The matter of meta in research data management: Introducing the CESSDA Metadata Office Project

André Förster[1], Kerrin Borschewski[2], Sharon Bolton[3] and Taina Jääskeläinen[4]

Abstract

Accompanying the growing importance of research data management, the provision and maintenance of metadata – understood as data about (research) data – have obtained a key role in contextualizing, understanding, and preserving research data. Acknowledging the importance of metadata in the social sciences, the Consortium of European Social Science Data Archives (CESSDA) started the Metadata Office project in 2019. This project report presents the various activities of the Metadata Office (MDO). Metadata models, schemas, controlled vocabularies and thesauri are covered, including the MDO's collaboration with the DDI Alliance on multilingual translations of DDI vocabularies for CESSDA Service Providers. The report also summarizes the communication, training and advice provided by MDO, including DDI use across CESSDA, illustrates the impact of the project for the social sciences and research data management community, and offers an outline regarding future plans of the project.

Keywords

Research data management (RDM), Metadata, Controlled vocabularies, Multilingualism, Consortium of European Social Science Data Archives European Research Infrastructure Consortium (CESSDA-ERIC)

1. Introduction

With the constant digitization of research and the ever-increasing relevance of open science, research data management (RDM) has become one of the most important indicators when assessing the quality of research. 'Good research data management is not a goal in itself, but rather the key conduit leading to knowledge discovery and innovation, and to subsequent data and knowledge integration and reuse' (European Commission 2016, p. 3). Thus, it is a substantial part in the research (data) lifecycle, which is, for instance, reflected by the highly regarded FAIR principles for scientific data management and stewardship (Wilkinson et al. 2016). According to the FAIR principles, guidelines concerning the documentation of research data and infrastructures for their long-term preservation should ensure that research data are *f*indable, *a*ccessible, *i*nteroperable and *r*eusable. These principles have become the guiding principles when handling research data, for research in general and for the social sciences in particular.

Accompanying these developments, research on RDM has become more important over the last years, although it is still considered to be at an early stage. For instance, some scholars investigate how information infrastructures at universities and research institutions have to be designed in order to optimally incorporate RDM processes (Blask and Förster 2019). Other scholars focus on how to encourage researchers to perform RDM (van den Eynden and Bishop 2014), since for many researchers, the disadvantages of doing RDM still dominate. They believe that doing RDM and fostering open data means having additional work without personal gain. They fear the misuse of their

data and that other researchers could beat them in publishing relevant findings their data provide, without integrating or citing the data producers. This means they would get less credit for their work, especially since publications are generally more appraised than the production of valuable data (Wolf 2017).

The most active field of research and services on RDM, however, is still the technical one. This is – among other things – due to the fact that many funding agencies have labelled RDM as means to fulfil specific mandatory funding requirements such as long-term preservation (Deutsche Forschungsgemeinschaft/German Research Foundation DFG 2013). Since the provision and maintenance of metadata – understood as *data about (research) data* – is cardinal to understand research data, metadata also play an important part in ensuring technical RDM requirements. Metadata describe publications and digital objects, make sure that research data can be contextualized, contribute to the implementation of the FAIR principles and thus help in following specific funding requirements (Gregory et al. 2009). Whereas in other sciences metadata are often not standardized according to reporting standards and information about the use of metadata schemas is not provided in the respective repositories, the quantitative social sciences have established the Data Documentation Initiative (DDI)[5] standard as the main standard of metadata documentation (Vardigan 2014). The DDI standard is a free and international metadata standard for the description of research data from the social, behavioral, economic, and health sciences. It allows for a detailed and semantically rich documentation. DDI is the most commonly used metadata standard for social sciences survey data on an international scale. As the DDI documentation is captured in extensible markup language (XML), it is machine-actionable and fosters interoperability, metadata exchange and reuse of data. Most of the social science data archives produce their metadata in the DDI format. Currently, two different DDI specifications (each having different versions) exist – DDI Codebook and DDI Lifecycle. The Codebook specification is focused on the after-the-fact-documentation. It includes information on document description, study description, variable description and file description. DDI Lifecycle is the more elaborate of both specifications. It covers all aspects of the research data lifecycle (defined by the DDI Alliance), starting with the planning and data collection right through to archiving (Green and Humphrey 2014; Hoyle et al. 2011; Rasmussen 2014; Vardigan et al. 2008; Zenk-Möltgen 2012).

Over the last decades, research infrastructures in the social sciences have been established to manage the technical aspects of RDM – including metadata – within the whole research (data) lifecycle (Kaase 2013; Renschler et al. 2013). The Consortium of European Social Science Data Archives (CESSDA)[6] was established in 1976 and started out as an umbrella organization of seven social science data archives (Mochmann 1998). As a leading research infrastructure for the social sciences CESSDA was awarded with the status of a European Research Infrastructure Consortium (ERIC) by the European Commission in 2017. The number of CESSDA member archives is increasing continually. The currently 20 CESSDA members work to improve data access for researchers. For this, CESSDA provides large-scale, integrated and sustainable data services to the social sciences and supports national and international research and cooperation (Consortium of European Social Science Data Archives CESSDA 2020). Acknowledging the importance of metadata in the social sciences and further backing this cause (van der Eycken et al. 2019), CESSDA has started the Metadata Office project (MDO)[7] in 2019. While CESSDA in general provides a full scale sustainable research infrastructure that supports social scientists in conducting high-quality research, MDO forms a core conceptual and strategic group consisting of six

partner institutions[8] to maintain and manage CESSDA's metadata-related material and monitors metadata developments and the metadata community. MDO not only oversees the strategic components and developments of all metadata related issues (including giving recommendations to CESSDA data archives, also named CESSDA Service Providers), but also manages and coordinates the content of the [European Language Social Science Thesaurus (ELSST)](9) and related multilingual vocabulary services. This project report summarizes MDO's various activities, illustrates its impact for the social science and RDM community, and finally offers an outline regarding MDO's future plans.

