How to create a data inventory

Article *in* F1000 Research · September 2018

DOI: 10.21955/gatesopenres.1114885.1

Tim Beale
Centre for Agricultural Bioscience International
8 PUBLICATIONS **31** CITATIONS

OPEN FOR FEEDBACK

WORK IN PROGRESS

How to create a data inventory

Open Data Institute

Contents

WORK IN PROGRESS

This is work in progress. It is likely to be updated as we continue our work. Keep an eye out for updates!

OPEN FOR FEEDBACK

How can it be improved? We welcome suggestions from the community in the comments.

Introduction

A data inventory is a list of datasets with metadata that describes their contents, source, licensing and other useful information.

A data inventory can be a useful tool for any organisation or project dealing with multiple types and sources of data. Creating a data inventory is also an important part of creating a data management plan for a research project.

An annotated list of datasets can help you to effectively locate, manage, use and share data. The context it provides can help users understand why data has been collected, what it contains, how it is managed and the ways it will be made available for others to use.

When published under an open licence, a data inventory can help people find and use the data they need.

This short guide provides

- a summary of the benefits of creating a data inventory
- an outline of the steps to create a data inventory
- some recommendations about what metadata to include in an inventory

Why create a data inventory?

A data inventory can be useful whenever it is helpful to be able to browse and compare a list of datasets. For example, a data inventory can help to:

- **Improve data discovery** – to understand the extent of the data that you organisation manages, uses or publishes. Publishing a data inventory under an open licence can help others to find, access and use the datasets that your organisation may be able to share or publish under an open licence. A data portal is an example of a public data inventory. An inventory might also be compiled to provide a list of datasets that are useful to tackle a particular problem or challenge.

- **Improve data governance –** the process of compiling and managing a data inventory can help you to take stock of the data that your organisation is managing. Creating an inventory is often the first step in improving your data governance. The inventory can help to identify duplicates, be used to improve best practices, and ensure that there are clear roles and responsibilities associated with managing data as an asset

- **Inform decision making around data management –** to understand the status of your data. A data inventory can help to prioritise resources, e.g. to improve data quality, rationalise technical platforms used to manage and publish data, or avoid duplication in collecting or purchasing data that is already available

- **Inform product development –** an inventory can help to understand which data sources are already available within an organisation. An inventory could help to identify which datasets could be published for others to reuse, or which could be used to develop new products or services.

- **Create a legal record** – an inventory can provide a legal record of the data that an organisation manages. You may have to do this for compliance reasons, such as maintaining a data asset register for the recently introduced General Data Protection Regulation (GDPR), or to maintain a list of third-party datasets your organisation accesses and the licensing and data sharing agreements which govern their use.

In short, a data inventory can help to provide useful information on the location, quality, technical and legal frameworks that will inform how data is managed, used and shared.

Creating and maintaining a data inventory is an important step in ensuring that data is treated as an asset, so that it is used and shared in ways that will help to maximise its value.

Publishing data inventories under an open licence can help other data users find data that is available to them to use.

A public inventory can also be used to create transparency around the data your organisation or project collects, either directly or through third-parties. This can help to build trust around how your organisation uses data.

Steps to building a data inventory

1) Plan the inventory

Decide on the purpose, scope and granularity of your data inventory.

- **Consider the inventory's purpose(s).** This will help define what metadata attributes you want to use for collecting information. For example, if you want to improve data management, you'll need to collect enough technical information to be able to understand how data is stored and formatted. If you want to identify business opportunities, you'll need to know about your rights to access and reuse the data, the costs of maintaining the data and its value. For compliance reasons you may need to identify which datasets contain sensitive personal information, etc.

- **Consider the inventory's sustainability.** Data inventories are most valuable if kept up to date. Think about how you can ensure this by adding inventory updates to your existing workflows (eg when a project is closed, the project manager or data steward should register new or updated information to the inventory). You should also consider where it will be hosted, how people will access it and who will update it.

