Project N°: 262608

**ACRONYM: Data without Boundaries**

## Integrated DELIVERABLE D7.2 - D7.3

Standards with future relevance for
European Social Science data infrastructure

Needs, Key Areas, Rules & Best Practices
in Metadata Standard selection and usage

## Work Package 7

Standards Development

| | | |
|---|---|---|
| **Reporting Period:** | **From: Month 18** | **To: Month 36** |
| **Project Start Date:** | **1st May 2011** | **DURATION: 48 Months** |
| **Date of Issue of Deliverable:** | **30th July 2014** | |
| **Document Prepared by:** | **26, 7, 15, 2, 9, 5** | **SCB, UGOT-SND, DESTATIS, UTA-FSD, NSD, GESIS** |

**Combination of CP & CSA project funded by the European Community**

**Under the programme "FP7 - SP4 Capacities"**

**Priority 1.1.3: European Social Science Data Archives and remote access to Official Statistics**

*The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 262608 (DwB - Data without Boundaries).*

---

This document has been prepared by: Claus-Göran Hjelm, task leader (SCB), Stefan Ekman, task leader (UGOT-SND), Hans Irebäck (SCB), Caspar Jordan (UGOT-SND), Björn Sjögren (UGOT-SND, Maurice Brandt (DESTATIS), Anja Croessmann (DESTATIS), Thomas Helmcke (DESTATIS) Mari Kleemola (UTA-FSD), Katja Moilanen (UTA-FSD), Ørnulf Risnes (NSD) and Uwe Jensen (GESIS).

# TABLE OF CONTENTS

**LIST OF FIGURES**



**LIST OF TABLES**

# 1 INTRODUCTION AND BACKGROUND

In DwB D7.1, in which the state-of-the-art in metadata usage in NSIs and DAs was reviewed, it was concluded that there are further needs for extensions and/or modifications of current standards as well as new issues that will become some relevance in the future.

Support for searching and locating OS microdata at NSIs is by far not as developed and coordinated at a European level as for research data at DAs. No structured metadata base exist that present national OS microdata from the NSIs at a European and international level, even if there are some degree of integration via Eurostat/European Statistical System, (ESS)[1], by which datasets from different countries are integrated through gathering data from common European Surveys coordinated by Eurostat/ESS.

This report is to identify and to discuss such subjects and questions about "future needs", "best practice" and "open issues" as related topics within the programmatic frame of WP7 on "Standards development". The report integrates the former separated deliverable DwB D7.2 and DwB D7.3 to achieve most possible coherence and integration in answering leading questions in the field and tackling respective work package tasks. Thus DwB D7.2/3 emphasizes the need for standards development from different points of view, and results from the work in T7.2, T7.3, T7.4 and T7.5, which is detailed in the following.

The audiences for this report are the statistical institutes and government agencies that would like to facilitate transnational discovery and access to official microdata.

## 1.1 Description of work

The Description of Work (DoW) concerning the following tasks T7.2 to T7.5 defines the basis of the work and respectively those research questions to explore, to identify and to answer in deliverable DwB D7.2/3.

> Task 7.2 will set up certain assessment criteria to establish which metadata standard meets the majority of needs and which related vocabularies and coding schemes may be beneficial across all sectors
>
> and task 7.3 is meant to explore and define a set of standards with future relevance for European social science data infrastructure needs, and to make an assessment of the different [standards'] applicability to specific purposes.
>
> For T7.4, the aim is to identify and discuss key areas, where the NSIs and data archives have issues that are not sufficiently covered by present standards
>
> and, for T7.5, to define rules and best practices for key areas of metadata standard selection and usage.

Based on these tasks, this report, combining deliverables 7.2 and 7.3, has the following descriptions in the DoW:

[1] http://epp.eurostat.ec.europa.eu/portal/page/portal/pgp_ess/about_ess

DwB D7.2. Standards with future relevance for European Social Science data infrastructure needs and key areas:
Report on set of standards with future relevance for European Social Science data infrastructure needs, and [an] inventory of key areas not sufficiently covered by present standards.

DwB D7.3. Metadata standard selection and usage – Rules and Best Practises:
Report on specific rules and best practices for key areas of metadata standard selection and usage.

The distinction between the work in T7.3 and T7.4 is not obvious. To find standards of future relevance, one has to look at a longer perspective to find what kinds of functionality that has to be supported by standards in the future, and which those standards are. The "key areas not sufficiently covered by present standards" can be identified by finding obvious issues at the NSIs and DAs. Robust standards for supporting those key areas are also most likely standards of future relevance, for example standards for versioning research data, a functionality not yet sufficiently covered by the standards in use.

This report deals with utilizing standards in building a federation that supports infrastructure for transnational discovery and access to European OS microdata in distributed resources. More specifically, the focus will be on metadata standards and controlled vocabularies for agencies that would like to facilitate transnational discovery and access to official microdata. For example, the producers of OS microdata and other agencies that are planned to cooperate with the data producers with search and access issues related to OS microdata. The set of involved agencies can differ, given what solution is chosen to accomplish discovery and access to the OS microdata.

## 1.2 Research questions and problem areas

Based on the work in T7.1 and DwB D7.1 and work in other parts of the DwB project and other research projects the following list of research questions was identified:

- Which standards are available for cross section use (NSI, NSA DA)?
- What is missing in available standards (mainly DDI, SDMX, ESMS)?
- What are the minimal requirements for a standard of future needs?
    - Interoperability,
    - Versioning,
    - Support for vocabularies,
    - Machine readability,
    - Business models
- How can solutions built on internal models and standards be mapped to a common standard?

In addition to these research questions, the following questions were also identified as relevant for this report:

- How can assessment criteria for metadata standards be established, and which would such criteria be?
- How to define metadata for/from administrative sources and Big Data?

- How do we handle RDF and Linked Data from a metadata perspective?
- How do we meet the need of metadata for process data?
- How do we deal with the lack of common terminology/semantic integration (further elaborated below)?
- How to make the metadata connectors between the layers of information, metadata and access?
- What particular key areas hold problems that require the implementation of rules and/or best practices and to what extent are there already rudimentary rules or best practices in place? (Controlled vocabularies and coding schemes should be taken to be included in the above.)

## 1.3 Metadata issues discussed in other DwB work packages

Beyond the specification of respective research questions WP7 has also to consider related work of further DwB work packages[2], involving issues about metadata and standards as well.

**DwB WP5 - "Servicing European Researchers in the use of Official Statistics Microdata" –** provides first overview of available national microdata in Europe with the metadata base CIMES (Centralizing and Integrating Metadata from European Statistics). The dedicated metadata schema of CIMES documents national OS microdata (D5.2). The schema describes necessary metadata about study series, singular studies, and adhered datasets. The documentation schema is compatible with DDI-Codebook and DDI-Lifecycle and was developed in close collaboration with WP8 and WP12. A report on a "Databank documenting integrated EU Official Statistics data" at Eurostat (MISSY metadata base) is forthcoming.

The work programme of **DwB WP8** concerns **"Improving Resource Discovery for Official Statistics Data"**. Metadata needed to support search and discovery across NSIs and DAs were reported in D8.1 "OS Object Model" based on DDI metadata available and suitable for describing disparate resources of OS data. Furthermore WP8 submitted a "Metadata model" (D8.2), which gives as well an overview of relevant standards (DDI and SDMX) and available controlled vocabularies. The report discusses also what is lacking, e.g. the linkage between aggregate data and microdata. D8.4 propose "... portal resource discovery functionality for a search/ browse portal interface". The report is based on user stories, which describes the functional requirements for the user interface of the portal to support resource discovery of Official Statistics.

Of interest for developing subjects and considerations in DwB D7.2/3 is the work in **DwB WP12** in charge of **"Implementing Improved Resource Discovery for Official Statistics Data"**. The published report D12.1 provides in chapter one "Metadata model and standards" an overview of developments since DwB started. Further topics regard e.g. the DDI Discovery RDF as a starting point, harvesting of ingested metadata and related standards to consider and finally the requirement of a minimal set of metadata to apply in such discovery

[2] http://www.dwbproject.org/about/deliverables.html

systems. The forthcoming report D12.2 will include a chapter about metadata processing and enhancements.

## 1.4 Outline of this report

Chapter 1 provide background information on the scope of work undertaken with this report and relationship to metadata issues discussed in other DwB work pages.

In chapter 2, we give a brief overview of the current situation, including the importance of a conceptual model, and looks at possible criteria for assessing standards.

Chapter 3 discusses important areas that, as far as we can see, are not covered by present standards. This regards selected metadata topics related with Big Data, Administrative Data as well as metadata on Versioning and Data Processing.

Chapter 4 raises the development of standards in the light of future developments. New topics in metadata development concern in particular DDI, SDMX. Respectively new challenge of Linked Data Initiative and application of RDF metadata are of high relevance for future planning of metadata systems and vocabularies.

Chapter 5 approaches, in the light of the previous chapters, related issue of rules and best practices.

Chapter 6 takes an analytical view and discuss the topics raised in the report at chapters two to five.

Chapter 7 concludes major finding as executive summary.

**Note on reference list**

To ease use of referenced information, in particular those on metadata standards, we organized the reference list by additional sections, when we use larger amount of references in the same metadata context. This regards information on ESSnet, statistical metadata from UNECE, DDI and presentations on Indexing and Classifications at several IASSIST conferences.

**Note on terminology and abbreviations**

Much of the terms used in this report have a variety of meanings depending on whether they are employed by NSIs or DAs, in one country or another, within one discipline or another. We have tried to provide the definitions that we use in the text.

For the reader's convenience, we have also added a Glossary of abbreviations at the end of this report.

# 2 PRESENT STANDARDS

## 2.1 Different standards for different purposes

Metadata standards that are used within NSIs, NSAs and DAs mirror procedures that are in focus in different reference frameworks and models that are followed within NSIs and DAs, and thereby they have native/built-in similarities and dissimilarities. The reference frameworks are results from efforts to describe organisations that traditionally have different purposes.

There are two metadata standards that are followed by a majority of the European NSIs and DAs. These are SDMX metadata standard and the DDI metadata schema specification. There are various reasons to why those two metadata standards are widely accepted.

The historical background that explains the driving forces for their development is described briefly here.

### *2.1.1 DDI and SDMX*

Deliverable 7.1 suggested that DDI and SDMX, as well-established metadata standards, could be extended to support relevant parts of the data management for OS microdata. In this section the differences between the two standards are described further and some conclusions are drawn regarding best use.

Even though Eurostat (CROS Portal. 2013. ESSnet on SDMX Phase II)[3] has evaluated and considered SDMX for handling microdata (i.e. in the data collection process), the usage (existing or planned) relates to the subsequent processes (process, analyse, disseminate). For this purpose, SDMX provides a rich data model for aggregated statistics. DDI, on the other hand, has strong support for the processes handling microdata. As described in DwB D7.1, possible benefits of using both DDI and SDMX as complementary standards as a joint metadata solution have been identified (see also UNECE. METIS. SDMX DDI Dialogue[4]).

It is important to notice that the objectives for using SDMX internally primarily are related to making the internal production processes at an NSI more effective. There are examples of NSIs who are planning to implement SDMX in the internal production process, but a majority of NSIs only employ SDMX in the data exchange process for reporting to Eurostat and other international organisations.

From our perspective, metadata standards for the European Social Science area should primarily focus on metadata for microdata[5] . The routines and metadata standards used by an

---

[3] http://www.cros-portal.eu/content/sdmx-ii-finished

[4] http://www1.unece.org/stat/platform/display/metis/SDMX+DDI+Dialogue+-+Overview+Page

[5] In the DwB context, aggregate data do not raise similar issues for transnational access like micro data. However, social science research like many economists use aggregate data for which metadata is also important. The chapter “Standardisation and Harmonisation” discusses related aspects.

NSI in the internal production processes for aggregating and disseminating statistics are, in this context, less interesting. In the DwB Deliverable 8.2, a scenario is mentioned, where ("ideally") a portal will index both microdata and aggregated datasets to help researchers to locate the microdata needed for research (p. 33). However, it is important to stress that statistical products/outputs are aggregated data that could have been produced from a variety of sources, such as survey data and register data, and some variables could have been derived from other variables. Very few NSIs have a metadata system that supports integrated documentation of the whole production process so that it is possible to link the aggregated output to the source variables on the microdata level.

Considering the above conditions, it seems like the best way forward in the short and medium term is an approach where internal standards are mapped to an agreed standard. As our main focus is on microdata, this agreed standard should be DDI.

For NSIs with internal metadata systems, it should be feasible to map their internal metadata to DDI. By doing so, some additional advantages will emerge, such as the possibility to use solutions and components that have been and will be developed by different communities. One example is the DDI-RDF Discovery Vocabulary for documenting research and survey data described later in this report.

### *2.1.2 Harmonisation and Standardisation on national and international level*

The Eurostat reported 2012 on the "Analysis of the future research needs for Official Statistics" (Eurostat. 2012)[6]. The report highlights metadata and standard related results from the ROS survey (Research needs in Official Statistics) under methodological and technological aspects:

> "Standardisation and harmonisation are important topics which are mentioned very often. These are seen as conditions for spreading collaborations within the European statistical system, e.g. the harmonisation of definitions and methods. The proposition to integrate between Statistical Data and Metadata Exchange (SDMX) and Data Documentation Initiative (DDI) has the same direction of impact." (Ibid, p. 51)
>
> "It is obvious that harmonisation is an important topic here as well. The target should be to promote technologies which help to standardise data exchange in order to enable data sharing and enhance comparability (common architecture). The data documentation initiative (DDI) which is oriented towards a specification of metadata in social and behavioural sciences is an example for harmonisation of information on data. Harmonisation and user-friendliness of software was also pointed out. Dynamic and flexible statistics via internet technologies raised interest here as well." (Ibid, p. 52)

Aspects of harmonisation and standardisation of metadata can be discussed under several aspects of interest. As it regards 'scientific content' of metadata organised by data documentation standards, it primarily a substantive social science research endeavour to harmonise such (national) content (including respective reporting) to enhance comparability of

[6] http://dx.doi.org/10.2785/19629

related OS data for itself. This issue is outside the scope of the reports as a whole. However, it even necessary to follows for example the discussions on harmonisation needs in the ESS research community to explore issues on metadata and standards development as general perspective in this chapter.

In this context is to bear in mind that for example the variable definitions at NSIs are to a large extent not harmonized. That holds true even for the definitions inside the respective NSIs where different definitions of the same variable in many cases have been created for different statistical products. The exception from this fact is statistical products that are regulated on the EU level.

> One example: the Household Budget Surveys in the EU[7] where there exists methodology and recommendations for harmonization[8] as well as an agreed classification (COICOP, Classification Of Consumption by Purpose[9]) (see Eurostat. 2007. 2010).
>
> An example from Statistics Sweden shows that as many as 90 statistical products have some kind of EU regulation. However, that does not mean that all the variables in these statistical products are harmonized but there is an ongoing work on harmonizing in population and welfare as well as in economic statistics.

In 2007, a Task Force on Core Social Variables (Eurostat. 2007; 2011[10]) presented a proposal for a limited number of core statistical variables for introduction in all EU social surveys. The objective is to make it possible to produce statistics, which are comparable across countries and across domains for different subpopulations. Related surveys of interest are among others: Labour Force Survey, Survey on Income and Living Conditions (EU-SILC), Household budget survey, Time Use Survey. Within the framework of EU cooperation Eurostat has together with the member states taken a decision on a gentlemen's agreement on the use of 16 core variables together with methodological guidelines, which were updated 2011 (Eurostat. 2007, 2011, p. 3).

For economic statistics standardised core variables has been one of the main objectives for the MEETS project "Modernisation of European Enterprise and Trade Statistics"[11]. Overarching objectives are

- The development of target sets of indicators for new areas and a review of priorities (objective 1).
- A streamlined framework for business-related statistics (objective 2).
- The support for the implementation of a more efficient way of producing enterprise and trade statistics (objective 3).

---

[7] http://epp.eurostat.ec.europa.eu/portal/page/portal/household_budget_surveys/Data/database

[8] http://circa.europa.eu/Public/irc/dsis/hbs/library?l=/rounds2005/transmission_2005pdf/_EN_1.0_&a=d

[9] http://unstats.un.org/unsd/cr/registry/regcst.asp?Cl=5

[10] http://epp.eurostat.ec.europa.eu/portal/page/portal/information_society/documents/Tab/CORE%20VARIABLES%20UPDATED%20GUIDELINES%20May%202011.pdf

[11] http://www.cros-portal.eu/content/37-meets

- Modernisation of the data collection system on trade in goods between Member States (Intrastat) (objective 4).

The MEETS programme (CROS Portal. MEETS) ran over a period of five years from 2009 to 2013 organised by 25 singular MEETS projects, which were carried out by ESSnet's, universities and statistical authorities with several subjects.

In the background description of the "Consistency", project (CROS Portal. Consistency)[12], which is of interest in the context of developing standards in the ESS, one can find the following analysis about the "Need for the project" to enhance the comparability of statistical data from different domains:

> … The development of the European Statistical System did not occur 'in one pour': different stakeholders, interests and user needs were involved; the Member States also had different views, history and experience. A comprehensive coordination of all these developments was not possible. Furthermore, the concepts and methods also evolved over time so that one specific statistical domain developed in the past might now be seen from a different perspective. Finally, not all statistical development took place under the ESS umbrella. … It is thus not surprising that concepts, definitions and methodologies as well as practices vary to some degree over the different statistical domains. This leads to a situation where the statistical outputs of these various domains cannot be compared because of the application of definitions, concepts and methodologies which are partially or even totally different. The user – who cannot (fully) compare the statistical data – will criticise this fact as a lack of (full) coherence."[Ibid]

The project considers that, from a high-level conceptual point of view, two kinds of consistency are relevant within the European OS data context, to enhance the comparability of data from different statistical domains.

- "Horizontal consistency

  Horizontal consistency refers to the comparability between the various statistical domains. Data between statistical domains can be compared if they are elaborated using the same statistical unit, the same coverage, the same classifications, the same definitions, the same frame and the same reference time and period. This is also valid as concerns the relations between monthly or quarterly data and the respective annual data.

- Vertical consistency

  Vertical consistency is the issue of comparability between the sum of Member States data and the European aggregate. Concepts developed for the national implementation may not be suited to derive the consistent European aggregate on the basis of such Member States data. This may occur in statistical domains where the statistical objects are of a cross-border nature."[Ibid]

[12] http://www.cros-portal.eu/projectdetail/1400

Work Package 3[13] of the Consistency project (being part of the MEETS programme) carried out their work with the objective to improve consistency by a proposal for better definitions of important variables. This includes

- a system approach (change from stovepipe approach to system approach),
- a comprehensive metadata system and
- a system of variables (as subsystem of the metadata base) allowing for unique coding system and standardised definitions.

A first list of variables and considering related metadata was put forward for further discussion and decision (CROS Portal. Consistency.WP3 minutes, p. 2-3)[14]. The proposal includes several (content and reference) metadata to document variables in the variable system:

- Standardised general definitions
- A unique variable identifier
- A unique variable name
- Objective and definition of the variable.
- Domain using the variable
- Relations to other variables
- Relation to other sub-systems of metadata
- Relations to National Accounts and to Accounting standards
- Description of already known consistency issues.

A new legal framework at the system level must also accompany all these activities (not on the level of the individual statistical products that is the case today). This includes legal acts providing a methodological and conceptual fundament for surveys and data compilations as well as legal acts governing surveys, data collections and statistical indicators. All these acts refer to and make use of the fundamental regulations, adding additional concepts by defining target populations, time schedules, defining variables etc.

These considerations on developing standardisation and extension of (OS variable) data/metadata are accompanied by the work and recommendations of the topical driven ESSnet project Standardisation (CROS-Portal. Standardisation[15]). Scope of this activities 'sponsoring standardisation' concern also development of metadata standardisation and related conceptual issues like

- data modelling,
- work-flow and process description,
- input/output harmonisation

involving respective existing metadata frameworks. Thus this work is undertaken in relationship to international standardisation activities (Australian bureau of Statistics, SDMX) and in cooperation with UN driven activities on GSBPM, GSIM and CSPA (Common Statistical Production Architecture).

---

[13] http://www.cros-portal.eu/content/final-workshop-essnet-consistency-wp3

[14] http://www.cros-portal.eu/sites/default/files//Del_32_WP3_Final%20workshop_Proposals_2.pdf

[15] http://www.cros-portal.eu/content/ess-standardisation

## 2.2 The importance of a conceptual model

The ESS discussions about Standardisations and Harmonisation to enhance comparability of OS data highlighted the relevance of conceptual models and metadata frameworks in the development of related standards for social science data in general.

