REPORT

The University of Nottingham

UNITED KINGDOM · CHINA · MALAYSIA

Data asset audit and the management of research data – a pilot study for ADMIRe and the School of Community Health Sciences

Author(s):	Laurian Williamson, Thomas Parsons, Jonathan Lamley, Julia Hippisley-Cox & David Williams.
Audience:	Project stakeholders & JISC MRD
Published:	05/06/2013

Contents

1. Introduction

The research community is now generating more data than ever before and the "data deluge" is likely to increase in the future and is creating significant research data management (RDM) challenges for institutions and subject disciplinary communities.

The drivers for RDM come from the research funders, legislation, the open data agenda, and UK HEIs wanting to have a better oversight of their research outputs, and increasingly of the research data produced within their institution.

It is important to note that whatever the drivers, good research data management is good for research: better management of research process, potentially more productive research, avoidance of data loss, benefits of data reuse.

In their revised policy on access to research outputs published in July 2012 the Research Councils UK (RCUK) have stated that peer reviewed research papers which result from research that is wholly or partially funded by the Research Councils must:

"Include details of the funding that supported the research, and a statement on how the underlying research materials – such as data, samples or models – can be accessed."

Compliance with the RCUK policy is expected from the 01st April 2013. The RCUK have not yet provided an indication of how compliance will be monitored but there are data expectations that need to be taken into account. Any 'non-compliance' may have a significant impact on the funding received by UoN from the RCUK and on the university's substantial research portfolio.

Researchers are faced with national mandates to both manage and share their research data. Most research funders require applicants to submit a statement on data management and sharing at the grant proposal stage. The coverage of this data management statement does vary by funder, but it is worth noting that the biomedical funders (BBSRC, Cancer Research UK, MRC, and the Wellcome Trust) focus heavily on data sharing.

2. Research data management drivers

From an institutional perspective, there is a clear need for better, centralised systems that can ensure that the research data produced in the University is known and managed in an infrastructure that meets the requirements of the major funding bodies and the needs of the researchers.

The University of Nottingham Research Data Management Policy[1] was approved by University Management Board in February 2013. It is acknowledged that it is an 'aspirational' policy and that implementation will take some years.

[1] http://www.nottingham.ac.uk/research/research-data-management/creating-data/policies.aspx

Some of the key strategic institutional challenges around managing research data include governance, data security, technical infrastructure, and ethical and risk-management issues.

Data security may be needed to protect intellectual property rights, commercial interests, or to keep personal or sensitive information safe.

The University's Internal Audit Service undertook an audit of Information Security in Community Health Sciences (CHS) which reported in August 2011. CHS volunteered for this audit and the audit investigated the types of information obtained, used and stored by the School, focusing on research data. One of the identified areas of possible risk was around sensitive and personal data.

In addition, the University Management Board (MB) considered papers on RDM submitted by Professor Saul Tendler in February 2013. One of the requirements endorsed by MB is as follows:

"A requirement that Schools and Departments create an inventory of confidential or highly confidential data which they create, process or store as part of their research activities and review the data security measures in place to protect these from inappropriate access, data loss or data breach."

In addition to the university-wide activities around research data security and data classification, it was felt that it would be useful and beneficial for both ADMIRe and CHS to 'pilot' a data inventory tool and do some requirements gathering on sensitive data within CHS. It was agreed that raising awareness of RDM practice would be useful for the School research staff, whether they were working on a funded or a non-funded internal project. In particular, the RDM issues around the management of sensitive and personal data were considered to be of great importance and the Data Asset Framework (DAF) was proposed as a potential candidate.

3. The Data Asset Framework

The DAF provides organisations with the means to identify, locate, describe, and assess how they are managing their research data assets. Frequently, institutions are unaware as to what data they hold and how (if at all) the data is being managed. DAF is beneficial for the institution and individual researcher or research groups, allowing them to quickly establish an overview of their data holdings - data has potentially huge value in improving access to it and its reuse. Raising awareness of data collections and monitoring data holdings is a valuable asset for an institution; this is one of the key benefits of auditing research data holdings. It is important to note that DAF can be applied to research data assets only, it is not deemed suitable for research publications/articles which were excluded from this pilot.

