



Citation for published version:

Patel, M, Lyon, L & Takeda, K 2014, 'Assessing requirements for research data management support in academic libraries: introducing a new multi-faceted capability tool', Paper presented at Libraries in the Digital Age (LIDA) 2014, Zadar, Croatia, 16/06/14 - 20/06/14.

Publication date:
2014

Document Version
Publisher's PDF, also known as Version of record

[Link to publication](#)

Publisher Rights
CC BY

University of Bath

Alternative formats

If you require this document in an alternative format, please contact:
openaccess@bath.ac.uk

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Assessing requirements for research data management support in academic libraries: introducing a new multi-faceted capability tool.¹

Liz Lyon²

School of Information Sciences, University of Pittsburgh, USA. elyon@pitt.edu

Manjula Patel

UKOLN Informatics, University of Bath, UK. M.patel@ukoln.ac.uk

Kenji Takeda

Microsoft Research Outreach. kenjita@microsoft.com

Abstract

This paper introduces the practice of assessing requirements for research data management (RDM) support in academic libraries, building on concepts of maturity, capability and readiness. An overview of existing RDM assessment methodologies, tools and instruments is presented, with institutional exemplars from the UK and the US. Drawing on consultations with the eScience community, we describe the development of the Community Capability Model Framework (CCMF), the derived capability factors and the CCM Profile tool. Finally, a Case Study for Agronomy research data is presented, showing how the CCM Profile tool can be applied to disciplinary research, to provide summaries and visualisations of data-intensive capability, which may inform planning for RDM support services in academic libraries.

Keywords: Research data management services, Capability assessment models, Data-intensive science

Introduction

The need for academic libraries to provide a range of research data management (RDM) services to large-scale disciplinary projects and local researchers, reflects the increasingly data-intensive research process described as the “*Fourth Paradigm*” by Hey, Tansley & Tolle (2009). Additional drivers come from research funder mandates for

data management plans to be submitted as part of the grant application process e.g. European Commission Horizon2020 Projects and the National Science Foundation in the United States, and from the reputational risks associated with problems of data access, data quality and integrity over time. Institutions such as universities and research units with substantive research portfolios, have significant data assets; this legacy data may be in a wide range of storage locations, formats and types. A key early activity in developing RDM services in libraries, is to understand the variety and state of this legacy data. In parallel, it is also crucial to understand the range of disciplinary practices and norms which underpin the research data lifecycle. This disciplinary evidence gives a perspective on the capability and readiness of the particular domain community for data-intensive research. Taken together, these two approaches provide a rich foundation to inform the development of RDM services. In this paper, we present a review of RDM assessment methodologies, and introduce a new tool for libraries, based on established capability and maturity models (Crowston & Qin 2011).

Theoretical Framework and Literature Review

Libraries now have a range of assessment methodologies to capture RDM requirements associated with legacy data and current research practice. One of the early tools to be developed was the Data Asset/Audit Framework (DAF). The need for such a tool was highlighted in a Recommendation in the “*Dealing with Data Report*” by Lyon (2007): “*JISC should develop a Data Audit Framework to enable all Universities and colleges to carry out an audit of departmental data collections, awareness, policies and practice for data curation and preservation (Rec 4).*”

¹ This work has been funded by Microsoft Research Outreach.

² For correspondence.

The DAF tool was developed by HATII at the University of Glasgow, UK and is described in detail by Jones, Ross and Ruusalep (2008). A four-stage methodology was conceived which involves a planning stage, an asset identification and classification stage, an asset assessment stage and a reporting stage. The tool is designed to be used by non-specialist staff. The DAF methodology was validated in pilot developments at each of the project partner sites at the School of GeoSciences, University of Edinburgh and the Innovative Design & Manufacturing Research Centre, University of Bath. Additional UK sites (King's College London, Imperial College London and University College London), also carried out DAF pilot studies. An overview of the methodology and pilot outcomes is available (Jones et al 2008) with full reports from the pilots at Edinburgh - Ekmekcioglu and Rice (2009), UCL – Polydoratou (2009) and Imperial College – Jerrome & Breeze (2009). Note that the tool is now known as the Data Asset Framework.

The UK Digital Curation Centre has developed the CARDIO (Collaborative Assessment of Research Data Infrastructure and Objectives) and CARDIO-lite tools, which provide a maturity-based assessment of research data collections and practices. The three-legged stool digital asset management model from Anne R. Kenney at Cornell University, has been used with foundational elements of technology, organization and resources. Statements relating to each of these aspects are then presented to the user, who will rate maturity on a five-point scale. An overall picture of the position for research data curation is then provided. The CARDIO-lite tool has been remodeled into a Mini Quiz by Fowler (2012) at the University of the West of England.