A conceptual model consists of the following components:

- A set of concepts, representing real world objects sharing structure and semantics.
- Each concept has a name and an optional set of properties.
- A set of relationships, representing semantic relationships between objects (Boukottaya, A. et al. 2004).

An important advantage of a suitable conceptual metadata model for European statistical institutes and agencies producing official microdata is improved communication, i.e. shared understanding, of these metadata. This improvement concerns communication

- among metadata users
- among metadata producers
- between metadata users and metadata producers,

where metadata users can also be metadata producers at the same time: For example, within an NSI production process metadata that comes from the first part of processing may be used in later processing stages.

The general means for achieving this shared understanding is the metadata comparability resulting from the model. It is therefore important that a conceptual metadata model supports the harmonization of metadata, including

- avoiding different terms for the same things by controlled vocabularies, by common classifications and by a framework of definitions,
- harmonizing metadata from different data producers, like NSIs, NSAs , DAs and other data producers.

Objects of such harmonization should be

- descriptions of the relationships between various types of metadata, like between variables and their possible values,
- descriptions of the data production processes,
- descriptions of ongoing data production processes, for example by statistics on imputations,
- data quality reports,
- structures of the textual part of information on the data,
- the kinds of documents to be included in the description of the data,
- and the independence of formats and technical implementation concerns.

Other particular means for improving metadata comparability are metadata standards. It might therefore be useful for the model to include the use of such standards.

The improved communication possible through metadata is an important advantage. It leads to better understanding and thus reduces misunderstandings and mistakes. This increases the quality of production and therefore of products, in several aspects:

- Researchers, in their role as *metadata users*, increase the quality of their research and therefore of their research output.
- Researchers, in their role as *metadata producers*, increase the quality of their research.
- Metadata producers, such as NSIs, NSAs and DAs, increase the quality of their products in the sense that their products better fit the research requirements.
- Metadata producers, in their role of *metadata users*, especially within their own producer institution, increase the quality of their production processes by using the model-based metadata as an input into the processes.

For conceptual metadata models, you can summarize in general:

A conceptual metadata model, independent of implementations and formats, will promote a understanding of metadata. It should provide a common and extensible semantic framework, a common and accepted language for domain experts that any metadata can be mapped to. It would thus be possible to turn the disparate, localized metadata of today into a coherent, high-quality and valuable global resource.

### *2.2.1 The layered model*

This report addresses metadata and standards issues also to support transnational discovery and access to European official statistics microdata.

In the process of finding metadata standards and recommending best practices to implement and apply them, it is essential for the data providers to be aware of researchers' needs.

However, metadata is created often only for producers' necessities, which can differ greatly from the requirements of those pursuing scientific goals. This is mainly because official surveys are based on legal regulations and the dissemination of the results is done primarily to the legally determined authorities or to the public.

The possibility of using confidential data for scientific purposes has been expanded only during the last few years. The recent Commission Regulation (EU) No 557/2013 (17 June 2013) says the following on this aspect:

(1) 'confidential data for scientific purposes' means data which only allow for indirect identification of the statistical units, taking the form of either secure-use files or scientific-use files;
(2) 'secure-use files' means confidential data for scientific purposes to which no further methods of statistical disclosure control have been applied;"
(3) 'scientific-use files' means confidential data for scientific purposes to which methods of statistical disclosure control have been applied to reduce to an appropriate level and in accordance with current best practice the risk of identification of the statistical unit; (European Commission. 2013. p. L 164/17)[16]

[16] Commission Regulation (EU) No 557/2013 of 17 June 2013. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2013:164:0016:0019:EN:PDF

For the different data providers of a nation one challenge is to offer standardised information for the scientific community with consideration of the peculiarities of each statistic. A second challenge is to harmonise these statistics across the different countries in Europe, also providing metadata about the harmonisation process.

There are different levels of information that researchers need to access during the course of the research process. Figure 1 shows a layered model of three levels of researcher needs: the general description, the metadata itself, and eventually real microdata access.

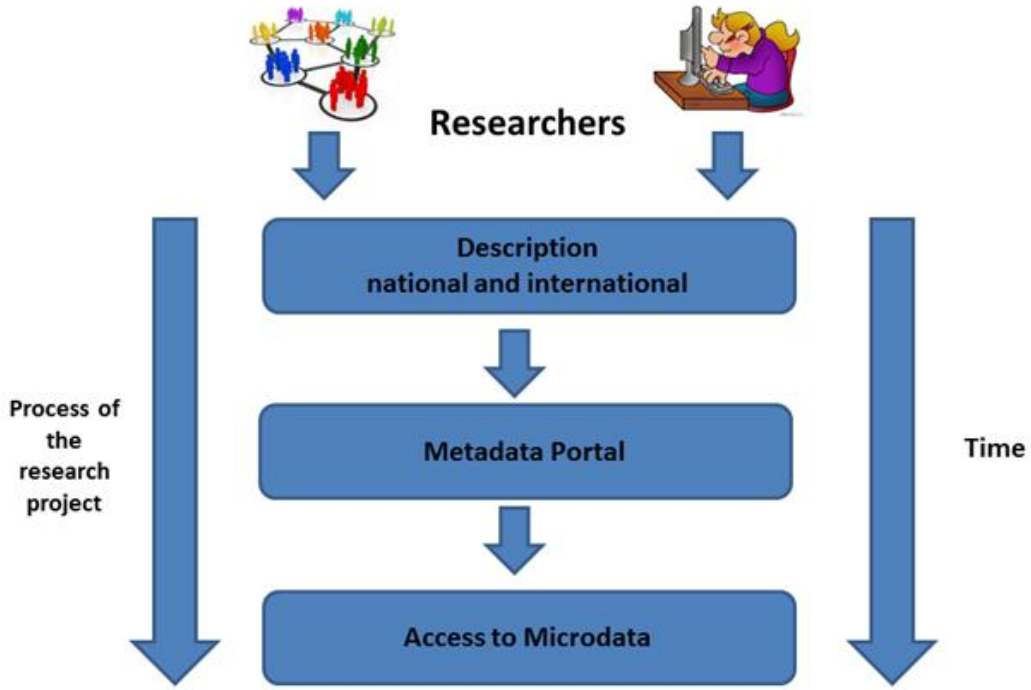


Figure 1: The three-layered model (by SCB/Destatis)

**Description - national and international level**

As part of surveying the field of their study, a research project will first need to have access to general information about what data are available. If researchers want to analyse microdata comparatively across Europe, a preliminary step is to know what data are available on national and European level. There has to be a description about the different ways of access to the microdata, and, depending on the way of access, an explanation of the different places where data can be accessed. To make microdata access as transparent as possible, an instruction of administrative procedures (such as contact people, legal issues, application forms, contract forms, a specification of fees) should be easily found and, if possible, offered

in a harmonised fashion[17]. One main requirement for international comparative research is to find all this information in a common, understandable language with a shared vocabulary and definitions.

**Metadata**

Having identified what data is needed for a particular research project, and thus carried out a first sifting of available data, the researcher now needs more detailed information. Documentation of the datasets in the form of metadata is essential for a researcher who means to carry out a comparative and/or secondary analysis. Such metadata has to capture all relevant parts of the research process, including the provenance of the dataset.

As described in DwB WP7 deliverable D7.1 (2013. pp. 13-15), there exist a number of definitions, typologies and/or nomenclatures of metadata used to organize the meaning, application or other metadata issues in differently specified contexts and communities at large.

The focus of this report is primarily, but not exclusively, on the following types of metadata, which are considered to be relevant to facilitate the finding, accessing and analysing of OS data.

- *Descriptive metadata* include descriptions of the produced data by a certain study and its producers (e.g. references to the producer, survey title, survey methodology, target population, geographical area, survey period, questionnaires, and further context information in terms of methods reports or codebooks). Thus these metadata act at the same time as reference metadata in retrieval systems to supply researchers with basic information about data and their context.
- *Structural metadata* deliver information about the relationship between items on variable level (e.g. variables, values, coding), the survey program, datasets, codebooks or syntaxes.
- *Process metadata* inform about the several facets in data processing ranging from creating the needed data file structure and respectively standardised variable definitions and/or controls, cleaning, imputations, calculations or harmonisation of data up to the integration of variables from different sources to a new data set.
- *Administrative metadata* regard information on file (data, documentations) and rights management like user access rights, study number, version number, persistent identifier, or (standardised) file names and format information.

It is worth noting how researchers have stressed the importance of metadata that provide context to the microdata – the provenance of the datasets. A definition of provenance useful in this context is that of the W3C PROV Data Model, a conceptual model that forms the basis for the W3C provenance specifications:

- "[Provenance] is defined as a record that describes the people, institutions, entities, and activities involved in producing, influencing, or delivering a piece of data or a thing. In

[17] CIMES (Centralising and Integrating Metadata from European Statistics) metadata base is a first attempt to provide such an overview. See DwB D5.2.

particular, the provenance of information is crucial in deciding whether information is to be trusted, how it should be integrated with other diverse information sources, and how to give credit to its originators when reusing it."(W3C Recommendation. 2013. PROV-DM, section 1, 1st. paragraph)

In the FP7 project Europeana Cloud, the need for provenance metadata for data material was particularly stressed by the participating researchers at the Expert Forums, as is evident from the respective Forum reports (Europeana Cloud. 2014. D1.5).[18]

The types of metadata mentioned here are important to varying extent for researchers and the requirements of a specific research phase like to find, review, and access existing data up to the preparation of the microdata analyses and the interpretation and publication of their results. Providing harmonised metadata standards will greatly alleviate data access for a transnational comparative research project or, indeed, make it at all possible.

**Microdata Access**

Access to the general information and the metadata should be managed through a joint platform where, optimally, every data producer or data provider provides the description on its own microdata. Presently platforms[19] already exists, e.g. for the access to European integrated data via Eurostat[20] or for the access to international public use census microdata via IPUMS international[21]. Examples of access channels include

- Public-Use Files,
- Scientific-Use Files,
- Secure-use files via Remote Execution,
- Remote Access or safe centre.

DwB WP5, WP8 and WP12 are currently developing an integrated service portal to provide researchers a single contact point for a transnational microdata access. Here, it is important to ensure the actuality of the provided information. It has to be defined who updates the metadata as well as when updates are due. Essential is also the provided technology as standardised software solutions (machine-actionable metadata-driven systems) and the possibility of metadata transmission.

### *2.2.2 GSBPM and GSIM*

The Generic Statistical Business Process Model GSBPM (UNECE. METIS. 2013a GSBPM) [22] and the Generic Statistical Information Model GSIM (UNECE. METIS. 2013b GSIM) are models

---

[18] These reports are part of the forthcoming Europeana Cloud Deliverable D1.5; see p. 8 in part 1; p. 23 in part 2; and p. 9 in part 3. Project page: http://pro.europeana.eu/web/europeana-cloud

[19] Currently, Eurostat and IPUMS do the work based on metadata provided by the different producers.

[20] http://epp.eurostat.ec.europa.eu/portal/page/portal/microdata/introduction

[21] https://international.ipums.org/international/

[22] www1.unece.org/stat/platform/display/metis/The+Generic+Statistical+Business+Process+Model

and frameworks primarily intended for making the internal production processes at NSIs and NSAs more effective (that is, industrializing the statistical production process). GSIM relates to GSBPM sub-processes in the way that it supports the definition of the information objects that flow between sub-processes.

Important advantages of combining GSBPM and GSIM that are highlighted by the GSIM community are how such as combination could:

- "Create an environment prepared for reuse and sharing of methods, components and processes
- Provide the opportunity to implement rule based process control, thus minimizing human intervention in the production process
- Facilitate generation of economies of scale through development of common tools by the community of statistical organizations."
  (METIS. 2014 GSIM Communication HLG. p.7) [23]

GSIM does this by defining objects related to statistical production, regardless of subject matter.

Another benefit of GSIM that is emphasised is that it provides a common language to improve communication at different levels:

- "Between the different roles in statistical production (business and information technology experts)
- Between the different statistical subject matter domains
- Between statistical organizations at national and international levels." (Ibid)

**GSBPM and GSIM in the context of WP7**

As both GSBPM and GSIM are focused at the internal production processes and information flows at NSIs, the question of what relevance they have for microdata access might be justified. GSIM is a conceptual framework, which means that implementations could be done with various different standards. In the context of access to microdata, it is not essential that every actor (NSIs, NSAs and DAs) use the same model/standards in their internal productions processes. The important thing is that each actor has an interface that can produce machine-readable output according to an agreed-upon standard. Statistics Sweden's metadata system (MetaPlus), for instance, is based on the Neuchatel model (ISO 11179) but it would be easy to map output from the system to any other agreed standard (e.g. DDI or some RDF-based format).

Therefore, in the context of WP7, the added value of GSBPM and GSIM is the possibility to improve communication at different levels by providing a common terminology for the basic information objects. For example, this could help to avoid the confusion that sometimes occurs between NSIs and DAs when they use different definitions or terms, such as "survey" and "study" or "universe" and "population".

[23] http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/2014/ECE_CES_2014_2-Generic_Statistical_Information_Model.pdf

A condition for this is that each partner involved agrees to implement GSIM as his or her standard conceptual model; this corresponds to the information level in our three-layered model.

For all other means of access to metadata and data, the implementation standards below the level of GSIM (metadata standards, vocabularies etc.) will be more relevant and necessary in a WP7 context.

## 2.3 Quality measurement and assessment criteria of metadata standards

This chapter will examine some different perspectives on quality assessment of metadata standards, looking at a number of attempts at setting up criteria for such assessment. There will then be a synthesis of the relevant parts from the examined perspectives in the form of a list of questions that need to be answered in the process of assessing any potentially useful metadata standard.

Assessment criteria comprise a collection of measures with which one can judge how well a defined purpose is achieved. In our case, we are proposing assessment criteria for metadata standards, and the purpose to achieve is fulfilling "the majority of needs" (DoW) that in DwB D7.1 (section 6.4) for NSIs and NSAs include retrieval, access and provision of research data and their respective metadata:

- To search for research data covering topics of interest on national or European level.
- The scope of data types range from a broad scope of aggregated data (statistics, indicators) to the underlying microdata from surveys, register data or process data and alike.
- Provision and retrieval of concepts is a key issue in dissemination to find and access relevant statistics at Eurostat or National Statistical Institutes for research purposes.
- With the provision of official statistics, metadata on the provided statistics has to cover in particular descriptive metadata and respective documentations of the used statistical concepts and definitions as well as metadata on the quality of the data.
- Furthermore, the provision of statistics and microdata must inform on confidentiality issues and access conditions according to data privacy regulation with respect to available dissemination and access channels (Public-Use Files, Scientific-Use Files, Secure-Use via Remote Execution or Remote Access or safe centre).

And for DAs:

- Public data services and related dissemination channels start for DAs in general from the data holdings catalogue of the national data archive, which have their (partly) integrated complement with the CESSDA data portal. With the study description scheme there exists the "container" to retrieve archived studies, datasets and documentations materials related to the study.
- The retrieval regards different types of majorly survey data and some statistics from complex microdata (like aggregate data from trend surveys, panels data, or alike) or partly also qualitative data.
- To support the retrieval of these data via the study description on European and national level thesauri vocabulary made available with the CESSDA topical classifications and ELSST thesaurus.

- Metadata for versioning (version history etc.) and persistent identifiers to support data citation and re-use of e.g. survey data (this is also of interest regarding disseminated OS microdata).
- The provision and access to survey data in relation to data privacy regulation is generally based on factual anonymised datasets (PUF/SUF).

  Few CESSDA members (UKDA with SDS in UK and RQ with CASD in France) provide access facilities to Secure-use files from OS via safe-centre environments.

  UKDA, RQ, DANS and SORS also provide access to SUF from NSIs, NSAs and support the documentation of OS data in cooperation with their NSIs / NSAs.

The importance of any single metadata element depends on how, where and for what purpose the metadata are to be used, and what are the goals that need to be achieved. For example, metadata that are important for long-term preservation could conceivably be useless for today's researcher.

In an ideal world, an organisation would be able to spend unrestricted amounts of time and resources to evaluate various different metadata standards in a detailed way. In reality, there are budget constraints, so simple yet sufficiently extensive criteria are needed. The criteria presented here will facilitate assessing the usability of a metadata standard on the conceptual level. Further work is necessary in order to create extensively detailed, task-specific criteria.

Zhu and Harris (2011, p. 129) explore the quality of data standards, defining a data standard as *"metadata that specifies the characteristics of data elements and their relationships"* and concluding that data quality concepts also apply to data standards. They also note that the quality of a data standard is its fitness for multiple users to produce highly interoperable data, that is, the quality of a data standard can be indirectly assessed by assessing the interoperability of the resulting data. Thus, one way to assess the quality of, for instance, DDI or SDMX, would be to use their method to examine the interoperability of a sample of metadata instances.

Bruce and Hillman (2004)[24] propose a framework of seven quality parameters for metadata:

- Completeness, Accuracy, Conformance to Expectations, Logical Consistency and Coherence, Accessibility, Timeliness, and Provenance.

Although the Bruce-Hillman framework is designed for evaluating the quality of metadata, the main principles are useful when assessing a metadata standard.

Table 1 summarizes the framework and suggests how each parameter could be applied to metadata standard evaluation.

[24] http://hdl.handle.net/1813/7895

Table 1: The framework from Bruce and Hillman applied to metadata standard quality assessment

| Parameter | Bruce & Hillman (2004) | In metadata standard quality assessment context |
|---|---|---|
| Completeness | Two sides of completeness:<br>- The element set used should describe the resource as fully as economically feasible.<br>- - The metadata elements should be filled in. | Ø The metadata standard should contain all the relevant information elements that are needed to describe the resource. |
| Accuracy | The information provided in the values of the elements needs to be correct and factual. In addition, typographical errors should be eliminated, standard abbreviations used etc. | Ø The metadata standard should have a way to control the contents of the elements (for example, a mechanism to restrict the values of the elements to match a given regular expression). |
| Provenance | Knowledge about who prepared the metadata, the level of expertise they have, what methods are used to create and handle the metadata, and the history of the metadata instance. | Ø The standard should be able to capture information about the provenance of the data.<br>Ø The background of the metadata standard needs to be documented (who created the metadata standard, what are the governance procedures, organisational viability and stability, what methods are used to update and maintain the standard, history of the standard). |
| Conformance to Expectations | Metadata should contain the elements that the user community would reasonably expect to find there. An agreed-upon compromise that is well executed and documented is better than an approach that aspires to all things to all people and ends up poorly and unevenly implemented. | Ø The metadata standard should contain the elements that the user community could reasonably expect to find there. |

| Parameter | Bruce & Hillman (2004) | In metadata standard quality assessment context |
|---|---|---|
| Logical Consistency and Coherence | Elements should be conceived in a way that is consistent with standard definitions and concepts used in the subject or related domains. | Ø The metadata standard should support use of vocabularies. |
| Timeliness | Two aspects of timeliness:<br>- Currency: metadata should change whenever the described object changes.<br>- Lag: a complete metadata instance should be available by the time the described object is disseminated. | Ø The metadata standard should support updating and versioning of both metadata and the described objects. |
| Accessibility | Metadata that cannot be read or understood by users has no value. Obstacles may be physical or intellectual. Barriers to physical access include that metadata and described objects are not properly keyed or linked, or that metadata is unreadable for technical reasons. Intellectual access requires that there are guides and documentation | Ø The metadata standard should have a user community (a standard that is not used has no value).<br>Ø The standard should support using PIDs.<br>Ø The standard should be openly available.<br>Ø The standard should be well documented, and there should be practice guides available. |

Beall (2007)[25] proposes twelve "chief points of comparison" to be used in comparing metadata schemes (see table 2 for a summary).

While metadata schemes are schemes and not standards per se, they nevertheless build on a standard and constitute a physical and therefore to some extent measurable manifestation of it. They are thus a natural unit to which to apply criteria of comparison and assessment.

