The International Journal of Digital Curation
Volume 8, Issue 1 | 2013

Research Data Management Education for Future Curators

Mark Scott, Richard Boardman, Philippa Reed and Simon Cox,
Faculty of Engineering and the Environment,
University of Southampton

Abstract

Science has progressed by "standing on the shoulders of giants" and for centuries research and knowledge have been shared through the publication and dissemination of books, papers and scholarly communications. Moving forward, much of our understanding builds on (large scale) datasets, which have been collected or generated as part of the scientific process of discovery. How will this be made available for future generations? How will we ensure that, once collected or generated, others can stand on the shoulders of the data we produce?

Educating students about the challenges and opportunities of data management is a key part of the solution and helps the researchers of the future to start to think about the problems early on in their careers. We have compiled a set of case studies to show the similarities and differences in data between disciplines, and produced a booklet for students containing the case studies and an introduction to the data lifecycle and other data management practices. This has already been used at the University of Southampton within the Faculty of Engineering and is now being adopted centrally for use in other faculties. In this paper, we will provide an overview of the case studies and the guide, and reflect on the reception the guide has had to date.

International Journal of Digital Curation (2013), *8*(1), 288–294. http://dx.doi.org/10.2218/ijdc.v8i1.261

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by UKOLN at the University of Bath and is a publication of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

Introduction

Hilbert and López (2011) estimated that the worldwide capacity for storing digital information in 2007 was 276 exabytes. A similar estimate from private sector research company International Data Corporation (IDC) put the figure at 264 exabytes and calculated that all the data created and replicated in the “digital universe” was 281 exabytes (Gantz et al., 2008). In fact, since 2007, more data is produced than can actually be stored, as much of the data is transient. Some data is deleted when it is no longer required, some might be transformed into other formats, and some might be processed and the raw data discarded.

As a large proportion of data is not kept, looking after your data is becoming much more important, and teaching the value of this to students early in their careers helps them to recognise its importance and start thinking about ways of managing data.

We looked at five researchers’ work from medicine, materials engineering, aerodynamics, chemistry and archaeology, and produced case studies showing the similarities and differences between the data types they produce. We created a guide containing the case studies and an introduction to research data management.

Student Guide

The case studies were written up into a glossy, introductory guide. The guide was broken down into the following three parts:

1. **Five Ways To Think About Research Data** – providing an introduction into data categorisation and data life cycles;
2. **Case Studies** from medicine, materials engineering, aerodynamics, chemistry and archaeology;
3. **Data Management Practices** – giving general tips on managing data.

Guide Part I: Five Ways To Think About Research Data

Combining some recognised definitions of research data, we introduced research data by showing the following five ways of considering it:

1. How research data is collected (Research Information Network, 2008)
2. The forms of research[1]
3. Electronic storage of the research data[1]
4. Electronic data volumes
5. Data life cycle

[1] Adapted from The University of Edinburgh: http://www.ed.ac.uk/schools-departments/information-services/services/research-support/data-library/research-data-mgmt/data-mgmt/research-data-definition

This helped to set the scene and introduce types of data and the research that data represented. The case studies in Part II were written using this framework.

Guide Part II: Case Studies

The aim of the case studies was to show the similarities and differences in research data between disciplines and how this was managed. Each case study was written using terminology from the discipline, whilst remaining accessible to non-experts.

Each case study begins with a table summarising the data categories used by the researcher, grouped using the framework introduced in the previous section. Each case study included a discussion of the researchers' practices when producing and using the data, broken down into three sections:

1. Obtaining the data
2. Using the data
3. Managing the data

These sections are taken from one of the data life cycles in section 1 of the guide, shown in Figure 1.

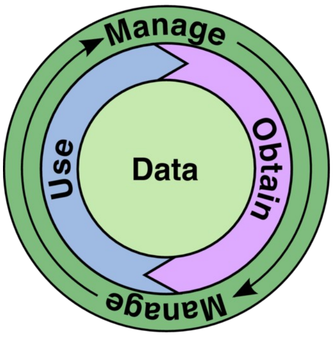
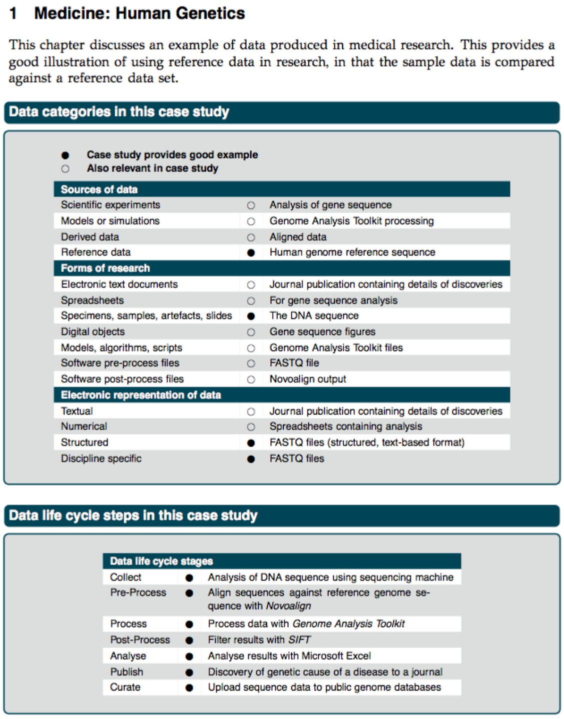


