

Aristotle User Guide: How to record data in a metadata registry

User story: I have one or more files of data that I want to record in my metadata registry, so I can inform people of what data I have and what it contains.

My data is stored outside of the registry, and I want to let people know this data exists, but I may not want to give others access to the data. Instead, I want to store details about my data in Aristotle such as the name, description, and what fields are available. I want to tell people who is authorised to access the data, or how to request access to the data. I also want to record what format the data is stored in, such as a spreadsheet, database, or an API, so others know how to read the data.

What metadata objects should I use: Datasets and Distributions

Data files are recorded in the Aristotle Metadata Registry using the **Dataset** and **Distribution** metadata types. Other metadata types can be linked to **Datasets** or **Distributions** to group metadata together or link data with information about how it was collected or what reports were generated from the data.

- A **Distribution** represents a specific available form of a dataset. Each dataset might be available in different forms, and these forms might represent different formats of the dataset or different endpoints. Examples of distributions include a downloadable CSV file, an API, or an RSS feed.
- A **Dataset** is a collection of Distributions of data, published or curated by a single person or group and is available for access or download in one or more formats.

One way to understand the difference between a data file and **Dataset** or **Distribution** in an Aristotle Metadata Registry is by using the analogy of a listing for a book in an online store like Amazon.com:

When browsing Amazon.com for a book, a product page tells us that the book exists *somewhere* and tells us information about the book such as its author, number of pages or its price. The product page tells us all the editions of a book that are available, such as in softcover, hardcover, an online copy or as an audio book – even though these are all different formats, they have similar information. A book may be out of print or not available for purchasing, so just because the page for the book is accessible it does not mean the actual book can be bought. Even if someone can see the product page and buy the book, it does not mean they are allowed to walk into the warehouse to get it – they have to tell the store where to ship the book instead.

Most importantly, a product page for a book in an online store is not the book itself.

Within the Aristotle Metadata Registry, a **Dataset** is like the product page for your data – it records that the data exists, including a name and description of the data, and may include information about if the information can be accessed and how people can request the data.

Each **Distribution** is a record for different formats of your data – as each copy of the data can have slightly different information, a **Distribution** records the different ways data in a **Dataset** is stored. a name for each version, as well the format of the data, date this copy was made, and a list of records within the data file.

For both a **Dataset** and a **Distribution**, just because a user can see the “product page” for your data, does not mean they will be able to access your data. What the metadata does do is helps others know what data is available, where it is stored and understand what information it contains.

Examples of Datasets and Distributions

Below are some examples of how data can be registered using **Datasets** and **Distributions** in the Aristotle Metadata Registry. Follow any of the links in the descriptions to see a complete example on the [Aristotle.Cloud](#) registry.

Using Datasets and Distributions to record data from different areas

This example of a **Dataset** is a “National Patient Experience Dataset” that records patient data about a health system. In this **Dataset**, each **Distribution** records data for a different state or region. This allows any changes in the data between states to be recorded with the data.

The screenshot shows the Aristotle Metadata Registry interface. At the top, there is a navigation bar with a search box containing 'Metadata', a search icon, and links for 'Browse', 'Help', 'Log in', and 'Sign Up'. Below the navigation bar, the breadcrumb trail reads 'Aristotle Fundamentals / Metadata / Data Set / 2019 National Patient Experience Dataset'. The main heading is '2019 National Patient Experience Dataset'. There are tabs for 'Item', 'History', 'Graphs', and 'Related', and a 'Download' button. A yellow warning box states: 'This item is under active development. Content may change frequently.' The 'Definition' section includes a description of the dataset and a link to the 'National Patient Experience Dataset Specification'. The 'Data source details' section is a table with the following information:

Release Date	1st June 2020
Landing page	http://data.example.com/patient-experience
Frequency	Annually. Conducted between February 1 and March 31 each year.
Spatial dimension	Australia
Temporal dimension	2019-01-01 to 2019-12-31

The 'Distributions' section lists one distribution:

- 2019 NSW National Patient Experience Data
NSW National Patient Experience Dataset data collection.
Format type: CSV

On the right side, there is a metadata box for the '2019 National Patient Experience Dataset' with the following details:

Type:	Data Set
Identifiers:	UUID - e5235a28-aa38-11ea-b901-0af53f8eba84
Stewardship Organisation:	Aristotle Fundamentals
Endorsed by:	<ul style="list-style-type: none">Not endorsed
Published to:	<ul style="list-style-type: none">Public from 10th June 2020
Last updated:	12th June 2020
Created:	9th June 2020
Supersedes:	<ul style="list-style-type: none">None
Superseded by:	<ul style="list-style-type: none">None
Collections:	<ul style="list-style-type: none">None

View the example 2019 National Patient Experience Dataset online at:

<https://aristotle.cloud/item/604112/dataset/2019-national-patient-experience-dataset>

Using Datasets and Distributions to record data from different time periods
A **Dataset** can be used to record data collected over different time periods, by using separate **Distributions** to record data captured at different time periods.

This example, the Longitudinal Employment Study **Dataset**, records employment data collected over a ten-year period. In this **Dataset**, each **Distribution** records a data file for each year so that changes to data across time can be recorded.

This example also shows how the formats used to store the data have change over time, so users can understand how the data can be access and analysed.

