Check for updates

RESEARCH ARTICLE

REVISED ECRIN – CESSDA strategies for cross metadata mappings in selected areas between life sciences and social sciences and humanities

[version 2; peer review: 1 approved, 2 approved with reservations]

Christian Ohmann[1], Katja Moilanen[2], Mari Kleemola[2], Steve Canham[1], Maria Panagiotopoulou[1]

[1]European Clinical Research Infrastructure Network, 30 Bd Saint-Jacques, Paris, 75014, France
[2]Finnish Social Science Data Archive, Tampere, 33014, Finland

v2 **First published:** 20 Oct 2023, **3**:180
https://doi.org/10.12688/openreseurope.16284.1
Latest published: 11 Dec 2023, **3**:180
https://doi.org/10.12688/openreseurope.16284.2

Abstract

Background

The recent COVID-19 (Corona Virus Disease 2019) pandemic dramatically underlined the multi-faceted nature of health research, requiring input from basic biological sciences, pharmaceutical technologies, clinical research), social sciences and public health and social engineering. Systems that could work across different disciplines would therefore seem to be a useful idea to explore. In this study we investigated whether metadata schemas and vocabularies used for discovering scientific studies and resources in the social sciences and in clinical research are similar enough to allow information from different source disciplines to be easily retrieved and presented together.

Methods

As a first step a literature search was performed, exemplarily identifying studies and resources, in which data from social sciences have been usefully employed or integrated with that from clinical research and clinical trials. In a second step a comparison of metadata schemas and related resource catalogues in ECRIN (European Clinical Research Infrastructure Network) and CESSDA (Consortium of European Social Science Data Archives) was performed. The focus was on discovery metadata, here defined as the metadata elements used

Open Peer Review

Approval Status ✓ ? ?

	1	2	3
version 2 (revision) 11 Dec 2023			
version 1 20 Oct 2023	✓ view	? view	? view

1. **Adrian Dusa**, University of Bucharest, Bucharest, Romania
2. **Veerle Van den Eynden**, KU Leuven, Leuven, Belgium
3. **Ricarda Braukmann**, DANS-KNAW, Nijmegen, The Netherlands

Any reports and responses or comments on the article can be found at the end of the article.

to identify and locate scientific resources.

Results

A close view at the metadata schemas of CESSDA and ECRIN and the basic discovery metadata as well as a crosswalk between ECRIN and CESSDA metadata schemas have shown that there is considerable resemblance between them.

Conclusions

The resemblance could serve as a promising starting point to implement a common search mechanism for ECRIN and CESSDA metadata. In the paper four different options for how to proceed with implementation issues are presented.

Keywords
social sciences and humanities, clinical research, clinical trials, metadata, crosswalk, cross-domains, contextual metadata, COVID-19

This article is included in the Research on Research gateway.

This article is included in the Horizon Europe gateway.

Corresponding author: Christian Ohmann (christianohmann@outlook.de)

Author roles: Ohmann C: Conceptualization, Methodology, Writing – Original Draft Preparation, Writing – Review & Editing; **Moilanen K**: Conceptualization, Data Curation, Methodology, Resources, Validation, Writing – Review & Editing; **Kleemola M**: Conceptualization, Data Curation, Methodology, Resources, Writing – Review & Editing; **Canham S**: Conceptualization, Data Curation, Methodology, Resources, Writing – Review & Editing; **Panagiotopoulou M**: Data Curation, Methodology, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This project has received funding from the European Union's Horizon Europe research and innovation programme (101046203).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2023 Ohmann C *et al*. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Ohmann C, Moilanen K, Kleemola M *et al.* **ECRIN – CESSDA strategies for cross metadata mappings in selected areas between life sciences and social sciences and humanities [version 2; peer review: 1 approved, 2 approved with reservations]** Open Research Europe 2023, **3**:180 https://doi.org/10.12688/openreseurope.16284.2

First published: 20 Oct 2023, **3**:180 https://doi.org/10.12688/openreseurope.16284.1

REVISED **Amendments from Version 1**

In the revised manuscript the following changes have been implemented according to the suggestions of the reviewers:

- Better description of the objectives
- Extension of the introduction (search, EOSC included)
- Clarification of the role of CESSDA DC and ECRIN MDR
- Mentioning of other platforms and search engines
- Better specification of end users and potential benefits
- Information about access to documents added
- Better specification of the different options
- Description of existing mappings to other metadata schemas
- Correction of links

Any further responses from the reviewers can be found at the end of the article

Introduction

The recent COVID-19 (coronavirus disease of 2019) pandemic dramatically underlined the multi-faceted nature of health research, requiring input from basic biological sciences (e.g. viral characterisation and genomic sequencing), from pharmaceutical technologies (e.g. development and manufacture of mRNA (messenger ribonucleic acid) vaccines), from clinical research (e.g. observational and interventional studies of candidate treatments), from social sciences (e.g. understanding vaccine hesitancy, examining political responses to the pandemic), and from public health and social engineering (e.g. promoting 'lockdown' behaviours and mask wearing).

The pandemic and efforts to treat it not only generated huge amounts of scientific work, results, and data, it also did so in a very wide variety of scientific disciplines. This raises the question of whether, in at least a proportion of cases, work in one discipline could have been more efficiently planned and executed, or analysed more usefully, by also considering results, insights and / or data from studies in another discipline – if, for example, recruitment to a clinical trial could have been improved by explicitly addressing concerns identified by social science surveys.

In a recent paper, the marginal role of the social sciences in the COVID-19 pandemic was stated[1]. It was demonstrated that policymaking during the COVID-19 pandemic has been biomedicine-centric in that its evidential basis marginalised input from non-biomedical disciplines. It was argued in particular that the social sciences could contribute essential expertise and evidence to public health policy in times of biomedical emergencies and that we should thus strive for a tighter integration of the social sciences in future evidence-based policymaking.

We are not talking about multi-disciplinary teams working on the same problem here, but rather whether knowledge and use of research outputs from one or more 'external' research disciplines can usefully be applied by a mono-disciplinary team. Specifically, we consider whether clinical researchers could have benefited from data and insights from the social sciences, and vice versa. As discussed above, prima facie, such extra-disciplinary 'awareness' ought to be potentially useful, but anecdotal evidence suggests it rarely occurs. This may be partly because researchers simply do not give consideration or are not encouraged by funders to use insights from other disciplines, or because it is difficult for people to know how and where to look for relevant resources outside of their own core discipline – i.e., the barrier to discovery is too high – or because researchers, being less familiar with the methodologies of other disciplines, are less confident in interpreting their results.