[25] http://docs.lib.purdue.edu/atg/vol19/iss1/7

**Table 2:** The twelve criteria proposed by Beall (Summary)

| Criterion | Description | Relevance for metadata standard quality assessment |
|---|---|---|
| Granularity | Some distinctions that are made in one scheme are not made in another one. For example, different types of authors (personal author, corporate author, conference author) are available in MARC but not in DC. | Ø The standard should make all the necessary distinctions. |
| Connection to Content Standards | Some standards come with an inherent set of controlled vocabularies whereas others require the user to choose relevant CVs. | Ø The standard should allow for the use of pertinent controlled vocabularies, either by incorporating them or by leaving the choice of CV to the user. |
| (Search) Tools | For some standards tools for searching already exist. Such tools may be more or less precise, allowing for varying quality of search results. There may also be other kinds of tools that facilitate implementation of the standard. | Ø Tools that fulfil the requirements of all potential users (researchers as well as metadata creators) should exist. |
| Specificity | Some standards are domain-specific and thus possibly more detailed in that particular domain whereas more general-domain standards may be less detailed. | Ø The standard should cover the relevant domains. |
| Interoperability | Interoperability includes the availability of mappings to other standards and of ways to harvest from the standard. | Ø The standard should be mapped to other relevant standards.<br>Ø The standard should be well-documented to allow its integration in larger automatised infrastructures. |
| Reputation | This includes popularity, often based on the results of earlier implementations. | Ø The standard should have a documented history of success in relevant fields. |

| Criterion | Description | Relevance for metadata standard quality assessment |
|---|---|---|
| Training Requirements | Standards that are linked to content standards (controlled vocabularies) may require much more training before they can be used correctly. | Ø There should not be an unreasonable amount of training required to be able to use the standard. |
| Organisational Criteria | Stability and activity of the organisation behind a standard are important for the standard to be held up to date. | Ø A stable and proactive organisation should maintain the standard. |
| Handling of Specific Functions | Standards cover different functions (discovery, long-term preservation, access rights, versioning etc.) to varying degrees. | Ø The standard should cover all relevant functions. |
| Adaptability to Local Needs | This includes the possibilities of adding custom fields to a standard and of not using certain parts of the standard. | Ø The standard should allow customisation without losing its integrity.<br>Ø Optionality should be a feature of most parts of the standard. |
| Scalability | The more detailed a standard is the better search results are achieved when the standard is applied to very large data collections. | Ø The standard should be detailed enough to provide the possibility of a unique description for every described unit. |
| Surrogacy | Surrogate metadata exists separately from (not embedded in) the object it describes and is thus more easily searchable and harvestable. Embedded metadata is safer as it is not easily separated from the object it describes. | Ø The standard should allow for both embedded and surrogate metadata. |

There is a lot of overlap between Bruce and Hillman on the one side and Beall on the other, even though their respective ways of categorising are rather different. Beall's criterion of "Handling of Specific Functions" can be taken to encompass a lot of what a metadata standard does and covers several of Bruce and Hillman's parameters. Nevertheless, there are some benefits in combining the two perspectives, especially when also taking into account other perspectives presented above.

The following list of questions tries to take into account all the points made in the different approaches to metadata standard quality assessment examined in this section, combining them into an organised set of questions that should be used systematically to assess the quality of any metadata standard under consideration.

The list does not, however, provide any methods or tools to be used in answering the questions. This has two reasons: firstly, the best ways to answer these questions are likely to be different depending on who is asking them and for what purpose, and secondly, coming up with these ways is beyond the scope of the present report.

**List of questions**

**Is the standard complete?**

- Does it cover the relevant domains?
- Does it make all relevant distinctions?
- Does it contain all fields that the user community expect?
- Does it cover all relevant functions?
    - Is there room for provenance data in the standard?
    - Does it support updating and versioning of both metadata and data?
- Is it detailed enough to allow for efficient retrieval?

**If not complete, does it allow for customisation?**

**Is the standard accurate?**

- Is there spell-check?
- Is the information content controlled, e.g. through use of controlled vocabularies?
- Are there optional fields (allowing more freedom, and risking useless entries)?

**Is the standard accessible?**

- Does it have a user community?
- Does it support PIDs?
- Is it openly available?
- Is it well-documented?

**Are the procedures involved in changing the standard openly available to stakeholders? Can stakeholders even influence them?**

- Is it mapped to other relevant standards?
- Does it allow for both embedding and surrogacy?
- Are there search and other tools available for the standard?
- Does it allow the data provider to choose which controlled vocabulary to use?

**Does the standard have a good reputation?**

**Are there documented cases of successful application in relevant fields?**

**Is there a stable and proactive organisation behind the standard?**

**Does using the standard require a lot of training?**

Of course, the success of a metadata standard will also depend on how well it is implemented. For example, a metadata record will not be useful if the creator does not actually fill in the relevant information in the available fields.

# 3 KEY AREAS NOT COVERED BY PRESENT STANDARDS

We have identified a number of key areas, and they will be described in no particular order in this chapter. It is worth pointing out that these areas concern metadata issues; other areas not covered by the present standards lie outside the scope of this report.

## 3.1 Big Data

**Definitional aspects**

A rough definition of Big Data could be vast amounts of data of which we do not have control over the origin, do not know how it has been processed and it is most likely unstructured.

Big Data usually includes datasets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process the data within a tolerable elapsed time. Big data sizes are a constantly moving target, as of 2012 ranging from a few dozen terabytes to many petabytes of data in a single data set.

In a 2001 research report META Group analyst Laney (2001)[26] defined data growth challenges and opportunities as being three-dimensional, i.e.

- (data) Volume, concerns increasing amount of data,
- (data) Velocity, regard speed of data in and out, and
- (data) Variety, covers the range of data types and sources.

Analyst groups like Gartner and the industry use this "3Vs" model for describing Big Data. In 2012, Gartner analyst Laney (2012)[27] updated the definition as follows:

- "Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization."

Others like IBM added additionally, a new V "Veracity", which includes "questions of trust and uncertainty with regards to data and the outcome of analysis of that data." (Ward and Barker. 2013)[28] If Gartner's definition (the 3Vs) is still widely used, the growing maturity of the concept fosters a more sound difference between Big Data and Business Intelligence, regarding data and their use:

- Business Intelligence uses descriptive statistics with data with high information density to measure things, detect trends etc.;

[26] http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

[27] http://www.gartner.com/resId=2057415

[28] http://arxiv.org/pdf/1309.5821

- Big data uses inductive statistics and concepts from nonlinear system identification to infer laws (regressions, nonlinear relationships, and causal effects) from large data sets to reveal relationships, dependencies and perform predictions of outcomes and behaviours.

Research funders like the NSF (National Science Foundation) promote an extended scientific view on the manifold dimensions of data intensive research and advancing big data sciences. They conclude unpretentiously:

> "The phrase "big data" refers to data that challenge existing methods due to size, complexity, or rate of availability." (NSF. 2014. P. 3)[29]
>
> As such data
>
> "come from many disparate sources, including scientific instruments, medical devices, telescopes, microscopes, satellites; digital media including text, video, audio, email, weblogs, twitter feeds, image collections, click streams and financial transactions; dynamic sensor, social, and other types of networks; scientific simulations, models, and surveys; or computational analyses of observational data.
>
> Data can be temporal, spatial, or dynamic; structured or unstructured; information and knowledge derived from data can differ in representation, complexity, granularity, context, provenance, reliability, trustworthiness, and scope.
>
> Data can also differ in the rate at which they are generated and accessed."[Ibid]

In what sense data intensive research and big data challenges affects NSIs and DAs is roughly explored in the following section.

**Big data in the Official Statics domain**

This chapter refers to Big Data from official statistics or producer of data sets that are so big and complex that they are not manageable on modern, local PCs, like for instance some health data or tax data, and they can be not analysed with traditional methods. Data from the private sector are not included, for instance from internet providers or search engines.

The official statistics are only partly concerned with Big Data. Some of the official surveys may meet the definition of Big Data, but the most of them do not (taking a particular definitional perspective). However, in the future the proportion of Big Data might increase, considering that also data from diverse statistical domain sources "challenge existing methods" (NSF. 2014) at large.

For example, RFID (Radio Frequency Identification)[30] codes will be used for the price statistics and the account information from enterprises will be extracted automatically from their financial controlling systems. This would enable real-time analysis but there is a lack of methods to analyse those data. How those data will be analysed might depend on the context and what information is needed, probably more and more with data mining methods.

---

[29] http://www.nsf.gov/pubs/2014/nsf14543/nsf14543.pdf

[30] http://www.centrenational-rfid.com/definition-of-rfid-article-71-gb-ruid-202.html

Especially in this constellation the metadata need to document the collection and source, the linking methods and the methods of analysis such Big Data as well.

In a way, the metadata become even more important than the data itself, because without good documentation it might not be possible to draw valid conclusions from the analyses. The Big Data is only the raw material where the information is drawn from. Not all dimensions of Big Data can be analysed simultaneously. The evaluation of the data might be selective and can change over time. This also needs to be documented in order to enable reproduction of results.

The particular design of metadata will depend on the developments in the nascent field of Big Data. It is worth bearing in mind that the metadata are at an early stage here, because they are more important for Big Data than for any other data type. At the moment, it is difficult to describe a concrete system for big-data metadata, but it is possible that the system of metadata will change in order to capture the whole picture of Big Data.

However, this is certainly a topic of future relevance, which researcher and staff at the data-producing authorities will have to deal with.

Heads of EU statistical offices recognise the relevance of Big Data for the European Statistical System (ESS) with the Scheveningen Memorandum (Eurostat. 2014a)[31]). The event "Big Data in Official Statistics" organised by Eurostat (2014b)[32], discussed corresponding challenges under several aspects. Among others issues regarded

- Opportunities and methodological challenges of Big Data for official statistics
- Big Data technologies and platforms for official statistics
- Transition plan to integrate Big Data in Official Statistics production

A related concept paper (Eurostat. 2014c)[33] provides some insight to the content discussed. Findings about of potential use and opportunities of Big Data based on studies from Statistics Netherlands on:

- "Dutch traffic loop detection records (e.g. for traffic statistics),
- Mobile phone data (e.g. for mobility and tourism statistics) and
- Dutch social media messages (e.g. for consumer confidence)." [Ibid., p. 7]

The session on "Methodology, quality issues and accreditation of Big Data sets" considered special research needs regarding:

- "social data [which] are highly unstructured, and often not accompanied by metadata hence a quality evaluation of such data is subject to a preliminary metadata enrichment phase (if possible).

---

[31] http://epp.eurostat.ec.europa.eu/portal/page/portal/pgp_ess/0_DOCS/estat/SCHEVENINGEN_MEMORANDUM%20Final%20version.pdf

[32] http://www.cros-portal.eu/content/big-data-event-2014

[33] http://www.cros-portal.eu/sites/default/files//Concept%20paper%20ESS%20Big%20Data%20Event.pdf

- hoc methodological developments [which] have to be undertaken. In addition, the lack of a "simple" structure requires pre-production treatment different from the usual one found in official statistics (from surveys or administrative registers). Data validation and editing may also need methodological developments, due to the volume and high frequency of Big Data. From the perspective of statistical representativeness, the coverage of Big Data and its capacity to be relevant for the study of the usual populations of interest for official statistics has to be studied.
- three different roles, namely:

  (i) "new" sources enabling access to data not yet collected by Official Statistics

  (ii) "additional" sources to be used in conjunction with traditional data sources and

  (iii) "alternative" sources fully or partially replacing traditional ones".[Ibid., p. 8]

Methodological challenges facing "Big Data in Public Opinion and Survey Research" were discussed for instance at the AAPOR conference 2014 under several scientific perspectives like

- How Can "Big Data" be the "Data" for Survey and Public Opinion Researchers?
- Big Data uses in Federal Statistics and Political Research
- Big Data, Ethics, Privacy and Confidentiality
- Big Data and how they relate to the Total Survey Error framework
  (AAPOR Conference. 2014)[34])

As a further Initiative in the field of OS data domain, UNECE launched 2013 the international collaboration project on "The Role of Big Data in the Modernisation of Statistical Production" to address several issues in this domain. The project started 2014 - overseen by the High-Level Group for the Modernisation of Statistical Production and Services (UNECE. 2010. HLG)[35] - and has three main objectives. It will

- "identify, examine and provide guidance for statistical organizations to act upon the main strategic and methodological issues that Big Data poses for the official statistics industry
- demonstrate the feasibility of efficient production of both novel products and "mainstream" official statistics using Big Data sources, and the possibility to replicate these approaches across different national contexts
- facilitate the sharing across organizations of knowledge, expertise, tools and methods for the production of statistics using Big Data sources." (UNECE. 2013. Big Data)[36]

A group of leading international experts under the supervision of the High-Level Group addressed issues and requirements around the use of Big Data in a paper entitled "What does Big Data mean for official statistics?"[37] (UNECE. 2013. HLG Papers).

---

[34] http://www.aapor.org/Abstract_Book.htm

[35] http://www1.unece.org/stat/platform/display/hlgbas/High-Level+Group+for+the+Modernisation+of+Statistical+Production+and+Services

[36] http://www1.unece.org/stat/platform/display/bigdata/Big+Data+in+Official+Statistics

The paper provide among other topics valuable descriptions about "Sources" of Big Data for official statistics, which are of high explanative interest in the context of the D7.2/3 report.

> "7. So far there are mainly two different ways that NSOs and International Organizations (IO) produce data: sample surveys and from administrative data sources including registers.
>
> The question that needs to be addressed is: how can big data help measure more accurately and timely economic, social and environmental phenomena?"
>
> 8. In general, large data sources can be classified as follows:
>
> a. Administrative (arising from the administration of a program, be it governmental or not), e.g. electronic medical records, hospital visits, insurance records, bank records, food banks, etc.
> b. Commercial or transactional: (arising from the transaction between two entities), e.g. credit card transactions, on-line transactions (including from mobile devices), etc.
> c. From sensors, e.g. satellite imaging, road sensors, climate sensors, etc.
> d. From tracking devices, e.g. tracking data from mobile telephones, GPS, etc.
> e. Behavioural, e.g. online searches (about a product, a service or any other type of information), online page view, etc.
> f. Opinion, e.g. comments on social media, etc.
>
> Administrative data is one of the main data sources used by NSO's for statistical purposes. Administrative data is collected at regular periods of time by statistical offices and is used to produce official statistics. Traditionally, it has been received, often from public administrations, processed, stored, managed and used by the NSOs in a very structured manner.
> Can one consider administrative data "Big" in accordance with the definition given above? For the moment the response would be, probably not. Administrative data can become "Big" when the velocity increases, e.g. using extensively administrative data where data is collected every day or every week instead of the usual once a year or once a month." (Ibid., p. 3)

The authors summarized particular "Challenges" of Big Data from official statistics under the following six categories, which are explained by further separate sections:

- Legislative, i.e. with respect to the access and use of data.
- Privacy, i.e. managing public trust and acceptance of data re-use and its link to other sources.
- Financial, i.e. potential costs of sourcing data vs. benefits.
- Management, e.g. policies and directives about the management and protection of the data.

---

[37] http://www1.unece.org/stat/platform/pages/viewpage.action?pageId=77170614

- Methodological, i.e. data quality and suitability of statistical methods.
- Technological, i.e. issues related to information technology." (Ibid., p. 3)

Furthermore, the authors provide some descriptive examples about "How Big data can be used in Official Statistics Communities", which regard (Ibid., pp. 6 - 8)

- "Traffic and transport statistics"
- "Social Media statistics."

"The following are a list of planned studies under Eurostat's programme of work and include a number of feasibility studies which aim at exploring the potential of Big data for official statistics." The major headers concern studies on

- Price Statistics
- Tourism Statistics
- Information and Communications Technology (ICT) usage Statistics

Finally, the report considers "three broad areas for experimentation:

- Combining Big data with official statistics.
- Replacing official statistics by Big data.
- Filling new data gaps, i.e. developing new 'Big data - based' measurements to address emerging phenomena (not known in advance or for which traditional approaches are not feasible).

The combination potential of Big data with official statistics represents some analogies with what has been done during the last several decades in respect of using administrative data with official statistics. What is likely to be slightly different though, and potentially attractive, is the possibility of applying, more extensively, statistical modelling for combining the two. In doing so, estimates may be obtained that maintain the strong quality properties of official statistics and enhance them with the power of near real time measurements obtained from Big data." (Ibid., pp. 6 - 8)

One next step in the related context of HGL Big Data project was in 2014 to review the draft paper about "How big is Big Data? Exploring the role of Big Data in Official Statistics" initialised by virtual sprint of the HLG Big Data projects, held four days in March 2014. (UNECE. 2014. Big Data.)[38]

In the ESSnet context, the CROS Portal provide an information space on Big Data like for relevant resources (e.g. conference paper), initiatives and relevant documents and papers in the context of official statistics (CROS Portal. Big Data).[39]

For example, Rey del Castillo (CROS Portal. Big Data. Publication. 2013)[40] explores and reflects "Big Data" issues in the light of main procedural aspects dealing with administrative data.

---

[38] http://www1.unece.org/stat/platform/download/attachments/99484307/Virtual%20Sprint%20Big%20Data%20paper.docx?version=1&modificationDate=1395217470975&api=v2

[39] http://www.cros-portal.eu/content/big-data

> "The experience on the use of another source that shares with Big Data this feature of not being designed for statistical purposes –as it is administrative data– may illuminate the road map, comparing the features in common and the ones that make a difference." [Ibid p.1].

Five aspects describe major concerns explored in 'handling' Big Data, but only one use the term 'metadata' explicitly:

> "Some Big Data have a certain structure related with the source of information and some are just unstructured text strings. Good metadata are not usually available and it seems that –in most of the cases– the tasks to harmonize or translate to statistical structures would be enormous."[Ibid p. 2]

A general impression is that investigations on uncovered issues in standards development in the field of Big Data has majorly to deal with high-levelled discussions and intensive explorations about conceptual, methodological, and technical concerns and requirements. These questions are discussed in the context of the statistical data domain along with conceptual models (GSIM; GSBPM) and several types of frameworks on legal, policy, technical and organisational aspects.

Beyond the development of conceptual areas and related standards, the UNECE HLG initiative provide information about the use of Big Data via the "Big Data Inventory" (UNECE. Big Data. Inventory). [41] The responsible team gathered a significant amount of information about the use of Big Data in various national and international statistical organizations.

## 3.2 Administrative data

There are basic terms used in the field of administrative data, to describe related aspects of the context. Different glossaries established in the statistical domain provide respective definitions.

The OECD Glossary of Statistical Terms provides the following definitions (OECD. Glossary):

- "Administrative data[42] is the set of units and data derived from an administrative source."
- "Administrative data collection[43] is the set of activities involved in the collection, processing, storage and dissemination of statistical data from one or more administrative sources.
  [It is] The equivalent of a survey but with the source of data being administrative records rather than direct contact with respondents."

---

[40] http://www.cros-portal.eu/sites/default/files//ReflectionsUseBigDataStatisticalProduction_0.pdf

[41] http://www1.unece.org/stat/platform/display/bigdata/Big+Data+Inventory

[42] http://stats.oecd.org/glossary/detail.asp?ID=6

[43] http://stats.oecd.org/glossary/detail.asp?ID=4328

- "Administrative source[44] is the organisational unit responsible for implementing an administrative regulation (or group of regulations), for which the corresponding register of units and the transactions are viewed as a source of statistical data."

The ESSNET Admin Data Wiktionary Glossary [CROS portal. AdminData][45] offer following definitions in the same terminological context

- Administrative Data: The data derived from an administrative source, before any processing or validation by the NSIs.
- Administrative Dataset: Any organised set of data extracted from an administrative source, before any processing or validation by the NSIs.

Beyond these definitions, the Administrative Data Liaison Service [ADLS. 2014] in the UK supplies comprehensive information and documentation on further contextual facets about creation and purpose of administrative data and the analytical potentials for social science research.[46]

> "Administrative data refers to information collected primarily for administrative (not research) purposes.
>
> This type of data is collected by government departments and other organisations for the purposes of registration, transaction and record keeping, usually during the delivery of a service.
>
> In the UK, government departments are the main (although not exclusive) purveyors of large administrative databases, including welfare, tax, health and educational record systems. These datasets have for many years been used to produce official statistics to inform policy-making.
>
> The potential for this data to be accessed for the purposes of social science research is increasingly recognised, although as yet has not been fully exploited. Two areas of research – education and health – have seen fairly extensive use of administrative data[1], but most other administrative datasets have not been widely used for research purposes."

The ADLS website introduction summarise also advantages and disadvantages of using administrative data in research. Furthermore, they inform about long experiences in Nordic countries utilising of administrative data resources like Denmark, Sweden and Finland.

