Research Data Discovery Manual

Contents

Background

In early 2014, The Centre for Hip Health and Mobility (CHHM) has partnered with the University of British Columbia (UBC) Library and UBC IT to establish and implement best practices in research data management. This manual outlines the visual discovery approach that was prepared to assess the research data management needs of researchers' in this interdisciplinary and multimodal bone and joint research centre.

Considering the entire lifecycle of data from inception to close of a research project and beyond, different approaches to determine how researchers' data management needs could be captured while identifying areas for immediate solution implementation and requirements for mid- to long-term solutions were investigated. The tool kit for Data Curation Profiles developed by Purdue University Libraries was considered. However, while comprehensive, its use of lengthy questionnaires and the estimated 15 hours required per project assessment suggested it being impractical for use in CHHM.

Pilot sessions were conducted for data curation profiling, confirming the impractical aspects of the Purdue tool kit and identifying possibilities for adaptation. It became clear that a visual and interactive approach for the discovery of research data management needs presents a viable solution. Different types of questions and topics contained within the Purdue tool kit and other sources, such as the UBC Research IT Support question catalogue, were extracted. The need for representing three types of objects, data collection instruments, processing steps, and outcomes, became apparent. These objects could then be modularly used as the three visual components with which to capture data flows and to effectively identify opportunities to improve data management processes. It was then possible to identify common workflows and needs across projects. This in turn led to determining the main data streams present at CHHM, including main challenges and needs related to them.

This manual outlines the steps for the visual approach developed. It was use to assess 11 research projects focusing on human subjects. 25 researchers and research operations team members were involved in this assessment. The steps outlined and their activities are seen to be recommendations. They should be adjusted according to the characteristics of the research environment.

Step 1 - Initial Contact with Research Groups

Whether the responsible person or team is from inside, e.g. research support function of a research centre, or outside, e.g. central research support function of a university, a research organization establishing initial contact with research groups will be necessary. While this can be done in different ways according to the circumstances, a written summary containing the following should be provided to the research groups' lead researcher:

- Aim of the data management assessment;
- Why their research project has been selected;
- The process of how the assessment is conducted;
- What kind of questions they can expect;
- A request for documents;
- A request for times they are available in selected weeks;
- And an invitation to invite queries;

Below an example of such a summary:

Dear <Name>,

As outlined in <a previous announcement/meeting/communication>, we are initiating a project to assess and address the data management needs at the <centre/department/institute>. We hope to address the following issues, among others:

- Finding information after a graduate student leaves and a grant ends.
- Finding data after publishing it, and comparing data with colleagues who are doing similar work.
- Ensuring that data and code are "frozen" after publication, in some cases to ensure accurate comparisons later.
- Interacting with researchers from other institutions by exchanging files or using software to collaboratively work on analysing data.

In order to develop solutions to address your needs and priorities, it is important for us to understand how your research group works with data.
We will be working to provide you with solutions that provide a value add in short order, as well as robust and scalable solutions that provide long-term benefits.

We would like to conduct an interactive discovery session with you. The aim of this discovery session, taking approximately 45 to 60 minutes, is to gain an overview of the role data plays in your major research project(s). The discovery session outcome will provide a visual representation of the data flow. Please find attached an example. The image file demonstrates how the data flow is being developed. The pdf file represents a concise summary of all the aspects we would like to look into. There are certainly other aspects that we can include as we move forward and refine our approach based on your feedback.

Your research group has been selected for the initial assessments due to <reason>.

