Unlocking Research
University of Cambridge Office of Scholarly Communication

Forget compliance. Consider the bigger RDM picture

March 8, 2016 Uncategorized conference, IDCC, open source, repository, research, research data management, sensitive data, text and data mining

The Office of Scholarly Communication sent Dr Marta Teperek, our Research Data Facility Manager to the International Digital Curation Conference *held in in Amsterdam on 22-25 February 2016. This is her report from the event.*

Fantastic! This was my first IDCC meeting and already I can't wait for next year. There was not only amazing content in high quality workshops and conference papers, but also a great opportunity to network with data professionals from across the globe. And it was so refreshing to set aside our UK problem of compliance with data sharing policies, to instead really focus on the bigger picture: **why it is so important to manage and share research data and how to do it best**.

Three useful workshops

The first day started really intensely – the plan was for one full day or two half-day workshops, but I managed to squeeze in three workshops in one day.

Context is key when it comes to data sharing

The morning workshop was entitled "A Context-driven Approach to Data Curation for Reuse" by Ixchel Faniel (OCLC), Elizabeth Yakel (University of Michigan), Kathleen Fear (University of Rochester) and Eric Kansa (Open Context). We were split into small groups and asked to decide **what was the most important information about datasets from the re-user's point of view**. Would the re-user care about the objects themselves? Would s/he want to get hints about how to use the data?

We all had difficulties in arranging the necessary information in order of usefulness. Subsequently, we were asked to re-order the information according to the importance from the point of view of repository managers. And the take-home message was that for all of the groups the **information about datasets required by the re-user was the not same as that required from the repository**.

In addition, the presenters provided discipline-specific context based on interviews with researchers – depending on the research discipline, different information about datasets was considered the most important. For example, for zoologists, the information about specimen was very important, but it was of negligible importance to social scientists. So **context is crucial for the collection of appropriate metadata information. Insufficient contextual information makes data not useful**.

So what can institutional repositories do to address these issues? If research carried out within a given institution only covers certain disciplines, then institutional repositories could relatively easily contextualise metadata information being collected and presented for discovery. However, repositories hosting research from many different disciplines will find this much more difficult to address. For example, Cambridge repository has to host research spanning across particle physics, engineering, economics, archaeology, zoology, clinical medicine and many, many others. This makes it much more difficult (if not impossible) to contextualise the metadata.

It is not surprising that information most important from the repository's point of view is different that the most important information required by the data re-users. In order to ensure that research data can be effectively shared and preserved in long term, repositories need to collect certain amount of administrative metadata: who deposited the data, what are the file formats, what are the data access conditions etc. However, **repositories should collect as much administrative metadata as possible in an automated way**. For example, if the user logs in to deposit data, all the relevant information about the user should be automatically harvested by feeds from human resources systems.

EUDAT – Pan-European infrastructure for research data

The next workshop was about EUDAT – the collaborative Pan-European infrastructure providing research data services, training and consultancy for researchers. EUDAT is an impressive project funded by Horizon2020 grant and it offers five different types of services to researchers:

- B2DROP – a secure and trusted data exchange service to keep research data synchronized, up-to-date and easy to exchange with other researchers;

- B2SHARE – service for storing and sharing small-scale research data from diverse contexts;
- B2SAFE – service to safely store research data by replicating it and depositing at multiple trusted repositories (additional data backups);
- B2STAGE – service to transfer datasets between EUDAT storage resources and high-performance computing (HPC) workspaces;
- B2FIND – discovery service harvesting metadata from research data collections from EUDAT data centres and other repositories.

The project has a wide range of services on offer and is currently looking for institutions to pilot these services with. I personally think these are services which (if successfully implemented) would be of a great value to Pan-European research community.

However, I have two reservations about the project:

- Researchers are being encouraged to use EUDAT's platforms to collaborate on their research projects and to share their research data. However, **the funding for the project runs out in 2018**. EUDAT team is now investigating options to ensure the sustainability and future funding for the project, but what will happen to researchers' data if the funding is not secured?
- Perhaps if the funding is limited it would be more useful to **focus the offering on the most useful services**, which are not provided elsewhere. For example, another EC-funded project, Zenodo, already offers a user-friendly repository for research data; Open Science Framework offers a platform for collaboration and easy exchange of research data. Perhaps EUDAT could initially focus on developing services which are not provided elsewhere. For example, having a Pan-Europe service harvesting metadata from various data repositories and enabling data discovery is clearly much needed and would be extremely useful to have.