2. Development of metadata products

MDO is not the first project dedicated to metadata within the CESSDA community. Since its establishment, CESSDA has been active in metadata matters, e.g. by establishing a common catalogue, engaging in the DDI Alliance, or in data exchange across Europe. Former projects include the EU funded projects Nesstar, MADIERA, MetaDater, Data without Boundaries and CESSDA-PPP (Jensen 2010; Jensen and Mochmann 2003; Mauer 2012; Silberman and Tubaro 2008). The latest finished project on metadata issues within CESSDA and MDO's forerunner is the CESSDA Metadata Management project. One of its main results was the development (Zenk-Möltgen et al. 2015) of a preliminary version of the CESSDA Metadata Model (CMM), which has been further developed and published as a first version within MDO (Borschewski et al. 2019). We present the CMM in the following chapter (2.1).

In general, it should be noted that projects on metadata accompany the developments of metadata standards such as DDI. While DDI is an elaborate metadata standard in the social sciences, its uptake and application vary between institutions, and even within CESSDA. Thus, there is a need for a common semantic understanding of metadata issues that helps institutions to align their conceptual and technical metadata requirements with these standards. Projects on metadata such as MDO help to establish this understanding within the CESSDA community.

2.1 The CESSDA Metadata Model (CMM)

The CMM is built from the viewpoint of quantitative social science data. Thus, it serves the purpose of helping CESSDA Service Providers to make their quantitative data more discoverable and comprehensible to users. The CMM is based on the DDI Lifecycle metadata standard, because it is currently the most elaborate standard for the social sciences and due to its objective of interoperability. Furthermore, most of the CESSDA archives use one of the DDI specifications for their metadata.

For the CMM, the CESSDA Metadata Management project agreed on elements concerning the quantitative social sciences which were considered important for CESSDA tools, such as the [CESSDA Data Catalogue (CDC)](10), the [CESSDA Euro Question Bank (EQB)](11), and for future tools of CESSDA. The aim of the CMM is to be an understandable, conceptual metadata schema. It is supposed to support CESSDA tools and also CESSDA Service Providers to have an overview of currently relevant metadata elements within CESSDA. Facing this challenge of being as comprehensive as possible, but also as easy to handle as possible, the CMM currently does not include elements that do not adhere to the above-mentioned characteristics. Following these requirements, the CESSDA Metadata Management project decided to include elements based on corresponding DDI Lifecycle 3.2 elements into the CMM.

However, for the CMM elements, the project chose more conclusive element names. As all elements follow the DDI Lifecycle 3.2 structure, the X-Path examples can be found within the CMM. The structure of the CMM is based on the principle of reusing metadata elements. This means that wherever possible, information is referenced and reused according to DDI Lifecycle 3.2, and the elements were stored within the DDI Lifecycle ResourcePackage. Within ResourcePackage information that is intended to be reused by multiple studies can be structured. The possibility to reference metadata information reduces the amount of work involved in the documentation process.

The first version of the CMM contains more than 450 elements. However, the attributes concerning a certain element (such as information on language) are counted as separate elements. The latest version of the CMM was published in November 2019. It includes a mapping to the current version of the CDC metadata schema and corresponding DDI 2.5 X-Paths, and it is also accompanied by a detailed documentation (Storviken et al. 2019). The new version also includes a sheet where users can track all the changes that have been made to the CMM from version 0.1 to the current version 1.0.

The CMM contains metadata elements, their definitions, examples of their use, and information on specific requirements, such as mandatories, repeatability and use of controlled vocabularies. Table 1 presents the CMM overview, listing the information covered by the CMM elements on different levels. These levels are captured in the sections
- information on study,
- information on person(s),
- information on organization(s),
- information on dataset,
- information on instrument,
- information on questions and responses,
- information on concepts,
- information on further documents,
- information on publication,
- information on group of studies and
- information on document description.

Furthermore, the CMM overview offers explanations of the CMM characteristics. These are used to define the CMM elements in detail. Specifically, the CMM includes information on the element number, on the name of an element, the definition on an element, and information on the status of an element, which is used to define whether the respective CMM element is mandatory, recommended or optional. Moreover, the CMM offers information on standardized and controlled content for each element. This informs the user whether to employ a controlled vocabulary, an ISO code or a thesaurus and also what type he or she should use. To give additional information to the status of an element and to make it easier for technicians to process the information, the CMM includes a column with the occurrences of an element. The last column of CMM contains DDI3.2 X-Paths examples. The DDI3.2 X-Path examples are exemplary DDI3.2 mappings to the element that are supposed to help the understanding of the respective element.

To make this clearer, we present an example. Figure 1 shows how the various characteristics are filled with information for each CMM element. For our example, we use the element 'Language of Study

Title'. The number of the element is 1.1.3.1. This shows that it is a sub element of 'Bibliographic information' (element number 1.1). The element 'Bibliographic information' itself is a sub element of 'Study' (element number 1). The definition of our example element with the element name 'Language of Study Title' is 'The language of the content of the element'. This means that the element is used to display the language in which the 'Study Title' was captured within the metadata. In the column for 'status' we find the information 'M (for DDI3.2)', meaning that if the top element of 'Language of Study Title', namely the element 'Study Title' is used, the use of the 'Language of Study Title' is mandatory for metadata captured in DDI3.2. The following column regarding standardization contains the text 'Use ISO 639-1 (Language Code)' for our example element 1.1.3.1. This means that the ISO language code 639-1 has to be used to capture the information of the language. Therefore, if the study title was in, say, Finnish, the content of element 1.1.3.1 would be 'fi'. The occurrence of our example element is '1', meaning that for each time the element study title is used, the information 'Language of Study Title' must be given. It also means that this information on the element can only be used once. The last column contains the DDI3.2 X-Path example, which in our case is 'ddi:DDIInstance/s:StudyUnit/r:Citation/r:Title/r:String/@xml:lang'.

