Research Data Alliance Plenary 4: Long tail of research data IG

Session notes drafted by Laura Molloy: all corrections to laura.molloy AT glasgow.ac.uk

Group webpage: https://www.rd-alliance.org/group/long-tail-research-data-ig.html

- Co-chair Kathleen Shearer welcomed participants to the session and clarified that the difference between our group and the Libraries for Research Data IG is that this group focuses on not just libraries, but any organisation that's collecting long tail data and ensuring we're sharing our practices across our group of diverse institution types and domain areas. Kathleen is also co-chair of the LRD IG so she can provide cross pollination where relevant.

- Strong representation at this group from librarians but there are also IT and domain-specific researchers.

- interoperability is a priority when we are considering solutions.

- Agenda: 1. case studies: environments that encourage deposit and use of repositories by the research community; 2. business models: what are the propositions, services provided, and revenue models for managing data?; 3. what work do we want to pursue in the context of the IG?; 4. Any other business.

Case studies were scheduled as follows:

* Veerle Van den Eynden, 'Incentives for sharing research data'; Knowledge Exchange & UK Data Archive
* Stefan Kramer, American University, US
* Dimitris Koureas, Natural History Museum London, UK
* Amy L. Nurnberger, Columbia University, US
* Kerstin Lehnert, Integrated Earth Data Applications, US
* Jochen Schirrwagen, Bielefeld University, Germany

Our first presentation was from Veerle Van Den Eynden, UK Data Archive:

- Veerle described a recently-completed Knowledge Exchange study on incentives, on which she worked with Libby Bishop (also UKDA). It is due to be published online in the next couple of weeks.

- Study looks at research data in general, though, not just long tail.

- A lot is already known about barriers for sharing research data. this is about incentives.

- They found there is a wide variety of data-sharing policies for various countries in Europe.

- Most researchers can see the benefits for general science, but experience problems sharing their own data.

- Study produced 5 case studies, based on interviews with 22 researchers.

- 5 countries involved were FI, DK, GE, UK, NL. Some interviews were in English, some in their local language.

- Research was from varied disciplines. Denmark: LARM Audio REsearch Archive (http://larm.blogs.ku.dk/about-larm/); Germany DFG-funded Adaptomics - evolutionary plant solution to ecological challenges (http://www.ruhr-uni-bochum.de/dfg-spp1529/Seiten/index.html); Netherlands: bioinformatics; Finland: ethnography; UK: chemistry department at the University of Southampton.

- The study found variation in what researchers consider to be data sharing - for example this can vary from sharing with immediate research team by way of email or common file storage, to deposit in an open repository licensed for re-use.

- Some data sharing styles provide mutual benefits, but sometimes the sharing is more in the form of data 'donation', i.e. placing data in a community repository without knowing exactly what will be done with it or by whom (although this is clearly also an important thing to do and one that is encouraged by RDM advice and guidance in general and many RDA activities in particular).

- Study found that data sharing is often already considered part of scientific process, happens in collaborative research, peer exchange, and in the creation and deposit of supplementary data to publications.

- Some sharing takes place early in the research process (this is often raw data); some later (this data more likely to be processed).

- There are well-established data sharing practice in some disciplines e.g. crystallography, genetics, which contribute to the development of community / topical databases such as BrassiBase, the LARM archive. Some sharing also happens already via public repositories in chemistry, ethnography, biology.

- What motivates sharing? Study found researchers recognise there are direct benefits:

1. for research itself - including sharing methods, participating in collaborative analysis, to perform research depending on data mining; to produce supplementary materials for publications; when funder is specifically funding creation of a data resource;

2. for research career - for visibility, reciprocity, reassurance e.g. invited to share, felt proud to be asked to share; for career progression;

3. for good of the discipline and for better science.

- Community or discipline 'norms' are also incentives: sharing happens more when it is the default in the domain.

- The study found hierarchical sharing throughout research career, generally taking place with researchers of a given level sharing their data with researchers one level 'higher up', i.e. master's students with PhD students; PhD students with PIs.

- Study noted that some / many researchers specifically want to challenge conservative non-sharing culture, recognising that openness benefits research. The view in some disciplines is, 'My data are well annotated anyway, so it's not much extra effort to deposit.'

