# File and folder naming convention World Bank Microdata Library Olivier Dupriez and Matthew Welch, DECDG

4/30/2021

# Folders

## Top Level – Country

* The top level folder name is the country name as used by the WB in the World Development Indicators

## Second Level - Study

* The second level folder will be the survey\study folder. This is comprised of the 3 letter ISO code then the survey year then the acronym for the survey. All separated by **underscores.**
* Example: the folder containing the 2002 LSMS from Albania would be **ALB\_2002\_LSMS**

## Third Level – Study Master and Adaptations

* The third level folder contains folders for all the versions of that study. The naming convention differs depending on whether the data is the Master version of the study or an adaptation and by version number.

### Master

* The original version of the study as deposited in the Microdata Library will always be designated as the Master File. In our Albania 2002 LSMS example the folder for the master file for version 1 of the data would be named as follows: **ALB\_2002\_LSMS\_v01\_M**
* The v01 indicates that this is version 1 of the master file and the M indicates that this is the Master file.
* If at some point a second version of this study is released then a new folder is created for this study with the same name but the version number would change to v02…v0#

### Study ID Components

* The study ID should include the following components in the following order separated by an underscore: a country code, a year, a study abbreviation, a master file version number, and M for the master file. If the study is an adaptation, the study ID also reflects a version and a type of the adaptation, after the part of the ID describing the master.
* The country is indicated by ISO 3166-1 three-letter country code. If the data was collected in multiple regions/countries, use the following codes: WLD for World, AFR for Sub-Saharan Africa, ECA for Europe and Central Asia region, EAP for East Asia and Pacific region, LAC for Latin America and the Caribbean, MENA for Middle East and North Africa, and SAR for South Asia.

Example:

ALB\_2002\_LSMS\_v01\_M – Albania Living Standards Measurement Survey 2002

WLD\_2012\_PISA\_v01\_M – World Programme for International Student Assessment 2012

* The year is indicated by the year when the data collection started. The exception of this rule are time series/panel datasets.

Example:

ALB\_2002\_LSMS\_v01\_M – the dates of collection are April 2002 to July 2002.

In some cases, the data collection could span from one year to another. Include the start year of the data collection in the survey ID, even if most of the data was collected in the subsequent year.

Example:

MAR\_2006\_ENNVM\_v01\_M – the dates of collection for this survey are December 2006 to November 2007.

Note, the study title is more flexible than the study ID. In this example, the producer of the survey, Morocco statistics office, refers to the survey as Enquête Nationale sur les Niveaux de Vie des Ménages 2007. We should follow the survey producer’s practice and include 2007 in the study title.

For the time series/panel datasets include the range of years from the year the data collection started for the first panel/round, to the year the data collection started for the last panel/round. The years should be separated by a hyphen.

Examples:

SLE\_2014-2015\_HFCPS\_v01\_M - Sierra Leone High Frequency Cell Phone Survey on the Socio-Economic Impacts of Ebola 2014-2015

This survey includes three rounds: the first round was conducted in November 2014 and the last one – in May 2015.

TZA\_2006-2013\_ES-P\_v01\_M - Tanzania Enterprise Survey 2006-2013, Panel Data

The dataset documented in this study includes panel data from Enterprise Surveys conducted in 2006 and 2013. The data collection period for the first survey is November to December 2006, for the second survey – January 2013 to February 2014.

* The study abbreviation is the acronym the survey is known by.

Most of the surveys have acronyms that are well-known and recognized: HBS - Household Budget Survey, LFS – Labor Force Survey, LSMS - Living Standards Measurement Survey, DHS – Demographic and Health Survey, MICS – Multiple Indicator Cluster Survey, HIS – Integrated Household Survey, and others.

If the documented study is one of the kind or doesn’t have a known acronym, abbreviate the survey title by combining first letters of the words in the survey title.

Example:

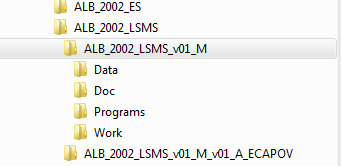
POL\_2014\_EDPFI\_v01\_M - Poland Entrepreneurial Discovery Process Project 2014, Firm Interviews

In some cases the study abbreviation could include additional components, separated from the acronym by a hyphen. Please refer to Special Cases section later in the document.

* The master study and adaptation studies’ version numbering is described later in the document.

### Adaptations

* Adaptations of the Master files are placed in their own folder on the same level in the folder tree as the master file.
* Adaptations are versions of the study that have been changed in some way. Generally to make them more useable for a particular project. They contain the same data, but will often have been reshaped, harmonized or contain derivations not found in the original Master files. Examples of adaptations within the World Bank are datasets such as the ECAPOV harmonized data and the SEDLAC harmonized data.
* The folder for version 1 of the Albania 2002 LSMS harmonized from the version 1 of the Master file for the ECAPOV project would look as follows:  
  **ALB\_2002\_LSMS\_v01\_M\_v01\_A\_ECAPOV**
* The following has been added to the study name given to the master file: **\_v01\_A\_ECAPOV**
* The **\_ v01** indicates that this is version 1 of the adaptation, the **\_A** indicates that this is an adaptation and the **\_ECAPOV** indicates that this adaptation forms part of the ECAPOV project here at the World Bank.