As background work a literature search was performed by one of the authors (CO) to exemplarily identify studies, in which data from social sciences (SS) have been usefully employed or integrated with that from clinical research (CR) and clinical trials (CT). This work was seen as necessary to motivate the approach of mapping metadata schemas between ECRIN and CESSDA. The (unsystematic) literature search revealed the following examples:

- Using studies from SS to support and optimise planning of CT (e.g., developing educational material, addressing key concerns of study participants, involving key participant groups)[2,3]
- Improving CT conduct and recruitment by considering data from SS (e.g., engagement of participants by CT staff, addressing participants concerns and misconceptions, providing targeted information and feedback)[4]
- Applying and generalising results from CTs to the real-world involving SS data (e.g., taking into consideration existing health systems, cultural norms and values, and the effects of low resource contexts)[4].
- Developing predictive models by combining studies from SS and CR (e.g., combining survey data on attitudes towards vaccination with quantitative immunity data from an observational study to develop a more accurate epidemic spreading model)[5].
- Performing systematic / scoping reviews combining SS and CR studies (e.g., combining social, behavioural, pharmacologic, and non-pharmacologic interventions in combat a disease or health issue) to provide more complete knowledge of diagnosis, treatment, and outcome[6].
- Increasing effectiveness of policy measures derived from CR by incorporating SS data.
- Increasing knowledge about diagnosis, treatment, and outcome of COVID-19 by combining SS and CT data (e.g., incorporating social and medical determinants into modelling of COVID-19 and other infectious diseases)[7].
- Increasing knowledge about associations and causal pathways between social capital and income inequality, and COVID-19 and other infectious diseases[8].

- Capturing psychological and behavioural benefits and pitfalls of intensive diagnostic testing to detect and prevent COVID-19 positive cases[9].

A major use case for mapping resources from SS and CR lies in the provision of generalised evidence synthesis[5]. For the exploration of the interventions in COVID-19 different research types and designs are applied, spanning both experimental and observational approaches. Studies may be prospective or retrospective, use hypothesis testing or hypothesis generating, cover a control group or not, be randomised or not, have individual citizens/patients or clusters/regions as primary target, be population-based or not, be primary research or based on secondary use of available data, use real data, or be based on simulations, etc. All these characteristics are of major importance in assessing the evidence for the efficacy and efficiency of interventions.

Platforms and search engines providing access to biomedical and life sciences literature (e.g., PubMed, Web of Science, Google Scholar) are useful in identifying similar research types and designs via free text search and through defined metadata elements. Due to the complexity in identifying resources cross disciplines and due to limits with the specific metadata schemas applied, this process is resource-intensive, error prone and does not guarantee a systematic discovery of research types and designs across domains (e.g., social sciences, life sciences). Systematic reviews are slow to produce. Academics must work hard even to identify relevant clinical trials in publication databases: the studies are not clearly tagged, and CT researchers are rarely in connection with researchers doing systematic reviews. In the standard paradigm of evidence-based medicine, researchers collect evidence on a therapy from randomised controlled trials until it gets a green or red light. But in many situations, such trials are unethical, impractical, or unfeasible. Often, researchers have to pragmatically assess a range of different evidence — surveys, natural experiments, observational studies and trials — and stitch them together to give a picture of whether something is worthwhile. A prerequisite for taking research design and approach into consideration in evidence synthesis would be to make this information explicit and to link it to research projects (e.g., via tagging). The type and spectrum of examples identified, confirmed the usefulness of our approach to attempt a metadata data mapping between ECRIN MDR (metadata repository) and CESSDA DC (data catalogue).

Several specialist COVID-19 portals were established, or additional pages were added to existing portals, to allow easier discovery of and access to COVID related resources, study details, results, and data. Most of these, however, were mono-disciplinary (WHO (World Health Organization) COVID page, ECRIN (European clinical Research Infrastructure Network) COVID page, CESSDA (Consortium of European Social Science Data Archives) COVID page, CLARIN (Common Language Resources and Technology Infrastructure) COVID page, others). Even when a specialist portal has been designed as multi-disciplinary and offers the ability to search across domains (e.g., the EU COVID-19 portal), the detailed metadata has remained separately organised by discipline, and often accessed via different pages. The European Open Science Cloud (EOSC) Portal Catalogue & Marketplace, an entry point to multitude of research services and resources, offers data discovery, but no detailed metadata. A user can therefore use these portals to discover and access research in a variety of disciplines, but they have to be comfortable with the details of the search methodology in each case. Accessing the actual datasets require several (manual) steps and data access information and procedures differ by domain and/or by dataset. In general, the current situation puts the emphasis on the individual researcher, to have both the motivation and the skills to find resources and results in disciplines outside their own, rather than having retrieval systems presenting this information automatically.

Systems that could work across different disciplines would therefore seem to be a useful idea to explore, though the initial question is whether such systems would ever be a realistic proposition. In other words, are metadata schemas and vocabularies used for discovering scientific studies and resources – whether published papers, datasets, or related material such as research protocols – similar enough across disciplines to allow information from different source disciplines to be easily retrieved and presented together, as a single search result? More accurately, are the likely costs of making metadata schemas and their underlying ontologies more congruent with each other, and thus more easily interrogated together, justifiable by the potential benefits of doing so?

It is worth noting that this question extends beyond simply having more effective and efficient searches – important though that is. Existing metadata schemas and systems were developed to meet specific and domain-centric needs, partly starting from a common basis like DataCite or Dublin core (e.g., MDR for clinical research) or as a separate domain-specific development (e.g., DDI for social, behavioral, economic, and health sciences. Differences between them partly reflect the different assumptions, vocabularies, research paradigms, methodologies, and available resources in those domains. If, however, metadata schemas and controlled vocabularies (CVs) were made more inter-operable, for example by mapping against each other, or by some degree of alignment (so that, for example, the same controlled vocabularies were used for some entity types), it would become easier to discover, discuss and understand scientific work across disciplines in general. This in turn could lead to an increased exchange of paradigms, metaphors, data patterns and theories that could go far beyond the simple sharing of data. In other words, a convergence of metadata could be a key enabling mechanism for a range of future work that is not easily possible or even identifiable now. While divergent metadata systems reinforce the division of knowledge, convergent schemas could help increase the chances of integrating not just data but crucially also the interpretation of that data, and the generation of new insights across disciplines.

The objective of the study is to investigate whether the metadata schemas and vocabularies used for discovery of

scientific studies and related resources across two scientific disciplines (social science, clinical research) are similar enough to combine retrieval and presentation of these resources.

The work was performed in the context of the BY-COVID project (BeYond COVID). BY-COVID is funded by the European Union's Horizon Europe Research and Innovation Programme under grant agreement number 101046203. Part of the results have been presented at the European DDI (Data Documentation Initiative) User Conference (EDDI) in Paris (1 December 2022) and are published in ZENODO[10].

Methods

In a first step a comparison of metadata schemas and related resource catalogues in ECRIN and CESSDA was performed. The focus is very much on *discovery metadata*, here defined as the metadata elements used to identify and locate scientific resources, with sufficient additional material being available for each resource to determine if it would be useful to use or at least investigate further in any particular context. A scientific resource can be a research project (or study) or a data object belonging to a research project (study), e.g., publication. *Descriptive metadata*, usually applied to datasets, includes a more detailed description of each data point, including where necessary its definition, and as such is often more domain specific. It is essential to understand a dataset once it has been retrieved or accessed, but as the focus of this report is on discovery it is not considered further here.