For example Statistics Finland

> "collects almost all (96 per cent[3] of its data from administrative sources. Since 1990 Finland has also been fully reliant on administrative data to produce the population and housing census. The Finnish Statistic Act (2004)[4] is based on the central tenet that

---

[44] http://stats.oecd.org/glossary/detail.asp?ID=7

[45] http://essnet.admindata.eu/

[46] http://www.adls.ac.uk/adls-resources/guidance/introduction/

wherever possible pre-existing administrative data should be sourced before making requests on the general public to provide information." [Ibid]

To reduce financial and operative burdens collecting census data, Statistics Finland has identified three key areas for broad usage of administrative data:

1. "Direct use of administrative data to produce national economic and social statistics, for example crime rates, election statistics and employment statistics[6].
2. Linking different complementary administrative datasets.
   Data linkage is facilitated through concerted collaboration efforts between data holding authorities, and a well established unified system of personal identity codes used across different datasets.
3. Combining survey and administrative data.
   Conducting surveys is still an important part of Statistics Finland's work and administrative data is used to provide sampling frames, improve the quality of survey data once it is collected, check for errors, impute missing data and supplement the data, allowing the survey to concentrate on information not available elsewhere." [Ibid]

In the context of the DwB D7.2/3 report, it is of interest, to recognise related requirements and needs for metadata to document this type of data (and respectively the production lifecycle), as well as issues for (further) standards development. The following report is applied for this purpose, as it provide comprehensive information about these aspects.

2007 the United Nations Economic Commission for Europe (UNECE) reviewed best practices on register-based in the Nordic countries. The report "Register-based statistics in the Nordic countries" outline major metadata and documentation needs under the following headers (UN. 2007. P. 38). [47]

**Metadata and documentation from administrative sources**

"Over the last few decades the need for metadata in the statistical production process has been increasingly evident. Most statistical offices are striving to introduce metadata systems, or improve existing ones."

"The nature of the metadata differs significantly between register-based data and surveys with their own data collection. It is also necessary to distinguish between the documentation of registers and the documentation of register-based statistics. When using existing registers to create new registers, register documentation is crucial. This type of documentation is characterised by:

- The volume of the metadata, which can be very high.
- Every administrative source must be documented.
- Changes in the administrative system must also be documented.
- The variables can be complicated so documentation must be precise.
- A large amount of data processing is done to create units and variables, and this processing should also be documented.

[47] http://unstats.un.org/unsd/dnss/docViewer.aspx?docID=2764

This means that the metadata system must suit the requirements of the register system and register-based statistics." (Ibid., p. 39)

**Documentation of administrative sources**

The "suppliers of data (register owners) submit record descriptions etc., which indicate the structure and content of the data being delivered. Furthermore, the NSIs and NSAs should have electronic access to the questionnaires, including instructions, to be stored in the metadata system." (Ibid)

For certain types of microdata from different sources, common documentation and classifications is needed. Such data are not necessarily produced by the data provider or disseminator of the data, but are combined from different sources over several years to complex, large data sets.

For such data, accurate documentation of the creation of the data is essential. Only if the production procedure for those microdata is well documented, the analysis can be correctly carried out. In this context also the imputation process also needs to be documented, in case as these data may include contradictory values for the same attribute or missing values in one source.

**Documentation and metadata about sources within a register system**

"In constructing a new statistical register the register holder often use variables from other statistical registers already available in the register system" (Ibid). In those cases, "it is essential to have easy access to the existing metadata to be able to search and select suitable variables. It is important to have the possibility to transfer metadata from the existing sources to the new register's documentation and metadata system. To avoid duplication of work and instead conveniently reuse metadata, documentations must be strictly formalised according to common rules and stored in a universally accessible way." (Ibid)

**Documentation and metadata about changes over time**

"Four types of events can affect register statistics, and in order to avoid incorrect interpretations of time series from register-based surveys we need to know the following:

- Have changes taken place in the administrative system that makes up the sources, whereby administrative concepts have been given new definitions?
- Have changes taken place in the way the register has been formed, e.g., new sources or new estimation methods?
- Have there been changes to the classifications that are used in the register?
- Have any external changes taken place that could have affected the statistics indirectly?" (Ibid)

**Classification and definitions database**

"Industrial classification, product category, education, occupation and regional codes are examples of important statistical standards and classifications. The administrative sources contain data on these hierarchical classifications and this information is used to create variables within the register system. These classifications are changed at regular

intervals. As value sets (sets of all codes/categories) are also large, a classification database is needed to manage all the codes and keys between the different versions. This classification database is an important resource when the variables in a register are documented." (Ibid., p. 40)

**Comprehensive metadata system**

"There should be a system that integrates all formalised metadata, including a calendar, and a classification and definitions database. In addition to this, a system is needed to manage documents with other types of metadata." (Ibid., p. 40)

ESSnet AdminData covers issues on administrative data in the ESSnet context, which is one of the MEETS (Modernisation of European Enterprise and Trade Statistics) projects. The project consists of 10 work packages covering different aspects of the use of administrative data for business statistics. All project information is available on a dedicated website (CROS Portal. AdminData)[48] [49]

## 3.3 Versioning

In a DA context, versioning is used to indicate how different versions of the same dataset differ, in terms of the data as well as metadata content. These versions are generally discrete in nature: a new version may indicate one or several changes to the data matrix or to the metadata. Principles for versioning vary between DAs. Below, the principles for microdata versioning of the Swedish National Data Service (SND) are used to illustrate the general idea.

A SND dataset version consists of two levels (for instance v1.2). The first level reflects changes in the data matrix: variables have been added, moved, recoded, combined, or renamed, for example. Reasons behind such changes include removal of inconsistencies (at the request of the primary investigator or data producer), combination of, for example, geographical variables to ensure secondary confidentiality, or the addition of specific SND variables (study number, dataset number, version number). If the primary investigator submits revised or extended versions of a dataset that also results in a higher first level.

The second level (the number after the full stop) changes if the metadata of the dataset is in some way updated, revised, or expanded. Basically, a given statistical analysis run on dataset versions that share the same first level (for instance v1.2 and v1.5) will always have the same results; something which is not necessarily the case if versions are of different first levels (for instance, v1.2 and 2.2).

To help researchers not only to refer to but also to find particular versions of the datasets, datasets can be supplied with permanent identifiers (PIDs). Such identifiers link permanently

[48] http://www.cros-portal.eu/content/use-administrative-and-accounts-data-business-statistics

[49] http://essnet.admindata.eu/

and uniquely to a particular object, to avoid confusion and to enable replication of research. As researcher interpretations of data and results may conceivably differ even between different second-level versions, each change in version, first *or* second level, thus results in new PID – in SND's case, the dataset is given a new DOI for each change in version: v1.2 would have a different DOI from v1.3.

As an alternative many NSIs and NSAs use the technique with timestamp on microdata in order to produce a version of a dataset for a given time period. From a metadata perspective this alternative demands that the classifications, including the domains for each variable is updated and accessible for the research users.

## 3.4 Metadata concerning data processing

Between data collection and publishing, dissemination or research output, data get changed through cleaning and other transformation processes during the phase of Data Processing according to the DDI life-cycle model (DDI Alliance. 2009) [50] and respectively in phase number 5 "Process" of the GSBPM model (UNECE. METIS. 2013a GSBPM) [51].

In this paper, we call information about these 'interventions' to the data in short "process metadata".

Process metadata is not necessarily captured in a structured manner, nor carried alongside data as they are transformed. This prevents both process reuse and full replicability in research. There are ongoing initiatives in both the SDMX and DDI communities to address this issue. Although both aim at describing data transformations in a generic way, the perspective and scope of the two are different.

The SDMX-related work focus mainly on dimensional data, where the goal is to have an *executable* language to perform data validations and transformations on data sets that can streamline the data production process. The DDI-related work, on the other hand, focuses on *documentation* of transformations for the sake of replicability and improved automation in the production of metadata along the lifecycle.

### *3.4.1 SDMX Validation and Transformation Language (VTL)*

The SDMX secretariat has initiated a process to design a Standard Validation and Transformation Language (VTL) (European Commission. 2014) [52]. VTL takes a mathematical approach, designing a language without bindings to any particular IT-system or existing platform. VTL operands are typically dimensional data sets, and VTL will support both common validation/comparison operators on these operands as well as transformation operators.

---

[50] http://www.ddialliance.org/what

[51] www1.unece.org/stat/platform/display/metis/The+Generic+Statistical+Business+Process+Model

[52] https://webgate.ec.europa.eu/fpfis/mwikis/essvalidserv/index.php/Standard_Validation_and_Transformation_Language_%28VTL%29

An important goal of VTL is that it should become an executable language. Programs written in VTL will be able to run directly on data residing in different systems (e.g. relational databases, files). This has several benefits:

- users will not have to learn e.g. SQL or the programming languages of statistical packages to carry out transformations and validations
- transformation and validation programs will be robust to changes in implementations/backend systems
- programs may be reused by other units/institutions with different IT infrastructure
- programs may be preserved in an unambiguous and purely logical fashion

To enable VTL to become executable, it is not sufficient to design the language itself. Indeed, a number of implementations also need to be created and kept in sync with emerging platforms and backend implementations.

Early drafts of VTL syntax exist, but have not been made publicly available at the time of writing this report.

### *3.4.2 DDI initiative to capture metadata on data transformations*

In a meeting in Ann Arbor (Michigan, US) 10-11 June 2014, ICPSR kicked off an initiative to capture and represent data transformations in DDI. The goal is to design a canonical, generic model (or language) that can represent data transformations in a platform-independent manner, not dissimilar to VTL.

However, the DDI-related initiative aims to build tools that can parse and convert existing data transformation scripts (written in e.g. SPSS, Stata, and SAS), and map/translate them into the canonical model/language, which then can be added to DDI metadata records.

Another goal is to be able to use the transformations to automate the generation of updated metadata records (e.g. remove variable metadata that have been removed in transformations, or change metadata for variables that have been re-coded).

The canonical model/language is yet to be defined. One option will be to align this with the data transformation portion of VTL. The alternative is to design a separate model/language.

The work on this will likely be carried onwards by both a DDI working group on data transformations (yet to be constituted) as well as concrete implementation projects funded by external bodies.

At the time of writing of this report, reports from the kick-off meeting are not yet published.

# 4 THE DEVELOPMENT OF STANDARDS: WHAT'S IN THE PIPELINE?

Metadata standards constitute an overwhelmingly complex field. In many cases, the various discipline, community and country specific metadata efforts have led to duplicate and overlapping standards as well as to artificial barriers to data discovery and reuse across disciplines.

One of the main reasons for this has been the lack of communication across communities. Another reason is a practical one: there is usually a pressing need to get a system running up to support a specific task (Willis, C., Greenberg, J., White, H. 2012).

In the field of metadata for research data, the Digital Curation Centre (DCC. 2014) provides links to disciplinary metadata standards[53], including information about tools to implement the standards as well as use cases of data repositories currently using the standards. At the time of writing, the DCC list contained 34 entries. Many of the listed standards are familiar to data archives and statistical institutes, for example DDI, ISO 19115, RDF and SDMX.

In 2013, the Research Data Alliance (RDA. 2013) established a Metadata Standards Directory Working Group[54]. Its goal is to develop a collaborative, open directory of metadata standards applicable to scientific data, and in the start, they will focus on widely used and domain community endorsed metadata standards.

## 4.1 DDI – future development

Over the last years, DDI has evolved to meet the requirements of new users like the statistical communities.

The DDI Alliance has decided that the next version of the specification will be based on an information model, and therefore the development of DDI has changed significantly from DDI3.x. The new DDI will be developed in a series of sprints that focus on a specific set of goals - in summary called the "DDI Moving Forward Process[55]".

The process began in 2012 with a expert workshop titled "DDI Lifecycle: Moving Forward" and held at Dagstuhl, Germany. A follow-up workshop, the first actual sprint, was held in 2013.

Conclusions from these two workshops include a dynamic description of the development process[56] as well as the following decisions:

- DDI developments shall be model-driven[57] (and not XML-driven as before)

---

[53] http://www.dcc.ac.uk/resources/metadata-standards

[54] https://rd-alliance.org/working-groups/metadata-standards-directory-working-group.html

[55] http://www.ddialliance.org/ddi-moving-forward-process-summary

[56] http://www1.unece.org/stat/platform/display/DDI4/DDI+4+Process

- DDI should broaden its scope into new research domains
- DDI should broaden its scope beyond data collected via other instruments than surveys (e.g. administrative registers, sensors, other types of observational data)
- The DDI specification should be simpler and more approachable for user groups outside the data archive community
- The model will be developed by Content Modelers in collaboration with (more technical) Data Modelers

Design principles under "DDI Moving Forward" include

- Interoperability and Standards – The model is optimized to facilitate interoperability with other relevant standards.[58]
- Simplicity – The model is as simple as possible and easily understandable by different stakeholders.
- User Driven – User perspectives inform the model to ensure that it meets the needs of the international DDI user community.
- Terminology – The model uses clear terminology and when possible, uses existing terms and definitions.
- Iterative Development – The model is developed iteratively, bringing in a range of views from the user community
- Documentation – The model includes and is supplemented by robust and accessible documentation.
- Lifecycle Orientation – The model supports the full research data lifecycle and the statistical production process, facilitating replication and the scientific method.
- Reuse and Exchange – The model supports the reuse, exchange, and sharing of data and metadata within and among institutions.
- Modularity – The model is modular and these modules can be used independently.
- Stability – The model is stable and new versions are developed in a controlled manner.
- Extensibility – The model has a common core and is extensible.
- Tool Independence – The model is not dependent on any specific IT setting or tool.
- Innovation – The model supports both current and new ways of documenting, producing, and using data and leverages modern technologies.
- Actionable Metadata – The model provides actionable metadata that can be used to drive production and data collection processes.

The second DDI Moving Forward Sprint was held in Paris in December 2013, the third in Vancouver in April 2014, and the fourth in Toronto in May 2014. In addition to technical topics, the sprints have discussed metadata issues related to, for example, access control, data management, provenance and discovery [DDI 4 Sprints – Overview] [59].

---

[57] Model driven implies that the model is separated from its representation. In DDI-C and DDI-L (3.x), the model itself is expressed in XML schemas. In the future, the model will be expressed in UML, and have representations in XML schemas, RDF schemas and potentially other forms

[58] http://www1.unece.org/stat/platform/display/DDI4/Guidance+for+relevant+standards

[59] http://www1.unece.org/stat/platform/display/DDI4/Sprints

The first draft of the model is scheduled to be published in spring 2015. It will most likely include coverage of foundational metadata such as concepts, codes, and universes as well as basic survey instruments and codebooks. The progress of the DDI Moving Forward project can be followed in the project website.[60]

In addition, in March 2014, the "The Copenhagen Mapping" (2014) [61] is an attempt to map GSIM 1.1 to DDI 3.2 It is still too early to predict the actual effects of this mapping, but it signals a close relation between the conceptual model and the metadata standard in the future.

## 4.2 SDMX – future development

In this report we have concentrated on microdata and since SDMX has its main focus on aggregated data we only mention SDMX briefly.

The work in this report focuses mainly on the SDMX Metadata Common Vocabulary, MCV[62]. The recommendation is to have a common terminology in order to facilitate communication and understanding. The philosophy is that if a term is used in the SDMX documentation, then its precise meaning should correspond to the MCV definition, and any reference to a particular phenomenon described in the MCV should use the appropriate term.

## 4.3 Linked Data and RDF

This section provides a short introduction to linked data and a discussion about why linked data are relevant to NSIs, DAs and the research community. Furthermore, this section includes a description of the activities and projects that are going on in the field of linked data, DDI and SDMX, and also a description of a use-case from a pilot study in Sweden where classifications were published as linked data. Some thoughts about recommendations are outlined for further discussion and conclusions in the chapters regarding Discussion and Conclusions.

### *4.3.1 Linked Data – Background*

The World Wide Web has provided us with a platform for the sharing of knowledge and has made it easy to publish and access information in a global information space. However, making data available and searchable in the same way as documents on the Web has not evolved as quickly. It is still common to publish data on the Web in formats such as HTML-

[60] http://www1.unece.org/stat/platform/display/DDI4/*Moving+Forward+Project

[61] http://cdn.colectica.com/TheCopenhagenMapping-Draft.pdf

[62] http://www1.unece.org/stat/platform/display/metis/SDMX+-+Metadata+Common+Vocabulary

tables, Excel-files and tab-separated files (CSV files). All these formats have serious limitations regarding the possibilities to describe structure and semantics of the data.

The concept of Linked Data offers a set of best practices for publishing and connecting structured data on the Web in a machine-readable way (compare W3C Sematic Web. Data. 2013[63]).

Another important feature is the possibility to link to and from other external data sets. In a similar way as web pages are connected by hyperlinks using HyperText Markup Language (HTML), Linked Data techniques make it possible to link data with the help of the Resource Description Framework [compare W3C Sematic Web. RDF.2014] [64]) format.

The Linked Data principles that have been defined by Berners-Lee (2006[65]) contain the following four rules for publishing data on the Web:

1. Use URIs as names for things (URI is an acronym for Uniform Resource Identifier. URIs are used for identifying names of web resources).
2. Use HTTP URIs so people can look up those names.
3. When someone looks up a URI, provide them with useful information, using the standards RDF and SPARQL (SPARQL is a query language that makes it possible to retrieve and manipulate data stored in RDF format).
4. Include links to other URIs, so that they can discover more things.

### *4.3.2 NSIs, NSAs, DAs and Linked Data*

NSIs, NSAs and DAs have an interest in having their data discoverable. The Web of Linked Data offers a new way of publishing data with a strong focus on machine-readability and linking between data sets. At the same time, there are concerns about the lack of standards and structured metadata in the linked-data cloud.

Several initiatives have been started with the aim to assess the possibilities to use Linked Data for disseminating statistical data. Of particular interest in this context are those initiatives that have developed Linked Data representation of the SDMX and DDI models.

From the perspective of data access for researchers, Linked Data can be seen as a candidate for a standard that could be used for data retrieval from sources that use different standards, such as SDMX and DDI. For instance, the work that has been done with vocabularies in the RDF community could be a way forward in the area of mapping internal standards to a common vocabulary.

One advantage with RDF-based tools is that RDF schemas focus on semantics instead of structure (like e.g. XML schemas). The strong semantic definitions enable not only the data to be machine-readable; it also means that machines can infer the meaning of the schema.

---

[63] http://www.w3.org/standards/semanticweb/data

[64] http://www.w3.org/standards/techs/rdf#w3c_all

[65] http://www.w3.org/DesignIssues/LinkedData.html

From a researcher perspective, it would be very useful to have the possibility to link related datasets and concepts in different organisations. Linked Data offer this possibility and also have a simple but powerful interface for data retrieval. RDF is also well suited for systems to be used in a distributed environment with decentralised metadata.

### *4.3.3 DDI and Linked Data*

The DDI community is committed to support RDF as well as XML schemas in the future versions of DDI. There are already existing approaches in place or close to production status for DDI-RDF-based metadata documentation (see DDI. RDF Vocabulary. Kramer, S., et al. 2012; Bosch, T., et al. 2013).

The following three DDI RDF vocabularies[66] are of particular interest in the context of the report.

**DISCO – DDI-RDF Discovery Vocabulary**[67]

The DDI-RDF Discovery Vocabulary is based on a subset of the DDI XML-specification of DDI Codebook and DDI Lifecycle. It allows discovery of data as well as metadata using RDF technologies.

**XKOS - Extended Knowledge Organization System**[68]

XKOS is an extension of the Simple Knowledge Organisation System (SKOS) which is widely used in the linked-open-data community. SKOS is a W3C standard, based on other Semantic Web standards (RDF, OWL) and provides a way to represent controlled vocabularies, taxonomies and thesauri (compare W3C Sematic Web. Ontologies the specification of RDF, SKOS and OWL (Web Ontology Language) [69]) The extension in XKOS provides support for statistical classifications.

**PHDD - Physical Data Description**[70]

PHDD is a relatively limited RDF-vocabulary for describing essential properties of rectangular data files (e.g. CSV-files or spreadsheets). The ambition of PHDD is to be able to describe a data file sufficiently for it to be imported automatically by programs that can understand PHDD.

---

[66] http://www.ddialliance.org/Specification/RDF

[67] http://rdf-vocabulary.ddialliance.org/discovery.html

[68] http://rdf-vocabulary.ddialliance.org/xkos.html

[69] http://www.w3.org/standards/semanticweb/ontology

[70] http://rdf-vocabulary.ddialliance.org/phdd.html

### 4.3.4 SDMX and Linked Data

The great value that SDMX could bring to the world of Linked Data is its very rich data model. It is therefore understandable that activities have been focused on creating a stable SDMX-compatible RDF vocabulary that can handle representation of statistical datasets.

Below follows a description of two of the initiatives who have been working with RDF vocabularies for statistical data.

