Opening doors to data

Grant Gibson

CRDCN and McMaster University

Date: 2025-04-24

Introduction

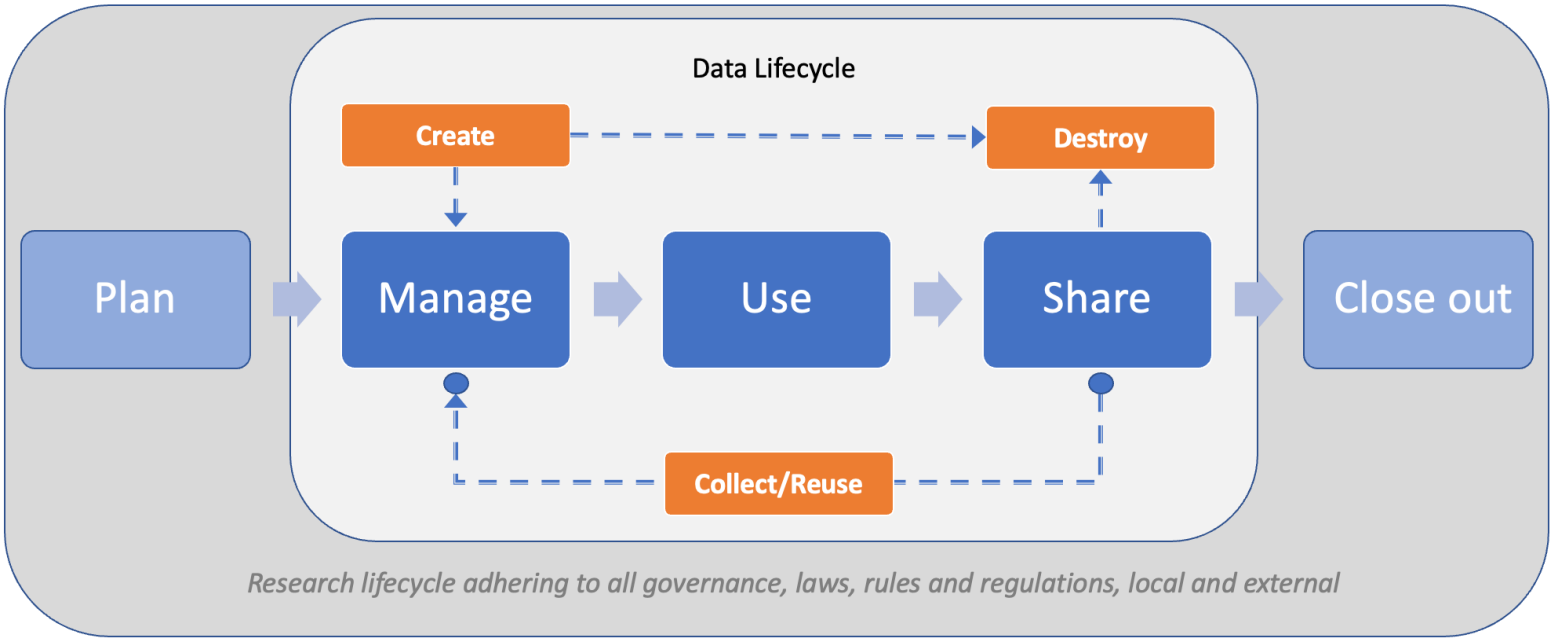
= Evolution of data discovery & open science - Journal pressures - Council pressures - University pressures

Containing this

- Today we're talking about restricted access data
- Not all sensitive data, and not open data, but data that have an access process or mechanism

Data lifecycle

- image of data lifecycle & description



Targets

- FAIR data
- “As open as possible, as closed as necessary”
- This is not binary, but a continuum

Researcher support for discoverability

- FAIR data F - Findable Can others even figure out where the data you used are? Concepts:
- Persistent Identifiers
- Indexing

Researcher support for discoverability

- FAIR data A - Accessible Can others access the data you used? Can they figure out HOW to do so? Concepts:
- Data accessibility statement
- Access metadata
- Transparent process

Researcher support for discoverability

- FAIR data I - Interoperable Concepts:
- Machine-readable
- Metadata
- Control vocabularies

Researcher support for discoverability

- FAIR data R - Reuseable Concepts:
- Provenance
- Licensing
- Archiving

Case study: CBS

- Canadian Blood Services offers secondary use research data (that is databases about donors that you can request)
- Suppose you were a researcher interested in research on data about donors

1. Think about a research question and evaluate whether you could answer it
2. Outline the process you would follow to get the data

- We'll take up these questions and discuss how "FAIR" we think the data are.

Takeup (GA)

FAIR Restricted data

- Findability doesn't **need** to be affected, but often is
- Accessibility might mean something different
- Interoperability can sometimes be tricky
- Reusability can be better, but this requires effort

Under the F

- Organizations providing data as a non-priority activity
- Resources to make it findable may not even be considered
- Knowledge of how to make data discoverable might not exist
- Even where a core dataset from an academic source exists there can be
  a. A research team primarily using the data and making it available is secondary (see item 1)
  b. Because it isn't open it's not posted and making it findable is a separate activity (where with open data it's often put into a service that manages both curation and discoverability)

Accessibility

- Is it even considered?
  - I would argue yes, very seriously, but not with a lens of FAIR
  - Some of this is foundational see Read et al. 2024

Interoperability

- This goes hand-in-hand with findability and suffers the same resourcing issue
- Metadata can have the potential to disclose individuals
- Restricted datasets don't generally have good metadata

Reusability

- Data often belong to an organization and so don't suffer the same risks that data held by individuals do
- Similar to Interoperability/Findability issues though, improper data management makes data less reusable.

What would success look like? (GA)

What gets us there?

- Data management planning
- Consider what anyone following on from you will be starting from. Everything that got you from that step to another point becomes part of your research data.
- Not every part of the project data will therefore be restricted.

What gets us there?

- Data accessibility statements
- Data discovery efforts
- Preservation efforts (both yours and your data source's) are important

Data Accessibility Statements

- Thorough description of the access process, links to info, financing considerations, and licensing info/terms-of-use.
- "Access available on request" NOT sufficient
- Encourage data source to template this language! If no, DIY with review.

Data discovery efforts

- Any datasource can now be indexed in Lunaris relatively easily.
- Metadata only deposits where there is some info to provide to potential users.
- Metadata only also creates a PID (persistent identifier) which makes the data a lot easier to find for someone who wants to use it in the future because YOU CAN CITE IT

Preservation efforts

- If a datasource overwrites my data with a new copy and doesn't say anything then someone trying to recreate my work will be very confused at best
- Ideally new versions get new PIDs and older versions point to the newest.
- Frequency of versioning will depend on frequency of access and preferences of organizations.

Test

- Test