Tom Pollard

Libraries and research data: exploring alternatives for services and partnerships

A workshop held on 28 June 2011 at the LIBER 2011 Conference: http://www.libereurope.eu/event/workshop-libraries-and-research-data-exploring-alternatives-for-services-and-partnerships

1. Introduction

Profound changes in scientific research place increasing emphasis on creating, using, and re-using data. This provides both a challenge and an opportunity for research libraries. In order to respond to this change, libraries are introducing new services around research data. This has required them to seek out new partnerships and work with new technologies. This workshop explored these new services and the approaches that research libraries have been taking to implement and deploy them.

The LIBER e-science group will organize four workshops addressing different views on the role libraries may play in the area of e-science. Two of the workshops will address strategic and managerial issues such as partnerships, service models and organisational consequences. Two further workshops will be more technical, addressing topics such as structure, interoperability and machine-usability. Each workshop will focus on a specific topic, with attendees being encouraged to take part in discussion and benefiting from talks on practical experiences. The outcome will be a report based on the results of the four workshops, to be presented at the LIBER 2012 Conference.

2. Brian Hole, Dryad UK

Brian introduced Dryad, which he described as a concrete solution for the management of data associated with research articles. Many datasets produced by researchers are not shared with the wider community, even after findings based upon them are reported in the literature. Dryad aims to address this problem by providing an international repository for data underlying peer-reviewed articles in the basic and applied biosciences. By accessing the data that underpins articles, researchers are able to validate published findings, explore new analysis methodologies, repurpose data for research questions unanticipated by the original authors, and perform synthetic studies.

Dryad UK, a JISC funded project run from the British Library, has been exploring how a Dryad-model service could be provided in the UK. It is exploring journal and publisher willingness to participate in the model, and has been working to develop a sustainable business plan.

The British Library has held several workshops to bring together stakeholders such as publishers, libraries, researchers, research institutions, and funding bodies to explore how such a service could be made sustainable. Valuable feedback has been gained from all stakeholders, and a report is due to be published in late 2011 that will capture this feedback and discuss possibilities for providing a Dryad-model repository in the UK.

The Dryad UK project is also helping to add value to the Dryad model by developing metadata standards. For example, work has been undertaken to develop MIIDI, a Minimal Information reporting standard for an Infectious Disease Investigation: http://imageweb.zoo.ox.ac.uk/wiki/index.php/MIIDI. The project is

also expanding the model into new disciplines, such as biomedicine and infectious disease.