- **Outline your definition of 'data' and decide what to include.** There are many ways of defining 'data' and so it's important to start with a common understanding of the scope of your inventory. For example, do you want to collect information on physical or paper-based assets? Is it for all information assets (eg reports) or just 'datasets'. Do you want to include data that was produced by third parties but forms part of your workflows? The following dataset definition may be a useful starting point:

> “*A dataset is a collection of data that relates to a common topic or was curated for a common purpose. A dataset has a consistent standard in terms of its format and structure. A dataset can contain ‘raw data’, analysed results or derived information.*

- **Decide what level of detail you need.** Do you want to record information about each single dataset, or a higher-level view of dataset collections grouped by subject, timeframe or creator? It is better to collect detailed information and then summarise the information for different audiences than have to return to add detail later.

2) Decide on the information you want to collect

Having considered your inventory’s purpose, decide on a set of attributes that you want to describe your datasets. We recommend the following attributes as a starting point:

Attribute	**Description**
ID	Unique identifier for the dataset
Title	The name of the data asset
Description	A description of the data asset
Purpose	Why was the data collected or produced?
Data creator	Who created the data?
Data manager/owner	Who manages the data?
Subject/keywords	What subjects/topics does this dataset cover? This will help discovery for users searching for this data. It is recommended to use a controlled vocabulary for this attribute (and others where possible) to improve future search and data linking potential eg finding related datasets
Location	Where is the data located or stored?
Creation date	When was the data created?
Update frequency	How often is the data updated?
Type	What type of data is it? Text, numbers, statistics, images, a database?
Format	What format is the data in? Eg MS Excel, CSV, JPEG, SQL DB
Rights and restrictions	What are the access and usage rights and restrictions? If you are publishing the data, what can users do with the data? Include a link to the relevant licence for use of the data (eg Creative Commons or a bespoke licence)

You can refine the above suggestions to include more attributes based on your needs. For example, you might want to record the spatial distribution of each dataset (e.g. which geographic areas it covers), or which time periods it covers. Other considerations are whether the dataset contains personal information, or contains commercially sensitive information.

See the links below for examples of other inventory templates and the Data Catalog Vocabulary for examples of other attributes you might want to collect.

3) Populate the inventory

Depending on the volume of data you'll be adding to the inventory, and the availability of existing sources of information, you might need to use some of following techniques to help you collect the data for your inventory

- **Delegate to existing dataset owners**
  - Identify dataset owners and ask them to complete the metadata directly
  - Build on information collected through other activities, audits, and IT documentation

- **Conduct interviews with data owners and product managers**
  - Individual or group interviews help to understand context by talking to the people who create and use the datasets
  - Collect all detailed information required and extra information such as opinions of the data assets and their potential use

- **Conduct surveys**
  - Electronic or web-based questionnaires help to reach many respondents at once. But sometimes questionnaires have the disadvantage of achieving low rates of return
  - Include commentary and guidance with questionnaires so information is collected following the same criteria

- **Create automated processes**
  - E.g. introduce a mechanisms so each time a document/ dataset/ database is added to a content management system, ensure that the user is prompted to complete basic metadata that is automatically added to the inventory. This is the most sustainable option in the long term

For anything other than a one-off audit, you will also need to plan for how to keep the data inventory up to date. Ensuring that the data inventory is embedded into internal data governance and management processes is one way to do this.

4) Publishing the inventory

The final stage is to collate findings and publish your inventory so that it can be used by others.

For smaller projects, the inventory could be stored in a spreadsheet that is periodically updated.

For larger quantities of information, a database might be more suitable. If the aim of the inventory is to improve discovery and data sharing, you might use the inventory as the basis of an online catalogue, which could allow users to search, browse and link to different data.

The inventory can also provide transparency about what data your organisation is collecting and using in its products and services.

Publishing an inventory does not imply that all of the datasets it contains must necessarily be available to others.

Data should be as open as possible, while still protecting people's privacy. Some data can be published as open data, for anyone to access, use and share. Other data might only be available via a data sharing agreement. Publishing an inventory can help users find the data they need.

Additional resources

UK Data Service – data inventories for research centres: Data Inventory Template

Data Asset Framework – a set of methods to identify, locate, describe and assess how organisations are managing research data assets: Data Asset Framework Methodology

Dublin Core – a set of vocabulary terms used to describe digital resources: DCMI Metadata Terms

What is a dataset? – a blog post reviewing dataset definitions