outputs shows that service providers need to speak to faculty members in their language, and stress the value of the repository to them (22). For researchers, the support they get from their organisation is a persuasive factor in sharing data (23), and especially support in the form of:

- Technical support for data management during their project
- Formal acknowledgement or attribution for data creators
- Selection to limit the amount of data shared
- Support for data storage
- Skills in searching and cataloguing data

For researchers considering how and what to share, the likelihood of achieving greater visibility for their research may be persuasive in the long run, and studies in some disciplines show correlations between data sharing and higher citations (24). The scale of collaboration in the research also affects this (25), so it makes sense for RDM services to provide tools and support that help researchers collaborate on ever increasing scales, build their trust in mechanisms for sharing more openly, and help to identify the costs and benefits of doing so.

4. Development phases

4.1 Envision

Developing RDM infrastructure beyond early pilots to a live service demands a shared vision at senior level in the organisation. Senior management will need to establish a steering group or advisory board to assist them, or co-opt an existing champion or group already taking initiatives towards improved RDM. Ideally a member of the senior management group responsible for research should chair the group (e.g. the Vice President or Pro Vice Chancellor for Research).

A steering group's overall goals may include establishing policy principles for RDM, to help communicate where it fits within the organisation's overall mission, and to define the roles and responsibilities outlined in this guide. The steering group will also devise a strategy or roadmap for high-level approval, considering the institution's research strategy, external policy drivers, service priorities and technology opportunities. The steering group's overall goal will be to identify a business case to be presented to whichever organisational body can commit funding towards establishing the necessary services. Once this case has been made, the steering group may be reconstituted to plan and guide a project through to an established service.

Key elements

- Establish management commitment and vision
- Discover research policy and strategic opportunities
- Identify technology drivers for change
- Scope initial investigation

4.2 Initiate

Having secured commitment for an initial investigation, the steering group's work will focus on raising stakeholder awareness and obtaining 'buy-in' for further development. Often RDM steering groups bring together service providers that communicate about their support for teaching and learning, but rarely about research support or anything directly relevant to data.

---

[20] *C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., Frame, M. (2011). Data Sharing by Scientists: Practices and Perceptions. PLoS ONE, 6(6), e21101. doi:10.1371/journal.pone.0021101*

[21] *Piwowar, H. A. (2011). Who Shares? Who Doesn't? Factors Associated with Openly Archiving Raw Research Data. PLoS ONE, 6(7), e18657. doi:10.1371/journal.pone.0018657*

[22] *See for example the University of Rochester (US) case study reported by Foster, N. Gibbons, S, Bell, S. and Lindahl, D. (2007) 'Institutional Repositories, Policies and Disruption'. Available at: http://hdl.handle.net/1802/3865*

[23] *See for example analysis of DataONE survey results in: Sayogo, D and Pardo, T (2013) 'Exploring the determinants of scientific data sharing: Understanding the motivation to publish research data' Government information Quarterly 30, pp. 519-531*

[24] *Piwowar, H., & Vision, T. (2013). Data reuse and the open data citation advantage. PeerJ PrePrints. doi:http://dx.doi.org/10.7287/peerj.preprints.1v1*

[25] *Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The Increasing Dominance of Teams in Production of Knowledge. Science, 316(5827), 1036–1039. doi:10.1126/science.1136099*

The group will need to gain mutual understanding and identify the salient issues. This will include identifying gaps in capabilities required to meet policy requirements of funding bodies and regulators, and to exploit opportunities e.g. for new research, or for efficiencies in data management. Dialogue about the benefits and risks – to researchers, the institution and to external customers and stakeholders – will help to identify the desired outcomes and criteria for success.

It will be vital to engage with researchers at all levels, and senior researchers should be included in the steering group. To understand the implications of making research data workflows more formalised, or opening them up, it is also valuable to get input from research users or knowledge exchange experts.

Scoping the real needs of data producers and users will begin with focus groups or workshops, surveys and structured interviews. This is likely to demand a broader resource than can be provided from the steering group. Operational teams will also be needed to supply effort, with input principally from the library, IT and research support services. Tools to support this stage include the Data Asset Framework (DAF) described in section 5. The KRDS Benefits Analysis Tool (26), which identifies potential returns to researchers, service providers and external stakeholders is relevant here. Generic project scheduling and budgeting techniques are also likely to be used at this stage.

Key elements
- Identify, consult and inform stakeholders
- Identify researcher & customer priorities
- Identify desired outcomes and success criteria
- Organise operational team(s)
- Conduct project planning

4.3 Discover

Working from priorities identified at a high level, the nest stage is to diagnose the need for change in current practice and discover the requirements. It is important to appreciate some of the disciplinary landscape and the prevailing norms around data production, sharing and use. The study will also consider the support service landscape, who that involves, and what issues researchers and other stakeholders encounter. The discovery phase should set out what the RDM service will need to do, in the form of use cases or user stories.

Research naturally involves highly specialised knowledge and non-standard techniques for data collection. Different research groups, even within similar disciplines, will have their own methods. Given the number and diversity of research groups in an institution, the RDM Steering Group or project manager will need to be selective in who they can seek to involve in scoping the service requirements. Ideally the groups engaged with should span different funding sources, data types and scale of research team (i.e. from lone researchers to large consortia), so that the study can account for variation across the factors highlighted in section 3. It is also useful to involve researchers from different career stages, as PhD students' needs and concerns will differ from those of senior professors.

It will be important to identify the appetite for change, how needs are framed and the likely barriers to aligning them with RDM strategy and regulatory requirements. The discovery phase may therefore include an assessment of the awareness of relevant policies, and chart the lifecycle of typical data assets and associated research objects (software, protocols, logs, etc). In practise the level of participation will depend on researchers' levels of interest or concern.

The selection of methods for requirements discovery should therefore include a spectrum of approaches that are quick and easy for researchers to engage with and yield an overview of needs and awareness of policy obligations, to those that take more time but yield more in-depth information on current assets and practices. Typically a project manager or group with operational responsibility will undertake this work in a series of short studies involving selected research groups and the current providers of any relevant services such as backup, storage or library support. Useful tools here include *Data Curation Profiles, DMVitals, Stakeholder Profiles*, and the Data Asset Framework (DAF). Each of these is detailed in the next section.

