I've collected my data, so what do I do with it now?

Research data management

Session 4

Data For Life - Digital Preservation for Health Sciences

Tutor Notes

DATUM for Health

www.northumbria.ac.uk/datum

Project funded by JISC

Session 4. Data For Life - Digital Preservation for Health Sciences
Notes for Tutors

SESSION DETAILS

Aims and Objectives/Learning Outcomes

By the end of this session participants will have:

- an awareness of digital preservation issues
- knowledge about digital preservation approaches
- awareness raising of the wider research data management agenda beyond the PGR student
- awareness raising of initiatives in other organisations
- networking opportunity.

Session Content

- A unique 'roadshow' style event organised in collaboration with the Digital Preservation Coalition (DPC) on digital preservation within the context of health.
- Morning focused on an introduction to digital preservation issues and approaches plus case studies. Afternoon focused on strategic issues of managing research data and further case studies.
- Afternoon session optional for PGR students.

Structure

Programme: 26th May, Room 204, Lipman Building, Northumbria University, Newcastle

1030 Registration and Coffee
1100 Welcome and Introductions
Prof Julie McLeod, Northumbria University (Datum for Health Project)
1110 Digital preservation and long term access: challenges, opportunities, approaches and tools
William Kilbride, DPC
1135 Managing qualitative data
Louise Corti, UK Data Archive, University of Essex
1200 Managing qualitative data for health research: A researcher's experiences
John Given, Researcher in narrative studies in health
1225 The challenges of digital preservation at the London School of Hygiene & Tropical Medicine
Victoria Cranna, London School of Hygiene and Tropical Medicine
1250 Questions
1300 Lunch
1345 Data – manage it, make the most of it!
Prof Charlotte Clarke, Associate Dean (Research), School of Health, Community and Education Studies, Northumbria University
1400 Strategic view of research data management for Higher Education
Dr Simon Hodson, JISC
1425 Managing qualitative data for health research: A case study
Ruth Sanders, The Health Experiences Research Group, University of Oxford

1450 Tea and coffee
1505 The Challenges of Digital Preservation in a Changing Environment
Andrew Pitt, Pfizer
1530 Roundtable: what is to be done, why and by whom?
Chair by Julie McLeod with Charlotte Clarke, Louise Corti, Simon Hodson, and William Kilbride
1600 Close

Directed learning tasks

- Make an entry in research diary / portfolio about research data management strategy and actions going forward
- Make an entry in training needs analysis document about research data management
- Meet with supervisory team to report on research data management learning and plans going forward

Handouts

- In line with the DPC's practice of making roadshow event presentations available in digital form only; DPC materials (e.g. case studies, guidance) were available for delegates to take

Presentations from the event are available via the DPC's website
http://www.dpconline.org/events/previous-events/730-data4life

NOTES TO ACCOMPANY THE PRESENTATIONS

PROF JULIE MCLEOD'S OPENING REMARKS

Good Morning. My name is Julie McLeod, I'm Prof in Records Management in the SCEIS here at Northumbria University, and it's my pleasure to welcome you all to today's event *Data For Life - Digital Preservation for Health Sciences.* This event is a new collaboration between Northumbria University and the Digital Preservation Coalition (the DPC), and part of the *DATUM for Health project*; which Northumbria University is conducting.

DATUM is one of five JISC funded projects which are "*promoting discipline-focussed research data management skills in Higher Education Institutions*" by designing and piloting discipline-focussed training programmes for postgraduate courses."[i] This training strand is one of several strands in JISC's *Managing Research Data Programme* a £2m initiative that is addressing a "strategic requirement for UK HE to improve its research data management capability and better to understand how this may be achieved" as well as encouraging the wider reuse and repurposing of research data for the further advancement of research. I won't say anymore about the initiative as Dr Simon Hodson, the Programme's Manager, is speaking this afternoon and I'm sure will share more about its aims, activities and achievements. But I will say a little more about the DATUM project to set the context for today's event.