The application of interview-based mechanisms has also been used to assess RDM practices. Three contrasting approaches are the Data Curation Profiles Toolkit developed at Purdue University by Witt et al (2009), scorecard approaches such as DMVitals, Sallans & Lake (2013) and research persona development e.g. Lage et al (2011). The Profiles Toolkit represents a well-documented suite of instruments used to gather information about disciplinary data collections and practice. There is a User Guide, Interviewer's Manual, Interview Worksheet and a basic template. The interviewer is prompted to probe particular data lifecycle areas and data management behaviours in some depth, to gather a full picture of the curation requirements of a particular domain. Examples have been collected in a Data Curation Profiles Registry. DMVitals developed at the University of Virginia, is an Excel-based tool with three types of worksheet: interview questions, data management categories and the report sheet. The latter contains sections for a sustainability index as percentage ratios which are grouped into five colour-coded levels of maturity; these are followed with recommendations and action statements. The personas approach developed at the University of Colorado Boulder, seeks to categorise researcher profiles based on their interview responses to a

fixed set of nine questions about their data curation practices. Results are then conflated into one of eight personas e.g. “*Very interested, has no support*”, “*Receptive, already has a repository*”, “*Not interested, competitive discipline with proprietary funders*”. This gives a landscape of perspectives and data requirements, which can subsequently inform the development of RDM services to researchers who share traits.

Survey instruments which may be based on Web software tools such as SurveyMonkey, have also been used to gather requirements about data curation practices and legacy data. Two examples which both provided rich and detailed quantitative and qualitative material, are those from the University of Bath, Pink et al (2013) and Knight (2013) at the London School of Hygiene & Tropical Medicine.

Research Questions

Given the character of existing tools, we sought to develop an assessment tool which addressed the widest range of parameters affecting data-intensive research: policy drivers, legal environment, researcher training, disciplinary practice/culture, technical infrastructure. The aim was to develop a simple self-assessment tool primarily for researchers to use, to catalyse the collection of disciplinary profiles to inform RDM service development in libraries, research funder investment decisions and policy-making in the data arena.

Methods

We gathered input and ideas from eScience researchers across a range of disciplines and data curation communities, to scope development of the community capability model (CCM) framework. This was achieved via a series of six international workshops (Cambridge MA, Melbourne Australia, Stockholm Sweden, Bristol and York, UK, and Amsterdam, Netherlands), which explored different maturity models and scoped the data requirements landscape. The workshops also helped to pinpoint the capability factors and the visual presentation of the concepts. In addition, three mini case studies were completed which introduced policy and practice perspectives from different stakeholder groups: an academic institution (University of Bath), a research funding body (Economic & Social Research Council ESRC) and the research community (a group of Principal Investigators from eResearch South). This collated evidence informed a CCMF White Paper (2012) which articulated the foundations and structure of the model and its dimensions. A visualization of the model was derived as a basis for the CCMF-Profile template.

The Research Data Alliance (RDA) has provided a further arena to engage with a wider group of data stakeholders and an RDA CCM Interest Group was established, meeting at the 2nd Plenary in Washington DC.

Results

The CCM Framework contains eight capability factors (Openness, Legal, Ethical & Commercial Considerations, Collaboration, Economic & Business, Skills & Training, Common Practices, Research Culture and Technical Infrastructure).

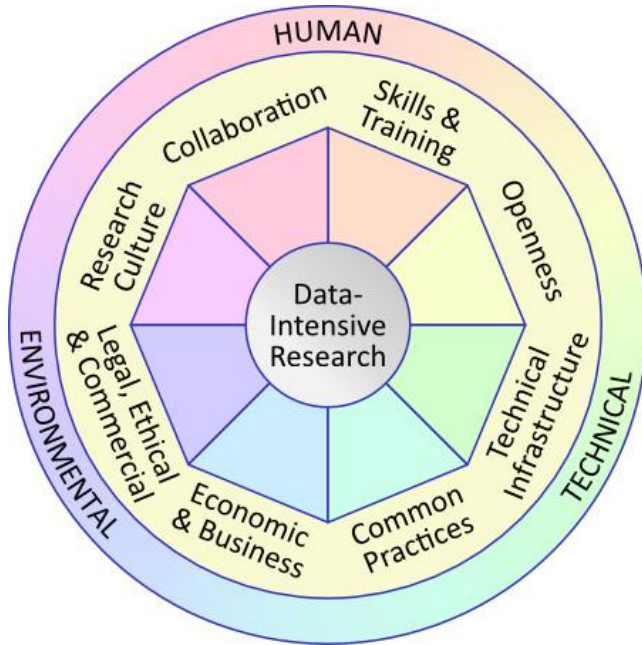


Figure 1: Community Capability Model Framework

We have developed an Excel-based CCMF-Profile template, which includes a range of characteristics associated with each data capability factor, allowing a deep analysis of each dimension. Five capability levels, ranging from “nominal” to “complete engagement”, are selected to describe the discipline, the project or researcher.