- External drivers: when funders directly fund data sharing projects; when journals expect supplementary data; when learned societies develop infrastructure and resources; when there are data support services; where there are publisher and funder policies and expectations.

- Where there are no requirements or mandates, as in some countries, study found that researchers were in favour and would like these to be implemented.

- Veerle reckons that funder data policies haven't achieved as much as they would like, but they are an important mechanism to slowly change general attitudes and norms.

- Future incentives for researchers:

- Researchers don't want to be outrun by competitors, so there needs to be an environment where they are reassured that data sharing won't give competitors an unfair advantage;

- Direct funding for RDM support;

- Student training in data sharing;

- Infrastructure and standards for data sharing;

- There is appetite for sharing failed experiments, coming from biochemistry and biology. Doing so helps avoid repetition of the same experiments;

- Need to be able to engage in micropublishing / micro-citation - need to attribute small chunks from a larger dataset;

- Broadening the disciplines where data sharing is the norm.

- Study contains recommendations for different stakeholder groups, with some of the key recommendations as follows:

- Discipline / domain norms are influential;

- Encouragement of direct benefits for sharing important;

- Leadership required - funders, learned societies, mixed economy of incentives phase in research data lifecycle and career stage.

- Also need to invest in rich data resources - data + context at European level.

- Recommendations for research institutions:

- data should be used to measure impact in phd career assessment, e.g. impact portfolio, data CV;

- Integrated RDM support Services ('one stop shop'-style offerings) should be developed;

- Data sharing training should be part of standard student research training.

- There are also recommendations for funders, learned societies, KE partners (inc Jisc, SURF).

------------

Our next presentation was from Stefan Kramer (skramer@american.edu), the research data librarian at American University, Washington DC. This is a medium-sized institution with c. 10K students, running a repository on the Dspace platform ('not a platform that makes data sharing very sexy'). Stefan's talk aimed to outline their efforts in incentivising research data sharing, starting with pertinent question, 'How many of you have never seen a RD lifecycle model?' (No-one put up their hand!) Stefan, in the awareness that we have all seen many such models, then showed his own, noting that the key difference with this one is that it explicitly acknowledges that archiving and publication could take place simultaneously or separately in any order.

- Dspace users need to know what they want to do with the data and have the software to do so; this is not ideal for all users.

- How to broaden impact of research? One potential target audience for this institution: DC policy makers, journalists, etc. However, these are not ideal users of Dspace.

- As a result, they have been trialling use of a data visualisation platform so that data can be interacted with by non-specialists. Is called Socrata open data portal - from a company primarily catering to government agencies making their data openly available, including at one point data.gov: www.socrata.com;

- Their instance is at opendata.american.edu. Faculty and graduate students can upload their data, and the product delivers hosting of the dataset plus visualisations.

- licensed until June 2015 as a test case to see if this incentivises users.

Q: Can datasets be found re. Google? That's what most users will employ. A: Our datasets hopefully indexed by Google. We're talking to Socrata re. providing DOIs for datasets. We want to provide archiving and visualisation from the same piece of software.

Q: Which file formats can be accepted? A: Data has to be in excel or .csv format, and there is no size restriction on individual files, but there is restriction on number of datasets. We can also ingest documents and image files as well.

Q; Did you try CKAN? A: We'd like to, but it would require development effort that we don't have sufficient staff to deliver.

-------------------

Dimitris Koureas, National History Museum, London

'Linking layers of biodiversity data: informatics challenges for long tail research'

- Biodiversity studies are very complex because this is a highly interdisciplinary domain;

- Challenge 1: find data and move it: these are heterogeneous, diverse. new data structures appearing all the time.

- Challenge 2: linking and aggregating data at different scales, e.g. community, national efforts, global;

- Challenge 3: synthesising data, e.g. modelling human pressures on biodiversity.

- We are dealing with very large data volume: around 17k new species are described every year.

- Using an informatician's view of (biodiversity?) data, we see that the majority of data are not available, e.g. investigator focused small data, locally generated visible data, incidental data: when we count these into overall mass of research data, 80% of data are more important mainly due to their volume (from Heidorn, PB in Library Trends 57, p280);

- Scholarly data publication is increasing exposure and citability of work. We don't just need to provide incentives - it is also crucial to lower barriers to participation in data sharing;

- Dimitris described a tool for the biodiversity domain, the Scratchpad VRE. An easy tool to collaborate, publish, expose, and leverage value of their data by linking it to large biodiversity resources.

