CESSDA ERIC Agenda 21-24, Tasks 21-22

Tools Pillar: Data Access Information Upgrade

D3 Specification for interoperable access conditions in CDC

Document info

Dissemination Level	PU
Due Date of Deliverable	31/05/21
Actual Submission Date	14/05/21
Type	Report
Approval Status	Approved by Tools Working Group Leader Mari Kleemola
Version	V2.0
Number of Pages	p.1 – p.13
DOI	10.5281/zenodo.7252394

Version history

Version	Date	Comment	Revised by
0.1	06.05.2021	First draft produced by NSD and UKDS in tools task 3	NSD/UKDS
0.2	10.05.2021	First version for peer-review	NSD
1.0	14.05.2021	First version produced by NSD and UKDS	NSD/UKDS
2.0	24.10.2022	Added information that specifications are no longer valid	Sikt

Author List

Organisation	Name	Contact information
UK Data Service	Sharon Bolton	sharonb@essex.ac.uk
NSD – Norwegian Centre for Research Data	Morten Jakobsen	Morten.jakobsen@nsd.no
NSD – Norwegian Centre for Research Data	Silje Storviken	Silje.storviken@nsd.no

Peer-review

Organisation	Name	Contact information
NSD – Norwegian Centre for Research Data	Trond Kvamme	Trond.kvamme@nsd.no

Contents

Executive Summary

Please note that the specifications presented in this report is no longer valid, updated specifications are published in `Guidelines and updated specifications for CESSDA Service Providers to make data access interoperable´[1]

This report is created for CESSDA ERIC Agenda 21-24 Tools pillar Task 3 Deliverable 3. The Deliverable text is as follows: "Specifications for interoperable access conditions in CESSDA Data Catalogue (CDC) (including changes needed to CDC)". This continues the work done in "Report on creating a CDC Publisher Names Vocabulary and investigating Data Access Interoperability" (Bolton and Jääskeläinen, 2021) which identifies common access categories for CESSDA SPs. This report examines the feasibility of these access conditions within the Data Documentation Initiative (DDI) framework, concluding that a flat three-part categorization can not function within current iterations of DDI. In considering the DDI landscape, and planning for future DDI versions, a revision is done of the proposed access categories. A top level two-part between **Open** & **Restricted** is defined, for implementation in current iterations of DDI. Sub-categories of **Restricted**, **Restricted - Safeguarded** & **Restricted - Secure Access** are also defined in preparation of newer versions of DDI. CDC Publishers were consulted with this redefinition and almost all feedback was positive, which indicates an agreement with the access categories presented here that are to be implemented in the CDC. This sets the foundation for the following work within Task 3 Data Access Information Upgrade.

Abbreviations and Acronyms

CDC	CESSDA Data Catalogue
CESSDA	Consortium of European Social Science Data Archives
CMM	CESSDA Metadata Model
CMV	CESSDA Metadata Validator
CV	Controlled Vocabulary
CVS	CESSDA Vocabulary Service
DDI	Data Documentation Initiative
DDI-C	DDI (see above) Codebook
DDI-L	DDI (see above) Lifecycle

[1] Jakobsen, Morten. (2022). Guidelines and updated specifications for CESSDA Service Provides to make data access interoperable (1.0). Zenodo. DOI: https://doi.org/10.5281/zenodo.5554488

GESIS	GESIS - Leibniz-Institute for the Social Sciences
MDO	CESSDA Metadata Office
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
SP	CESSDA Service Provider (an organisation, usually a national data archive, which is responsible for providing relevant services in the framework of CESSDA)
UKDS	UK Data Service

1. Introduction

The CESSDA Data Catalogue[2] (CDC) includes metadata on the data collections of CESSDA's Service Providers (SPs) and associated organisations (referred to as CDC Publishers within this report). It allows search and discovery across all these collections, enabling access to the catalogue records of over 30,000 pan-European social science research studies. To make sure the research needs of CESSDA and its data user community are met, the CDC needs to be easy to use and able to provide precise results. To facilitate this, a range of search filters has been developed to ensure that the many thousands of records in the CDC are easily navigable and browsable.

