8th International Data Curation Conference - Amsterdam 14-17 January 2013
Extensive conference visit report by Kasper Abcouwer, Driek Heesakkers, Henriette Reerink & Mariëtte van Selm (University of Amsterdam, Library).
Please note that it was edited and translated by Driek, all mistakes are mine!

The International Data Curation Conference is an annual conference organised by the Digital Curation Centre (UK). This year's edition took place in Amsterdam, giving several colleagues from the University of Amsterdam a chance to attend. With around 275 participants and speakers from Great Britain, US, Canada, Australia, as well as Germany, Holland and other European countries, it was a varied and international affair.

Programme (with links to presentations): http://www.dcc.ac.uk/events/idcc13/programme
Keynote videos: http://www.dcc.ac.uk/events/idcc13/video-gallery
Social media summary: http://eventifier.co/event/idcc13/
Searchable archive of #idcc13 tweets: http://t.co/enEPBgU8

Overall impression
Being able to support researchers in handling research data is rapidly becoming very important for university libraries. Researchers and students are working intensely with data, and do not make much use of the physical library. If we, as libraries, do not focus our attention on the scientists, we will become obsolete in a few years. This was the urgent message that was repeated time and time again, from different points of view, but always with a sense of now or never.

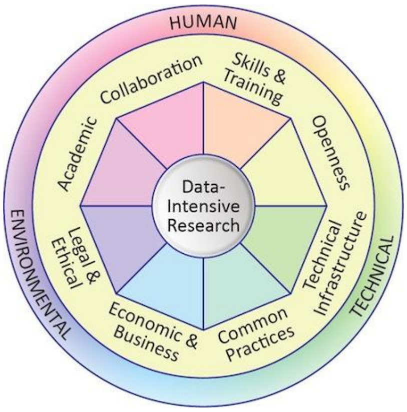
Key points to take home:

- The UK clearly is taking the lead. The government's open data policy and the Freedom of Information Act have a big influence.
- The UK has seven major research councils that finance research projects (as opposed to two in the Netherlands). After a few years where each experimented with different requirements on data management, these efforts are now coordinated, and are to be fully harmonized in 2015: "All publications should include a statement on how data (but also analogue parts of the research such as samples) can be obtained".
- Of these seven, the EPHRC (Engineering Higher Research Council) most radically places the responsibility for managing data produced in EPHRC-funded projects with the institutions and not with the research groups. This is a big driver for the rapid introduction of Research Data Management policies and services in the UK. Institutions with significant EPHRC funding were among the first. This is a major lesson for Dutch and European research funding agencies!
- Institutions with a well-run RDM programme have in common that the policy is clearly endorsed from above.
- Apart from the UK, Monash is the leading example. Their strategy starts from a very clear statement of intention: https://confluence-vre.its.monash.edu.au/display/rdmstrategy/Research+Data+Management+Strategy+and+Strategic+Plan+2012-2015 - It's worth remembering that this did not happen overnight, but is the result of a long campaign, led by the head librarian who persistently put data management on the agenda.

- Monash has great facilities and support for RDM, both in the form of the MeRC as well as in support through the regular library and ICT staff. This works because on all levels there are plenty of meetings to bridge the gap between institutions.
- A notable point of concern: a decade ago, innovations in our field were mostly developed by institutions. It seems that the capacity to do this has dramatically dropped as in-house ICT knowledge has been cut. Innovations now come from commercial companies. For example, the incubator digital-science.com (backed by MacMillan and Nature) has an impressive portfolio: figshare, LabGuru (inventory system for lab supplies, which incorporates regulations), 1degreeBio (online ratings for lab), Symplectic elements (altmetrics). Each of these companies were started by frustrated researchers who saw an opportunity, first tried within their institution but when they couldn't get it off the ground there, went independent. There is a lesson in here…

Workshops 14 January 2013

Community capability model framework (CCMF) for data-intensive research
Microsoft Research Connections and UKOLN (funded by JISC and the University of Bath) are working in partnership to develop the CCMF model.
The model aims to provide a framework useful for researchers, funders and institutions (ao. libraries) in modelling a range of disciplinary and community behaviours with respect to the adoption, usage, development and exploitation of cyber-infrastructure for data-intensive research.
The model is described in a white paper: http://communitymodel.sharepoint.com/Documents/CCMDIRWhitePaper-24042012.pdf).
Eight capability factors are described by means of relevant questions and are visualised in the diagram below.
UKOLN and Microsoft are working on an online generic checklist tool based on this diagram that represents the data intensive research life cycle. By clicking on the capability factors, one arrives at a series of fundamental questions that are related to the topic, to check whether all requirements have been taken into account. The tool (CCMF) is threefold as it can be used from the researcher's, institution's as well as the funder's perspective. The CCMF was applied and reviewed during the workshop by the attendants. Its opportunities were explored 1. in relation to the EU Horizon 2010 Programme from a funders perspective by the Head of Sector Research Data Infrastructures EU and 2. from a domain practitioner's perspective (Geosciences), with a focus on the cross-domain interoperability of the data.



