Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 33 (2014) 215 – 222

Procedia Computer Science

CRIS 2014

Emerging good practice in managing research data and research information within UK Universities

Joy Davidson (a), Sarah Jones (a), Laura Molloy (a), Ulla Bøgvad Kejser (b)

(a) Humanities Advanced Tehcnology and Information Institute (HATII), Glasgow G12 8QJ, UK
(b) Royal Library of Denmark, Copenhagen DK-1016, Denmark

Abstract

Sound data intensive science depends upon effective research data and information management. Efficient and interoperable research information systems will be crucial for enabling and exploiting data intensive research however it is equally important that a research ecosystem is cultivated within research-intensive institutions that foster sustainable communication, cooperation and support of a diverse range of research-related staff. Researchers, librarians, administrators, ethics advisors, and IT professionals all have a vital contribution to make in ensuring that research data and related information is available, visible, understandable and usable over the mid to long term. This paper will provide a summary of several ongoing initiatives that the Jisc-funded Digital Curation Centre (DCC) are currently involved with in the UK and internationally to help staff within higher education institutions prepare to meet funding body mandates relating to research data management and sharing and to engage fully in the digital agenda.

Keywords: research data management; Jisc; Digital Curation Centre; Collaboration on Clarifying the Costs of Curation; Faciliate Open Sceince Training for European Research;

doi:10.1016/j.procs.2014.06.035

1. UK context

The notion of access to research data as a public good has been around for some time[1] and continues to pick up momentum[2] in the UK and globally. Over the past decade, individual research councils and funding bodies in the UK have issued specific policies relating to research data management (RDM) and sharing. Most RCUK funding bodies place the responsibility for RDM and data sharing with the individual researcher however the Engineering and Physical Sciences Research Council (EPSRC) instead specifies requirements[3] aimed at the institution as a whole. EPSRC has stated that any UK university seeking future funding should be able to comply with their nine key requirements by May 1st, 2015. As a major source of research funding in the UK, the EPSRC requirements have prompted many UK universities to review their current research data management practices and research environments to identify what additional systems and support will be needed to ensure that compliance can be realised. RCUK are keen to align the individual research data policies and in 2011 they released their Common Principles on Data Policy[4]. This paper introduces some of the work being undertaken by the Jisc-funded Digital Curation Centre (DCC) to assist UK HEIs in improving their capacity for research data management and sharing.

2. Understanding funding bodies' policies

As various research data management and sharing policies emerged from individual funding bodies in the UK and internationally, the DCC undertook a period of desk-based research to compare the various policies and to produce a report summarising the current state of the art. It soon became clear that the amount of time and effort needed to trawl through these individual policy documents was substantial and represented a major barrier for any single UK institution wanting to get to grips with what was required by each of the main funders. To help make the individual requirements of each funder easier to understand and ultimately easier to comply with, the DCC produced a single table summarising the key requirements of nine UK funders[5].

● Full Coverage ◐ Partial Coverage ○ No Coverage

	Policy Coverage		Policy Stipulations					Support Provided			
Research Funders	Published outputs	Data	Time limits	Data plan	Access/ sharing	Long-term curation	Monitoring	Guidance	Repository	Data centre	Costs
AHRC	●	●	●	●	●	◐	○	●	○	◐	◐
BBSRC	●	●	●	●	●	●	●	●	●	◐	●
CRUK	●	●	●	●	●	●	●	◐	●	○	○
EPSRC	●	●	●	◐	●	●	●	◐	○	○	●
ESRC	●	●	●	●	●	●	●	●	●	●	◐
MRC	●	●	●	●	●	●	○	◐	●	○	◐
NERC	●	●	●	●	●	●	●	●	●	●	◐
STFC	●	●	●	●	●	●	●	◐	●	◐	◐
Wellcome Trust	●	●	●	●	●	●	●	●	●	◐	●

Figure 1 - DCC Overview of funders' data policies

The table has proven to be a valuable resource for research support staff and researchers alike as they prepare new bids for funding. This table is updated regularly and represents a real efficiency for research administrators working

[1] OECD Principles and Guidelines for Access to Research Data from Public Funding http://www.oecd.org/science/sci-tech/38500813.pdf
[2] Panton Principles http://pantonprinciples.org/
[3] EPSRC Policy Framework on Research Data expectations http://www.epsrc.ac.uk/about/standards/researchdata/Pages/expectations.aspx
[4] http://www.rcuk.ac.uk/research/datapolicy/
[5] DCC overview of Funders' Data Policies Table http://www.dcc.ac.uk/resources/policy-and-legal/overview-funders-data-policies

within UK universities who would have otherwise needed to compile this information locally. The DCC are keen to work with other support providers across the international community to expand this resource as international funding bodies release similar data management and sharing policies.

