Introduction to Metadata

March 2018

Stats4SD

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

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	SITEID	BLOCKID	VILLID	HHID	RESPSEX	RESPREL	hhethnic	HHTYPE	HHSIZE	PRODCRAW	PRODCPRC	PRODFRUT	PRODFODD	FARM
2	IN09	28	0008	141	1	0	29	1	4	0	0	0	0	
3	IN09	28	0008	142	1	0	29	1	5	1	0	0	0	
4	IN09	28	0008	143	1	0	6	1	6	1	0	0	0	
5	IN09	28	0008	144	1	1	6	1	6	1	0	0	0	
6	IN09	28	0008	145	1	0	34	1	3	0	0	0	0	
7	IN09	28	0008	146	1	0	8	1	5	1	1	0	0	
8	IN09	28	0008	147	1	0	8	1	6	1	1	0	1	
9	IN09	28	0008	148	1	0	8	1	28	1	1	1	1	
10	IN09	28	0008	149	1	0	8	1	3	1	0	0	1	
11	IN09	28	0008	150	1	0	8	1	15	1	1	1	0	
12	IN09	28	0008	151	1	0	8	1	5	1	1	0	1	
13	IN09	28	0008	152	1	0	8	1	4	1	1	0	1	
14	IN09	28	0008	153	1	0	8	1	15	1	1	1	1	
15	IN09	28	0008	154	1	2	8	1	8	1	0	0	0	
16	IN09	28	0008	155	1	0	8	1	5	1	0	1	0	
17	IN09	28	0008	156	1	0	8	1	3	1	0	1	0	
18	IN09	28	0008	157	1	1	8	1	8	1	1	0	0	
19	IN09	28	0008	158	1	0	8	1	4	1	1	1	1	
20	IN09	28	0008	159	1	0	8	1	8	1	1	0	1	

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 2 - 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'.

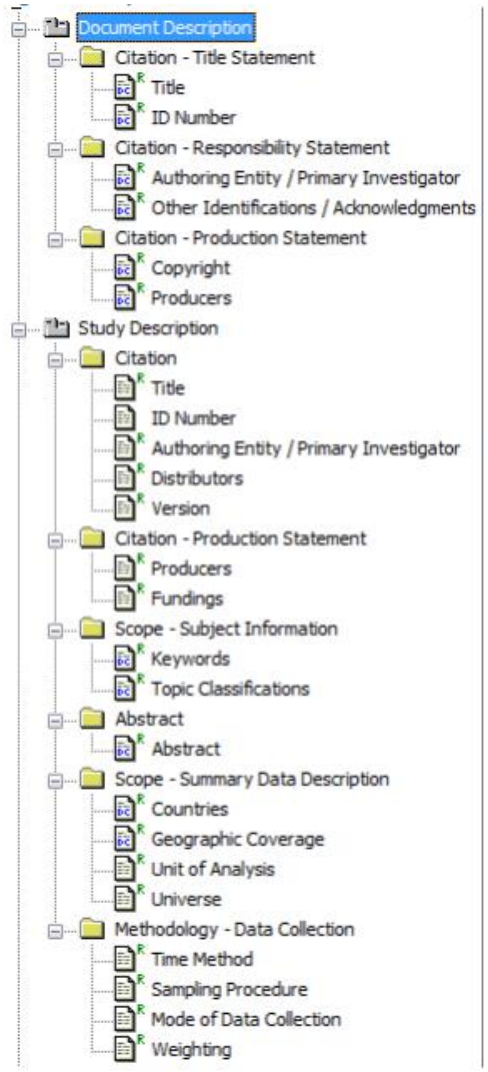
I'm still not clear – what exactly are these two levels you talk about?

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 3 - 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 are allowed in the item. We can also see whether or not decimal places are allowed.

