Exploring Research Data Management

Exploring Research Data Management

Andrew M. Cox and Eddy Verbaan

facet publishing

Published by Facet Publishing
7 Ridgmount Street, London WC1E 7AE
www.facetpublishing.co.uk

Facet Publishing is wholly owned by CILIP:
the Library and Information Association.

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library.

ISBN 978-1-78330-278-9 (paperback)
ISBN 978-1-78330-279-6 (hardback)
ISBN 978-1-78330-280-2 (e-book)

First published 2018

Text printed on FSC accredited material.

FSC
www.fsc.org
MIX
Paper from
responsible sources
FSC® C013604

Cover design by Kathryn Beecroft
Typeset from author's files in 11/14pt Revival 565 and Frutiger by Flagholme Publishing Services.
Printed and made in Great Britain by CPI Group (UK) Ltd, Croydon,
CR0 4YY.

Contents

List of tables and figures

Tables

Figures

CHAPTER 3

What are research data?

Aims

The aim of this chapter is to explore what research data are. It will help you start to have productive conversations with researchers about their data.

Research data are important to (some) researchers

For many researchers in the sciences and social sciences research data are of central importance to their work. Planning the collection of appropriate data is a key part of research design. The term 'data creation' may often be a more accurate term than 'data collection' or 'data capture', which imply that data are something existing before the researcher intervenes to actively construct them. But the language in use varies between fields of study and not all research projects actively create data. Creating data may take up many hours of work, and can be one of the most exciting parts of the research process, where the researcher gets into the laboratory or out into the field in the hope of finding something new. Then, processing the data and analysing them are central to creating new knowledge.

Skill and innovation in eliciting and then analysing data is central to one's success as a researcher. The researcher's deep relationship to data is strongly linked to their methodological commitments about how they believe science builds knowledge. A common understanding of methods is a central aspect of their subject discipline. Thus they have a strong investment in research data and a concern with their quality.

Actually, in many fields there is a kudos attached to collecting data oneself. It could be seen as a rite of passage for the novice researcher in some subjects. When talking about their work researchers often talk about 'my data . . . my stuff'. This points to the strong relationship between research, research data and identity. The conversation about research data is a deep one. We have even heard researchers talk about their data as their 'life's work'. Material they are gathering is part of building a legacy. Imagine the researcher who has pursued their interests over multiple projects throughout a long research career. To them they have an intimate connection to the various datasets that they have accumulated and pored over for many hours. Often it can be this, as much as pragmatic concerns such as fear of being beaten to publication, that inhibits research data sharing.

Having said all this, for many researchers data are essentially a means to an end: they are the foundation for gaining understanding of a phenomenon and then for publishing one's results. It is the understanding and the publications that matter more than the actual data.

Furthermore, some researchers would deny collecting 'data' at all. This might be because they see the term data as implying quantitative material such as survey data, when they deal with qualitative material such as interviews and observations. Or it may be that in their field one simply does not refer to evidence as 'data'. Thus, historians typically differentiate primary sources (original documents such as archival material) and secondary sources (interpreting the phenomenon that is studied, usually published works). Their primary sources are their data. So do not always start the conversation by talking about 'data'. If you do you run the risk of alienating humanities scholars.

Furthermore, some researchers genuinely don't collect data, e.g. in a purely theoretical field such as philosophy, arguably there are no data.

Talking to a researcher about the data they collect and analyse is a key conversation to have if you are working in the RDM field. But one has to be careful to use the terminology that researchers in that particular field relate to.

Exploring further

Start reading some papers produced by the researchers in the institution you work for, if you work for one, or an institution where you are studying. What are the data sources they are using? For example, in a social science or science paper the methodology section should describe in a fair amount of detail what the data were and how they were handled and analysed. The paper should reflect a particular methodological position.

Have a look at some research methods books for the same field. These also give you a sense of the typical sorts of research going on in that area and the data types in use.

Talk to a researcher about their work. Make a conscious effort to listen out for the terms they use to describe the research process and to categorise data. It may well be hard to understand the exact meaning of some of the measurements they make, if one does not have a related background. But one can begin to explore the issues that the researcher has about data quality.

Types of research data

Institutional surveys for RDM (see Chapter 9) often ask questions about research data such as how much data individuals have in gigabytes and what sort of data it is, e.g. whether it is in Word files, images, spreadsheets and so on.

Even if it is important to them and even if they do have lots of data, one should probably not rely too much on researchers' own estimates of the quantity of data they hold, or even the order of magnitude of data they have. Do you know how many megabytes of files you have on your work computer? Probably not; because there is no real need to know. When we ran an RDM survey at Sheffield in 2014 around a quarter answered 'don't know'. Many more who did answer may simply have been guessing.

