Research Data Management Forum 12

Linking Data and Repositories (and other systems)

Jisc

University of Leicester, 18-19 November 2014

DCC

Contents

- Housekeeping
- Event overview and context
- Event goals
- Programme
- Event outputs

Housekeeping

- Fire alarm
- Bathrooms
- Bedrooms
- Drinks, dinner, breakfast and lunch
- Twitter hashtag: #RDMF12

Context

- Our last event took the topic of **Workflows and Lifecycle Models for Data Management**. This event will hopefully follow nicely on from that, as we start putting these models and workflows (and policies, and…) to the test by passing content between systems
- Seamless flow of content, both internally and across institutional boundaries, is one of the core goals of information management
- At a basic level, there is both a shared desire and a growing impetus to link scholarly publications to the datasets which underpin them, and to sustain these links for the longer-term
- But while the publications will generally be held in comparatively stable repositories, data (and metadata) may be created, held in, and accessed via, a variety of different systems...

Infrastructure overview

"**The local infrastructure does not exist in a vacuum and interacts with, and is dependent upon a range of other services and processes in an information ecosystem**. At one end of the continuum of research data curation is the local storage of data and metadata (data identification, description and documentation), usually accessible to the project team only. At the other end are international discipline-based data repositories or national data centres that publish research data, facilitating its discovery and access. Research institutions lie in the middle of this continuum and provide the means to move research data and metadata from their local, unpublished state to an international published state. **Some institutions now publish research data, either by modifying the institutional repository (IR) to accommodate datasets in addition to research papers or through a data repository, being a new instance of repository system running alongside the IR.** However, discipline-based repositories are considered the most appropriate facility for data publishing, due to their configuration for the data types and metadata formats associated with the research community they serve." - Lewis (2014)

Technical components

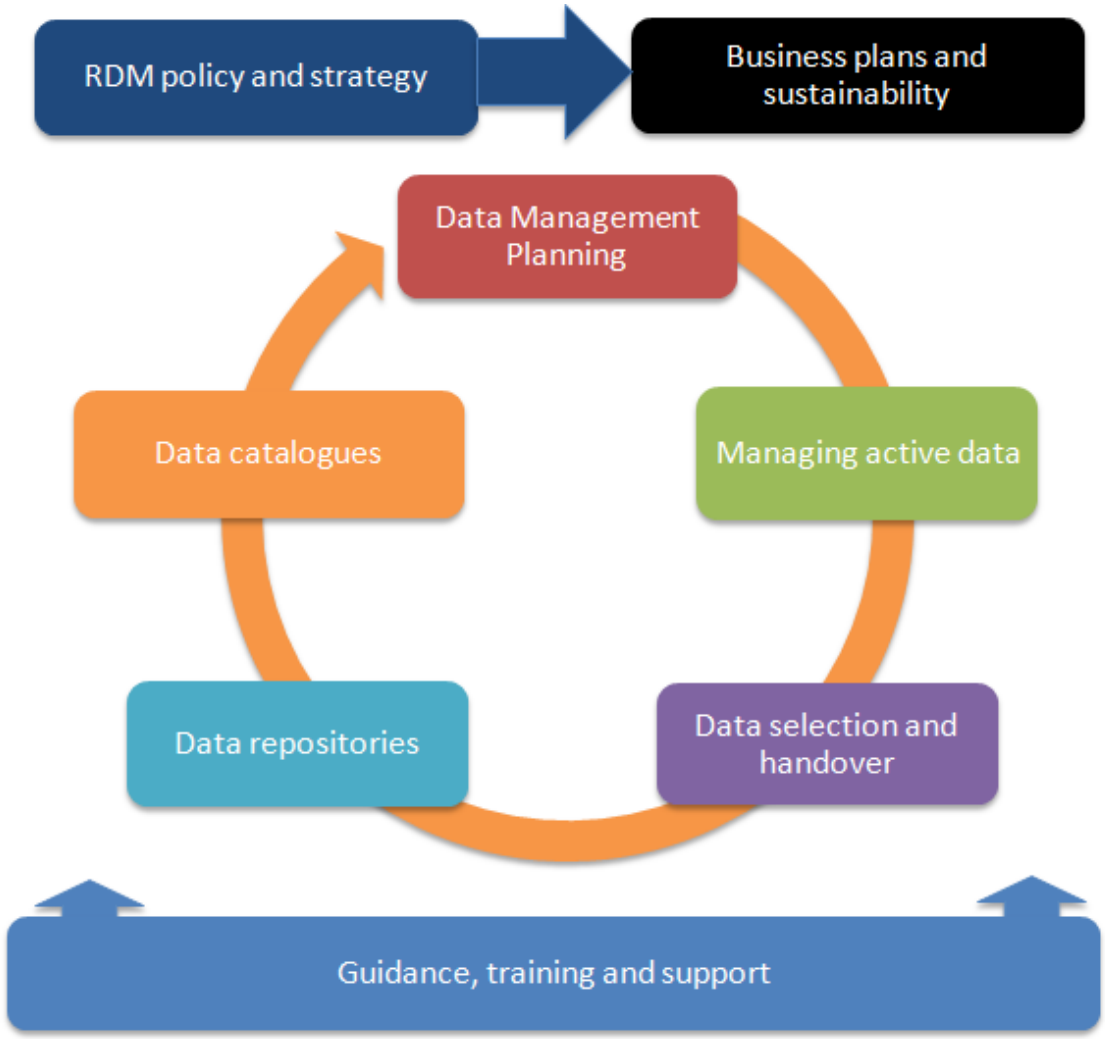
The major functional components of the RDM technical infrastructure for the institution are:

