META-DATA SPECIFICATION FOR THE DESCRIPTION OF SOCIAL SCIENCE DATA RESOURCES – CESSDA METADATA MODEL

Aleksandra Bradić-Martinović*,
Jelena Banović

Data Centre Serbia for Social Sciences,
Institute of Economic Sciences,
Belgrade, Serbia

Abstract:

Finding the necessary digital objects on the internet poses increasing challenges, and one way to overcome this problem is to use metadata that describes digital objects in a specific way. This paper aims to explain the importance and role of metadata and metadata standards/schemes, with particular reference to metadata used to describe data sets that researchers collect, archive, and disseminate in the Social Sciences. The paper describes the metadata model developed by the Consortium of European Social Science Data Archives (CESSDA ERIC) for the digital archiving of data sets in the European Research Area (ERA).

Keywords:

Metadata, Metadata Scheme, Social Sciences, CESSDA Metadata Model, Data.

INTRODUCTION

We are witnessing a digital revolution that has caused a flood of information. In the last 20 years, a considerable amount of data has been generated. The Covid-19 pandemic has caused an even more significant increase in the volume. The internet has become an almost infinite source of documents, images, e-books, music files, web pages and other data formats. Consequentially, searches are becoming increasingly challenging, especially since sources are numerous and diverse - governments, businesses, science, education, IoT, AI and similar. For that reason, most digital objects are described by metadata. Numerous software uses metadata as the basis of their functionality. The best examples are social media (Facebook, Twitter), video and music content providers (YouTube, Spotify) and many others. Metadata allows users to find the content they need.

In the field of scientific research, data plays a crucial role in analysing and testing scientific hypotheses. Researchers use different data types depending on scientific discipline, while the data collected and shared in social sciences and humanities could be particularly sensitive. The first reason is the inability for replication – the intersection of social phenomena is unique at a specific time, and the second is the possibility

Correspondence:

Aleksandra Bradić-Martinović

e-mail:

abmartinovic@ien.bg.ac.rs

DOI: 10.15308/Sinteza-2021-193-199

of compromising the privacy of respondents, which is regulated by law in most countries. For these reasons, data in social sciences and humanities have great value, so it emphasises the importance of their availability.

The paper is dividing into several sections. After the introduction, we covered the basics definitions and division of metadata and metadata standards/schemes. The third section contains a more profoundly explanation of metadata schemes in Social Sciences, while the fourth part introduces the CESSDA metadata model, which aims to describe digital objects containing primary data collected in scientific research.

2. METADATA AND METADATA STANDARDS/ SCHEMES

2.1. METADATA

The term "metadata" was first introduced by Jack E. Myers back in 1969 and became popular through the name of his company – "The Metadata Company".

The most straightforward definitions of metadata are "Metadata is data about data" or "The digital catalogue card" or "Information about the object [1], but they are too general to explain the essence of this concept. The National Information Standards Organization (NISO) provides a complete explanation through a more technically accurate definition – "metadata are structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage information".

The concept of metadata has proven to be very useful in various fields, especially in computer and communication sciences, libraries, statistics and numerous databases [1]. Technically, metadata contains basic information about data or digital content (or a digital object). The metadata structure is not universal but adapts to the specific content or object they describe. Metadata that describes photos, music or videos, e-books, files that contain information about a person will be different in scope and structure, but the most important thing is to describe the object as accurately as possible - to contain essential information. Typically, the metadata contains answers to the following questions: What?, When?, Where? Who? How? Which? and Why? [2]

Depending on the type of content or object that describes, the metadata is divided into several categories – Descriptive metadata, Structural metadata, Preservation metadata, Provenance metadata and Administrative metadata.

Descriptive data, as the name suggests, this type of metadata has the purpose of describing the content or a digital object, even though all metadata are descriptive. They are the most commonly used metadata. A simple example is an electronic description of a book, and the metadata contains the name of the book, the Name of the writer, the Year of publication, the Name of the publishing house and similar. There are certain situations when descriptive data become complex structures, and these are websites and code-driven projects.