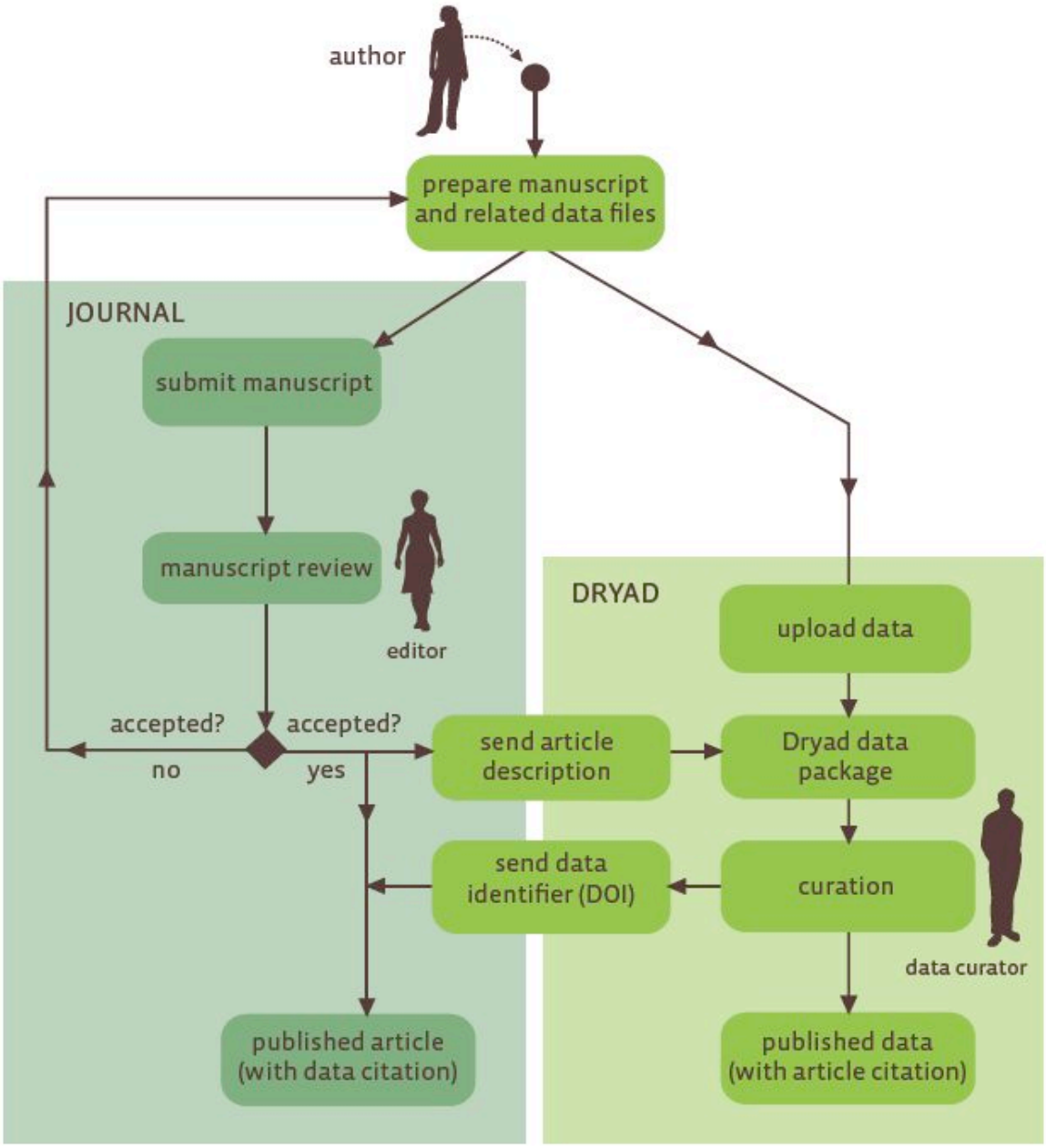


Image based on Lowry, R., E. Urban, and P. Pissierssens (2009), A New Approach to Data Publication in Ocean Sciences, Eos Trans. AGU, 90(50), doi:10.1029/2009EO500004.

Dryad UK has been well received by UK publishers, including the British Medical Journal and PLoS One, who recognise potential for increased revenue and impact. There may be a need for institutional level support that could be met by libraries as a service to the research community.

The resources required to offer a Dryad-model repository include maintenance of servers, provision of 'digital curators' (a new group in the library as we move away from analogue to digital), training costs, and customer support.

Libraries, Brian suggested, might be well placed to provide such a service. Libraries can offer a brand that can reassure the community of long-term stability, commitment and security. In addition, libraries have a tradition in curation and cataloguing skills, and often provide a hub for communication between the relevant stakeholders, such as publishers and the research community.

By providing such a service, libraries could gain: a broader and more useful collection; new skills and expertise in the digital environment; increased relevance in the digital environment; and perhaps a financial income. The future of data is unclear but it is likely to become increasingly important, and so it is crucial that libraries are proactive and develop skills and services now to ensure future relevance in this domain.

3. Merce Crosas, Dataverse, Harvard University

The Dataverse project started six years ago at Harvard University to meet the increasing demands from researchers for a place to manage and share data outputs from research projects.

Dataverse is a software application to enable data to be published, shared, and analysed. It recognises the differing needs of stakeholders such as researchers, data libraries, and publishers, and aims to provide not only tools for data management but also incentives for stakeholders to use the system.

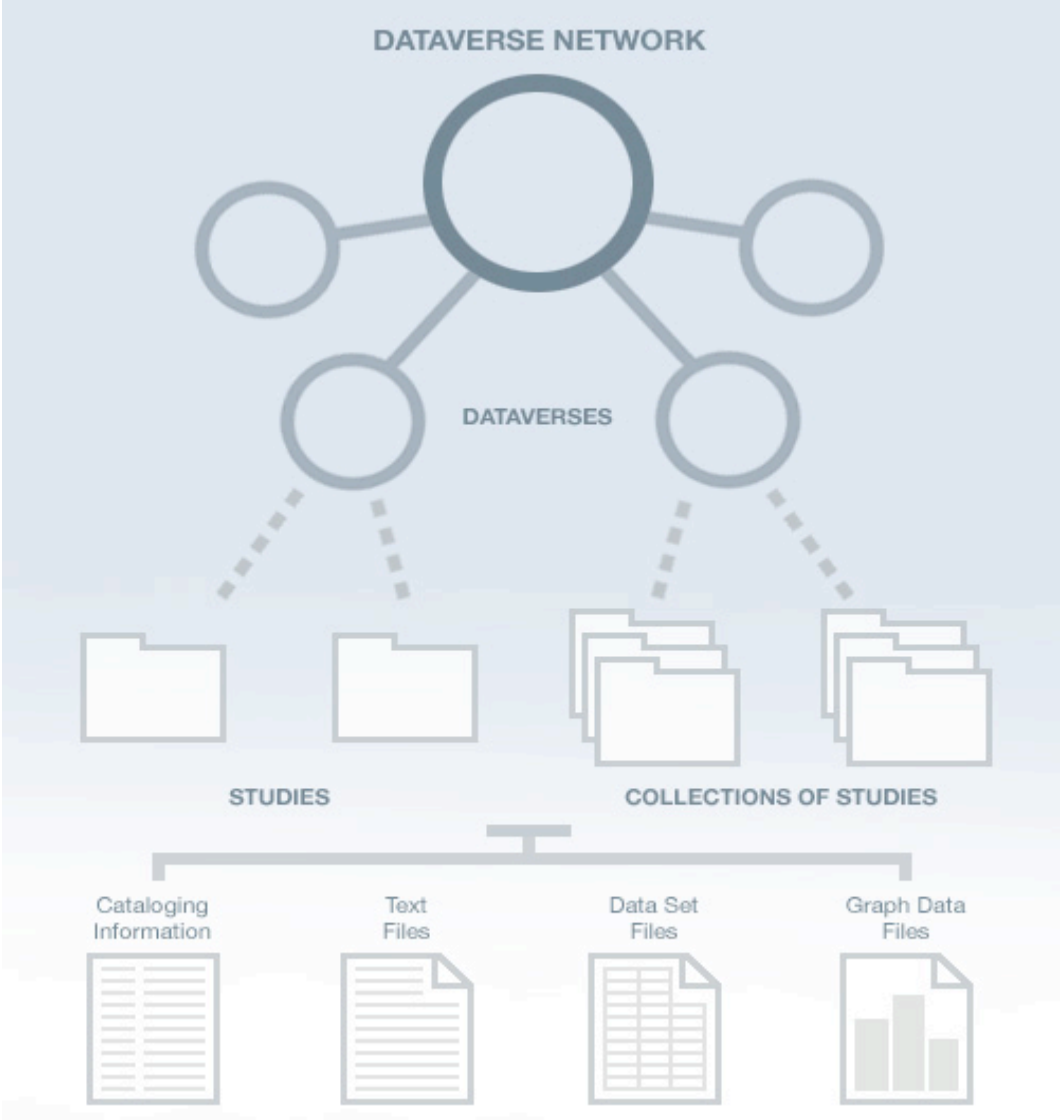
For example, Dataverse gives datasets unique identifiers allowing citations in literature to be tracked, potentially creating an infrastructure where researchers can be given credit and attribution for their data. Dataverse also helps to facilitate cataloguing and preservation of data, by offering centralised professional archiving, curation, and distribution control.

