User-friendly framework for metadata and microdata documentation based on international standards and PCBS Experience

Haitham Zeidan (Haitham@pcbs.gov.ps)[1] , Geoffrey Greenwell (geoffrey.greenwell@oecd.org)[2]

Keywords: Metadata, Microdata, DDI, DCMI, SDMX, RDF, XML, Semantic web, NADA, PCBS, OECD.

Abstract

This paper discussed and investigated the experience of Palestinian Central Bureau of Statistics (PCBS) [1] in designing documentation model and user-friendly framework for metadata and microdata documentation. PCBS uses two metadata specifications: the Data Documentation Initiative (DDI) [2] and the Dublin Core Metadata Initiative (DCMI) [3]. Both are defined in the Extensible Mark-up Language (XML) and the Resource Description Framework (RDF). This paper focused also on the DDI and DCMI as well as its relationship to other relevant metadata standards (e.g., The Statistical Data and Metadata Exchange (SDMX)) [4] and the semantic web technologies, we addressed the features of these standards as Richer content, Coverage, On-line analytical capability, Search capability and Interoperability since these standards are defined in the Extensible Mark-up Language (XML).

1. Introduction

Good documentation has a number of features. It should accurately describe the data. The information should be clear so that the data are not incorrectly used. It should also be comprehensive, so that the statistical agency is not dependent on the institutional memory of staff. A basic principle is that all information that can foster the effective and accurate use of datasets by secondary users should be preserved and disseminated.

Unfortunately, documentation is often the last step of the survey process, and it is then often too late to capture all metadata produced during the life cycle of the survey activities. This results in the loss of useful information generated at early stages, such as the comments received from various stakeholders at the stage of questionnaire design, problems encountered during pilot-testing of the questionnaire, etc. Treating documentation as an ongoing part of survey activity will reduce the documentation costs and increase its quality. Using the international metadata standards, such as the Data Documentation Initiative (DDI) and the Dublin Core Metadata Initiative (DCMI) specifications, can reduce the burden considerably, because they provide a rigorous framework for organizing the process and will help to address the technical issues related to documentation, preservation and dissemination process of the surveys, in addition to improve management and use of microdata.

2. Objectives

The objectives of this study is to display the user-friendly framework for metadata and microdata documentation that introduced in Palestinian central bureau of statistics (PCBS) for better documenting, preserving, anonymizing and disseminating of existing

[1] Palestinian Central Bureau of Statistics (PCBS)