- **Metadata capture system** – in order to identify, describe and document the research data as they are created, captured and processed and record the context, conditions, variables and instrument settings. This may be accomplished manually, by the researcher filling in forms, or automatically, concurrent with data capture, by using appropriate equipment.
- **Active research data management system** – Active research data needs to be accessed rapidly, may require large computational resources and may require stringent security and access arrangements. A number of collaborative systems and virtual research environments have been developed to fulfil these requirements. These can be considered to comprise of a filestore and a **data registry** (sometimes known as a metadata store or asset registry).

Technical components (cont.)

- **Research Data Repository** – will be an appropriate place for preservation and publishing of archive research data for which there is no discipline-based repository or data centre available. The catalogue and archive functions of the repository may be separated.
- **Research Data Catalogue** – holds the metadata records of published (but not necessarily open access) research data. The data themselves may be held in a discipline-based data repository outside the institution or in an institutional data archive.
- **Research Data Archive** – preserves data not, or not yet, submitted to discipline-based data repositories. The associated metadata records will be held in the research data catalogue.
- **Current Research Information System (CRIS)** – manages the metadata associated with researcher identity, project information, research costing, grant applications and awards. (Lewis 2014)

Technical and nontechnical components



Components of a data management service : **http://www.dcc.ac.uk/resources/how-guides**

Interrelation

- **These components may overlap in function**, but need to be interoperable to provide seamless RDM. Alternative approaches to an infrastructure composed of diverse components, where ensuring interoperability may be problematic, are provided by data grids and micro-services**. The technical infrastructure must fit into the researcher workflow, making RDM processes automatic and virtually invisible to the researcher as far as possible**. This is so as not to burden the researcher with additional work or changes to their practice. Products, processes and practices that have been developed by a researcher community should be adopted, adapted and developed for the needs of other researchers, rather than new solutions developed.
- **The choice of technical infrastructure components and the approach of implementation will need to be considered with regard to the infrastructure and expertise already present.** Integrating and modifying existing components may be as expensive, in terms of development work, as implementing new infrastructure. Installing and configuring free open-source software may prove expensive in terms of development, compared with proprietary systems.
  - *John Lewis (May 2014) "A Review of Options for the Development of Research Data Management Technical Infrastructure at the University of Sheffield"*

Various positions / perspectives

- As data becomes increasingly accepted as a first-class research output in its own right, types of (re-)user become more diverse: research funders and university administrators, for example, take a growing interest in usage statistics and impact, as well as monitoring compliance with the relevant policies. Each of the stakeholder groups has its own outlook on what they need research data to 'do', and the kinds of system they need to interface it with.
- So who do we have at this event? Speakers from universities, funders and professional/scholarly associations, as well as delegates from the following stakeholder communities: repository managers, data centre staff, librarians and cataloguers, RDM service coordinators, research managers, software providers/developers…

What's new for #rdmf12?