### Version numbering for adaptations

Irrespective of the number of master file versions archived, the adaptation version number should always move a level up and follow the sequence of already existing adaptation versions. This helps to keep track of the number of adaptations, and to inform on the latest adaptation version.

Example:

EST\_2003\_HBS\_v01\_M\_v01\_A\_ECAPOV

EST\_2003\_HBS\_v01\_M\_v02\_A\_ECAPOV

EST\_2003\_HBS\_v01\_M\_v03\_A\_ECAPOV

EST\_2003\_HBS\_v02\_M\_v04\_A\_ECAPOV

However,

The version numbering for adaptations of a new series of a harmonization (for example, SEDLAC-02, SEDLAC-03) will start with v01 for a new series.

Example:

PRY\_2001\_EIH\_v01\_M\_v03\_A\_SEDLAC-02

PRY\_2001\_EIH\_v01\_M\_v01\_A\_SEDLAC-03

### Grouped Data

Grouped data, data consisting of frequencies, mid-points, shares, is assigned a study ID as an adaptation with the word “GROUPED” at the end of the study ID:

Example:

CHN\_1990\_HBS\_v01\_M\_v01\_A\_GROUPED

The version numbering for grouped data follows the same rule as adaptations versions numbering.

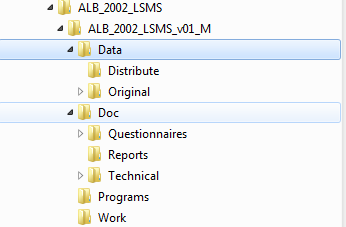
### Public Use Versions

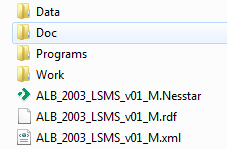
* Data which has had confidential information suppressed or removed from it in order to allow for public distribution is considered an adaptation of the Master.
* Naming for such a study has **\_A\_PUF** added at the end. For example if the Albania study we used as an example above had identifying information in it and we removed it in order to allow for data distribution the folder name would be: **ALB\_2002\_LSMS\_v01\_M\_v01\_A\_PUF**

### The same study in different languages

* There may be cases where the same version of a study is archived but in a different language. For example: you may have and Indonesian study in the archive which has been documented in Bahasa and also in English.
* Under the system above the studies would have the same ID. To prevent this it is necessary to indicate the difference by adding the two letter ISO code for the language as part of the versioning section. \_v01-ID (Bahasa Indonesia). Note: there is a hyphen between the version and the language code.
* It is not necessary to add the language code to all studies – only to those ID’s would conflict due to a translated version being included. It is also not necessary to put the EN code on the English one.
* Example: IDN\_1990\_SUSENAS\_**v01-ID**\_M indicates that this version is the Bahasa Indonesia version.
* For a list of ISO language codes you can look here: <http://en.wikipedia.org/wiki/List_of_ISO_639-1_codes>

## Fourth Level – Study Folders

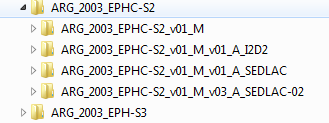
* All study folders contain the following sub-folders  
  
* A **Data** Folder which contains two folders. A folder for the **Original** data and a folder for the data that should be **Distributed** to users.
* A **Doc** folder which contains three folders. One each for **Questionnaires**, **Reports** and **Technical** documents.
* A folder for statistical **Programs**
* And a folder called **Work** to be used for temporary work files during the process of documenting the study.
* The Nesstar file, the DDI and the RDF files all live in the root of this folder.



## Special Cases

### Surveys conducted multiple times a year

* Sometimes the same study with the same name is conducted in the same year. For example every semester or every quarter.
* In such cases each ‘WAVE’ of the study should be placed in its own second level folder with the Master files and Adaptations within folders under that.



* The example above is for the Argentinean EPH surveys. These are conducted every semester. Note: that the first and second semester studies each have their own second level folder. The Semester is indicated as part of the Survey Acronym and separated by a hyphen. **\_EPH-S1** for the first semester and **\_EPH-S2** for the second semester. You would use **Q** for quarter or **W** for wave etc.
* A hyphen is also used in cases where a study abbreviation would be broken by a space. E.g. AUT\_2012\_EU-SILC\_v01\_M\_v02\_A\_UDB-C,
* The folder names for the Master and Adaptations then follow from there – see example above.

### Higher Frequency surveys

* Some surveys are conducted at a higher frequency e.g. monthly, weekly etc. We can archive them in the same way we would other time periods described above.
* The example below illustrates how we archive a 12 wave 2015 Listening to Tajikistan high frequency survey, using W for wave.