DATUM is a 10-month collaborative project between 3 Schools and 2 external organisations. The internal schools are:

- School of Computing, Engineering & Information Sciences bring the information/data management subject expertise
- School of Health, Community & Education Studies provides the discipline focus with Prof Charlotte Clarke the named scholar in health, and
- The Graduate School, responsible for PGR students and their training (the Director Prof John Dean is on the team)

Externally: Digital Curation Centre and the DPC are 2 leading organisations specialising in the areas of digital curation (i.e. management of digital information throughout its life) and digital preservation (focusing on the long term preservation of digital information). They also provide training/awareness and that is their main input. [Dr Kevin Ashley, Director DCC & colleagues; Dr William Kilbride, Director DPC]

This is an exciting new collaboration which brings together complementary knowledge and experience in all key aspects of the project [research / information management training & education, research, research data management, proposed methods and the important discipline-focus of the project]. We also have an Advisory Panel which brings independent advice on RDM requirements & helps raise awareness (includes staff from our Library / Institutional Repository, supervisors, Research, Business & Innovation & the Grad School responsible for training PGR students).

DATUM is focusing on:

- health studies discipline AND
- postgraduate research (PGR) students (i.e. doctoral students), as an integral part of a doctoral training programme covering generic & discipline-specific issues

and concentrating on

- qualitative, unstructured data AND

- covering the whole data management lifecycle not one particular aspect e.g. planning or preservation

Our (andragogic) approach to developing the training programme is problem-oriented so that the content is immediately relevant to participants' work, with activities, discussion, sharing of experiences. Its four sessions have:

- introduced Research Data Management - why it's important, goals & how to develop a data management plan
- the data curation lifecycle
- problems, practical strategies and solutions AND (today)
- digital preservation

Having been to a DPC Roadshow I was keen to include something like that but focused on the health discipline. And was delighted that William Kilbride was keen to collaborate.

Although the DATUM training programme participants are PhD students at different stages (from Northumbria and other universities in the region), we wanted to open today's event to a wider audience (a) because a lot of health related research takes place in the region and (b) there hasn't been a DPC event in Newcastle before. So, a special welcome to those who are here and not part of the DATUM project.

The aims of today are to:

- introduce key concepts of digital preservation
- learn from case studies about emerging tools and technologies for managing data
- provide a forum to review and debate current issues & future developments in preserving digital qualitative research data in health, and
- start a discussion on how the necessary skills can most effectively be developed

We've therefore invited a range of speakers who will share their knowledge & experience about services, solutions, strategy & policy for RDM in health. I'd like to thank them all for accepting out invitation to speak. As you can see from the programme we have a very full day covering:

- Effective management of qualitative data in health
- Emerging tools and services for long term access to research data
- Research data management: policy and practice
- Practical research data management skills for health professionals]

We will try to have time for 1 or 2 questions at the end of each presentation and have built in time for questions just before lunch and in the panel session at the end of the day. For the roundtable we have put some post-it notes on the table for you to write a question you'd like me to ask the panel; we will also take questions from the floor.

Prof. Peter Golding, the University's Pro Vice-Chancellor (Research & Innovation), had hoped to open today's event but unfortunately an overseas commitment, the timing of which was outside his control, means that he can't be here. However, he has sent these words:

> *"I am very sorry not to be able to join you at this important and exciting conference. To anybody working in research, in whatever discipline, it is obvious that generating new data means knowing how to manage, maintain, and disseminate it. Research is the art of creating new knowledge and understanding, but without the means to harvest that knowledge, and make it available to others, research is in vain. Your research is focussed on health, and all research bodies are agreed that the development of health management, and the contribution of research to the health*

and well being of this and future generations, is increasing all the time. It is vital that this new knowledge is not wasted, and this makes the management and maintenance of data arising from research imperative and crucial.

I know this fundamental work cannot go forward without the attention to training of those whose task it is to manage health research data. In this your work, and that of the DATUM project, are crucial. Data4Life seems to me a wonderful crisp way of capturing what this work is all about. I am very sorry not to be able to join you, but wish you well in your conference and the very best of success in your future work."

And now I'd like to introduce our first speaker, Dr William Kilbride, Director of the DPC.