One catalogue filter yet to be constructed covers data access criteria, so that users can see easily whether they are likely to be able to gain access to the data in their chosen study. Some data have severely restricted access conditions, meaning that they are unlikely to be widely available to researchers, whereas other data may be fully open to all without any restrictions. A filter that allows researchers to select data on the basis of availability is very useful, as it will allow them to see quickly whether they are likely to be able to access them or not and can plan their research accordingly. For example, if time is limited, researchers or students may only be interested in open data they can access quickly and easily.

To make a data access filter work effectively, it needs to work well for all organisations who provide records for the CDC. CDC Publishers vary in size and nature and may have very different data access criteria. To make sure all records can be included in the filter, a controlled vocabulary (CV) must be built that is interoperable; i.e. it is able to encompass all the different access criteria across organisations and allocate each record to a suitable CV category. The intention to implement data access interoperability and a timescale in which it should be completed, is built into the CESSDA Data Access Policy[3], under Principle 11: 'Access conditions to data shall, by 2022, be fully interoperable.'

[2] "CESSDA Data Catalogue: https://datacatalogue.cessda.eu/"; [accessed 5 May 2021]
[3] Woollard, M., L'Hours, H. and Beedham, H. (2016). CESSDA Data Access Policy (Version 1). Zenodo. DOI: http://doi.org/10.5281/zenodo.4054793

2. Background

In 2020, the CESSDA Work Plan Task Metadata Office (MDO) Task 2 project undertook background work to begin the process of creating a data access filter. This work is fully described in the MDO Task 2 Deliverable 'D3: Report on creating a CDC Publisher Names Vocabulary and investigating Data Access Interoperability'[4], but briefly, the work included:

1. A comparison of access conditions between two large CESSDA SP data archives with varied collections that include data across the full spectrum of likely access conditions: the UK Data Service (UKDS) and GESIS – Leibniz-Institute for the Social Sciences (GESIS), based in Germany;
2. the construction of a table of access conditions that encompassed the holdings of both organisations, and;
3. future recommendations on how the work could be taken forward. This included consulting other CESSDA SPs to see how their holdings would fit into the suggested categories, and working towards eventual addition of an element to the CDC metadata DDI profiles and the CESSDA Metadata Validator[5] (CMV) that would enable data access interoperability.

The table of access conditions developed in MDO Task 2 is provided below. The following access categories were proposed:

- open data, with no restrictions on access;
- data with some restrictions on access ('safeguarded/accountable', though terminology may vary across organisations);
- Restricted data, accessible only under very limited conditions and/or secure locations.

Table 1: Table of proposed access conditions

Category	**Access credentials**	**Registration**	**Licence(s) (non-commercial use always allowed)**
Open	Free download	No	Open (e.g. Creative Commons) Commercial use allowed
	Free download and click-through terms of use	No	Open Commercial use allowed
	Free download and click-through terms of use	Yes	Open Commercial use allowed

[4] Bolton, Sharon, & Jääskeläinen, Taina. (2021). Report on creating a CDC Publisher Names Vocabulary and investigating Data Access Interoperability (Version 2.0). Zenodo. DOI: https://doi.org/10.5281/zenodo.4524429

[5] "CESSDA Metadata Validator: https://cmv.cessda.eu/#!validation"; [accessed 7 May 2021]

Safeguarded / Accountable	Secure download and signed user contract/licence (registration or application)	Yes	Signed user contract/licence Commercial use allowed
	Secure download and signed user contract/licence (registration or application)	Yes	Signed user contract/licence Commercial use not allowed
	Secure download and signed user contract/licence (registration or application)	Yes	Signed user contract/licence Commercial use allowed Additional conditions
	Secure download and signed user contract/licence (registration or application)	Yes	Signed user contract/licence Commercial use not allowed Additional conditions
Restricted	Secure remote access and signed user contract/licence (registration or application)	Yes	Signed user contract/licence Commercial use not allowed Additional conditions
	Secure remote access and signed user contract/licence (registration or application)	Yes	Signed user contract/licence Commercial use allowed Additional conditions
	On-site use only and signed user contract/licence (registration or application)	Yes	Signed user contract/licence Commercial use not allowed Additional conditions
	On-site use only and signed user contract/licence (registration or application)	Yes	Signed user contract/licence Commercial use allowed Additional conditions

It is a complex table, but the MDO Task 2 team felt it would be understandable and reflect all access conditions that were likely to be in place across SPs. It was not intended to make SPs change any of their access conditions, but to provide a useful framework where they would be able to map their existing access conditions to one of the three broad categories.