As basis for the comparison and mapping of the metadata schemas under investigation a generic characterisation of discovery metadata was applied. To be fully effective, discovery metadata needs to include the following metadata items[11]:

a) A name (or names) of the resource

b) A description of the resource providing information about what the resource is, and which can therefore be used as the basis of text-based searches.

c) A persistent identifier, so that the resource can be referenced concisely and unambiguously from other systems.

d) A collection of keywords (or 'topics', or 'subjects') that indicate what the resource is about, or covers, allowing them to be used as the basis of a keyword-based search. Ideally, these keywords are constrained and / or mapped to specified controlled vocabularies (CVs), that can be used for both storing and querying the keyword data more efficiently.

e) The 'type' of resource, in terms of its role within the research process. Resources can be, for example, data, published papers, biological samples, result media files, study summary documents, analysis plans or web pages.

f) The 'source entities' described by the resource. For human related resources this means an indication of the source population, e.g. in gender, age and geographical distribution, as well as defining characteristics – e.g. a particular diagnosis, socio-economic group. Such data is useful in filtering resources within a search process.

g) The time the resource was constructed, and / or the time the underlying material (e.g. data) for the resource was assembled. Again, useful in filtering resources within a search process.

h) The methodologies used in creating the resource, in particular the nature and design of the generating study (if there is one), and the data collection methodologies used. An important part of understanding the 'context' of a resource, including the strengths and weakness of any resource and the conclusions generated from it.

i) The people and organisations that contributed to the resource's creation, or that contributed to the study or activity that generated the resource.

j) Country in which the resource was collected

k) An indication of how the resource can be accessed – e.g., if it is publicly available, its location (usually as a URL (Uniform Resource Locator)) and if not publicly available how access can be requested, for example whether specific prerequisites are required, and whether such access allows downloads or is *in-situ* only. This information is especially crucial, if the resource is involving sensitive data and GDPR (General Data Protection Regulation) needs to be obeyed in detail.

l) The physical nature of the resource – its file type, size etc – and any technical details (e.g. file formats) that will need to be taken into account by resource consumers.

Most of the systems or metadata schemas include only a part of these metadata items– they represent a summary of what, ideally, should be present rather than the current reality. Resources can also be distinguished according to how they are generated and organised. For example, there is a distinction between

a) Resources originating in studies, assembled to help answer the specific set of questions that the study is designed to examine, and

b) Resources – usually data or physical samples - collected or assembled outside of any specific study, assembled as a general resource for later work, the nature of which may not be known during the resource gathering phase.

Examples of digital objects that may be related to a specific study are study protocol, informed consent form, data management plan, statistical analysis plan, a data set, a report, or a publication. A problem arises when these digital objects are not stored together with the study data or are not directly linkable to a study via a PID.(personal identifier). In that case, searching for digital objects related to a study or vice versa identifying the study related to a digital object, may get difficult. This is of major relevance because the link between a study and its related digital objects preserves the context for any type of further analysis.

A comparison between ECRIN and CESSDA resources is complicated by a different interpretation of "study". The difference between the two systems might be summarised by saying that the MDR is *study-led*, and identifies digital objects linked to those studies, whereas the CDC is *resource-led*, and identifies (or sometimes even defines) a study as the activity that created the resource. This is not surprising given that the MDR starts with study registry data, whilst the CDC starts with data archives and their contents. There is no issue with finding all study linked resources, when they are packaged together in a catalogue entry or when the linkage can be provided by a PID.

In discovery terms, study originated resources may not be easily identified outside of the context of the study – they are 'Study X final dataset', or 'study Y protocol' – and often do not have persistent identifiers or even unique titles of their own (unless they are a published paper). Discovering such *study linked* resources therefore tends to use the characteristics of the generating study to identify relevant resources – in a search context the relevant studies are found first, and the associated resources can then be examined. This blurs the distinction between the data and metadata describing studies and those describing resources. In the social sciences, particularly datasets related to studies are archived with the social science data archives/repositories. These datasets receive PIDs (personal identifiers) and all the metadata needed. It is rarer to archive datasets that are formed outside a study (e.g., register or administrative data that are usually stored and disseminated by the agencies collecting them,[12]).

For resources collected without reference to a set of specific research questions – e.g., census data taken at regular intervals, or population level health surveys – the nature of the resource itself, usually a dataset, needs to be used in determining its relevance, e.g., its geographical and temporal distribution, and the methodology used to collect it. Such study independent resources are also much more likely to have their own unique names and descriptions, and often persistent identifiers as well. It is possible, however, to define the resource collection process itself as a 'study' (it depends how a 'scientific study' is defined), so that these resources can also be viewed as 'study-linked'. That makes it much easier to organise the metadata when some resources are study-linked, and some are not. In that situation, as discussed above, attributes such as 'period of collection' can be linked to the data, the 'study' or both, and the distinction between study and resource metadata again becomes blurred.

There are some obvious exceptions to the description of discovery metadata above – for example review papers are resources derived from other resources rather than a study or data collecting activity. Some disciplines organise data into aggregate knowledge systems (e.g. genomic or taxonomic databases) rather than retaining study results as discrete 'packets' and therefore use different metadata schemas. Historical and cultural factors can also make expectations about resource, especially data, sharing different in different scientific disciplines, which in turn influences the amount and quality of metadata available. For example, in clinical research making data available to others – despite extensive pressure from editors and funders – still only occurs in a minority of cases and, even when it is possible, it is often only advertised by a brief statement stating, 'reasonable requests will be considered by the investigator'. In social sciences, the first data archives were established in the early 1960s and were conceived as survey data archives although from the very beginning, the secondary analysis of the data was an important motive together with verification and long-term historical value[13]. The summary as provided above does, however, apply to much of the metadata as collected and organised by ECRIN and CESSDA, and is therefore provided as a background for the analysis.

Work in the study is performed according to the following steps:

a) Characterisation of the metadata schemas and initial assessment of similarity

b) Examples of retrieval

c) Detailed comparison of metadata between CESSDA (CDC), ECRIN (MDR) and fundamental discovery metadata items

d) Mapping of clinical research "study types"

Results

Characterisation of the metadata schemas and initial assessment of similarity

The ECRIN metadata schema was first developed in 2016 and has been revised several times since[14]. Almost all the resources in clinical research are study-linked, and there are often multiple resources per study. Therefore, the schema has to include substantial data about the generating studies as well as the linked resources, as described above.

In fact, the ECRIN schema consists of two inter-related schemas, one for studies and one for data objects (where 'data object' means any electronically available object, e.g., a file or web page). The *study schema* is based on the main data points used by ClinicalTrials.gov, the largest trial registry in the world, containing about 440,000 existing study entries (from a total of about 740,000). Those data points are themselves based around the core dataset required by the WHO and so – in broad terms – are also supported by the other 18 globally recognised trial registries. Trial registry data, particularly from Clinicaltrials.gov, represents the de facto standard data model for describing clinical research studies.

The MDR collects study metadata from the study registries covered by the system. So, fields in the MDR are available

when the metadata fields in the original registries are mandatory. This is, for example, the case for CliniocalTrials.gov, were most of the metadata elements are mandatory but not all (e.g., IPD data sharing statement). Even if fields are mandatory, quality checks are limited. Nevertheless, some fields are almost always present (e.g., registry ID, title).

The *data object schema* is based on the DataCite standard, extended to cover the needs of clinical researchers, specifically to provide additional data points covering:

- Location, ownership, and access arrangements for data objects, many of which would not be immediately or publicly available, and instead require an application process, usually to the study investigator or sponsor, for access to be granted.

- Links to the generating studies. Apart from journal articles most of the data objects generated by clinical research are closely linked to the study or studies that generated them, and are usually discovered using the study's name or identifiers.