Aristotle Fundamentals / Metadata / Data Set / 2009-2019 Longitudinal Employment Study Data

2009-2019 Longitudinal Employment Study Data

Item History Graphs Related Download

Warning: This item is under active development. Content may change frequently.

Definition

The 2009-2019 Longitudinal Employment Study was a decade long study on employment in the population, funded by the Federal Department of Labour.

Survey respondents were kept in the Employment study for 3 annual waves to determine long-term impacts of federal employment policy.

Data source details

Release Date	9th April 2020
Updated Date	30th April 2020
Frequency	Annually.
Temporal dimension	2009-01-01 to 2019-12-31

Distributions

- 2019 Longitudinal Employment Study collection (v v0.1)
Format type: SQL

2009-2019 Longitudinal Employment Study Data

Type: Data Set
Identifiers: UUID - b105dab8-aa42-11ea-a8c4-0212b740f8f0
Stewardship Organisation: Aristotle Fundamentals
Endorsed by:

- Not endorsed

Published to:

- Public from 10th June 2020

Last updated: 12th June 2020
Created: 9th June 2020
Supersedes:

- None

Superseded by:

- None

Collections:

- None

View the example 2009-2019 Longitudinal Employment Study Data record online at:
<https://aristotle.cloud/item/604118/dataset/2009-2019-longitudinal-employment-study-data>

Using a Distribution to record standalone data within a data registry

A **Distribution** can be a standalone datafile in a single format not associated with a Dataset.

This example **Distribution** is a single datafile about student demographic data. Where only a single version of a data collection is available, a single distribution makes it easier to create and updated records.

This **Distribution** also includes links to **Data Elements** that describe what each column in the data means. This example shows how the “dob” column records a Persons date of birth, and corresponds to the [Person-date of birth, DDMMYYYY](#) Data Element. This also lets users of the data know that the date of birth is stored in Day-Month-Year format, which is important when analysing the data.

By recording the fields and linking them to **Data Elements** in the metadata registry, it’s easier to understand how the data is stored and find related data – without having access to the data directly.

The screenshot shows the Aristotle Fundamentals metadata registry interface. At the top, there is a navigation bar with a search box, a 'Metadata' dropdown, and links for 'Browse', 'Help', 'Log in', and 'Sign Up'. Below the navigation bar, the breadcrumb trail reads 'Aristotle Fundamentals / Metadata / Distribution / 2019 Student Census Datafile'. The main content area is titled '2019 Student Census Datafile' and includes tabs for 'Item', 'History', and 'Graphs', along with a 'Download' button. A yellow warning box states: 'This item is under active development. Content may change frequently.' The 'Definition' section explains that the 2019 Student Census was conducted by all students at nationally funded universities to gather key statistics on tertiary students. The 'Data source details' section lists various attributes: 'Contained in dataset' (not linked), 'Release Date' (5th January 2020), 'Updated Date' (no updated date), 'Format type' (unknown), and 'License' (restricted access). A metadata box on the right provides details for the '2019 Student Census Datafile' distribution, including its type, UUID, stewardship organization (Aristotle Fundamentals), and publication dates. At the bottom, a table lists the data elements included in the distribution:

Path name	Data Element	Definition	Data Type	Format
dob	Person—date of birth, DDMMYYYY	The date of birth of the person.	Date/Time	DDMMYYYY
hours_worked	Person (employed)—hours worked, total hours NNN	The total number of hours worked by a person over a specified period.	String	NNN

View the example 2019 Student Census Datafile Distribution online at:

<https://aristotle.cloud/item/604117/distribution/2019-student-census-datafile>

References and Links

To ensure consistency between user stories across different organisations, the Aristotle Metadata Registry uses standard terminology when describing metadata types. Using standard terms makes it easier for people communicate by giving them a common vocabulary for describing how they manage their data and metadata.

The terms **Dataset** and **Distribution** are both used in two international data standards – the ISO/IEC 11179 standard for metadata registries and the W3C DCAT Data Catalog format. Links to these standards are included below.

ISO/IEC 11179: <http://metadata-standards.org/11179/>

DCAT: <https://www.w3.org/TR/vocab-dcat-2/>

When browsing or creating metadata within the Aristotle Metadata Registry, **Datasets** and **Distributions** are included under “**Data set registration and management**” section.

Data set registration and management			
These metadata types can be used to record the existence of data catalogs, dataset and metadata specifications for data using models defined by the ISO11179 Part 7 and World Wide Web Data Catalog standard.			
Create new	Help	Basic description	
Data Catalog		A data catalog is a curated collection of metadata about datasets.	+ Create
Data Set		A collection of data, published or curated by a single agent, and available for access or download in one or more formats.	+ Create
Data Set Specification		A collection of Data Elements for collecting data.	+ Create
Distribution		Represents a specific available form of a dataset. Each dataset might be available in different forms, these forms might represent different formats of the dataset or different endpoints. Examples of distributions include a downloadable CSV file, an API or an RSS feed	+ Create

For more help on creating metadata in the Aristotle Metadata Registry, visit our Knowledge base: <https://desk.zoho.com.au/portal/aristotlecloudservicesaustralia/kb/articles/creating-metadata-items>

For more information on Aristotle.Cloud or the Aristotle Metadata Registry, visit us online at www.aristotlemetadata.com, contact the Aristotle team at hello@aristotlemetadata.com.



ARISTOTLE
metadata registry