The DCC and the University of Glasgow Humanities Advanced Technology and Information Institute HATII[2] developed the DAF methodology as part of a JISC-funded project. All outputs, including methodology, implementation guide, pilot projects and their findings and lessons learned, presentations and core documents, and an online tool, can all be found on the DAF Website[3].

The DAF methodology has been implemented in several UK HEIs, for example, Glasgow, Edinburgh, UKOLN, King's, Imperial, UCL, Oxford, and Southampton. Lessons learned from the 4 DAF audit pilots[4] (School of GeoSciences, Edinburgh; the Innovative Design and Manufacturing Research Centre (IdMRC) at the University of Bath; Glasgow University Archeological Research Division (GUARD); Centre for Computing in the Humanities at King's College London) were reviewed and taken into account.

The four steps DAF recommends for effective data management are as follows:

Step One: Planning the audit (appointing and auditor; establishing a business case for senior management buy-in; desk-based research/planning).

Step Two: Identifying and classifying assets (types of data; producing a structured list of assets; surveys/interviews, doing an inventory and classification).

Step Three: Assessing the management of data assets (finalize and complete documentation on every on every data asset held, and document minimum information (core elements) required for each asset).

Step Four: Reporting results and recommendations (creating a final audit report and outline recommendations for on-going research data management).

Rather than seek to carry out a full data audit within the timescales of ADMIRe, it was proposed that a DAF template could be created using Word or Excel and the questions and usefulness of this would be assessed during the pilot. Therefore the end result would be a subject-specific version of DAF in a questionnaire format, for later incorporation into which ever research data catalogue system Nottingham eventually deploys.

3.1. Piloting DAF

After initial project discussions a questionnaire was created using the DAF methodology. And for the purposes of this quick data asset audit, we defined research data as: *"That which underpins a research assertion".*

[2] http://www.gla.ac.uk/subjects/informationstudies/
[3] http://www.data-audit.eu/
[4] http://www.data-audit.eu/docs/DAF_lessons_learned.pdf

Even though both staff from ADMIRe and CHS viewed DAF as a useful audit tool it was felt that for the purposes of the 'pilot' it would be better to simplify the data audit template, which currently has more than 50 elements. By minimising the required elements it was felt that researchers were more likely to engage with this audit. Hence after initial feedback 14 key elements were condensed and reviewed with CHS (see Appendix 5.1). As well as helping to identify research data, it was also suggested researchers may find the information they provide here useful when reviewing their own data management plans (DMPs).

The draft data audit questionnaire, which only includes mandatory core elements was piloted and revised following further discussions. The revised questionnaire briefly covers issues such as provenance, ownership, location, management, archiving and preservation (see Appendix 0).

3.2. Refinement

After an initial review period in March 2013, feedback from CHS indicated that further refinement was needed and this resulted in a version tailored to clinical data. The original purpose of the pilot was to assess DAF as a means of identifying sensitive data in CHS, so adopting a subject-specific approach was a welcome bonus.

A number of suggestions were made, which although may go beyond the scope of the initial pilot are still valid requirements:

1. It was suggested this the audit was best done via a web site data entry form ideally with as many closed questions as possible (i.e. pick options).
2. Each dataset should be given a unique ID
3. The data holder should update the register with any changes or on an annual basis.
4. There should be an option to identify data which are licensed to the university e.g. CPRD[5].
5. There should be a question to ascertain whether there are any external audit rights for the licensing organisation and if so, what these are and the associated terms etc.

Clinicians from CHS also noted that in the Caldicott Report published on 26th April 2013[6] data can be classified in three ways:

1. Identifiable data
2. Anonymised data
3. Pseudonymised data or de-identified data for limited access

This was viewed as a helpful classification because the issues that need to be documented vary according to these groups.

[5] http://www.cprd.com/intro.asp
[6] https://www.gov.uk/government/publications/the-information-governance-review

3.2.1. Identifiable data

Any research data that contains identifiable data should have the following recorded at a minimum:

- Who is the data controller?
- Who is the data processor?
- The source of original data
- Which identifiers are held? (i.e. name, address, postcode, phone number, email, NHS number, hospital number, etc)
- What clinical or other data is held?
- What the lawful basis for holding the data are.