**SCOVO – The Statistical Core Vocabulary**

SCOVO was one of the first initiatives that developed an RDF vocabulary for statistical data. Although proven to be compatible with the core elements of the SDMX model, SCOVO has limitations concerning the possibilities to make queries (with SPARQL) against data since it does not contain description of the structure of the data cube. Thus, the SCOVO vocabulary is deprecated and it is strongly advised to use the Data Cube Vocabulary instead.

**The RDF Data Cube Vocabulary**

The Data Cube Vocabulary is an RDF vocabulary for multidimensional data cubes. The vocabulary is compatible with SDMX and supports extension vocabularies that open up for domain-specific extensions. Like SCOVO it also focuses on a subset of the SDMX Information model but it is better suited to support SPARQL queries.

The RDF Data Cube Vocabulary[71] was developed by the W3C Government Linked Data (GLD) Working Group and has been reviewed and endorsed as a W3C Recommendation 16th January 2014 (W3C Recommendation. 2014[72]).

### 4.3.5 Use Cases, projects and pilots

**LATC - Linked Open Data Around The Clock[73]**

LATC was a Specific Support Action (SSA) in the context of the FP7 with partners all over Europe and clients around the world.

The LATC activities were primarily aimed at the following target groups:

- original data owners, such as government agencies;
- researchers who deal with large-scale data;
- web developers who build applications with Linked Data;
- and SMEs that want to benefit from the lightweight data-integration possibilities of Linked Data.

[71] http://www.w3.org/TR/2014/REC-vocab-data-cube-20140116/

[72] http://www.w3.org/2011/gld/wiki/Data_Cube_Vocabulary

[73] http://latc-project.eu/

The objective for the LATC project was to publish data sources of Institutions and Bodies of the European Union as Linked Open Data to seed the EU data cloud. The project ended in September 2012.

**LOD2 - Creating Knowledge out of Interlinked Data**[74]

LOD2 is a large-scale integrating project co-funded by the European Commission within the FP7 Information and Communication Technologies Work Programme. This four-year project comprises leading Linked Open Data technology researchers, companies, and service providers from across eleven European countries (and one associated partner from Korea) and is coordinated by the AKSW research group at the University of Leipzig.

Some examples of developments are:

- enterprise-ready tools and methodologies for exposing and managing very large amounts of structured information on the Data Web
- a testbed and bootstrap network of high-quality, multi-domain, multi-lingual ontologies

In the context of DwB and WP7, one of the LOD2 work packages (WP9) could be of special interest as the use case embeds the national level of statistical offices. The WP9 Use Case 3: LOD2 for Citizen – PublicData.eu[75], has the purpose

> "to increase public access to high-value, machine-readable data sets generated by the European, national as well as regional governments and public administrations. Although this effort will be similar to developments in other parts of the world, for the case of Europe it will be more challenging due to the larger organizational and linguistic diversity and thus represent an ideal application scenario for Linked Data technologies."[76]
>
> LOD2 WP9 deliverable D9.5.1[77]
>
> "describes the activities of the Serbian LOD2 team related to the establishment of a Serbian CKAN[78] metadata repository (http://rs.ckan.net) that serves for publishing open governmental data from Serbia, as well as a source catalogue for the PublicData.eu portal.
>
> In order to explain the need for integrating public data at the European level, the "Country Statistical Office" Use Case has been introduced. National statistical offices across Europe responsible for the development, production and dissemination of European statistics are obliged to harmonize their standards, classifications and methodologies with the guidelines provided by EUROSTAT, a Directorate-General of the European

---

74 http://lod2.eu/Welcome.html

75 http://publicdata.eu

76 http://lod2.eu/WorkPackage/wp9.html

77 http://static.lod2.eu/Deliverables/LOD2_D9.5.1_Serbian_CKAN.pdf

78 CKAN (Comprehensive Knowledge Archive Network) is open source data portal software: http://ckan.org

Commission. In this sense, this Deliverable shows how the CKAN software can be customized and how the LOD2 technologies can be used to publish and disseminate statistical data as linked data, thus providing easy access to official statistics information."(LOD2. 2012. P. 3)

#### *4.3.5.1 LOD and LODify – A Use case on linked statistical data from Statistics Sweden*

During 2012, Statistics Sweden participated in a project financed by the Swedish Governmental Agency for Innovation Systems (VINNOVA). The project had two main objectives:

- to bring forward the benefits of linked data and spread knowledge about this technology, as well as
- to publish a selected part of the basic data from our society, which could be useful and reusable by a large group of actors.

The datasets constitute examples of how linked data makes it possible to link and reuse data from various different data sources and use them together, and they can thereby act as inspiration and examples for other organizations, as well as provide a concrete reusable resource.

The project published two datasets from the metadata-classification databases of Statistics Sweden:

- data about counties, municipalities, and parishes in Sweden, as well as
- data about the Swedish industry codes (SNI codes) for classification of industry activities.

By giving these entities unique and dereferenceable URIs, and publish data about them as RDF, the project enables data reuse and linking over the web. LODify[79] is used to convert SCB's statistics to RDF which then are provided in the cloud service, which, among other things, allows to perform advanced queries using SPARQL.

Through a web portal[80] the project also started the promotion of a Swedish linked data community that will hopefully continue to grow in the years to come

## 4.4 Trends in vocabularies and coding schemes

Any prescribed set of terms used for standardisation of (metadata) values can be referred to as a vocabulary (Hider. 2012, p. 251).

A controlled vocabulary is more than a prescribed list of terms. It is the formally managed collection of terms, which is defined and accepted in the community (Neiswender. 2009)[81].

---

[79] LODify is a cloud service to facilitate the publication of Linked Open Data; http://lodify.com/en/showcases/

[80] http://www.linked-data.se

[81] http://marinemetadata.org/guides/vocabs/vocdef

A controlled vocabulary forms a semantic structure that is e.g. designed to control synonyms, distinguish among homographs and link together terms whose meaning are closely related (Lancaster. 1991, pp. 13-14). Controlled vocabularies are a widely used approach for instance to organise and describe data in a consistent way by restricting the possible values of the metadata elements both in study-level (Figure 2) and variable-level.

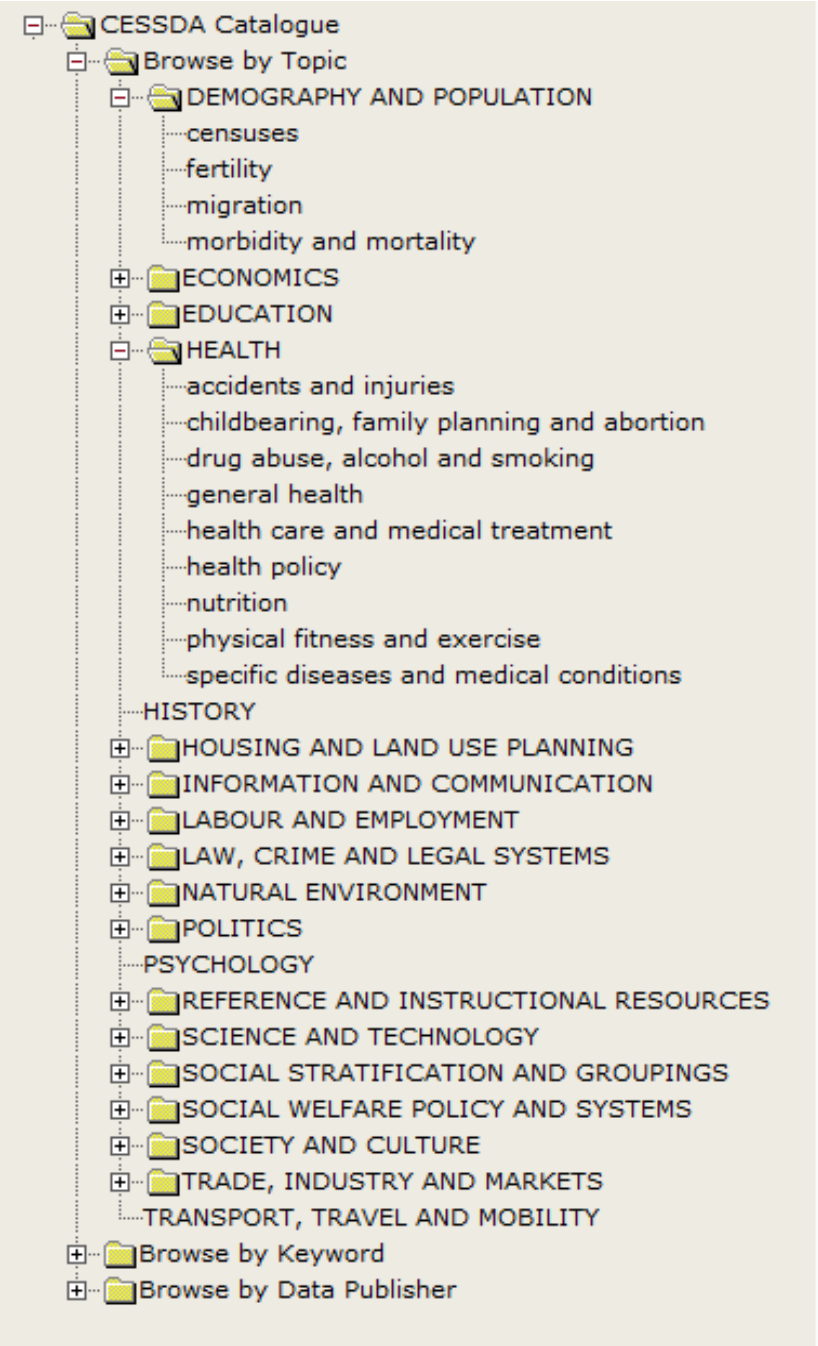


Figure 2: CESSDA topic classification - a controlled vocabulary used by many DAs[82]

[82] http://www.cessda.net/catalogue/

Coding schemes and statistical classifications are controlled vocabularies used in variable-level.

According to GSIM Specification [UNECE. METIS. 2013b GSIM, p. 22][83] coding schemes (lists) merely include categories combined with a code representation whereas in the statistical classification (Figure 3) the categories must be mutually exclusive and jointly exhaustive.

| Broad field | Narrow field | Detailed field |
|---|---|---|
| 00 Generic programmes and qualifications | 001 Basic programmes and qualifications<br>002 Literacy and numeracy<br>003 Personal skills and development | 0011 Basic programmes and qualifications<br>0021 Literacy and numeracy<br>0031 Personal skills and development |
| 01 Education | 011 Education | 0111 Education science<br>0112 Training for pre-school teachers<br>0113 Teacher training without subject specialisation<br>0114 Teacher training with subject specialisation |

Figure 3: Statistical classification of Fields of Education and Training (ISCED-F. 2013)

ISCED-F (2013), the statistical classification of Fields of Education and Training[84], is part of the International Standard Classification of Education (ISCED).

More detailed definitions and examples of different types of controlled vocabularies can be found in DwB D7.1 where some classifications used by NSIs and some vocabularies used by DAs are presented in depth (DwB D7.1. 2013, pp. 44-53).

Data-intensive science, the "fourth paradigm" as described by Gray (see Hey, T. et al.,2009), Big Data and linked data "hype" together with various national and international infrastructure policies like Horizon 2020 have lead to a growing interest in metadata by the research community.

When more and more data and metadata sources are made available, there's also a growing need for interoperability between metadata. Using common or compatible vocabularies is one way to achieve the required interoperability. As Greenberg (2014 p. 59)[85] states

> "metadata vocabularies promote greater consistency across data grids, repositories, and hubs, and can contribute to an architecture supporting an unified set of services and interfaces".

Greenberg et al. (2013) have also introduced the concept of "metadata capital", by which they refer to re-using good quality metadata. Investing resources to metadata generation

[83] http://www1.unece.org/stat/platform/display/gsim/GSIM+Specification

[84] http://www.uis.unesco.org/Education/Documents/isced-fields-of-education-training-2013.pdf

[85] https://journals.lib.washington.edu/index.php/acro/article/view/14680/12320

and development can yield positive returns if metadata is reused. Especially linked and open vocabularies can be seen as valuable metadata capital. Most users have a need to modify or tailor existing vocabularies for their own purposes. Instead of creating a totally new vocabulary, if an open vocabulary exists, it can be reused and relevant properties added.

A glance through the latest IASSIST conferences reveals continuous and possibly even growing interest in vocabularies and classifications. For example,

- in 2011, Schaible (2011) presented how free text from DDI Codebook documentation can be categorized when converting the metadata to DDI Lifecycle,
- in 2012 the cognitive aspects of classification (Gillman, Bosley and Fricker. 2012) as well as a DDI resource package for the ISCED classification (Wackerow and Orten. 2012) were discussed,
- 2013 saw a presentation from El-Haj (2013) about indexing with a SKOS version of HASSET Thesaurus, and
- in 2014 a whole session with five presentations was dedicated to harmonization, thesauri and indexing discussing the following projects and topics:
  - Taxonomy / Lexicon Project at the US Bureau of Labor Statistics (Gillman, D. 2014)
  - Linking thesauri: ELSST as a hub for social science data terms (Balkan, L. 2014)
  - Improving precision and recall in study retrieval: a concept for thesaurus-based syntactic indexing (Friedrich, T. 2014)
  - The case of CharmStats or how the process of harmonization can document itself using the right tool (Katsanidou, A. 2014)
  - DDA Indexing platform (Jensen, J. 2014).

(Details see Reference List section Presentations on Indexing & Classifications at IASISST)

# 5 THE NEED FOR RULES AND BEST PRACTICES

Like many terms in common usage, Best Practice is often used without an explicit definition. As this chapter is specifically concerned with best practices, it would seem prudent to clarify our understanding of the term.

A basic description of Best Practice is a method that is accepted and/or prescribed as being the correct or most effective one.

In this context, we take Best Practice to mean an established way effectively to manage a process or situation with the benefit of not having to resort to creating a method from scratch ("re-inventing the wheel") and thus risking introducing a plethora of ways to solve similar problems or set up similar processes.

With this definition, there may be several best practices, which address the same problem. A rule might then be formulated that favours one practice over the others, giving a definitive guideline to the preferred way of solving the problem in question.

This chapter examines what rules and best practices have bearing on a number of key areas related to the application of DDI to microdata for researchers, including the use of controlled vocabularies and classifications. First, problematic areas identified in DwB D7.1 are discussed (5.1), followed by various other key areas (5.2). The chapter concludes by presenting a fictitious use-case to illustrate where and how the various rules and best practices come into play in the process of finding and accessing research data (5.3).

## 5.1 Key areas identified in DwB D7.1

This section contains what appear to be key areas of concern from DwB D7.1, inferred and/or extrapolated from sections 6.3 through 6.6.

### *5.1.1 Metadata issues*

The issues within this key area concern practical metadata management and application in everyday practice, although on a structural level. No single metadata standard covers all information relevant to secondary users of data. Instead, different standards are used for different purposes. For the project at hand, however, a de facto standard involving DDI seems to be more or less implicitly promoted, though obviously DDI would have to be accompanied by complimentary best practices and rules that cover those pieces of information not available in DDI. But even if such a de facto standard is agreed upon there are still some practical issues to be addressed concerning the corresponding application of such standards, based on best-practice guidelines and implementation concepts to be developed by an organisation.

**Problem 1: Varieties of metadata-standard usage**

Central to an understanding and cross-analysis of data from DAs as well as NSIs from several countries is a consistent use of metadata standards. Although researchers can be expected

to deal with some variation, all variations in terms of standards make research more difficult and increase the risk of mistakes. Today, variations in how DAs and NSIs use metadata standards are ubiquitous. These variations can be broken down into two main types:

A. Different metadata standards for different purposes.

   Dušan Praženka and Peter Boško (2011, p. 1) observe that NSIs generally have adopted a standard different from that of international statistics organisations in their production of statistics, and that they therefore have to translate (through, for instance, mapping) data and metadata to required standards when delivering data to those organisations.

   A European example would be those NSIs that use SDMX for (aggregated) data that are sent to Eurostat but who employ a different metadata standard for their microdata.

B. Different versions of the same standard.

   Example DDI Codebook (DDI 1.x, 2.x) and DDI Lifecycle (DDI 3.x).

   In the survey of DAs referred to in DwB D7.1, it was observed how both the Codebook and Lifecycle versions of DDI were employed by the data archives; most archives use one DDI version or the other, and a few use both versions.[86]

   There is a basic conceptual difference between the two versions, with the result that there is a difference in the amount of metadata that can be attached to a material. It is therefore not possible to make a complete transformation from DDI Lifecycle to DDI Codebook. Partial transformation of DDI Lifecycle to Codebook is possible and integrated in some of the tools available.[87]

Any standardised usage under A would require rules and best practices that guide the future usage of DDI and additional standards or parts of standards are agreed upon. Presumably, such usage should be considered in the context of relevant processes on the basis of GSBPM and objects derived from GSIM, e.g. in case of planning a dedicated "data/metadata conversion project" to keep or enhance – step by step – the usability of an important collection of data for future use. The organisational situation to service for different but combined metadata standards might either be replaced by a more – jointly agreed – integrated, efficient standard, or there must be a way to map the utilised standard(s) on the expanded DDI system in the case of complex microdata.

(It is also worth pointing out parenthetically that DAs that are CESSDA Service Providers need to fulfil the obligations in ANNEX 2 of the CESSDA Statutes: "be fully compliant with the elements of the DDI metadata standard that are required to enable the member/observer to contribute fully to CESSDA activities". (CESSDA. 2014)[88]

---

[86] DwB D7.1 pp. 66–67, 76.

[87] For a list of tools and the DDI versions they support, see the DDI Alliance website: http://www.ddialliance.org/resources/tools

[88] http://www.cessda.net/export/sites/default/about/docs/Annexes-to-Statutes-for-CESSDA-210213-Final-Version-brand.pdf

However, changing existing processes, rules and standards require long-term strategic and operational planning among the involved and cooperating organisations.

For B, the most obvious solution to this problem would be the universal introduction of Lifecycle as the only DDI instance, but in reality, a situation with mixed usage of versions (within as well as between DAs) is likely for several years to come.

The development of a strategy for "best practices on pros and cons" in applying one of the current DDI production-lines where sensible – considering e.g. minimal (risk for) loss of information and required investments – would likely increase the internal efficiency to produce and manage the data and metadata and to provide sufficient "products" that support research needs for high quality data.[89]

Relevant considerations and planning to apply a certain DDI solution might use the DDI Working Paper Series[90], which currently includes papers on new and emerging themes as well as papers on best practices and use-case literature. Topics concern for instance papers on "Best Practices Across the Data Lifecycle"[91] and "Best Practices for Longitudinal Data"[92].

For local use of DDI 3x corresponding Best Practices describe the creation of a local DDI 3.0 Profile, which is a subset of DDI 3.0 fields to be used by an organization or shared by a community of users.

The use of DDI-Profiles is a recommended solution in case where two GSIM implementations using DDI might not interoperate. GSIM concludes:

> "organization or application can specify exactly how it uses the DDI XML formats. It is anticipated that as organizations implement GSIM, these standard profiles can be used to form a base set of DDI elements for interoperable use."[93]

Further DDI Metadata Resources[94] provide mapping or transformation solutions of DDI-C and/or DDI-L, Dublin Core, and MARC21 XML formats and inform on recommended DDI elements used at organisations like ICPSR or CESSDA.

**Problem 2: Need for data versions and persistent identifiers**

In DwB D7.1, the need is addressed for

> "metadata for versioning (version history etc.) and persistent identifiers to support data citation and re-use […] require complementary efforts in the domain of OS data" (p. 79).

---

89 For a more detailed comparison between DDI Codebook and DDI Lifecycle, see DwB D7.1 pp. 23–25.

90 http://www.ddialliance.org/resources/publications/working/all

91 http://www.ddialliance.org/resources/publications/working/BestPractices/DataLifeCycle

92 http://www.ddialliance.org/resources/publications/working/BestPractices/LongitudinalData

93 http://www1.unece.org/stat/platform/display/gsim/Implementing+GSIM

94 http://www.ddialliance.org/metadata-resources

The DDI Alliance Best-Practice paper on Versioning and Publication underscores this need, pointing out that

> "[clear] and consistent versioning processes for metadata and data must be adhered to" and that the "versioning process should be clear and transparent to end users so they always know which version of the metadata and data they have acquired."(DDI. Publications. DDI Best Practices. 2009. No.8, p. 7)

Wherever versioned data are provided there is need for standardised and accessible information on which version the data is in, the version date, connection to and differences from previous versions and the persistent identifier (PID) that refers to that specific version. Persistent identifiers are currently not used extensively at the NSIs and not by all DAs, which makes it more difficult for the researcher to identify, retrieve and cite datasets (Jackson. 2012)[95].