3. Working with individual UK universities to scope RDM and data sharing challenges and opportunities

From mid-2011 until mid-2013, the DCC engaged with over twenty individual UK universities on a one-to-one basis to help them review their current RDM practices, to streamline their workflows and to consider how their research systems might be integrated ahead of the EPSRC deadline of May 2015. Our engagement programme[6] aimed to increase data management capabilities amongst a range of HEI staff and to develop a suite of good practice examples that can be reused by other institutions. Findings and examples distilled from our engagements are available from the DCC website. Each of our engagements was designed in close consultation with senior management within the institution to ensure that we could provide specific support where it was most needed to improve their RDM and data sharing capacity. This included participation in steering groups, developing draft research data policies, and developing RDM training for institutional staff. A key finding from our work with universities has been the importance of initially identifying the range of disparate institutional and external systems being used by researchers and support staff, pinpointing the numerous stakeholders who should be involved in RDM activities, visualising research workflows, and collating the often vast array of relevant institutional policies that govern the research lifecycle.

In many cases, our engagement work started with a review of RDM practices using the Data Asset Framework (DAF) methodology[7]. DAF helps users to identify, locate, describe and assess how they are managing their research data assets and provides a set of methods along with an online tool to enable data auditors to gather this information. The DCC are currently working with Purdue University to explore the feasibility of developing an international question bank based on surveys conducted using the DAF methodology, questions developed through DAF use by Jisc Managing Research Data (MRD) programme projects, and through the Purdue Data Curation Profile toolkit[8]. We anticipate that this resource will be beneficial to any research intensive organisation wishing to develop survey instruments to assess their staff's current RDM practices and holdings. We have also been helping HEIs to assess their readiness to support research data management and curation using our Collaborative Assessment of Research Data Infrastructure and Objectives (CARDIO) tool[9]. To make readiness easier to assess, we have developed a specific CARDIO lens that reflects RCUK's Common Principles and the EPSRC's requirements.

4. Fostering RDM skills development

A core objective of the DCC is to increase institutional capacity for managing and curating research data. To this end, we deliver training targeted to researchers and PhD students as well as training specifically aimed at librarians[10] and research administrators. While generic, high level training is a valuable first step, discipline specific training is often vital to delve more deeply into research data management workflows and tools. Between 2010-2013, the Jisc Managing Research Data programme, under the RDMTrain umbrella, funded a number of projects to develop discipline specific training materials[11]. A condition for Jisc funding was that these training materials be made freely available for reuse within Jorum which is a Jisc-funded service to collect and share Open Educational Resources (OER). The DCC, in partnership with the Research Information Network (RIN) and Jorum provided support to the RDMTrain projects through the Jisc-funded Research Data Management Skills Support Initiative - Assessment,

[6] http://www.dcc.ac.uk/tailored-support/institutional-engagements
[7] Data Asset Framework http://www.dcc.ac.uk/resources/repository-audit-and-assessment/data-asset-framework
[8] Data Curation profiles Toolkit http://datacurationprofiles.org/
[9] http://cardio.dcc.ac.uk/
[10] http://www.dcc.ac.uk/training/rdm-librarians
[11] http://www.jisc.ac.uk/whatwedo/programmes/mrd/rdmtrain.aspx

Benchmarking and Classification (DaMSSI-ABC) project[12]. DaMSSI-ABC tested the Research Information and Digital Literacies Coalition (RIDLs)[13] draft assessment criteria for describing and evaluating training offerings[14], identified benchmarks for learning outcomes through the mapping of outputs to Vitae's Researcher Development Framework, and developed a classification scheme for RDM resources to assist content re-users to select the training materials which best meets their needs. As part of the DaMSSI-ABC project work, the draft classification scheme was built into a new Research Data Management user community within Jorum. DaMSSI-ABC also aimed to engage with professional bodies to endorse resources and promote good data management practice as we believe that this will be crucial for ensuring that RDM skills are embedded into disciplinary workflows.