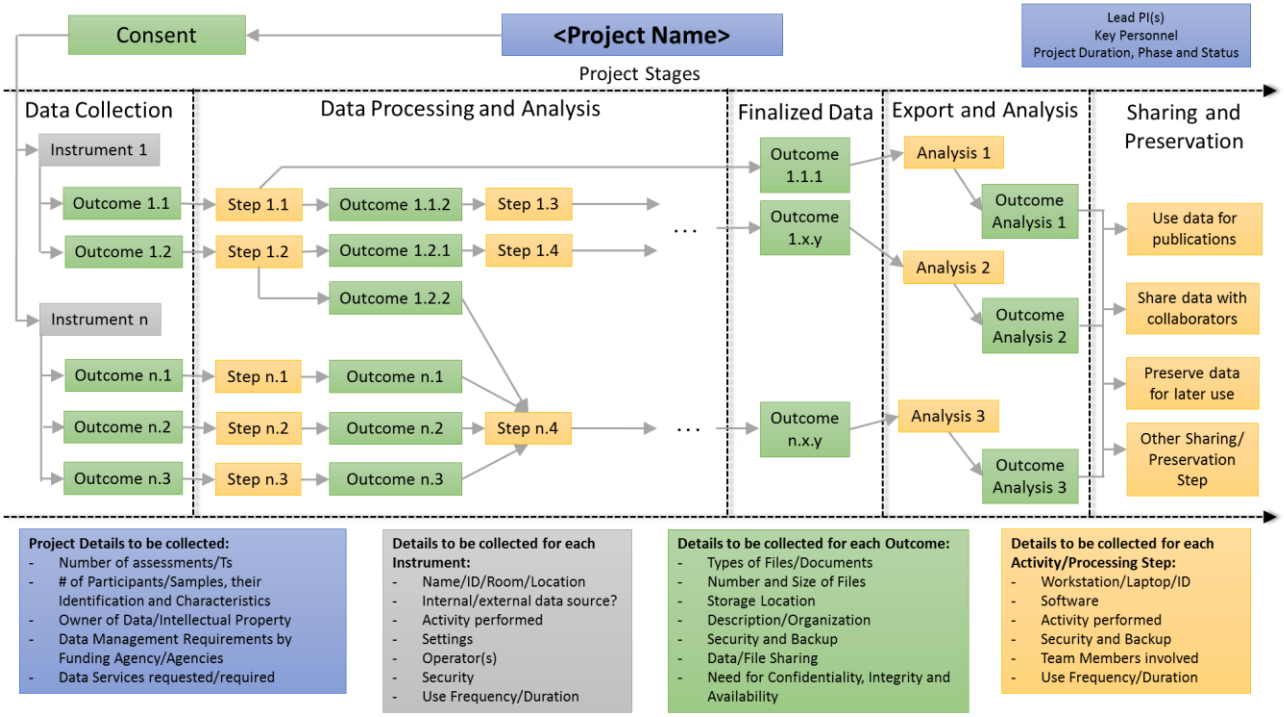
Key questions we will exploring to discover your needs are:
- What happens to the data and how would you characterize the data at each stage of the research process?
- What do you consider barriers and enablers for your work with research data?
- What kind of services, software, and hardware are you using or would like to use in terms of data processing, accessing, storing, organizing, and sharing?

Prior to meeting with your team we would like to prepare by reviewing material related to your projects. If possible could you please send me the grant proposal, a copy of the UBC IRB (ethics) application, and a copy of all documents submitted as part of the UBC IRB application for the research projects. Publications resulting from the projects and any other documents you think might be helpful.

We would also like to set up a schedule to meet with you and members of your team in <month/week>. Can you identify who the key team members are that we should meet with? Should we meet with them as a group or on an individual basis? Would you want to be part for all or some of these meetings? Should we meet with you first or gather data from the team and then meet with you? During which timeframe would you and your team members be available?

Let us know if you have any questions.

Example of Flow Chart – Whiteboard



Example of Flow Chart

Step 2 - Preparation

1) Upon receipt of project documents, review them to establish a basic understanding of the aims and needs of the project, as well as the data flow.
2) Create the start of a flow chart to be used in the discovery session.
3) Create sticky notes for the identified equipment/instruments being used for this project. (if not already created)
4) Schedule meeting (can be done in parallel with 1, 2 and 3).

Step 3 - Just before discovery session

1) Print: Start of data flow for research project discussed in session
2) Print*: Sheet with session questions - see appendix A
3) Print*: Data classification definitions - see appendix B
4) Print*: Equipment lists (if available and needed)
5) Print*: Paper templates for session participants to draw on (see templates folder)
6) Make sure markers in different colors are available and are working.
7) If required: Make sure the audio recording device is fully charged and has sufficient storage space left.
8) Bring laptop, and audio recording device if required

*If no more printouts are available.

