

**EOSC Future/RDA Call:
Optimising (RDA) Open Science Frameworks and Guidelines in the context of EOSC Provider**

**Project:
Framework for Increased Discoverability of Social Science Data Objects in the EOSC Portal
Service Catalogue**

**Report on descriptors of data types in popular generic descriptors, most important
distinct types of social science data objects, and most relevant metadata fields for
discovering social science data objects**

Version 1.1

Authors

Vaidas Morkevičius¹
Andrius Blažinskas¹
Antanas Štreimikis¹
Giedrius Žvaliauskas¹



This report is licensed under a Creative Commons Attribution-ShareAlike 4.0 (CC BY-SA 4.0) International License.

Recommended Citation: Vaidas Morkevičius, Andrius Blažinskas, Antanas Štreimikis, Giedrius Žvaliauskas (2022). Report on descriptors of data types in popular generic descriptors, most important distinct types of social science data objects, and most relevant metadata fields for discovering social science data objects. EOSC Future/RDA Call Report. DOI: 10.5281/zenodo.7125597

Kaunas, Lithuania
2022

¹ Kaunas University of Technology

Table of Contents

Abbreviations and acronyms.....	3
Introduction.....	4
Descriptors of data and data types in popular generic descriptors.....	6
Recommendations.....	12
Types of social science data objects.....	14
Milestone 1: Recommendation of the vocabulary for standardized description of most important data types for SSD objects.....	20
Metadata fields for detailed description of social science data objects.....	22
Survey data.....	22
Aggregated data.....	24
Recommendations.....	27
References.....	28

Abbreviations and acronyms

CESSDA ERIC	European Research Infrastructure Consortium of European Social Science Data Archives
COAR	Confederation of Open Access Repositories
DCMI	Dublin Core Metadata Initiative
DDI	Data Documentation Initiative
EOSC	European Open Science Cloud
FAIR Guiding Principles	Guidelines to improve the Findability, Accessibility, Interoperability, and Reuse of digital assets formulated in Wilkinson et al. (2016)
LiDA	Lithuanian Data Archive for Social Sciences and Humanities
OpenAIRE	Open Access Infrastructure for Research in Europe
RDA	Research Data Alliance
SSD object	Social science data object
URI	Uniform Resource Identifier

Introduction

Both the RDA and the EOSC are strongly committed to the advocacy and practical implementation of the FAIR data principles (Wilkinson et al., 2016) when developing data curation infrastructures. F2² principle of the FAIR Guiding Principles requires that data be described with rich metadata as digital resources and objects that are “not well-described cannot be accurately discovered” (Jacobsen et al., 2020). Importantly, implementation of this principle requires that appropriate generic and domain-specific descriptors are included into data catalogues and search engines, so that users are able to easily find required data. Generic descriptors (such as Dublin Core,³ DataCite,⁴ or OpenAIRE⁵) provide basic and very general information about datasets and can be used for describing various types and formats of the data in most of the domains. However, they provide little detailed information about the specifics of the data objects that are pertinent to the specific domains and are used in search queries by the users. For example, finding survey data about attitudes of Lithuanian population towards immigrants in 2007 would be hardly possible without the detailed metadata provided alongside the survey data (for example, in the DDI Codebook⁶).

EOSC Portal Service Catalogue⁷ as envisioned in its architecture would be a Web portal that facilitates searching, discovering and ordering of services from various providers across domains in European countries (EOSC Executive Board, 2021: 7). This requires harvesting not only the generic metadata from the research data providers, but also richer domain-specific metadata. For example, certain fields/elements from DDI Codebook/Lifecycle⁸ descriptors in case of SSD objects have high importance for discovery of the data. Thus, information about the fieldwork dates, participating countries, sample sizes and questions included are among the most important elements that a user wishing to compare trends of trust in the EU among European citizens from the beginning of this century would be looking for in the metadata describing various international survey data sets. Without quick access to this information (possibly, on a single search platform) it would be much more difficult and time-consuming to collect the required data. More generally, the whole idea of secondary data analysis depends on the availability of rich metadata, so that researchers aiming to analyze data created

2 www.go-fair.org/fair-principles/f2-data-described-rich-metadata.

3 oai_dc, <https://dublincore.org/specifications/dublin-core>.

4 oai_datacite, <https://schema.datacite.org/oai/oai-1.0>.

5 oai_openaire, https://openaire-guidelines-for-literature-repository-managers.readthedocs.io/en/latest/use_of_oai_pmh.html.

6 <https://ddialliance.org/Specification/DDI-Codebook>.

7 <https://marketplace.eosc-portal.eu>.

8 <https://ddialliance.org/Specification/DDI-Lifecycle>.

by other people or for other purposes be able to identify context and details of the data collection in order to decide on the quality and suitability of the data for their research purposes.

The project “Framework for Increased Discoverability of Social Science Data Objects in the EOSC Portal Service Catalogue” (further – the *Framework*) supported by the EOSC Future/RDA Call “Optimising (RDA) Open Science Frameworks and Guidelines in the context of EOSC Provider” aims to recommend a framework for harvesting and delivering for discovery rich metadata of SSD objects for the EOSC Portal Service Catalogue. The proposed framework intends to provide guidelines to enrich generic descriptors of data (such as such as oai_dc, oai_datacite or oai_openaire) with relevant additional information depending on the type of the SSD object. This conditional model would allow for flexibility and comprehensiveness at the same time, as its main operating framework would be based on the most widely used generic metadata descriptors, at the same time, integrating additional elements from domain specific metadata descriptors. This **report** is intended to attain **three objectives**:

1. Identify, describe and recommend elements within the popular generic descriptors (such as oai_dc, oai_datacite or oai_openaire) that allow to differentiate SSD objects into distinct types that require differing additional metadata blocks for enriching generic descriptions.
2. Identify, describe and recommend major distinct types of SSD objects as included into the biggest social science repositories.
3. Identify, describe and recommend most relevant metadata fields for harvesting and delivering for discovery of the identified distinct types of SSD objects based on the CESSDA Metadata Model (Akdeniz et al., 2021).

