



Introduction to Metadata

**February 2018**

# 

This document is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License](http://creativecommons.org/licenses/by-nc-sa/3.0/deed.en_GB)

# Introduction

Researchers are often confused by metadata and what exactly should be included in the metadata. In this document we define what we understand by “metadata” and why it is an essential part of documenting the data. We will introduce some popular metadata standards which we will try to explain in simple terms. We’ll look at the difference between study level metadata which we will refer to as the “Study Catalogue” and data level metadata which we will call the “Data Dictionary”. Finally we will reference the CG Core Metadata Schema.

# What is Metadata?

Metadata is often described as “*data about data*”. In basic terms it is information that enables others to fully understand a dataset or a resource. This includes where the dataset came from, how it originated, who collected or generated the data, how and where the data were collected and why they were collected.

Metadata is also used to help locate and reference a particular dataset. When a dataset is archived, most public repositories will generate a unique citation based on the metadata provided. This unique citation enables researchers to easily locate both their own datasets and datasets from other researchers within the archive.

# Metadata Standards

Over the years there have been many metadata initiatives that have attempted to standardise metadata content and format. The two most well-known schemas are the Dublin Core and the Data Documentation Initiative.

## Dublin Core Metadata Initiative

The Dublin Core Metadata Initiative (DCMI) is an open organisation supporting innovation in metadata design and best practices across metadata. The Dublin Core schema is a generic and widely-adopted schema that has been in use since the mid-1990s. For further information, go to their website at <http://www.dublincore.org/>.

## Data Documentation Initiative

The Data Documentation Initiative (DDI) is an international standard for describing the data produced by surveys and other observational methods in the social behavioural, economic and health sciences. It is a free standard that can document and manage different stages in the research data lifecycle such as conceptualisation, collection, processing, distribution, and archiving. Find out more about the DDI by going to the website of the DDI Alliance at <https://www.ddialliance.org/>.

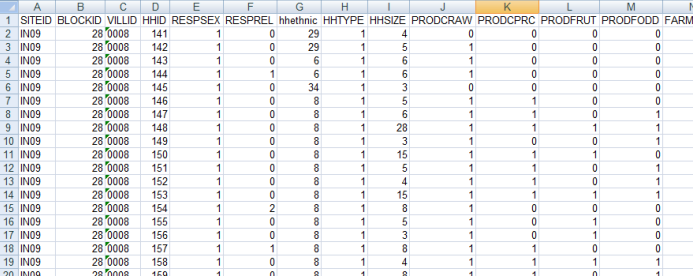
There are also several metadata editors available that help you to implement these standards. One example is the DDI Metadata Editor from the International Household Survey Network (IHSN). Further information can be found at the following site: <http://www.surveynetwork.org/software/ddi-metadata-editor>.

# Metadata in layman’s terms

While the initiatives mentioned above are very useful and can help, looking at the details can also be very daunting for the typical researcher. So let’s take a slightly different and hopefully easier approach.

Imagine you were presented with the following dataset:

Figure 1 - Typical dataset



What information would you need before you could use this dataset? Well you might need to ask the following questions:

* Where did this dataset come from? What study does it belong to? Can I have details of the study?
* Where were the data collected and when?
* Who carried out the study?
* What does each column represent?
* Are these coded data or real values – if coded, what do each of the codes mean?
* Do I have permission to use this dataset? Who should I contact for information?
* Where can I find further information?
* How were the data collected?
* Etc.

You can see here that there are top level or study level questions, but also dataset level questions. In simplistic terms the top level questions can often be answered by the Principal Investigator (PI), whereas the dataset level questions are often answered by the data manager.

As another example consider the following photo:

|  |  |
| --- | --- |
| Figure - Fieldwork photo | To fully understand what this photo is showing you might ask:   * Where was this photo taken? * When was it taken? – date, time of day * Who took the photo? * Who is the person in the photo? * What is it intended to show? * What sort of camera was used? * Etc. |

# Levels of Metadata

In the above examples we mentioned ‘study level’ and ‘dataset level’. Thinking in terms of these two levels can help you in the process of gathering together your metadata. The study level data would be cataloguing type information such as the title of the study, names of individuals and/or organisations that carried out the work, where the study was done and when, etc. We will refer to this as the ‘Study Catalogue’.

The dataset level would include detailed information about the individual datasets and what they contain. This should include a description of each variable, the type of data, and labels for coded variables. This should also include details of the format of the dataset, details of how derived variables were calculated, missing value codes, etc. We will refer to this as the ‘Data Dictionary’.

*I’m still not clear – what exactly are these two levels you talk about?*

Well you can think of this in terms of describing the entire study or activity as the top level, and describing the individual data files as the lower level. As we mentioned earlier you might find it easier to think in terms of the **“study catalogue”** and the **“data dictionary”**. Let’s start with the catalogue.