Interviews and workshops can yield a great deal of qualitative description that will need to be distilled to identify the activities most in need of support. The CARDIO tool (Collaborative Assessment of Research Data Infrastructure and Objectives) can usefully complement this analysis. This provides the RDM project manager with a generic model of desirable capabilities for an RDM service. The model is applied in the form of rating scales, so that stakeholders can provide a numeric rating of current support provision and discuss different perceptions, drawing on the evidence garnered from interviews and workshops with researchers or other service users. Such models can help map the service to be designed, by providing an initial framework for summarising the large amounts of qualitative information that emerges from requirements gathering.

This analysis should then feed into more standard design approaches that are used to document user needs, such as use cases and user stories (27). The

[26] *KRDS Benefits Framework Worksheet (2011). Retrieved from: http://www.beagrie.com/KRDS_BenefitsFramework_Worksheetv1word_July2011.doc*
[27] *Cockburn, A. (2001). Writing effective use cases (Vol. 1). Addison-Wesley Reading. Retrieved from http://alistair.cockburn.us/get/2465*

acceptance criteria that will be used to judge how well use cases are fulfilled are an equally important output from this stage. Benefits frameworks for RDM are also likely to be helpful here; e.g. the KRDS Benefits Analysis Tool mentioned earlier (27).

Key elements:
- Document existing data practices and support
- Analyse existing data practices and support
- Identify required organisational, technical and resource capabilities
- Identify user needs and acceptance criteria relevant to data.

4.4 Design

Designing an RDM service is an iterative process, starting from early prototypes established in the discovery phase and taking these further through alpha and beta stages. As in any other service design project the basic concepts of the new service will be identified through the discovery phase. The design phase should identify with progressively greater clarity what purpose each service will fulfil for its users/ customers, what functions will be needed to do that, what value is provided as a result, and how that value will be known. These elements (purpose, functions, quality and performance) can be used to describe the services to be provided (28).

Agile design methods are likely to pay dividends. An early step in drafting use cases or user stories is to establish what roles and responsibilities are needed to provide the required functions and level of support e.g. online only or with some degree of face-to-face advice or consultation. These can be aligned with high-level description of services (e.g. Figure 2) before refining them to identify a modular set of services and interactions that fulfil the required use cases. The 'curation micro services' model (29) is an example of this modular approach, and may help to minimise the impact of changes in costs and availability of individual services.

The alpha stage will require a detailed operational plan for taking any online RDM system through to beta and live phases. Whichever design approach is taken the alpha phase should implement a basic working prototype from the use cases and any wireframe or paper prototype, aiming to solicit user feedback. If this indicates that a workable approach can be established with the resources available, the beta stage will take forward what has been learned and produce a fully working prototype of the online service (see, for example, the Government Service Design Manual (30).

The design phase will also address requirements for integration with other services. These are likely to include, *inter alia,* a grant costing system, Current Research information system (CRIS), and research output repository. The design process will also need to take account of existing or planned lower-level infrastructure such as network-attached storage or external cloud-based storage-as-a-service.

Research groups in disciplines with mature infrastructure for RDM (such as astronomy, or genomics), may have well-established platforms and workflows for using and depositing data in externally based archives and virtual research environments (VREs). There may be home-grown specialist repositories, presenting opportunities to integrate these with any central data repository the RDM service is to provide. Bringing all existing platforms within the design prospect will not happen overnight but a step-by-step approach can be taken, working with willing research groups to identify opportunities for integrating the workflows for metadata management with those for publishing metadata in an institutional catalogue. Tools available for workflow modelling include Research Activity Information Development (RAID) diagrams (31), Web Curator and MyExperiment (see entries in DCC Tools and Services Catalogue (32)).

Key elements
- Define and analyse new service concepts
- Prototype and detailed design of new service
- Design human resource structures
- Analyse and design data management tools and infrastructure

4.5 Implement

The design beta phase should establish which support functions (e.g. IT, Library or Research Office) are the 'owners' of which services, and any needs for restructuring of individuals' roles within these services. It should also identify the need for new relationships and workflows to be established between these roles. Implementing these changes will need advocacy, training and professional development. As new workflows are put in place, any published guidance may need to be updated to communicate these changes. Implementation may need careful negotiation, as it is liable to disrupt the existing norms of service providers as well as impose on practices that have long been the sole responsibility of Principal Investigators (PIs).

For data management tools and software-based services the critical issues are likely to be around integration with other systems and compliance with

[28]Taylor, S. (2011). Service Intelligence: Improving Your Bottom Line with the Power of IT Service Management. Retrieved from http://www.informit.com/store/product.aspx?isbn=0132692074
[29]Abrams, S., Kunze, J., & Loy, D. (2010). An Emergent Micro-Services Approach to Digital Curation Infrastructure. International Journal of Digital Curation, 5(1), 172–186. doi:10.2218/ijdc.v5i1.151
[30]UK Government Digital Service (2013) Government Service Design Manual. Retrieved from https://www.gov.uk/service-manual
[31]Darlington, M., Ball, A., Howard, T., Culley, S., & McMahon, C. (2011). RAID Associative Tool Requirements Specification. Retrieved from http://opus.bath.ac.uk/22811/
[32]DCC Tools & Services Catalogue. Digital Curation Centre. Retrieved from http://www.dcc.ac.uk/resources/external/tools-services[25]Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The Increasing Dominance of Teams in Production of Knowledge. Science, 316(5827), 1036–1039. doi:10.1126/science.1136099

other systems and compliance with standards. For many universities the institutional systems to which RDM support will interface, such as research information systems and output reporting, are both new and the subject of shifting policy demands from funders (e.g. in the case of delivering Open Access). This will in particular affect the requirements for metadata exchange between these systems. Tools should also be flexible enough to allow support for standards-based profiles, particularly those based on CERIF for research information (see EuroCRIS (33)), as these become more widely used for information about research datasets.

There may be a need for the institutional RDM service to comply with standards for web accessibility, or for ISO 27000 information security management (34). Of course all compliance needs should be identified in the discovery phase. However these will change as standards in the RDM domain develop further. For example at the time of writing few institutions have sought accreditation for their RDM services to the 'trusted repository' standards, such as the Data Seal of Approval and ISO 16363 (see APARSEN, 2012), and this may well change.

Implementation of data management tools and infrastructure should be to a defined level of service, reflecting expectations for availability and reliability. Acceptance measures and testing plans for these and other performance criteria will have been defined in the beta phase. These need to be included in the business model and operating procedures for the service in its live phase.

Key elements
- Reorganise support services
- Implement data management tools and services
- Ensure compliance with relevant standards

4.6 Evaluate

A decision to 'go live' with an RDM service will depend on the case being made to senior management for offering a fully operational service. Decisions on resourcing may hang on the availability of evidence showing measurable benefits to users and other stakeholders. If these are identified early, in the discovery phase, they can be refined through later phases in light of the practicalities of gathering meaningful data. This should provide the groundwork for setting in place feedback mechanisms and analytics that will allow continuous improvements to be made to the service.