A Dataverse Network is installed at institutional level, and multiple Dataverses are created within the institution for individual research groups for example. Each Dataverse contains collections of studies, with each study including cataloguing information about the data and complementary files.

The service supports conversion to preservation formats, and visualisation tools are being developed. Dataverse also supports Universal Numerical Fingerprints that allow integrity of files to be checked. Workflows are customisable; including permission levels and access management, meaning the user can choose a personalised data management solution.

The Harvard Dataverse Network contains over 300 Dataverses. An example of one of these Dataverses is the Murray Research Archive:
http://dvn.iq.harvard.edu/dvn/dv/mra

The Murray archive is a permanent repository for quantitative and qualitative research data and comprises over 125 terabytes of data, audio, and video.



Another example of a Harvard Dataverse is the Harvard Election Data Archive: http://projects.iq.harvard.edu/eda/data. This archive is collecting election data from the past 20 years, enabling analysis of election results and potentially providing new insights into the election process.

4. Max Wilkinson, DataCite

Developments in technology are enabling researchers to generate ever-increasing quantities of data. However, currently our data infrastructure is fragmented and responsibilities are unclear. The scholarly communication system caters well for articles, but not for the underlying data. There is currently no clear method for identifying and citing data, meaning data is becoming increasingly difficult to discover or to access and finally data is being lost. Data is

treated as a 2nd class citizen, when in fact it forms a crucial part of scholarly communication.

These problems have been recognised for some time and were the driver for the formation of DataCite. In December 2009, building on preliminary work by the German Technical Library, several libraries and information organisations came together as DataCite to create a citation framework for data. Just as research is global, DataCite is global, with member institutions offering services and advice directly where they are needed by the researchers.

DataCite is a member of the International DOI Foundation and provides a service for data centres to mint DOIs and register associated metadata. Through this service DataCite is building a Metadata Store, which contains discovery metadata about the datasets that are being assigned DOIs. Services are being developed around the Metadata Store to allow third parties, such as Web of Science, to harvest the metadata.

Establishing a data citation framework involves both technical and social change, and so DataCite is also creating a data citation community to develop standards, guidance, and good practice. DataCite holds an annual meeting that brings together the data citation community, holds workshops to address the challenges of data citation, and participates in related projects such as ORCID.

5. Ed King: British Library Newspaper Project

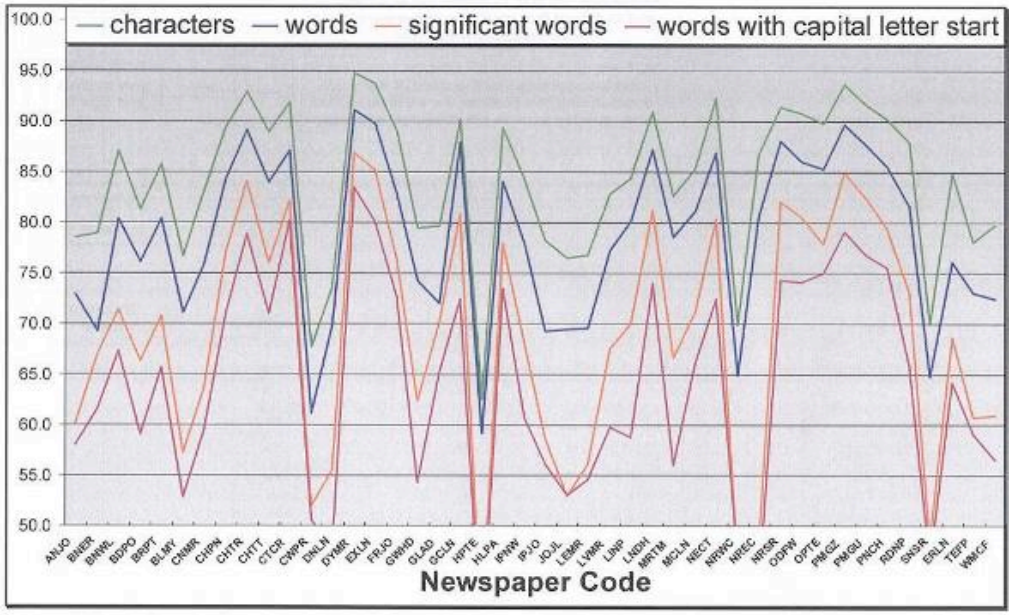
Focusing on the development of a single collection rather than broader services as previous talks, Ed explained that the British Library was working hard to increase access to its collection via digital technology. In 2004 the British Library received funding from JISC to digitise 2 million newspapers within its collection.