- Dimitris added the very important point that some tools try to do lots of things in one discipline, and others focus on one activity across many disciplines, highlighting the usefulness of tools with clear focus.

- They are now hoping to get funding for linking of data services and communities for predictive modelling of the biosphere. This is their long term goal.

Q: Kenji T asked about the sustainability model. A: We want community input from users, not necessarily in money but in time, expertise.

Q: Kathleen S asked re. data integration. D says they put the effort on trying to ensure the data is well described, but full integration would be very difficult.

------------------

Amy L. Nurnberger, Columbia University, US: 'Encouraging deposit and use of repositories'

- Columbia University institutional repository is called Academic Commons, accepts publications, datasets, images. They focus on data that don't have disciplinary homes elsewhere.

- What can be done to incentive deposit (i.e. with current resources)?

- They found that offering free storage space, the volume of which has grown from 2gb > 5gb > 10gb, seems to encourage to a certain extent.

- They have also provided DMP templates for the institution.

- They also provide boiler plate about the repository (uses?)

- Activities include advertising that the data exist and are available for re-use, search engine optimisation, promotions of datasets tied to historical anniversaries, national events, etc.

- The repository service provides DOIs through easyID from CDL so that resources are citable;

- Slide: 'Developing a community around research data management' explains the development of a community interest group at university level. Coming out of this: stories of success about research data use and re-use.

- The team is developing guidelines rather than policies, and integrating RDM good practice within training and guidance for good research practice in general. This involves librarians, compliance office, and the running of RDM workshops for disciplines,

- The team is looking at DMPs to see which mention the repository specifically, and then connecting directly with those researchers;

- The team is having conversations with review boards about the balance of risks of sharing vs protection in human research (for example);

- Amy is clear of the value of librarians in the data management and sharing conversation: We are librarians; we're in a learning organisation. We can learn those skills. We have a lot of relevant skills already.

Q: I asked whether the 'good data sharing stories' will be made publicly available. A: Yes, they will be. On the website for the Centre for Digital Research and Scholarship.

http://cdrs.columbia.edu/cdrsmain/?q=index.php

Q: How do you measure success? A: Project hasn't been in place long enough to measure that effectively.

Q: How to entice those researchers who are resistant? A: We've tried to established relationships with the faculty. e.g. designated librarian. We're going with gathering the low-hanging fruit first.

Q: Sarah Jones asked: How do you follow up on DMPs? A: We have built a relationship with grants office, allows us to look at successful DMPs. They'll become part of public record anyway.

Comment: Even in 'really hard' sciences there are data with no obvious home (i.e. it's not just the case in the arts and humanities).

A: We encourage researchers to look at discipline-specific data repositories. We're doing mirror deposits in domain repositories and IRs.

Comment: It's a myth there is no interoperability between IRs and subject repositories. Just that there are many cases where researchers don't have a domain repository that they can use.

Kerstin Lehnert, Integrated Earth Data Applications, US NSF IEDA

- Kerstin discussed strategies for incentivising data sharing through provision of small amounts of money for data sharing activities.

- New initiative for developing data best practices for earth science publications. talking about consistent policies and procedures.

- Visualisation can be a powerful tool for incentivisation as it provides a new way of looking at datasets and understanding them which can be very attractive both to researchers and other users - we want to do what we can to encourage people saying 'yes, I want to submit my data to this for use in these syntheses.'

- IEDA data rescue mini awards provided, for example, USD5k to digitise and preserve a set of lunar glass samples. The researcher was going to retire, and these would have otherwise been destroyed.

- Main funding from NSF, with the goal to establish as necessary tool for researcher.

- DMPs don't work as they are not enforced. Nobody is checking whether the researchers did what they say they will do in the DMP and so they are not currently effective in improving actual practice.

Q: Small grants are so important. Sometimes the advocacy is totally successful but there just aren't the funds to do the work.

- They are currently evaluating systems including CKAN and Dspace.

Jochen Schirrwagen, Bielefeld university library, Openaire

- Jochen discussed how the library coordinated to set up RDM services in the university

- Bielefeld is medium sized, with 5 main strategic research areas: in humanities, human development, conflict and violence, plus sciences.