## Study Catalogue

When you come to putting your data and documents into the public domain you will often be faced with a detailed form to complete with questions requesting authoring information, study title, subject & keywords, study abstract, scope (e.g. geographical coverage), methodology (e.g. sampling procedures), etc. The following structure is taken from the Metadata Editor which is part of the IHSN’s Microdata Management Toolkit:

|  |  |
| --- | --- |
| Figure - Study information from Metadata Editor | So this would include   * The full title of the data collection. * The person or organisation responsible for the data collection’s intellectual content; * The source of funds for the data collection; * A summary describing the purpose, nature, and scope of the data collection; * Basic units of analysis – e.g. individuals, households, institutions, etc. * The timing of the data collection; * Type of sample design used including sample size; * Method of data collection; |

We recommend you consider how and where you will archive your study and see if there are any specific requirements. Dataverse for example, includes a detailed form for cataloguing information. Please see the separate document about Dataverse for details.

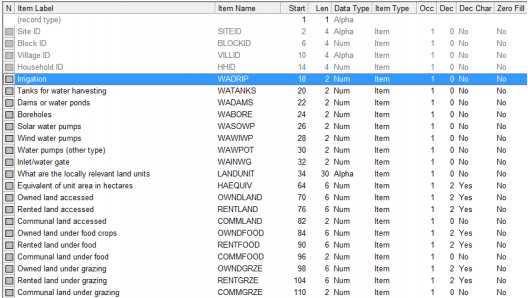
We strongly recommend you start to build your study description from the outset of your project. This ensures the information is documented and is available when needed. From experience we can tell you that waiting until the end of the project to gather this information is very time-consuming, frustrating and at times almost impossible as those with the necessary knowledge have moved onto other projects. The PI is generally the person with the necessary knowledge to complete the study catalogue – however, it is likely to be the data manager who is assigned the task of developing the metadata document.

## Data Dictionary

The data dictionary includes the data type, variable and value labels, details of missing value codes, etc. This can be a separate document or, depending on the software you use for your data, it can be included in with the data file itself.

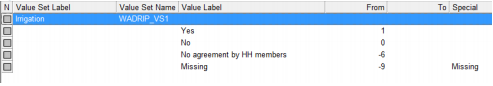
In CSPro for example, when you create a data entry system, the first thing you need to do is to create the data dictionary. Below is an example from a CSPro dictionary. Each item (or variable) has a label and a short name. We can see the data type (Alphanumeric or Number) and the length of the item which gives us an indication as to what data is allowed in the item. We can also see whether or not decimal places are allowed.

Figure - Data Dictionary from CSPro



For the coded variables CSPro has value sets similar to the one shown below:

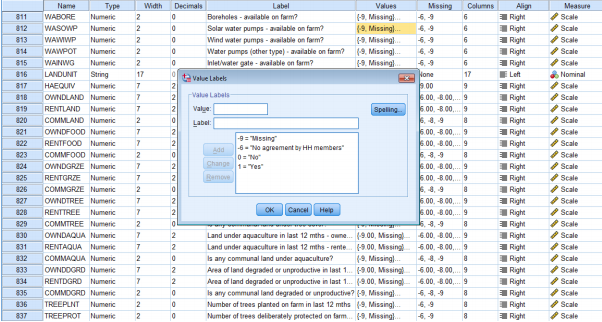
Figure - Value Set from CSPro



This also includes the missing value code for the item.

SPSS has a similar system for labelling the data and this is shown in Figure 6 below:

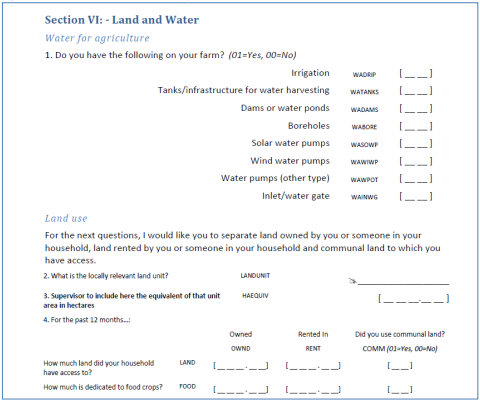
Figure - Variable View in SPSS



Remember though that this isn’t automatic. We would recommend creating a syntax file to do all the necessary labelling. In SPSS this would include **variable labels, value labels** and **missing values** commands; for other software packages you would need to find out the equivalent commands.

For survey data we would also suggest adding the variable names to the questionnaire as shown below in Figure 7.

Figure - Variable names within the questionnaire



This makes it much easier for researchers to match the data to the questions in the survey.

# CG Core Metadata

The CGIAR have developed a CG Core Metadata Schema which is a set of metadata elements used by CGIAR research centres and CRP repositories. The aim was to facilitate cross-repository searching and enhance discovery of CGIAR information products through Open Access.

The schema is closely aligned with the Dublin Core and with the Data Documentation Initiative. The generic nature of these initiatives make them ideally suited to be adopted for a wide variety of purposes.

The CG Core is designed to follow DCMI and DDI as much as possible with additional elements or attributes incorporated to capture and share CGIAR-specific administrative information. A separate document – CG Core Metadata Schema and Application Profile – provides details of the elements of the CG Core that are required.

# Summary

The key point to take away from this document is that you should start developing your meta-data documents from the start of your project. This will make it easier when you come to archiving. Develop a checklist relevant to your study/activity and to your datasets so that you can check you have all the necessary information. Think in terms of What, When, Where, Who, Why and How when considering what should be included.