Descriptors of data and data types in popular generic descriptors

Generic data catalogs and repositories, such as Harvard Dataverse Repository, Dryad Digital Repository, or Figshare Repository,⁹ commonly use generic descriptor metadata schemes for curating their data. And even if descriptions in more specific metadata standards are available in some of them (for example, DDI Codebook metadata format in the Harvard Dataverse Repository), they are just transformations from and/or additions to generic formats. This practice is largely predetermined by the nature of these repositories – they aim to store very different data sets from variety of disciplines. Therefore, they need to keep metadata as simple as possible so that tabular, textual, coordinate, or visual data could be uniformly described.

Most commonly generic data repositories employ the Dublin Core terms and/or DataCite Metadata Schema for describing their data. The former was first conceptualized in 1995 as Dublin Core Metadata Element Set (later expanded into the DCMI Metadata Terms) and is the most common metadata standard used by libraries (Weibel 1995, 1997). The version 1.1 of the **Dublin Core Metadata Element Set** contains 15 elements:¹⁰

1. **Title**: Name of the resource.
2. **Subject**: Topic of the resource.
3. **Description**: Essential information about the resource (usually, abstract).
4. **Creator**: Author(s) of the resource.
5. **Publisher**: Publisher of the resource (making the resource available).
6. **Contributor**: Subjects making any contributions to produce the resource.
7. **Date**: Date when the resource was created/modified.
8. **Type**: The nature or genre of the resource.
9. **Format**: The file format, physical medium, or dimensions of the resource.
10. **Identifier**: An unambiguous reference to the resource (digital object identifier).
11. **Source**: Any related resource from which the described resource is derived.

⁹ For a list of major repositories, see <https://www.nature.com/sdata/policies/repositories>.

¹⁰ <https://www.dublincore.org/specifications/dublin-core/dcmi-terms/#section-3>. It is important to note that DCMI Metadata Terms is a more extensive (qualified) version of the Dublin Core introduced in 2001 and contains three additional elements (Audience, Provenance and RightsHolder), as well as element qualifiers (sub-elements/properties).

12. **Language:** Language(s) of the resource.
13. **Relation:** Any related resource (if relevant).
14. **Coverage:** The spatial or temporal topic of the resource, spatial applicability of the resource, or jurisdiction under which the resource is relevant.
15. **Rights:** Information about rights held in and over the resource.

As can be seen this scheme is well-suited for describing both digital and physical objects and was primarily used for describing library resources (such as, books). One can find two fields in the scheme, which are most relevant for the intended *Framework*. First, there is an element *Type*, which is used to describe the type (genre) of the resource. Importantly, it has a recommendation to employ DCMI Type Vocabulary (developed and maintained by the DCMI¹¹) in order to describe the type or genre of the resource. However, the DCMI Type Vocabulary contains only 11 values of a very general nature: Collection, Dataset, Event, Image, InteractiveResource, MovingImage, PhysicalObject, Service, Software, StillImage, and Text. Therefore, Dublin Core element *Type* could be employed in the intended *Framework* only in case of being extended to include more nuanced and detailed taxonomy of resource types.

Another field relevant for developing the intended *Framework* and wherein more detailed information about the resource could be included is the *Description* element. Importantly, it is not restricted in terms of content and varied information can be added (comment states: description may include but is not limited to: an abstract, a table of contents, a graphical representation, or a free-text account of the resource). However, it does not contain a specification to be used with any relevant controlled vocabulary¹². Therefore, its use is rather loose and information included may vary from resource to resource considerably.

DataCite Metadata Schema was first introduced in 2009 and was specifically aimed at the registration and description of research data (Brase et al. 2009). DataCite is also an entity issuing DOI identifiers for data objects, therefore, it requires certain standards to be followed when data object is registered and published. The version 4.4 of the Schema contains 20 properties (elements):¹³

1. **Identifier:** unique string that identifies a resource (with identifierType sub-property).

11 <https://www.dublincore.org/specifications/dublin-core/dcmi-terms/#section-7>.

12 It has to be noted that DCMI Metadata Terms has two elements (subproperties of *Description* term): *abstract* and *tableOfContents*. However, they are still either very generic (*abstract*) or more suitable to describe books (*tableOfContents*) than research data sets.

13 https://schema.datacite.org/meta/kernel-4.4/doc/DataCite-MetadataKernel_v4.4.pdf.

2. **Creator:** The main researchers involved in producing the data, or the authors of the publication, in priority order (with creatorName, nameType, givenName, familyName, nameIdentifier, nameIdentifierScheme, schemeURI, affiliation, affiliationIdentifier, affiliationIdentifier, and SchemeURI sub-properties).
3. **Title:** A name or title by which the resource is known (with titleType sub-property).
4. **Publisher:** The name of the entity that holds, archives, publishes prints, distributes, releases, issues, or produces the resource.
5. **PublicationYear:** The year when the data was or will be made publicly available.
6. **ResourceType:** A description of the resource (with resourceTypeGeneral sub-property).
7. **Subject:** Subject, keyword, classification code, or key phrase describing the resource (with subjectScheme, schemeURI, valueURI, and classificationCode sub-properties).
8. **Contributor:** The institution or person responsible for collecting, managing, distributing, or otherwise contributing to the development of the resource (with contributorType, contributorName, nameType, givenName, familyName, nameIdentifier, nameIdentifierScheme, schemeURI, affiliation, affiliationIdentifier, affiliationIdentifier, and SchemeURI subproperties).
9. **Date:** Different dates relevant to the resources (data collection date and other relevant dates, with dateType and dateInformation sub-properties).
10. **Language:** The primary language of the resource.
11. **AlternateIdentifier:** An identifier other than the primary Identifier applied to the resource being registered (with alternateIdentifierType sub-property).
12. **RelatedIdentifier:** Identifiers of related resources (with relatedIdentifierType, relationType, relatedMetadataScheme, schemeURI, schemeType, and resourceTypeGeneral sub-properties).
13. **Size:** Size (e.g., bytes, pages, inches, etc.) or duration (extent), e.g., hours, minutes, days, etc., of the resource.
14. **Format:** Technical format of the resource.
15. **Version:** The version number of the resource.