In 2021, CESSDA Tools Task 3, Data Access Information Upgrade, has taken over data access interoperability, and the work done towards implementing the recommendations from MDO Task 2, and the construction of the filter, is described below.

3. Interoperable access conditions in CESSDA Data Catalogue

3.1 Challenges

In theory, the three part categorization presented in table 1 is an excellent foundation, but when practical considerations are taken into account difficulties arise. DDI[6] is the main metadata standard CDC Publishers use, and it is also the metadata standard CDC accepts. In order to make data access interoperable in CDC it must be defined within this framework.

[6] "Data Documentation Initiative: https://ddialliance.org/"; [accessed 29 April 2021]

D3 in MDO Task 2 2020 points out that DDI does not support a CV for relevant elements, and this is still the case at the time of this report. Development work on DDI Codebook (DDI-C) indicates that this is planned for version 2.6[7], but the situation is unknown for DDI-Lifecycle[8] (DDI-L). Without support for a CV it becomes difficult, if not impossible, to specify or implement a CV in relevant elements defined in the CESSDA Metadata Model[9] (CMM) expressed in DDI-L or the CDC DDI-C profiles[10]. Updating these profiles is also an objective of this task, and the practical constraints observed need to be considered in defining the CV.

3.2 Specifying a controlled vocabulary for data access

Three objectives are taken into consideration when defining the CV: (1) The CV must encompass all access categories used by CDC Publishers. (2) The CV must work within the current iterations of DDI. (3) The CV can be implemented in future iterations of DDI that support the use of CVs in the relevant elements.

In order to achieve these goals a redefinition of the three-part categorization is necessary. Instead of a flat three way distinction, a two-part CV with subcategories is introduced as shown in table 2.

Table 2: Data Access Categories

Category	**Definition**
Open	Free download with or without click-through terms of use and/or registration
Restricted	Secure download, remote access or on-site use only, signed user contract/licence (registration or application)
Restricted - Safeguarded	Secure download and signed user contract/licence (registration or application)
Restricted - Secure access	Secure remote access or on-site use only, signed user contract/licence (registration or application)

[7] "'Data access vocabulary element?': https://ddi-alliance.atlassian.net/browse/DDILIFE-3687"; [accessed 29 April 2021]
[8] "'Data access CV element': https://ddi-alliance.atlassian.net/browse/DDILIFE-3687"; [accessed 29 April 2021]
[9] Akdeniz, Esra, Borschewski, Kerrin, Moilanen, Katja, Jääskeläinen, Taina, Jakobsen, Morten, Storviken, Silje, … Bolton, Sharon. (2021). CMM CESSDA Metadata Model (Version 2.0). Zenodo. DOI: https://doi.org/10.5281/zenodo.4672414
[10] Bell, Darren. (2021, March 11). CESSDA Data Catalogue - DDI Codebook Profile (Version 1.0.4). Zenodo. DOI: http://doi.org/10.5281/zenodo.4580376

At the top level a two-part distinction is made between **Open** & **Restricted**. These categories represent the most basic information on data access. Users of the CDC will be informed on whether the data are freely available or if there are conditions that need to be met to get access. Detailed information on the exact conditions should always be available from the CDC Publisher. By keeping this level two-part it gives flexibility in specifying mappings for CDC Publishers in current DDI iterations by allowing a **True** or **False** representation. This will allow CDC Publishers to tag studies that are **Open** and everything else will be treated as **Restricted**.

The sub-categories of **Restricted** indicate to the users that there is an application process for both, but the circumstances under which they can access data are different. After meeting the criteria in the category of **Restricted - Safeguarded** data is transferred to the user. When meeting the criteria in the category **Restricted - Secure access** users are either given remote access or are able to access data on-site. The exact criteria that need to be met should always be available from the CDC Publishers.