Figure 1. A simplified data life cycle, used as section headings for the case studies.

Finally, each researcher was asked to provide some images that showed their research or data in use.

The case study was formatted to fit on three to four pages. An example of one of the case studies is shown in Figures 2–4. As can be seen, the format of the guide was presented clearly and used colour extensively to make it more approachable and easier to digest.



1 Medicine: Human Genetics

This chapter discusses an example of data produced in medical research. This provides a good illustration of using reference data in research, in that the sample data is compared against a reference data set.

Data categories in this case study

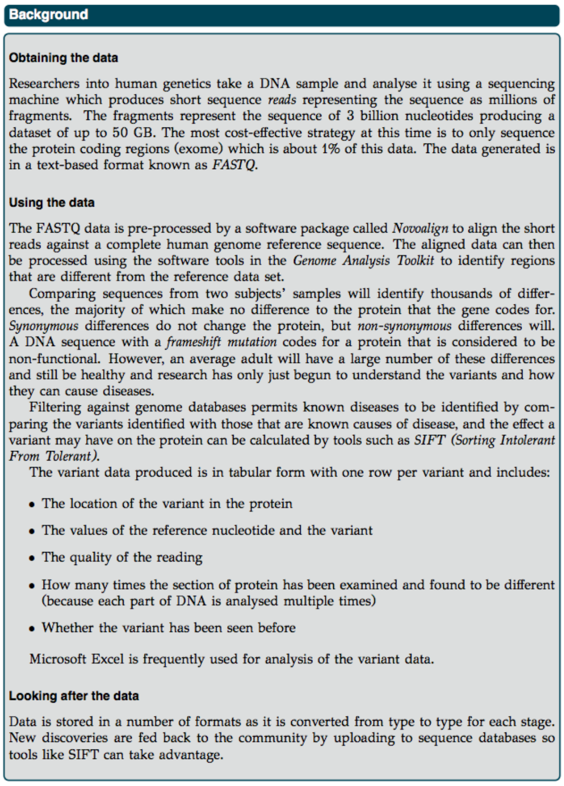
● Case study provides good example
○ Also relevant in case study

Sources of data		
Scientific experiments	○	Analysis of gene sequence
Models or simulations	○	Genome Analysis Toolkit processing
Derived data	○	Aligned data
Reference data	●	Human genome reference sequence
Forms of research		
Electronic text documents	○	Journal publication containing details of discoveries
Spreadsheets	○	For gene sequence analysis
Specimens, samples, artefacts, slides	●	The DNA sequence
Digital objects	○	Gene sequence figures
Models, algorithms, scripts	○	Genome Analysis Toolkit files
Software pre-process files	○	FASTQ file
Software post-process files	○	Novoalign output
Electronic representation of data		
Textual	○	Journal publication containing details of discoveries
Numerical	○	Spreadsheets containing analysis
Structured	●	FASTQ files (structured, text-based format)
Discipline specific	●	FASTQ files

Data life cycle steps in this case study

Data life cycle stages		
Collect	●	Analysis of DNA sequence using sequencing machine
Pre-Process	●	Align sequences against reference genome sequence with *Novoalign*
Process	●	Process data with *Genome Analysis Toolkit*
Post-Process	●	Filter results with *SIFT*
Analyse	●	Analyse results with Microsoft Excel
Publish	●	Discovery of genetic cause of a disease to a journal
Curate	●	Upload sequence data to public genome databases

Figure 2. Case study data usage summary.



Background

Obtaining the data

Researchers into human genetics take a DNA sample and analyse it using a sequencing machine which produces short sequence *reads* representing the sequence as millions of fragments. The fragments represent the sequence of 3 billion nucleotides producing a dataset of up to 50 GB. The most cost-effective strategy at this time is to only sequence the protein coding regions (exome) which is about 1% of this data. The data generated is in a text-based format known as *FASTQ*.

Using the data

The FASTQ data is pre-processed by a software package called *Novoalign* to align the short reads against a complete human genome reference sequence. The aligned data can then be processed using the software tools in the *Genome Analysis Toolkit* to identify regions that are different from the reference data set.