The project team will have a roadmap or operational plan to monitor progression against milestones. Beyond this, benchmarking can help the project team and others to maintain an overall picture of how well capabilities are improving, and guide decisions on readiness to move from alpha to beta stages and make the business case for more funding. The CARDIO tool (see section 6) and other generic models are available (e.g.35) to guide a development team on the range of capabilities needed.

Projects should consult widely on the anticipated benefits, and agree realistic indicators as evidence of their accomplishment. Carrying this out from the beginning and throughout a project, rather than only as an end-of-project activity, should help to ensure that evaluation serves the practical needs of the service and its users. For example, in the JISC Managing Research Data programme a team of 'evidence gatherers' consulted individual projects on the evidence of benefits they could realistically produce from narratives and short case studies as well as quantitative metrics such as downloads (36).

Key elements
- Evaluate benefits and costs of service improvements
- Identify the metrics for a continuous improvement programme

5. Getting Started and Discovering Requirements

The approaches featured in this section have been adapted to the RDM domain, and there is a wide range of more generic tools and methodologies for requirements discovery that can also be drawn upon (see e.g.37).

5.1 In-depth case studies

Case studies of research groups' data practices can be of great value to understanding their requirements in context, although this can also be a very time-consuming and therefore costly approach. The case studies may be 'immersive', involving information or data science specialists following researchers and the story of the data they produce from observation of their work. This can require weeks or months of

[33]*Eurocris. Main Features of CERIF. Retrieved from http://www.eurocris.org/Index.php?page=featuresCERIF&t=1*
[34]*A useful example was developed in the 'Data Management Planning for Secure Services' project, see http://www.ucl.ac.uk/ich/research-ich/mrc-cech/data/projects/dmp_ss*
[35]*Crowston, K., & Qin, J. (2010). A Capability Maturity Model for Scientific Data Management. Retrieved from http://crowston.syr.edu/content/capability-maturity-model-scientific-data-management-0*
[36]*For more details see Whyte, A., Mollow, L, Beagrie, N and Houghton, J. (2013) 'What to Measure? Towards Metrics for Research Data Management' in Ray, J. (Ed) Research Data Management: Practical Strategies for Information Professionals. Purdue University Press*
[37]*Alexander, I and Beus-Dukic, L. (2009) Discovering Requirements: How to Specify Products and Services Chichester: Wiley*

participation and observation of research, if justified by the complexity of the data management issues. For example, some recent studies point to substantial challenges in the reuse of data, especially for new purposes or across disciplines (11,38). Case studies can also involve workshops or focus groups with researchers on data policy issues and support requirements.

Lighter-weight approaches outlined below also use interviews and focus groups or workshops. Clearly, balance of effort and return is needed. Lightweight approaches will often be preferable in a development context, but in-depth studies may be feasible through collaboration with researchers specialising in social aspects of information or computing science. These may have a better chance of yielding new insights and models for RDM, or addressing inconvenient truths.

In-depth study can be easiest to justify where substantial changes are planned to workflows and can be anticipated to have large and significant impacts. For example, automating the procedures around depositing data into a large repository might justify an in-depth analysis of workflows to ensure that requirements are properly understood. If such studies cannot be resourced within an RDM development programme, relevant examples may be available e.g. from the *International Journal of Digital Curation* (www.ijdc.net), *Data Curation Profiles Directory* (http://docs.lib.purdue.edu/dcp/), or from the DCC and DataONE websites.

5.2 Surveying data practices and benchmarking service capabilities

Many institutional RDM projects benefit from online surveys of researchers and other stakeholders in data management. These typically draw on similar questions to larger scale surveys of common practices and attitudes towards data sharing and reuse. For example, the DataONE project surveyed thousands of scientists in 2009-10 to produce a 'baseline assessment', following this up with more detailed work on stakeholder personas (see below).

Data Asset Framework

DAF (Data Asset Framework) offers a quick and lightweight approach to discovering data management practice though online and face-to-face surveys and interviews. The main stages in DAF (39) are:

- Stage 1 Planning, defines the purpose and scope of the survey, e.g. institution-wide or a specific faculty group or support function, and conducting preliminary research. In some institutions DAF studies have supported consultation on draft RDM policy, feeding into the university's formal decision-making processes through committees dealing with information governance and research.

- Stage 2 Identifies what data assets exist and classifies them to determine where to focus efforts for more in-depth analysis. Depending on the scope this stage may simply be to characterize data types and volumes, or in detailed reviews of a group's data practices a more thorough appraisal of a research group's assets can be carried out.

- Stage 3 identifies and follows the typical life cycle of the research to characterize researchers' workflows and identify opportunities and threats/risks in data creation and curation practices.

- Stage 4 pulls together the information collected and provides recommendations for improving data management.

Stages 2 and 3 use a combination of *online surveys* and semi-structured interviews. The online surveys are typically 10-20 questions covering research-active staff awareness of policy requirements, responsibilities for data management planning, expectations of benefits, needs for training and guidance, current practices for backup and storage, providing access to working data, and sharing data of longer-term value, plus their priorities for support service provision.

The *semi-structured interviews* typically span a number of pilot groups, across disciplines, funding sources and scale of research teams. Studies normally involve 3-6 researchers depending on group size, from PhD students to group leaders at professorial level. The interviews cover similar topics to survey questions, but aim for a more conversational approach to understand the researchers' current field of research and context. This covers aspects as outlined in section 3 e.g. instruments and artefacts used in data collection, the tools, standards and infrastructure used to work with data, views on data reuse and policy drivers towards that.

Resourcing these studies demands careful planning. The outcome will be a high-level analysis of the scope of RDM requirements, with recommendations to the RDM steering group on which service development priorities to address. A DAF study can entail similar protocols to a qualitative research case study. For example, although the purpose is development rather than research, ethical review may be required of interview questions, informed consent forms and any steps taken to anonymize the results.