The size and scope of the project presented new challenges to the British Library and significant preparation was needed before the digitisation process could begin. A particular challenge was carrying out an inventory of the 2 million newspapers to understand the content. It was important to have a strong understanding of the content before it could be digitised – for example, knowing titles, characteristics of type, quality of paper etc.

The decision was made to create microfilms of every paper and then digitise from the microfilms. On average papers contained 3000-4000 words per page, so there was a lot of work to be done. The libraries criteria were that once a title was selected for digitisation, every article within the title had to be digitised.

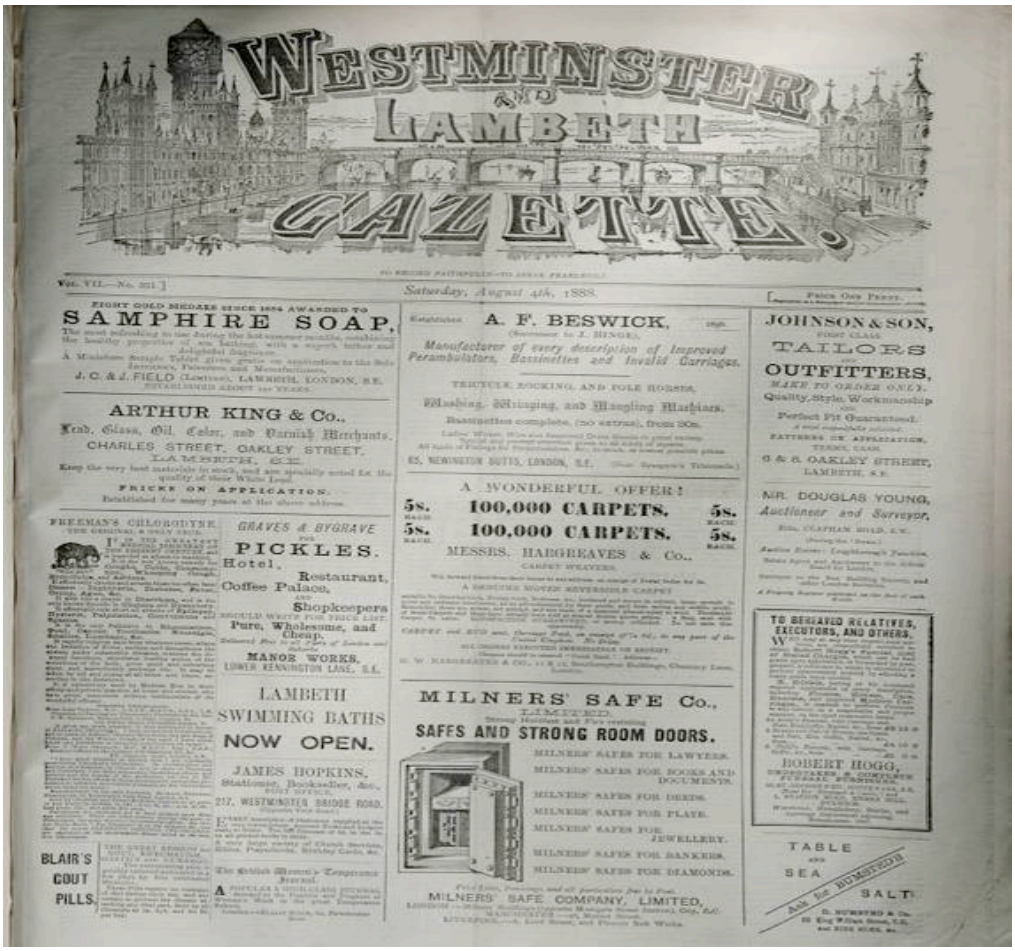
A commercial organisation was commissioned to carry out the digitisation work, and the costs of digitisation meant that the organisation was granted a fixed licence period where the company could sell content and recover costs. The result of this agreement is that newspapers are behind a paywall for a 10 year period.

The diversity of text within the newspapers (for example, font sizes as small as 4pt and decorated fonts) made OCR a significant challenge. An external review of the OCR work suggested an overall accuracy of approximately 78%.



When the results are mapped in a graphical mode by title a relationship between accuracy factors becomes very clear.

There was a strong relationship between accuracy and newspaper title, and classified sections were particularly challenging to OCR because of the breadth of content.


WESTMINSTER AND LAMBETH GAZETTE.

Saturday, August 4th, 1888.

Price One Penny.

EIGHT GOLD MEDALS SINCE 1884 AWARDED TO
SAMPHIRE SOAP.

ARTHUR KING & Co.,
Lead, Glass, Oil, Color, and Varnish Merchants,
CHARLES STREET, OAKLEY STREET,
LAMBETH, S.E.

PRICES ON APPLICATION.

FREEMAN'S CHLOROIDYNE.

GRAVES & BYGRAVE
PICKLES.
Hotel,
Restaurant,
Coffee Palace,
and
Shopkeepers
SHOULD WRITE FOR PRICE LIST.
Pure, Wholesome, and Cheap.
MANOR WORKS,
LOWER KENNINGTON LANE, S.E.