Defining data by format, as in Table 3.1, may be useful for data management purposes, but it tells us little about what is in the document or spreadsheets.

Table 3.1 *Some formats of data*

Documents (Text, PDF, Microsoft Word)
Spreadsheets (for example: Microsoft Excel)
Websites
Notebooks/diaries
Databases (for example: Access, MySQL, Oracle)
Questionnaires, transcripts, codebooks
Audiotapes, videotapes
Film, photographs
Artefacts, slides, specimens, samples
Collection of digital objects acquired and generated during the process of research
Raw data files generated by software, sensors or instrument files
Models, algorithms, scripts
Contents of an application (input, output, log files for analysis software, simulation software, schemas)

Table 3.2 gives us an immediate sense of the range of types of data. Virtually anything could be data. It could be non-digital: it could be a material object or a completed printed questionnaire. If it is digital, it could be vast and complex; or small. Categories such as 'images' disguise the huge range of visual material used in research, from works of art and historical photos to satellite imagery and medical photography. One project might produce multiple forms of data.

Table 3.2 *Some types of data*

Results of experiments
Measurements collected in the field
Software programmes and their outputs
Interview audio recordings and transcripts
Focus group transcripts
Questionnaire responses
Government surveys
Images
Moving images
Historical documents
Physical objects
Social media data: tweets
Logs of web server traffic or another activity

From an RDM point of view this proliferation of data types is central to the challenge. For example, we may need to run a repository that handles at least part of this range of types of material. Inevitably the descriptive standards and documentation of data are also widely variable across subjects, and so similar types of data might be described in rather different ways.

Some definitions of research data

Read these definitions carefully, and consider their strengths and weaknesses:

> Factual records (numerical scores, textual records, images and sounds) used as primary sources for scientific research, and that are commonly accepted in the scientific community as necessary to validate research

findings. A research dataset constitutes a systematic, partial representation of the subject being investigated. (OECD, 2007, 14)

Data are facts, observations or experiences on which an argument or theory is constructed or tested. Data may be numerical, descriptive, aural or visual. Data may be raw, abstracted or analysed, experimental or observational. Data include but are not limited to: laboratory notebooks; field notebooks; primary research data (including research data in hardcopy or in computer readable form); questionnaires; audiotapes; videotapes; models; photographs; films; test responses. Research collections may include slides; artefacts; specimens; samples. (University College London, 2013)

Qualitative or quantitative statements or numbers that are (or assumed to be) factual. Data may be raw or primary data (e.g. direct from measurement), or derivative of primary data, but are not yet the product of analysis or interpretation other than calculation. (Royal Society, 2012, 12)

Research data are defined as recorded factual material commonly retained by and accepted in the scientific community as necessary to validate research findings; although the majority of such data is created in digital format, all research data are included irrespective of the format in which it is created. (EPSRC, n.d.)

The data, records, files or other evidence, irrespective of their content or form (e.g. in print, digital, physical or other forms), that comprise a research project's observations, findings or outcomes, including primary materials and analysed data. (Monash University, 2010)

Research data are the evidence that underpins the answer to the research question, and can be used to validate findings regardless of its form (e.g. print, digital, or physical). These might be quantitative information or qualitative statements collected by researchers in the course of their work by experimentation, observation, modelling, interview or other methods, or information derived from existing evidence. Data may be raw or primary (e.g. direct from measurement or collection) or derived from primary data for subsequent analysis or interpretation (e.g. cleaned up or as an extract from a larger dataset), or derived from existing sources where the rights may be held by others. Data may be defined as 'relational' or 'functional' components of

> research, thus signalling that their identification and value lies in whether and how researchers use them as evidence for claims. They may include, for example, statistics, collections of digital images, sound recordings, transcripts of interviews, survey data and fieldwork observations with appropriate annotations, an interpretation, an artwork, archives, found objects, published texts or a manuscript. The primary purpose of research data is to provide the information necessary to support or validate a research project's observations, findings or outputs. (Concordat on Open Research Data, 2016)

> The output from any systematic investigation involving a process of observation, experiment or the testing of a hypothesis, which when assembled in context and interpreted expertly will produce new knowledge. (Pryor, 2012, 3)

> Anything you perform analysis on. (Briney, 2015, 6)

There are many definitions of data. In policy documents it may be useful to try and define research data in a fairly formal way, but some definitions seem to work much better for certain disciplines or meta-disciplines than others. For example, the Royal Society and EPSRC definitions apply more for science subjects. The UCL definition is useful for making it clear to all sorts of researchers that that 'stuff' they are creating is indeed data. It is more of a definition through listing examples than by focusing on what 'data' is conceptually. Briney's (2015) definition has the value of being simple and direct. The Monash definition perhaps confounds data and the findings based on that data. Pryor's definition focuses on the systematic process and the purpose, creating new knowledge, though the range of methods feels a little narrow. The most comprehensive definition is from the Concordat. The start of the Concordat definition focuses on purpose. The purpose of research data is to provide an answer to research questions. The Concordat also usefully differentiates various states of data e.g. raw, primary, derived data.