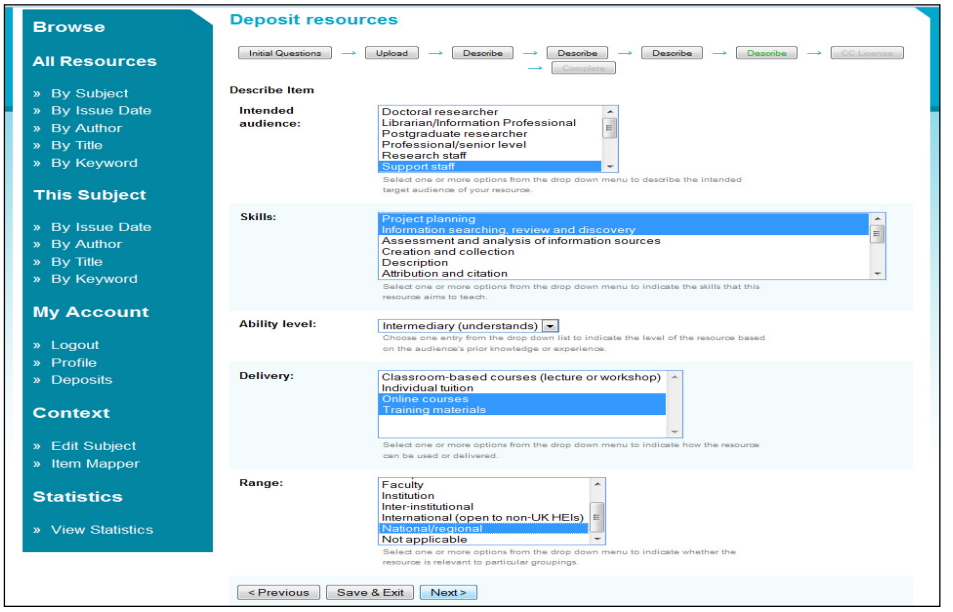

Figure 2 - Jorum resource deposit interface

In addition to a growing body of reusable training materials in the UK, there are a number of excellent training courses and materials emerging internationally. However, it can often be difficult to find and assess these resources. The DCC gathers and shares links to good training resources via our website [15] and, in February 2014, started work on the European Commission funded Facilitate Open Science Training for European Research (FOSTER) project[16]. FOSTER aims to accelerate knowledge and practices of open access and open science with a particular emphasis on enabling compliance with the open access policies and rules of participation set out for Horizon 2020. FOSTER will work to increase open science capacity among young researchers, project managers, administrators, and librarians. A key objective for the project is to equip a growing body of trainers capable of supporting researchers within European universities. The FOSTER project is currently in the process of identifying and describing reusable training content, using a version of the DaMSSI-ABC classification scheme, which will make the materials more visible for reuse by trainers and researchers alike through a dedicated portal. As part of our work, we are supporting the development and delivery of new training content to augment that which already exists. The first call for

[12] http://www.dcc.ac.uk/training/damssi-abc
[13] http://www.researchinfonet.org/infolit/ridls/
[14] http://www.dcc.ac.uk/training/damssi-abc
[15] http://www.dcc.ac.uk/training/materials-for-trainers
[16] http://www.fosteropenscience.eu/

FOSTER co-sponsored training events closed in March 2014. The second call for co-sponsored training events will be issued in October 2014.

5. Supporting data management planning

Data management planning is a topic of increasing importance to researchers, their host institutions and research funders alike. Many research funders now require a Data Management Plan (DMP) as part of grant proposals, and several UK universities have also mandated them for all research activity carried out within the institution. The DCC has done significant work in this area for some years. As noted above, we initially compared funders' data policies, examining requirements for DMPs and using this analysis to author a Checklist for a Data Management Plan. We also produced DMPonline - the first online tool to assist in the data management planning process. DMPonline is a web-based tool which helps researchers and research support staff to produce data management and sharing plans. It was first demonstrated at the Jisc conference in London in 2010 and has since undergone regular updates and some major redevelopment. Registered users of DMPonline are able to select from a range of templates that guide them through the process of developing a data management plan that reflects the selected funders' expectations. The tool provides general tips for each question along with pointers to any funder specific guidance. The tool has received international recognition, and we were very pleased when it was shortlisted along with the work of US colleagues on DMPTool for the DPC's Digital Preservation Awards at the end of 2012.