Europe and the Research Data Alliance
http://rd-alliance.org/

De Research Data Alliance is een Australisch-Europees-Amerikaans initiatief, bedoeld om het delen en uitwisselen van onderzoeksdata te stimuleren en faciliteren. Beoogde werkwijze is die van *working groups*, die in 12 tot 18 maanden concrete resultaten presenteren. In drie presentaties werd de noodzaak van de RDA onderschreven door Carlos Morais Pires (Scientific Officer bij de Europese Commissie), Peter Wittenberg (Max Planck Instituut/EUDAT) en Simon Lambert (Science and Technology Facilities Council/ODE). Vervolgens beoordeelden de deelnemers aan de workshop in groepen twee voorstellen (*case statements*) van kandidaat working groups: *Legal Interoperability* en *Repository Audit and Certification*. Criteria waren Focus, Impact & Engagement, Timeframe en Scope/Fit.

Designing Data Management Training Resources
Na een korte introductie door Simon Hodson (Programme Manager van het Managing Research Data Programme van JISC) deelden achtereenvolgens Hannah Lloyd-Jones (Open Exeter) en Cathy Pink en Jez Cope (Research360) hun ervaringen met het opzetten van data management training voor onderzoekers.
OpenExeter signaleert zeven stadia in de ontwikkeling van training: *survey, requirements analysis, develop, test, feedback & evaluation, embed, maintain*. Tips: stel duidelijke doelen, ken je publiek, gebruik onderzoekers bij het maken van trainingsmateriaal, wees beknopt, zorg dat materiaal makkelijk te lezen is, vermijd jargon, bied woordenlijsten (glossaries) aan, zorg voor *take-away materials* (iets dat onderzoekers meenemen en eenvoudig kunnen ophangen), en vul de *face-to-face* training aan met *online guidance*.
Research360 (slides): twee uur is onvoldoende voor training in een groot en complex onderwerp als data management. Training is een startpunt en heeft twee doelen: *raise researcher's awareness of his/her responsibilities* en *try to prevent worst data management mistakes*. Onderzoekers hoeven niets extra's te doen, het gaat erom dat zij dat wat ze nu al doen *goed* doen. Aandacht voor 'why' (*demonstrate relevance*) én 'how' (ondersteuning van onderzoekers met praktische adviezen), 'why' zonder 'how' is zinloos.
Voor *data management planning* zijn diverse vragenlijsten en tools in omloop; die kunnen worden gebruikt om na te gaan of er zaken in de training ontbreken. Workshopdeelnemers evalueerden http://bit.ly/idcc13-dcc (DCC), http://bit.ly/idcc13-dmppgr (Jez Cope) en http://bit.ly/idcc13-20q (David Shotton's Twenty Questions).

15 January 2013, conference day 1

The conference opened with two keynotes by researchers in data-intensive fields.

First **Ewan Birney**, associate doctor of European Bioinformatics Institute (EBI), "say we're a sort of CERN for molecular biology". This field has a tradition of sharing: "a gift of the previous generation of scientists foresight of putting all in one database". However, as DNA sequencing costs implode, data volumes explode. Effective infrastructure is a must: "effective infrastructures raise the bar for scientists and make new research possible - like electricity, they'll only notice it when it goes wrong".
Birney advocates a pragmatic approach for international collaboration.
Both centralized and heavily distributed approaches have shown severe disadvantages, on both scaling and experience. The only way forward is a robust network with a strong hub. In this field the EBI is that central hub, with national and domain-specific hubs connecting to it.
youtube

Second came **Hans Pfeiffenberger** on stewardship of data in marine sciences. Pfeiffenberger, head of IT infrastructure at Alfred Wegener Institute for Polar and Marine Research, started by stating that science has always been based on data, at least since the renaissance. The astrologer Tycho Brahes published data tables on the movement of the moon. Kepler formulated his laws based on these tables. Newton's laws then explained those of Kepler. Data should not be separated from publications: publications "are the best 'metadata' of the data we can have".