Jisc Shared RDM Services for UK institutions

I then attended the second half of Jisc workshop on shared Research Data Management services for UK institutions. The University of York and the University of Cambridge are two of 13 pilot institutions participating in the pilot. Jenny Mitcham from York and I gave presentations on our institutional perspectives on the pilot project: where we are at the moment and what are our key expectations from the pilot. Jenny gave an overview of an impressive work by her and her colleagues on addressing data preservation gaps at the University of York. Data preservation was one of the areas in which Cambridge hopes to get help from the Jisc RDM shared services project. Additionally, as we described before, Cambridge would greatly benefit from solutions for big data and for personal/sensitive data. My presentation from the session is available here.

Presentations were followed by breakout group discussions. Participants were asked to identify the areas of priorities for the Jisc RDM pilot. **The top priority identified by all the groups seemed to be solutions for personal/sensitive data and for effective data access management**. This was very interesting to me as at similar workshops held by Jisc in the UK, breakout groups prioritised interoperability with their existing institutional systems and cost-effectiveness. This could be one of the unforeseen effects of strict funders' research data policies in the UK, which required institutions to provide local repositories to share research data.

As a result of these policies, many institutions were tasked with creating institutional data repositories from scratch in a very short time. Most of the UK universities now have institutional repositories which allow research data to be uploaded and shared. However, very few universities have their repositories well integrated with other institutional systems. **Not having the policy pressure in non-UK countries perhaps allowed institutions to think more strategically about developing their RDM service provisions** and ensure that developed services are well embedded within the existing institutional infrastructure.

Conference papers and posters

The two following days were full of excellent talks. **My main problem was which sessions to attend**: talking with other attendees I am aware that the papers presented at parallel sessions were also extremely useful. If the budget allows, I certainly think that **it would be useful for more participants from each institution to attend the meeting to cover more parallel sessions**.

Below are my main reflections from keynote talks.

Barend Mons – Open Science as a Social Machine

This was a truly inspirational talk, raising a lot of thought-provoking discussions. Barend started from a reflection that more and more brilliant brains, with more and more powerful computers and with billions of smartphones, created a single, interconnected social super-machine. This machine generates data – vast amount of data – which is difficult to comprehend and work with, unless proper tools are used.

Barend mentioned that with the current speed of new knowledge being generated and papers being published, it is simply impossible for human brains to assimilate the constantly expanding amount of new knowledge. Brilliant brains need powerful computers to process the growing amount of information. But in order for science to be accessible to computers, we need to move away from pdfs. **Our research needs to be machine-readable**. And perhaps if publishers do not want to support machine-readability, we need to move away from the current publishing model.

Barend also stressed that if data is to be useful and correctly interpretable, it needs to be accessible not only to machines, but also to humans, and that effort is needed to make data well described. Barend said that research data without proper metadata description is useless (if not harmful). And how to make research data meaningful? Barend proposed a very compelling solution: **no more research grants should be awarded without 5% of money dedicated for data stewardship**.

I could not agree more with everything that Barend said. I hope that research funders will also support Barend's statement.

Andrew Sallans – nudging people to improve their RDM practice

Andrew started his talk from a reflection that in order to improve our researchers' RDM practice we need to do better than talking about compliance and about making data open. **How a researcher is supposed to make data accessible, if the data was not properly managed in the first place?** The Open Science Framework has been created with three mission statements:

- Technology to enable change;
- Training to enact change;
- Incentives to embrace change.

So what is the Open Science Framework (OSF)? It is an **open source platform to support researchers during the entire research lifecycle**: from the start of the project, through data creation, editing and sharing with collaborators and concluding with data publication. What I find the most compelling about the OSF is that is allows one to easily connect various storage platforms and places where researchers collaborate on their data in one place: researchers can easily plug their resources stored on Dropbox, Googledrive, GitHub and many others.

To incentivise behavioural change among researchers, the OSF team came up with two other initiatives:

- Promoting transparency and openness among publishers;
- Cash rewards for researchers who want to pre-register their research data with OSF: there are one thousand $1000 cash rewards.

Personally, I couldn't agree more with Andrew that **enabling good data management practice should be the starting point**. We can't expect researchers to share their research data if we have not helped them with providing tools and support for good data management. However, I am not so sure about the idea of cash rewards.