16. **Rights:** Any rights information for the resource (with rightsURI, rightsIdentifier, rightsIdentifierScheme, and schemeURI sub-properties).
17. **Description:** All additional information that does not fit in any of the other categories (may be used for technical information, with descriptionType sub-property).
18. **GeoLocation:** Spatial region or named place where the data was gathered or about which the data is focused (with geoLocationPoint, pointLongitude, pointLatitude, geoLocationBox, and other sub-properties).
19. **FundingReference:** Information about financial support (funding) for the resource (with funderIdentifier, funderIdentifierType, SchemeURI, awardNumber, awardURI, and awardTitle sub-properties).
20. **RelatedItem:** Information about a resource related to the one being described, e.g., a journal or book (with relatedItemType, relationType, relatedItemIdentifier, relatedItemIdentifierType, relatedMetadataScheme, schemeURI, schemeType, Creator and other sub-properties).

Generally, DataCite Metadata Schema extends Dublin Core terms and includes more detailed specification for geographic, funding, versioning and size information related a resource. Again, two fields in the Schema are relevant for the intended *Framework*. First, there is an element *ResourceType*, which is used to describe the type of the resource. In addition, the element has a sub-property *resourceTypeGeneral*, which allows for controlled description of the type of the resource. However, the list of values includes rather generic types, such as Audiovisual, Book, Dataset, Event, Software, Text etc. Therefore, it does not include a list of values for more detailed description of data sets (or any other generic objects). Moreover, it recommends adding free-format textual description of the resource, if one would like to provide more detailed description. Thus, DataCite Metadata Schema element *ResourceType* could also be employed in the intended *Framework* only in case of being extended to include more nuanced and detailed taxonomy of resource sub-types.

DataCite Metadata Schema also includes *Description* field relevant for developing the intended *Framework*, which is similar to *Description* element in the Dublin Core terms. However, this Schema allows for more detailed specification of the description types with sub-property *descriptionType*, which (currently) may contain six controlled values: Abstract, Methods, SeriesInformation, TableOfContents, TechnicalInfo, and Other. This sub-property is essential for making more detailed metadata available for potential secondary users of data sets as it would allow including different descriptors from discipline specific metadata schemes if properly specified.

In 2010 European Open Science infrastructure OpenAIRE started developing guidelines for its content providers (van Berchum, Rodrigues 2010). The developed **OpenAIRE Guidelines**¹⁴ and **Application Profile** has different versions (the OpenAIRE Guidelines for Literature Repository Managers, the OpenAIRE Guidelines for Data Archive Managers, the OpenAIRE Guidelines for CRIS managers, the OpenAIRE Guidelines for Software Repository Managers, and the Guidelines for Other Research Products Repository Managers). However, the OpenAIRE Guidelines for Literature Repository Managers is the main version, which is actively developed. It fits not only “Literature” repositories but is also used for describing research data sets. The latest release candidate (4.1) of these Guidelines contains 32 fields (elements), which are directly taken from either Dublin Core (7) or DataCite (12), or developed by the OpenAIRE itself (13).¹⁵ The latter ones are:

1. **Funding Reference**: Information about financial support (funding) for the resource (with several attributes).
2. **Resource Type**: The type of scientific output the resource is a manifestation of. It describes the genre of the resource (with attributes *resourceTypeGeneral* and *uri*).
3. **License Condition**: Information about license rights held in and over the resource (with several attributes).
4. **Resource Version**: Depending on the resource type this property is used to indicate: a) the version number of a dataset or software, b) the status in the publication process of journal articles (with attribute *uri*).
5. **File Location**: An unambiguous reference to the files, e.g. fulltext, the resource is associated with (with several attributes).
6. **Citation Title**: The title name of the container (e.g. journal, book, conference) this work is published in.
7. **Citation Volume**: The volume, typically a number, of the container (e.g. journal).
8. **Citation Issue**: The issue of the container (e.g. journal).
9. **Citation Start Page**: The start page is part of the pagination information of the work published in a container (e.g. journal issue).

14 <https://guidelines.openaire.eu/en/latest>.

15 https://openaire-guidelines-for-literature-repository-managers.readthedocs.io/en/latest/application_profile.html.

10. **Citation End Page**: The end page is part of the pagination information of the work published in a container (e.g. journal issue).
11. **Citation Edition**: The edition the work was published in (e.g. book edition).
12. **Citation Conference Place**: The place where the conference took place.
13. **Citation Conference Date**: the date when the conference took place.

Clearly, only the first five elements are relevant for publishing data sets. Importantly, the element *Resource Type* (*oaire:resourceType*) is an extended version of the property *ResourceType* in the DataCite Metadata Schema and is relevant for developing the intended *Framework*. It has a generic attribute *resourceTypeGeneral* for identifying the general type of a resource with a controlled list of values: literature, dataset, software, and other research product. It also requires using controlled list values from the COAR Resource Type Vocabulary¹⁶ and includes *uri* attribute for the vocabulary term linking. These properties make it the best available candidate for the development of the intended *Framework* and employing it as an element allowing to differentiate types of datasets.