Exploring further

Collect some more definitions of research data and analyse them. Look at relevant national or international policy statements and see how they seek to define data. If your institution has a research data policy, how is data defined there? From your conversations with researchers which is the most useful definition?

Data collections

Individual scholars or projects may produce data collections: coherent bodies of data that others might want to re-use.

Thinking about collections of research data might be a rather 'library' way of looking at data, as if it's a coherent body of material with clearly defined scope. This might not be quite how a researcher would see their 'stuff'. They are probably more likely not to have really thought of it as a coherent body of material, just something they use and that grew organically. Nevertheless it can be a useful perspective for thinking systematically about the scope and content of a body of research data.

Carlson has advocated a structured interview for capturing a profile of a research dataset (http://datacurationprofiles.org). The data curation profile technique constitutes a rather comprehensive and systematic approach to finding out all about the data produced in a project or series of projects. The structure is itself a very useful way of thinking about the different aspects of data, even if it is actually something smaller or less tidy than a 'collection'. Some of the headings include:

- overview of the research, including the topic and funding source
- data kinds and stages – in the form of a narrative about the data collection, and including a data table itemising data collected by size, format and number of objects
- intellectual property rights relating to the data
- organisation and description of the data – including metadata standards in use
- target repository
- sharing and access – who can use the data and on what basis, including any desire for an embargo
- discovery – including target audiences
- tools – tools used in the research that others may need to use to re-use the data
- measures of impact – what usage measures would be appropriate to this material
- data management – practical issues, including back-up and security
- preservation – which material should be preserved and for how long.

If you are thinking of talking to a researcher about their data this approach gives you a systematic way of thinking.

Exploring further

Read some of the data curation profiles http://datacurationprofiles.org/.

Try and relate the proposed structure and some of your early conversations with researchers. Some questions may not feel relevant to a particular area of research. Work on a set of questions you feel comfortable asking a researcher.

Look at some datasets in the local data repository, a subject data repository or a general one like Dryad or Figshare and examine some of the deposits and how they have been described.

Data lifecycles

As well as data being created within the research lifecycle, data could be considered to have their own lifecycle. Data tend to go through a process of creation, cleaning, combination, storage, analysis, and possibly then preservation, sharing and re-use.

The metaphor of a lifecycle, be that of life to death or life to rebirth, has a strong resonance in the world of RDM. It has always been central to archival and preservation work. For example, the UK Data Archive (www.ukdataservice.ac.uk/manage-data/lifecycle) proposes a simple model with six stages:

- Creating data – this stage involves such activities as planning data collection, locating existing data sources and the actual data collection tasks, including documenting the data. In research involving human subjects it is highly likely to include the important ethics clearance stage.
- Processing data – validating and cleaning data, prior to the serious business of analysis.
- Analysing data – this the stage at which data are analysed and includes publication.
- Preserving data – this is about getting them into the right format for preservation and documenting them.
- Giving access to data – this includes making data discoverable, setting up conditions of re-use and promoting such re-use.
- Re-using data – including follow-up research and others re-analysing the data.

These are more like logical steps than the ones we might observe in any actual project. By definition such lifecycles are a simplification of real life, which is far less linear and more iterative in practice. Some commentators talk about research workflows rather than lifecycles: but this may make the complex and contingent patterns of research sound a bit too much like a defined administrative flow of work. Having said that, looking for temporal patterns is likely to be rewarding. Also, this is a data perspective on research. Most researchers would be more preoccupied with gaining grants, outputs and publication, than the life of the data. Again, this realisation needs to be borne in mind when trying to use the model.

Another rather famous representation of research data is the DCC curation model. Again, this is more like the data curator's vision of the lifecycle of data, than something a researcher would relate to strongly.