The differentiation of these criteria is important since this can affect the analytical possibilities for the CDC user, either in limiting software usage or the possibilities of combining the accessed data with other data.

Implementation of the sub-categories is in major part intended for DDI versions that support the use of CV in relevant elements. If there are CDC Publishers who need to be able to make this distinction in their metadata records in current DDI iterations, the DDI profile will support this. Development of this capability will only be done based on explicitly expressed interest from CDC Publishers. Before this can be assessed, CDC Publishers were given the chance to review the proposed categories.

4. Data Access CV and CDC Publishers

4.1 Reaching Out

Task 3 Data Access Information Upgrade reached out to CDC Publishers through email to inform about the project and referring to the previous work done in "Report on creating a CDC Publisher Names Vocabulary and investigating Data Access Interoperability"[11] to get some feedback on the proposed access categories (appendix 1). The project also contacted SPs who are preparing Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) for harvesting to the CDC to get their perspectives.

[11] Bolton, Sharon, & Jääskeläinen, Taina. (2021). Report on creating a CDC Publisher Names Vocabulary and investigating Data Access Interoperability (Version 2.0). Zenodo. DOI: https://doi.org/10.5281/zenodo.4524429

The main focus in deliverable 3 of the Data Access Information Upgrade project was to get CDC Publishers to evaluate if they are in need of a three-part CV consisting of the categories **Open, Safeguarded/Accountable & Restricted** or if its sufficient for their organisation with a two-part CV consisting of **Open & Restricted** with the possibilities of sub-categories (appendix 1). In the reminder sent out it was informed that no feedback indicates no issues with choosing a two-part CV for data access.

4.2 Feedback

The project contacted 22 different organizations (which includes SPs that are not represented in CDC) and received feedback from 10 CESSDA metadata contacts. For most of these organizations it is sufficient to have a two-part CV with sub-categories. One CDC Publisher gave feedback that they would prefer to have a three-part CV. Having Safeguarded/Accountable and Restricted visible is an important distinction from the users' point of view. The project's point of view is that a top-level two-part CV is easiest for CDC Publishers to map to initially. The project informed the CDC Publisher of this, and no further questions were raised. The difference between Safeguarded/Accountable and Restricted can be expressed in the sub-categories (see table 2 Data Access Categories).

There was a question from a CDC Publisher about how to handle embargoes but that's not something being addressed in this project as this task focuses on the data access CV itself.

5. Moving Forward

5.1 Future Plans

After receiving feedback from CDC Publishers that the proposed categorization of data access meets their access conditions the project now needs to define the CV in CESSDA Vocabulary Service[12] (CVS). The technical capacities of OAI-PMH endpoints must also be examined, configuring OAI-PMH headers[13] to set up a **True** or **False** filter seems promising. After the most reasonable approach has been found, this needs to be translated to the DDI profiles used in CMV for validation-gating to the CDC. The task will get in touch with CDC Publishers who expressed the need to be able to distinguish between **Restricted - Safeguarded** & **Restricted - Secure access,** to see if they intend to do it using DDI 2.5.

Even though CDC Publishers use DDI as a metadata standard, the experience is that it can be challenging to harmonize metadata within this framework. CDC Publishers are encouraged to use their roles, both as users of DDI and members of the DDI Alliance, to enhance this metadata standard. CVs are crucial for harmonization, and increasing the

[12] "CESSDA Vocabulary Service: https://vocabularies.cessda.eu/"; [accessed 7th May 2021]
[13] "OAI-PMH Record: http://www.openarchives.org/OAI/openarchivesprotocol.html#Record"; [accessed 4th May 2021]

flexibility of implementing these in DDI will be beneficial internally for SPs and for cross-national organizations such as CESSDA.

5.1.1 Future Discussion

How to handle data under embargoes and licenses in alignment with data access are important and something that needs to be addressed in future projects. Another theme for future discussion is how to handle research projects that consist of multiple data sets with different kinds of data access.