Since the CMM has a conceptual nature and is independent of actual implementation, its further development by MDO supports CESSDA Service Providers and other research infrastructure institutions in the provision and maintenance of metadata while still offering them many possible ways to actually store, manage, organize and present metadata. Thus, the final implementation of the CESSDA CMM remains in the Service Providers' and institutions' own responsibility.

However, CESSDA has great interest in enabling its Service Providers to produce high-quality metadata, and one way to ensure this is fostering the standardization of metadata delivered by the Service Providers and received by CESSDA tools (such as CDC and EQB). Standardization is achieved by giving clear instructions on how to use specific CMM elements, controlled vocabularies etc. More specifically, CESSDA's long-term goal regarding standardization is that all Service Providers use DDI3.2. However, many providers lack resources to develop an editor being able to handle DDI3.2 references, leading to the majority of Service Providers still using DDI2.5. Since CESSDA tools need to be able to harvest metadata even in the current situation, it would not make sense to define elements as mandatory that cannot be produced using the DDI2.5 specification. In general, fostering the process of standardization also means that metadata have to be checked continuously, errors have to be found and those errors have to be communicated and corrected. At the moment, MDO enters differences detected in harvested metadata into an issue tracker and assigns these issues to the respective Service Providers, asking them to amend their data (see also chapter 3). Additionally, CESSDA is preparing a strategy on how it can help Service Providers to move towards DDI3.2.

Since different services require information in different languages, the CMM supports multilingualism by requesting users to produce language tags in their metadata and by providing controlled vocabularies in CESSDA languages (see also chapter 2.2). Specifically, the CMM allows users to provide metadata in different languages, requesting the use of ISO language tags in metadata. In doing so, different services are able to detect in which language the metadata and its elements actually are. This is particularly important when harvesting multilingual metadata files. In addition, CESSDA provides controlled vocabularies (e.g. the DDI and CESSDA vocabularies) included in the CMM in different CESSDA languages. The controlled vocabularies included in the CMM are currently provided

in eleven different languages (see chapter 2.2). As MDO works closely together with the respective experts on controlled vocabularies within the DDI Alliance, most vocabularies – with the exception of the CESSDA-specific Topic Classification – are DDI controlled vocabularies. Both DDI and CESSDA vocabularies can be reused via the Creative Commons BY 4.0 license[12].

Multilingualism in the CMM is not only reflected in the controlled vocabularies, but also in certain text elements. Language information is especially important for the CESSDA EQB. For elements of which EQB requires the language information, the use of the language attribute is mandatory. For instance, CMM includes the element 'Question Item Text', which is a text element. Since EQB also uses this element and requires information about language, the element has a mandatory attribute 'Language of Question Item Text'. Many language attribute elements are mandatory within CMM, apart from elements where this does not necessarily make sense. For instance, this accounts for the element 'Variable Name', since a variable name can also be an alphanumeric code not specific to a certain language and will not be translated, if a dataset is distributed in another language.

Apart from the maintenance and conceptual enhancement of the CMM and its documentation, MDO's task is also to improve its technical representation. Regarding a technically sophisticated form of the CMM, MDO is currently developing an XML Schema Definition (XSD) and a DDI profile derived from the model, accompanied by application profiles for other CESSDA tools and services, such as the EQB and the CESSDA Vocabulary Service (CVS)[13].

Contained Information and Numbering:	
Complete List of Elements	Information on Study: 1 Information on Person(s): 2 Information on Institution(s): 3 Information on Dataset: 4 Information on Instrument: 5 Information on Questions and Responses: 6 Information on Concepts: 7 Information on further Documents: 8 Information on Publication: 9 Information Group of Studies: 10 Information on Document Description: 11

Spreadsheet column headlines	Signification
No.	Number of Elements; represents the structure (1.1 means "is child element of" 1)
Element	Name of element
Definition	Definition of element
Status (Mandatory / Recommended / Optional)	Is this element mandatory (M), recommended (R) or optional (O).
Condition (if applicable) for M / R / O	If applicable: Under which condition is the element mandatory / recommended / optional?
Standardized/ Controlled content for this element	Which CV / standard is used for the element (if DDI CV: http://ddialliance.org/controlled-vocabularies), or ISO / Thesaurus etc. or default values
Occurence	Occurence of element
DDI 3.2 Element	X-Path for DDI 3.2 elements. Important remark: the X-Paths are only preliminary and exemplary mappings. They are supposed to help with understanding the meaning of the elements. However, the final technical implementation of the elements will be up to the CESSDA SP. There will be no constraint to adopt them.
Mapping information: CDC Element Property Name	Mapping CDC to CMM (version September 2018 https://docs.google.com/spreadsheets/d/1u9NSMvCwH1EMcpgKRDomZV9UW0tbDTrDZnr74vIgKY8/edit#gid=0) . CDC Element Property Name
Mapping information: CDC Element Name for Interface	Mapping CDC to CMM (version September 2018 https://docs.google.com/spreadsheets/d/1u9NSMvCwH1EMcpgKRDomZV9UW0tbDTrDZnr74vIgKY8/edit#gid=0) . CDC Element Name for Interface
Mapping information: Mandatoriness (MustShouldCould)	Mapping CDC to CMM (version September 2018 https://docs.google.com/spreadsheets/d/1u9NSMvCwH1EMcpgKRDomZV9UW0tbDTrDZnr74vIgKY8/edit#gid=0) . Mapping information: Mandatoriness (MustShouldCould) of CDC
Mapping information: Schema element in DDI 2.5	Mapping CDC to CMM (version September 2018 https://docs.google.com/spreadsheets/d/1u9NSMvCwH1EMcpgKRDomZV9UW0tbDTrDZnr74vIgKY8/edit#gid=0) . Mapping information: Schema element in DDI 2.5 by CDC
Mapping information: Notes for DDI 2.5 Schema	Mapping CDC to CMM (version September 2018 https://docs.google.com/spreadsheets/d/1u9NSMvCwH1EMcpgKRDomZV9UW0tbDTrDZnr74vIgKY8/edit#gid=0) . Mapping information: Notes for DDI 2.5 Schema by CDC