MEETING NOTES

Digital preservation and long term access: challenges, opportunities, approaches and tools. William Kilbride, DPC

Summary of tweets

A link to a guide about the OAIS reference model mentioned at #data4life: http://bit.ly/ljY2QI

A link to an article explaining more about #LOCKSS - http://bit.ly/ln8vwo

Link for PLATO preservation planning tools / methodology / library http://bit.ly/sm5XA

Link for pronom/droid for file characterisation and management http://bit.ly/34gvkh

PRONOM/DROID, PLATO, LOCKSS part of a preservation architecture

Preservation tools, PRONOM: technical registry of file formats

Preservation tools, DROID: tool to identify files by extension, provides checksum

In the "technology/organization/resources" triangle there's been too much emphasis on "technology"

4 key digpres elements are: migration; emulation; hardware preservation; and exhumation

Emphasising obsolescence of digital research data storage & preservation techniques themselves

Data storage is cheap, (re)discovery is expensive

Digpres should focus on people and opportunities rather than 'data', 'risk', etc

Managing qualitative data. Louise Corti, UK Data Archive, University of Essex

Discussion after the presentation

Ownership of data. Who owns it? Participants own their own words, though researchers own the recording / transcript. Health records are owned by the Secretary for State. If data from elsewhere is used in the research, then the copyright of that data is held by someone else.

Summary of tweets

UKDA model consent form http://www.data-archive.ac.uk/media/210661/ukdamodelconsent.doc

Data management lifecycle needs to identify the critical intervention points

Mentions file format issues - here's a useful introduction and guide to the topic http://bit.ly/mSbo2u

Important that researchers document their datasets to ensure context is understood by other users

Researchers need to be aware that data destined for sharing must be anonymised

Important to consider: informed consent, protecting identities, restricting/regulating access where needed

Confidentiality and documentation are critical for social science research data

UKDA focuses on *sharing* data - digpres underpins this

Data protection act: nota bene anonymised data does not fall under the data protection act

Kilbride: preservation tools, PRONOM: technical registry of file formats

Managing qualitative data for health research: A researcher's experiences. John Given, Researcher in narrative studies in health

Discussion after the presentation

Some researchers are reticent about making data publicly available. The 'academy' focuses on confidentiality. Consumer groups are desperate to tell their stories publicly, so John chooses to work with them.

There is a deluge of data, and secondary analysis of this data. And this is all public.

Research councils want to measure the impact of research; the impact on communities, and use of videos, workshops, plays, working with community groups, are all ways suggested to make impact. But the two strands will exist in parallel: (1) this community / impact strand, and (2) the academic strand published in peer-reviewed journals.

Summary of tweets

Good point about the e-thesis being more than just text placed online

Highlights problems in recording personal narratives that the subjects later wish to change/suppress

Understandably questions how guarantee confidentiality of digitised human subject data but maybe argues case for

Talks about challenge of ongoing consent narrative in research. People want (have a right?) to change their narratives.

Capturing "narrative performance" of identity. Overcoming limitations of text only transcripts via streaming

Notes Robson 2007: notion of hypertext or multimedia thesis linking to electronic sources, data etc

Performance of identity means we need digital resources not transcription for research - and thus #digitalpreservation

Therefore need data collection approach that captures performance

Interested in the narrative performance of identify, rather than the construction of identity

Introducing digital preservation issues in narrative research in #data4life - making sense of his collisions with technology

First mention of data deluge in #data4life - notes also that HE is a small part of the research community, and not even typical of research?

The challenges of digital preservation at the London School of Hygiene & Tropical Medicine. Victoria Cranna, LSHTM

Discussion after the presentation

Different views about metadata: metadata at the technical/variable level (researchers, IT) and metadata at the descriptive level (records managers, archivists). Records managers etc. need understanding of researchers' methods, researchers need understanding of the need for descriptive metadata. On the Web there are numerous different versions of the LSHTM's name, so a controlled vocabulary is necessary.

Summary of tweets

email archive issues mentioned at #data4life - might be interested to know that #dpc has an event on email archiving late July

Q to Cranna: I can't find data retention policy on LSHTM page listing research data policies - where is it?

Challenge of communication between archivist and researcher: catalogue, collection level metadata vs research variables metadata

New LSHTM WG on making research data accessible to comply with funder requirements

2003 JISC project at LSHTM led to adoption of data retention policy, data kept for at least 10 years

Digpres terminology an issue - researchers have different views of 'metadata' (and thus whose responsibility it is)

Lon Schl of Hygiene & Tropical Medicine now has mandate of minimum 10 year preservn of research data - reusable?