The DCC released version 4 of DMPonline in late 2013 and is currently working with a number of UK universities to develop and incorporate tailored local guidance to ensure that data management plans generated using DMPonline reflect the local infrastructure, policies, and procedures of the university. Access to local guidance saves researchers a great deal of time when completing new data management plans that will accompany new grant applications and allows institutions to help define and communicate preferred practices in terms of short and longer-term data storage and access mechanisms. As well as providing local guidance, we are working on the development of APIs to integrate DMPonline with other systems such as institutional repositories and research information management systems (CRIS). The DCC is participating in the Jisc-funded CASRAI UK pilot project which has recently defined a data management plan profile based on the DCC's DMPonline taxonomy and work will continue to develop further version(s) of the data management plan profile as new use cases emerge. We expect this data management plan profile will be of benefit to institutions as they work to integrate their disparate research information management and costing systems to support RDM and data sharing. By pushing and pulling data between DMPonline and other universities systems, we aim to prepopulate as much administrative information as possible to lessen the burden on researchers and support staff and to ensure join-up across systems. DCC will be launching an admin interface to DMPonline in Spring 2014 which will allow universities to develop their own customisations of the tool. They will be able to add templates, provide custom guidance (both at the institutional level and for individual schools, units or groups), provide example and suggested answers, and offer dropdown options to help researchers select appropriate answers. The DCC has run a number of webinars on customising DMPonline[17] and will be trialing the admin interface with an initial cohort of 20 universities from late Spring 2014.

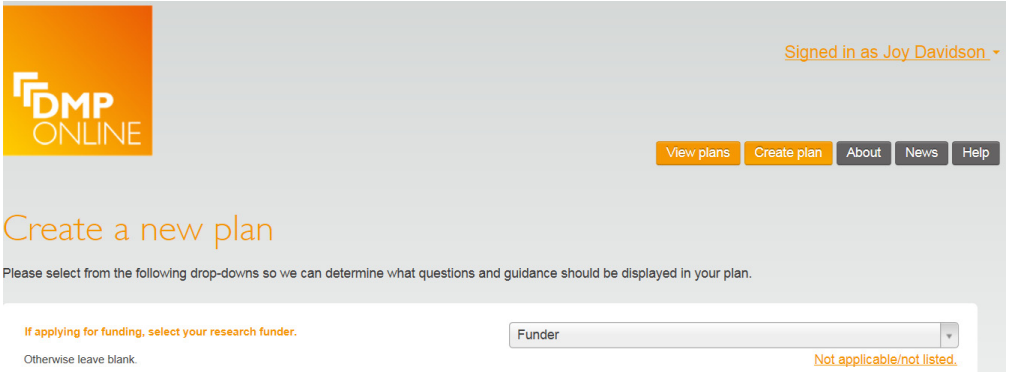

Figure 3 - DMPonline interface

Although requirements for Data Management Plans often focus on the grant application stage, DMPonline allows multiple phases/stages to be set to encourage DMPs to be regularly updated during the course of the research. The requirements from Natural Environment Research Council (NERC) and the European Commission Horizon 2020 Open Data Pilot are a case in point here, as both have differing requirements pre- and post-award. DMPonline represents these as consecutive phases in one project so the information the researchers provide in the initial stages can be fleshed out and built upon as details are agreed. The EC requests that new versions of DMPs are provided whenever major changes occur within projects, or at least in line with mid-term and final-review reporting. The DCC endorses this approach and encourages researchers to consider DMPs as active documents that are updated throughout the course of the project and to make use of DMPs as a communication tool between partners to help ensure effective data management.

6. Facilitating data discovery

While UK HEIs are beginning to understand the value of identifying, tracking and sustaining access to their research data along with other institutional assets, the development of institutional collections of research datasets are still generally in their infancy in UK universities. Many research disciplines already have international, discipline-specific databases and data archives (including the UK Data Service[18] and the various NERC- and STFC-funded data centres[19]) to meet researchers' needs. However - and as recognised by the EPSRC and other research councils - there is still a crucial role for HEIs to play in facilitating access to research data. Several UK HEIs are currently in the process of developing research data catalogues as a means of recording the research data generated through externally funded research within their institution and making it visible at the institutional level. The Jisc research data registry and discovery service (RDRDS)[20], currently in its pilot phase, aims to improve the discoverability of institutional research datasets to support their use nationally and across research disciplines. The RDRDS will not itself hold research data or metadata. Rather, solutions provided by the project must allow the RDRDS to harvest and expose metadata from a variety of institutional repositories and discipline specific data archives. RDRDS is working with participating institutions to test and appraise various software solutions, develop agreement around the metadata profile to be used and the harvesting workflow, and tackle feasibility and implementation challenges. The RDRDS project builds on lessons learned from earlier work carried out by the Australian National Data Service (ANDS)[21] in the development and implementation of their registry and discovery service Research Data Australia[22], which aggregates simple but textually rich metadata records for research data assets held in Australian universities and datacentres. The RDRDS pilot work includes 'test-driving' an instance of the Research Data Australia software, ORCA[23], amongst other solutions in order to discover, by practical experiment, the advantages and difficulties of various technical solutions. RDRDS has focused on fostering communication between universities and discipline-specific research datacentres from across the UK as this mix of varied experience and input will be crucial in seeding uptake and refining user requirements as the pilot is progressed.