Table 1: Information covered by the CMM.

No.	Element	Definition	Status (Mandatory / Recommended / Optional)	Standardized/Controlled content for this element (if DDICV: http://www.ddialliance.org/controlled-vocabularies) (or ISO / Thesaurus etc.) or default values	Occurence	DDI 3.2 Element [remark: the X-Paths are only preliminary and exemplary mappings. They are supposed to help the understanding of the meaning of the elements. However, the final technical implementation of the elements will be up to the CESSDA SP. There will no constraint to adopt them]
Information on Study:						
1	Study	[no metadata element] Information on the study/studies	M		1	
Information on Study: Bibliographic Information						
1.1	Bibliographic information	[no metadata element]	M		1	
1.1.1	Study ID	Identifier of the Study according to DDI 3.2 structure.	M		1-2	Either URN or triple Agency, ID, Version is mandatory in DDI3.2 URN: ddi:DDIInstance/s:StudyUnit/r:URN Triple Agency, ID, Version: ddi:DDIInstance/s:StudyUnit/r:Agency ddi:DDIInstance/s:StudyUnit/r:ID ddi:DDIInstance/s:StudyUnit/r:Version
1.1.2	Study Number	Unique archival number - if there is no unique archival number available, a PID may be used alternatively (if a PID is used as study number, the information for this PID should however still be captured in the element 1.8 'PID Study')	M		1-n	ddi:DDIInstance/s:StudyUnit/r:UserID Example: r:UserID typeOfUserID="StudyNumber">ZA4800</r:UserID
1.1.2.1	Type of Study Number (for DDI3.2)	Required attribute containing the local user identification system that maintains and defines the UserID.	M (for DDI3.2)	UniqueArchivalNumber	1	ddi:DDIInstance/s:StudyUnit/r:UserID/@typeOfUserID
1.1.3	Study Title	Study title	M (CDC)		1-n	ddi:DDIInstance/s:StudyUnit/r:Citation/r:Title/r:String
1.1.3.1	Language of Study Title	The language of the content of the element	M	Use ISO 639-1 (Language Code)	1	ddi:DDIInstance/s:StudyUnit/r:Citation/r:Title/r:String/@xml:lang
1.1.3.2	Translation status of Study Title	Is the content of the element translated?	R	true, false	0-1	ddi:DDIInstance/s:StudyUnit/r:Citation/r:Title/r:String/@isTranslated
1.1.4	Subtitle	Subtitle of the study	O		0-n	ddi:DDIInstance/s:StudyUnit/r:Citation/r:SubTitle/r:String
1.1.4.1	Language of Subtitle	The language of the content of the element	M	Use ISO 639-1 (Language Code)	1	ddi:DDIInstance/s:StudyUnit/r:Citation/r:SubTitle/r:String/@xml:lang
1.1.4.2	Translation Status of Subtitle	Is the content of the element translated?	R	true, false	0-1	ddi:DDIInstance/s:StudyUnit/r:Citation/r:SubTitle/r:String/@isTranslated
1.1.5	Alternative Title	Alternative title of the study	O		0-n	ddi:DDIInstance/s:StudyUnit/r:Citation/r:AlternateTitle/r:String
1.1.5.1	Language of Alternative Title	The language of the content of the element	M	Use ISO 639-1 (Language Code)	1	ddi:DDIInstance/s:StudyUnit/r:Citation/r:AlternateTitle/r:String@xml:lang
1.1.5.2	Translation Status of Alternative Title	Is the content of the element translated?	R	true, false	0-1	ddi:DDIInstance/s:StudyUnit/r:Citation/r:AlternateTitle/r:String@isTranslated
1.1.6	Funding information	[no metadata element]	O		0-n	
1.1.6.1	Funding Agency Reference	Reference to the funding agency (the agency which funded the described entity).	O		0-n	Either URN or triple Agency, ID, Version and the type of referenced object are mandatory in DDI3.2 URN: ddi:DDIInstance/s:StudyUnit/r:FundingInformation/r:AgencyOrganizationReference/r:URN Agency, ID, Version: ddi:DDIInstance/s:StudyUnit/r:FundingInformation/r:AgencyOrganizationReference/r:Agency ddi:DDIInstance/s:StudyUnit/r:FundingInformation/r:AgencyOrganizationReference/r:ID ddi:DDIInstance/s:StudyUnit/r:FundingInformation/r:AgencyOrganizationReference/r:Version

Figure 1: Characteristics of the CMM elements. Note that only a selection of elements is shown.