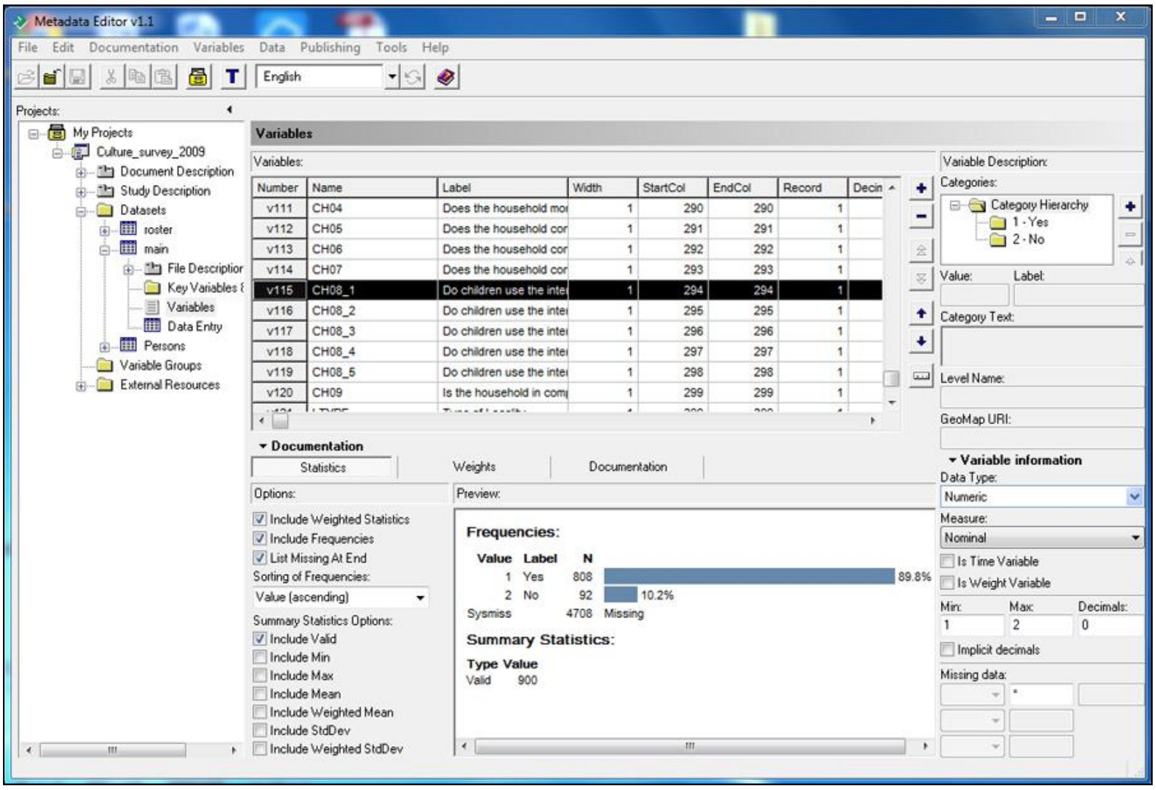
[2] The Organization for Economic Cooperation and Development (OECD)

microdata, this framework produced based on international standards: Data Documentation Initiative (DDI) and the Dublin Core Metadata Initiative (DCMI).

3. USER-FRIENDLY FRAMEWORK FOR METADATA AND MICRODATA DOCUMENTATION

Our framework produced based on international standards: Data Documentation Initiative (DDI) and the Dublin Core Metadata Initiative (DCMI). We use in PCBS The IHSN Metadata Editor, also known as the Nesstar Publisher [5] as shown in figure (1), which is a rich editor for the preparation of metadata and data for publishing in an online catalog, such as the IHSN-developed National Data Archive (NADA) [6]. The metadata produced by the Editor is compliant with the Data Documentation Initiative (DDI) and the Dublin Core XML metadata standards. The application is developed by Nesstar at the Norwegian Social Science Data Archive (NSD) and is distributed as freeware. The features of metadata editor are:

- Easy editing/creation and export of DDI documented datasets with XML experience needed.
- Tools to validate metadata and variables.
- The ability to include automatically generated frequency and summary statistics for each variable.
- Tools to compute/recode/label new, or existing, variables to be added to a dataset before publishing.
- The ability to import and export data to the most common statistical formats, including delimited files.
- Multilingual - Arabic, Chinese, English, French, Portuguese, Russian and Spanish.



Variables:

Number	Name	Label	Width	StartCol	EndCol	Record	Decin
v111	CH04	Does the household mo	1	290	290	1	
v112	CH05	Does the household cor	1	291	291	1	
v113	CH06	Does the household cor	1	292	292	1	
v114	CH07	Does the household cor	1	293	293	1	
v115	CH08_1	Do children use the inte	1	294	294	1	
v116	CH08_2	Do children use the inte	1	295	295	1	
v117	CH08_3	Do children use the inte	1	296	296	1	
v118	CH08_4	Do children use the inte	1	297	297	1	
v119	CH08_5	Do children use the inte	1	298	298	1	
v120	CH09	Is the household in com	1	299	299	1	

Options:

☑ Include Weighted Statistics
☑ Include Frequencies
☑ List Missing At End
Sorting of Frequencies: Value (ascending)
Summary Statistics Options:
☑ Include Valid
☐ Include Min
☐ Include Max
☐ Include Mean
☐ Include Weighted Mean
☐ Include StdDev
☐ Include Weighted StdDev