Step 4 - At the beginning of session

1) Introduce the assessment process, elaborate on the big picture and hand out the example data flow and paper templates.
2) Invite queries.
3) If required: Ask whether it is ok to record the session with an audio recording device.
4) Briefly summarize your understanding of the research project, and ask for confirmation that it is correct.
5) Go to whiteboard and start drawing the start of the data flow while explaining what you are doing. (Important: Make no assumptions about anything!)
6) Once finished with drawing ask whether this understanding is correct and ask someone to come up and make corrections if necessary.
7) Point at a particular object (e.g. a piece of equipment/instrument) and ask what is happening to the data related to this object.
8) If no one steps up to the whiteboard, encourage this activity.

Step 5 - During the main part of the session

1) Ask applicable questions from question sheet.
2) Highlight any spots for which clarification is need and take a note about it.
3) Take notes about details that cannot be covered or it does not seem possible at this moment to visualise them.
4) Pay attention to any challenges or needs mentioned and add them to the flow chart as well.

Step 6 - At the end of the session

1) Review the flow chart for any obviously missing items.
2) Summarize what has been accomplished and any open items.
3) Ask whom to contact for sending a draft of the data flow chart and the open questions.
4) If in use: Stop the audio recording.
5) Take images of the flow chart.
6) Copy images to laptop and check if they are readable.
7) Clean the whiteboard.

Step 7 - After the session

1) Use the images taken to recreate the data flow in PowerPoint or a similar software. Highlight the spots that require missing information.
2) Compile a list of the missing information.
3) Send the draft and the list of missing information to the main contact and ask for a review of the draft and for determining the missing information.
4) If applicable: Also ask whether a second meeting might be useful to clarify any misinterpretations in the draft or related to missing or complex information.
5) Wait for response. If none received after a week, determine if there is any event/activity that might result in this. If not, contact researchers.
6) While waiting, generate (summary) annotations for images and audio (if applicable), and store them in a secure, but accessible place.

Step 8 - After all information are available

1) Generate a final draft of the data flow diagram.
2) Obtain confirmation from the research group that it is accurate.
3) Send final version to research group and PI. If applicable: Also send list of high-level project questions (See appendix B) to PI or meet with PI to discuss them.
4) Review the data flow for needs to be addressed. Classify the needs according to urgency and importance. Also determine if a need could be addressed in the short-term by solutions already available or need to be considered in the long-term.

Appendix A – List of Questions/Aspects to Determine

Summary of aspects to determine

Project Details to be collected:
- Project Name
- Lead PI(s) and Key Personnel
- Project Stage
- Project Duration
- # of Participants/Samples, their Identification and Characteristics
- Owner of Data/Intellectual Property
- Data Management Requirements by Funding Agency/Agencies
- Data Services requested/required

Details to be collected for each Activity/Processing Step:
- Workstation/Laptop/ID
- Software
- Activity performed
- Security and Backup
- Team Members involved
- Use Frequency/Duration

Details to be collected for each Instrument:
- Name/ID/Room/Location
- Internal/external data source?
- Activity performed
- Settings
- Operator(s)
- Security
- Use Frequency/Duration

Details to be collected for each Outcome:
- Types of Files/Documents
- Number and Size of Files
- Storage Location
- Description/Organization
- Security and Backup
- Data/File Sharing
- Need for Confidentiality, Integrity and Availability

Project Details

- Title, Abbreviation, Number and characteristics of participants/samples
- Lead PI(s)
- Key personnel
- Project status and phase
- ID for participants/samples and if the ID is cryptic then what its individual parts represent (How does anonymization work?)
- Determine how to classify the data according to the classification system selected.
- Who is the owner of the data?
- What are the funding source(s) for this research?
- Do any of the funding source(s) require that:
    - A data management plan is drafted as a condition of funding?
    - Data is shared with others, published, or deposited into a data repository?
    - Data is preserves beyond the life of funding?
    - Data set is bound by any privacy or confidentiality concerns?
- In the journals, or places the researcher(s) publish most often, are data, or other supplemental information, accepted for publication? (If yes, which places of publication?)
- Would researcher(s) be interested in?
    - The ability to see the usage statistics on how many people have accessed this data.
- The ability to gather information about who has accessed or made use of the data.If needed, classify the overall project data according to the classification criteria outlined in appendix B.

Consent

- How is consent recorded?
- Are participants asked whether they are willing to participate in other studies?
- What kind of files/medium is it stored in?
- How are the consent forms stored and secured?
- Is there any backup?