6 Appendix

6.1 Appendix 1

Dear all,

NSD and UKDS are currently working on developing a controlled vocabulary (CV) for data access through the Tools Pillar, Task 3 “Data access information upgrade” in Agenda 21-24.

The CV will be implemented in the CESSDA Data Catalogue (CDC) to make data access interoperable and provide users with information on the access restrictions of the studies listed.

We are building on the work done previously in Metadata Office (MDO), Task 2, Deliverable 3 “*Creating a CDC Publisher Names Vocabulary and investigating Data Access Interoperability*” (not yet published). The table below is an excerpt from the report (full final version of report attached).

Category	**Access credentials**	**Registration**	**Licence(s)** *(non-commercial use always allowed)*
Open	Free download	No	Open (e.g. Creative Commons) Commercial use allowed
	Free download and click-through terms of use	No	Open Commercial use allowed
	Free download and click-through terms of use	Yes	Open Commercial use allowed

Safeguarded / Accountable	Secure download and signed user contract/licence (registration or application)	Yes	Signed user contract/licence Commercial use allowed
	Secure download and signed user contract/licence (registration or application)	Yes	Signed user contract/licence Commercial use not allowed
	Secure download and signed user contract/licence (registration or application)	Yes	Signed user contract/licence Commercial use allowed Additional conditions
	Secure download and signed user contract/licence (registration or application)	Yes	Signed user contract/licence Commercial use not allowed Additional conditions
Restricted	Secure remote access and signed user contract/licence (registration or application)	Yes	Signed user contract/licence Commercial use not allowed Additional conditions
	Secure remote access and signed user contract/licence (registration or application)	Yes	Signed user contract/licence Commercial use allowed Additional conditions
	On-site use only and signed user contract/licence (registration or application)	Yes	Signed user contract/licence Commercial use not allowed Additional conditions
	On-site use only and signed user contract/licence (registration or application)	Yes	Signed user contract/licence Commercial use allowed Additional conditions

The plan is to use the categories in the data access CV for the CDC.

We would kindly ask you all to evaluate these categories in relation to your specific access conditions and answer the following questions:

1. Do these 3 categories: **“Open, Safeguarded/Accountable & Restricted”** cover your access conditions?
2. Is a 2 part access conditions CV consisting of: **“Open, Restricted** (includes **Safeguarded/Accountable**)**”** sufficient for your organization?

We believe a 2 part access condition structure would be easier to map to for Service Providers and others, it also opens up the possibility for sub-categories, either now or in the future.

If you could provide an answer within a month, 18.03.21 at the latest, we would be grateful.

If anyone has questions about this, please do not hesitate to get in contact with us.

All the best,
Morten Jakobsen, NSD - morten.jakobsen@nsd.no
Silje Storviken, NSD - silje.storviken@nsd.no
Sharon Bolton, UKDS - sharonb@essex.ac.uk

7. References

Akdeniz, Esra, Borschewski, Kerrin, Moilanen, Katja, Jääskeläinen, Taina, Jakobsen, Morten, Storviken, Silje, … Bolton, Sharon. (2021). *CMM CESSDA Metadata Model* (Version 2.0). Zenodo. DOI: https://doi.org/10.5281/zenodo.4672414

Bell, Darren. (2021, March 11). *CESSDA Data Catalogue - DDI Codebook Profile* (Version 1.0.4). Zenodo. DOI: http://doi.org/10.5281/zenodo.4580376

Bolton, Sharon, & Jääskeläinen, Taina. (2021). *Report on creating a CDC Publisher Names Vocabulary and investigating Data Access Interoperability* (Version 2.0). Zenodo. DOI: https://doi.org/10.5281/zenodo.4524429

Jakobsen, Morten. (2022). *Guidelines and updated specifications for CESSDA Service Provides to make data access interoperable* (1.0). Zenodo. DOI: https://doi.org/10.5281/zenodo.5554488

Woollard, M., L'Hours, H. and Beedham, H. (2016). *CESSDA Data Access Policy* (Version 1). Zenodo. DOI: http://doi.org/10.5281/zenodo.4054793