Preview:

Frequencies:

```
Value  Label   N
    1  Yes   808   89.8%
    2  No     92   10.2%
Sysmiss     4708   Missing
```

Summary Statistics:

```
Type   Value
Valid  900
```

Data Type: Numeric
Measure: Nominal
☐ Is Time Variable
☐ Is Weight Variable
Min: 1 Max: 2 Decimals: 0
☐ Implicit decimals
Missing data: *

Figure 1. The Features of Metadata Editor

3.1. Data Documentation Initiative (DDI)

The DDI developed standards that provide a structured framework for organising the content, presentation, transfer and preservation of metadata in the social and behavioural sciences. It enables documenting even the most complex microdata files in a way simultaneously flexible and rigorous.

The DDI seeks to establish an international XML based standard for microdata documentation. Its aim is to provide a straightforward means of recording and communicating to others all the salient characteristics of micro-datasets. The DDI specification is a major transformation of the once-familiar electronic ‘codebook’: it retains the same set of capabilities but greatly increases the scope and rigour of the information contained therein.

3.1.1 DDI Features

Interoperability: DDI-compliant documentation can be exchanged and transported seamlessly, and applications can be generically written, because the documents are homogeneous.

Richer content: The DDI provides data analysts with broader knowledge about data content, because the DDI initiative provides a comprehensive set of elements that can describe micro-datasets as completely and as thoroughly as possible.

Multipurpose documentation: A DDI codebook can be restructured to suit different applications, because it contains all the information necessary to produce different types of output.

On-line analytical capability: DDI documents can be easily imported into on-line analysis systems, rendering datasets more readily usable by a wider audience. This is made possible because the DDI mark-up extends down to the variable level and provides a standard uniform structure and content for variables.

Search capability: Field-specific searches across documents and studies are made possible, because each of the elements in a DDI-compliant codebook is tagged in a specific way.

3.1.2 DDI Coverage

The DDI specification has been designed to fully encompass the kinds of data generated by surveys, censuses, administrative records, experiments, direct observation, and other systematic methodologies for generating empirical measurements. In other words, the unit of analysis could be individual persons, households, families, business establishments, transactions, countries, or other subjects of scientific interest. Similarly, observations may consist of measures taken at a single point in time in a single setting, such as a sample of people in one country during one week, or they may consist of repeated observations in multiple settings, including longitudinal and repeated cross-sectional data from many countries, as well as time series of aggregate data. The DDI specification also provides for full descriptions of the methodology of the study (mode of data collection, sampling methods if applicable, universe, geographical areas of study, responsible organization and persons, and so on).

3.1.3 DDI Structure

The DDI specification permits all aspects of a survey to be described in detail: the methodology, responsibilities, files and variables. It provides a structured and comprehensive list of hundreds of elements and attributes that may be used to document a dataset, although it is unlikely that any one study would use all of them. However, some elements, such as “Title,” are mandatory (and must be unique). Other elements are optional and can be repeated, for example “Authoring Entity/Primary Investigator”, since it includes information on the person(s) and/or organization(s) responsible for the survey.

The DDI elements are organized in five sections:

Section 1.0: Document Description: A study (survey, census or other) is not always documented and disseminated by the same agency as the one that produced the data. It is therefore important to provide information (metadata) not only on the study itself, but also on the documentation process. The Document Description consists of overview information describing the DDI-compliant XML document, or, in other words, “metadata about the metadata”.

Section 2.0: Study (Survey) Description: The Study Description consists of overview information about the study. This section includes information about how the study should be cited, who collected, compiled and distributes the data, a summary (abstract) of the content of the data, information on data collection methods and processing, and so on.

Section 3.0: Data File Description: This section is used to describe each data file (Microdata) in terms of content, record and variable counts, version, producer, and so on.