Instrument

- What is the exact name of the equipment/instrument, its ID, and its location?
- What is being done with this equipment/instrument?
- What settings are used in the process of performing this activity?
- Who is responsible for operating the equipment/instrument? (title/name)
- What security measures are in place for this piece of equipment/instrument?
- How much time does it take to perform the activity using this equipment/instrument and how often is this activity performed?
- Are any external data sources used for data collection?

Outcomes

- What kind of files are generated?
- How big are the files and how many are there?
- How is the data stored and secured? Is there a backup?
- How is the data organized? Are there any descriptions or metadata, such as a code book?
- Is this data set bound by any privacy or confidentiality concerns?
- Is this data being shared? If so, with whom?
- Who do you think would be interested in this data?
- Would you want to share the data with others (with collaborators, others in the Centre, others within affiliated institutions, others in your field, others outside your field, anyone)?
- Under which conditions would you share the data?
- Would you be willing to submit this data to a repository or similar central storage space, and if so under which conditions? What preparations would be needed for the data to be transferred into a repository (e.g. remove identifiers, annotate it)?
- Would you like to have one of these services? (If so, ask about priority for the service.)
    - The ability to cite this dataset in my publications.
    - A requirement that others cite my data set if they want to use it for research.
    - The ability to access a backup copy of the data set.
    - The ability to restrict access to the data set to authorized individuals.
- Classify the data according to the classification criteria outlined in appendix B.

Processing Steps

- What workstation, software, and tools are being used in this stage? (if possible determine exact name or identifier and version of item, respectively)
- What is being done with this workstation, software, or tool? Where is the data generated, where is it processed, where is it stored?
- How is the use of the workstation, software, or tool secured? Is there any backup for them?
- Who is responsible for performing this processing step? (title/name)
- How much time does it take to perform the activity and how often is this activity performed?

Appendix B – Data Classification Definitions

Confidentiality-Integrity-Availability Data Classification

Adapted from: http://security.utexas.edu/policies/data_classification.html

Data can be classified according to Confidentiality, Integrity, and Availability. The combined classification of these three helps to determine how important managing the classified data is. Most of the legal and regulatory requirements are driven by confidentiality and integrity concerns, while availability mostly relates to the degree of accessibility to the data. Definitions for the three terms.

- Confidentiality: The need to strictly limit access to data to protect the university and individuals from loss.
- Integrity: Data must be accurate, and users must be able to trust its accuracy.
- Availability: Data must be accessible to authorized persons, entities, or devices.

The combined classification results in generating 3 classes of data with Class 1 being the data that needs to be prioritised, Class 2 being the data that should be addressed, and Class 3 being the data has a low priority.

	Class I	**Class II**	**Class III**
Need for Confidentiality	Required (High)	Recommended (Medium)	Optional (Low)
	AND/OR	AND/OR	AND/OR
Need for Integrity	Required (High)	Recommended (Medium)	Optional (Low)
	AND/OR	AND/OR	AND/OR
Need for Availability	Required (High)	Recommended (Medium)	Optional (Low)

Data Classification According to the C-I-A Classification

Appraisal and Selection of Research Data for Curation

Adapted from: http://www.dcc.ac.uk/resources/how-guides/appraise-select-data

Beyond considering confidentiality, integrity, and availability, these criteria should be considered when determining the importance of which data to prioritise:

- **Relevance to Mission:** The resource content fits the centre's remit and any priorities stated in the research institution or funding body's current strategy, including any legal requirement to retain the data beyond its immediate use.
- **Scientific or Historical Value:** Is the data scientifically, socially, or culturally significant? Assessing this involves inferring anticipated future use, from evidence of current research and educational value.
- **Uniqueness:** The extent to which the resource is the only or most complete source of the information that can be derived from it, and whether it is at risk of loss if not accepted, or may be preserved elsewhere.
- **Potential for Redistribution:** The reliability, integrity, and usability of the data files may be determined; these are received in formats that meet designated technical criteria; and Intellectual Property or human subjects issues are addressed.
- **Non-Replicability:** It would not be feasible to replicate the data/resource or doing so would not be financially viable.
- **Economic Case:** Costs may be estimated for managing and preserving the resource, and are justifiable when assessed against evidence of potential future benefits; funding has been secured where appropriate.

- **Full Documentation:** the information necessary to facilitate future discovery, access, and reuse is comprehensive and correct; including metadata on the resource's provenance and the context of its creation and use.