Comparing sequences from two subjects' samples will identify thousands of differences, the majority of which make no difference to the protein that the gene codes for. *Synonymous* differences do not change the protein, but *non-synonymous* differences will. A DNA sequence with a *frameshift mutation* codes for a protein that is considered to be non-functional. However, an average adult will have a large number of these differences and still be healthy and research has only just begun to understand the variants and how they can cause diseases.

Filtering against genome databases permits known diseases to be identified by comparing the variants identified with those that are known causes of disease, and the effect a variant may have on the protein can be calculated by tools such as *SIFT (Sorting Intolerant From Tolerant)*.

The variant data produced is in tabular form with one row per variant and includes:

- The location of the variant in the protein
- The values of the reference nucleotide and the variant
- The quality of the reading
- How many times the section of protein has been examined and found to be different (because each part of DNA is analysed multiple times)
- Whether the variant has been seen before

Microsoft Excel is frequently used for analysis of the variant data.

Looking after the data

Data is stored in a number of formats as it is converted from type to type for each stage. New discoveries are fed back to the community by uploading to sequence databases so tools like SIFT can take advantage.

Figure 3. Case study text description.

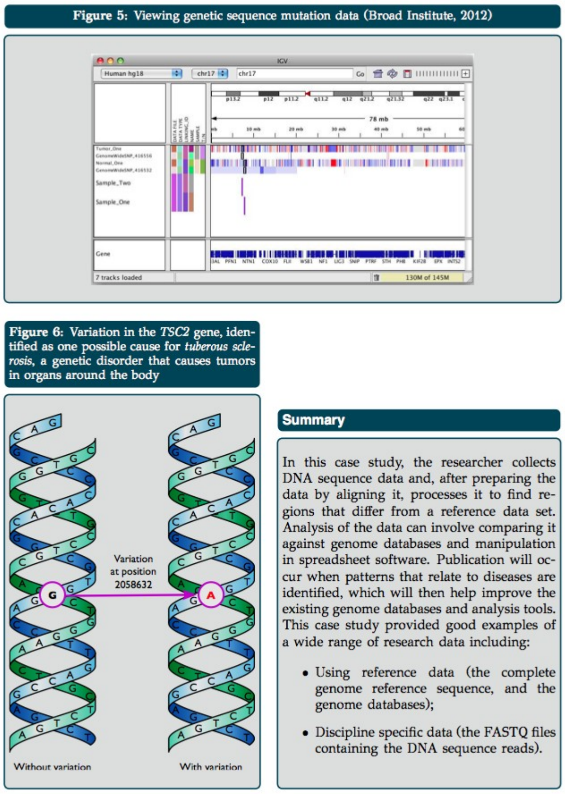


Figure 5: Viewing genetic sequence mutation data (Broad Institute, 2012)

Figure 6: Variation in the *TSC2* gene, identified as one possible cause for *tuberous sclerosis*, a genetic disorder that causes tumors in organs around the body

Summary

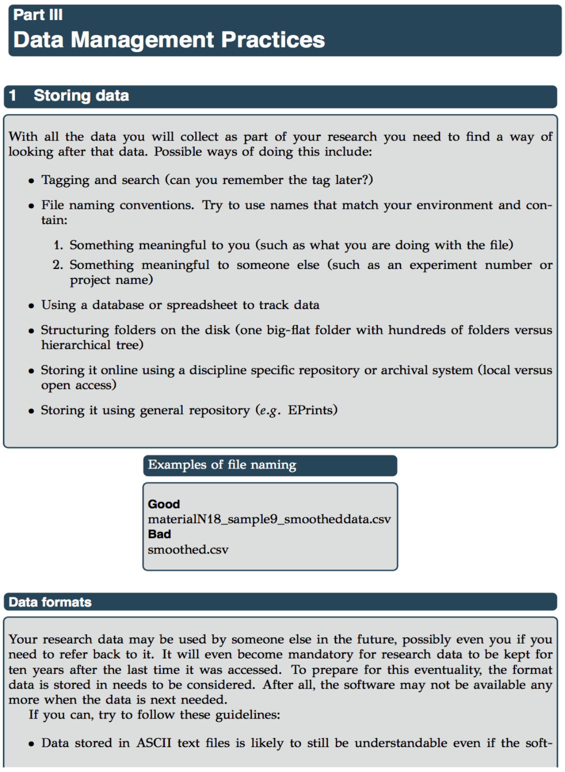
In this case study, the researcher collects DNA sequence data and, after preparing the data by aligning it, processes it to find regions that differ from a reference data set. Analysis of the data can involve comparing it against genome databases and manipulation in spreadsheet software. Publication will occur when patterns that relate to diseases are identified, which will then help improve the existing genome databases and analysis tools. This case study provided good examples of a wide range of research data including:

- Using reference data (the complete genome reference sequence, and the genome databases);
- Discipline specific data (the FASTQ files containing the DNA sequence reads).