The metadata schema is primarily used in ECRIN and related scientific communities. There is no similar comprehensive schema for clinical research metadata from any other source, with the possible exception of the schema used internally by Vivli (https://vivli.org/) and a few other specialist clinical data repositories. The WHO's ICTRP (International Clinical Trial Registry Platform) provides basic data on clinical research studies, using the WHO core dataset mentioned above. Nevertheless, the MDR remains the only portal, as far as we are aware, offering metadata on both clinical studies and clinical research outputs and other data objects.

CESSDA's metadata schema for the CESSDA data catalogue are available as enhanced DDI profiles (https://cmv.cessda.eu/documentation/profiles.html). and are based on the DDI Codebook versions DDI 2.5 and DDI 1.2.2 and DDI Lifecycle 3.2. CESSDA also offers tools for ensuring that the metadata is CDC valid.

The CESSDA Data catalogue profile is a subset of the CESSDA Metadata Model (CMM)[15], which itself is mostly a subset of DDI (https://ddialliance.org/Specification/DDI-Lifecycle/3.3/XMLSchema/FieldLevelDocumentation/ and https://ddialliance.org/Specification/DDI-Codebook/2.5/). Full DDI has more contextual metadata fields than CMM – for example DDI is able to describe the process of study development as a series of development activities, categorised using a controlled vocabulary. The CMM in turn has more contextual information than the CDC (CESSDA Data Catalogue) profile (e.g., 1.1.6 funding information, 1.1.10 study version, 1.3.4 universe, and 1.3.6 type of data source).

It could be argued that any general examination of social science metadata should start with DDI, or at least the CMM. In this case, we are focused on the subset of the DDI currently in use in the common CESSDA data catalogue, as represented by the CDC profile. Therefore, only the CDC profile is used for most of the comparison with the ECRIN clinical research metadata schema.

The Catalogues. ECRIN provides MDR, which is a database and associated web portal that brings together metadata data at a global level about a) clinical research – the studies – and b) the resources linked to them – the data objects and makes that metadata searchable in various ways (https://crmdr.org/; https://newmdr.ecrin.org/). The MDR and its metadata schema are both described further at (https://wiki.crmdr.org/index.php?title=Project_Overview). Currently, the MDR and portal are being rewritten and the material will be updated. The rewrite is also intended to add a public API (Application Programming Interface) to the MDR. The system currently has data on about 740,000 studies and 1.1. million objects.

CESSDA maintains a multilingual Data Catalogue CDC (https://datacatalogue.cessda.eu/), which contains metadata relating to more than 40,000 datasets held by CESSDA's Service Providers (SPs). The Service Providers are national data archives and repositories from CESSDA member countries. The CDC is designed as a 'one-stop shop' for searching and finding European social science data.

In both systems resources are linked to studies. In the MDR all studies are registered clinical research investigations – about 80% are interventional (i.e., clinical trials), and almost all the other are observational studies of various kinds. A very small proportion (several hundred in total) are listed as 'expanded access' (case studies of compassionate use of novel products) or 'registries', (collections of routine health data). In CESSDA, most resources are also linked to studies. Even when this is not strictly the case, the resource is treated as if it were linked to a study – for example entries labelled as censuses are linked to a 'study' that is the name and year of the census. This allows the same study-resource metadata schema to be applied to all entries in the system.

There are also similarities in the way in which the data is presented. In both systems a search using words or categories provides search engine result page (SERP) listing all suitable studies, paginated to 30 by default in the CDC, 10 in the MDR. In both cases users can navigate from any listed study in SERP (Search Engine Result Page) to a record/item level details page that provides a more complete description of the study and the linked data object(s). In the CDC, the user can navigate from the details page or the SERP to the service provider page, i.e., external to the CDC, that provides further details about the dataset and how to access the dataset itself. Accessing the dataset often requires the user to log into the information system of the CESSDA Service Provider, implying the need to set up an account for that specific system. In the MDR, the user can navigate, from the details page or the SERP, directly to publicly available resources. If a resource is under a controlled access, the user should normally be able to navigate to a page giving the information about how the resource can be accessed. A key public resource in the MDR is the study registry page (sometimes two or more registry pages are present)

that includes the public details of the study. In some ways, these pages are analogous to the SP page that can be accessed via the CESSDA Data Catalogue, although the detailed (meta)data available on each is very different.

A key difference between the two systems is that in the MDR a study can be linked to several resources – a study registry page as a minimum, often a results summary, papers published about the study, a study protocol etc. Of the 1.1 million objects in the system, about 750,000 are study registry pages, leaving about 350,000 that are some other types of resource – over 200,000 are associated journal papers. The number of linked data sets, however, is very small, about 1200, because data sharing in clinical research is a recent phenomenon and many researchers are still uncertain about exactly how and when it should be done. A recent survey by ECRIN of COVID-19 studies found that although about 15% of researchers in those studies made some form of commitment to data sharing, only a tiny proportion of that number had actually provided any data[16].

In contrast all entries in the CDC are linked to a single dataset, though that might consist of several files, and in some cases associated with information about publications e.g., journal papers where the dataset has been used. In some cases, the 'study' is to be defined by the dataset, to maintain that one-to-one link. For example, there are seven studies named 'Effects of the Trades Union Studies Project, 1977–78' followed by a specific sub-study name, and each has an associated dataset of the same name. In MDR terms, this would be one study with 7 datasets, each corresponding to sub-studies, in CDC terms it is 7 studies each with a single dataset.

Another difference between MDR and CDC is that CDC has multilingual metadata. The metadata is in the languages of the Service Provider, though in many cases it is provided in both English and in the local languages. Currently, about 75% of study descriptions are available in English in CDC while MDR contains metadata only in English.

The difference between the two systems might be summarised by saying that the MDR is *study-led*, and identifies data objects linked to those studies, whereas the CDC is *resource-led*, and identifies (or sometimes even defines) a study as the activity that created the resource. This is not surprising given that the MDR starts with study registry data, whilst the CDC starts with data archives and their contents.

A crosswalk between CESSDA Data Catalogue (CDC) Metadata Profile and ECRIN Metadata Schema has been performed and is published on ZENODO[17]. In another ZENODO document a mapping of metadata elements of the CESSDA Data Catalogue to the CESSDA Metadata Model (CMM), to OpenAire, to B2Find, to schema.org and to Dublin Core is described[18]. The ECRIN metadata schema is based on DataCite with some extensions (additional access details, de-identification & consent details) ana subset of the metadata elements of ClinicalTrials.gov (de facto standard). Mapping between CESSDA DC and OpenAIRE has already been performed and is under way for ECRIN MDR.

Examples of retrieval

Many of these advantages mentioned in the introduction were particularly identified in the context of COVID-19 but most of them would apply to any condition. For this reason, there does seem to be genuine added value in being aware of, and using, CR and SS data together. To investigate this question further, a series of quick searches were carried out – in the CDC using health-related terms, and in the MDR using social science terms (taken from the CESSDA topic list).

Table 1 shows the numbers of studies retrieved from the CDC using the selected health-related terms. As one would expect the more specific the term the fewer the studies found. The studies found would obviously need to be investigated further, but in all cases the numbers represent a non-trivial set of potential resources.

Table 2 shows similar results, showing the number of studies found with 'social science terms' in their title or listed keywords in the MDR. The results here are perhaps less impressive given the much greater size of the MDR, and

Table 1. Results of searching (in English only) for health-related terms in the CDC. (accessed 13 February 2023).

Term	CESSDA (from 26,362 entries in English)
health	8271
health service	3710
covid	781
heart	355
heart disease	104
nutrition	677
drug abuse	886
schizophrenia	21

Table 2. Results of searching for social science related terms in the MDR.