LAMBETH
SWIMMING BATHS
NOW OPEN.

JAMES HOPKINS,
Stationer, Bookseller, &c.,
217, WESTMINSTER BRIDGE ROAD.

BLAIR'S
GOUT
PILLS.

Established 1856. A. F. BESWICK,
Manufacturer of every description of Improved Perambulators, Bassinettes and Invalid Carriages.
TRICYCLE, ROCKING, AND POLE HORSES,
Mashing, Stringing, and Mangling Machines.
Bassinettes complete, (no extras), from 30s.
65, NEWINGTON BUTTS, LONDON, S.E.

A WONDERFUL OFFER!
5s. 100,000 CARPETS. 5s.
5s. 100,000 CARPETS. 5s.
MESSRS. HAIGREAVES & Co.,
CARPET WEAVERS.

MILNERS' SAFE Co.,
LIMITED.
SAFES AND STRONG ROOM DOORS.
MILNERS' SAFES FOR LAWYERS.
MILNERS' SAFES FOR BOOKS AND DOCUMENTS.
MILNERS' SAFES FOR DEEDS.
MILNERS' SAFES FOR PLATE.
MILNERS' SAFES FOR JEWELLERY.
MILNERS' SAFES FOR BANKERS.
MILNERS' SAFES FOR DIAMONDS.
MILNERS' SAFE COMPANY, LIMITED,

JOHNSON & SON,
FIRST CLASS
TAILORS
AND
OUTFITTERS,
MAKE TO ORDER ONLY.
Quality, Style, Workmanship
AND
Perfect Fit Guaranteed.
PATTERNS ON APPLICATION.
TERMS, CASH.
6 & 8, OAKLEY STREET,
LAMBETH, S.E.

MR. DOUGLAS YOUNG,
Auctioneer and Surveyor,

TO BEREAVED RELATIVES, EXECUTORS, AND OTHERS.

ROBERT HOGG,

TABLE
AND
SEA
SALT.

Benefits of the project for the Library included making the newspapers available to a wider audience, allowing researchers to access the content remotely, reducing the need for original papers to be handled, and allowing multiple users to search the same content at any given time.

Approximate aggregate costs of the project were £1 per page of digitisation (in the years 2005-2006), but it should be noted that the costs of image capture and processing are continuing to drop. The cost estimate does not include preservation of the digital content, and this is difficult to estimate because existing infrastructure at the Library is being utilised for storage and preservation.

Where an external digitisation route is taken, local costs includes staffing and resource for (1) performing condition checks on materials and logging fragile material before sending to supplier (2) ensuring sufficient metadata are captured, including rights, provenance, and preservation (3) ensuring regular QA inspections and responding as necessary (4) ensuring data is ingested into the library's digital store. Good project management is also crucial.

Perhaps the biggest lesson from the project was the importance of having a strong understanding of the content to be digitised. Any exceptions within the content (such as incorrect dates, inserted pages, supplements etc) must be

noted and the implications must be assessed. Where an exception arises during the digitisation process, they should be dealt with as soon as possible and the decision recorded to ensure consistency and to avoid delays.

6. Kevin Ashley: Digital Curation Centre

What is curation?

Kevin discussed the meaning of curation: “Maintaining, preserving and adding value to research data throughout its lifecycle”. He said that curation could be both more and less than preservation. For example it could require active management in dealing with change, whilst also requiring decisions to be made on what to discard. Curation can add value to a digital object, and can also involve sharing, publication, and citation.

Why care?

Kevin said that there were many reasons why libraries should care about curation of data. Data is often expensive to produce and may be irreplaceable, so it is an investment that needs to be maintained. Data is not just an issue for research, but also from government and industry. Reuse potential is far reaching, from education advancement to financial forecasting and planning.