2.2 The CESSDA Vocabulary Service (CVS)

The [CVS](#)[11] provides a user-friendly source of standardized controlled vocabularies. These controlled vocabularies can be defined as hierarchically organized lists of codes with descriptive terms and definitions in one or more languages. Currently, there are 24 controlled vocabularies available in the CVS. The service contains both an editor for creating and updating vocabularies and a user interface where users can search and browse the published vocabularies and download them in different formats. The service also provides Uniform Resource Names (URNs) for both the controlled vocabulary and for each version of it, as well as an Application Programming Interface (API). At concept level, code value is the identifier that stays the same across all language versions. Figures 2 and 3 present the CVS search interface and a detailed view of a specific vocabulary.

Figure 2: CVS search interface.

Figure 3: Detailed view of a specific vocabulary.

Vocabulary management is done in the CVS Editor. Access to the Editor is governed by agency and language specific user roles. Source language administrators can create source vocabularies of their agency and translators translate them into their language. All administrators can publish and version their vocabularies, as well as browse all draft, unpublished agency vocabularies in any language. Registered users receive training to ensure that they are following best practice. Training includes a training webinar, power point slides and an extensive online user guide, thus making the use of the tool easy and straightforward.

The CVS provides a respected and authoritative source for CESSDA controlled vocabularies which data producers can use to ensure systematic metadata and description for data assets. Since DDI vocabularies form an important part in standardizing metadata, the DDI Alliance uses the CVS Editor for managing its vocabularies and their translations. Their vocabularies are published on the DDI website, but the vocabularies are also published in the CVS user interface, where they can be browsed. The vocabulary service can also be used for managing other agency vocabularies, for instance those of other research infrastructures or data repositories. Organizations interested in using the tool for creating their agency vocabularies or in producing and maintaining a new language version of DDI or CESSDA vocabularies can contact the MDO team.[8] Since DDI is an international standard, there is some incentive to add languages from outside of CESSDA.

Thus, the CVS provides CESSDA and other users with a robust multilingual service for the standardized description of social science data and associated materials. Vocabulary information includes detailed documentation of any changes in published vocabularies, which enables users to update their legacy metadata.

The CVS was created as an internal CESSDA project in 2017 to 2018, with participants from GESIS, the Finnish Social Science Data Archive (FSD), the United Kingdom Data Service (UKDS) and the Swedish National Data Service (SND). When the project ended, FSD and UKDS continued their work on the CVS within the MDO project. In 2019, they provided user feedback regarding the CVS system, tested some amendments and compiled an online user guide. FSD and UKDS act as content managers of the CVS, handling user management, access and training. Currently there are vocabularies available in eleven languages: Danish, English, Finnish, French, German, Italian, Norwegian, Portuguese, Serbian, Slovenian and Swedish. Japanese and Estonian are expected to be added next.

2.3 The European Language Social Science Thesaurus (ELSST)

The ELSST[9] is a broad-based, multilingual thesaurus for the social sciences. A thesaurus is also a controlled vocabulary. Rather than a list, however, a thesaurus comprises a structure that consists of terms and the relationships between them. Terms may be related hierarchically (broader/narrower) or non-hierarchically (related and synonymous). This is a more complex structure to the controlled vocabularies held within the CVS described above and means that ELSST is not managed via the current CVS system but is held in a separate ontology management system that enables sophisticated hierarchical editing. Also, their release schedule is currently different; ELSST has an annual new version release including all languages. Controlled vocabularies in CVS, on the other hand, are published as soon as they have been finalized, and each language is versioned and published separately with their own time schedule. Controlled vocabularies and ELSST keywords are utilized in separate ways within the metadata record. Controlled vocabularies are used or are planned to be used as filters in search interfaces which places even stricter requirements for the harvested metadata to be standardized.

CESSDA Service Providers are required to update their legacy metadata after changes regarding the controlled vocabularies. They need human-readable, detailed version history to see what they need to change and for deciding which changes can be done by machine and which require manual updates. ELSST keywords are different, as they describe in detail the actual subjects and concepts covered by the data, hence the more complex structural relationships between terms and editing facilities that ELSST requires. In the medium- to long-term, it is planned that the management of CESSDA metadata ontologies including the CVS and ELSST will be done within the same system, which will enable synergies between where possible although they are intended for different purposes.

The ELSST was originally based on the monolingual Humanities and Social Science Electronic Thesaurus (HASSET), developed by the UK Data Archive at the University of Essex. In 2000, the EU-funded Language Independent Metadata Browsing of European Resources (LIMBER) project developed the first multilingual version of ELSST, translating the English HASSET into French, German and Spanish.

The thesaurus was further enhanced and extended through subsequent EU grants (such as MADIERA) and additional UK funding. In 2018, the CESSDA Vocabulary Services Multilingual Content Management (VOICE) project took over development and management of ELSST, moving forward with editorial work, software enhancements, and additional translations. The CESSDA Multilinguality Policy was also developed in VOICE with a view to managing ELSST in the context of other multilingual controlled vocabularies, such as the CVS. From 2019, it was therefore logical to manage ELSST within the MDO alongside CESSDA's other key metadata assets, the CMM and CVS. The ELSST is currently available in 14 languages and is managed by a dedicated team within the MDO project, in close cooperation with expert translators.

In order to improve the functionality, currency, and utility of ELSST, its multilingual content must be maintained and updated. Therefore, regular collaboration takes place between MDO partners to review ELSST content (with input from subject specialists) and updates are released annually. The 2019 release included a new language translation (Dutch) and a new set of translated scope notes (Slovenian), revised terms and hierarchy structures (groups of terms, held within a specific relationship to each other, that describe aspects of a social science concept), selected to strengthen subject coverage and reflect emerging topics in social science.