- Breakout group approach
- Greater efforts to structure the event, and give it a stronger 'narrative'
- More institutional case studies
- Increased focus on concrete event outputs

Overall event goals

- To understand the environment in which repositories and data archives and interact, overlap or intersect
- To gain an improved understanding of the different perspectives and interests that different stakeholder groups may have
- To get more detailed information about specific software solutions
- To identify gaps and opportunities for collaboration and further work
- Others? That's up to you…

Programme (Tuesday)

Time	Presentation	Speaker
16.00	Welcome and introduction	Martin Donnelly, DCC
16.15	Keynote: “I’m leaving you… my data!” Practical research data sharing within your institution and the wider community	Jonathan Tedds, Senior Research Fellow and Director of Health and Research Data Informatics, University of Leicester, and Elizabeth Newbold, STM Content and Collections Leader, British Library (BL) and co-chair of the Research Data Alliance working group on Publishing Data Workflows
17.15	Group discussion	Facilitator: Martin Donnelly, DCC
18.00	End of sessions	
19.00	Drinks reception / dinner at 19.30	

Our keynote speakers

- Jonathan Tedds

- Elizabeth Newbold

Group discussion

Some questions and themes that the event might seek to address…

- How do we capture requirements for integrated RDM/IM systems?
- Beyond data and repositories, what other types of system are relevant?
- How do we cope with crossing institutional boundaries to communicate with systems we can't control?
- How do we hide/tailor the experience to suit the user?
- What are the opportunities for pooling of effort and resources? What are the barriers to this?
- What are the necessary infrastructural modifications (training, funding etc)?

There is a choice of three breakout groups tomorrow, each to be either facilitated or reported on by a DCC person. Discuss potential topics tonight over drinks/dinner, and we'll agree them in the morning.

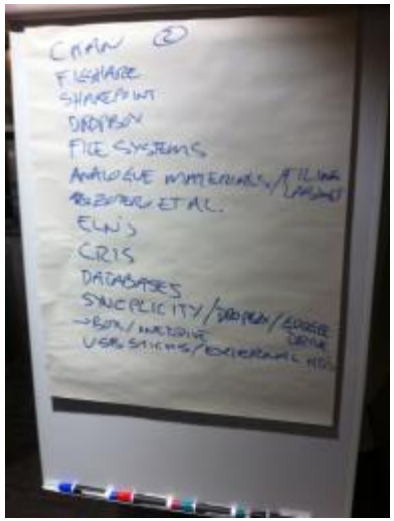
Programme (Wednesday a.m.)

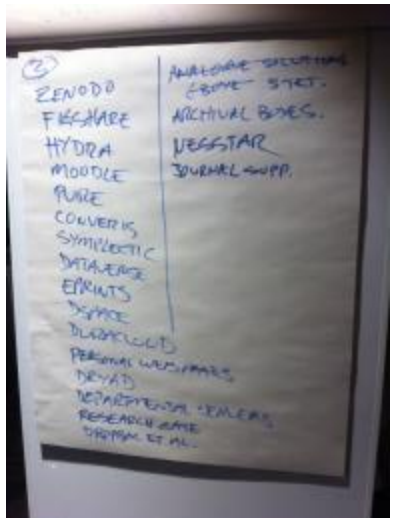
Time	Presentation	Speaker
09.00	Recap of yesterday's discussions, agree goals for remainder of event	Chair: Martin Donnelly, DCC
09.30	Presentation – Building long-term connections: from data to publications and vice versa	Catherine Jones, Information Systems Project Manager, Science and Technology Facilities Council
10.00	Institutional case study 1 – The University of Edinburgh RDM Programme: service interoperation, inc. integrating the RSpace electronic lab notebook within a university research infrastructure	Stuart Macdonald, RDM Coordinator and Associate Data Librarian, University of Edinburgh, and Rory Macneil, CEO, Research Space
10.45	Coffee break	
11.15	Institutional case study 2 – "Your flexible friends? Adapting and linking Eprints publications and data repositories in response to researcher needs"	Stephen Grace, Research Services Librarian, University of East London
11.45	Institutional case study 3 – Linking Hydra with other systems	Neil Stewart, Digital Library Manager, London School of Economics and Political Science
12.15	Lunch	

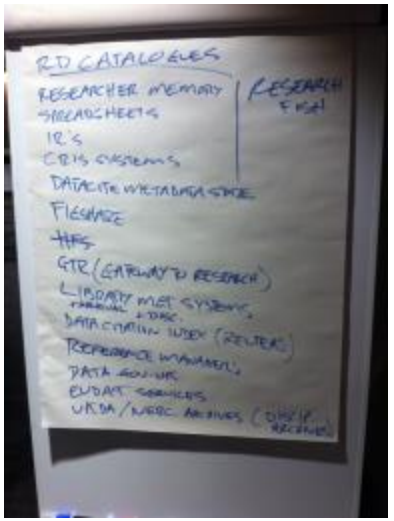
Programme (Wednesday p.m.)