[38]Faniel, Ixchel, Eric Kansa, Sarah Whitcher Kansa, Julianna Barrera -Gomez, and Elizabeth Yakel 2013 "The Challenges of Digging Data: A Study of Context in Archaeological Data Reuse" JCDL 2013 Proceedings of the 13th ACM/IEEE- CS Joint Conference on Digital Libraries, 295-304. New York, NY: ACM http://dx.doi.org/10.1145/2467696.2467712 Pre print available online at http://www.oclc.org/content/dam/research/publications/library/2013/faniel-archae-data.pdf

[39]For more on Data Asset Framework, see http://www.data-audit.eu/

Notes or any transcripts of recordings will need to be analysed, and while this will not be to the same level of rigour as research (which typically takes around 6 hrs analysis per hour of interview) it will still involve at least as much time as carrying out the interviews. You can reduce the effort needed by involving more researchers per interview, reducing the length to 15-30 minutes, or using fewer open-ended than closed questions, although the latter two approaches will also reduce the richness of the results.

Using the DAF at Georgia Tech

Susan Wells Parham, Research Data Project Librarian

Although known for its engineering programs, Georgia Tech has a strong research presence in a range of technology and data-rich fields within the science, social science, and humanities disciplines. Our goal in conducting a research data assessment was to develop a broad understanding of the research data environment across these varied fields.

Rather than conduct a comprehensive audit of a single school or research group, we used the DAF to create an online survey to gather basic data holding and management information from at least one researcher in each of the institute's schools, and from multiple research centers. We recruited researchers working with a wide range of research methodologies and practices, as well as receiving various levels of technical and financial support (63 total responses).

This snapshot of the research data environment at Georgia Tech allowed us to modify existing services and to plan for future data curation. Survey results revealed a gap in services for unfunded research – many respondents not participating in sponsored research expressed a particular need for data sharing and preservation services, as well as information about data management best practices. We immediately began the development of promotion and outreach targeted to faculty, graduate students, and research administrators including presentations, articles, web guides, and print materials. We also developed training and consultation for researchers writing data management plans.

Results from our DAF-based survey also revealed that many researchers create data sets in easily accessible formats that are relatively small in size, and can be made publicly available for an indefinite amount of time. These criteria fit the collecting policies for our institutional repository, SMARTech, so we developed policies for research data sets and began working with researchers to collect and preserve data via our repository.

We asked respondents to indicate their interest in any number of data curation services by selecting them from a predefined list. 73% of our respondents indicated an interest in data storage and preservation; 67% in data sharing tools; and 52% in data management best practices. Roughly 40% indicated an interest in: information about developing a formal data management plan; assistance meeting funding agency data management requirements; and selecting data for long-term preservation. These responses informed our work of partnering with other units at Georgia Tech, and our strategy for coordinating the stewardship of research data to provide the long-term access and preservation of data.

Collaborative Assessment of Research Data Infrastructure and Objectives

CARDIO (Collaborative Assessment of Research Data Infrastructure and Objectives) is a benchmarking approach that can be used to assess the gaps between current and required support capabilities. Both CARDIO and DAF seek information on current support for research data management. Where DAF gathers semi-structured information, CARDIO employs rating scales to assess the support offered, with a view to comparisons over time or across different groups.

The steps in a CARDIO benchmarking exercise initially involve a coordinator or project manager making an initial assessment by assigning ratings on a range of support elements, then recruiting the participants and inviting them to contribute their ratings. The assessments are then compared with a view to agreeing a consensus, for example by discussing areas where responsibilities are sufficiently clearly defined to merit a high rating. The participants' ratings and comments can be elicited in interviews, workshops or using survey techniques. Additionally, an online CARDIO tool (40) may be used to gather online responses, optionally employing a chat facility.

CARDIO uses a rating scale to assess organisational, resourcing and technology elements of RDM service provision. These elements can be assessed to different degrees of granularity according to the level of engagement required:

1. A simple 'CARDIO quiz' to help initiate a review of service provision. The quiz consists of 13 questions, with a choice between three statements representing their institution's current position. Individual responses are then used to offer an assessment of strengths and weaknesses that may be followed up by more in-depth assessment.

2. A 'roadmap matrix' to assess progress towards delivering RDM service capabilities that UK funding

[40] *CARDIO tool. Available at: http://cardio.dcc.ac.uk/*

bodies expect institutions to develop (10). This comprises a 9 x 5 table used in workshop settings. The nine rows are grouped under three headings:

- Policy, strategy development and sustainability
- Data management support and staff development
- Research data storage, preservation and sharing

In each of the nine rows in the table, a user can choose between five statements representing stages towards embedding the required services in the institution. These five steps broadly match those in section 3 of this guide:

1) Envisioning & initiating : a need for change is recognised and acted on
2) Discovering: requirements are being investigated and scoped
3) Designing & piloting: solutions are being tested through small-scale pilots
4) Rolling out: solutions are being funded and piloted more widely
5) Embedding: service in place with process for continuous improvement

3. Services may comply with funding body expectations without necessarily being optimal. So the third option is to assess the maturity of service provision against broadly defined elements of good practice. The full CARDIO model uses a set of 30 element descriptions, covering organisational, technical and resource aspects of RDM service provision. These elements and statements describe good practice, and can be assessed for the institution as a whole or at the level of faculties or smaller units, depending for example on whether service provision is devolved or centralised.

The CARDIO tool can help to gather assessments online. Face-to face workshop settings may also be worthwhile, especially where the individuals making the assessment are from diverse research and service backgrounds and do not ordinarily interact. A blend may be appropriate - e.g. a face-to-face workshop to introduce the tool and the participants to each other, with a project manager following this up using the online tool to facilitate more detailed ratings and discussion.

Using DAF and CARDIO at University of Warwick

DAF and CARDIO may be used to complement each other. For example DAF can fulfil an 'intelligence gathering' role to inform CARDIO assessments, collecting evidence of researchers' current practices and views on service provision, which can then be used to re-assess the level of readiness these services have reached. Alternatively, a CARDIO workshop can help to scope a DAF survey by identifying issues to prioritise.

DAF and CARDIO informed workshops at several of the universities DCC engaged with in 2012-2013. One of these was University of Warwick. DCC staff worked with information professionals in academic liaison, repository management and research support to deliver requirements gathering workshops. These involved academics and support providers, and examined factors governing institutional interest in RDM support, challenges and gaps in capabilities. The workshops combined presentations and discussion groups. Three groups were formed around the three categories mentioned previously, i.e.;

- Policy, strategy development and sustainability
- Data management support and staff development
- Research data storage, preservation and sharing

Each rated the three capabilities in their category, first as individuals, and then discussed reasons for their ratings and the issues they saw as priorities, and then agreed a consensus rating for each capability. From notes taken, DCC provided a summary report of the ratings and reasons given for them.