On the other hand, OpenAIRE Guidelines employ *Description* element from the Dublin Core terms, which is a bit limited as it does not allow more nuanced and detailed attributes (especially, allowing for controlled values), such as those available in the DataCite alternative (namely, sub-property *descriptionType* with controlled list values). Therefore, it could be employed in the intended *Framework* only in case of being extended to include more nuanced and detailed taxonomy of resource types.

To sum up, all the three analysed generic metadata standards – Dublin Core terms, DataCite Metadata Schema, and OpenAIRE Guidelines – do contain two elements (terms) that could be employed in developing the intended *Framework*:

1. **Type (Resource Type)**. Suitable for differentiating the types of datasets. Dublin Core terms recommends using DCMI Type Vocabulary for differentiating types of resources. However, 11 values of this vocabulary are more suitable for identifying the most general types, such as, book or dataset. Similarly, DataCite Metadata Schema also allows for differentiating among the general types of resources with sub-property *resourceTypeGeneral* of this element. Again, values of resources are rather generic (Books, Software, Dataset). OpenAIRE Guidelines contain most advanced specification of resource types. It both requires to identify the general

16 http://vocabularies.coar-repositories.org/documentation/resource_types.

type of the resource in the attribute *resourceTypeGeneral*, and also demands to use COAR Resource Type Vocabulary (with *uri* attribute linking to the term) for describing the data set more precisely. These characteristics make Resource Type element from the OpenAIRE Guidelines the best exemplar of how the resource types should be specified in the intended *Framework*.

2. **Description.** Suitable for including detailed information about the various important aspects related to the context, conditions and process of data collection. In Dublin Core Terms and OpenAIRE Guidelines it is defined very loosely and quite varied information is allowed to be added. DataCite Metadata Schema allows for more detailed specification of the description types with sub-property *descriptionType*. This sub-property is essential for including different descriptors from discipline specific metadata schemes and needs to be specified, preferably, as a controlled vocabulary.¹⁷

Recommendations

1. **For identifying and differentiating dataset types** the best present option in the analyzed metadata standards is the one implemented in the OpenAIRE Guidelines. Therefore, recommendation for **Type (Resource Type)** element in metadata schemes describing in detail types of the resources is, first, to have attribute sub-element for general type, which differentiates resources into broad categories, such as, datasets, images, books, texts, software etc. In addition, the more detailed taxonomy of the resources within the more general sub-type of resources should be included into the field following an existing (updated, if needed) or newly developed controlled vocabulary with appropriate linking via URIs. Among the former the most comprehensive existing option is the COAR Resource Type Vocabulary.
2. **For including detailed information about important aspects related to the context, conditions and process of data collection** the best present option in the analyzed metadata standards is implemented in DataCite Metadata Schema. Therefore, recommendation for **Description** element in metadata schemes describing the contents of the resources is to have attribute sub-element for the type of the description following an existing (updated, if needed) or newly developed controlled vocabulary with appropriate linking via URIs. The vocabularies for

¹⁷ Also, *language* attribute should be present in both elements, as European (and other non-English-speaking) social science data archives usually curate their datasets in national languages, providing some support for English language. Therefore, the metadata is multilingual, and it would be important to differentiate between languages when harvesting metadata.

different disciplines and types of datasets would be different, however, some of the values may overlap as, for example, abstract, study notes, series and series information. The best option for SSD objects would be to employ mandatory and recommended fields from the CESSDA Metadata Model, especially, those related to “Information on Study: Methodical Information” (1.3 Methodical Information), “Information on Dataset: Content Information” (4.2 Content Information), “Information on Dataset: Variables” (4.3 Variable), “Information on Instrument: Content Information” (5.3 Content Information), “Information on Instrument: Technical Information” (5.4 Technical Information), “Information on Questions and Responses” (6 Questions and Responses), “Information on Concepts” (7 Concept), and “Information on Group of Studies” (10 Study Group).

Types of social science data objects

Since the *Framework* is specifically targeted at better discoverability of SSD objects we first of all planned to investigate the Social Science datasets in the Harvard Dataverse Repository,¹⁸ datasets in the CESSDA Data Catalogue,¹⁹ and datasets in the ICPSR Catalogue²⁰ in order to determine the most important distinct types of social science data objects. We also envisioned to consult the existing controlled vocabularies that may be relevant in classifying types of SSD objects: 1) those supported by the CESSDA Vocabulary Service,²¹ those available among the DDI Controlled Vocabularies,²² and 3) the taxonomy of resource types in the COAR Resource Type Vocabulary.²³

SSD objects are very diverse and this diversity depends on three aspects related to data collection:

1. ***Methodological approach***: qualitative vs. quantitative studies.
2. ***Instruments used***: interviews vs. surveys vs. observation vs. text analysis.
3. ***Data collected***: structured vs. unstructured.

These aspects are interrelated, as for example, in qualitative studies researchers would usually resort to less structured interviews and collect unstructured data (interview transcripts from audio or video recordings). And in quantitative studies survey collecting highly structured data would be employed.

From the preliminary analysis of the types of SSD objects available in the major data archives around the world it appeared that quantitative studies are most frequently submitted for curation, and survey as well as compiled (aggregate) data dominate the submissions. This may be related to the fact that far less data archives have implemented modules for curating qualitative data, and qualitative data curation standards are still rather underdeveloped. In addition, secondary analysis of qualitative data is still performed very rarely and the need for storage of data collected in qualitative studies is comparatively less pronounced than for various data collected in quantitative studies. Even among quantitative studies only those having a larger scale (usually, internationally comparable and longitudinal) are employed frequently in secondary analyses. Therefore, analysis of only these sources

18 <https://dataverse.harvard.edu>.

19 <https://datacatalogue.CESSDA.eu>.

20 <https://www.icpsr.umich.edu/web/pages/ICPSR/index.html>.

21 <https://vocabularies.CESSDA.eu>.

22 <https://ddialliance.org/controlled-vocabularies/all>.

23 https://vocabularies.coar-repositories.org/resource_types.