As the next phase of the pilot project begins, we will explore the potential for RDRDS to be made interoperable with equivalent services in other countries and to expand the stakeholder community as more institutions build and launch data catalogues and repositories. We will also aim to engage in partnerships with developers of various research administration tools such as the RCUK Gateway to Research and the re3data registry of data to ensure that the discovery service allows researchers to find relevant datasets for reuse and also enhances the ability of research administrators to track research outputs and to trace and record impact.

---

[18] UK Data Service http://ukdataservice.ac.uk/
[19] NERC data centres http://www.nerc.ac.uk/research/sites/data/
[20] Jisc Research Data Registry and Discovery Service http://www.dcc.ac.uk/projects/research-data-registry-pilot
[21] Australian National Data Service http://www.ands.org.au/
[22] Research Data Australia http://researchdata.ands.org.au/
[23] ORCA registry software described in the ANDS context at http://www.globalregistries.org/orca.html

7. Assessing research data management costs and benefits

Supporting RDM and data sharing requires sustained investment from the institution. With increasing mandates to retain research data for the mid to long term, many research intensive institutions must now make informed decisions at the grant application stage about both the in-project data management costs and their potential to retain the data for as long as their funder requires, in most cases a period of at least ten years. This will involve being able to identify and estimate directly incurred costs during the active phase of the research project as well as assessing the potential budgetary impact on retaining the data over the longer term. The DCC is currently working as part of the the EC-funded 4C project (Collaboration to Clarify the Cost of Curation)[24] which aims to to ensure that a wide range of stakeholders are aware of and understand how to employ existing, relevant work on assessing curation costs. An additional aim of the 4C project is to examine how these resources might be made more fit-for-purpose, relevant and useable by a wider range of organisations operating at different scales in both the public and the private sector. These objectives will be achieved by a coordinated programme of outreach and engagement that will identify existing and emerging research and analyse user requirements. This will inform an assessment of where there are gaps in the current provision of tools, frameworks and models. The project will support stakeholders to better understand and articulate their requirements and will clarify some of the complexity of the relationships between cost and other factors. The outputs of this project will include various stakeholder engagement and dissemination events (focus groups, workshops, and a conference), a series of reports, the creation of models and specifications, and the establishment of an international Curation Costs Exchange framework (CCEx).

Figure 4 - mock-up of the CCEx interface

The CCEx is intended to be an online, virtual community platform for the exchange of curation cost information. We anticipate that the CCEx will be used in two key ways. Firstly, it will provide a starting point for a range of stakeholders interested in costing curation activities within their organisations. The CCEx will package and disseminate information gathered through the project including comparisons of current models, tips for making the case for investing in curation through a better understanding of economic determinants, risks, and benefits and examples of what peer organisations are doing. Secondly, CCEx will gather real cost information and data from

[24] Collaboration Clarify the Costs of Curation (4C) http://4cproject.eu/

partner organisations and a range of stakeholders, submitted to the exchange using a submission form. The form will capture calculation processes, metrics, effort statistics, and value calculations, from stakeholders in order to underpin future activity with empirical knowledge. The 4C project is currently seeking input on the CCEx and welcomes feedback from a range of stakeholder communities.

8. Conclusions

Being prepared to meet research data management and sharing requirements and to fully exploit the emerging data-driven research environment depends upon establishing and sustaining an integrated and supportive institutional research ecosystem. UK universities, many with the help of Jisc and the DCC, have made good progress in developing fledgling RDM services and support systems. As UK funding bodies mandated RDM and data sharing earlier than many other international funders, the lessons we've learned and the examples of good practice and emerging shared services will be of benefit to research intensive organizations outside the UK and may help to reduce costs and avoid duplication of effort. The DCC has just started its fourth phase of funded activity and will continue to work with international partners such as the Research Data Alliance (RDA), euroCRIS and CASRAI to identify examples of good practice and to disseminate these as widely as possible.