The workshop comments were then used to frame a set of interview questions, adapted from previous DAF studies. DCC then undertook interviews with research staff from two pilot groups, shadowed by the Warwick information professionals who had also taken part in the workshop. The interviews covered these topics: -

- Planning for data management and sharing
- Data collection
- Data Processing and Analysis
- Making data accessible, safeguarding valued data
- Support and training
- Data management challenges
- Expectations about Services and Priorities

The interviews were transcribed and summarised, and these summaries fed back to participants to check views were accurately represented. Highlighted themes were then presented in a follow-up workshop with service providers to help formulate an implementation plan to put into practice the University's Research Data Management Policy.

DMVitals

The DMVitals approach developed at University of Virginia has similarities to both DAF and CARDIO. Like DAF it is structured around interviews with individual researchers, and aims to record details of their data management practices with a view to understanding how these may be improved. DMVitals uses a more highly structured interview protocol, consisting of closed questions rather than the open-ended questions used in the other approaches in this guide.

The DMVitals questions are based around eight 'components' of data management practice, which are:

1) File Formats and Data Types
2) Organizing Files
3) Security/Storage/Backups
4) Funding Guidelines
5) Copyright & Privacy/Confidentiality
6) Data Documentation & Metadata
7) Archiving & Sharing
8) Citing Data

The available options for each question relate to statements of good practice that are derived from University of Virginia guidelines and the Australian National Data Service's (ANDS) model for rating long-term storage options. (41)

Similarly to CARDIO a "sustainability level" rating is given to the range of possible responses for each component. The five-point rating scale is derived from a Capability Maturity Model for research data management by the Australian National Data Service (ANDS); ranging from 'initial', through 'development, 'defined', 'managed' and, finally, 'optimizing'. (42)

While CARDIO ratings are used as a basis for dialog between stakeholders on how support may be improved, the DMVitals tool uses the scoring to assess the sustainability of individual researcher DM practices, and rates the quality of these from worst to best. This draws on pre-defined 'action statements' corresponding to the sustainability level that the researcher's responses indicate. To provide a framework for defining and improving researchers' DM practices, the DM sustainability ratios are averaged to define a data management maturity level. The DMVitals tool thereby provides immediate and actionable feedback from a data management interview, and may be used to automate advice provision for data management planning.

The strength of the DMVitals tool is the creation of a data management report, which generates tasks customized to each researcher. These tasks can then easily be grouped into phases, creating a data management implementation plan for each researcher based on his or her personal data interview and subsequent information gathering. Combining this tool with assessment and planning methods helps to expedite the recommendation report process, and provide valuable actionable feedback that the researcher can use immediately to improve the sustainability of his or her data.

The automated response approach aids requirements gathering where the requirements fit a known range of parameters, but is more limited in its ability to take into account an individual's research goals or circumstances. That does not necessarily make it a 'one size fits all' approach to requirements. If used as a benchmarking tool alongside a more open-ended interview, DM Vitals should help researchers make explicit any 'good' research reasons for 'bad' data management practice, and focus their thinking on how the tool's recommendations fit to their particular data lifecycle.

5.3 Documenting data lifecycles with Data Curation Profiles

The Data Curation Profile (DCP) is another tool used to gather information about a researcher's data set, what they are doing with it, and what they would like to do. Use of the tool often presumes that a researcher has not had the time to think about what it would take to deposit data in a repository. The Profile can help researchers articulate various aspects of the data, such as which experimental output should be shared (e.g., raw or analyzed data). It also asks researchers to describe needs or requirements for use, citation, rights, etc. As such, a Profile can provide insight into the workflow of scientific data and barriers to making it available to others.

The Data Curation Profile as an interview instrument is the result of a research project undertaken to explore "who is willing to share what with whom and when." (43) The project originally approached 21 researchers to discuss their data, and through an iterative process developed a series of probes that the researchers believed to be important. Among these were: what data would be important to share, what format or form the data should be shared in, parameters related to ownership, and conditions related to how the data would be used.

A Data Curation Profile (DCP) outlines the 'story' of a dataset or collection, describing its origin and

[41] *ANDS and Data Storage. Available at: http://ands.org.au/guides/storage.htm*
[42] *Australian National Data Service (ANDS). (2011). Research data management framework: Capability maturity guide. Available at: http://ands.org.au/guides/dmframework/dmf-capability-maturity-guide.html.*
[43] *Witt, M., Carlson, J., Brandt, D. S., & Cragin, M. H. (2009). Constructing data curation profiles. International Journal of Digital Curation, 4(3)*

and lifecycle within a research project. The approach was developed to address the challenge of determining "…which disciplinary and sub-disciplinary distinctions need to be attended to in shaping curation requirements and services" (44).

Data Curation Profiles can inform decisions applied to particular collections, for example in the selection of datasets for retention, and in the provision of metadata. The approach has a primary focus on analysing differences in data sharing practices across specific research communities, and how these are affected by, for example, the data types used and the stage of the data lifecycle.

The DCP approach is close to DAF in some respects, including the scope of the questions asked. Like DAF, the profiles are intended for librarians and others to use to inform decisions. However, where DAF aims for a general awareness of datasets and practices across a department or institution, DCP takes a more rigorous look at the data associated with a sub-discipline.

The process of using a DCP in a librarian-researcher 'data interview' can be useful in training and development, to build the participants' confidence to engage in discussions about data management – e.g. see MANTRA (45). Data Curation Profile interviews are designed to take place across two one-hour sessions. Like DAF, the approach uses a semi-structured interview format, with questions about the interviewees' research process, types of data collected, accessibility and ownership issues, how data is transformed through the research process, and any practices relating to sharing the outputs of the various stages.

A toolkit is available, comprising a profile template, a worksheet that researcher and interviewer will complete during the interview, a manual for interviewers, and a user guide (46). Completed Profiles may be submitted for publication in a reference resource, the *Data Curation Profiles Directory* (47). The DCP Directory provides a suite of services to support the publication of Profiles, including assigning a DOI for each published DCP, improved visibility for Profiles through inclusion in indexing and discovery tools, and a commitment to the preservation of DCPs through CLOCKSS and Portico.

The User Guide is designed to walk librarians through the process of doing a data interview, from initiating contact, to recording interview, to synthesizing a Profile from a transcript. The User Guide also gives recommendations for problems that might occur, such as keeping a researcher on track when discussing a pertinent data set, and which sections to focus on if a researcher can't spare very much time.