ELSST provides an indexing resource for data producers to aid in the curation and publishing of their data. In particular, updates to different language versions of ELSST allow a common approach to indexing, which will benefit the CDC and the catalogues of CESSDA Service Providers, as ELSST terms can be used for search and filtering purposes. Guidance for users and translators is kept updated, ensuring that all the information they need to use the system is readily available.

During 2020, ELSST will move to a new technical software platform that will enable better linked open data capabilities and interoperability. The MDO's ELSST team is currently in consultation with the producers of other key international thesauri, such as EuroVoc, the European Union Publications Office official thesaurus, and the Food and Agriculture Organization (FAO) of the United Nations' AGROVOC multilingual thesaurus, to exchange knowledge about multilingual thesaurus management and explore potential mappings between ELSST and other key thesauri.

Due to ELSST's development history across various projects, database rights in the organisation and collection of the data, and the underlying application, are currently held by the University of Essex.

Copyright in the natural language translations are held either by the individual translators or the relevant translating organisation. Organisations can obtain a licence from the UK Data Service to use or adapt ELSST. Once ELSST moves to its new technical platform, a more streamlined licencing model will be put in place using Creative Commons (similar to the CVS), administered by CESSDA.

3. Communication, coordination, training and outreach

An important result from previous research on RDM and from previous metadata related CESSDA projects is that communication, guidance and training are needed to achieve more involvement in the matter of research data services in general and specifically in metadata standards (Tenopir et al. 2014). Providing these resources also positively affects the implementation of standards within the CESSDA community and its institutions. Since CESSDA has always been active in teaching and learning, its involvement established in the CESSDA Training[14] pillar, MDO has acted accordingly, for instance by integrating rather small and not yet well-established CESSDA data archives into the MDO project. In particular, the Data Centre Serbia for Social Sciences (DCS) and the Portuguese Social Information Archive (APIS) have worked on the translation of controlled vocabularies and on further developing the additional documentation materials of the CMM. Furthermore, proper coordination as part of these efforts is also required when it comes to maintaining technical metadata issues (e.g. harvesting of metadata, discovery of studies etc.) and implementing CESSDA's metadata standards in close cooperation between MDO and CESSDA Service Providers. Therefore, MDO and the CESSDA Main Office have established a CESSDA internal metadata issue tracking system via a bitbucket[15] issue tracker. Since possibly sensitive technical questions might also be addressed there, access rights to the tracker are managed by the CESSDA Main Office, while MDO currently is responsible for managing the participation of the community. This way, we ensure that metadata experts at the CESSDA Service Providers are put in charge of solving the respective issues.

In order to further increase the visibility of metadata issues within the CESSDA community, MDO has undertaken several additional outreach efforts. For instance, MDO has installed an MDO Newsletter that informs the CESSDA community about the latest activities and future plans of the project. Currently, it is published at irregular intervals via the CESSDA basecamp[16] groups 'CESSDA Metadata Office' and 'Service Providers' Forum' and is also available on request from MDO. In general, the CESSDA community and its Service Providers take part in the further development of MDO's activities in various ways. While a core group of six Service Providers manages the MDO project, all Service Providers can reach out to the MDO project team to give feedback regarding their metadata requirements and the use of the respective documents and tools (e.g. CMM, CVS). Additionally, the MDO project has contacted each Service Provider and all the tools separately, asking for feedback concerning their needs on the MDO tasks. MDO is also exploring contacts within other CESSDA projects such as EQB and the CESSDA Training Group to raise awareness of MDO and ask what it can do for them. Further activities include the participation in CESSDA events (such as the CESSDA Expert Seminar 2019) and the possibility to host additional new Service Providers within the project (such as DCS and APIS).

Regarding public services and outreach beyond CESSDA, MDO has liaised with various metadata experts from other organizations. For instance, FSD has arranged a seminar on metadata, data catalogues and tools for findability, with participants from SND and the Japanese national project on

developing a joint data catalogue for the social sciences (Laaksonen 2019). Furthermore, MDO members of UKDS attended the IASSIST 2019 conference and presented 'Sustainable European Multilingual Vocabularies: A Model for Cooperation in Metadata Management among European Data Archives' (Barbalet and Bolton 2019). Based on this presentation, UKDS is working together with the Australian Data Archive and the Department of Library and Information Science of Punjabi University, Patiala on further opportunities for cooperation. Presentations on CVS, DDI vocabularies and their use in CESSDA were held at the EDDI 2019 conference (Bolton and Jääskeläinen 2019; Jääskeläinen and Bolton 2019).

Furthermore, the CDC is also now included in the European Open Science Cloud (EOSC) marketplace[17], which offers another source of (non-CESSDA) contacts and their feedback. Additionally, MDO is participating in the Research Data Alliance (RDA)[18] metadata interest group. All feedback is used to enhance the quality of the MDO metadata materials. More specifically, the CESSDA-external feedback enables MDO to expand its horizons. For instance, we can get an idea whether the CMM is used in more general contexts apart from CESSDA and which challenges other institutions encounter when using it. In general, we can learn about new questions and ideas in the metadata world.

4. Future plans

The future plans of MDO include a systematic review of international metadata developments in relevant consortia, institutions and projects. Furthermore, since one of the general purposes of CESSDA's metadata materials is the support of the FAIR principles, MDO will further develop these openly published materials, so that they clearly state in which way they contribute to aligning with the FAIR principles and how they support the obtainment of trust badges such as the Core Trust Seal[19]. For instance, MDO will define a specific set of metadata profiles that are the standard for its Service Providers and monitor their application within the CDC.