The final question then splits to gives rise to three options, namely:

1. Subject consent
2. Section 251 approval (S251)[7]
3. Not known

If identifiable data is catalogued that has neither subject consent nor S251 approval, then there would need to be an urgent investigation to ensure the legality of the collection.

Further information would include the dates and scope of the S251 support documentation. An audit of any S251 records would then require verification of the S251 application form and approval letter.

For consented data, then a copy of the patient consent form is required.

The data holder must also be able to demonstrate compliance with the 8 principles of the Data Protection Act[8] (DPA) so this should also be documented. It is important to note that a key consideration is how the DPA requirement for fair processing and subject access requests are met.

3.2.2. Anonymised data

Anonymised data can potentially be published, as patients cannot be identified by this by definition and therefore does represent a risk, if anonymisation is insufficient. If such data is identified, then it is recommended that the holder should refer to the ICO code of anonymisation[9] and the HSCIC code of confidentiality when published (expected June 2013)[10].

3.2.3. Pseudonymised or de-identified data for limited access

---

[7] http://www.hra.nhs.uk/hra-confidentiality-advisory-group/what-is-section-251/

[8] http://www.ico.org.uk/for_organisations/data_protection/the_guide/the_principles

[9] http://www.ico.org.uk/for_organisations/data_protection/topic_guides/anonymisation

[10] http://www.hscic.gov.uk/dataflowstransitionmanual

The final category is where the data per se are not identifiable in the context that the data are held (i.e. within a strong governance framework). Patients are assigned a pseudonym and all their data is attributed to this, rather than their original identifying data.

The risk which has to be mitigated is the risk of patient re-identification should the data be linked with other data sources or information. This would then render the data as identifiable data and a lawful basis for that must be identified as above.

A final version of the DAF questionnaire was designed to incorporate the suggestions raised and is available within Appendix 5.3.

4. Next steps

As of June 2013, the revised DAF questionnaire is undergoing testing within CHS and it is expected that feedback will be provided after the end of the ADMIRe project. This raises a number of outstanding actions that should be taken forward as part of a Nottingham RDM service:

1. Convert the template into a web based format with drop-down lists as appropriate
2. Consider creating a template within the proposed Nottingham Data Catalogue to collect sensitive data
3. Store the metadata as records within the proposed Nottingham Data Catalogue

It is expected that these actions will be carried out by a Nottingham RDM service, but at present, the details of this how this will occur are not known. What is clear, is that there is a growing need for audits such as these from both a legal and research data management perspective – hence every effort should be made to ensure tools such as these, are available to researchers who work with sensitive or medical data.

5. Appendix

5.1. Draft version 1 data audit questions

No	Parameter	Comment
1	Data creators (s)	*Person, group or organization responsible for the intellectual content of the data asset*
2	Title	*Title of the data asset*
3	Description	*A description of the information contained within the data asset*
4	Keywords	*Relevant keywords that describe the data asset*
5	Creation/start date	*Date when the data asset was created/started*
6	Completion date	*Date when the data asset was completed/ data collection ceased (if data is no longer being added)*
7	Original Purpose	*Description of what was the main reason for the data asset's creation*
8	Updating frequency	*Estimated frequency for updating the data asset*
9	Format	*Text, numerical, models, software, multimedia, discipline specific, instrument specific, or any other*
10	FoI, data protection, personal privacy issues	*Description of any potential data protection or ethical issues related to content of the data asset and if any restrictions based on these are currently applied*
11	Access	*Describe who can typically access the data asset*
12	Back-up and storage	*Number of copies of the data asset that are currently stored, frequency of back-up and storage location*
13	Usage constraints	*Access restrictions applied to the data asset*
14	Research data management to date	*History of maintenance and integrity of the data asset*