A reliable PID system implicates "real persistence of all metadata and data ever published,"[96] but there are currently a variety of possible PID systems available, such as "handles" (hdl) and "digital object identifiers" (DOI).

In the current DDI Lifecycle version (DDI Lifecycle 3.2. 2014), there is a number of elements that allow for versioning metadata, such as[97]

- DataFileVersion, which provides the version information for the data file related to this physical instance;
- VersionDistinction, which describes the data versioning scheme(s) used by an organization; and
- VersionRationale, which is a textual description of the rationale/purpose for the version change and a coded value to provide an internal processing flag within an organization or system.

  There is no dedicated element to use for PIDs, but DDI offers certain specific elements (for example the UserID element, which can be used with any relevant ID type) that can be employed for PIDs as well.

However, not all data come in versions. At NSIs it is commonplace practice to put together a dataset from varied sources at the request of some project, rendering the idea of a dataset's version useless since the sources of the included data may in turn have different versions. It is a more dynamic way of organising data than the common data archive division into rigid datasets that do evolve but are not commonly divided and put together in different combinations.

---

[95] http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=STD/CSTAT/MICRO%282012%2914&docLanguage=En

[96] CESSDA PPP D9.4, p. 6. See also DwB D7.1 pp. 28-29, 76, 79.

[97] http://www.ddialliance.org/Specification/DDI-Lifecycle/3.2/XMLSchema/FieldLevelDocumentation/

**Problem 3: Documentation of contextual information/provenance**

DwB D8.4 found another interesting problem that may belong to the area of metadata:

> "Some interviewees indicated a more fundamental complexity with comparative research. One needs to have a basic kind of knowledge about the national context of a country to be able to interpret the results of the analysis. Reports with key indicators are essential in this respect." (p. 14 in DwB D8.4)[98]

This problem results from the fact that different countries, cultural, legislative and political systems divide the world into varying sets of concepts and this division leaves traces in the data collected in the different countries.

For example, one kind of data may be publicly available in one country but not in another, because the two legislative systems differ. This will in turn result into different labelling of what is essentially the same kind of data when it comes to its content. Another example is data collected by an official agency that exists in one country but not in another. These data may be superficially comparable to data collected by a research institute in the other country, but on closer scrutiny it may turn out that the data collected by the agency cover the entire population while the research institute's data cover only a subset, resulting in comparative statistical analyses of samples and of populations.

Considerations and questions are part of substantial scientific work of the Cross-cultural Research community. For instance, the Cross-Cultural Survey Guidelines (CCSG)[99] informs on related issues and formulates Best Practises for the conduct of comparative survey research across cultures and countries. The guidelines cover a large number of metadata-driven documentation aspects ranging from Study Structure to Data Dissemination, and support the analyses and re-use of such complex data in future comparative research.

Communication between social-science researcher and metadata-standard developers is therefore required in order to bring together research needs and well-documented data. For example, the session "The Role of Structured Metadata in Cross-national Surveys"[100] discussed this issue at RC33 – 8. Conference on Social Science Methodology.

### *5.1.2 Controlled vocabularies and classification systems*

Within this key area issues connected to controlled vocabularies and other classification systems used in metadata organisation are addressed. Having a metadata standard for the description of data is essential but is generally not enough as the standard does not provide detailed guidance on how the elements are to be filled out. Instead, it is necessary to employ some sort of classification.

---

[98] http://dwbproject.org/export/sites/default/about/public_deliveraples/dwb_d8_4_portal-resource-discovery-functionality_final-report.pdf

[99] http://ccsg.isr.umich.edu/index.cfm

[100] http://conference.acspri.org.au/index.php/rc33/2012/schedConf/schedule

**Problem 1: Inconsistent usage**

The data can hardly be described in a consistent way, when NSIs and DAs are inconsistent in their usage of controlled vocabularies (henceforth *CVs*). This will in turn adversely affect the researchers' work at several levels, for example making it harder to retrieve, understand and compare the research data. Other disadvantages involve diminishing interoperability and cost-effectiveness.

For consistency to prevail, several aspects need to be considered. One of the most important features to deal with concerns the use of controlled vocabularies. Initially, the most appropriate CV must be chosen. On the one hand, this might be an apparently easy matter for one NSI/DA alone. On the other hand, other NSIs/DAs who want to represent the same concepts may choose different CVs. DwB D7.1 (2013, p. 45) underscores how their use of vocabularies is not homogenous but diverse.

Research data should ideally be described by different national and international organisations using a set of standardised controlled vocabularies covering all applicable concepts. While this might be considered a Best Practice in an ideal world, certain factors may make it impracticable (under some circumstances) to agree on a common set of vocabularies.

Best practice, then, cannot always involve a standardisation process by which all NSIs/DAs end up using the same vocabularies. The challenge is how to achieve the highest possible metadata quality without being able to rely on a unique set of shared vocabularies.

A first step to resolve this problem would be to state explicitly, which are the chosen vocabularies, as this will allow the researcher to understand and compare the same concepts across different data providers. Even if vocabularies and classification systems are used consistently, the metadata elements may still be unclear for the researcher as long as it is not obvious which the chosen standards are.

At the moment, it is not always obvious where a particular metadata item comes from:

- was it taken from a controlled vocabulary,
- did it occur in some other documentation or
- was it conveniently made up on the fly?

This sort of ambiguity can be removed by, regardless of the origin of the item, expressing its source in a transparent manner directly in the portal.

When analysing obtained search results, it would also help if the data portal indicates terms related to the search term by displaying "broader, narrower and related terms of the term the user has used in his/her search". (Jääskeläinen, et al. 2009, p. 36) [101]

Furthermore, transparency is not only beneficial *after* having obtained a search result, but also *before* carrying out the search for relevant data as it will enable the researchers to improve their query formulation.

[101] http://www.iassistdata.org/downloads/iqvol3312wackerow_0.pdf

While letting researchers know what CVs have been used is good, letting them know how to map them is better. Explicitly stating what CVs have been employed is a fairly practical solution, but the obvious drawback is that researchers will have to go through and compare the datasets manually.

A more efficient solution (for the researchers) would be to implement a harmonisation process in which the vocabularies' different concepts are properly mapped onto each other. This only works in some cases, however; but when the reasons for using different CVs depend on structural or conceptual differences, it may be hard or impossible to map two CVs to each other.

**Problem 2: Need for coordinated management of controlled vocabularies**

It is fair to assume that different metadata professionals in many situations will identify and choose an identical set of CVs to describe a particular dataset. In all probability, however, they would sometimes end up having chosen disparate CVs.

Perhaps this will most often occur when there are two or more CVs, which appear to be (more or less) equally appropriate in the process of characterising a dataset, and a rule needs to be formulated favouring one of the CVs.

In this regard, OECD (2007, p. 20) has emphasised the importance of senior management within each organisation

> "to ensure that appropriate practices and principles involving the use [of] consistent terminology are developed and adopted across the organization".[102]

Without such guidelines,

> "the costs of producing, updating and maintaining vocabularies might become greater than the benefits". (DwB D7.1, p. 45)

Moreover, the management must be a process coordinated at an international level involving the relevant NSIs and DAs.

Just as metadata professionals within an organisation may end up using different CVs for a particular set of metadata items, lack of international coordination will lead to the same type of heterogeneous usage *between* different data providers. However, having agreed on a Best Practice for a particular metadata item does not imply that the selected CV will permanently be the most appropriate one.

NSIs and DAs have to be prepared to revisit their current best practices as a result of societal changes, emerging research questions, or any other essential factor that have an effect on what is perceived as the best way to describe a particular data category.

[102] http://www.oecd.org/std/37671574.pdf

### *5.1.3 Some practical aspects*

This key area is the most practical in that it gathers issues that require less planning on the level of abstract models and more work on a practical technical level.

**Problem 1: Non-machine-actionable formats**

DwB D7.1 repeatedly accentuates the importance of and need for machine-processable metadata.[103] Such requests are also easily found elsewhere in literature dealing with (Best Practice of) metadata (e.g. Neiswender et al. 2013; DDI Best Practices. 2009. Controlled Vocabularies; Jääskeläinen et al., 2009).

There are several reasons why it is a good idea to store and present metadata in machine-processable formats, and it is beyond the scope of this report to give a detailed review of them all. In short, it brings advantages both to the data provider (e.g. cost efficiency, interoperability) and the user (e.g. retrievability, comprehensibility).

Storing metadata according to either a single standard or a set of standards mapped to each other is necessary in order to obtain machine-processability, but the storage also has to build on certain technical formats that make the data elements machine-processable. Unfortunately, data providers sometimes still handle their metadata in non-machine-processable formats such as PDF and Word.[104]

This usage results in suboptimal work within the NSIs and DAs as well as among researchers (e.g. Problem 2 below), and thus cannot be considered a part of a Best Practice for metadata handling. On the contrary, it is essential that every NSI and DA endeavour to revise all their working procedures so that they comprise machine-processable data formats. There are a number of existing formats of this kind, such as XML, which constitutes the basis for both DDI Codebook and DDI Lifecycle. Hence, there is little need for developing new formats. The issue rather concerns an adaptation process, in which NSIs and DAs conform their microdata handling to the XML Schemas, prescribed either by the current DDI versions or to another machine-processable format that can be mapped against those.

**Problem 2: Non-integrated search capabilities**

Another practical aspect discussed in DwB D7.1 concerns the need for European social scientists to have easy access to research data across the European Union.[105] Importantly, the need goes beyond mere accessibility; it is equally crucial to promote understanding and comparability of the data in order to let researchers conduct efficient and high-quality research.[106] The current situation does not fully meet these requirements as much of the

[103] DwB D7.1, pp. 14, 41, 44-45, 50, 78-80.

[104] DwB D5.2, pp. 3-4; Deliverable 7.1, pp. 66, 78.

[105] DwB D7.1, pp. 6, 8-9, 70, 79, 81.

[106] DwB D7.1, p. 11.

European social-science microdata are not accessible within a pan-European integrated system. Even if researchers ultimately may gain access to the data via local systems of NSIs and DAs, an integrated search system would bring about significant advantages at several levels.

One way to achieve this would be a solution similar to, or an extension of, the existing CESSDA portal which

> "is an example of integration of data in heterogeneous, autonomous resources (data archives) by using harmonised descriptive metadata represented in a common metadata standard, and using controlled vocabularies and code schemes."[107]

Visiting the CESSDA portal, researchers can carry out an integrated and multilingual search among microdata held at European DAs thanks to a common metadata standard agreement. Having obtained interesting lists of datasets, the researcher has the possibility to continue the exploration at the respective DAs, which is referred to directly in the portal.

In contrast to the integrated search function among the resources of the DAs via the CESSDA portal, each NSI keeps most of their OS microdata exclusively in their own systems; an issue that is further addressed in Work Packages 8 and 12 of the DwB project but beyond this report.

## 5.2 Use-case

This use-case is designed to give the reader the researcher's perspective on the implementation of the rules and best practices discussed above. It aims at showcasing a "best case" scenario, that is, how would access to NSI and DA data from several European countries work if each of the identified problems were satisfactorily addressed. The use-case will be the story of a researcher interacting with a future infrastructure for accessing multi-national social science data interspersed with comments on the rules and best practices at work in each step of that interactive process.

Dr. Danica Kovac from the Charles University in Prague is working in a research project investigating correlations between the financial situations of national education systems and the youth unemployment rates in a selection of European countries. She suspects that relevant data are available from both NSIs, NSAs and DAs across the EU.

However, before beginning to check the web pages of all the institutions that are relevant to her project she first visits the web page of the Czech Social Science Data Archive in order to test the search functionality and develop a strategy for searching and retrieving data using her mother tongue.

At the archive's search page, Dr. Kovac discovers that the search can be extended to include the Czech Statistical Office. However, she also makes another discovery: there is a recently implemented one-stop search portal that provides access to data from NSIs, NSAs and DAs

[107] DwB D7.1, p. 9.

from all over the European Union. This could have major implications for the process of gathering (and analysing) data for her project.

She begins by looking for general descriptions in the portal, in order to identify relevant datasets. To get a first impression of how the portal works, Dr. Kovac enters some English-language search-terms relating to economic factors that are known to have an impact on education systems. As her search terms are rather broad she gets a great many hits, a considerable number of which are not relevant to her project. To narrow down her search she uses the portal's dynamic faceting functionality, which, among other things, enables her to select only the countries, and the time frame she is interested in.

She now takes a closer look at a few hits in her list and realises that she only gets very basic information about the studies and datasets as long as she is not logged in. She then discovers that she may use the login to her own university network to log in on the portal and during the process of logging in she is informed that she may use Czech or any other European language to conduct her searches as the portal is multilingual. Once logged in, she can access all the metadata and documentation for the studies and datasets she is interested in while also getting higher precision by using Czech search terms.

Using her own language brings out search term creativity and she begins applying more complex text search techniques such as Boolean and other search operators, yielding more precise search results. There is also a drop-down menu to select the types of data she is looking for.

After some trimming by means of the different search tools she has found, among others, a Portuguese study that looks promising. The information about the study can be retrieved in Czech thanks to the portal's multilinguality feature, but she soon realises that even when she is logged in the information is still too rudimentary for her to decide on whether the study is really worthwhile taking a closer look at. Now she finds that there is a section for related documentation where she finds a codebook that can be downloaded. The codebook turns out to be available in Portuguese and English and by means of the latter, she is soon able to determine that in fact the study is not relevant to her.

Continued browsing of her search results reveals that most studies have questionnaires, codebooks, methodology reports and/or quality reports available for download directly from the portal, most of them including an English-language version.

After some further browsing she has found a number of studies from five European countries that she thinks might be comparable and therefore of high relevance to her comparative project. Time to examine their metadata more closely. She selects them for comparison employing the portal's built-in compare function. After some processing she finds that the variables she is interested in examining closer are present and comparable in four of the studies she had selected. This is a great result for just over an hour of work.

But in order to be able to assess the validity of the comparison she does not rely on the portal's compare function alone, she needs to know more about the national context where the data were gathered. She turns to the portal's background section where she finds a helpful starting point for her background research: most of the NSIs have linked to various reports containing information on a range of key indicators relevant to the content of the studies in question. From there she can single out the key indicators relevant to her own understand-

ing of the data and by following references and beginning a new query at the portal she can find publications related to these key indicators.

Now Dr. Kovac wants to download the datasets she has singled out to begin analysing them in her own setup of analytic software. The portal, being a portal and not an archive, does not host the files but only the information about them. There are, however, download links in each study description that lead directly to the national service provider's download page or ordering form. Before too long, she has access to the microdata that she needs and can proceed with her comparative data analysis.

# 6 DISCUSSION

The focus in this report is the importance of common metadata standards to facilitate access to data for research purposes. Besides this, the report initially emphasizes the importance of continued work of harmonization at European level in order to increase the comparability of data from different data sources in different countries. This work is outside the scope of this report but consistency in the form of comparable statistical units, classifications and definitions is an important prerequisite for the successful introduction and use of technical metadata standards.

In this chapter, we will revisit the central objects of discussion in the report and briefly examine them in terms of what their current status is and what development we can expect in the near future, but also what we believe will happen in the long-term, including some visions of what the future might hold.

## 6.1 Standards of future relevance

Ultimately, this report explores the many facets involved in selecting, implementing, managing, and employing the wide variety of metadata and related standards available to NSIs, NSAs, and DAs that produce, archive, or disseminate microdata of interest for researchers. Much of the report boils down to a discussion about standards – mainly but not only metadata standards. How to select a standard capable of living up to all the demands that can be placed on it, now and in the future, is therefore an issue of central concern. Before we turn to the main discussion, we will therefore raise a point regarding the quality assessment of metadata standards (applicable to standards in general).

In section 1.3 we list questions that can be used as a basis for assessing metadata standards. Although broad, there is no guarantee that the list is comprehensive; but it is certainly extensive enough to warrant a strategy when approached. Answering every question in a way that ensures reliable answers would require a great deal of work and depending on the resources available to the assessing organisation, it is recommended to begin any assessment process by ranking the questions in terms of relevance.

This report provides no methods for how to answer these questions; what is a suitable method depends on who assesses and for what purpose. It is also important to be aware of the fact that the questions do not touch upon the implementation of a standard but only help to evaluate the standard as a tool. The implementation of a standard will have a major impact on the quality of the data product and must be guided by rules and best practices that have no connection to metadata-standard quality-assessment.

## 6.2 A temporal perspective: from current situations to future visions

In any discussion that purports to examine what the future might hold, it is worthwhile to start by looking at the situation today before turning to the situation tomorrow and projec-

tions about an even more distant future. Table 3 summarizes the key points of discussion in this report and what can be said about their current and future situation. Each object is then discussed in more depth below.

Table 3: Central objects of discussion in a short- and long-term perspective

| Objects | Current and Short-Term | Visions and Long-Term |
|---|---|---|
| Conceptual/Information models and terminology | DDI, SDMX (some GSIM implementations) | GSIM is established as conceptual model for describing implementations of SDMX and DDI. |
| Metadata models | NSAs and NSIs have different models.<br><br>Minimal common metadata model and harvesting according to that agreed minimal model.<br><br>DDI is the best way forward for microdata. Harvesting via API.<br>NSIs convert internal format to DDI (e.g. from Neuchatel model to DDI).<br>Only some combinations possible (e.g. surveys that are standardised by EU regulation). | Metadata exposed with semantic web approach (RDF triplets).<br><br>SDMX-based metadata on macro level can be used for finding microdata for research<br><br>Combination of data sets possible through community adopted models and vocabularies. |
| Big Data | We don't know the origin, the format or the domain values. | Most likely the same as today. |
| Administrative data | Data is (sometimes) delivered with metadata from the origin. | The delivered data is metadata-driven. |

| Objects | Current and Short-Term | Visions and Long-Term |
|---|---|---|
| Versioning | Versioning of datasets and vocabularies widespread in DAs.<br><br>Metadata standards have extensive capacity to document versions and versioning information. | Thoroughly documented versioning is standard at all data archives. |
| Timestamp<br>of microdata | Existing technique that is used at NSAs and NDAs. | More spread and the foundation for a Data Warehouse operation. |
| PIDs | Existing technique that is used mainly at DAs to refer to specific version of a dataset. | Ubiquitous use among DAs, and more widespread use within NSIs/NSAs where appropriate. A single PID standard is adopted. |
| Process data | Different solutions in different organisations. | A standard has developed that many organisations are using. |
| Linked data | Some data producers are testing it today. | The most of the data are available in this format. |
| Vocabularies | ELSST (language-independent classifications),<br>DDI-Discovery RDF,<br>EU- and UN-regulated classifications. | Well established and agreed vocabularies via SKOS, XKOS and OWL. |

## Conceptual/information models and terminology

There are different levels of information that researchers need to access during the course of the research process. A layered model of three levels of researcher needs comprise the general description of the data, the metadata itself, and eventually real microdata access.

One problem that is described in this report is the use of different terminology within NSIs, NSAs and DAs. We have identified GSIM as the most promising standard in this area. GSIM

has the potential to become a common language that can enhance communication at different levels. However, it is important to note that GSIM is still in an early stage of implementation and mainly used within the NSI domain. Being a conceptual model, GSIM is also dependent on standards like DDI and SDMX for implementation in a production environment. The models for mapping between GSIM and standards such as SDMX and DDI already in existence can be expected to be developed further as people from the SDMX, DDI and GSIM communities take part in the GSIM development.

Based on the long-term vision for GSIM, we conclude that solutions to transnational data access must be based on the premise that data producers will have different internal metadata standards for a long time yet. Data-access solutions must therefore be based on the mapping of internal standards to a common standard. We suggest that DDI is the most viable such standard, taking into account that our focus is access to microdata. This would also bring benefits to the data producers, who can expect to reap benefits from developments of DDI standards and tools for documenting research and survey data.

## Metadata models

As many producers of data have a production model that goes back a long way – predating the models we are discussing here – there is no overall model that applies for all NSIs. There is a general trend, which, in combination with common production models, points to DDI as the future standard for microdata.

### DDI as a standard for microdata

Some promising standards have already been developed and among these, the DDI Discovery Vocabulary appears to be particularly interesting as it supports metadata linkage in the microdata domain. Initial work that has been carried out in another DwB work package (WP12) also confirms that RDF has desirable properties regarding semantics, extensibility and linkage. WP12 therefore plan to use DDI Discovery Vocabulary for the internal metadata model that will support the Discovery Portal soon to be developed.