Term	MDR (from about 740,000 studies)
attitude	1603
minorities	531
trust	244
crime	85
quarantine	76
occupational health	38
social change	35

as above any preliminary results would need investigating further. Both searches were, however, very quick exercises using single terms rather than a cluster of synonyms, and a more rigorous search would almost certainly yield increased numbers, in both cases.

If we look specifically at COVID-19 in the MDR, we get more than 10000 studies with a search of "COVID-19" in the title. If this is restricted by filtering to "interventional trials", 7185 resources remain (22.02.2023). In the CESSDA data catalogue a search for "COVID-19" finds 750 studies from a total of 36758 studies in English. If we look further down to specific interventions, we can identify with a search for the term "lockdown" 95 resources in CESSDA and 76 studies in the MDR. This example shows that even for a specific intervention, quite a few biomedical and social studies can be identified.

Detailed comparison of Metadata between CESSDA (CDC), ECRIN (MDR) and fundamental discovery metadata items

Given the clear potential for using metadata from both systems, the next question is how easy it would be to do using an integrated search mechanism of some kind including the discovery metadata items defined in methods section. As already mentioned before, CESSDA Data Catalogue contains study-level metadata at present. Therefore, it is possible only to compare metadata items in the context of study but not in the context of a data object. One of us (SC) performed a 'cross walk' between the MDR metadata schema and the CDC metadata schema, later annotated and clarified by a colleague from CESSDA (KM). The crosswalk is openly available in ZENODO[17]. The walk was performed in both directions, i.e. if and how CDC metadata could be integrated into MDR, or MDR metadata integrated into the CDC. Then we checked if MDR and CDC metadata has all the discovery metadata items listed in methods.

The findings of the ECRIN (MDR) and CESSDA (CDC) metadata mapping to the discovery metadata specified in methods are shown in Table 3.

It is considered how each of the main elements of discovery metadata are treated in these two systems and therefore the potential interoperability between them. The main points to emerge are:

- The first three discovery metadata items a) study name(s), b) description and c) identifiers appear in

Table 3. ECRIN – CESSDA discovery metadata.

Reference	Discovery Metadata Item	MDR (ECRIN)	CDC (CESSDA)
a	A name (or names) of the resource	Yes (Usually public and scientific titles, often also acronyms)	Yes (Study title)
b	A description of the resource	Yes (Study description, outcomes, endpoints)	Yes (Abstract)
c	A persistent identifier	Partly (Identifiers, persistent in practice but not formally guaranteed / structured as PIDs)	Partly, some of the records include PID some only identifier (Study number / PID)
d	A collection of keywords (or 'topics', or 'subjects')	Yes, but vocabularies vary with source and keyword type	Yes (Keywords, Topics)
e	The 'type' of resource	Yes	Not explicitly
f	The 'source entities' described by the resource	Yes, (Participants specified by gender, age, inclusion- and exclusion criteria)	Yes (Analysis unit)
g	The time the resource was constructed	Partly (Start of the study is included - end dates may be be present but often approximate)	Yes (Data collection period)
h	The methodologies used in creating the resource	Yes (e.g., study phase, allocation type, intervention model, masking; but not always complete)	Yes (Time dimension, Sampling procedure, Data Collection mode)
i	The people and organisations that contributed to the resource's creation	Yes (Trial sponsor and study leads / contacts usually identified, also additional funders)	Yes (Creator)
j	Country in which the resource was collected	Yes (Country)	Yes (Country)
k	An indication of how the resource can be accessed	Yes, (Through data sharing statement but access for datasets not always explained	Yes (Terms of data access and link to the resource)
l	The physical nature of the resource	No (Only in a small minority of cases)	No

these two systems broadly similar, and it would be straightforward to (for example) use this metadata from one system inside the other or access it in both systems with a common interface. A major difference is that the datasets referenced by CESSDA usually have a PID (persistent identifier), often a DOI (digital object identifier), whereas the objects in the MDR, except journal papers, usually do not. An identifier should be unique, persistent and machine-actionable. For MDR this is achieved only partly, trial identifiers are persistent in practice but not formally guaranteed. Nevertheless, study identifiers are actionable and can connect to the original resource. But the URL may vary as the sites evolve (e.g., ClinicalTrials.gov, Dutch registry). The exception comes from the ISRCTN, where each study web page has also a DOI. For CESSDA DC, a dataset PID is mandatory and CESSDA data archives follow the CESSDA PID Policy ((https://doi.org/10.5281/zenodo.3611324). In addition, objects in the MDR (again the major exception is journal papers) often do not have a pre-specified name, and so one must be constructed for display and listing purposes, whereas the objects in the CDC all have a title, usually the same as or related to the study name.

- The discovery metadata item d) a collection of keywords occurs both in MDR and CDC. Nevertheless, the most significant difference between these two systems lies in the *controlled vocabularies* (CVs) used to generate keywords or topics. While the ways CVs are used are very similar in both systems, the CVs themselves are not. Within the MDR most topics are taken from MeSH (Medical Subject Headings), reflecting their use in two important US data sources – Clinicaltrials.gov and PubMed. If MeSH cannot be used the topics are usually as generated by the authors / study leads, though a few use MedDRA (Medical Dictionary for Regulatory Activities) or ICD (International Classification of Diseases) for 'condition under study'. In CDC there are both 'Topics', using the relatively high level CESSDA topic classification, and keywords using the ELSST (European Language Social Science Thesaurus), SPs' national thesauri or keywords given by authors.

  The ELSST includes about 3,300 concepts covering social sciences with the most of them available in each of 16 languages. In contrast, MeSH has about 30,000 heading terms covering all aspects of medicine, and another 300,000+ supplementary terms (although a lot of these are simply the names of chemicals and medicines) but it is almost entirely in English. MedDRA, for comparison, covering medical diagnosis and symptomology, has about 25,000 concepts ('preferred terms') and about 85,000 synonyms, and claims to be translated into 15 languages. Finally, the WHO's International Classification of Diseases (ICD), again designed to be used across languages, has a full system of about 35,000 terms, but by using only the 4 character 'stem' codes, this can be reduced to about 4,800, and that can be reduced further, to a few hundred terms, by considering only the 'blocks' of related codes.

- The type of the resource (discovery metadata e) is not part of the metadata in CDC but it is explicitly stated that CDC is Data Catalogue. In MDR, the type of resource is part of the metadata and included under "type" of data object attributes.

- The discovery metadata item f), the source entities, which is the description of the population appears both in MDR and CDC. Still, it is handled in different ways in these two systems, reflecting the different levels of specificity that are usually required.

- Study location, country in which the resource was collected (j), is present in both systems, but it is more explicit in the CDC

- The time the resource was constructed (g), the start and end times of data collection are captured more clearly in the CDC. In the MDR they often are estimates in the prospective study registration records.

- There are several metadata items related to discovery metadata h), the methodology. One of them, occurring both in CDC and MDR, is study design type: whether the resource is produced in an observational or interventional manner. In the CDC, almost all the studies are observational (mostly surveys) while in MDR most of the studies are interventional. This is reflected in both metadata schemas.