- Example Properties: Title, Author, Subject, Genre, Publication date
- Primary Uses: Discovery, Display, Interoperability

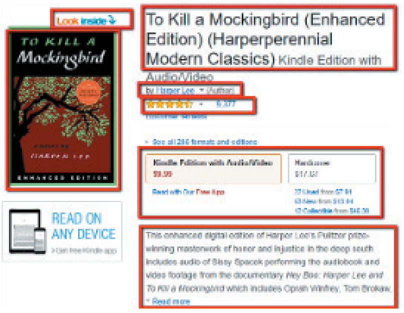
> Interoperability definition:
>
> "*Enabling information that originates in one context to be used in another in ways that are as highly automated as possible*" [3]

Technical metadata is a subgroup of descriptive metadata. These metadata contain information about the technical characteristics of digital objects, such as ownership, object type (database, text file, music or video, and similar).

- Example Properties: File type, File size, Creation date/time, Compression scheme
- Primary Uses: Interoperability, Digital object management, Preservation

Structural metadata is more complex than descriptive ones and is most commonly used when it is required to describe how a digital object or resource is sorted. An example is a video material with specific duration sections, which fit in precisely the specified order. Structural metadata carries information that is important to users to place sections on the memory space properly.

- Example Properties: Sequence, Place in a hierarchy
- Primary Uses: Navigation



To Kill a Mockingbird (Enhanced Edition) (Harperperennial Modern Classics) Kindle Edition with Audio/Video

by Harper Lee (Author)

This enhanced digital edition of Harper Lee's Pulitzer prize-winning masterwork of honor and injustice in the deep south includes audio of Sissy Spacek performing the audiobook and video footage from the documentary *Hey Boo: Harper Lee and To Kill a Mockingbird* which includes Oprah Winfrey, Tom Brokaw

Figure 1 - An example of an Amazon book metadata [4]

Preservation metadata provide the information necessary in the process of maintaining digital objects. Preservation metadata has the function to record and store all changes that occur with a digital object during use to preserve its integrity. To track changes, the most commonly used form is Preservation Metadata Implementation Strategies, which tackles all activities conducted on a digital object and access rights.

- Example Properties: Checksum, Preservation event
- Primary Uses: Interoperability, Digital object management, Preservation

Provenance metadata are used in cases where digital objects duplicate, i.e. when copies are made. This phenomenon is very prevalent in the digital world and provenance metadata stores data on the object's earliest history. History information is vital in tracking the digital object's lifecycle. Provenance metadata may also contain information about users who made changes to the files.

Administrative metadata provide instructions to users about rules and restrictions regarding the use of a particular digital object. They are primarily intended for administrators, who use them to restrict access to files relative to the defined level of access - qualifications (guest, internal user, external user, administrator, and similar). This metadata is comprehensive and provides the ability to manage objects. They can also be seen as a basic version of a piece of data. Their role is also to make complex objects easier to understand by users.

- Example Properties: Copyright status, Licence terms, Right holder
- Primary Uses: Interoperability, Digital object management

Metadata are typically stored as a HTML, XML or MARC 21 document linked to the resource it describes.

```
<mets:amdSee ID=”AMD_OTHER”>
<mets:techMD ID=”D09002ee180affcca-TEC”>
<mets:mdRef ID=”D09002ee180affcca-tdiv”
MDTYPE=”PREMIS” MIMETYPE=”text/xml”
LOCTYPE=“URL”
xmlns:xlink=http://www.w3.org/1999/
xlink
xling:href=”file/premis.xml” />
</mets:techMD>
</mets:amdSec>
```

Listing 1 - An example of administrative metadata from the XML file.

2.2. METADATA STANDARDS/SCHEMES

Bearing in mind that the types and purposes of digital objects are very diverse, appropriate standards need to be established to provide a unique set of rules. The primary purpose of these standards is to ensure the consistency of metadata and to enable interoperability.

In a specific sense, metadata standards, or schemas, define the necessary fields to describe a particular digital object. Therefore, the target fields are the essential elements of each schema metadata, and each of these fields contains the necessary information about the object. Depending on the type of object, the metadata scheme varies. In practical terms, any community that describes digital objects with metadata can have its metadata scheme [5].