Metadata services and research on RDM and metadata management serve the advancement of data-based research, especially in the quantitative social sciences. By further developing CESSDA's metadata products, MDO will continue its work within the CESSDA community and beyond, relying on the excellent working relationships built up in the first months of the project. With our work, we hope to contribute that data will always be more than just collections of numbers or codebooks.

References

Barbalet, Suzanne; Bolton, Sharon (2019): Sustainable European Multilingual Vocabularies: A Model for Cooperation in Metadata Management among European Data Archives. Available online at https://ukdataservice.ac.uk/media/622472/iassist19_sbarbalet_sbolton_02.pdf, checked on 7/23/2019.

Blask, Katarina; Förster, André (2019): Designing an information architecture for data management technologies: Introducing the DIAMANT model. In *Journal of Librarianship and Information Science* online first. DOI: 10.1177/0961000619841419.

Bolton, Sharon; Jääskeläinen, Taina (2019): *DDI Vocabularies and Language Versions*. DOI: 10.5281/zenodo.3596792.

Borschewski, Kerrin; Förster, André; Friedrich, Tanja; Zenk-Möltgen, Wolfgang; Miranda, Patrícia; Moura Ferreira, Pedro et al. (2019): *CMM CESSDA Metadata Model*. DOI: 10.5281/zenodo.3236171.

Consortium of European Social Science Data Archives (CESSDA) (2020): The CESSDA Consortium. Available online at https://www.cessda.eu/About/Consortium, checked on 1/31/2020.

Deutsche Forschungsgemeinschaft (DFG) (2013): Vorschläge zur Sicherung guter wissenschaftlicher Praxis. Proposals for Safeguarding Good Scientific Practice. Available online at https://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/download/empfehlung_wiss_praxis_1310.pdf, checked on 7/18/2019.

European Commission (2016): Guidelines on Data Management in Horizon 2020. Available online at https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf, checked on 8/7/2019.

Green, Ann E.; Humphrey, Chuck (2014): Building the DDI. In *IASSIST Quarterly* 37 (1-4), pp. 36–44. DOI: 10.29173/iq500.

Gregory, Arofan; Heus, Pascal; Ryssevik, Jostein (2009): Metadata. In *RatSWD Working Paper Series* 57, pp. 1-22. DOI: 10.2139/ssrn.1447866.

Hoyle, Larry; Castillo, Fortunato; Clark, Benjamin; Kashyap, Neeraj; Perpich, Denise; Wackerow, Joachim; Wenzig, Knut (2011): *Metadata For The Longitudinal Data Life Cycle*. DOI: 10.3886/DDILongitudinal03.

Jääskeläinen, Taina; Bolton, Sharon (2019): *CESSDA Vocabulary Service – For managing vocabulary content*. DOI: 10.5281/zenodo.3600174.

Jensen, Uwe (2010): Data and Metadata Extensions of the CESSDA RI. Enhancement of data and metadata infrastructures for the CESSDA RI. Available online at https://ppp.cessda.eu/doc/D8.3_Data_metdata_enhancement.pdf, checked on 8/7/2019.

Jensen, Uwe; Mochmann, Ekkehard (2003): MetaDater: Towards standards and tools for the description of comparative surveys. In *ZA-Information* 52 (1), pp. 191–198. Available online at https://www.gesis.org/fileadmin/upload/forschung/publikationen/zeitschriften/za_information/ZA-Info-52.pdf, checked on 8/7/2019.

Kaase, Max (2013): Research infrastructures in the social sciences: The long and winding road. In Brian Kleiner, Isabelle Renschler, Boris Wernli, Peter Farago, Dominique Joye (Eds.): Understanding Research Infrastructures in the Social Sciences. Zürich: Seismo Press, pp. 19–27.

Karjalainen, Merja; Kleemola, Mari; Jensen, Uwe (2013): Metadata Standards - usage and needs in NSIs and Data Archives. Available online at http://www.dwbproject.org/export/sites/default/about/public_deliveraples/dwb_d7-1_metadata-standards-usage_report.pdf, checked on 8/16/2019.

Laaksonen, Helena (2019): FSD's multilingual and qualitative data expertise brings in international visitors. Available online at https://tietoarkistoblogi.blogspot.com/2019/04/jsps-seminar.html, checked on 7/23/2019.

Mauer, Reiner (2012): Das GESIS Datenarchiv für Sozialwissenschaften. In R. Altenhöner, C. Oellers (Eds.): *Langzeitarchivierung von Forschungsdaten. Standards und disziplinspezifische Lösungen*. Berlin: Scivero, pp. 197–215. URN: nbn:de:0168-ssoar-46476-7.

Mochmann, Ekkehard (1998): European Co-Operation in Social Science Data Dissemination. In Rachel Walker, Marcia Freed Taylor (Eds.): Information Dissemination and Access in Russia and Eastern Europe. Problems and Solutions in East and West. Amsterdam: IOS Press, pp. 33–42.

Rasmussen, Karsten Boye (2014): Social Science Metadata and the Foundations of the DDI. In *IASSIST Quarterly* 37 (1-4), pp. 28–35. DOI: 10.29173/iq499.

Renschler, Isabelle; Kleiner, Brian; Wernli, Boris (2013): Concepts and key features for understanding social science research infrastructures. In Brian Kleiner, Isabelle Renschler, Boris Wernli, Peter Farago, Dominique Joye (Eds.): Understanding Research Infrastructures in the Social Sciences. Zürich: Seismo Press, pp. 11–18.

Silberman, Roxane; Tubaro, Paola (2008): Data Archives and Access to Government Data for Researchers. State of the Art and Future Developments in Europe in the ERI Perspective. Available online at https://gala.gre.ac.uk/id/eprint/5437/1/6_1_CESSDA.pdf, checked on 8/7/2019.