(major data archives around the world) would provide biased view on which types of SSD objects are most frequently collected and archived in repositories.

Therefore, we resorted to the analysis of existing social science related controlled vocabularies that include terms for research instruments, data types, and modes of data collection. CESSDA Vocabulary Service and DDI Controlled Vocabularies have at least three relevant controlled variables:²⁴

1. **Data Source Type**, which specifies a typology of data sources. Among the major categories we find: Registers/Records/Accounts, Events/Interactions, Processes, Communications, Research data, Population groups, Geographic areas, Physical objects, Biological samples, and Other.
2. **Mode Of Collection**, which specifies procedures, techniques, or modes of inquiry used to collect the data. The major terms are: Interviews, Self-administered questionnaires, Focus groups, Self-administered writings and/or diaries, Observations, Experiments, Recordings, Automated data extraction, Content codings, Transcriptions, Compilations/Syntheses, Summaries, Aggregations, Simulations, Measurements and tests, and Other.
3. **Type of Instrument**, which specifies a typology of data collection instruments. It contains these major values: Questionnaire, Interview scheme and/or themes, Data collection guidelines, Participant tasks, Technical instrument(s), Programming script, and Other.

In addition, we also examined the COAR Resource Type Vocabulary, which is of a more general nature (most recent version 3.1). However, it contains relevant classification of dataset types, which includes thirteen terms: aggregated data, clinical trial data, compiled data, encoded data, experimental data, genomic data, geospatial data, laboratory notebook, measurement and test data, observational data, recorded data, simulation data, and survey data. Importantly, it is already used as a list value for the *Resource Type* element in the OpenAIRE Guidelines. Therefore, ideally it should be extended/amended (if need be) in order to include important additional types of datasets (if missing) relevant for SSD objects.

The following table aggregates the information about data/instrument types from the above mentioned resources and provides a taxonomy of major types of SSD objects.

Table 1. Major types of social science data objects.

²⁴ Vocabularies for *Data Type* and *General Data Format* were excluded as they describe more technical aspects (format) of data, such as, whether the data is numeric, textual, or of string, integer, date and/or similar type.

Type	Subtype	Description
Experimental data	Laboratory Field Web	Data resulting from the experimental research in laboratories, field (natural) settings (including quasi-experiments), or on the Web, wherein a researcher tries to manipulate certain conditions relevant for changes in participant behavior and attitudes. Important characteristics that have to be recorded and specified for experimental data are: specific design of experiment, recruitment, numbers and random allocation of participants, number of groups, stimulation materials used, recordings and surveys conducted (if any) etc.
Survey data	Cross-sectional Longitudinal cross-sectional Panel	Data resulting from a survey, where a sample for a population of subjects is studied by means of answers to a set of researcher pre-planned questions. Important characteristics that have to be recorded and specified for survey data are: specific design of survey, sample design and sampling, unit of analysis, mode of collection (web, personal, telephone etc.), response rate, sample attrition, questionnaire, questions and answer scales, interviewing, anonymisation rules etc.
Interview data	Semi-structured Unstructured	Data resulting from an interview, a pre-planned communication between two people - the interviewer and the interviewee - in which interviewer aims to obtain some information from the interviewee. Important characteristics that have to be recorded and specified for interview data are: level of structuration, type of sampling, interviewees and their characteristics, mode of data collection, themes and/or questions asked, anonymisation rules etc.
Focus group data		Data resulting from a focus group interview on a particular topic, organized for research purposes. The discussion is moderated by researcher or moderator. Important characteristics that have to be recorded and specified for focus group data are: information about moderator, number of discussions, recruitment of participants, themes and/or questions asked, stimulation materials used, recordings (if any) made, transcription information etc.
Observational data	Structured Unstructured	Data resulting from collecting information as it occurs (for example, observing behaviors, events, development of condition or disease, etc.) without attempting to manipulate anything. Important characteristics that have to be recorded and specified for observational data are: sampling of observation sites, observation site information, observation targets, observation protocols, field notes etc.

Type	Subtype	Description
Aggregated data	Population statistics table Cultural statistics table Economic statistics table Educational statistics table Health statistics table Historical statistics table Political (including legal and administrative) statistics table Social statistics table	Data aggregated or assembled from other sources into a new table, sometimes called data cube (macro level data). Data could be averaged, totaled, or otherwise derived from individual-level data, such as, available in census/enumeration data or voting results. Data could be also collected or assembled from multiple, often heterogeneous sources that have some reference points in common. The data are compiled into a new entity, which is a table (matrix) with at least one dimension devoted to geographic or temporal aspect of the data. If temporal aspect of the data is present dataset may be called time-series data. For example, crime statistics in regions of a country by years (two-way table) or different types of criminal offenses in regions of a country by years (three-way table). Important characteristics that have to be recorded and specified for aggregated data are: sources of the data, data collection (extraction) methods, information about any transformations, pre-coding/re-coding and anonymisation measures applied, information about dimensions of the data, information about geographical and time-series units etc.
Textual data	Structured Unstructured	Data resulting from collecting textual information from various sources, such as, printed and internet media articles, social media posts content, blogs, legal, administrative and political documents etc. Important characteristics that have to be recorded and specified for textual data are: producers of the texts, sources of the data, data collection (extraction) methods, information about level of structuration, information about any transformations, pre-coding/re-coding and anonymisation measures applied etc.
Administrative records data		Data originating in official, formal, or semi-formal records listing for example items, names, occurrences, actions, or results, and preserved in written or digital form. Data is derived from information collected on individuals or other entities (micro level data) as part of the routine administrative procedures of an agency, business, or institution. Such data are not usually collected with research purposes in mind, usually is voluminous, and may require preparation such as coding in order to be usable by researchers. Important characteristics that have to be recorded and specified for administrative records data are: sources of the data, size of the data, update intervals, information about any transformations, pre-coding/re-coding and anonymisation measures applied etc.