- With respect to i) study contributors (people and organisations), these two systems are mainly similar but have differences in detail. At the present, there is only Creator (author, principal investigator) in CDC. In MDR, for example, the sponsor, i.e., the organisation that has legal responsibility for a clinical trial, exists. In social sciences, the controller has legal responsibility on data protection but usually this information is not recorded in the public metadata.

- The crucial discovery metadata item for the users, an indication how the resource can be accessed (k) is available in both of the systems, MDR and CDC. They both have a metadata field for describing data access (CDC "Terms of data access" and MDR "Data sharing statement" and "location and access details"). In the MDR this covers the managing organisation, the access type (e.g., open, controlled), the access details and resources. Rights are also included in the schema but not yet practically applied. In addition, there is ongoing work on documenting access and use rights in EU-funded projects (e.g., BY-COVID, FAIR-(Findable, Accessible, Interoperable, and Re-usable Impact) and also in CESSDA.

- The physical nature of the resource (l) is missing both from the CDC and MDR.

Mapping of clinical research "study types" between CESSDA and ECRIN

In the paper of Sim *et al.*[19], a human study design ontology for clinical research is presented. Through iterative consultation

with statisticians and epidemiologists, the typology of study designs has been based on discriminating factors that define mutually exclusive and exhaustive study types In Table 4, the classes belong to study design in OCRE are presented (https://bioportal.bioontology.org/ontologies/OCRE/?p=classes&conceptid=root):

Considerations regarding documentation of "study type" in ECRIN and CESSDA

The considerations should help to clarify whether specific clinical research study design can be documented in CESSDA and as well in the MDR. This could be of value for generalised evidence-synthesis cross social sciences and clinical research.

Interventional studies. In the MDR, there is a primary distinction between interventional trial and observational study. This is reflected in the MDR filter variable "Type" (interventional, observational). In CESSDA, most of the studies are observational without intervention. Interventional studies are characterised by using DDI Alliance Controlled Vocabulary for Mode Of Collection and the term "Experiment " or even more specific "Field/Intervention experiment", "Laboratory experiment" and "Web-based experiment". In the term definition of " Field/Intervention experiment", interventions/clinical studies are explicitly mentioned.

In clinical research participants are selected according to defined inclusion and exclusion criteria, which are defined in the study protocol. This is usually a non-randomised procedure with no random sampling from a larger population, so it is not explicitly documented. Instead, for clinical research, the type of allocation of research participants to the interventions is of major importance. This can be documented in the MDR, for example, by "allocation type = randomised". In CESSDA, the selection of research participants would be documented from the beginning. There should be at least two different terms for the sampling procedure (using DDI Sampling Procedure CV) and also some free-text explanations. First there would be a term telling which kind of Non-probability (non-randomised) sampling is done e.g. "Non-probability: Purposive" and additional text e.g. "At the first selection phase of research participants, the research participants were chosen based on their suitable medical history." Secondly there would be a term telling that this group of participants selected in the first phase of sampling, is in the second sampling phase divided in two (or more) different groups using Probability (randomised) methods e.g. "Probability: Simple random" and adding again explanation text e.g. "In the second selection phase, the selected group of research participants were divided into control group and intervention group with simple random sampling."

Table 4. Classes belonging to "study design" in OCRE. (*from: https://bioportal.bioontology.org/ontologies/OCRE/?p=classes&conceptid=root).

Ontology of Clinical Research (OCRE*)

• Case-only study design
• Qualitative study design
  ◦ Case series study design
• Quantitative study design
  ◦ Interventional study design
  ▪ Crossover study design
  ▪ N-of-1 crossover study design
  ▪ Parallel group study design
  ▪ Single group study design
  ◦ Observational study design
  ▪ Case-control study design
  ▪ Case-crossover study design
  ▪ Cohort study design
  ▪ Natural history study design
  ▪ Cross-sectional study design

Another important aspect may be the observation unit of the study. In clinical research, most of the trials are with individuals but some trials work with clusters, e.g., cluster-randomised trials. This is not used in the MDR but in CESSDA, the analysis unit can be e.g. "Individual", "Event/Process/Activity", "Organization/Institution", "Group" or "Family" as the DDI Alliance Controlled Vocabulary for Analysis Unit is used.

Observational studies. In the MDR, observational studies can be identified by "type = observational". In CESSDA, this can be documented by using the DDI Alliance Controlled Vocabulary for Mode of Collection and the term "Observation" or with more detailed terms e.g. "Field observation", "Participant laboratory observation". Here specifically, development of condition or disease is mentioned. The narrower term "Laboratory observation" is relevant when comparing to clinical research.

In clinical research, there are several subtypes of observational studies. For clinical trials, of primary interest are **cohort studies**. In MDR, this is covered by "observational model = cohort". In CESSDA this can be documented via using DDI Alliance Controlled Vocabulary for Time method and the term "Longitudinal: Cohort/Event-based". Other types for clinical trials are **case-control**, **cross-section studies and case series**. In MDR, case-series are covered by "observational model = cases only". There is no counterpart for case-series in the vocabularies used by CESSDA. The term with closest match is "Non-probability: Purposive" from DDI Sampling Procedure vocabulary describing that the research participants were selected for the information they can provide on the research topic. To be able to identify case-series studies, there is a need to have some additional text explaining the study design/sampling. MDR has included "observational model = cross-sectional", which is described in CESSDA by using DDI Alliance Controlled Vocabulary for Time method and the term "Cross-section". With respect to case-control study, there is "observational model = case-control" in the MDR. Case-control cannot be directly described with vocabularies

used by CESSDA. The best option would be the same as with case-series, to use the term "Non-probability: Purposive" from DDI Sampling Procedure CV and to add again additional text explaining the study design/sampling.

Trials with simulation can be documented in CESSDA by using DDI Alliance Controlled Vocabulary for Mode Of Collection and the term "Simulation" but not in the MDR.

The possibility to document specific clinical research "study types" in CESSDA and ECRIN MDR is summarised in Table 5:

Discussion

Given the differences and similarities between the metadata schemas and systems of ECRIN and CESSDA, and the underlying studies and data objects, the question is what might be desirable and possible in discovery metadata integration, and what further work might be needed for clarifying these questions. Here in the discussion, we give tentative suggestions which will require further discussion. These suggestions are made largely with reference to the existing CDC profile, but the crosswalks demonstrated that the bulk of the potential interoperability between the clinical research and social science metadata exists within this profile. Using the CMM instead of the CDC profile would therefore add comparatively little to the current discussion.

The following suggestions are made:

- With respect to names (a), identifiers (c) and descriptions (b) the metadata schemas of CESSDA and ECRIN appear broadly similar, and it should be straightforward to map these metadata items from one schema to another. Interpretations and definitions of each of the metadata items would need to be clarified and checked but no substantial problems are foreseen.

Table 5. Documentation of specific clinical research "study types" in CESSDA and ECRIN.

Study type	CESSDA	ECRIN MDR
Interventional	+	+
- Randomised	+	+
Observational	+	+
- Cohort study	+	+
- Cross-section	+	+
- Case-series	-	+
- Case-control	-	+
- Randomised sample	+	-
Simulation studies	+	-

- The explicit type of the resource (e) is implicit in CDC and explicit in MDR. Mapping of the type of the resource should be effortless and machine-actionable.

- The source entities (f) are described both in MDR and CDC, but the level of detail is different. This might cause troubles, if we wished to harmonise this metadata. MDR has more detailed fields for describing the source entities but they are rarely filled in. CDC has only one field describing source entities: the unit of analysis (e.g. individual, media unit). This is difficult to handle for legacy data, but for future use one option could be to align this metadata discovery item between MDR and CDC.