In the end **researchers become researchers because they want to share the outcomes of their research with the community**. This is the principle behind academic research – the only way of moving ideas forward is to exchange findings with colleagues. Do researchers need to be paid extra to do the right thing? I personally do not think so and I believe that whoever decides to pursue an academic career is prepared to share. And it is our task to make data management and sharing as easy as possible, and the use of OSF will certainly be of a great aid for the community.

Susan Halford – the challenge of big data and social research

The last keynote was from Susan Halford. Susan's talk was again very inspirational and thought-provoking. She talked about the growing excitement around big data and how trendy it has become; almost being perceived as a solution to every problem. However, Susan also pointed out the problems with big data. **Simply increasing the computational power and not fully comprehending the questions and the methodology used can lead to serious misinterpretations of results.** Susan concluded that when doing big data research one has to be extremely careful about choosing proper methodology for data analysis, reflecting on both the type of data being collected, as well as (inter)disciplinary norms.

Again – I could not agree more. Asking the right question and choosing the right methodology are key to make the right conclusions. But are these problems new to big data research? I personally think that we are all quite familiar with these challenges. **Questions about the right experimental design and the right methodology have been known to humankind since scientific method is used**.

Researchers always needed to design studies carefully before commencing to do the experiments: what will be the methodology, what are the necessary controls, what should be the sample size, what needs to happen for the study to be conclusive? To me this is not a problem of big data, to me this is a problem that needs to be addressed by every researcher from the very start of the project, regardless of the amount of data the project generates or analyses.

Birds of a Feather discussions

I had not experienced Birds of a Feather Discussions (BoF) before at a conference and I am absolutely amazed by the idea. Before the conference started the attendees were invited to propose ideas for discussions keeping in mind that BoF sessions might have the following scope:

- Bringing together a niche community of interest;
- Exploring an idea for a project, a standard, a piece of software, a book, an event or anything similar.

I proposed a session about sharing of personal/sensitive data. Luckily, the topic was selected for a discussion and I co-chaired the discussion together with Fiona Nielsen from Repositive. We both thought that the discussion was great and our blog post from the session is available here.

And again, I was very sorry to be the only attendee from Cambridge at the conference. There were four parallel discussions and since I was chairing one of them, I was unable to take part in the others. I would have liked to be able to participate in discussions on ‘Data visualisation’ and ‘Metadata Schemas’ as well.

Workshops: Appraisal, Quality Assurance and Risk Assessment

The last day was again devoted to workshops. I attended an excellent workshop from the Pericles project on the appraisal, quality assurance and risk assessment in research data management. The project was about how an institutional repository should conduct data audits when accepting data deposits and also how to measure the risks of datasets becoming obsolete.

These are extremely difficult questions and due to their complexity, very difficult to address. Still, the project leaders realised the importance of addressing them systematically and ideally in an (semi)automated way by using specialised software to help repository managers making the right preservation decisions.

In a way I felt sorry for the presenters – their project progress and ambitions were so high that probably none of us, attendees, were able to critically contribute to the project – **we were all deeply impressed by the high level of questions asked**, but our own experience with data preservation and policy automation was nowhere at the level demonstrated by the workshop leaders.

My take home message from the workshop is that **proper audit of ingested data is of crucial importance**. Even if there is no automation of risk assessment possible, repository managers should at least collect information about files being deposited to be able to assess the likelihood of their obsolescence in the future. Or at least to be able to identify key file formats/software types as selected preservation targets to ensure that the key datasets do not become obsolete. For me the workshop was a real highlight of the conference.

Networking and the positive energy

Lots of useful workshops, plenty of thought-provoking talks. But for me one of the most important parts of the conference was meeting with great colleagues and having fascinating discussions about data management practices. I never thought I could spend an evening (night?) with people who would be willing to talk about research data without the slightest sights of boredom. And the most joyful and refreshing part of the conference was that due to the fact we were from across the globe, our discussions diverted away from the compliance aspect of data policies. **Free from policy, we were able to address issues of how to best support research data management**: how to best help researchers, what are our priority needs, what data managers should do first with our limited resources.

I am looking forward to catching up next year with all the colleagues I have met in Amsterdam and to see what progress we will have all made with our projects and what should be our collective next moves.

Summarising, I came back with lots of new ideas and full of energy and good attitude – ready to advocate for the bigger picture and the greater good. I came back exhausted, but I cannot imagine spending four days any more productively and fruitfully than at IDCC.

Thanks so much to the organisers and to all the participants!

Published 8 March 2016
Written by Dr Marta Teperek