Type	Subtype	Description
Network data	Social media Communication Collaboration Interactions	Data consisting of a square matrix of measurements of connections/interactions/exchanges. Both rows and columns of the matrix represent the same set of cases, subjects, or other entities. And each value in the cells of the matrix describes a relationship (or its degree) between the entities. Important characteristics that have to be recorded and specified for network data are: network entities (subjects), sources of the data, data collection methods, information about any transformations, pre-coding/re-coding and anonymisation measures applied etc.
Recorded data	Audio Still images Moving images	Data registered by mechanical or electronic means, in a form that allows the information to be retrieved and/or reproduced. For example, images or sounds on disc or magnetic tape. Examples could be recorded sound, including voice, music etc.; moving images, such as, films, animation, digital recordings, visual output from simulations, recorded television programs. Important characteristics that have to be recorded and specified for recorded data are: producers of the recordings, sources of the data, data collection (extraction) methods, information about level of structuration, information about any transformations, pre-coding/re-coding and anonymisation measures applied etc.
Encoded data		Data derived from content coding applied to qualitative data (textual, video, audio or still-image) originally produced for other purposes into quantitative data (expressed in cases-by-variable matrices) in accordance with pre-defined categorization (coding) schemes. Important characteristics that have to be recorded and specified for encoded data are: producers and types of the original qualitative data, original qualitative data collection (extraction) methods, information about any transformations, pre-coding/re-coding and anonymisation measures applied, content coding rules (machine or human coding), categorization (coding) schemes applied, categorization (coding) schemes construction description, units of context and coding, coding validity and reliability information etc.

Type	Subtype	Description
Other	Summaries Time budget diaries Narratives or essays written by study participants (including life-stories) Conversational data Event data Measurement and test data Simulation data (including program source code and programming scripts)	Other data types that are less frequently used and/or collected in the social science research. The sub-type list is not comprehensive and additional types may be considered.

As it can be seen from the Table 1 at least 11 major types of SSD objects could be identified. Important aspect of these dataset types is that they require some specific information to be included in order that the data be described as fully as possible. For example, encoded data and textual data may have the same original set of textual data (say, party manifestos). However, textual data would differ from the encoded data since the former would include textual data more or less “as-it-is”, and the latter one may not include the original textual data at all, just a matrix of codes applied to textual units. Consequently, information about the original sources of data would be included in both instances (such as, producers, time, collection methods etc.), and information about different coding strategies and instruments applied only in the case of encoded data.

The presented taxonomy could be a starting point²⁵ for developing an actual controlled vocabulary of types of SSD objects, which could eventually be included into a more general taxonomy of data types (not just social science). It could also be integrated into the existing COAR Resource Type Vocabulary. Most importantly, some version of it should be included into the *Type (Resource Type)* element of metadata standards for more detailed and standardized description of the type of resources.

25 We do not suggest that our taxonomy is a final say in the field, just a proposition of a possible draft version.

Milestone 1: Recommendation of the vocabulary for standardized description of most important data types for SSD objects

1 Social science dataset

1.1 Experimental data

1.1.1 Laboratory experiment

1.1.2 Field experiment

1.1.3 Web experiment

1.2 Survey data

1.2.1 Cross-sectional survey

1.2.2 Longitudinal survey

1.2.3 Panel survey

1.3 Interview data

1.3.1 Semi-structured interview

1.3.2 Unstructured interview

1.4 Focus group data

1.5 Observational data

1.5.1 Structured observation

1.5.2 Unstructured observation

1.6 Aggregated data

1.6.1 Population statistics

1.6.2 Cultural statistics

1.6.3 Economic statistics

1.6.4 Educational statistics

1.6.5 Health statistics

1.6.6 Historical statistics

1.6.7 Political statistics

1.6.8 Social statistics

1.7 Textual data

1.7.1 Structured

1.7.2 Unstructured

1.8 Administrative data

1.9 Network data

1.9.1 Social media

1.9.2 Communication

1.9.3 Collaboration

1.9.4 Interactions

1.10 Recorded data

1.10.1 Audio recordings

1.10.2 Still images

1.10.3 Moving images

1.11 Encoded data

1.12 Other social science data

1.12.1 Summaries

1.12.2 Time budget diaries data

1.12.3 Narrative/essay data

1.12.4 Conversational data

1.12.5 Event data

1.12.6 Measurement and test data

1.12.7 Simulation data

Metadata fields for detailed description of social science data objects

Since at least eleven distinct major types of SSD objects were identified in the previous section of the report, it is not possible to include extended specifications of metadata fields relevant for their detailed descriptions. We chose two most common types – *survey data and aggregated data* – for more detailed specification of metadata fields relevant for detailed descriptions of these types of datasets. As our source of metadata fields we used CESSDA Metadata Model (Akdeniz et al., 2021), one of the most authoritative guides for creating metadata in the social science domain. Also, we consulted DDI Codebook and DDI Lifecycle, as these two metadata standards are primarily devoted to describing SSD objects. Description of datasets involves a lot of generic information, which is also relevant for other types of resources, such as, books, software or video recordings. Elements *Title*, *Subject*, *Creator*, *Publisher*, *Contributor*, *Date*, *Identifier*, *Language*, *Rights* (if taken from the Dublin Core terms) are the most obvious examples. In what follows we detail only the other important characteristics of survey and aggregated datasets, in addition to general descriptor fields.

Survey data

Surveys are among the most frequently used instruments in data collection for social sciences, at least in quantitative studies. They may be of various types themselves: single country or international, cross-sectional, repeated cross-sectional or panel, highly structured (with only closed answer scales) or less highly structured (with some (only very few) open answer scales). However, the most important other characteristics are similar to all of them. Table 2 enumerates them and gives more detailed descriptions.