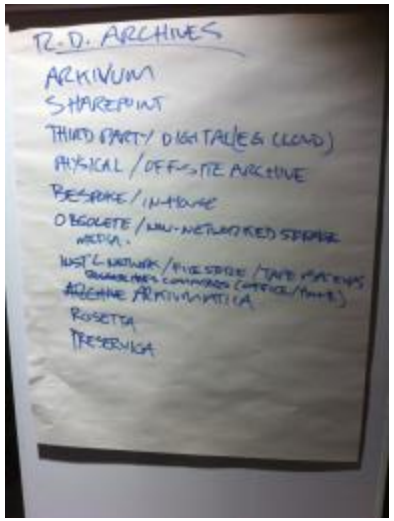
Time	Presentation	Speaker
13.00	Presentation – A Research Manager and Administrator's Perspective: From REF to Eternity	Simon Kerridge, Director of Research Services, the University of Kent and Chair of the Association of Research Managers & Administrators (ARMA)
13.30	Break-out groups: see separate slide (tea/ coffee also available)	Facilitators: Jonathan Tedds, Helen McEvoy, Jez Cope, Simon Kerridge
14.30	Report back from groups / discussion	Chair: Martin Donnelly, DCC
15.10	Jisc's "Research at Risk": From innovation to shared services	Daniela Duca, Jisc
15.30	Summary / parting remarks	Simon Hodson, Executive Director, CODATA
16.00	Ends	

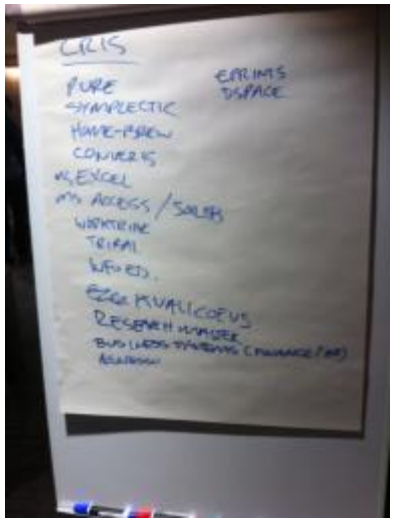
Yesterday's list of systems











- 100 systems across 7 categories (inc. duplicates)
- ~80 when duplicates removed

Systems list

Metadata capture system	Active research data management	Research Data Repository	Research Data Catalogue	Research Data Archive	Current Research Information System (CRIS)	OTHER:
MS Excel	CKAN	Zenodo	Researchers' memories	arkivum	Pure	email
Electronic lab notebooks	figshare	Figshare	MS Excel	sharepoint	Symplectic	zendto (southampton)
Paper records	sharepoint	hydra	IRs	third party digital (e.g. cloud)	"Homebrew" CRIS systems	twitter
Digital cameras	dropbox	moodle	CRISs	physical / offsite archive	Converis	publications
brisskit	file systems	pure	Datacite metadata store	bespoke / in-house systems	MS Excel	altmetrics
Laboratory equipments	Analogue storage (e.g. folders and archival boxes)	converis	Figshare	obsolete / non-networked storage media	MS Access	dmp online
Figshare	Reference managers (Zotero et al)	symplectic	Gateway to research	institutional network / filestore / tape backups	Worktribe	datacite
SPSS / Nudist (?)	Electronic lab notebooks	dataverse	Library and archival management and discovery systems	researcher's computer (at work or at home)	Tribal	orcid
Nvivo	CRISs	eprints	Reuters data citation index	arkivmatica	Info Ed	Je-S
Matlab	Databases	dspace	Reference managers (Zotero et al)	rosetta	Kuali Coelus	Re3data
Reference managers (Zotero etc)	Dropbox et al	duracloud	Data.gov.uk	preservica	Research master	Citation indexes
	USB sticks, external HDs	personal/work web pages	EUDAT services		Aggresso	The web
		dryad	Disciplinary archives (UKDA, NERC etc)		all other institutional information systems (e.g. hr, finance etc)	Google scholar
		departmental servers			Eprints	data mining tools
		research gate			Dspace	all other institutional information systems (e.g. hr, finance etc)
		Dropbox et al				OpenAire
		Analogue storage (e.g. folders and archival boxes)				Data visualisation tools
		nesstar				Data wrangling tools
		journal supplementary Information				Collaboration tools (e.g. Google Docs?)