Section 4.0: Variable Description: This section presents detailed information on each variable, including literal question text, universe, variable and value labels, derivation and imputation methods, and so on.

Section 5.0: Other Material: This section allows for the description of other materials related to the study or survey. These can include resources such as documents (questionnaires, coding information, technical and analytical reports, interviewer's manuals, and so on), data processing and analysis programs, photos, and maps. However, the Dublin Core Metadata Initiative (described below) is better suited for the framework requirements.

3.2. The Dublin Core Metadata Initiative (DCMI)

The DCMI Metadata Element Set (ISO standard 15836), also known as the Dublin Core metadata standard, is a simple set of elements for describing digital resources. This standard is particularly useful to describe resources related to microdata such as questionnaires, reports, manuals, data processing scripts and programs, etc. It was initiated in 1995 by the Online Computer Library Center (OCLC) and the National Center for Supercomputing Applications (NCSA) at a workshop in Dublin, Ohio. Over the years it has become the most widely used standard for describing digital resources on the Web and was approved as an ISO standard in 2003. The standard is maintained and further developed by the Dublin Core Metadata Initiative - an international organization dedicated to the promotion of interoperable metadata standards.

3.2.1 DCMI Elements

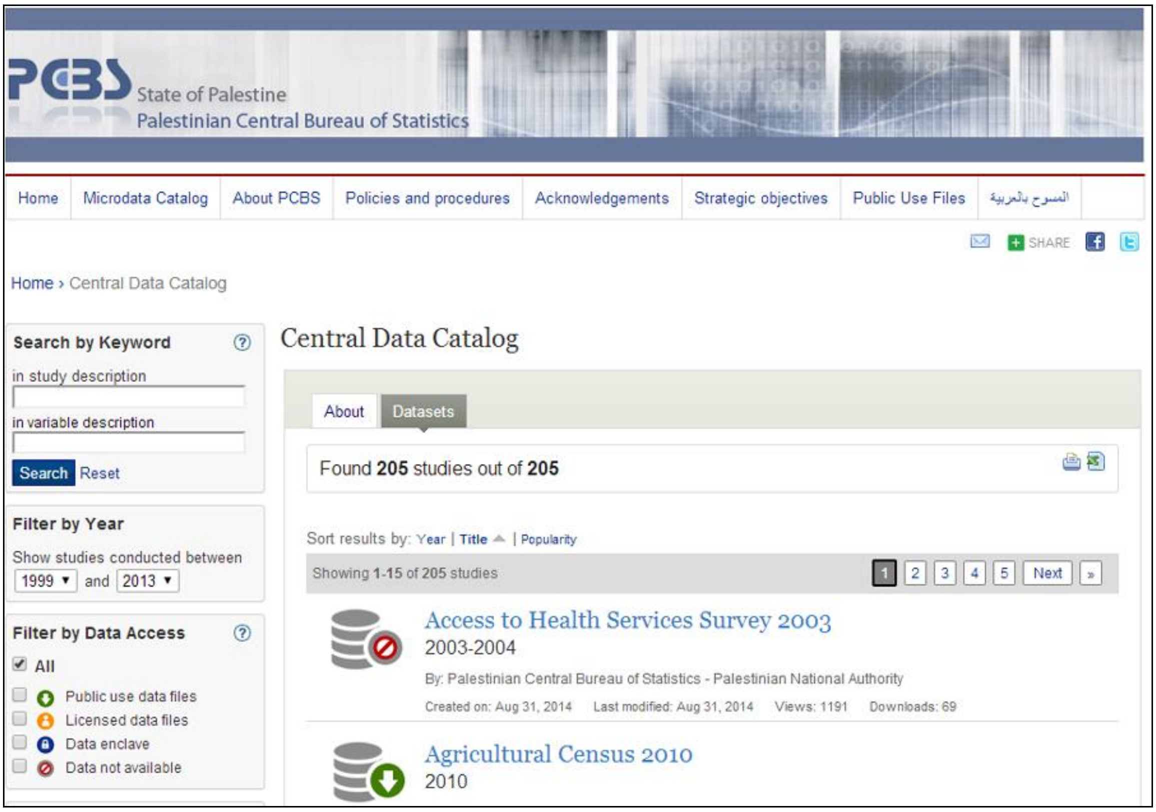
The Dublin Core metadata standard is based on the same principles as the DDI specification. It consists of a set of elements (or "tags"), organized to form an XML file. The Dublin Core standard includes two levels: Simple and Qualified. In the framework, only the Simple Dublin Core elements are used. They include the following fifteen elements as shown in Table (1) below:

Table 1. The Dublin Core Metadata Initiative (DCMI) Elements

Element	Details
Title	The name by which the resource is formally known.
Subject	The topic of the resource.
Description	An abstract, a table of contents, or a free-text account of the content.
Type	The nature or genre of the content of the resource (e.g., a survey questionnaire, a data processing syntax program, a map).
Source	A reference to a resource (e.g., a PDF filename, or a website URL).
Relation	A reference to a related resource (this element will rarely be used).
Coverage	The extent or scope of the content of the resource. Coverage will typically include spatial location (e.g., a country), or a temporal period (a date or date range).
Creator	The person(s), organization(s), or service(s) responsible for making the content of the resource.
Publisher	The person(s), organization(s), or service(s) responsible for making the resource available.
Contributor	The person(s), organization(s), or service(s) having contributed to the content of the resource.
Rights	A rights management statement for the resource.
Date	A date associated with an event in the life cycle of the resource. Typically, Date will be associated with the creation or availability of the resource.
Format	For use in determining the software, hardware or other equipment needed to display or operate the resource (e.g., "STATA Version 8"; or "MS-Excel 2000").
Identifier	An unambiguous reference to the resource within a given context. Examples of formal identification systems include the Uniform Resource Locator (URL), and the International Standard Book Number (ISBN).
Language	A language of the intellectual content of the resource.

3.3. Dissemination Surveys Using National Data Archive (NADA)

After documentation process we in PCBS disseminate the surveys on National Data Archive (NADA) portal as shown in figure (2), NADA is a web-based cataloging application that allows for the creation of portals that allows users to browse, search, compare, apply for access, and download relevant census or survey information. It was originally developed to support the establishment of national survey data archives. The application is used by a diverse and growing number of national, regional, and international organizations. NADA, uses the Data Documentation Initiative (DDI), XML-based international metadata standard.



PCBS State of Palestine
Palestinian Central Bureau of Statistics

Home | Microdata Catalog | About PCBS | Policies and procedures | Acknowledgements | Strategic objectives | Public Use Files | المسوح بالعربية

SHARE

Home › Central Data Catalog

Search by Keyword

in study description

in variable description

Search Reset

Filter by Year

Show studies conducted between 1999 and 2013

Filter by Data Access

☑ All
□ Public use data files
□ Licensed data files
□ Data enclave
□ Data not available

Central Data Catalog

About | Datasets

Found 205 studies out of 205

Sort results by: Year | Title | Popularity

Showing 1-15 of 205 studies 1 2 3 4 5 Next »

Access to Health Services Survey 2003
2003-2004
By: Palestinian Central Bureau of Statistics - Palestinian National Authority
Created on: Aug 31, 2014 Last modified: Aug 31, 2014 Views: 1191 Downloads: 69

Agricultural Census 2010
2010

Figure 2. National Data Archive (NADA) portal used to disseminate documented surveys.

4. DDI AND SDMX STANDARDS

Recently, two technical standards for statistical and research data and metadata have been receiving much attention. Particularly for those working with both micro-data and time-series aggregates, there can be some confusion as to the relationship between these standards, and questions about which may be more appropriate for use in a particular application or institution. We describes the basic scope of each standard.