Table 2. Metadata fields for detailed description of survey data.

Metadata field	Description of metadata field
Study Series	Name and history of the dataset series to which the dataset belongs, and summary of features that apply to the series as a whole.
Study Group	Information on group(s) of studies.
Time Method	Describes the time dimension of the data collection and frequency of data collection. Controlled vocabulary available: “DDI Time Method”.
Time Period Covered	Time period to which the data refer. This item reflects the time period covered by the data, not the dates of coding or making documents machine-readable or the dates the data were collected.

Survey Period	Start of data collection and end of data collection, data collection single date, and data collection period.
Geographic Coverage	Information on the geographic coverage of the data. Includes the total geographic scope of the data, e.g., Country / Nation, State / Province, City, Other
Geographic Unit	Lowest level of geographic aggregation covered by the dataset, e.g., village, county, region.
Completeness of Study Stored	This item indicates the relationship of the data collected to the amount of data coded and stored in the data collection. Information as to why certain items of collected information were not included in the data file stored by the archive should be provided.
Universe	Description of the population covered by the data in the file; the group of people or other elements that are the object of the study and to which the study results refer. The universe may consist of elements other than persons, such as housing units, court cases, deaths, countries, and so on. Also known as the universe of interest, population of interest, and target population.
Unit of Analysis	Describes the entity being analyzed in the study or in the variable, such as individuals, families/households, groups, institutions/organizations, administrative units, and more. Controlled vocabulary available: "DDI Analysis Unit".
Sampling Information	Sample frame information, type of sampling procedure used for data collection, target sample size, major deviations for sample design, response rate, estimates of sampling error. Controlled vocabulary available: "DDI Sampling Procedure".
Mode of Data Collection	Description of the method of data collection - the procedure, technique, or mode of inquiry used to collect. Controlled vocabulary available: "DDI Mode of Data Collection".
Type of Research Instrument	Type of data collection instrument used. Controlled vocabulary available: "DDI Type of Instrument".
Instrument Development	Describes any development work on the data collection instrument.
Data Collector	Individual, agency or organization responsible for administering the questionnaire or interview.
Collector Training	Type of training provided to the data collector.
Characteristics of Data Collection Situation	Description of noteworthy aspects of the data collection situation. Includes information on factors such as cooperativeness of respondents, duration of interviews, number of call backs, or similar.
Actions to Minimize Losses	Summary of actions taken to minimize data loss. Include information on actions such as follow-up visits, supervisory checks, historical matching, estimation, and so on.
Control Operations	Methods to facilitate data control performed by the primary investigator or by the data archive.

Cleaning Operations	Methods used to clean the data collection, such as consistency checking, wildcode checking, or other.
Other Forms of Data Appraisal	Other issues pertaining to the data appraisal. Describe issues such as response variance, nonresponse rate and testing for bias, interviewer and response bias, confidence levels, question bias, or similar.
Weighting	Describes the criteria for using weights in analysis of the dataset, since the use of sampling procedures might require to apply weights to produce accurate statistical results.
Imputation	Describes procedures used by which missing values were estimated for items that a survey respondent failed to provide.
Data Processing	Describes various data processing procedures not captured elsewhere in the documentation, such as derivation, topcoding, recoding, suppression, anonymization etc.
Missing Data	This element can be used to give general information about missing data, e.g., that missing data have been standardized across the collection, missing data are present because of merging, etc.
Variable Information	Variable description, variable types, measurement level, answer category names, variable derivation instructions.
Information on Questions and Responses	Literal questions and responses for the relevant variables. Showcards, if used.
Data File Information	Number of data files, file descriptions, number of cases and number of variables in data files.
Study Notes	General notes on the study and data collection methods.
Survey Data Notes	Additional notes on dataset, data files, any study level errors, variables etc.

These 30 metadata fields could be used as a starting point for developing an actual controlled vocabulary of additional description types for survey data, which could eventually be included into a more general taxonomy of description types (not just social science). Importantly, some version of it should be included as an attribute sub-element for the *Description* element of metadata standards for more detailed and standardized description of SSD objects.

Aggregated data

Another very frequent type of SSD objects is aggregated data. This type of SSD objects may represent data aggregated or assembled from other sources into a single new table, sometimes called a data cube or matrix, and generally, contain two types of data: 1) averaged, totaled, or otherwise derived from individual-level data, such as, available in census/enumeration data or voting results; 2) collected

or assembled from multiple, often heterogeneous sources that have some reference points in common. The data are compiled into a new entity, which is a table (data cube, matrix) with at least one dimension devoted to geographic or temporal aspect of the data. If temporal aspect of the data is present, dataset is called time-series data. Very different information may be assembled into aggregated data tables relevant for social science research: population statistics, cultural statistics, economic statistics, educational statistics, public health statistics, political (including legal and administrative) statistics, social statistics, and historical statistics. Though purposes, methods and contents of the data contained in aggregated datasets may vary, most of the important descriptors are relevant to all of them. Table 3 enumerates them and gives more detailed descriptions.

Table 3. Metadata fields for detailed description of aggregated data.

Metadata field	Description of metadata field
Study Series	Name and history of the dataset series to which the dataset belongs, and summary of features that apply to the series as a whole.
Time Method	Describes the time dimension of the data collection. Controlled vocabulary available: “DDI Time Method”.
Time Period Covered	Time period to which the data refer. This item reflects the time period covered by the data, not the dates of coding or making documents machine-readable or the dates the data were collected.
Geographic Coverage	Information on the geographic coverage of the data. Includes the total geographic scope of the data, e.g., Country / Nation, State / Province, City, Other
Geographic Unit	Lowest level of geographic aggregation covered by the dataset, e.g., village, county, region.
Completeness of Study Stored	This item indicates the relationship of the data collected to the amount of data coded and stored in the data collection. Information as to why certain items of collected information were not included in the data file stored by the archive should be included.
Universe	Description of the population covered by the data in the file; the group of people or other elements that are the object of the study and to which the study results refer. The universe may consist of elements other than persons, such as housing units, court cases, deaths, countries, and so on. Also known as the universe of interest, population of interest, and target population.
Dimension	This element identifies dimensions of the table, and should be repeated to describe each of the table dimensions.
Measure	The element measure indicates the measurement features of the table cell content: type of aggregation used, measurement unit, and measurement scale. Two tables may be identical except for their measure - for example, a count of persons by age and percent of persons by age.