Storviken, Silje; Hagen, Sunniva; Bockaj, Brigita; Bolko, Irena; Vipavc Brvar, Irena; Fink Kjeldgaard, Anne Sofie et al. (2019): User Guide for the CESSDA Metadata Model. DOI: 10.5281/zenodo.3236193.

Tenopir, Carol; Sandusky, Robert J.; Allard, Suzie; Birch, Ben (2014): Research data management services in academic research libraries and perceptions of librarians. In *Library & Information Science Research* 36, pp. 84–90. DOI: 10.1016/j.lisr.2013.11.003.

van den Eynden, Veerle; Bishop, Libby (2014): Sowing the Seed: Incentives and motivations for sharing research data, a researchers' perspective. Available online at http://repository.jisc.ac.uk/5662/1/KE_report-incentives-for-sharing-researchdata.pdf, checked on 7/18/2019.

van der Eycken, J.; Styven, D.; Gheldof, T.; Depoortere, R. (2019): Trust and Understanding. The value of metadata in a digitally joined-up world: Conclusion, a Vision for the Future. In R. Depoortere, T. Gheldof, D. Styven, J. van der Eycken (Eds.): Trust and Understanding: the value of metadata in a digitally joined-up world. Brussels: Archives et Bibliothèques de Belgique, pp. 135–144. Available online at https://hal.archives-ouvertes.fr/hal-02125062/document, checked on 8/7/2019.

Vardigan, Mary (2014): The DDI Matures: 1997 to the Present. In *IASSIST Quarterly*, pp. 45–50. Available online at https://iassistquarterly.com/pdfs/iqvol371_4_vardigan.pdf, checked on 7/19/2019.

Vardigan, Mary; Heus, Pascal; Thomas, Wendy (2008): Data Documentation Initiative: Toward a Standard for the Social Sciences. In *International Journal of Digital Curation* 3 (1), pp. 107–113. DOI: 10.2218/ijdc.v3i1.45.

Wilkinson, Mark D.; Dumontier, Michel; Aalbersberg, IJsbrand Jan; Appleton, Gabrielle; Axton, Myles; Baak, Arie et al. (2016): The FAIR Guiding Principles for scientific data management and stewardship. In *Sci. Data* 3: 160018. DOI: 10.1038/sdata.2016.18.

Wolf, Christof (2017): Implementing Open Science: The GESIS Perspective. Talk given at Institute Day of GESIS, 28 September 2017. In *GESIS Papers* (26), pp. 1–12. Available online at https://www.ssoar.info/ssoar/bitstream/handle/document/54950/ssoar-2017-wolf-Implementing_Open_Science_The_GESIS.pdf?sequence=1&isAllowed=y&lnkname=ssoar-2017-wolf-Implementing_Open_Science_The_GESIS.pdf, checked on 7/16/2019.

Zenk-Möltgen, Wolfgang (2012): Metadaten und die Data Documentation Initiative. In R. Altenhöner, C. Oellers (Eds.): Langzeitarchivierung von Forschungsdaten. Standards und disziplinspezifische Lösungen. Berlin: Scivero, pp. 111–126. Available online at https://www.ssoar.info/ssoar/bitstream/handle/document/46679/Zenk-Moeltgen_Metadaten%20und%20die%20Data%20Documentation%20Initiative.pdf?sequence=1, checked on 8/7/2019.

Zenk-Möltgen, Wolfgang; Kleemola, Mari; Etheridge, Anne; Fink Kjeldgaard, Anne Sofie (2015): Introducing the CESSDA Metadata Management Project. Available online at http://www.eddi-conferences.eu/ocs/index.php/eddi/eddi15/paper/view/190/174, checked on 8/7/2019.

End-notes

1 André Förster was the head of the CESSDA Metadata Office project at the GESIS – Leibniz Institute for the Social Sciences in 2019. He now works for the National Coordination Agency in Education Monitoring in Germany and can be reached by email: andre.foerster@kommunales-bildungsmonitoring.de (version: March 2020)

2 Kerrin Borschewski works in the CESSDA Metadata Office project and in CESSDA Training. She is a doctoral candidate at the GESIS – Leibniz Institute for the Social Sciences and can be reached by email: kerrin.borschewski@gesis.org

3 Sharon Bolton is head of the CESSDA Metadata Office project at the UK Data Service. She can be reached by email: sharonb@essex.ac.uk

4 Taina Jääskeläinen works at the Finnish Social Science Data Archive and participates in the CESSDA Metadata Office project. She can be reached by email: taina.jaaskelainen@tuni.fi

5 Data Documentation Initiative; https://www.ddialliance.org/

6 https://www.cessda.eu/

7 https://www.cessda.eu/About/Projects/Work-Plans/Work-Plan-2019#metaoff

8 The partners are the GESIS – Leibniz Institute for the Social Sciences, the United Kingdom Data Service (UKDS), the Finnish Social Science Data Archive (FSD), the Norwegian Centre for Research Data (NSD), the Data Centre Serbia for Social Sciences (DCS) and the Portuguese Social Information Archive (APIS). MDO can also be reached by email: metadata-office@cessda.eu

9 https://elsst.ukdataservice.ac.uk/

10 https://datacatalogue.cessda.eu

11 https://www.cessda.eu/About/Projects/Work-Plans/Work-Plan-2019#eqb19

12 https://creativecommons.org/licenses/by/4.0/

13 https://vocabularies.cessda.eu/

14 https://www.cessda.eu/Training

15 https://bitbucket.org/product/

16 https://basecamp.com/welcome-back

17 https://marketplace.eosc-portal.eu/

18 https://www.rd-alliance.org/

19 https://www.coretrustseal.org/