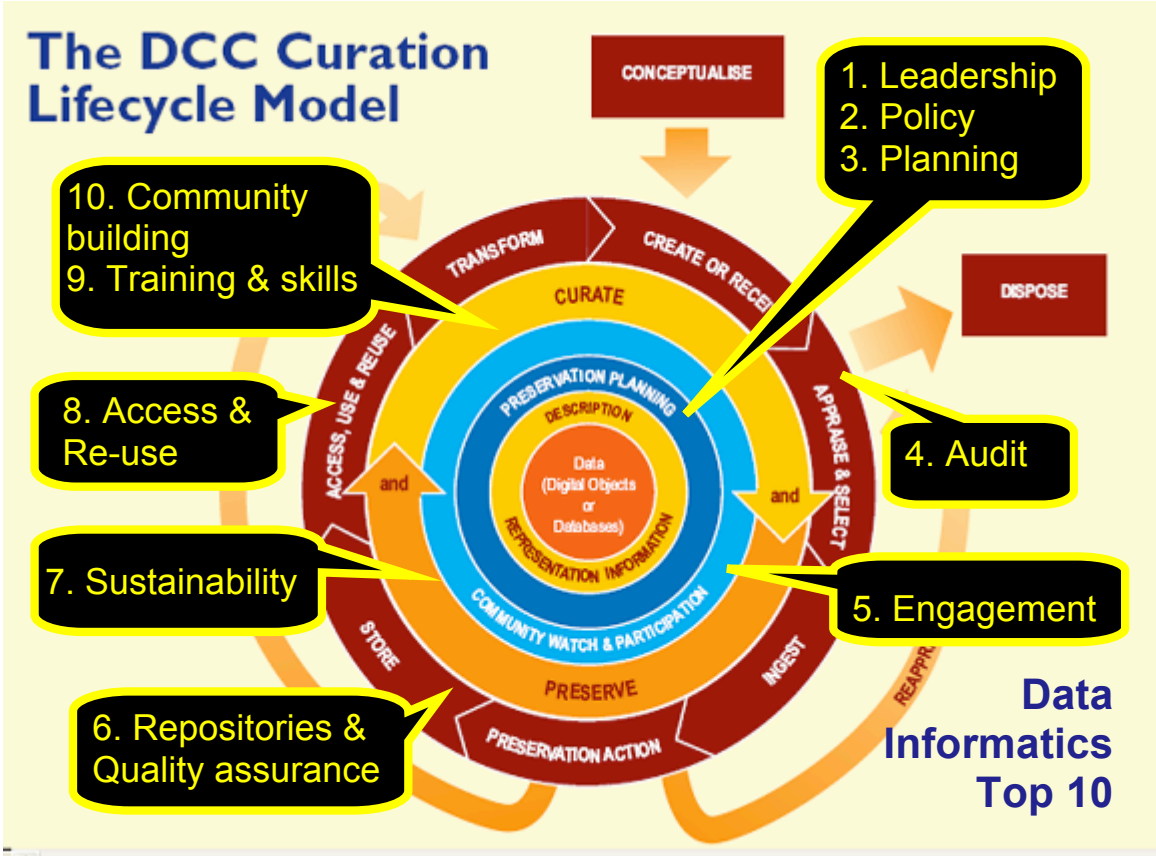
Data is managed at numerous levels. Data may be well managed within domain-specific repositories, and by institutional, national, and global organisations, but the infrastructure is often fragmented. Below the institutional level, data is often poorly managed.

Role for libraries

Libraries are traditionally associated with selection and access to content and they could continue to provide these roles in the data landscape. Libraries could also provide leadership and coordination in the data management infrastructure and provide services for example at researcher and research group level.

If research centres are not managing data, then libraries should take responsibility to coordinate action. Libraries may also have a role in auditing the data landscape to ensure that there is awareness of what needs to be preserved and to understand the data management landscape.

As summarised in the diagram below, roles for libraries may include: leadership and coordination; audit (who has what, where does it go?); advice on access and reuse of data, preservation and permanence; citability and discovery of data; data management education and training, and community building.

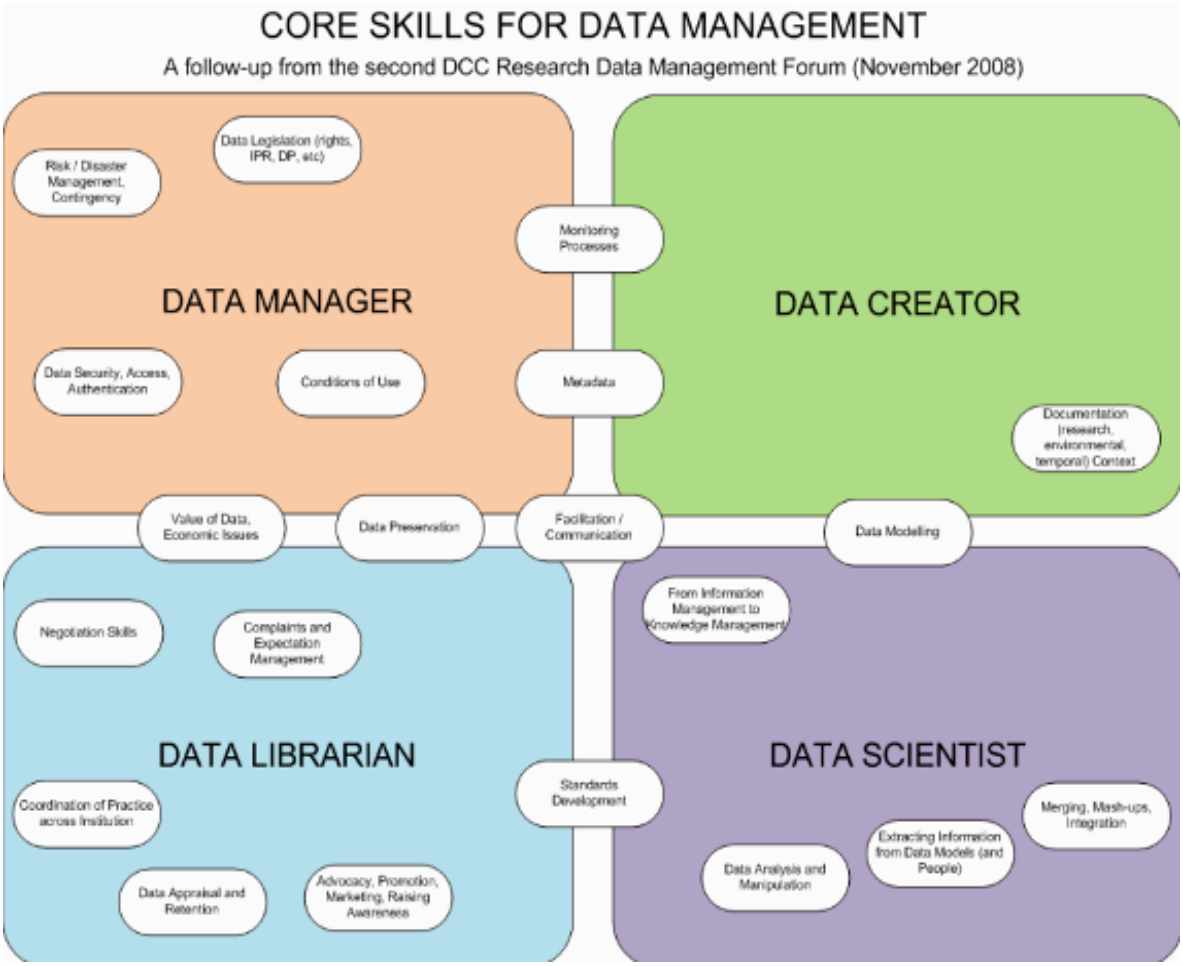


How to get there?