Metadata standards are commonly associated with specific areas. Some examples are [6]:

- General metadata standards - Dublin Core (DC), Metadata Object Description Schema (MODS), Metadata Encoding and Transmission Standard (METS);
- Arts - Categories for the Description of Works of Art (CDWA), Visual Resources Association (VRA Core);
- Astronomy - Astronomy Visualization Metadata (AVM);
- Biology - Darwin Core;
- Ecology - Ecological Metadata Language (EML);
- Geographic - Content Standard for Digital Geospatial Metadata (CSDGM);
- Social Sciences - Data Documentation Initiative (DDI).

In the next part of the paper, Social Science metadata standards will be explained more broadly.

3. METADATA STANDARDS/SCHEMES FOR SOCIAL SCIENCES DATA SETS

Data Documentation Initiative (DDI) is an international standard for describing data sets obtained through different social and behavioural sciences observation methods. Standard is based on an XML format for content, presentation, transfer and preservation of documentation and data caps [7]. Initially, the standard was conceived as support in describing metadata in social sciences, but in later versions, it included data and other scientific fields.

DDI's goal is to anticipate key descriptive elements for data sets, which can be understandable to all parties, data creators, developers, librarians, and researchers. DDI encourages a comprehensive description for finding and analysing data. It is structured to enable machine find, functioning and interoperability of data (FAIR data) [8]. DDI provides a standard structure for all metadata that follows a data set, thus helping users interpret what is in the set. It is of great importance to everyone who uses a data set(s). Since metadata is expensive to produce, standardising metadata through DDI enables less time and money consumption and promotes interoperability. Also, DDI supports creating and using coders that are interactive, structured, and enable users to navigate more easily through metadata collections. DDI standard is continuously evolving and is actively working on customising its use in more complex data sets. In social sciences, it is very applicable because the creation of quality metadata is enabled to the maximum.

Encoded Archival Description (EAD) is the standard for coding information that comes from archive records. Archival timber is a specific form of timber. The main difference with the library structure is that the vast majority of the material is unpublished and unavailable online or elsewhere. With the development of the internet and the enabling of machine-readable records, it has become possible to consider developing digital aids that would help search archive timber. Work on the EAD standard began in 1992 at Berkeley, and the first version was released in 1998. After that, the second version came in 2002 and the last one in 2015 [9]. Today, this standard has wide use in archives, libraries, museums, and historical organisations worldwide. EAD enables users to find the primary sources they need through a standardised system for coding archive timber descriptions. The EAD uses a standard XML schema that determines the elements for describing the handwriting collection and the layout of those elements.

MIDAS Heritage. As the historic environment is an essential source of knowledge, it is clear that historical records are even more critical today because digitisation has enabled the transfer of most of the material to a digital format. MIDAS Heritage is the standard for historical data, i.e., data from the historic environment. It outlines what information should be recorded and which should not to enable effective exchange and long-term preservation of knowledge about the historical environment [10].

The MIDAS Heritage standard was created in 2007 in order to substantiate these needs. The standard creates records of buildings, monuments, archaeological sites, landscapes, parks, etc.. The standard is based on minimality - a minimum amount of information is required to describe cultural goods and includes all procedures involved in understanding, protecting, and managing goods. According to the formal text of the standards, its primary mission is to "share the knowledge of the past" [11] Government organisations use it, as well as local authorities, research communities and everyone else who deals with cultural goods in some capacity. Today, this standard facilitates modern life and enables the sustainability of records, ensuring that the same knowledge can be used and reused by future generations.

Statistical Data and Metadata Exchange (SDMX) is an international initiative aimed at modernising and standardising all mechanisms and processes to exchange statistical data and metadata between international organisations. Several organisations have teamed up to facilitate more efficient exchange of data and metadata in the field of statistical organisations, which are the Bank for International Settlements (BIS), the European Central Bank, Eurostat, the International Monetary Fund, the Organisation for Economic Co-operation and Development, the United Nations Statistics Division, and the World Bank [12]. SDMX has focused on facilitating the exchange and processing of data and metadata among organisations, which means that no typical data structure is exchanged among users. There are several different data formats and metadata: for time series, for cross-sectional data, for describing the structures of independent metadata sets, for structural metadata [13]. The standard focuses on statistical macroaggregates and is developed to support both microdata and unstructured data formats. Unlike other standards, SDMX focuses on increasing efficiency and ability to use and exchange data and metadata, not on metadata during the life cycle.