Figure 2 shows the Excel worksheet for the Collaboration capability factor, which has four characteristics exemplified by four questions. A participant scores each of these depending on where they estimate the current state to be (Nominal activity (1); Pockets of activity (2); Moderate activity (3); Widespread activity (4); or Complete engagement (5)). The sum of the scores provides an indication of how well the Researcher believes that the relevant community is engaged in terms of that particular capability factor.

	Nominal Activity (1)	Pockets of Activity (2)	Moderate Activity (3)	Widespread Activity (4)	Complete Engagement (5)	Comments (6-8)
1. Collaboration						Below pertinent, please be explicit of collaboration for N activities (e.g., # of labs, # of staff/group)
1.1 Collaboration within the discipline e.g. Use of guidelines and tools for data management	None or Limited	Departmental research groups	Collaboration across research groups within or between organisations	Discipline organised at a national level	International collaboration and consortia	
1.2 Collaboration and interaction across disciplines	None or limited	Individual researchers occasionally collaborate outside their disciplines	Disciplines collaborate through joint conferences or publications	Bilateral collaborations	Formal collaboration between research groups from several different disciplines	
1.3 Collaboration and interaction across sectors (e.g. public, private, government)	None or limited	Attempts have been made but are not considered successful	Despite successful exercises working with other sectors is not the norm - some barriers are generated	A discipline or group has gained experience of working closely with one or two sectors	Work successfully with several other sectors on different problems	
1.4 Collaboration with the public	None or limited	The public's involvement is limited to acting as subjects of study, user testing, etc.	Contact with the public is only through occasional appearance in the media e.g. news bulletins, TV programmes	Discipline sometimes participates, targeted media programmes are organised to engage the public e.g. science fairs	Dedicated programmes involving the public in research. Crowdsourcing/open science	

Figure 2: Worksheet for Collaboration Capability factor

The CCMF profile tool is being applied to assess RDM capability by academic libraries, disciplines and projects.

Case Study – Agronomy

At Purdue University, a Principal Investigator in the area of agronomy spent several hours customising the profile tool (iteratively), to make it more appropriate and accessible to other agronomists. For example, in a number of cases the original language was changed to make it more (sub) discipline-specific and the original examples used in the profile were changed to be more relevant to agronomy. In addition, the language in several of the questions was modified to clarify what was being asked for.

Following this process, the researcher and two other research agronomists spent less than one hour completing the localized profile tool. The results are shown in Figure 3.

In this case, all three researchers are agronomists, working in areas that are similar. Their research foci are different, but they do collaborate - all three share graduate students, methods and resources. Researcher 1 oversees a field station for studying water quality and works with other groups on campus; Researcher 2 works with environmental Nano materials with other groups regionally; and Researcher 3 works in crop physiology/ecology and works with national groups.

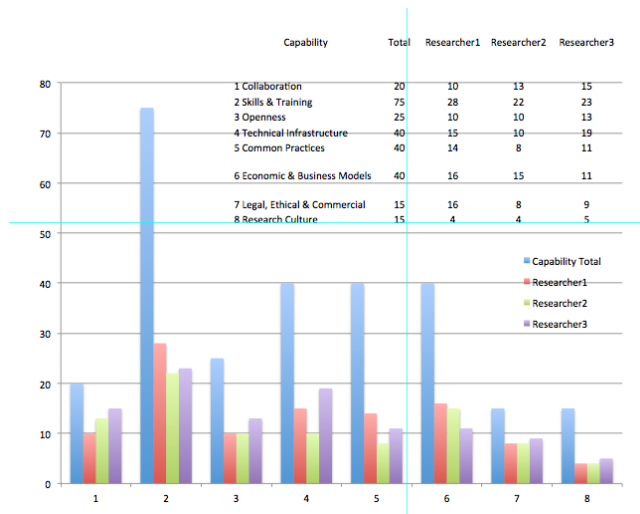


Figure 3: Results from Agronomy case study (adapted from a slide by Scott D. Brandt)

Even a cursory analysis comparing the three scores of the Researchers from least amount of activity or engagement in an area, to what would be most activity or engagement (indicated in blue), can be quite revealing. As can be seen, capability 2, Skills and Training, is the area in which there are the largest gaps, but there are relative gaps throughout. For Researcher 1, the highest score was 3 (8 times) and the mode was 2, or Pockets of Activity. For Researcher 2 Nominal Activity (1), was the mode, and the highest rating was a 4 (Widespread Activity) in Openness of literature. Researcher 3 had much more variety, selecting 4 (Widespread Activity) 7 times, and gave Openness of Literature a 5, indicating complete engagement.