Figure 4 - Data Dictionary from CSPro

N	Item Label	Item Name	Start	Len	Data Type	Item Type	Occ	Dec	Dec Char	Zero Fill
	(record type)		1	1	Alpha					
	Site ID	SITEID	2	4	Alpha	Item	1	0	No	No
	Block ID	BLOCKID	6	4	Num	Item	1	0	No	No
	Village ID	VILLID	10	4	Alpha	Item	1	0	No	No
	Household ID	HHID	14	4	Num	Item	1	0	No	No
	Irrigation	WADRIP	18	2	Num	Item	1	0	No	No
	Tanks for water harvesting	WATANKS	20	2	Num	Item	1	0	No	No
	Dams or water ponds	WADAMS	22	2	Num	Item	1	0	No	No
	Boreholes	WABORE	24	2	Num	Item	1	0	No	No
	Solar water pumps	WASOWP	26	2	Num	Item	1	0	No	No
	Wind water pumps	WAWIWP	28	2	Num	Item	1	0	No	No
	Water pumps (other type)	WAWPOT	30	2	Num	Item	1	0	No	No
	Inlet/water gate	WAINWG	32	2	Num	Item	1	0	No	No
	What are the locally relevant land units	LANDUNIT	34	30	Alpha	Item	1	0	No	No
	Equivalent of unit area in hectares	HAEQUIV	64	6	Num	Item	1	2	Yes	No
	Owned land accessed	OWNDLAND	70	6	Num	Item	1	2	Yes	No
	Rented land accessed	RENTLAND	76	6	Num	Item	1	2	Yes	No
	Communal land accessed	COMMLAND	82	2	Num	Item	1	0	No	No
	Owned land under food crops	OWNDFOOD	84	6	Num	Item	1	2	Yes	No
	Rented land under food	RENTFOOD	90	6	Num	Item	1	2	Yes	No
	Communal land under food	COMMFOOD	96	2	Num	Item	1	0	No	No
	Owned land under grazing	OWNDGRZE	98	6	Num	Item	1	2	Yes	No
	Rented land under grazing	RENTGRZE	104	6	Num	Item	1	2	Yes	No
	Communal land under grazing	COMMGRZE	110	2	Num	Item	1	0	No	No

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

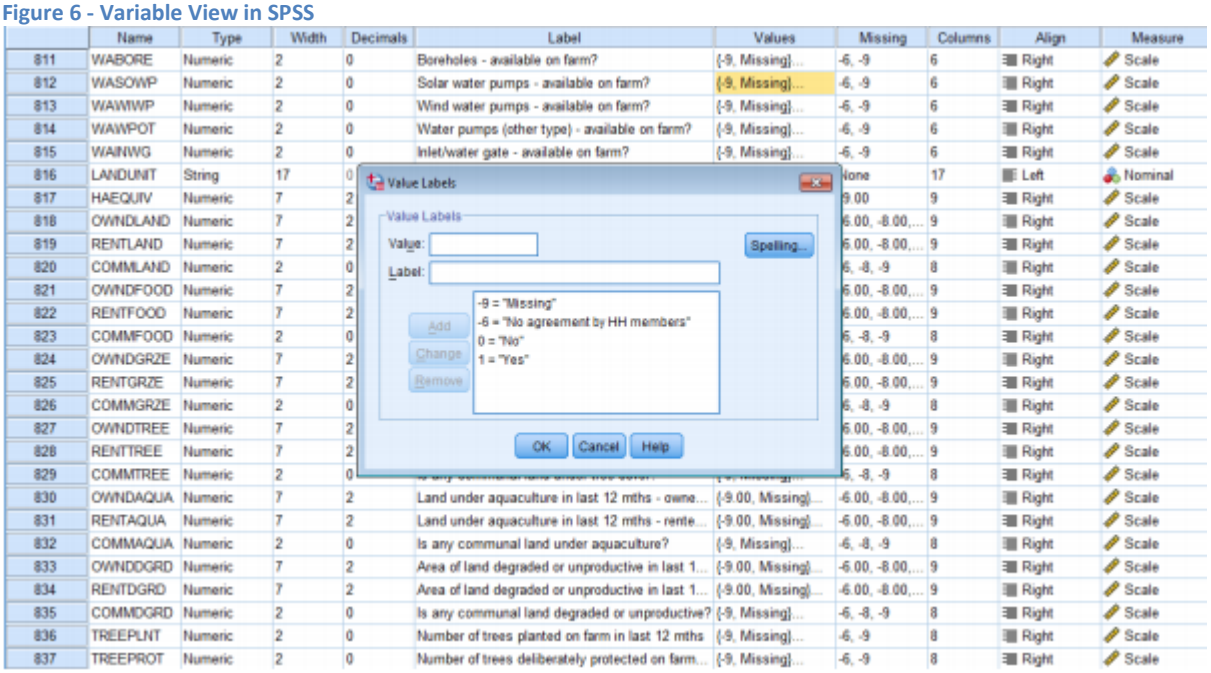
Figure 5 - Value Set from CSPro

N	Value Set Label	Value Set Name	Value Label	From	To	Special
	Irrigation	WADRIP_VS1				
			Yes	1		
			No	0		
			No agreement by HH members	-6		
			Missing	-9		Missing