### SDMX for aggregated data

In a long-term perspective, the development of RDF representations of SDMX can also be expected to add value for the research community. Since the SDMX model focus on aggregated data, RDF-based representations could be valuable when no microdata is available or when a researcher wants to examine the provenance and/or get a picture of the processes that were executed in creating the aggregated statistics from the microdata. However, as we have described previously in this report, for the time being very few NSIs and NSAs have a metadata system that supports integrated documentation of the entire production process, thus linking the aggregated output to the source variables on the microdata level.

## Big Data

One question many organisations are working with today is the new and future use of Big Data. These “new” data can be used either as a new source of information or as a comple-

ment to existing sources. From a metadata and quality perspective these Big Data have unknown origin, formats and source and these problems will also occur in the future. We also notice that the future use of Big Data not only is a metadata issue but also a statistical methodological issue.

### Administrative data

Administrative sources for microdata in the production of statistics and/or in a research project demands the knowledge of that the data has been produced for other purposes and that data on a microlevel maybe not is totally accurate. Today many primary NSIs have or are in the process of receiving microdata from administrative sources. To improve the quality of these data, we see that they have to be metadata-driven so that the producer of the data clearly shows what data there are.

### Versioning of datasets

The rather rigid datasets commonly held at data archives lend themselves to traditional X.X-type versioning, which also works well with PID systems. However, the dynamic sets of data produced at NSIs change too much to be considered versions of each other in the traditional sense (see Timestamp of microdata, below). DAs have developed rules and best practices for handling versioning of datasets, and metadata standards can handle dataset versions. It is fair to assume that such capacity will only get better in the future.

### Timestamp of microdata

Many NSIs and NSAs use the technique with timestamp on microdata in order to produce a version of a dataset for a given time period. From a metadata perspective this alternative demands that the classifications, including the domains for each variable are updated and accessible for the research users.

### Persistent identifiers

Presently, the use of persistent identifiers is widespread if not ubiquitous among data archives. This is no doubt partly due to the work done by for instance DataCite to encourage the use of PIDs (in their case, DOIs) to facilitate citation of datasets.

Currently (June 2014), over 3.4 million datasets have been provided with DOIs through DataCite (Statistics Beta)[108]. Starting in 2012, Thomson Reuters will include the Data Citation Index in their Web of Science (Thomson Reuters. 2012, p. 2)[109]. It will allow data publication to affect bibliometric impact studies and their use of DataCite's metadata for the index provides an added incentive for researchers to use repositories that provide DOIs (other PIDs

[108] DataCite Statistics Beta. June 25, 2014. http://stats.datacite.org/

[109] http://wokinfo.com/media/pdf/DCI_selection_essay.pdf

have similar incentive structures) (Robinson. 2013)[110]. Taken together, this suggests that the use of PIDs will keep spreading in the future.

### Process data

This report observes how many data producers implement metadata standards mainly in order to create a more efficient production process and to increase the possibilities of sharing tools and methods. The self-interest of this approach notwithstanding, in the long run this development also creates better conditions for researchers to understand the data content. The reason for this is that metadata that is used to drive the production process can also be provided as process data that can be of great interest in a research context, as the researcher can see the changes.

### Linked Data: an interface on the rise

Linked Data has a great potential for the research community when it comes to discoverability and linking of disparate datasets and also non-quantitative resources like papers and publications.

However, achievement of the goal to make microdata and statistical data available as linked data, is not preliminary a technical challenge. The challenge is to get the global statistical community and DAs to agree on common naming and metadata standards.

### Vocabularies

Today, the importance of controlled vocabularies is well-established in the (meta) data-producing communities. Standard CVs are being improved or are under development: the DDI CV has released a number of vocabularies for its elements and has stepped up the development and publication progress. Language-independent or multilingual classifications are also under constant development, such as the European Language Social Science Thesaurus (CESSDA-ELSST. 2012-2014)[111].

Although XML formats dominate today, turning vocabularies into linked-data format is on the planning or initial stage (for example the DDI-RDF Discovery Vocabulary [Disco]).

EU- and UN-regulated classifications are available in several areas. In the long term, we can expect key vocabularies to become ubiquitous in metadata production for microdata, and vocabularies and classifications will be defined through linked data, using standards such as SKOS, XKOS and OWL.

---

[110] http://www.slideshare.net/datacite/2013-datacite-summer-meeting-thomson-reuters-data-citation-index-cooperation-nigel-robinson-thomson-reuters

[111] http://ukdataservice.ac.uk/about-us/projects/cessda-elsst/details.aspx

# 7 CONCLUSION

This report discusses metadata standards with future relevance with respect to needs and key areas of the evolving European Social Science data infrastructure. The report also includes an overview of rules and best practices to handle metadata needs in respective key areas. As is made clear from the Description of Work, the scope of this deliverable does not include operative work on particular metadata schema, information models, or resource discovery functionalities; such issues are addressed by WP5, WP8 and WP12.

As is widely agreed, DDI and SDMX are the major metadata standards to consider when developing and enhancing the internal metadata-driven production-systems at NSIs, NSAs, and DAs at large. However, to support the development of an extended "European database" for comparative research, it seems necessary to develop a more compatible metadata system that allows standard-oriented processing and documentation. It also appears necessary to establish advanced access and retrieval capabilities to existing and upcoming data types of interests in the future. Examples include register data from national agencies or data compiled from different internal data sources with rich potential for analysis.

In addition to the need of metadata standards there is also need for international standardisation of metadata values, in other words international controlled vocabularies that can be used both on the level of studies and on that of variables. These international controlled vocabularies should contain URIs for each term so they could be straightforwardly used to form linked data. To have internationally standardised metadata values both in the (meta) data of NSIs and DAs will require a great deal of work to be done. Especially on the variable level (except for the variables that are EU-regulated) there is need for contribution of registrars and researchers who collect the datasets.

The prerequisite to develop enhanced and integrated systems to access data and metadata from both NSIs and DAs is a common understanding about central objects and production processes. Therefore GSIM and GSBPM act as "communicators and translators" in discussion of conceptual-level needs. An example would be the options and requirements of a potential project that aims to improve common access to data from both NSIs, NSAs and DAs to support research needs. Furthermore, looking at the current metadata situation it appears sensible to enhance the usability of available national data resources (at NSIs, NSAs and DAs) for comparative research purposes as well. There are several means and tools available that foster a more standardised, interoperable, machine-actionable management and usage of metadata for various purposes.

This report covers major findings on three levels of abstraction. On the conceptual level it examines the needs and options in *standards development* with respect to internal and/or federated metadata production-systems in general. While GSBPM and GSIM already play an important role in system designs, these models are complemented in this report by a set of assessment criteria for metadata standards. These criteria are offered as an aid to systematic selection of metadata standards and as a reflection on quality issues in standards development beyond daily practices.

On the level of applying dedicated metadata standards for specific purposes, both DDI and SDMX are core standards with specific capacity to document and manage metadata on microdata and aggregate data respectively. The upcoming RDF constitutes a new key area of

future relevance to tackle by those data providers who need or want to prepare their collections for the Linked Open Data network.

Although technical and research-driven standards are not discussed in detail in this report, they were extensively described in DwB D7.1. They are part of the application of metadata standards in general, though, and subject of the next level of abstraction in particular.

The third level in this report focuses on metadata issues and related needs for best practices for employing standards in a running data and metadata production-system. Drawing on the work in Deliverable 7.1, the report has identified a number of key areas and needs to be solved by best practices in a broader sense, whether such are already available, still in need of improvement, or remain to be developed.

- The first key area concerns how metadata and standards are used for specific purposes such as versioning and persistent identifiers for datasets; capturing the changes to data during the production process; conveying the cultural, political, or other particular context of datasets; and the general use of interoperable metadata standards in production systems.
- A second key area concerns how metadata and standards are used to manage complex and/or new types of data, including the need for standardised data management on Study and Variable level compiling integrated datasets on space and /or time dimensions for comparative analyses. The area also includes the need for preparing and managing administrative data for re-use by researchers, as well as to meet the methodological and metadata-driven challenges of Big Data.
- The third key area takes up issues about vocabularies and classification systems, including the need for harmonization to allow consistent use and the need for (senior) coordination and management of vocabularies – and in the cases where harmonization is impossible, the need for interoperability.
- A fourth key area deals with the functional aspects to improve metadata usage. This concerns improved metadata production in order to achieve fully machine-actionable formats, which can support metadata retrieval on the study and variable levels when sensible and to allow for mapping that enables integrated search capabilities.

In addition to this bundle of metadata-related key areas and issues in standards development and Best Practises, there are specifications and objectives raised in this report that may become relevant in planning forthcoming projects that assemble common needs and interests of DAs and NSIs in this field.

The outcomes of the report may also be of interest for further communication and cooperation about standards development within for instance CESSDA, the ESSnet, and the SDMX and DDI communities. To support such endeavours, this report provides an overview of needs related to data, metadata and best practices with future relevance for a European data infrastructure supporting the requirements for improved discovery and access to data for comparative scientific research.

# REFERENCE LIST

**Note**: [last visit] indicate last access to a linked resource in the reference list.

## Literature and metadata resources


AAPOR Conference. 2014.
Topic: Big Data in Public Opinion and Survey Research
http://www.aapor.org/Abstract_Book.htm [23.06.2014]

ADLS. 2014
Administrative Data Liaison Service.
Administrative data introduction
http://www.adls.ac.uk/adls-resources/guidance/introduction/ [25.07.2014]

Beall, Jeffrey. 2007.
Discrete Criteria for Selecting and Comparing Metadata Schemes.
Against the Grain, Vol. 19, Issue 1, Article 7.
http://docs.lib.purdue.edu/atg/vol19/iss1/7 [22.07.2014]

Berners-Lee, Tim. 2006.
Linked Data
http://www.w3.org/DesignIssues/LinkedData.html [22.07.2014]

Boukottaya , C. Vanoirbeek , F. Paganelli , O. Abou Khaled. 2004.
Automating XML documents transformations: a conceptual modelling based approach.
In Proceedings of the First Asian-Pacific Conference on Conceptual Modelling - Volume 31.
Pages 81 - 90.
http://crpit.com/confpapers/CRPITV31Boukottaya.pdf [22.07.2014]

Bruce, Thomas R. and Hillman, Diane I. 2004.
The Continuum of Metadata Quality: Defining, Expressing, Exploiting.
In: Hillman, Diane & Westbrooks, Elaine (eds.): Metadata in Practice, 239-256. Chicago:
American Library Association. 2004.
http://hdl.handle.net/1813/7895 [22.07.2014]

CESSDA. 2014.
Statutes for CESSDA: Annexes, CESSDA AS, 12 June 2014.
http://www.cessda.net/export/sites/default/about/docs/Annexes-to-Statutes-for-CESSDA-210213-Final-Version-brand.pdf [22.07.2014]

CESSDA- ELSST. 2012 - 2014
CESSDA European Language Social Science Thesaurus.
http://ukdataservice.ac.uk/about-us/projects/cessda-elsst/details.aspx [24.07.2014]

CCSG.
Cross-Cultural Survey Guidelines
http://ccsg.isr.umich.edu/index.cfm [19.06.2014]
DataCite.
Statistics Beta.
http://stats.datacite.org/ [25.06.2014]

DCC. 2014.
Digital Curation Centre. Disciplinary Metadata
http://www.dcc.ac.uk/resources/metadata-standards [22.07.2014]

DwB. 2013.

- Overview of all DwB deliverables:
  http://www.dwbproject.org/about/deliverables.html [22.07.2014]

- WP7 D7.1. 2013. Metadata Standards – usage and needs in NSIs and Data Archives.
  http://www.dwbproject.org/export/sites/default/about/public_deliveraples/dwb_d7-1_metadata-standards-usage_report.pdf [22.07.2014]

Ehrenström, Birgitta; Netterstrøm, Søren; Gro Hustoft, Anne; Macchi, Claude; Held, Dominique and Karge, Reinhard. 2004.
Neuchâtel Terminology Model: Classification database object types and their attributes.
Version 2.1. August 19. 2004.
http://www1.unece.org/stat/platform/download/attachments/14319930/Part+I+Neuchatel_version+2_1.pdf?version=1 [22.07.2014]
http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2013:164:0016:0019:EN:PDF
[25.07.2014]

European Commission. 2014.
Eurostat Info space. Standard Validation and Transformation Language (VTL).
https://webgate.ec.europa.eu/fpfis/mwikis/essvalidserv/index.php/Standard_Validation_and_Transformation_Language_%28VTL%29 [22.07.2014]

European Commission. 2013.
Commission Regulation (EU) No 557/2013 of 17 June 2013 implementing
Regulation (EC) No 223/2009 of the European Parliament and of the Council on European Statistics as regards access to confidential data for scientific purposes and repealing Commission Regulation (EC) No 831/2002
http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2013:164:0016:0019:EN:PDF
[22.07.2014]

Europeana Cloud
http://pro.europeana.eu/web/europeana-cloud [24.07.2014]

- D1.5 Expert Forum Tools & Content for Social Science Research
  http://pro.europeana.eu/documents/1414567/2240207/D1.5+Expert+Forum+Tools+%26+Content+for+Social+Science+Research [24.07.2014]

Eurostat. 2014a
Scheveningen Memorandum
http://epp.eurostat.ec.europa.eu/portal/page/portal/pgp_ess/0_DOCS/estat/ SCHEVENIN-GEN_MEMORANDUM%20Final%20version.pdf [25.07.2014]

Eurostat. 2014b
Big Data in Official Statistics 2014.
http://www.cros-portal.eu/content/big-data-event-2014 [25.07.2014]

Eurostat. 2014c
CONCEPT PAPER (DRAFT VERSION v0.3).
Modernization of European Official Statistics through Big Data methodologies and best practices: ESS Big Data Event Roma 2014
http://www.cros-portal.eu/sites/default/files//Concept%20paper%20ESS%20Big%20Data%20Event.pdf
[25.07.2014]

Eurostat. 2012
Analysis of the future research needs for Official Statistics
Theme: General and regional statistics
Collection: Methodologies & Working papers
European Commission. Luxembourg: Publications Office of the European Union. 2012
http://dx.doi.org/10.2785/19629 [24.07.2014]

Eurostat. 2007. 2011.
Task Force on Core Social Variables.

- Final Report 2007.
  http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-RA-07-006/EN/KS-RA-07-006-EN.PDF [22.07.2014]

- Implementing core variables in EU social surveys. Methodological guidelines. 2011.
  http://epp.eurostat.ec.europa.eu/portal/page/portal/information_society/documents/Tab/CORE%20VARIABLES%20UPDATED%20GUIDELINES%20May%202011.pdf
  [22.07.2014]

Eurostat. 2007. 2010.
Household Budget Surveys in the EU
http://epp.eurostat.ec.europa.eu/portal/page/portal/household_budget_surveys/Data
[24.07.2014]
Related sources

- COICOP – Classification Of Consumption by Purpose
  http://unstats.un.org/unsd/cr/registry/regcst.asp?Cl=5 [24.07.2014]

- Improving data Comparability for the next HBS round (2010)
  http://epp.eurostat.ec.europa.eu/cache/ITY_SDDS/Annexes/hbs_esms_an5.pdf
  [24.07.2014]

Greenberg, Jane. 2014.
Metadata Capital via a linked data HIVE.
Advances in Classification Research Online, North America, 24, Jan. 2014.
https://journals.lib.washington.edu/index.php/acro/article/view/14680/12320 [23.05.2014]

Greenberg, J., Swauger, S., and Feinstein, E. 2013.
Metadata Capital in a Data Repository.
in DC 2013: Proceedings of the International Conference on Dublin Core and Metadata Applications. Lisbon, Portugal, September 2-6, 2013.
http://dcpapers.dublincore.org/pubs/article/view/3678/1901 (10.6.2014)

Gregory, Arofan. 2011.
The Data Documentation Initiative (DDI): An Introduction for National Statistical Institutes.
Open Data Foundation.
http://odaf.org/papers/DDI_Intro_forNSIs.pdf [22.07.2014]

Hey, Tony; Tansley, Steward and Tolle, Kristin. 2009.
Jim Gray on eScience: A Transformed Scientific Method.
In Hey, Tony; Tansley, Steward and Tolle, Kristin, Eds.: The Fourth Paradigm. Data-Intensive Scientific Discovery.
http://research.microsoft.com/en-us/collaboration/fourthparadigm/ [23.5.2014]

Hider, Philip. 2012.
Information Resource Description. London: Facet Publishing.
http://www.facetpublishing.co.uk/title.php?id=6671&category_code=102 [22.07.2014]

IPUMS international.
Integrated Public Use Microdata Series, International
https://international.ipums.org/international/ [22.07.2014]

Jackson, Paul. 2012
Microdata Exchange and the challenges of open data and transparency.
Paper presented for the OECD Expert Group for International Collaboration on Microdata Access, Paris, December 2012.
http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=STD/CSTAT/MICRO%282012%2914&docLanguage=E [21.07.2014]

ISCED. Standard Classification of Education

- UNESCO Overview page
  http://www.uis.unesco.org/Education/Pages/international-standard-classification-of-education.aspx [22.07.2014]
- ISCED-F (2013). ISCED Fields of Education and Training.
  http://www.uis.unesco.org/Education/Documents/isced-fields-of-education-training-2013.pdf [22.07.2014]

Jääskeläinen, Taina; Moschner, Meinhard and Wackerow, Joachim. 2009.
Controlled vocabularies for DDI3: Enhancing machine-actionability.
In IASSIST Quarterly Spring - Summer 2009, 2009.
http://www.iassistdata.org/downloads/iqvol3312wackerow_0.pdf [25.06.2014]

LOD2 – Creating Knowledge out of Interlinked Data

- LOD2. WP9 Use Case 3: LOD2 for Citizen: PublicData.eu
  http://lod2.eu/WorkPackage/wp9.html [22.07.2014]

- LOD2. 2012. WP9. Deliverable 9.5.1. Establishment of the Serbian CKAN
  http://static.lod2.eu/Deliverables/LOD2_D9.5.1_Serbian_CKAN.pdf [22.07.2014]

Lancaster, F.W., 1991.
Indexing and Abstracting in Theory and Practice.
London: The Libarary Association.