It is also valuable to mention **Open Archives Initiative Object Reuse and Exchange (OAI ORE) and Qualitative Data Exchange Format (QuDEx)**. OAI ORE defines standards for the description and exchange of web resource aggregations, sometimes called complex digital objects, to combine resources with multiple media types, including text, pictures, data, and videos [14]. QuDEx is an XML schema for documenting metadata for qualitative data sets. The QuDEx has been developed by the UKDA in 2006 [15], and it is intended for standard coding of metadata of qualitative collections. The scheme is entirely complementary to the DDI scheme.

4. CESSDA METADATA MODEL (CMM)

CESSDA ERIC is a vital element of the European Research Area in data management in social sciences. Bearing in mind that the Republic of Serbia is a Consortium member since 2019, the CESSDA recommendations are also an obligation for the Data Center Serbia for Social Sciences (DCS), the national research infrastructure and CESSDA's Service Provider for our country.

CESSDA aims to enable all national digital repository that collects, store and share primary data sets, a simple method for increasing visibility through its data catalogue (CESSDA Data Catalogue - CDC). In this way, data collected as part of national surveys can gain international visibility.

As part of the CESSDA Metadata Office project, which covers related topics, the CESSDA Metadata Model (CMM) has been created to introduce European digital archives into best practice in this subject. The broader concept, CESSDA Metadata Portfolio, consists of the "CESSDA Metadata Model, User Guide, CESSDA Vocabulary Service, European Language Social Science Thesaurus (ELSST), CESSDA Data Catalogue Profiles, CESSDA Metadata Validator, UML model, Supplementary Materials and Management and Maintenance Plan" [16].

The purpose of CMM is to describe every data set that researchers deposit into a repository and has a formal structure in this sense. It consists of primary and auxiliary elements. The main elements are Information on Study; Information on Persons; Information on Organisations; Information on Dataset; Information on Instrument; Information on Questions and Responses; Information on Concepts; Information on further Documents; Information on Publications (publications where data have been used); Information on Group of Studies and Information on Document Description ('metadata about metadata'). It is relying on DDI Lifecycle 3.2. The simplest way to understand CMM is through example. In this case, we will describe the first element – Information on Study, i.e. metadata about the study in which the data was collected.

Number and element	1 Study
Child element	1.1 Bibliographic Information 1.2 Content Information 1.3 Methodical Information 1.4 Access Information
Description	Information on the study/studies. No metadata element.
Mandatory/ Recommended/ Optional	Mandatory
Occurrence	1
Controlled vocabulary	-
Usage notes	.

Table 1 – Information on Study – The first level

Number and element	1.1 Bibliographic Information
Child element	1.1.1 Study DDI Identifier 1.1.2 Study Number 1.1.3 Study Title 1.1.4 Subtitle 1.1.5 Alternative Title 1.1.6 Funding Information 1.1.7 Principal Investigator Reference 1.1.8 Publisher 1.1.9 Publication Date (controlled) 1.1.10 Study Version 1.1.11 Contributor Reference 1.1.12 Reference Study to Document
Description	Bibliographic information. No metadata element.
Mandatory/ Recommended/ Optional	Mandatory
Occurrence	1
Controlled vocabulary	-
Usage notes	.

Table 2 – Bibliographic information – The second level

Number and element	1.1.1 Study DDI identifier
Child element	None
Description	Identifier of the study according to the DDI 3.2 structure.
Mandatory/ Recommended/ Optional	Mandatory for DDI 3.2, not for 2.5
Occurrence	1-2 for DDI 3.2, 0 for DDI 2.5
Controlled vocabulary	-
Usage notes	It is recommended to have both the URN and the combination of subclasses Agency, ID and Version as an identifier in DDI-L 3.2. However, it is possible to use only the URN or only the combination of Agency, ID and Version as an identifier.

Table 3 – Bibliographic information – The third level

All of the above primary elements are expanded by levels (in some cases up to five levels) to describe the data set properties in more detail. By applying CMM, all digital archives included in CESSDA ERIC become interoperable, and the data they store becomes internationally available and easily searched.