LSHTM research data working group on making data accessible to comply with funders' requirements and aspirations

Most LSHTM staff very reluctant to acknowledge problems in storing/managing their data

LSHTM, 2003 JISC-funded project on data retention policy: researchers' reluctance to engage, problems of storage media

Data – manage it, make the most of it!. Prof Charlotte Clarke, Associate Dean (Research), School of Health, Community and Education Studies, Northumbria University

Discussion after the presentation

The cardboard box - the power of physical objects!

We've talked of technical and archival metadata but we also need the 'research question' metadata. Risk that research data management is seen as a separate activity with nothing to do with research. Methodology and research data management are inextricably linked.

Summary of tweets

Cardboard boxes we've known and loved - http://www.ahds.ac.uk/creating/case-studies/newham/index.htm

The Newham kneecap has had greater value as a teaching example than archaeological artefact, I think!

The cardboard box: this is a ‘heart-sink’ moment for me, a jigsaw, a challenge to reconstruct the data

The "cardboard box of stuff" is a great way of explaining data management

Typically researcher's data collection is a Pandora's Box lacking a theoretical framework

Speakers emphasise importance for real world human research studies of capturing research questions as well as data

Important point: Methodology and research data management are inextricably linked!

Strategic view of research data management for Higher Education. Dr Simon Hodson, JISC

Discussion after the presentation

When will we see significant change? After the next stage of the programme. JISC funding is often to enable the activation hurdle to be overcome.

Summary of tweets

Lots of useful info & links in slides; summarising state of play in #jiscmrd with health sciences emphasis

Call for 2nd phase of JISC's Managing Research Data prog now due in June 2011

Notes wide-ranging value of JISC's Incremental project (Cambridge @theUL + /Glasgow)

Introduces the foggy bottom accord and other statements on research data access and integrity

http://bit.ly/jWARHn Research data management policy at University of Edinburgh University http://www.ed.ac.uk/schools-departments/information-services/about/policies-and-regulations/research-data-policy

Edinburgh's new 10-point res data management policy places major responsibility on PIs

Mentions UK Res Councils' data management policies for researchers - but do these "requirements" have teeth?

Managing qualitative data for health research: A case study. Ruth Sanders, The Health Experiences Research Group, University of Oxford

Discussion after the presentation

Keenness of people to volunteer: very keen for some topics e.g. motor neurone disease, hard to recruit for other topics, e.g. psychosis.

Differences in volunteering between classes: often more volunteers from the middle class, but for the topic of the menopause. more working class women volunteered.

Summary of tweets

Systems: how and where to store the data; need for training researchers on the systems; e.g.s of filename conventions

HERG holds 2500 interviews + '000s of transcripts etc

http://www.herg.org.uk/ 61 projects in healthtalk

Oxford HERG focus is on what patients want to discuss (not what doctors think they should discuss)

The Health Experiences Research Group - Partners include Monash, Australia

The Challenges of Digital Preservation in a Changing Environment. Andrew Pitt, Pfizer

Summary of tweets

Numerical data with 46 decimal places, checking data integrity with a hexadecimal editor. Yikes

Stresses importance of digital curation to maintain institutional memory, counter effects of staff turnover, system change

Virtualisation & emulation approach described by Pfizer incl compression & economies in architecture

Global corp may have data with multiple external attachments (e.g. 21 million scanned tif docs!)

Validation a key issue for Pfizer in their migration processes

Fascinating talk about digital preservation at Pfizer - cultural, technical, regulatory, business and processes

Dig pres in the private sector - challenges are business and technical

ROUNDTABLE - SUMMARY OF DISCUSSION

Opening Remarks

Welcome back to 4 of our speakers (represent organisations explicitly involved in promoting / providing DP services & Charlotte Clarke who represents the stakeholder community). They have agreed to answer questions and give their views on the future for ***digital preservation specifically and RDM more broadly in health.*** Might even pose some questions.

We're taking notes of this session and will post a summary of the questions & discussion. If we don't have time to discuss all the questions we'll address them in our report.