Exploring further

If you can, ask a researcher about the detailed steps in their research process. This will help you get to grips with the life of data. You might want to take the approach used by Mattern et al. (2015), who asked researchers to produce a hand-drawn diagram of the research process, and then asked them to add to the picture notes about actions relating to data. It might be best to focus on a particular research project, because there may be differences across different projects. Capturing more about the flow of the research process can help you map out where support is needed or can be offered within the research process. Comparing diagrams produced by researchers in the same field will give you a fascinating insight into the commonality and variation within a single research area.

Mattern, E., Jeng, W., He, D., Lyon, L. and Brenner, A. (2015) Using Participatory Design and Visual Narrative Inquiry to Investigate Researchers' Data Challenges and Recommendations for Library Research Data Services, *Program*, **49** (4), 408–23, http://doi.org/10.1108/PROG-01-2015-0012.

Research data is complex

There have already been some hints that research data is not simple; this section further explores the complexity of research data.

Commentators often refer to the five Vs of big data:

1 Volume
2 Variety
3 Velocity
4 Veracity
5 Value.

These aspects can serve as headings for thinking about research data, too. We have already discussed that researchers may not have a precise idea of the *volume* of data they have collected. Much of the early discussion of RDM was linked to the concept of a 'data deluge', vast quantities of data being created in big science and challenging to store and document for re-use. Researchers working in astronomy, for example, might well be involved in work generating truly vast quantities of data. Not all research does have great volume, though. The title of Christine Borgman's (2015) book *Big Data, Little Data, No Data* neatly captures the fact that not all research data has huge 'volume'. But just because the amounts concerned are not vast does not mean that they are easy to manage.

We have also already discussed the *variety* of data. This applies within individual projects as well as between disciplines. In their study of data in the life sciences, Williams and Pryor (2009) mapped researchers using a complex array of data sources and different tools. A single study may draw in multiple forms of data that have to be managed collectively. Some are actively created for the project, other data are background or reference data that does not necessarily get cited. Each sub-field they studied was very different.

By *velocity*, commentators on big data are referring to the continuously updated streams of data that might be produced by such sources as sensors or internet traffic. Of course, this very same material could well become research data. Not all researchers will be trying to handle such dynamic data, but it is useful to think about most research data as dynamic. Data should not be seen as a thing, like a specific spreadsheet. Data are multiple and changeable. Thus, a researcher might gather some measurements in the field. Then they would enter this raw data into a spreadsheet back in the office. After some quality checking there might be a new version (derived data). Manipulating the data might lead to new versions. Combining this with other data would produce further versions. Tangible examples of such processes are described in the next chapter. But the key conclusion is that data changes. This lies behind Borgman's (2015) question: 'When is data?' Something plays a role of being data at a particular moment.

To take another example, a researcher might conduct interviews, making audio recordings of them, and taking down some field notes to be associated with each one. They might well then transcribe the interview. Transcripts can be imported into some qualitative data analysis software for coding. Later one might tidy up an anonymised version of the interview transcript for sharing. Thus there are likely to be multiple versions of each interview, not necessarily a single definitive one. This creates data management (such as file naming) issues.

Such patterns of change are suggested by the lifecycle model. Yet processes of creation may be far more contingent, iterative and non-linear than a lifecycle model implies. Research is often akin to a craft skill. It is often not based on formulaic following of a closely defined recipe.

Data are also mobile across contexts. Some of the fascinating research on data journeys explores how data change meaning across different domains (Bates, Lin and Goodale, 2016). For example, weather station measurements (temperature, rainfall, etc.) are created by local, often volunteer, work in the spirit of creating open data. Combined with other similar weather measurements a weather dataset becomes a valuable commercialised source for the weather forecasters. It may also turn up as a valuable asset in the futures market on the stock exchange. Thus the meaning and value attached to data can alter as it moves between contexts.

Veracity is about the reliability of data. For the researcher the reliability of their data, data quality, is always a key issue. This could be about calibrating instruments or checking for outliers in a dataset. For a linguist it could be about precisely transcribing pauses in the transcription of a conversation. So what defines quality will be different, but it will always be a concern for researchers.

Again, research data vary in terms of their *value*. Some data are irreplaceable. They're the measurement of a unique event. They cannot be recreated. In other cases the cost would prohibit collecting them again. On the other hand, data produced from a computer model or simulation (e.g. in engineering or economics) can easily be reproduced. It is the model that needs to be preserved.

It is also worth reminding ourselves that digital data are fragile. They lose meaning out of the context of their collection if not documented. A spreadsheet without a key to the headings is more or less useless. The archival concept of provenance captures the importance of knowing about how an information object was created. A researcher wanting to re-use data will want to understand critical features of how the data were generated to be sure they can use them with confidence. What those critical features are, however, may vary.