Typically a researcher is identified based on a number of reasons: for instance, their needs for data curation, previous relationship, or because their research data is otherwise deemed significant. It is essential as part of preparation to identify a specific project on which to focus. Some homework should be done to familiarize the interviewer with the researcher's work by scrutinizing portfolio websites and identifying a recent publication or preprint on the research. From this "intel" an interviewer should be able to identify both the researcher to interview and data set to discuss. (Note: while this interview does not collect personal information about the researcher, if the interviewer wishes to publish outcomes an approved protocol for human subject research may be needed.)

Typically two hours are needed to get through the interview process, usually divided up into two sessions. However, if for some reason a researcher is not able or willing to commit that much time, the interviewer should be prepared to focus on sections of the Profile which are considered required: an overview of the research, a breakdown of the data by kinds and stages, description of the organization of data, and researcher perspective on sharing and access. While the Profile will suffer from not including other sections (e.g., tools used, mechanisms for discovery, means for interoperability, intellectual property, etc.), it is crucial to capture required information at a minimum.

The Interview Worksheet is just that—a form intended for the researcher to fill out, either in advance or during the interview. This is important as the Data Curation Profile is meant to capture information from the researcher's perspective. The Interviewer's Manual helps both the interviewer and interviewee walk through the questions. The Manual suggests prompts to add depth and breadth to answers, as this makes for a rich Profile. Because of the likelihood of great detail, it is recommended to tape the interview, with the researcher's permission. It is very difficult to take useful notes while engrossed in a data discussion. Interviews are then transcribed or "indexed" to facilitate synthesis of information into the Profile Template. In its current form the Profile is meant to be semi-structured, yet flexible enough to represent a researcher's perspective of various data attributes. (For instance, data collected may be termed "initial" or "raw," data massaged for analysis may be called "processed" or "anonymized," etc. depending on discipline, lab practice, or a specific project.)

While primarily designed to review general research data attributes for a specific project, Profiles can be used in other situations. Some of the questions may help in designing data workflows that account for needs of curation, sharing and preservation "downstream." Or, the Profiles may help enhance

[45] *MANTRA (2013) MANTRA Research Data Management Training. Retrieved from http://datalib.edina.ac.uk/mantra/index.html*
[46] *Data Curation Profiles Toolkit. Retrieved from http://datacurationprofiles.org/*
[47] *Data Curation Profiles Directory. Retrieved from http://docs.lib.purdue.edu/dcp/*

curation requirements for data collections that need to be made available (e.g., to other researchers, publishers, or the public). But one of the greatest benefits seen so far is awareness and relationship building—researchers are appreciative of someone helping them by asking the right questions.

5.4 Stakeholder profiles, personas and scenarios

Stakeholder Profiles

Stakeholder profiles are designed to look in a little more depth at the data management knowledge stakeholders already have and how they are interacting with data. The profiles can also help identify how tools and services may be tailored to better serve stakeholders, and can provide insight into tools and services that can extend or improve the data practices of stakeholders.

The first step in constructing a stakeholder profile is to identify the primary and secondary stakeholder communities and their relationships to one another. Knowing who your stakeholders are helps tools and services developers prioritize the allocation of resources to the requirements of each community. Primary stakeholders are those who are the key focus of the tools or services being developed. For example, in the case of DataONE, the goal is to serve as a foundation for integrative biological and environmental research therefore its primary stakeholders are scientists.

Identifying the secondary stakeholders can be more challenging, since there may be a wide variety of organizations and individuals who regularly interact with the primary stakeholders during the research process. An efficient way to identify secondary stakeholders is to visualize what a successful research environment for the primary stakeholders would look like, and then to note the stakeholders who would need to be involved for this to be realized. As mentioned above in section 4.3, for DataONE the approach was to identify five key science research environments that scientists worked in, then to note what other stakeholders were in those research environments. For example, libraries and librarians were identified as important secondary stakeholders in every one of these research environments.

The next step is to learn about the attitudes, practices, perceptions, and requirements for each stakeholder community. Conducting an assessment survey is valuable for learning about the current attitudes and practices of the stakeholder communities and also for providing a baseline that can be used to assess how tools and services have been adopted and how, over time, they have changed research and data management practices. By surveying each stakeholder community separately, it is possible to build a rich picture of how the stakeholder community interacts with data throughout the different phases of the data lifecycle. Importantly, it provides an ability to judge the prevalence of different attitudes, beliefs and practices. This allows tools and services developers to better identify and prioritize requirements.

The DataONE data lifecycle shown in Figure 4 is a useful aid here. This lifecycle is developed from the data perspective and the expectation is that multiple individuals will be interacting at different points and engaging in different activities.

In DataONE, a scientist assessment was completed early in the project, and then scheduled to be repeated at regular intervals. Assessments for the secondary stakeholders in the five science research environments were developed and deployed in priority order. Through these assessments a more complete picture of the complex environmental science research process emerged, facilitating the work of cyberinfrastructure developers as well as those involved with community engagement.

Figure 4. DataONE data lifecycle (from http://www.dataone.org/best-practices)

Personas and usage scenarios

The stakeholder assessments discussed above provide a picture of the prevalence of current attitudes, perceptions, practices, and requirements. However, understanding how a member of each stakeholder group would engage in the research process is especially useful when developing tools and services for data services and management.

One way for user-focused service designers to characterise the stakeholders in a proposed service is to build 'personas' representing key characteristics likely to affect their use of the service, and develop scenarios representing typical use cases (48). In terms of the design phases referred to earlier, *stakeholder profiles* help to scope user needs in the discovery phase. In the design phase *personas* help align the archetypal user characteristics with the emerging design concepts. *Scenarios* then describe in more detail how the service will work for users that the personas represent.

These approaches are exemplified by the DataONE project's design process (Michener et al, 2012). The DataONE network supports scientists and other stakeholders to engage with the relevant science, data, and policy communities. It also facilitates easy, secure, and persistent storage of data, and disseminates tools for data discovery, analysis, visualization, and decision-making. As mentioned in Section 3.2 DataONE followed up their stakeholder analysis with 'baseline assessments' of the stakeholders' current practices, perceptions and needs relevant to all stages of the data lifecycle. These were online surveys, similar in scope to DAF surveys and data curation profile questions. Relative to these the DataONE survey is more detailed on service roles and on stakeholders' demographic characteristics.

Based on the survey findings and other materials such as the interviews from the DCP, personas were written to describe 'typical' stakeholders in both primary and secondary categories. They included research scientists in early, mid and late career; a scientist at a field station; a data modeler; a data librarian; a citizen scientist; and a university administrator' (49). The personas were then related to a set of use cases representing the intended functions of DataONE tools and services, which were then fleshed out into scenarios describing how particular capabilities would be used at a specific point in the data lifecycle, such as enabling searching across multiple data sources.