Let's start with a question (from post-it notes) ...

Questions Asked

Why is the wider health / medicine community not more engaged with digital preservation and related? Surely this is not just a question for the academic community?

Should we keep all research data?

Research council 'requirements' for research data seem to lack teeth - what's the stick (or carrot) to secure compliance?

Are research council policies enforceable or enforced?

How can I persuade senior managers to engage with digital preservation topics?

Should data from PhD studies be kept for 10 years (RCUK requirement)? If yes, who will keep it and where?

Questions Debated

Why is the wider health / medicine community not more engaged with digital preservation and related? Surely this is not just a question for the academic community?

The health community doesn't appear very much in the digital preservation world.

Not many good, published case studies of digital preservation in the health community.

Cause could be the nature of the data. Research data is very complex re ethics and governance. Ethics - very specific what can be done with the data. Bars you from sharing and public access. More emphasis on deletion. The health ethics system needs to become more permissive.

Health and medical community is huge and diverse. Genomics led the way in digital preservation; other areas are lagging behind. The health funders ??accord on data sharing should help.

Get around the governance issues with anonymised data; not always trivial to anonymise data.

NHS side - sheer number, variety of different technical systems, and even within one trust. So real overheads for integration and preservation. Complicated picture.

Deep rooted cultural differences. Universities - long term, build on knowledge. NHS - sharp end is patient care. ??NHS R&D programme on research access but this has lost funding. There are some good examples of University and health linked research programmes, e.g. The Cambridge Health Alliance http://www.challiance.org/home/index.shtml.

Is the cost issue really able to be resolved.

Challenging. No easy answers. Solution is a collection of answers.

Some research councils have a principle that cost for data management can be included in bid. But resistance to this from some researchers, as affects the funding for the actual research. Other research councils do not allocate such funding; it's up to institutions to provide the resource. There is a strong case for discussion across the Research Councils to develop consistent requirements and guidelines on RDM requirements.

What are all the stakeholder views?

Lifecycle - funding for research infrastructure - where does that come in?

Significant benefits and impacts to improve research data management drivers.

In order to achieve cost cutting for RDM: cheaper storage solutions, efficient management procedures, particularly for appraisal.

What equipment for storage does a project / research group need? This might be available through project funding. But longer term, who has the responsibility? The long term is the issue as RDM may be budgeted for the life time of the project but longer term funding is more difficult to cost and source. Services range from in house to an outsource specialist. 'Trust' issues are also raised. A 'trusted' and equipped digital repository is a complex commitment.

So who are involved? What are the solutions?
For some aspects it's collaboration, e.g. file format registry. Data storage, data protection issues, etc. need to be addressed at more localised levels. Trust issues, e.g. in the digital repository, how can the supplier prove this?

Lack of case studies to show to senior levels. People are needed not just the physical storage. Provide a cost-benefits case. JISC has started work in this area http://www.jisc.ac.uk/events/2011/03/jisc11/programme/1researchdata.aspx

Are a small number of specialists in a central storage facility the answer? Centralisation / decentralisation is an interesting issue.

Make sure we do not see RDM as an add on. RDM is core to delivering a high quality project. It's an integral part of the research, a particular activity, and responsibility has to be taken for it.

Relates to outsourcing. You can't outsource the knowledge and expertise in research and institutional responsibility to ensure well managed data can't be outsourced.

For PhD students these are core research skills. Where can they go outside the research domain to get these skills? Expertise and support is available, e.g. IT, Ethics Committees.

There is a whole discussion on vocabulary and this also needs to be put in place to ensure consistency and understanding.

Capacity is a big challenge; staff skills, embedding of skills. Development of tools and services. People with skills to use these and embed them. The challenge for the next 10 years.

DPC is a membership organisation. Activities are aimed at moving people on, helping them over the hump, sharing expertise and experience. Join the DPC!

Summary of tweets

Q: why more awareness of digpres in HE than in NHS? A: cultural? - HE prizes accumulated knowledge, NHS thinks 'now'.

[i] Jisc, Research data management training materials (RDMTrain) , http://www.jisc.ac.uk/whatwedo/programmes/mrd/rdmtrain.aspx