Exploring further

Read the RIN report on life sciences (Williams and Pryor, 2009).

Continue to talk to researchers about their work. Try and understand more about the whole process of data creation and analysis – in the wider context of the research lifecycle. A visual analysis like the ones in the RIN report could be useful in mapping out how data is assembled and used. You might also think in terms of drawing up your own lifecycle model to capture the dynamic changes in the research life course.

Williams, R. and Pryor, G. (2009) *Patterns of Information Use and Exchange: Case studies of researchers in the life sciences*, Report by the Research Information Network and the British Library, www.dcc.ac.uk/projects/life-science-case-studies.

Information management and RDM

These complex aspects of research data help us start to grasp the importance of the information management aspects of RDM.

- Sometimes the researcher is battling to find storage for the sheer volume of 'active' data they are working on right now. They may also be forced to make choices about which data are preserved, because they have so much.
- Researchers need to manage different types of data in a coherent way over the lifetime of the project.
- Researchers are having to manage a flow of data sources which are themselves changing and malleable. When we talk about data re-use we are thinking about a complex contextual transition that is challenging to manage.
- Data quality is key to the reliability of the outcome of the research, and so its credibility. But what is critical to quality will vary.
- Data has central value to much (though not all) research. Empirical research turns on the interpretation of data. It takes up the researcher's time to produce. It may also literally cost money to license or have potential monetary value.

Research does have a strong information management component.

Further reading

Carol Tenopir's series of studies about scientists' views on research data are a key reference point for the field.

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., Manoff, M. and Frame, M. (2011) Data Sharing by Scientists: Practices and perceptions, *PLOS ONE*, **6** (6), https://doi.org/10.1371/journal.pone.0021101.

Tenopir, C., Dalton, E. D., Allard, S., Frame, M., Pjesivac, I., Birch, B., Pollock, D. and Dorsett, K. (2015) Changes in Data Sharing and Data Re-use Practices and Perceptions Among Scientists Worldwide, *PLOS ONE*, **10** (8), https://doi.org/10.1371/journal.pone.0134826.

If you want to read more look for the latest years' citations of these works.

At a more conceptual level Shankar's (2007) paper describes the element of judgement and convention that go into turning scientific work into something labelled data. Recognising the element of judgement and choice in this process helps us understand the way that any data is a construction in a particular context, rather than an objective object.

Shankar, K. (2007) Order from Chaos: The poetics and pragmatics of scientific recordkeeping, *Journal of the Association for Information Science and Technology*, **58** (10), 1457–66.

References

Bates, J., Lin, Y.-W. and Goodale, P. (2016) Data Journeys: Capturing the socio-material constitution of data objects and flows, *Big Data & Society*, **3** (2), http://doi.org/10.1177/2053951716654502.

Borgman, C. L. (2015) *Big Data, Little Data, No Data: Scholarship in the networked world*, MIT Press.

Briney, K. (2015) *Data Management for Researchers: Organize, maintain and share your data for research success*, Pelagic Publishing.

Concordat on Open Research Data (2016) www.rcuk.ac.uk/documents/documents/concordatonopenresearchdata-pdf.

EPSRC (n.d.) EPSRC Policy Framework on Research Data: Scope and benefits, www.epsrc.ac.uk/about/standards/researchdata/scope/.

Mattern, E., Jeng, W., He, D., Lyon, L. and Brenner, A. (2015) Using Participatory Design and Visual Narrative Inquiry to Investigate Researchers' Data Challenges and Recommendations for Library Research Data Services, *Program*, **49** (4), 408–23, http://doi.org/10.1108/PROG-01-2015-0012.

Monash University (2010) *Research Data Policy*, http://policy.monash.edu.au/policy-bank/academic/research/research-data-

management-policy.html.
OECD (2007) OECD Principles and Guidelines for Access to Research Data from Public Funding, www.oecd.org/sti/sci-tech/oecdprinciplesandguidelinesforaccesstoresearchdatafrompublicfunding.htm.
Pryor, G. (ed.) (2012) *Managing Research Data*, Facet Publishing.
Royal Society (2012) *Science as an Open Enterprise*, https://royalsociety.org/topics-policy/projects/science-public-enterprise/report.
University College London (2013) *UCL Research Data Policy*, www.ucl.ac.uk/isd/services/research-it/documents/uclresearchdatapolicy.pdf.
Williams, R. and Pryor, G. (2009) *Patterns of Information Use and Exchange: Case studies of researchers in the life sciences*, Report by the Research Information Network and the British Library, www.dcc.ac.uk/projects/life-science-case-studies.