Use cases and scenarios that focus on the tasks in which users engage can provide a clear picture of how someone may work with the tools or services during the research and data curation cycles. A use case enumerates the actions a user takes when working with the system which helps identify needed features, and is a tool commonly used by software developers. A scenario is similar to a use case, but is presented as a narrative that is developed by looking at the potential goals and motivations of the user to describe how the system might be used.

Use cases and scenarios focus on the tasks users undertake rather than on the users, therefore they can be augmented by creating personas. These are a representation of the archetypal user including her requirements as well as affective perspectives such as emotions. Personas are usually created from data collected from the users either from assessments, interviews or through ethnographic means. Contextualizing this data makes personas powerful tools for project development in terms of identifying value propositions and scenarios to guide tools and services development even when users cannot be directly consulted. For example, in DataONE, members of the socio-cultural and usability and assessment working groups developed personas for individuals in the primary and secondary stakeholder communities, then compiled a list of tools researchers would like to have as well as a list of value-added items that DataONE provides.

5.5 Development workshops - from hackathons to mashups

Scenarios, user stories or lightweight statements of functional requirements are still relevant even in agile development methods like SCRUM that abandon the traditional approach of a separate 'requirements' stage leading to a lengthy specification. In any software development methodology the overriding goal is to enable the 'owner' of the product or service to clearly articulate what needs to be built, and to define what high quality means to them (50). The emphasis of agile development is on turning around useful products or services quickly, by developing something that works to the extent that the intended users are able to give feedback on how it can be improved.

Workshops are another approach to quickly identifying development priorities and prototyping the solutions, especially event formats that are loosely structured for intensive collaboration between developers, users and other stakeholders. Web development has spawned a growing range of these 'unconference' style events, and the digital curation and preservation community has been quick to adopt and adapt them.

[49] *Michener, W. K., Allard, S., Budden, A., Cook, R. B., Douglass, K., Frame, M., Vieglais, D. A. (2012) Participatory design of DataONE—Enabling cyberinfrastructure for the biological and environmental sciences. Ecological Informatics, (0). doi:10.1016/j.ecoinf.2011.08.007*

[50] *Moccia, J. Agile Requirements Definition and Management. Retrieved from: http://www.scrumalliance.org/community/articles/2012/february/agile-requirements-definition-and-management*

One example is CURATEcamp (51), inspired by the similar Hackfest (52) and BarCamp (53) event formats. Their common aim is to encourage a diverse group to attend and take part in solving a problem by sharing ideas and approaches, with more emphasis on learning from each other than on the software produced. CURATEcamp events have helped build a community around the 'curation microservices' approach (28). An event will typically involve:

- An expectation that everyone gives a demo, presents a talk, drives a discussion, or participate in a panel or roundtable
- An open agenda of parallel sessions to identify topics or 'projects' that different groups may tackled over the course of the event
- Participants vote with their feet, participating in whichever group they feel they can contribute to
- One session reflects on how the event could have worked better
- An emphasis on socialising and shared meals

This style of event has become broader in appeal, originally catering for the more technically-oriented. For example from 2009-12 the DevSci workshop series (54) helped establish a software development community around UK institutional repositories. More recent events with similar aims have broadened the discussion scope to digital content issues, opportunities for preservation, and reuse of data or tools. For example the AQuA and SPRUCE digital preservation projects organized 3-day 'mashup' events. Practitioners were invited to these as well as developers. In the mashup workshop format, practitioners give short talks as 'collection owners', presenting examples of data and their preservation goals and issues. Tasks and challenges are identified from discussion, and the facilitators provide wiki templates to record how solutions are addressed over the course of the event (55,56).

This format can be extended beyond preservation issues, for example a similar format has been used in the life sciences data management project BRISSkit. Here a 'Community Meet and Hack' event was organised around clinical data management and governance issues, looking to match these with appropriate open source software solutions (57).

Digital Curation Centre has used 'roadshow' events to bring together institutional stakeholders . These combine researcher-led case studies on data management issues affecting their projects, with group sessions to frame and prioritise institutional support improvements. These are oriented to planning institutional RDM services, though not to developing specific solutions in 'hack day' style. Some institutions however have the development of discipline-specific RDM platforms as a core part of their strategy. Monash University for example has a record of embedding development in specific research communities, with a view to ensuring the sustainability of institutional RDM.

Events that bring together specific research units and service providers to understand their data issues and identify working solutions may well become essential to ensure take up of institutional RDM services.

6. Next Steps and Future Challenges

6.1 Managing the requirements

The perennial challenge in delivering and developing RDM services will be to identify which of the requirements that have been identified are generic enough to be supported across the institution, yet specific enough for the institution to provide local support, rather than draw on external infrastructure and services. Generally these decisions will be about balancing priorities and resources, guided by a forward-looking strategy or roadmap for taking the RDM service forward. Some requests for support received or elicited may be met without any requirement for service changes, and satisfied by drawing on existing expertise and guidance resources. Of those that do call for a change in the service, some may imply a need for reconfiguring the software platforms used, other might require application development, while others can best be met by updating local online guides or using externally-sourced tools and resources.

In the above respects RDM is no different from other academic support services institutions deliver. The methods used to manage change in these will be useful for RDM. Many organisations manage change in online services and support using the ITIL (IT Infrastructure Library) approach to IT service management ( ). The CMMI (Capability Maturity Model Integration) approach to process improvement is also widely used ( ). CMMI in particular describes Requirements Management steps. Commercially

[51] *About CURATEcamp. (n.d.). Retrieved December 3, 2013, from http://curatecamp.org/about*
[52] *Tennant, R. (2003, November 15). Where Librarians Go To Hack. Library Journal. Retrieved January 3, 2014, from http://lj.libraryjournal.com/2003/11/ljarchives/where-librarians-go-to-hack/*
[53] *BarCamp. (2013, December 27). In Wikipedia, the free encyclopedia. Retrieved from http://en.wikipedia.org/w/index.php?title=BarCamp&oldid=583513559*
[54] *About | DevCSI | Developer Community Supporting Innovation. (n.d.). Retrieved from http://devcsi.ukoln.ac.uk/about/*
[55] *Wheatley, P., Middleton, B., Double, J., Jackson, A., & McGuinness, R. (2012). People Mashing: Agile digital preservation and the AQuA Project. In http://ipres2011.sg/conference-procedings. Retrieved from http://eprints.whiterose.ac.uk/43837/*
[56] *Wheatley, P. (2013). Just what is a SPRUCE Mashup and what's in it for me? Retrieved December 2, 2013, from http://wiki.opf-labs.org/pages/viewpage.action?pageId=13041673*
[57] *BRISSkit Community Meet and Hack. (n.d.). Retrieved December 2, 2013, from https://www.brisskit.le.ac.uk/civicrm/event/info*

available tools support requirements management, e.g. to help ensure that use cases and requirements statements can be traced to service changes, and vice versa.