To ensure libraries continue to stay relevant within the data landscape, it is important for them to take a lead in developing services and skills necessary. There are many areas where libraries can be involved, including: creating and developing policy in collaboration with data centres, funders, publishers, etc; implementing existing digital services locally where appropriate; learning and using audit tools made available by organisations such as the Digital Curation Centre.

Libraries should also take a lead in learning about data and its sources, taking into account more than simply scholarly data where appropriate (for example, social media data and commercial data). Libraries should ensure that data literacy is promoted internally, for example by ensuring subject librarians understand existing data resources and use them where appropriate. In addition, libraries are in a strong position to act as a bridge between stakeholders in the data landscape such as publishers and researchers.

In 2008 a report for JISC ("The Skills, Role and Career Structure of Data Scientists and Curators: An Assessment of Current Practice and Future Needs", Key Perspectives) explored core skills and roles for data management and identified a need for specialised data librarians.



Considerations

It is important to recognise that data centres are already established in many domains, and libraries should seek to build on this infrastructure rather than offer competing services, Kevin suggested.

The nature of datasets is very different to the nature of publications such as monographs and research articles, and so thinking of data solely as a publication is too narrow a view. For example, datasets may continually change and may have indistinct boundaries.

Kevin suggested that the three key questions that libraries should consider are: (1) How does data management align with institutional mission (2) what skills must you acquire? (3) Why you? When coordinate and when take action yourself?

7. Ricky Erway, OCLC Research

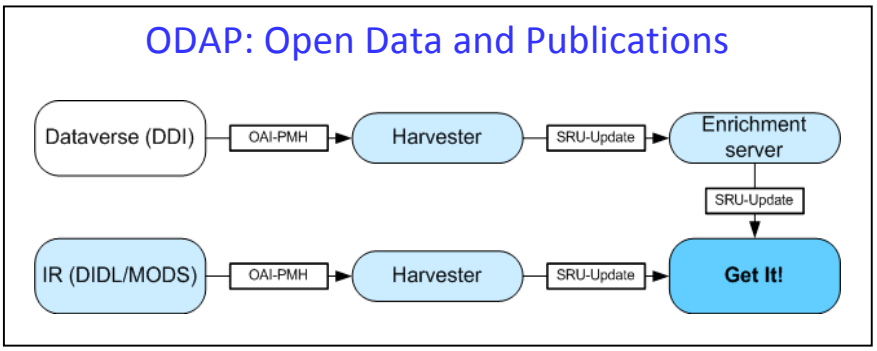
Linking funding to research output can create incentives for researchers to share their work. Ricky suggested funding in the UK is more closely linked to research output in the UK than in countries such as the US, and suggested that the US needed to close this gap.

She noted that in the US, certain subject areas were well funded (for example the sciences), while other areas, such as humanities, were largely ignored. She asked what the funding/business model should be for the 'crazy quilt' of data repositories and how they should fit together. This question is being researched by the OCLC.

8. Marc van den Berg, Tilburg University Library

Marc gave an overview of the data projects at Tilburg University Library.

Data projects at Tilburg University Library



CARDS project acronym: Controlled Access to Research Data Storage.
The project targets *best practices in research support* for Netherlands Academia.
Tilburg University use case: the World Tax Indicators (WTI) portal.
Challenge: The CARDS project aims to fill in data gaps that limit taxation research and fosters the visibility of taxation research methods and concepts.
Core technologies and tools to be used in CARDS: SDMX, DataVerse and R.
Deliverables:

- SDMX gateway for tax indicator discovery and data dissemination.
- Documentation of estimation procedures in R.

More info: rob.grim@uvt.nl

The first project, Open Data and Publications (ODAP), links Dataverse to data within institutional repositories. Researchers will be able to link their datasets to their publications, allowing an integrated front-end to deliver an enhanced publication, available for reuse.

The second project, Controlled Access to Research Data Storage (CARDS) aims to offer specific assistance to researchers who are independently managing their research data. Researchers experience challenges in storing and controlling sharing of data on a daily basis. The Tilburg University use case focuses on the World Tax Indicators portal. International tax data are incomplete and difficult to compare. Key variables for international comparative research on taxation are still missing. The world tax portal needs to be aware of the data that is already

available elsewhere. The project therefore builds on standards, data formats and tools that are used i.e. at the IMF, OECD, Worldbank and ECB. The deliverables enable the research community to easily store, discover and reuse tax data indicators and research methods for comparative research.