He cited two examples of large data gathering projects to learn from, one that works, and one that hasn't. With Argo, Marine science has a large scale data collecting project, larger than CERN. 3500 floating buoys worldwide gather ocean data (temperature, current and other variables) between sea level and 2000m deep. This data is complemented by ship expeditions, which is where it gets messy as there are different standards between countries and ships. But overall they manage.

The Polar year 2007-2008 is an example of what happens when data management is not thought through for a large scale project. The goal was to take 'a snapshot of the poles' and 63k scientists participated, and 1 billion euro of research was funded. The data is supposed to be used for decades to come, and preservation is intended. Full access to IPY data is however still years off, and it only comes trickling through, publication by publication, and fragmented.

youtube

Anthony Beitz, Monash e-research Center

Presentation; youtube

Monash University is often described as an institution with a successful research data management strategy. Anthony Beitz, head of the Monash e-Research Center gave a good overview that drives the point home.

The secret? "At Monash senior management 'get it'." The institution aims to have a national leadership role in RDM, and published this in a statement of intent as part of the 'Monash Futures' program. The institution aspires to be a leader in research, and RDM is seen as a key aspect for the quality of the research: research data that is better managed, discoverable and available for reuse will improve research impact and outcomes, reduce legal risk, attract the best researchers and attract additional research income.

Of the many other interesting points, I want to highlight one: institutional data collection is good for many things (showcasing, peer review) but does not enable reuse. For that, researchers tend to look at community resources. Researchers and institutions have fairly divergent needs, which cannot be ignored. This does not mean the institution has to fulfil all aspects of the RDM landscape, on the contrary. Beitz stresses that it's important to carefully choose where to put your energy.

Three areas stand out. For supporting researchers MeRC works closely with library and IT services (collaboration is fostered on all levels in small groups). Then there is the data storage service, free for all researchers, currently at 4PB (and growing), which though intended for work-in-progress data, is quite secure with 4 copies of each file spread across 2 data centres. Finally there is a specific tool for large data sets (TB's), MyTardis, which started out as domain-specific for raw bio-science data, and is now deployed for many data-intensive disciplines.

Monash RDM strategy document: https://confluence-vre.its.monash.edu.au/display/rdmstrategy/Research+Data+Management+Strategy+and+Strategic+Plan+2012-2015
UKDCC case study on Monash: http://www.dcc.ac.uk/news/rdm-monash-university

Adam Farquhar, British Library
Presentation; youtube
"We treat digital content a lot like print content." Digitale content is een fractie van de totale content die de British Library in huis heeft, o.a. tot stand gekomen door digitaliseringsprojecten (handschriften, kranten, Google Books), maar er kan vele malen meer mee dan met print content. Er gaapt een kloof tussen wat curatoren/bibliothecarissen kunnen en de ondersteuning die zij aan onderzoekers willen geven: *building capacity* is hard nodig. Transitie van individuele items naar collecties: "Reading individual works is as irrelevant as describing the architecture of a building from a single brick, or the layout of a city from a single church" (Franco Moretti, Stanford).

Paul Miller
Presentation; youtube

Patricia Cruse, University of California Curation Center
Presentation; youtube
The University of California Curation Center (UC3), which is a division of the California Digital Library, has developed five services: DMPTool (creating and sharing data management plans); EZID (persistent identifiers); Merritt (data repository); WAS (webarchiving); DataUp (describing, managing and sharing data).

Kaitlin Thaney, Digital Science
Presentation, youtube
Kaitlin Thaney talks fast to cover all services offered by Digital Science. Digital Science, a division of Macmillan Publishers (publisher of Nature), has eight products at the moment. The best known one is undoubtedly figshare. Figshare enables the publishing of datasets, presentations, images, as well as complete articles. Figshare ensures linking to a DOI (digital object identifier), which makes it easy for others to refer to your publication.
Figshare is free and therefore the business model is looked upon with some suspicion. On the other hand, Macmillan as a large publisher is present in the background, and the growing amount of collaborations inspire confidence. One example of such a collaboration is the use of figshare as the research data repository for the seven PLoS journals.
The four more discipline specific products (Labguru, SureChem, BioRAFT en 1DEGREEBIO) are more or less 'laboratory' bound. 1DEGREEBIO (unbiased reviews on Life Science products and service providers) is used by the Radboud University Nijmegen and Utrecht University in the Netherlands, and is an 'open' alternative for the paid services offered by ACS.
Symplectic is a CRIS (Current Research Information System) and scientific output manager. It helps managers and scientists to gain insight into scientific output in the broadest sense. Plus, it also looks great (slick). If we were policy makers at faculty level, we would adopt and use Symplectic as a CRIS .
AltMetric is a social media software product, which can help you gain insight into the (social media) impact of an article. ReadCube has the slogan: "your research simplified". And this is

true. Managing your PDFs in an easy way, annotating and referring to them, and searching full text in your library are functions that are within the bounds of the possibilities. Your pdf library can also serve as a basis on which new articles are recommended. And last but not least, the support from Digital Science is enthusiastic and excellent.