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 6 - Variable View in SPSS

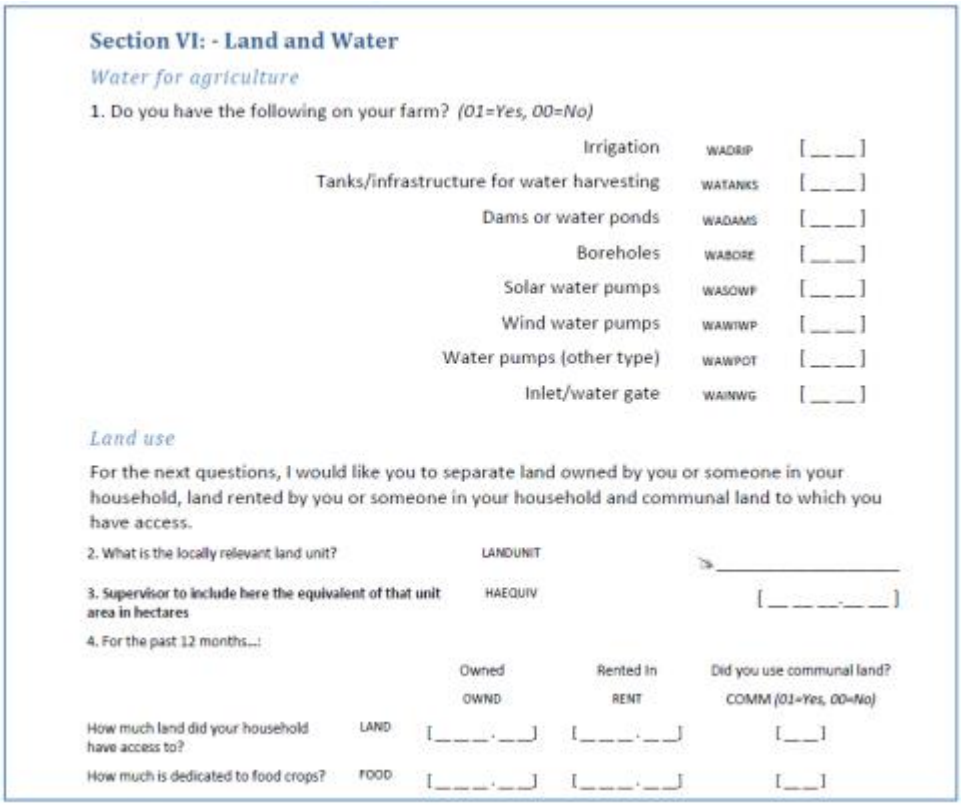


	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
811	WABORE	Numeric	2	0	Boreholes - available on farm?	{-9, Missing}...	-6, -9	6	Right	Scale
812	WASOWP	Numeric	2	0	Solar water pumps - available on farm?	{-9, Missing}...	-6, -9	6	Right	Scale
813	WAWIWP	Numeric	2	0	Wind water pumps - available on farm?	{-9, Missing}...	-6, -9	6	Right	Scale
814	WAWPOT	Numeric	2	0	Water pumps (other type) - available on farm?	{-9, Missing}...	-6, -9	6	Right	Scale
815	WAINWG	Numeric	2	0	Inlet/water gate - available on farm?	{-9, Missing}...	-6, -9	6	Right	Scale
816	LANDUNIT	String	17	0			None	17	Left	Nominal
817	HAEQUIV	Numeric	7	2			9.00	9	Right	Scale
818	OWNDLAND	Numeric	7	2			6.00, -8.00,...	9	Right	Scale
819	RENTLAND	Numeric	7	2			6.00, -8.00,...	9	Right	Scale
820	COMMLAND	Numeric	2	0			6, -8, -9	8	Right	Scale
821	OWNDFOOD	Numeric	7	2			6.00, -8.00,...	9	Right	Scale
822	RENTFOOD	Numeric	7	2			6.00, -8.00,...	9	Right	Scale
823	COMMFOOD	Numeric	2	0			6, -8, -9	8	Right	Scale
824	OWNDGRZE	Numeric	7	2			6.00, -8.00,...	9	Right	Scale
825	RENTGRZE	Numeric	7	2			6.00, -8.00,...	9	Right	Scale
826	COMMGRZE	Numeric	2	0			6, -8, -9	8	Right	Scale
827	OWNDTREE	Numeric	7	2			6.00, -8.00,...	9	Right	Scale
828	RENTTREE	Numeric	7	2			6.00, -8.00,...	9	Right	Scale
829	COMMTREE	Numeric	2	0			6, -8, -9	8	Right	Scale
830	OWNDAQUA	Numeric	7	2	Land under aquaculture in last 12 mths - owne...	(-9.00, Missing)...	-6.00, -8.00,...	9	Right	Scale
831	RENTAQUA	Numeric	7	2	Land under aquaculture in last 12 mths - rente...	(-9.00, Missing)...	-6.00, -8.00,...	9	Right	Scale
832	COMMAQUA	Numeric	2	0	Is any communal land under aquaculture?	{-9, Missing}...	-6, -8, -9	8	Right	Scale
833	OWNDDGRD	Numeric	7	2	Area of land degraded or unproductive in last 1...	(-9.00, Missing)...	-6.00, -8.00,...	9	Right	Scale
834	RENTDGRD	Numeric	7	2	Area of land degraded or unproductive in last 1...	(-9.00, Missing)...	-6.00, -8.00,...	9	Right	Scale