9. Summary and discussion

Involving researchers

Researchers are key stakeholders who must be engaged in the discussion. Are there examples of strong collaborations between the research community and the library community, one member asked? It was felt that some libraries are better placed to engage with researchers than others. For example, university libraries sometimes have a pool of local researchers. A potential strength of libraries is their close association with researchers. Libraries should make use of this association to understand their needs and assess what services are needed.

Incentives for researchers

A key question was the importance of creating incentives for researchers for sharing data.
Kevin said that sometimes, increased discoverability (of data) was seen as a positive incentive for researchers. There is some evidence to suggest that making data available may increase citations of associated research articles (Piwowar). In addition, ANDS have shown that while acting at a national level they have made Australian research more visible to the world than it was. So you can have a big impact even if you operate at a national level.

Merce said Harvard was an example of a library that closely involves researchers. She noted that researchers measured prestige on citations (& salary is correlated linked to number of publications), which was a big incentive for researchers to engage with the library. She noted that if the same was true for data citation, this would be a strong incentive for researchers. She said that initiatives such as DataCite were paving the way for this by creating a citation infrastructure.

It was suggested that libraries should begin collecting positive examples of data sharing (for example, researchers who have benefited from sharing data), to help us answer the question ‘why should I share my data?’

Supporting reuse

We also need ways of ensuring that data is as reusable as possible, for example, providing methods for visualisation of data, and standards for ensuring interoperability and reuse.

Ownership was also felt to be a key issue here. A licence cannot be asserted unless ownership is clear, so it is still critical to know who owns the data. Only an owner can create the licence.

Funding and sustainability of services

A representative from Open University said that sustainability of services was a key concern when deciding whether to provide a service for data. They are keen to provide services, but are also wary of managing expectations and being able to offer something that is sustainable.

Ed said that from a project management perspective, it was important to ensure that senior staff at the library were aware of the benefits of data management, because these staff had a key role in allocation of funding.

Birte commented on a presentation she had seen from CERN, which had suggested that 10-15% of grant allocation would need to be set aside just for data management activities. This is a challenge! Ed suggested it is just a continuation of the current challenges for archives and libraries.

Differing nature of libraries

A representative from a Belgium library noted that the nature of libraries varied widely. It was clear that national libraries might have a different role to university libraries for example. The environment was also important, with funding infrastructure and political will being particularly important. Question is whether some services, such as the DataCite services, could be provided by local libraries, or whether this role should be solely for the national libraries.

Ensuring libraries remain relevant

There are concerns that without finding a role within the data landscape, libraries could increasingly become disengaged and unnecessary for researchers. Max suggested that libraries should begin to think of data as a first class research object, on the same level as traditional articles and monographs. He said that this information will increasingly be seen as an asset, and libraries will need to begin to identify how these assets can be put to best use for the library and for its researchers.

Prepare to make mistakes

Kevin said that it was necessary for libraries to be bold in this new environment. He said that libraries are traditionally cautious, for good reasons, but it was important to take a lead and explore possibilities. He gave the example of DataCite and Dryad UK, which he said did not offer perfect solutions, but which were addressing the problems now, rather than waiting for a perfect solution.

Libraries and data centres

Wolfgang suggested that there was a tendency for some to say that local libraries don’t have a role in the data landscape, and that data centres should take the lead in this area. Wolfgang suggested this idea should be challenged. Reality is that both local and national repositories of data are needed, though this might provide new challenges.

Birte said that if libraries are putting together a data management plan, then libraries should push hard to be involved. If we lead, then we create an opportunity or even retake a role. Libraries should be places of education and learning, and if this is being lost then it needs to be taken back.

It was suggested that greater definition was needed between the role of the library and the role of the data archive. Should libraries be the place for data deposit? Maybe the question is not ‘should’ but ‘could’, said Birte.

Kevin gives one example where the University of Edinburgh had produced a data management policy approved by highest levels of university. To make this happen, both the archive and library needed to work together. Archives needed to make sure policy captured need, Kevin suggested, whereas libraries represented access, learning, and teaching.

Should libraries play a role in promoting good practice and standards for data sharing and management? Merce said curation sometimes required specific knowledge about domains, so could be better suited as a role for research institutions. Kevin said that mistakes were inevitable and that it was important to accept that sometimes important data would be discarded. Kevin gave an example of the UK BSE-crisis, where information about slaughtering of animals had been discarded shortly before the outbreak.

Partnership and collaboration

Birte highlighted the importance of partnership and collaboration. Libraries and data centres need to identify areas of responsibility, and researchers should be closely involved in the decisions. Actors and roles need to be identified and are unclear at present.

We also need to recognise that different libraries will have different specialist areas, and we shouldn’t set out to do everything ourselves. Through partnerships we can offer a complete range of services. Some libraries might offer identifiers, some might offer support etc. We need to be able to work together.

We also need to improve engagement between people who are solving different problems, for example by connecting the data citation community and the linked data community.

Business models and service provision

What we haven’t covered is how to pay for services. One example is newspaper where there is pay per view. Business models will change and we will need to touch upon this in one of the next workshops. Important because there is a need for sustainable services – can’t afford for even one to fail, because it endangers how you could get to data. (E.g. identifier fails so can’t get data, repo fails etc…) This ties to reproducibility of results in papers (hence nature require data associated with articles), so this will also need to be specifically addressed.

Technology versus social challenges

It was important to separate the technical challenges from the social challenges. Software such as Dataverse addresses technical issues, but social challenges requires organisational discussion and collaboration, with engagement of all parties. Could libraries have a role in leading this discussion?