The Statistical Data and Metadata Exchange (SDMX) technical specifications come out of the world of official statistics and aim to foster standards for the exchange of statistical information. They have been created by the Statistical Data and Metadata Exchange Initiative. The initiative is a cooperative effort between seven international organizations: the Bank for International Settlement (BIS), the International Monetary Fund (IMF), the European Central Bank (ECB), Eurostat, the World Bank (WB), the Organization for Economic Cooperation and Development (OECD), and the United Nations Statistical Division (UNSD). The output of this initiative is not just the technical standards, but also addresses the harmonization of terms, classifications, and concepts which are broadly used in the realm of aggregate statistics.

The Data Documentation Initiative (DDI) is a specification for capturing metadata about social science data. It is maintained by the Data Documentation Initiative Alliance, a membership-driven consortium including universities, data archives, and national and international organizations. The specification was originally created to capture the information found in survey codebooks, which remains the focus of the first two versions. The DDI 3.0 version covers the whole data lifecycle, from the survey instrument design to archiving, dissemination and repurposing, allowing for a description of re-codes, processing, and comparison of studies by design or after-the-fact [7].

SDMX and the latest version of the DDI have been intentionally designed to align themselves with each other as well as with other metadata standards. Because much of the micro-data described by DDI instances is aggregated into the higher level data sets found at the time-series level, it is not surprising that the two have been designed to work well together. Although there is some overlap in their descriptive capacity, they can best be characterized as complementary, rather than competing [8].

4.1. DDI/SDMX Overlap

SDMX provides XML formats for describing data and independent metadata structures, which can be user-configured to hold any concepts desired. They also provide XML formats based on these configurations. The concept of exchanging a data set or a metadata set is the primary focus in SDMX, which is optimized for the exchange of aggregate data. The typical case is the exchange of time series data.

DDI also provides the ability to describe a rich set of metadata in an XML format, with an emphasis on micro-data, but also allowing for tabular formats and multidimensional cubes. In the 3.0 version, DDI supports all phases of the lifecycle from a description of concepts and the survey instrument used to collect data to the end product held in a data archive and used for analysis. DDI 3.0 also provides an XML format for micro-data and tabular/multi-dimensional data, but very often the data is held in text or statistical software specific binary files. The user-configurable aspects of DDI ("variables") are mixed with specific metadata fields.

These two standards are well aligned means that they can be combined in powerful ways, and that users of the two standards can move data from one standard format to the other fairly easily.

5. Conclusion and Future Work

This research aimed to introduce and to display the user-friendly framework for metadata and microdata documentation in PCBS based on international standards DDI and DCMI, the features of these standards and relations with other standards like SDMX. We introduced also dissemination surveys using national data archive (NADA).

Future work includes extending our framework and using it in other ministries and agencies to build centralized nada portal for all documented surveys, this will enhance and support the national statistical system (NSS) and the national strategy and will improve the documentation and dissemination policy.

References

[1] Palestinian Central Bureau of Statistics (PCBS): http://pcbs.gov.ps/site/lang__en/1/default.aspx

[2] Data Documentation Initiative (DDI): Available at: http://www.ddialliance.org/

[3] Dublin Core Metadata Initiative (DCMI): Available at: http://dublincore.org/

[4] The Statistical Data and Metadata Exchange (SDMX): Available at: http://sdmx.org/

[5] The IHSN Metadata Editor, also known as the Nesstar Publisher: Available at: http://www.ihsn.org/home/software/ddi-metadata-editor.

[6] IHSN-developed National Data Archive (NADA): Available at: http://www.ihsn.org/home/software/nada.

[7] M. Vardigan, P. Heus, W. Thomas, Data Documentation Initiative: Toward a Standard for the Social Sciences, The International Journal of Digital Curation. ISSN: 1746-8256 (2008), Vol. 3, No. 1, pp. 107-113.

[8] A. Gregory, P. Heus, DDI and SDMX: Complementary, Not Competing, Standards, Open Data Foundation (2007).