- As part of the methodology (h), the type of the study is essential especially for the clinical research. At a high level, it ought to be possible to categorise study *types* into a single comprehensive system that is able to indicate

  a) whether or not a specific research question or hypothesis was being examined, as opposed to data collection on a routine, periodic basis, (be that in the general population or in a specified subgroup and / or setting such as a hospital). This actually can be a bit challenging because there is not a proper field for this kind of information in CDC profile (or DDI). Most of the studies/datasets available from CESSDA are collected because of the specific research question and only a very limited amount has another origin, e.g., census.

  b) whether the methodology used was interventional, observational or something else. In ECRIN this information is available, in CESSDA "mode of collection" could be a starting point for doing this classification.

  The difficulty here is that this would involve additional work in classifying the CESSDA studies, the distinction already being made within the MDR, and thus is unlikely unless the effort was thought to be worthwhile.

  At a lower level of 'study type', ECRIN may benefit from exploring the usage of the CESSDA categorisations for observational studies, referring to the mode of collection vocabulary (though the methodological details are often not available in the source data). CESSDA could utilize some of the interventional study classifications in the MDR schema (which are taken from ClinicalTrials.gov), but again it would depend on the work involved and whether the details were present in the source data. It should be taken into consideration that interventional studies in social sciences are very rare.

- In MDR, other metadata items concerning methodology are dependent on the type of the study, whether it is

observational or interventional. In CDC, both these study types have the same metadata items. For observational studies MDR uses vocabularies "Observational study model" (metadata item "Observational model") and "Time perspective". In CDC, the same kind of information can be found from the "Data Collection Mode", "Time Dimension", and "Analysis Unit". There is also "Sampling procedure" in CDC, but it does not have a counterpart in MDR.

For interventional studies MDR has "Phase", "Primary purpose", "Allocation", "Intervention model" and "Masking". In CDC, the "Sampling procedure" includes the same kind of information as "Allocation" in MDR, but other MDR metadata items are out of scope of the CDC. In CDC, also "Data collection mode", "Time dimension" and "Analysis unit" are used for interventional studies too.

The difference of these two disciplines can be noticed very clearly in the methodology. Some of the information can be mapped, but not all. The ideas of how to do the mapping is included in the appendix.

- Metadata items representing person and organisation (i) contributors could probably be mapped without too much difficulty. Persistent identifiers (PIDs) e.g., ROR (research organization registry) for the organisations and ORCID (Open Researcher and Contributor ID) for persons, could help to make the metadata more consistent and searchable. Persistent identifiers are useful but the proportion of source data containing them continues to be relatively low. This is at least partly due to the costs and effort needed to update a considerable amount of legacy (meta)data.

- Metadata items for study dates (g) could also be mapped but not the date when the resource was constructed because it is often missing from MDR. Unfortunately, the date when the data became available is the only one present in both schemas. Other dates than the year of publication / availability, would be difficult to map.

- Study location could probably be mapped at the level of the country (j) but not more detailed than that. A consistent set of identifiers for countries (e.g., from Geonames) could help with metadata consistency.

- Both MDR and CDC have keywords (d). To achieve a more consistent use of keywords, it might be possible to explore an approach where different CVs (controlled vocabularies) were used for different types of keywords. Thus,

    - For terms describing diagnosis or 'relevant medical condition', it might be possible to use ICD (International Classification of Diseases) 11 as the master CV and map other terms to that. This is currently being explored in the context of the MDR, where the plan is to provide an ICD browser as part of the search mechanism.

    - For terms covered by the ELSST (European Language Social Science Thesaurus), it might be considered as master CV, with other terms mapped to that. Again, ideally, a browser to navigate the thesaurus should be available as part of a search mechanism.

    Similar approaches could be used for other 'domains' of terms. Quite common approach is to do the mapping between the different CVs used but it is both difficult and expensive. Rather than trying to use a single CV that suited everyone in every case the longer-term solution is to push the source systems into using a common set of CVs. In the past, that has proven to be very difficult, and certainly ECRIN has no power to do that with respect to the source trial registries it uses. It is suggested that exploring the feasibility and practicality of 'CV convergence' would be a useful next step in examining how metadata can be made more inter-operable, even if the conclusions are that the resources required would need substantial amount of additional funding.

- As stated previously, the physical nature of the resource (l), is missing from the both (MDR and CDC).

If we wished to integrate the search mechanism of both systems, and assumed that the metadata schemas themselves could be made sufficiently interoperable, there are at least four implementation possibilities:

a) Include CDC search result metadata within the MDR.

b) Include MDR search result metadata within the CDC.

c) Provide a new, single search interface 'over the top' of both systems, with the results integrated to some extent, but with each returned study linking back to details in the source system.

d) Include the metadata of both systems into an overarching system (e.g., COVID-19 Data Portal, OpenAIRE, EOSC Portal Catalogue & Marketplace)

First three approaches would require an effective search API to be available for both the MDR and the CDC – there is no suggestion that the metadata should or could be integrated at the level of the source data. Instead, it would be the search results that would be integrated. The difference between these three approaches is therefore mainly in the location and nature of the search mechanism and the location of the listed results.

The CDC has recently developed a search API (https://api.tech.cessda.eu/) and https://api.tech.cessda.eu/#/DataSets/findRecordsByQuery). This API supports better interaction with external collaborators and makes it possible to search over all the records in the catalogue. CESSDA also has an OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) endpoint that exposes the metadata in DublinCore,

DataCite and DDI 2.5 Codebook formats. Currently, around half of the resources is covered by OAI-PMH. CDC is part of EOSC Portal & Marketplace.

The CDC metadata generally appears more consistent than the MDR metadata. CDC is currently focused on datasets, while MDR has studies with only other types of linked objects. Therefore, including MDR metadata into CDC search engine result page could confuse users. Conversely, integrating the CDC search result metadata within a MDR SERP (search engine results page) might be easier. MDR already contains a wide range of different data objects attached to studies. Links in such a case could lead back to the CDC details page, as now in the CDC, or possibly back to the SP page, so that access to the data was more immediate. Accessing the dataset would, in many cases, require logging into the CESSDA service provider's information system.

The option of having a different search portal altogether, in a sense overlaying both systems, is probably the most attractive. It would allow search results to be returned either as a single integrated list or as two lists within the same page. The important thing is that both lists would be automatically visible to the user. The major difficulty is that it would require some developer input, as well as separate and additional hosting arrangements. The hosting should not be too difficult, but developer input is often difficult to resource within academic research infrastructures, so setting up such a portal might be more difficult than simply extending one of the existing systems.

As already mentioned, the intended audience of the paper are professionals that have prior knowledge about metadata and research data. The end users envisaged for the common search system for which the metadata comparison was made, are researchers involved in secondary use health-research related data. Two major scenarios are foreseen:

a) Searching complementary study designs in CESSDA DC and ECRIN MDR and perform cross-analysis

b) Searching similar study designs in CESSDA DC and ECRIN MDR for systematic reviews and scoping reviews

In summary, the discovery of studies with complementary designs in CESSDA DC and ECRIN MDR and a cross-analysis could support the generalisability of results from randomised clinical trials to a larger population. In systematic reviews and in scoping reviews usually studies with similar study designs are combined. Especially for COVID-19, there is a variety of interventions from social sciences and clinical research, which are used to combat the disease, or a health issue related to it. This covers social, behavioural, pharmacologic, and non-pharmacological interventions. Therefore, it would make sense to use the application to simultaneously identify similar study types in the CESSDA DC and ECRIN MDR[20].