18 January 2013 conference day 2

Herbert van de Sompel
Presentation; youtube
The presentation was meant as a wakeup call, and discussed the changes in scholarly communication and views on web infrastructure over the last 15 years, and the resulting consequences for us now.
Formerly, scholarly communication consisted of fixed articles, now we deal with all different sorts of dynamic research results (articles, video, datasets, pdf etc.), components which change over time (different versions) and changed software/infrastructure with which they were generated. This has consequences for the reproducibility of research.
OAI-ORE (resource maps) describes what (versions, formats, id's, enrichments etc) belongs together. It is used by Europeana. We're on the right track. Where OAI-ORE is about grouping assets, Memento is about versioning assets. Using Memento, one can access different versions through both the original uri and date. Vd Sompel illustrates this by checking the bibliography (url's) of an article he published in 2004; do the mentioned references still exist on the web? And if so, are the versions still the same? The check illustrates that some still do exist and some don't. He warns archives and repositories to not forget to submit http uri's for web archiving. Too often this does not happen. Use repository software with solid versioning mechanisms. Archive linked context at the date of publication. Archive at the moment of use, and proactively expose versions to web crawlers.
Unfortunately there was no time left to discuss ResourceSync, a project about resource synchronization, which is all about uri's again. OAI-ORE, Memento and ResourceSync illustrate the potential of influencing the web infrastructure for scholarly communication.
Vd Sompel's conclusion is based on the evolution of 15 years' worth of journals and PDF archives, to a network of interconnected web assets and actors:

- there is potential to influence existing web structures to tackle the problems of scholarly communication such as compounding and versions,
- don't think of interoperability in terms of between systems (repositories, etc.) but of interoperability with the web infrastructure,
- build on the infrastructure the entire world already depends upon, it is not going to go away.

Vd Sompel has come to trust the internet infrastructure to a certain point.

Parallel sessions

Session 1a - four institutional approaches to RDM

Summary: four institutions share their findings on getting a RDM policy in place. All are from the UK, where research data management is high on the agenda due to serious pressure from government and funding agencies for open data. However, it is still a good roundup on what works well and what to avoid for those outside Great Britain. And all have one thing in common: early and intense collaboration between all involved departments of the institution is key.

University of Bath, Cathy Pink (who also works for UKOLN).

As an engineering school, Bath's largest funder is the EPSRC. That makes their task relatively straightforward: implement the rules as set out by this agency.
Three important lessons learned:
- it's important to define what falls under policy, and what does not. Contract research for instance generates data for third parties and therefore does not fall under institutional digital assets.
- involve your institution early and often: management, legal team, research office.
- when dealing with funder-driven policy, get someone from the agency on your project board.

RDM work at Edinburgh, Robin Rice.

EPSRC strikes again! Their requirement to put responsibility for data sharing and preservation with the research organization was very influential for Edinburgh (perhaps even the tipping point).

RDM at Edinburgh is a major IS-investment-led program. One of the goals is a common resilient storage infrastructure and file store. Storage needs defined at 0,5tb/user for 2012-14. This leads to a roadmap (University of Edinburgh RDM roadmap), which is a living document, published on site and changed when needed.

Like in Bath, the project requires broad interdepartmental collaboration, and early and intensive involvement of all parties is important. Implementation committee: library, data library, IT infra, user services, DCC (!). This committee reports to the academic steering group (high-profile scientists), who do in fact have real influence and steered the direction of the project several times.
Difference between researchers and the institution: researchers want a dark archive, and the institution wants more openness. Not yet decided. Also undecided is long term sustainability. For now, storage is offered for free but researchers are encouraged to add budget for storage when applying for grants.
In 2002, a fire destroyed a lot of data in the School for Informatics. This made them very motivated. An unfortunate event, but is shows that raising awareness on the subject is much needed.

Creating an RDM service at Nottingham, Tom Parsons.

Lessons learned:
- have the highest management involved from the start. They weren't and this led to a delay of more than a year.
- numerous other services exist within an institution that reference data - research ethics guidelines especially - make sure you know them.
- We expected data to be more digital than what we found (lots of analog lab books etc.). Made decision to concentrate on digital only. Estimate: average 56GB/user - 10% of

Edinburgh! Conservative estimate, and without extrapolation (large variations, some top 50TB). Typically, researchers store data in at least five places (i.e. on as many USB sticks...). Only 24% create any metadata.