5. CONCLUSION

Metadata, structured according to the needs of specific scientific fields, helps researchers to locate digital objects, such as e-books, scientific publications, video materials, and similar. In addition to general metadata standards, many scientific and professional organisations have created their own standards and schemes to enable interoperability within the scientific field. Knowledge of standards and schemes is beneficial from two points of view. The first is the possibility for researchers to find the necessary digital material for their research, and the second is to make their scientific publications or data sets available and easily accessible to other researchers. For the needs of researchers in the social sciences, CESSDA ERIC has created the CESSDA Metadata Model intending to harmonise the meta-fields describing the data sets collected in the primary surveys, which are available in the public repositories of the national providers of the countries participating in this European infrastructure.

6. ACKNOWLEDGEMENTS

The authors acknowledge support from the Ministry of Education, Science and Technological Development of the Republic of Serbia.

REFERENCES

[1] J. Greenberg, “Understanding Metadata and Metadata Schemes,” *Cataloging & Classification Quarterly,* vol. 3/4, pp. 17-36, 2005.

[2] J. Hare, “What is Metadata and Why is it as Important as the Data itself?,” opendatasoft.com, [Online]. Available: https://www.opendatasoft.com/blog/2016/08/25/what-is-metadata-and-why-is-it-important-data. [Accessed 21 03 2021].

[3] “The <indecs> metadata framework: Principles, model and data dictionary,” 06 2000. [Online]. Available: https://www.doi.org/topics/indecs/indecs_framework_2000.pdf. [Accessed 22 03 2021].

[4] “Author Imprints,” [Online]. Available: https://www.authorimprints.com/metadata/. [Accessed 2021].

[5] I. Smith, A. Breytenbach and R. Groenewald, “Metadata, Metadata Schemas & Metadata Standards, IGBIS Seminar "Digital Library Standards & Metadata - The Basics",” 20 06 2007. [Online]. Available: https://repository.up.ac.za/bitstream/handle/2263/2823/igbis_breytenbach_up.pdf?sequence=5. [Accessed 23 05 2021].

[6] “Libraru Guides: Research Data Management: Metadata Schemas and Examples,” Victoria University, Melbourne Australia, 2015. [Online]. Available: https://libraryguides.vu.edu.au/research-data-management/metadataschemasandexamples. [Accessed 21 05 2021].

[7] K. Miller and M. Vardigan, “How Initiative Benefits the Research Community - the Data Documentation Initiative,” [Online]. Available: https://ddialliance.org/sites/default/files/miller.pdf. [Accessed 29 04 2021].

[8] “DDI Profiles,” DDI, 2021. [Online]. Available: https://ddialliance.org/resources/ddi-profiles. [Accessed 29 05 2021].

[9] E. A. D. W. G. o. t. S. o. A. A. a. t. N. D. a. M. S. O. f. h. L. o. Congress, “Encoded Archival Description Tag Library,” 2002. [Online]. Available: https://www2.archivists.org/sites/all/files/EAD_TagLibrary_2002.pdf. [Accessed 29 05 2021].

[10] F. o. I. S. i. Heritage, “MIDAS Heritage,” [Online]. Available: http://www.heritage-standards.org.uk/midas-heritage/. [Accessed 13 05 2021].

[11] T. F. o. I. S. i. Heritage, “MIDAS Heritage - The UK Historic Environment Data Standard, v1.1,” The Forum on Information Standards in Heritage, 2012.

[12] S. Community, “SDMX,” 2019. [Online]. Available: https://sdmx.org/?page_id=2705. [Accessed 29 05 2021].

[13] A. Gregory and P. Heus, “DDI and SDMX: Complementary, Not Competing, Standards,” Open Data Foundation, 2007.

[14] O. A. Initiative, “Open Archives Initiative Object Exchange and Reuse,” [Online]. Available: http://www.openarchives.org/ore/. [Accessed 13 05 2021].

[15] L. Corti, “Publishing digital qualitative data,” UK Data Archive, [Online]. Available: https://www.ukdataservice.ac.uk/media/604312/lc_metqualdata_part2mar16.pdf. [Accessed 13 05 2021].

[16] CESSDA, “User Guide for the CESSDA Metadata Model,” CESSDA, 2019.