835	COMMDGRD	Numeric	2	0	Is any communal land degraded or unproductive?	{-9, Missing}...	-6, -8, -9	8	Right	Scale
836	TREEPLNT	Numeric	2	0	Number of trees planted on farm in last 12 mths	{-9, Missing}...	-6, -9	8	Right	Scale
837	TREEPROT	Numeric	2	0	Number of trees deliberately protected on farm...	{-9, Missing}...	-6, -9	8	Right	Scale

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 7 - Variable names within the questionnaire



Section VI: - Land and Water

Water for agriculture

1. Do you have the following on your farm? *(01=Yes, 00=No)*

Irrigation	WADRIP	[__ __]
Tanks/infrastructure for water harvesting	WATANKS	[__ __]
Dams or water ponds	WADAMS	[__ __]
Boreholes	WABORE	[__ __]
Solar water pumps	WASOWP	[__ __]
Wind water pumps	WAWIWP	[__ __]
Water pumps (other type)	WAWPOT	[__ __]
Inlet/water gate	WAINWG	[__ __]

Land use

For the next questions, I would like you to separate land owned by you or someone in your household, land rented by you or someone in your household and communal land to which you have access.

2. What is the locally relevant land unit? LANDUNIT ➔ _____

3. **Supervisor to include here the equivalent of that unit area in hectares** HAEQUIV [__ __ __.__ __]

4. For the past 12 months...:

		Owned OWND	Rented In RENT	Did you use communal land? COMM *(01=Yes, 00=No)*
How much land did your household have access to?	LAND	[__ __ __.__ __]	[__ __ __.__ __]	[__ __]
How much is dedicated to food crops?	FOOD	[__ __ __.__ __]	[__ __ __.__ __]	[__ __]

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 makes 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.

Associate Videos

Videos accompanying the original release of the CCAFS Data Management Support Pack in 2013 are available as a playlist on the Statistical Services Centre YouTube Channel at https://www.youtube.com/channel/UCs7EU95YMjhvNozJKCD92xQ/playlists. These videos have not been updated since the original release but are mostly still relevant.

In particular the playlist includes a video on “Metadata” which is available at: https://www.youtube.com/watch?v=AdX5OUJY9P0&index=11&list=PLK5PktXR1tmNRaUPsFiYlyhg2lui0xgpj