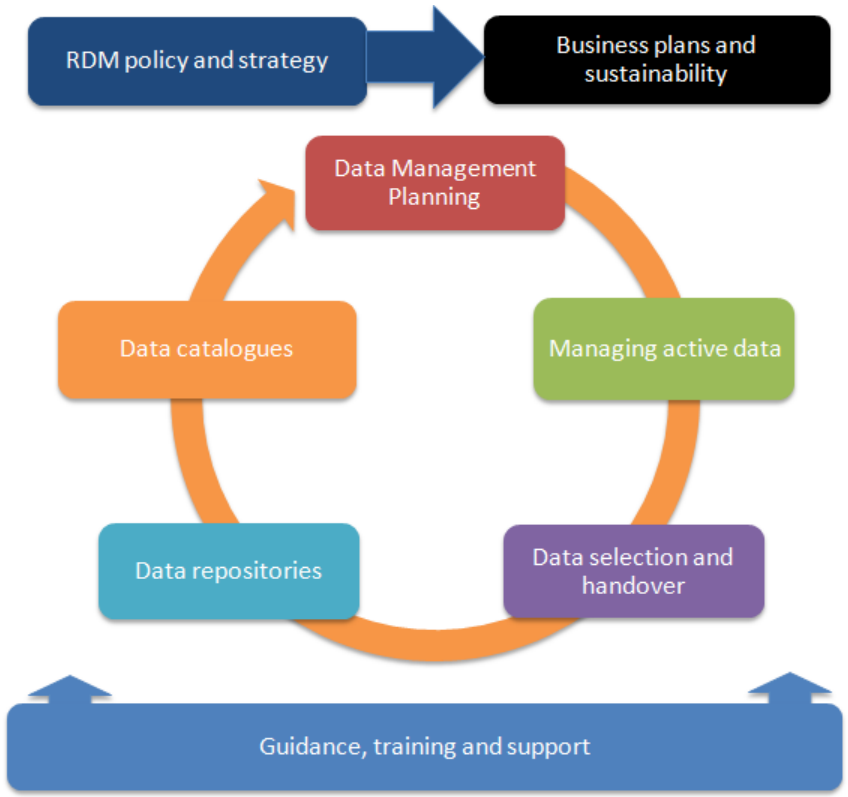
Systems list (overlaps)

Metadata capture system	Active research data management	Research Data Repository	Research Data Catalogue	Research Data Archive	Current Research Information System (CRIS)	OTHER:
MS excel	Ckan	Zenodo	Researchers' memories	Arkivum	Pure	Email
Electronic lab notebooks	Figshare	Figshare	MS excel	Sharepoint	Symplectic	Zendto (Southampton)
Paper records	Sharepoint	Hydra	Irs	Third party digital (e.G. Cloud)	"Homebrew" CRIS systems	Twitter
Digital cameras	Dropbox	Moodle	Criss	Physical / offsite archive	Converis	Publications
Brisskit	File systems	Pure	Datacite metadata store	Bespoke / in-house systems	MS excel	Altmetrics
Laboratory equipment	Analogue storage (e.G. Folders and archival boxes)	Converis	Figshare	Obsolete / non-networked storage media	MS access	DMP online
Figshare	Reference managers (zotero et al)	Symplectic	Gateway to research	Institutional network / filestore / tape backups	Worktribe	Datacite
SPSS / nudist (?)	Electronic lab notebooks	Dataverse	Library and archival management and discovery systems	Researcher's computer (at work or at home)	Tribal	Orcid
Nvivo	Criss	Eprints	Reuters data citation index	Arkivmatica	Info ed	Je-s
Matlab	Databases	Dspace	Reference managers (zotero et al)	Rosetta	Kuali coelus	Re3data
Reference managers (zotero etc)	Dropbox et al	Duracloud	Data.Gov.Uk	Preservica	Research master	Citation indexes
	USB sticks, external hds	Personal/work web pages	EUDAT services		Aggresso	The web
		Dryad	Disciplinary archives (UKDA, NERC etc)		All other institutional information systems (e.G. Hr, finance etc)	Google scholar
		Departmental servers			Eprints	Data mining tools
		Research gate			Dspace	All other institutional information systems (e.G. Hr, finance etc)
		Dropbox et al				Openaire
		Analogue storage (e.G. Folders and archival boxes)				Data visualisation tools
		Nesstar				Data wrangling tools
		Journal supplementary information				Collaboration tools (e.G. Google docs?)