Such analyses and comparisons based on self-assessment can be used to highlight trends and gaps within an area or discipline, which can then be used to plan or modify goals, priorities, policies and resource distribution, depending on the desired outcome.

We envisage that the CCMF-Profile tool will be particularly useful for undertaking longitudinal studies over a period of several years, to monitor progress in specific areas.

References and Citations

- Community Capability Model Framework White Paper (2012) <http://communitymodel.sharepoint.com/Documents/CCMDIRWhitePaper-24042012.pdf>
- Crowston, K. & Qin, J. (2011) A capability maturity model for scientific data management: evidence from the literature. *Proc. Am. Sci. Info. Technol.* 48(1), 1-9. DOI: 10.1002/meet.2011.14504801036.
- Ekmekcioglu C. & Rice R. (2009) University of Edinburgh: Data Audit implementation Project Final Report http://repository.jisc.ac.uk/283/1/edinburghDAFFinalreport_version2.pdf

Fowler S. (2012) UWE Maturity Modelling Case Study <http://blogs.uwe.ac.uk/teams/mrd/archive/2012/05/04/uwe-maturity-modelling-the-next-steps.aspx>

Hey, T., Tansley, S. Tolle, K. (2009). The Fourth Paradigm: Data-intensive Scientific Discovery. Microsoft Research. <http://research.microsoft.com/en-us/collaboration/fourthparadigm/>

Jerrone N. & Breeze J. (2009) Imperial College Data Audit Framework Implementation Final Report <http://repository.jisc.ac.uk/307/>

Jones, S. Ross, S. & Ruusalepp, R., (2008) DAF: a toolkit to identify research assets and improve data management http://www.data-audit.eu/docs/DAF_iPRES_paper.pdf

Jones S., Ball, A., Ekmekcioglu, C., (2008) The Data Audit Framework: a first step in the data management challenge <http://www.ijdc.net/index.php/ijdc/article/view/91/62>

Knight G. (2013) London School of Hygiene & Tropical Medicine: Research Data Management at LSHTM Web Survey Report <http://blogs.lshtm.ac.uk/rdmss/files/2013/04/LSHTM-RDM-Web-Survey-Report.pdf>

Lage K. Losoff B. & Maness J. (2011) Receptivity to Library Involvement in Scientific Data Curation: A Case Study at the University of Colorado Boulder. http://www.press.jhu.edu/journals/portal_libraries_and_the_academy/portal_pre_print/current/articles/11.4lage.pdf

Lyon, L. (2007). Dealing with Data: Roles, Rights, Responsibilities and Relationships. http://www.jisc.ac.uk/media/documents/programmes/digitalrepositories/dealing_with_data_report-final.pdf

Pink C. et al (2013) University of Bath: Research360 Faculty-Industry Data Requirements Report <http://opus.bath.ac.uk/36361/>

Polydoratou, P. (2009) UCL Data Audit Framework pilot implementation Final Report <http://discovery.ucl.ac.uk/15053/>

Sallans, A. & Lake S. (2013) Data Management Assessment and Planning Tools. In *Research Data Management: Practical Strategies for Information Professionals*, ed. Joyce Ray, pp 87-98, Purdue University Press.

Witt M., Carlson, J., Brandt, D.S. and Cragin, M.H. (2009) Constructing Data Curation Profiles. *IJDC* 4(3), 93-103. <http://ijdc.net/index.php/ijdc/article/view/137/165>

Curriculum Vitae

Dr Liz Lyon is a Visiting Professor in the School of Information Sciences (iSchool), University of Pittsburgh, where she teaches classes on Research Data Management and Data Infrastructures. Previously, Dr Lyon was Associate Director of the UK Digital Curation Centre and Director of UKOLN Informatics at the University of Bath. Liz is a frequent international speaker, giving keynotes in North America, Europe and Australasia. She has served on various Boards and Committees and is currently Co-Chair of the DataONE International Advisory Board.