5.2. Draft version 2 data audit questions

No	Parameter	Comment
1	Data creators (s) / PI	*Person, group or organization responsible for the intellectual content of the data asset*
2	Title of the research activity/project	*Title of the data asset*
3	Project code and funder	*UoN project code and main funder of the research; also includes internally funded projects*
4	Description	*A description of the information contained within the data asset*
5	Keywords	*Relevant keywords that describe the data asset*
6	Creation/start date	*Date when the data asset was created/started*
7	Completion date	*Date when the data asset was completed/ data collection ceased (if data is no longer being added)*
8	Original Purpose	*Description of what was the main reason for the data asset's creation*
9	Updating frequency	*Estimated frequency for updating the data asset*
10	Format	*Text, numerical, models, software, multimedia, discipline specific, instrument specific, or any other*
11	FoI, data protection, personal privacy issues, anonymisation	*Description of any potential data protection or ethical issues related to content of the data asset and if any restrictions based on these are currently applied. Indicate if anonymisation is required.*
12	Access	*Describe who can typically access the data asset*
13	Back-up, storage and encryption	*Number of copies of the data asset that are currently stored, frequency of back-up; storage location and if the data asset is encrypted*
14	Archiving	*Description of any archiving activities planned or applied to the data asset*
15	Preservation	*Description of any preservation activities planned or applied to the data asset*
16	Usage constraints	*Access restrictions applied to the data asset*
17	Research data management to date	*History of maintenance and integrity of the data asset*

5.3. Final data audit questions

No	Parameter	Comment
1	Data creators (s) / PI	*Person, group or organization responsible for the intellectual content of the data asset*
2	Title of the research activity/project	*Title of the data asset*
3	Project code and funder	*UoN project code and main funder of the research; also includes internally funded projects*
4	Data set ID	*A unique ID for the data set*
5	Description	*A description of the information contained within the data asset*
6	Keywords	*Relevant keywords that describe the data asset*
7	Licence details	*Relevant licences if this is a commercial or external dataset*
8	Audit rights for the data	*Details of external rights to audit this data and associated terms*
9	Creation/start date	*Date when the data asset was created/started*
10	Completion date	*Date when the data asset was completed/ data collection ceased (if data is no longer being added)*
11	Original Purpose	*Description of what was the main reason for the data asset's creation*
12	Updating frequency	*Estimated frequency for updating the data asset*
13	Format	*Text, numerical, models, software, multimedia, discipline specific, instrument specific, or any other*
14	FoI, data protection, personal privacy issues, anonymisation	*Description of any potential data protection or ethical issues related to content of the data asset and if any restrictions based on these are currently applied. Indicate if anonymisation is required.*
15	Identifiable data	*Does the data contain identifiable data? If yes, then please complete questions 16-23*
16	Who is the data controller?	*Please provide details*
17	Who is the data processor?	*Please provide details*
18	The source of original data	*Please provide details*
19	Which identifiers are held? (i.e. name, address, postcode, phone number, email, NHS number, hospital number, etc)	*Please provide all relevant details*
20	What clinical or other data is held?	*Please provide details*
21	Are there records of subject consent?	*For consented data, then a copy of the patient consent form is required.*
22	Is there Section 251 approval (S251)?	*Further information would include the dates and scope of the S251 support documentation. An audit of any S251 records would then require verification of the S251 application form and approval letter.*
23	Has the Data Protection Act been considered?	*The holder must be able to demonstrate compliance with the 8 principles of the Data Protection Act (DPA).How DPA requirement for fair processing and subject access requests should also be documented.*

24	Anonymised data	*Does the data contain* anonymised *data? If yes, then please describe and complete questions 25-27*
25	Is the data sufficiently anonymised?	*The holder should refer to the ICO code of anonymisation and the HSCIC code of confidentiality when published (expected June 2013) .*
26	Pseudonymised or de-identified data	*Does the data contain* pseudonymised *data* or de-identified data*? If yes, then please describe and complete question 27.*
27	How have the risks of subject or patient re-identification been mitigated?	*Patient re-identification can occur should the data be linked with other data sources or information. If not, then this should be treated as identifiable data.*
28	Access	*Describe who can typically access the data asset*
29	Back-up, storage and encryption	*Number of copies of the data asset that are currently stored, frequency of back-up; storage location and if the data asset is encrypted*
30	Archiving	*Description of any archiving activities planned or applied to the data asset*
31	Preservation	*Description of any preservation activities planned or applied to the data asset*
32	Usage constraints	*Access restrictions applied to the data asset*
33	Research data management to date	*History of maintenance and integrity of the data asset*
34	Review date	*An annual review of the information is advisable*
35	Significant changes	*A description of changes to the data or information i.e. dataset deleted or location changed*