- The German Rectors' Conference made a Resolution on13 May 2014 to acknowledge availability, heterogeneity and growth of digital research data, its relevance for most disciplines, and how it is important to open up new research avenues using IT tools.

- The Bielefeld policy is at data.uni-bielefeld.de/resolution (2011), based at library

- 2013: built their own tool for DMPs; they have a data repository.

- Jochen is working in liaison with international initiatives including Iassist, DDI, Openaire and collaborative DFG-funded research centres.

- They are providing templates for DMPs including for H2020.

- Their repository provides each researcher with a profile page listing the Google scholar profile, Scopus author id, Orcid, Github ID, to connect these identities.

- They also provide cross links with other repositories, e.g. Web of Science, PubMed Central.

- They are going to apply for the Data Seal of Approval. They want to move towards reproducible research - empirical and experimental sciences data.

In questions, we established they built their own DMP tool on Drupal which was developed onsite.

--------

Kathleen Shearer, IG co-chair

Outcomes from the IG so far:

- Completed survey about a year ago in metadata discovery practice in long tail repositories. Asked re. what people are doing to improve discovery of their datasets. The results are on the IG webpage.

- What can we do as a group that would be good to help your work now?

- Suggested work areas: interoperability, discoverability, incentives for deposit and use, costs and funding models, evidence about the long tail. Others?

- Perhaps some of us could go back to our campuses and try to find out the kind of data holdings we have.

- To add to the list of outcomes: if there are tools that help long tail data, these should be identified, shared and reviewed.

- Johan Bielefeld: suggests getting those responsible for management practices and service implementation together (to compare practice?)

- Q re. interoperability of tech platforms and data sets. Wolframm: there are hundreds and maybe thousands of different repositories. How can you have a discovery layer or exchange between these repositories if you don't have any interoperability? How can you link data together that are in different repositories? We need to automate that.

- Dimitros: how do you mobilise the data that isn't in managed storage? The entry point is the most important. (However, this is not long tail specific.)

- James Wilson: a useful outcome would be a list of various tools and APIs that are available already. Then we could do a gap analysis of what's lacking in terms of the tools researchers already use.

- It would be useful to cross check with data publishing workflows, and identify 20-25 examples of how data gets into the workflow.

- Robin Rice: Edinburgh data repository seen as successful. percentage of university researchers? less than 1%. does that mean a failure? this group could be useful to set that context.

- Anna Clements: Incentivisation is important. Where there are league tables re. who is performing well, these tend to incentivise university departments. RDM is not serious yet but open access is, so perhaps we need to be a bit further down the path before we see a substantial improvement in practice.

- The EC open data pilot has the potential to improve awareness and practice. We could prepare a briefing paper of incentives to help stop people backing out early from EC open data pilot.

- We should make a distinction between incentives and workflows. It's not just about depositing. For our collection, incentives to reuse and for researchers of other disciplines to find and re-use the data is important. It's not all about getting stuff into the repository.

- Susan Reilly: Language of empowerment is sometimes more useful than mandate, when motivating researchers.

- It would be useful to have examples of people looking for data and not being able to find it. Kerstin mentioned the Earthcube (http://earthcube.org) initiative. They carried out a survey to establish whether researchers want to have access to other people's data. There is a publication coming out on this survey from Illinois (Urbana-Champaign). (In the meantime, see slide 16 onwards at http://www.slideshare.net/EarthCube/ec-stakeholder-alignment-it-software-assembly-boulder-mar-2014).

- James: people know their collections are not being used.

- Robin: not every researcher has a project finished and ready to go. Sometimes they're looking for data. Librarians can be the bridge for these datasets that are wonderful but nobody knows about.

Wolfram:

- We should decide our communication strategies, do some desk research, produce a document.

Andrew Treloar:

- big data gets plenty of airtime. Plenty of significant stuff can be described as long tail data. Would it be worth coming up with a set of recommendations: ten simple things you or your institution could do? It won't solve the problem overnight, but it'll improve things. You'd be better off for doing them. So institutions can say, are we doing this? And if not why not? It would be a very useful first step. And you don't have to become a WG to do this - interest groups can have outputs.