Systems appearing in more than one category

All other institutional information systems (e.g. HR, finance, etc)	Electronic lab notebooks
Analogue storage (e.g. folders and archival boxes)	Eprints
Bespoke / in-house systems (inc. CRISs)	Figshare
3rd party CRISs	MS Excel
Converis	Pure
Databases	Reference managers (Zotero et al)
Dropbox et al	Sharepoint
Dspace	Symplectic

Are these of any special significance with regard to the components model?



Breakout groups

1. Which research data solutions do institutions or researchers pick & for what / where / when? (Suggested by Jonathan Tedds on Twitter)
2. Capturing requirements (inc. appraisal?), and matching these with solutions
3. ???

Other potentials

- How do we cope with crossing institutional boundaries to communicate with systems we can't control?
- How do we hide/tailor the experience to suit the user?
- What are the opportunities for pooling of effort and resources? What are the barriers to this?
- What are the necessary infrastructural modifications (training, funding etc)?

Okay, on to the presentations…

- Starting with Catherine Jones of STFC

Breakout groups: final

Topic	Chair	Rapporteur (& blogger)	Room
Use cases: Which research data solutions do institutions or researchers pick, and for what/where/when?	Jonathan Tedds	Laura Molloy	Main room (Chestnut)
Shared services: what are the opportunities for pooling of effort and resources? What are the barriers to this?	Helen McEvoy	Angus Whyte	Willow
Crossing boundaries: How do we cope with crossing institutional boundaries to communicate with systems we can't control?	Jez Cope	Martin Donnelly	Sycamore
Tracking uses: Tracking who uses/reuses data (in academia and industry), and what they use it for	Simon Kerridge	From group	Bar / coffee area

Wrapping up

- Next steps
  - Slides will be uploaded to the DCC website
  - Twitter archive – will Tweet link to this next week (#rdmf12 hashtag)
  - Blog posts, inc. breakout group reports
  - Others TBC?
- Feedback forms
- RDMF13
  - The next meeting will be a one-day event, to be held in central London in the Spring
  - Everyone is invited to suggest topics via the RDMF social space. Topics suggested at and after #rdmf11 include…

Future topics?

- **From last event's feedback**
  - Metadata / capturing metadata III
  - Licensing and copyright II
  - Costs I
  - Advocacy / dealing with culture change I
  - Data publication. I
  - Institutional Data Catalogues - examples, advice, tips, problems. I
  - Revisiting the funders I
  - Humanities data I
  - Crossing divisional boundaries within institutions (e.g. IT, library, etc) I
  - Data quality / suitability for long-term preservation I
  - Improving DMPs I
  - Promoting DOIs I

We also received a few pleas for institutional case studies/approaches *as an RDMF topic*. We've attempted to build these into the programme as a standing feature; let us know if you think this has been successful or if we need a different approach.

Previous RDMF events

#	Date	Location	Topic
*	19 November 2007	London	Planning meeting
1	19-20 March 2008	Manchester	Inaugural workshop
2	26-27 November 2008	Manchester	Roles and Responsibilities for Effective Data Management
3	30 April-1 May 2009	Manchester	Value and Benefits of Data Sharing and Management
4	10-11 March 2010	Manchester	Dealing with Sensitive Data
5	27-28 October 2010	Manchester	Economics of Applying and Sustaining Digital Curation
*	13 January 2011	London	RDMF Review Meeting
6	5-6 May 2011	Leicester	Planning for Research Data Management
7	2-3 November 2011	Coventry	Incentivising Data Management and Sharing
8	29-30 March 2012	Southampton	Engaging with the Publishers
9	14-15 November 2012	Cambridge	Shaping the Infrastructure
*	25 April 2013	Birmingham	Special event: Funding Research Data Management
10	3-4 September 2013	Oxford	Research Data Management in the Arts and Humanities
11	20 June 2014	London	Workflows and Lifecycle Models for Data Management
12	17-18 November 2014	Leicester	Research Data and Repositories (and other systems)

Research Data Management Forum 12

Linking Data and Repositories (and other systems)

Thank you and safe home

See you at #rdmf13