While ITIL, CMMI and other generic project/ process management approaches offer concepts applicable to RDM services, this does not necessarily mean that RDM support services should be fully integrated with those for IT support, Library helpdesks, or any other 'helpdesk' service. There may be an efficiency case for such integration, but it is also important to consider the differences between research practice and other kinds of institutional activity that were mentioned in Section 3. Project managers will need to consider whether the local research culture would take up services embedded in other support areas, and assess the risk of front line support staff being unable to differentiate RDM issues from other kinds of support request.

6.2 Challenges of scale and complexity

> *"…comprehensive data management will require more than depositing data into repositories. It will require an understanding of the overall context regarding data, engagement with the researchers who produce the data and the provision of services that account for data as primary, compound objects within a broader scholarly communication landscape"*
> Sayeed Choudhury (61)

Whatever approach you take to discovering requirements for your RDM project, the approach will need to be regularly reviewed to make sure it can accommodate new challenges of scale and complexity. As we have seen in sections 2 and 3 these are challenges that cut across research and data lifecycles. The value to researchers of a more formal and coordinated approach to research data management often arises indirectly, because of the research opportunities of working at scale and the complexities that technology brings to their practice. So it is important that institutions can similarly adapt their support systems to reflect changes in the environment for digital research. These include rapid growth in scholarly communication services, and in cyber-infrastructures or research infrastructures that bypass institutional boundaries.

Scholarly communication is changing to accommodate new forms of data publishing, driven by changing demands to peer review data, and to make more content accessible for mining and integration with new 'big data' sources emerging across disciplines. Research infrastructures are important because they are growing in scale to meet these challenges, and providing new kinds of tools and middleware services to underpin broader access by more people to yet more kinds of data.

Currently institutions' RDM practices are shaped by those in place for managing conventional scholarly outputs, especially journal articles. In the UK, research council policies encourage this. Research quality assessment processes continue to focus on conventional scholarly outputs, which need institutional repositories to manage them. Often data management policies emphasise retention of data 'underlying' these outputs, with a similar institutional repository favoured as the place of deposit if no disciplinary repository can be found.

That may change faster than we think. For example, data mining may highlight the value of data repository contents, stimulating competition between institutional and domain-based repositories to attract depositors. Also, the complex relationships between published data and software must be addressed to respond to a growing 'crisis of reproducibility' (62). Current RDM models frame data sharing in terms of choosing the most appropriate repository to deposit in. However, we may see requirements quickly emerge to coordinate records of 'what data has been used where' that span a range of repositories, published articles and other online sources. The assumed norm of research findings depending on a well-bounded dataset is already changing, e.g. where researchers are deriving findings from very large datasets that are already shared, and cannot be deposited in a single repository. Policy models are also changing. The US National Science Foundation (NSF) policy for example has already changed to make all scholarly outputs including datasets credit-worthy (63).

A short-term challenge for requirements discovery is to move from a 'one off, one size fits all' model of interviewing focused on the researcher's Data Management Plan or (at a broader scale) informing the institution's RDM policy. The approaches mentioned in the guide all identify 'good questions', and examples

[61] *Choudhury, S. (2013). 'Case Study 1: John Hopkins University Data Management Services' In. Pryor, G., Jones, S., & Whyte, (Eds.) A. Delivering research data management services: fundamentals of good practice. [S.l.]: Facet Publishing.*
[62] *Lynch, C. (2013) 'The Next Generation of Challenges in the Curation of Scholarly Data' In. Ray, J. (Ed.) Research data management: practical strategies for information professionals. Purdue University Press*

of their use can be found from the sources identified below. There is much scope to refine these further, to help research support professionals to easily identify questions appropriate to different contexts.

RDM support professionals will need good 'requirements questions' that suit different stages in the research lifecycle and recognise variations in that lifecycle. For example an, RDM consultancy service offering researchers support with data sharing close to the end of their project may need to ask more detailed questions than were asked when the pre-award Data Management Plan was being written. The questions may also need to be different depending on how and where they are asked (online, one-to-one discussion, group workshop). They should match varying levels of maturity across different areas of support. For example the institution will likely have strengths in particular areas e.g. metadata advice, or guidance on using external data repositories; and want to ask more detailed questions about requirements for those areas.

Change in the RDM environment is a certainty whatever shape that change takes. The requirements methods described in this guide can help your organisation respond to the most basic needs. To really provide a continually improving service, providers need to listen and respond to changes in research lifecycles and gradually accommodate more complex needs.

[63] *Piwowar, H. (2013). Altmetrics: Value all research products. Nature, 493(7431), 159–159.*

Further Information

General: Alexander, I., Beus-Dukic, Ljerka. (2009). Discovering requirements: how to specify products and services. Chichester, England; Hoboken, NJ: Wiley.

Data Asset Framework: Digital Curation Centre. Available at: http://www.dcc.ac.uk/resources/repository-audit-and-assessment/data-asset-framework

CARDIO Collaborative Assessment of Research Data Infrastructure and Objectives. Digital Curation Centre. Available at: http://www.dcc.ac.uk/projects/cardio

DM Vitals: University of Virginia Library Data Management Consulting Group. Available at: http://dmconsult.library.virginia.edu/

Data Curation Profiles Toolkit: Available at: http://datacurationprofiles.org/

Stakeholder profiles, personas and scenarios: DataONE Sociocultural Issues Working Group. Available at: http://www.dataone.org/working_groups/sociocultural-issues-working-group

Mashup workshop format: Paul Wheatley (2013, Nov. 28) Open Planets Foundation Wiki 'Just what is a SPRUCE Mashup and what's in it for me?' Available at: http://wiki.opf-labs.org/pages/viewpage.action?pageId=13041673

Acknowledgements

The authors wish to thank colleagues for their input, particularly Rebecca Koskela, Executive Director of DataONE, and Kerry Miller, RDM Coordinator at the University of Edinburgh.

The DCC programme of institutional engagement is funded by Jisc (and from 2009-2011 also by the Higher Education Funding Council for England). DataONE is funded by the US National Science Foundation under a Cooperative Agreement (#083094)

Follow the DCC on Twitter: @digitalcuration and #ukdcc