Figure 4. Case study figures and summary.

Guide Part III: Data Management Practices

The final part of the guide provided general advice on how to manage data. Topics include file naming, data preservation, file tracking, file formats, backups and file versioning. Figure 5 shows an example of how this was presented.



Part III
Data Management Practices

1 Storing data

With all the data you will collect as part of your research you need to find a way of looking after that data. Possible ways of doing this include:

- Tagging and search (can you remember the tag later?)
- File naming conventions. Try to use names that match your environment and contain:
  1. Something meaningful to you (such as what you are doing with the file)
  2. Something meaningful to someone else (such as an experiment number or project name)
- Using a database or spreadsheet to track data
- Structuring folders on the disk (one big-flat folder with hundreds of folders versus hierarchical tree)
- Storing it online using a discipline specific repository or archival system (local versus open access)
- Storing it using general repository (*e.g.* EPrints)

Examples of file naming

Good
materialN18_sample9_smootheddata.csv
Bad
smoothed.csv

Data formats

Your research data may be used by someone else in the future, possibly even you if you need to refer back to it. It will even become mandatory for research data to be kept for ten years after the last time it was accessed. To prepare for this eventuality, the format data is stored in needs to be considered. After all, the software may not be available any more when the data is next needed.

If you can, try to follow these guidelines:

- Data stored in ASCII text files is likely to still be understandable even if the soft-

Figure 5. An extract from Part III, giving advice on data management best practices.

Lessons Learned

The guide has been presented to students twice as part of a training lecture given to first year postgraduates. To gauge its reception, students were asked three questions at the end of the lecture, as shown in Table 1.

Question	Lecture 1	Lecture 2
	80 students	30 students
	Month 7	Month 2
Would students change what they were doing as a result of the lecture?	60/80	15/30
Should the talk be given to future postgraduates?	78/80	20/30
Did any students feel it was a complete waste of time?	2/80	1/30

Table 1. Feedback from research data management introduction lectures.

A few students did not feel the guide was relevant to their research but the majority believed it was useful and should be given to first year postgraduates in future. Interestingly, the feedback was more positive when the lecture was given in month seven, once postgraduates had settled in. This is perhaps because they had begun their research and had started to collect data.

The guide itself proved very popular and an additional print run was required due to unexpected demand.

This was our first attempt at this type of guide. It was felt that future versions should extend the third section (Data Management Best Practices) to provide more tips on data management, as students found this part the most useful as reference material and reported that it helped them to think about how to look after their own data. The case studies in Part II provided perspective and encouraged students to think about data in general. Part I (Five ways to think about research data) is useful, but should be kept brief in a lecture to maintain the students' interest.

Conclusions

Research data management has become a requirement for all researchers. Ensuring students start considering these problems early in their careers should help them as the amount of data used in research continues to rise. Producing a guide on research data for students and training them on the issues of data management is one approach. The feedback received from students suggests that thinking about these issues partway through Year 1 of their research study is helpful and necessary.

Teaching about the management of research data is now part of the University's data management strategy.[2] In the wider context of the University, we have a data management plan to cover research data coming from science and engineering projects.

Acknowledgements

The following people helped with the preparation of this document:

- Andy Collins (Human Genetics case study).
- Thomas Mbuya and Kath Soady (Materials Fatigue Test case study).
- Gregory Jasion (CFD case study).
- Simon Coles (Chemistry case study).
- Graeme Earl (Archaeology case study).
- Mark Scott, Richard Boardman, Philippa Reed and Simon Cox (overall content).

We acknowledge ongoing support from the University of Southampton, Roberts' funding (ECR training), Microsoft, EPSRC, BBSRC, JISC, AHRC and MRC.

[2] See the University of Southampton's Research Data Management web site for further details: http://www.southampton.ac.uk/library/research/researchdata/

References

Gantz, J. F., Chute, C., Manfrediz, A., Minton, S., Reinsel, D., Schlichting, W. & Toncheva, A. (2008). *The diverse and exploding digital universe: An updated forecast of worldwide information growth through 2011*. IDC white paper sponsored by EMC. Retrieved from http://www.emc.com/leadership/programs/digital-universe.htm

Hilbert, M. & Lopez, P. (2011). The world's technological capacity to store, communicate, and compute information. *Science*, 332(6025), 60–65. doi:10.1126/science.1200970

Research Information Network (2008). *Stewardship of digital research data: A framework of principles and guidelines.* Author. Retrieved from http://www.rin.ac.uk/system/files/attachments/Stewardship-data-guidelines.pdf