**TJK\_2015\_** **L2TJK-W1\_v01\_M**

**TJK\_2015\_** **L2TJK-W2\_v01\_M**

**TJK\_2015\_** **L2TJK-W3\_v01\_M**

**TJK\_2015\_** **L2TJK-W4\_v01\_M**

**TJK\_2015\_** **L2TJK-W5\_v01\_M**

**TJK\_2015\_** **L2TJK-W6\_v01\_M**

**TJK\_2015\_** **L2TJK-W7\_v01\_M**

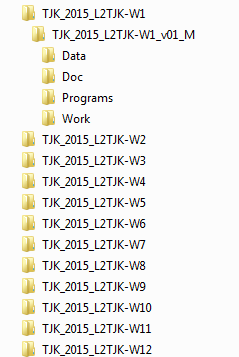
**TJK\_2015\_** **L2TJK-W8\_v01\_M**

**TJK\_2015\_** **L2TJK-W9\_v01\_M**

**TJK\_2015\_** **L2TJK-W10\_v01\_M**

**TJK\_2015\_** **L2TJK-W11\_v01\_M**

**TJK\_2015\_** **L2TJK-W12\_v01\_M**



We could also use M for month e.g. **MEX\_2002\_ENCO-M12\_v01\_M**

The metadata will describe the time interval. (10 days or whatever it ends up being).

“Is there an intention to merge these waves at any point? How will it be most usefully presented to the user? We have other data like this that is consolidated into yearly files at the end of each year. How is this data most conveniently used will determine how it is presented to the user?” Encourage producers to think about this.

**Survey Modules**

* Sometimes a module of a survey needs to be documented separately, for example, when a data access to the module is different from an access to other modules, or when only separate modules are available for documentation.
* In such cases, the survey ID and the survey title should indicate which module is documented.
* In the survey ID, an abbreviation of the module (for example, MG for migration, CON for consumption, LB for labor, IN for income, DEM for demographics, SP for social protection, etc.) should be added to the survey abbreviation after a hyphen.

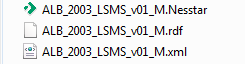
Example:

KGZ\_2015\_KIHS\_v01\_M (Kyrgyz Integrated Household Survey 2015)

KGZ\_2015\_KIHS-MG\_v01\_M (Kyrgyz Integrated Household Survey 2015, Migration Module)

# File names

* The Nesstar, DDI\xml, and rdf files all take on the same name as given to the study folder in which they are stored
* Example the Albania 2003 LSMS version 1 of the Master file will have the following file names for the nesstar, ddi\xml and rdf



# Study ID and DDI naming convention

* The naming convention for the **DDI Document ID Number** and the Study Description **ID Number** which is entered in the Nesstar Publisher follows the same name as those used for the Nesstar file name.
* The **DDI Document ID Number** is however always preceded by **DDI\_** and followed by the acronym of the DDI producer, e.g. **\_WB** if the DDI was produced by the World Bank
* **Examples**: For version 1 of the Albania 2003 LSMS master file **the DDI Document ID Number** would be **DDI\_ALB\_2003\_LSMS\_v01\_M\_WB**
* For the same study it’s **ID Number** would be **ALB\_2003\_LSMS\_v01\_M**
* For the LSMS for the 2002 ECAPOV adaptation the **DDI Document ID Number** would be **DDI\_ALB\_2002\_LSMS\_v01\_M\_v01\_A\_ECAPOV\_WB** and the **ID Number ALB\_2002\_LSMS\_v01\_M\_v01\_A\_ECAPOV**

# Geospatial data

We document Geospatial data in the same way that we document microdata in the Microdata Library. Within the Geospatial Library (GeoLib), the files are organized in the same way as the microdata files in DataLib, except that the master file is denoted by a G instead of an M suffix.

## Top Level – Country

* The top level folder name is the country name as used by the WB in the World Development Indicators

## Second Level - Study

* The second level folder will be the survey/study folder. This is comprised of the 3 letter ISO code then the survey year then the acronym for the survey. All separated by **underscores.**
* Example: the folder containing the 2010 WPOP-POP data from Algeria would be

**DZA\_2010\_WPOP-POP**

## Third Level – Study Master and Adaptations

* The third level folder contains folders for all the versions of that study. The naming convention differs depending on whether the data is the Master version of the study or an adaptation and by version number.

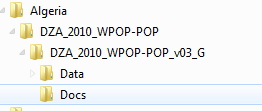
### Master

* The original version of the study as deposited in the Geo Spatial Library will always be designated as the Master File (denoted by G). In our Algeria 2010 WPOP-POP example the folder for the master file for version 1 of the data would be named as follows:

**DZA\_2010\_WPOP-POP\_v01\_G**

* The v01 indicates that this is version 1 of the master file and the G indicates that this is the Master file.
* If at some point a second version of this study is released then a new folder is created for this study with the same name but the version number would change to v02…v0#

Example



### Adaptations

* Adaptations of the Master files are placed in their own folder on the same level in the folder tree as the master file.

As with microdata, adaptations are versions of the study that have been changed in some way, generally to make them more useable for a particular project. If for instance, the Algeria geospatial data was anonymized for public distribution to make a public use file, an adaptation of the Algeria 2010 WPOP-POP Master file would be **DZA\_2010\_WPOP-POP\_v03\_G\_v01\_A\_PUF**