Data Sources	List of books, articles, serials, or machine-readable data files that served as the sources of the data collection.
Origin of Sources	Information about the origin of the sources and the rules followed in establishing the sources.
Characteristic of Sources Noted	Assessment of characteristics and source material.
Documentation and Access to Sources	Level of documentation of the original sources.
Mode of Data Collection	Description of the method of data collection - the procedure, technique, or mode of inquiry used to collect. Controlled vocabulary available: "DDI Mode of Data Collection".
Type of Research Instrument	Type of data collection instrument used. Controlled vocabulary available: "DDI Type of Instrument".
Instrument Development	Describes any development work on the data collection instrument, including data collection reliability checks.
Date of Collection	Start of data collection and end of data collection, data collection single date, and data collection period.
Data Collector	Individual, agency or organization responsible for aggregating or assembling the data.
Coding Instructions	Describes specific coding instructions used for data aggregating, assembling, processing, or tabulation, including coder instructions.
Collector Training	Type of training provided to the data collector.
Characteristics of Data Collection Situation	Description of noteworthy aspects of the data collection situation.
Actions to Minimize Losses	Summary of actions taken to minimize data loss. Include information on actions such as historical matching, estimation, and so on.
Cleaning Operations	Methods used to clean the data collection, such as error checking, summation checking, or other.
Other Forms of Data Appraisal	Other issues pertaining to the data appraisal.
Imputation	Describes procedures used by which missing values were estimated for table cells for which information was not available.
Data Processing	Describes various data processing procedures not captured elsewhere in the documentation, such as derivation, topcoding, recoding, suppression, anonymization etc.
Missing Data	This element can be used to give general information about missing data, e.g., that missing data have been standardized across the collection, missing data are present because of merging, etc.
Data File Information	Number of data files, file descriptions, file dimensions.
Study Notes	General notes on the study and data collection methods.

Aggregated Data Notes	Additional notes on the dataset, data files, any data errors etc.
-----------------------	---

Again, these 30 metadata fields could be used as a starting point for developing an actual controlled vocabulary of additional description types for aggregated data. It can be seen that most of the types recommended for both survey and aggregated data overlap. However, some of them are relevant for only one type. Therefore, an integrated version extending to all the SSD objects identified in previous section of this report should be included as an attribute sub-element for the *Description* element of metadata standards for more detailed and standardized description of SSD objects.

Recommendations

1. Metadata fields relevant for detailed description of **both survey and aggregated data** are: Study Series, Time Method, Time Period Covered, Geographic Coverage, Geographic Unit, Completeness of Study Stored, Universe, Mode of Data Collection, Type of Research Instrument, Instrument Development, Data Collector, Collector Training, Characteristics of Data Collection Situation, Actions to Minimize Losses, Cleaning Operations, Other Forms of Data Appraisal, Imputation, Data Processing, Missing Data, Data File Information, Study Notes. These metadata fields are relevant for most of the SSD objects (at least produced in quantitative studies), thus, they could be used as a starting point for developing controlled vocabulary of SSD objects' *Description* types.
2. Metadata fields relevant for detailed description of **survey data** are: Study Group, Survey Period, Unit of Analysis, Sampling Information, Control Operations, Weighting, Variable Information, Information on Questions and Responses, Survey Data Notes. These metadata fields should be added to controlled vocabulary of SSD objects' *Description* types, as they are essential for extended description of this type of data.
3. Metadata fields relevant for detailed description of **aggregated data** are: Dimension, Measure, Data Sources, Origin of Sources, Characteristic of Sources Noted, Documentation and Access to Sources, Date of Collection, Coding Instructions, Aggregated Data Notes. These metadata fields should also be added to controlled vocabulary of SSD objects' *Description* types, as they are essential for extended description of this type of data.

References

- Akdeniz, E., Borschewski, K., Moilanen, K., et al. (2021). *CMM CESSDA Metadata Model (2.0)*. Zenodo. doi: 10.5281/zenodo.4751455.
- Brase, J., Farquhar, A., Gastl, A. et al. (2009). Approach for a joint global registration agency for research data. *Information Services & Use*, 29(1), 13-27. doi: 10.3233/ISU-2009-0595.
- EOSC Executive Board (2021). *EOSC Architecture Working Group View on the Minimum Viable EOSC. Report from the EOSC Executive Board Working Group (WG) Architecture*. European Commission.
- Jacobsen, A., Kaliyaperumal, R., Bonino da Silva Santos, L.O. et al. (2020). A generic workflow for the data FAIRification process. *Data Intelligence* 2, 56-65. doi: 10.1162/dint_a_00028.
- van Berchum, M., Rodrigues, E. (2010). OpenAIRE Guidelines 1.0 : Guidelines for content providers of the OpenAIRE information space. Zenodo. doi: 10.5281/zenodo.59204.
- Weibel, S.L. (1995). Metadata: The Foundations of Resource Description. *D-Lib Magazine*, 1(1). Available online at: <http://www.dlib.org/dlib/July95/07weibel.html>.
- Weibel, S.L. (1997). The Dublin Core: A Simple Content Description Model for Electronic Resources. *Bulletin of the American Society for Information Science*, 24(1), 9-11.
- Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Nature: Scientific Data* 3, 160018. doi: 10.1038/sdata.2016.18.