Laney, Douglas. 2001
3D Data Management: Controlling Data Volume, Velocity, and Variety. Meta Group
http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf [25.07.2014]

Laney, Douglas. 2012
The Importance of 'Big Data': A Definition. Gartner.
http://www.gartner.com/resId=2057415 [25.07.2014]

NSF. 2014
National Science foundation. Critical Techniques and Technologies for Advancing Big Data Science & Engineering (BIGDATA). PROGRAM SOLICITATION - NSF 14-543.
http://www.nsf.gov/pubs/2014/nsf14543/nsf14543.pdf [25.07.2014]

Neiswender, Caryn. 2009.
What is a Controlled Vocabulary?
In The MMI Guides: Navigating the World of Marine Metadata.
http://marinemetadata.org/guides/vocabs/vocdef [22.05.2014]

Neiswender, Caryn. 2011.
Semantic Network (Relational Vocabularies).
In The MMI Guides: Navigating the World of Marine Metadata.
http://marinemetadata.org/guides/vocabs/voctypes/voccat/semanticnetwork [13.05.2014]

Neiswender, Caryn and Stocks, Karen. 2011.
Machine Readability.
In The MMI Guides: Navigating the World of Marine Metadata.
http://marinemetadata.org/guides/mdataintro/machinereadability [09.06.2014]

OECD. Glossary of Statistical Terms
http://stats.oecd.org/glossary/index.htm [25.06.2014]

OECD. 2007
Data and metadata reporting and presentation handbook
http://www.oecd.org/std/37671574.pdf [25.6.2014]

Pellegrino, Marco. 2013.
Standards landscape for micro and aggregated data: How the standards-based industrialization of statistical production fits into the picture (SDMX, DDI, GSBPM, GSIM, …).
Paper presented at the SDMX Global Conference, Paris, September 2013.
http://www.oecd.org/std/SDMX%202013%20Session%204.5%20-%20Standards%20landscape%20for%20micro%20and%20aggregated%20data.pdf [22.07.2014]

Pellegrino, Marco, and Denis Grofils. 2013.
E.S.S. cross-cutting project on Information Models and Standards.
Presentation at the METIS Work Session on Statistical Metadata, Geneva, May 2013.
http://www.unece.org/fileadmin/DAM/stats/documents/2006/WP5_slides.pptx.
[22.07.2014]

Praženka, D., Boško, P. 2011.
Combining technical standards for statistical business processes from end-to-end.
Paper presented at New Techniques and Technologies for Statistics (NTTS), Brussels, February 2011.
http://www.cros-portal.eu/sites/default/files/S4P4_0.pdf [22.07.2014]

RC33 – 8. Conference on Social Science Methodology
Session: The Role of Structured Metadata in Cross-national Surveys
http://conference.acspri.org.au/index.php/rc33/2012/schedConf/schedule [19.6.2014]

RDA. 2013.
Research Data Alliance: Metadata Standards Directory Working Group
https://rd-alliance.org/working-groups/metadata-standards-directory-working-group.html
[22.07.2014]

RFID - Radio Frequency Identification
Definition of RFIDat French National RFID Center
http://www.centrenational-rfid.com/definition-of-rfid-article-71-gb-ruid-202.html
[21.07.2014]

Robinson, Nigel. 2013
Discovery, Access, and Citation of Published Research Data: The Data Citation Index – Partnership with DataCite
Presentation, DataCite Summer Meeting, Washington D.C., 19–20 December, 2013.
http://www.slideshare.net/datacite/2013-datacite-summer-meeting-thomson-reuters-data-citation-index-cooperation-nigel-robinson-thomson-reuters [25.06.2014]

SDMX Standards.
Section 1. Framework for SDMX Technical Standards.Version 2.1 (April 2011).
http://sdmx.org/wp-content/uploads/2011/04/SDMX_2-1_SECTION_1_Framework.pdf
[21.07.2014]

Studman, Brian. 2010.
A Collaborative Development Approach to Agile Statistical Processing Architecture: Australian Bureau of Statistics (ABS) Experience and Aspirations
Paper presented at the Meeting of the Management of Statistical Information Systems (MSIS), Daejeon, Republic of Korea, April 2010. Accessed April 14, 2014.
http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.50/2010/wp.3.e.pdf
[21.07.2014]

The Copenhagen Mapping. 2014.
Mapping GSIM 1.1 to DDI 3.2. May 2014. Version 0.9
http://cdn.colectica.com/TheCopenhagenMapping-Draft.pdf [10.06.2014]

UN 2007.
United Nations Economic Commission for Europe.
Register-based statistics in the Nordic countries. Review of best practices with focus on population and social statistics. NEW YORK AND GENEVA, 2007
http://unstats.un.org/unsd/dnss/docViewer.aspx?docID=2764 [21.07.2014]

Thomson Reuters. 2012
Repository Evaluation, Selection, and Coverage Policies for the Data Citation Index within Thomson Reuters Web of Knowledge.
Philadelphia, Thomson Reuters, 2012.
http://wokinfo.com/media/pdf/DCI_selection_essay.pdf [25.06.2014]

Vale, Steven. 2010
Exploring the relationship between DDI, SDMX and the Generic Statistical Business Process Model.
Paper presented at the 2nd Annual European DDI Users Group Meeting (EDDI), Utrecht, December 2010.
http://www.iza.org/conference_files/eddi10/EDDI10_Presentations/EDDI10_P2_StevenVale_Slides.ppt [21.07.2014]

W3C Recommendation. 2014.
The RDF Data Cube Vocabulary.
Richard Cyganiak and Dave Reynolds (eds.). W3C recommendation January 16, 2014.
http://www.w3.org/TR/vocab-data-cube [21.07.2014]

W3C Recommendation. 2013.
PROV-DM: The PROV Data Model - 30 April 2013. [The provenance data model]
http://www.w3.org/TR/2013/REC-prov-dm-20130430/#dfn-provenance [21.07.2014]

W3C Sematic Web. Data. 2013
http://www.w3.org/standards/semanticweb/data [24.07.2014]

W3C Sematic Web. Ontologies. 2013
http://www.w3.org/standards/semanticweb/ontology [24.07.2014]

W3C Sematic Web. RDF.2014
http://www.w3.org/standards/techs/rdf#w3c_all [24.07.2014]

Willis, Craig; Greenberg, Jane and White, Holie. 2012.
Analysis and synthesis of metadata goals for scientific data.
Journal of the American Society for Information Science and Technology 63(8):1505-1520.
http://dx.doi.org/10.1002/asi.22683 [21.07.2014]
Ward, Jonathan Stuart and Barker, Adam. 2013
Undefined By Data: A Survey of Big Data Definitions
http://arxiv.org/pdf/1309.5821 [25.07.2014]

Zhu, Hongwei & Wu, Harris. 2011.
Quality of data standards: framework and illustration using XBRL taxono-my and instances.
Electronic Markets June 2011, Volume 21, Issue 2, pp 129-139.
http://dx.doi.org/10.1007/s12525-011-0060-4 [21.07.2014]

## Presentations on Indexing and Classifications at IASISST

- IASISST 2014. Session 2F: Harmonization, Thesauri and Indexing
  http://www.library.yorku.ca/cms/iassist/program/sb2/#sb2f [10.6.2014]

- Gillman, Daniel. 2014.
  Taxonomy / Lexicon Project at the US Bureau of Labor Statistics.
  Presentation IASISST 2014, Session 2F
  http://www.library.yorku.ca/binaries/iassist2014/2F/2014_2F_Gillman.pptx
  [10.6.2014]

- Balkan, Lorna. 2014.
  Linking thesauri: ELSST as a hub for social science data terms
  Presentation IASISST 2014, Session 2F
  http://www.library.yorku.ca/binaries/iassist2014/2F/2014_2F_Balkan.pdf
  [10.6.2014]

- Friedrich, Tanja. 2014.
  Improving precision and recall in study retrieval: a concept for thesaurus-based syntactic indexing.
  Presentation IASISST 2014, Session 2F
  http://www.library.yorku.ca/binaries/iassist2014/2F/2014_2F_Siegers.pptx
  [10.6.2014]

- Katsanidou, Alexia. 2014.
  The case of CharmStats or how the process of harmonization can document itself using the right tool.
  Presentation IASISST 2014, Session 2F
  http://www.library.yorku.ca/binaries/iassist2014/2F/2014_2F_Charmstats.pptx
  [10.6.2014]

- Jensen, Jannik. 2014.
  DDA Indexing platform.
  Presentation IASISST 2014, Session 2F
  http://www.library.yorku.ca/binaries/iassist2014/2F/2014_2F_Jensen.pptx
  [10.6.2014]

- El-Haj, Mahmoud. 2013.
  Keyword Indexing with a SKOS Version of HASSET Thesaurus.
  Presentation at IASSIST 2013.
  http://www.iassistdata.org/downloads/2013/2013_poster_el-haj.pdf [10.6.2014]

- Gillman, Daniel W., Bosley, John and Fricker, Scott. 2012.
  Research on Cognitive Aspects of Classification: Effects on Metadata Practice and Standards. Presentation at IASSIST 2012.
  http://www.iassistdata.org/downloads/2012/2012_g1_gillman_etal.pdf [10.6.2014]

- Wackerow, Joachim and Orten, Hilde. 2012.
  A DDI resource package for the International Standard for Classification of Education [ISCED].
  Presentation at IASSIST 2012.
  http://www.iassistdata.org/downloads/2012/2012_g1_wackerow_etal.pdf
  [10.6.2014]

- Schaible, Johann. 2011.
  Structuring Unstructured Data Using Controlled Vocabularies.
  Presentation at IASSIST 2011.
  http://www.iassistdata.org/conferences/2011/presentation/2862 [10.6.2014]


## DDI Information


DDI Alliance. 2009.
DDI-Lifecycle Graphic "what-is-ddi-diagram_small.jpg". What is DDI?
http://www.ddialliance.org/what [21.07.2014]


### DDI 4. Moving Forward


- DDI Moving Forward Process Summary
  http://www.ddialliance.org/ddi-moving-forward-process-summary [10.6.2014]


### DDI 4. Moving Forward Project at UNECE wiki


- DDI 4. Guidance for relevant standards
  http://www1.unece.org/stat/platform/display/DDI4/Guidance+for+relevant+standards (12.6.2014)

- Process for the DDI 4 Development Project. Version 0.3 – 28 April 2014
  http://www1.unece.org/stat/platform/display/DDI4/DDI+4+Process [12.6.2014]

- DDI 4 Spints – Overview
  http://www1.unece.org/stat/platform/display/DDI4/Sprints [21.07.2014]
- DDI Lifecycle: Moving Forward: Schloss Dagstuhl, 21. – 26. October 2012
  http://www.dagstuhl.de/de/programm/kalender/evhp/?semnr=12432 [21.07.2014]
- DDI Sprint 1: Schloss Dagstuhl, October 28-November 1, 2013
  http://www.ddialliance.org/ddi-moving-forward-process [21.07.2014]
- DDI Sprint 2: Reseau Quetelet, December 5-6, 2013
  http://www.ddialliance.org/ddi-moving-forward-process [21.07.2014]
- DDI Sprint 3: Vancouver, March 2014
  http://www1.unece.org/stat/platform/display/DDI4/Vancouver+Sprint [21.07.2014]
- DDI Sprint 4: Toronto, May 2014
  http://www1.unece.org/stat/platform/display/DDI4/Toronto+Sprint [21.07.2014]

**DDI 3. DDI Lifecycle 3.2. 2014**

- DDI Lifecycle 3.2: XML Schema Documentation: Field Level Documentation
- http://www.ddialliance.org/Specification/DDI-Lifecycle/3.2/XMLSchema/FieldLevelDocumentation/

**DDI. Publications**

- Working Paper Series. Overview
  http://www.ddialliance.org/resources/publications/working/all [22.07.2014]
- DDI Working Paper Series.
  Best Practices Overview.
  http://www.ddialliance.org/resources/publications/working/BestPractices
  [22.07.2014]
- DDI Best Practices. Overview
  Best Practices Across the Data Life-cycle. Overview
  http://www.ddialliance.org/resources/publications/working/BestPractices/DataLifeCtaLifeCycle [22.07.2014]
- DDI Best Practices.
  Best Practices for Longitudinal Data. Overview
  http://www.ddialliance.org/resources/publications/working/BestPractices/LongitudinalData [22.07.2014]

- DDI Best Practices. 2009.
  Best Practices No. 8. Versioning and Publication.
  http://www.ddialliance.org/sites/default/files/bp/DDIBestPractices_VersioningAndPublication.doc.pdf [22.07.2014]

  Best Practices No. 5. Controlled Vocabularies.
  http://dx.doi.org/10.3886/DDIBestPractices05 [22.07.2014]

**DDI. RDF Vocabulary**

- DISCO - DDI-RDF Discovery Vocabulary. 2013
  A vocabulary for publishing metadata about data sets (research and survey data) into the Web of Linked Data.
  Unofficial draft September 23, 2013.
  http://rdf-vocabulary.ddialliance.org/discovery.html [22.07.2014]
- XKOS - Extended Knowledge Organization System. 2014.
  An SKOS extension for representing statistical classifications
  Unofficial Draft 28 May 2014
  http://rdf-vocabulary.ddialliance.org/xkos.html [22.07.2014]

- PHDD - Physical Data Description. 2014
  Version 0.1.
  http://rdf-vocabulary.ddialliance.org/phdd.html [22.07.2014]

**Related Publications**

- Kramer, S., et al. 2012.
  Stefan Kramer, Amber Leahey, Humphrey Southall, Johanna Vompras, and Joachim Wackerow. 2012
  DDI Working Paper Series – Semantic Web
  No. 1. Using RDF to Describe and Link Social Science Data to Related Resources on the Web.
  http://dx.doi.org/10.3886/DDISemanticWeb01 [22.07.2014]

- Bosch, T., Cyganiak, R., Gregory, A and Wackerow, J. 2013.
  DDI-RDF Discovery Vocabulary:
  A metadata vocabulary for documenting research and survey data.
  In: Proceedings of the WWW2013 Workshop on Linked Data on the Web. 2013.
  http://events.linkeddata.org/ldow2013/papers/ldow2013-paper-12.pdf [22.07.2014]

## ESSnet Information

CROS Portal.
Portal on Collaboration in Research and Methodology for Official Statistics
The CROS Portal is dedicated to the collaboration between researchers and Official Statisticians in Europe and beyond.
http://www.cros-portal.eu/page/about-cros-portal [22.07.2014]

**ESSnet Topics**

- ESSnet on SDMX - Phase II. Project 2011 - 2012
  http://www.cros-portal.eu/content/sdmx-ii-finished [22.07.2014]

- ESSnet on SDMX - Phase II.
  D3.4: DDI and SDMX Analysis
  http://www.cros-portal.eu/content/wp-3-support-sdmx-application-micro-data-handling [22.07.2014] Note: Server hosting D3.4 was not available:
  http://www.ecollect-x.eu/en/deliverables-of-essnet-project.aspx]

- Standardisation
  http://www.cros-portal.eu/content/ess-standardisation [24.07.2014]

- Big Data.
  http://www.cros-portal.eu/content/big-data [24.07.2014]
  Big Data. Publication 2013
  Pilar Rey del Castillo, Eurostat
  Reflections on the use of Big Data for statistical production
  http://www.cros-portal.eu/sites/default/files/ReflectionsUseBigDataStatisticalProduction_0.pdf
  [25.07.2014]

**ESSnet MEETS Programme**

Modernisation of European Enterprise and Trade Statistics
http://www.cros-portal.eu/content/37-meets [24.07.2014]

Overview. MEETS Projects
http://www.cros-portal.eu/content/meets-essnet-projects [24.07.2014]

- Consistency. MEETS Project.
  http://www.cros-portal.eu/content/consistency-0 [24.07.2014]

  WP3 - Minutes from the Final workshop (Deliverable 32)
  Del_32_WP3_Final workshop_Proposals.pdf

- AdminData. MEETS Project.
  http://www.cros-portal.eu/content/use-administrative-and-accounts-data-business-statistics [25.07.2014]

  AdminData. MEETS Project. Separate Project Website:
  http://essnet.admindata.eu/ [25.07.2014]

  ESSNET Admin Data Wiktionary Glossary
  http://essnet.admindata.eu/ [25.07.2014]

## METIS Statistical Metadata Information

**UNECE information / METIS-wiki**

**Statistical Metadata**

- METIS. 2011a. The Common Metadata Framework.
  Part B: Metadata Concepts, Standards, Models and Registries.
  **Statistical Classifications**
  Steven Vale (Australian Bureau of Statistics). Last edited February 25, 2011.
  http://www1.unece.org/stat/platform/display/metis/Statistical+Classifications
  [23.06.2014]

- METIS. 2011b. The Common Metadata Framework.
  Part B: Metadata Concepts, Standards, Models and Registries.
  **SDMX - Metadata Common Vocabulary**
  Steven Vale (Australian Bureau of Statistics). Last edited February 25, 2011.
  http://www1.unece.org/stat/platform/display/metis/SDMX+-+Metadata+Common+Vocabulary [23.06.2014]
- **METIS**. 2013a. GSBPM
  Generic Statistical Business Process Model, GSBPM, Version 5
  www1.unece.org/stat/platform/display/metis/The+Generic+Statistical+Business+Process+Model [11.06.2014]

- **METIS**. 2013b. GSIM Specification
  Generic Statistical Information Mode – GSIM Specification
  Version 1.1, December 2013
  http://www1.unece.org/stat/platform/display/gsim/GSIM+Specification [11.06.2014]

- **METIS**. 2013c. GSIM communication
  Generic Statistical Information Model (GSIM). Version 1.1, December 2013
  http://www1.unece.org/stat/platform/display/gsim/GSIM+Communication+Paper
  [11.06.2014]

- **METIS**. 2013d. SDMX DDI Dialogue
  http://www1.unece.org/stat/platform/display/metis/SDMX+DDI+Dialogue+-+Overview+Page [24.07.2014]

- **METIS**. 2014. GSIM Communication HLG
  Generic Statistical Information Model (GSIM): Communication paper for a general statistical audience.
  High-Level Group for the Modernization of Statistical Production and Services- HLG.
  Paper presented at Conference of European Statisticians, Paris, April 9–11, 2014.
  http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/2014/ECE_CES_2014_2-Generic_Statistical_Information_Model.pdf [11.06.2014]

**UNECE – Big Data**

- UNECE. 2013. Big Data
  Big Data HLG project**:** Big Data in Official Statistics. 2014.
  http://www1.unece.org/stat/platform/display/bigdata/Big+Data+in+Official+Statistics [23.06.2014]

  Inventory. Big Data use
  http://www1.unece.org/stat/platform/display/bigdata/Big+Data+Inventory
  [25.06.2014]

  UNECE. 2014. Big Data. Publications
  "How big is Big Data? Exploring the role of Big Data in Official Statistics"
  http://www1.unece.org/stat/platform/download/attachments/99484307/Virtual%20Sprint%20Big%20Data%20paper.docx?version=1&modificationDate=1395217470975&api=v2

- UNCE. 2013. HLG Papers
  What does Big Data mean for official statistics? - HLG Paper, March 2013
  http://www1.unece.org/stat/platform/pages/viewpage.action?pageId=77170614
  [25.06.2014]
- UNECE. 2010. HLG
  High-Level Group for the Modernisation of Statistical Production and Services - HLG

http://www1.unece.org/stat/platform/display/hlgbas/High-Level+Group+for+the+Modernisation+of+Statistical+Production+and+Services [23.06.2014]

**UNECE – Further Resources**

- 7. Lessons learned (Australian Bureau of Statistics).
  Thérèse Lalor (Australian Bureau of Statistics). Last edited May 1, 2013.
  http://www1.unece.org/stat/platform/pages/viewpage.action?pageId=78677819 [22.07.2014]

- The ABS Information Management Transformation Program: A statistical metadata perspective.
  Alistair Hamilton (Australian Bureau of Statistics). Last edited July 9, 2011.
  http://www1.unece.org/stat/platform/display/metis/ABS+IMTP [22.07.2014]

## GLOSSARY OF ABBREVIATIONS

| | |
|---|---|
| CESSDA | The Council of European Social Science Data Archives |
| CESSDA ERIC | CESSDA European Research Infrastructure Consortium |
| CESSDA PPP | CESSDA Preparatory Phase Project |
| CROS | **C**ollaboration in **R**esearch and Methodology for **O**fficial **S**tatistics |
| CROS Portal | Dedicated to the collaboration between researchers and Official Statisticians in Europe and beyond |
| CSV | Comma Separated Value. File format |
| CV | Controlled Vocabularies |
| D.x.y | Deliverable / x = Work package nr. / Y = D Deliverable. nr. |
| DA | Data Archive |
| DDI | Data Documentation Initiative |
| DDI C | DDI Codebook (DDI v.2x) |
| DDI L | DDI Lifecycle (DDI v.3x) |
| DOI | Digital Object Identifier |
| DoW | Description of Work |
| DwB | Data without Boundaries |
| ELSST | European Language Social Science Thesaurus |
| ESMS | Euro SDMX Metadata Structure |
| ESS | European Statistical System |
| ESSnet | The Centres and Networks of Excellence (Cenex, now called ESSnet) |
| ESSnet CORE | Continues the work of a previous ESSnet called CORA |
| EU | European Union |
| EU- SILC | European Union Statistics on Income and Living Conditions |
| EUROSTAT | The statistical office of the European Union |
| GSBPM | Generic Statistical Business Process Model |
| GSIM | Generic Statistical Information Model |
| HLG | High-level group for the Modernisation of Statistical Production and Services (UNECE) |
| HLG BAS | High level group for strategic development in business architecture in statistics |
| ICPSR | Interuniversity Consortium for Political and Social Research |
| ISCED | International Standard Classification of Education |
| LOD | Linked Open Data |
| LFS | Labour Force Survey |
| MCV | Metadata Common Vocabulary |
| MetaPlus | A component of Statistics Sweden's metadata system |
| METIS | Statistical Metadata; Information system (METIS-wiki) provided by UNECE |
| MEETS | Modernisation of European Enterprise and Trade Statistics |

| | |
|---|---|
| MS | Member States |
| NSI | National Statistical Institute |
| NSA | National Statistical Agency |
| OECD | Organisation for Economic Cooperation and Development |
| OS | Official Statistics |
| OWL | Web Ontology Language |
| PID | Persistent Identifier |
| PUF | Public Use File |
| RDF | Resource Description Framework |
| SDMX | Statistical Data and Metadata eXchange |
| SDMX MCV | SDMX Metadata Common Vocabulary |
| SDMX/DDI dialogue | Dialogue engages the two standards bodies |
| SIOPS | Standard International Occupational Prestige Scale |
| SKOS | Simple Knowledge Organization System |
| SUF | Scientific Use File |
| T | Task |
| UN | United Nations |
| UNECE | United Nations Economic Commission for Europe |
| W3C | World Wide Web Consortium |
| WP | Work Package |
| XML | Extensible Markup Language |