Oxford, DaMaRo (Data Management Rollout), James Wilson

DaMaRo is both a JISC project (now finished) as well as long term internal project.
Strong even collaboration between various divisions (Bodleian, IT, research services), none leading. Researchers needs are at the core. Started opportunistic; as more funders require RDM, more traction and RDM policy in place, more coordination.
Cost model they aim for is divided. Cost recovery model for highly specific services that will give researchers tangible benefits, and for research areas where funders are active. Elsewhere aim for centrally funded model, and include the cost in the overhead.

In November 2012, the project held a survey among 400+ Oxford scholars. Two highlights::
- one third of researchers work alone, and even more in small groups. Don't forget the smaller research groups or individual researchers, they are so easily overlooked in large projects!
- tabular data -excel- is the most often used data format. Puts importance of tabular data in perspective! Then comes Word, which "may be a container for all kinds of unstructured data we don't even want to start thinking about"... surprisingly, SQL databases come third.

Continued by Salley Rumsey and Neil Jefferies, on DataBank + DataFinder
DataFinder is the heart of DaMaRo. "Think of this as a discovery tool for data that is hosted elsewhere, a Data Catalogue". All kinds of data: both the results of research (residing in subject or institutional repositories) as well as licensed datasets. Researchers are strongly encouraged to at least register the existence of data (BTW the name 'data registry' was considered but rejected as too formal). And anything can be data, digital and analogue: you can register a specimen in a jar here. Integrated in DataFinder is an ID minting service, which also provides citation for dark archive objects. (one interface, mints DOIs for trusted digital objects, UUID's for others).
Pragmatic minimum metadata model: the five fields required for datacite, plus location, a small abstract, 12 fields in total.
The look and feel are similar to DataBank and DataStage and other DaMaRo parts, to make it come across as one service. Off note: "The visuals have a soft 'online questionnaire' approach, not librarian style with bold field labels".

Repositories / Data Archives
Inna Kouper (Indiana University). SEAD Virtual Archive. Building a federation of institutional repositories for long term data preservation.

The SEAD (Sustainable Environment - Actionable Data) Virtual Archive addresses requirements and proposes policies and architecture to address the needs of sustainability science researchers, who study the physical, biochemical, and social interactions that affect our planet. Sustainability science is an example of long tail science. SEAD is not only about the preservation of scientific data using institutional repositories today, but also on its rich access and use in the future.

SEAD VA is a federation layer over multiple institutional repositories which offers the community of sustainability scientists a coherent view on their collective data. In addition to serving as a federated deposit service, the SEAD VA performs another crucial function in data services. One could envision a repository supporting multiple lightweight federation services like SEAD VA, each of which serves a particular scientific community. If each federation service supports its outward interfaces via record-exposing protocols such as the ones developed by the Open Geospatial Consortium (OGC) or DataONE, a network of federated services would become a scientific search portal with a rich discovery interface.

Natascha Schumann (Leibniz Institute for the Social Sciences). The GESIS Data Archive for the Social Sciences: a widely recognised data archive that is on its way towards a new level of trustworthiness.

The presentation is about experiences in evaluating an established data archive (Leibniz Institute celebrated its 50th anniversary this year) with a longstanding commitment to preservation and dissemination of social science research data, against recently formulated standards for trustworthy digital archives. As stakeholders need to be sure that the data they produce, use, or fund is treated according to common standards, the GESIS Data Archive decided to start a process of audit and certification within the European Framework of Certification and Audit: starting with the Data Seal of Approval (DSA). An overview of workflows within the archive is presented and illustrated by some of the steps necessary to obtain the DSA, as well as to optimize some of its services.

E. Yakel (University of Michigan, School of Information). Trust in digital repositories.

Audit and certification of trustworthy digital repositories outline actions a repository can take to be considered trustworthy. But research that examines whether the repository’s designated community of users associates such actions with trustworthiness has been limited. Findings from interviews with 66 archaeologists and quantitative social scientists are presented. Similarities and differences across the disciplines and among the social scientists were found. Both disciplinary communities associated trust with a repository’s transparency. However, archaeologists mentioned guarantees of preservation and sustainability more frequently than the social scientists who talked more about institutional reputation. Repository processes were also linked to trust, with archaeologists more frequently citing metadata issues, and social scientists discussing data selection and cleaning processes. Among the social scientists, novices mentioned the influence colleagues have on trust in repositories almost twice as often as the experts.