Dependent on the type of end-user, both systems have strengths and weaknesses. The researcher as end user might be interested in the individual data. Due to the different construction of ECRIN MDR and CESSDA CDC, there is a major difference in the ability to assess the data underlying a study. For the MDR direct links to the individual datasets are rarely available, currently this is the case in less than 1% of the studies included to the MDR. According to data sharing statements, that are now required by the ICJME journals upon study registration, around 5 to 16% indicate their willingness to make IPD available for secondary use. So, the MDR does not allow the access to IPD in most of the cases, but it provides information about the data access process. Instead, in the CDC the end user can find the datasets and often datasets are directly available for re-use.

Other end-users, such as policymakers, might be more interested in research output of studies (e.g., publications). Here the ECRIN MDR has advantages because PubMed is integrated in the system and publications related to a study are stored as digital objects. In CESSDA CDC this is only possible if publications are included in the metadata description of a resource or easily identifiable in the references linked to a resource.

For the fourth option a distinction should be made between an overarching system that is specific to a certain subdomain or topic, like the COVID-19 portal compared to an overarching system that is domain-agnostic like the OpenAIRE graph. The COVID-19 Data Portal facilitates data sharing and analysis by bringing together and continuously updating information about relevant COVID-19 datasets and tools. The Portal allows users to search across all different data types presented on the Portal. Data sources include from molecular biology and clinical data to social sciences, empowering researchers' discoveries of relevant data objects from potentially unfamiliar resources. OpenAIRE harvests metadata not only about publication but about research products in general, covering datasets, software, and other research products. In the OpenAIRE Research graph, research products are related to projects, data sources, funders, organisations, beneficiaries, etc. It would be beneficial to add metadata from domain-specific repositories, such as the metadata repository (MDR) of ECRIN. CESSDA DC is already part of the OpenAIRE Research graph.

With respect to this fourth option, some work has already been done. CESSDA Data Catalogue is one of the COVID-19 Data Portal's sources. The BY-COVID project has implemented a scalable multi-tiered indexing system for the COVID-19 data portal. At the lowest level, Tier 3, users can simply find relevant resources, for example a database, through the portal. At Tier 2, they can find the individual records of the database, for example specific surveys referenced in a database. At Tier 1, comprehensive and harmonised metadata allow relevant records to be found across multiple resources, for example

all studies from different resources involving a specific virus variant. CESSDA DC is part of the "social sciences and humanities" resources in the COVID-19 data portal and has been indexed according to tier 2. The MDR currently does not provide public APIs or endpoints, but this is under development and will come soon. Meanwhile an API endpoint on the MDR has been set up, which provides data for the COVID-19 portal in the required format. It is planned to provide a RESTful API and a GraphQL API.

Development of an API for the MDR will allow interoperability with OpenAIRE. There is also an ongoing mapping of the MDR with the OpenAIRE RG (Research Graph). CESSDA DC is already harvested by OpenAIRE. Therefore, the progress of the integration with OpenAIRE is expected soon. However, not all metadata items from CESSDA DC and ECRIN MDR have a direct counterpart in OpenAIRE and have to be handled as descriptive text. Unfortunately, this will result in some loss of structured information for discovery. The OpenAIRE Graph includes metadata and links between scientific products (e.g. literature, datasets, software, and "other research outputs"), organizations, funders, funding streams, projects, communities, and (provenance) data sources. As such, the OpenAIRE graph already includes some of the basic entities to model contextual metadata (e.g., funder, project, organisation). Unfortunately, the research process, covering "research projects" and "research activities" (e.g., studies) is not modelled explicitly. If mapped, many of the MDR fields would be amalgamated into a textual description, which greatly reduces the value of the data.

The full potential of a combined system for a researcher as user could only be elucidated if the discovery metadata of both systems could be assessed by a single search interface based upon the mapped discovery metadata (option c). If the search must be performed separately in both systems, even if starting from an overarching system (e.g. COVID-19 data portal – option d), this is suboptimal because the metadata mapping is not considered. A link to an overarching agnostic system, such as OpenAIRE,(option d) would certainly be beneficial but not optimal because of some loss of structured information for discovery. Similarly, option a) and b) do not offer real advantages, so for a researcher which wants to identify studies and data sources, option c) would certainly be the best.

Conclusions

A close view at the metadata schemas of CESSDA and ECRIN and the basic discovery metadata items presented in methods, as well as a crosswalk between ECRIN and CESSDA metadata schemas have shown that there is considerable resemblance between them. The resemblance serves as a promising starting point to implement a common search mechanism for ECRIN and CESSDA metadata. In this paper, we present four different options for how to proceed with the implementation issue. Either way, further discussions and the provision of use cases demonstrating the benefit of the approach to be selected will be needed. Regardless of the selected implementation, there could be support for discoverability of research types and designs. This approach has a huge potential to improve predictive models, to increase knowledge and to provide better evidence synthesis with data coming from different domains, in this case from life sciences and social sciences.

List of abbreviations

API =	Application Programming Interface
BY – COVID =	BeYond-COVID
CDC =	CESSDA Data Catalogue
CESSDA =	Consortium of European Social Science Data Archives
CLARIN =	Common Language Resources and Technology Infrastructure
CMM =	CESSDA Metadata Model
COVID-19 =	Corona Virus Disease 2019
CR =	Clinical research
CT =	Clinical trial
CV =	Controlled vocabulary
DC =	Data catalogue
DDI =	Data Documentation Initiative
ECRIN =	European Clinical Research Infrastructure Network
EDDI =	European DDI
ELSST =	European Language Social Science Thesaurus
EOSC =	European Open Science Cloud
FAIR =	Findable, Accessible, Interoperable, and Reusable
GDPR =	General Data Protection Regulation
ICD =	International Classification of Diseases
ICTRP =	International Clinical Trial Registry Platform
MDR =	Metadata repository
MedDRA =	Medical Dictionary for Regulatory Activities
MeSH =	Medical Subject Headings
mRNA =	Messenger ribonucleic acid
OAI – MPH =	Open Archives Initiatives – Protocol for Metadata Harvesting
OCRE =	Ontology of Clinical Research
ORCID =	Open Researcher and Contributor ID
PID =	Personal Identifier
REST =	Representational state transfer
ROR =	Research Organization Registry
SERP =	Search Engine Result Page
SS =	Social Sciences
URL =	Uniform Resource Locator
WHO =	World Health Organisation
XML =	Extensible Markup Language

Data availability

The crosswalk underlying the analysis between CESSDA Data Catalogue (CDC) DDI2.5 Metadata Profile (https://cmv.cessda.eu/profiles/cdc/ddi-2.5/1.0.4/profile.html) and ECRIN Metadata Schema for Clinical Research Data Objects Version 6.0 (August 2021)[14] has been registered with the DOI 10.5281/zenodo.8129621 and with the licence Creative Commons Attribution 4.0 International in ZENODO[17].

References

1. Lohse S, Canali S: **Follow *the* science? On the marginal role of the social sciences in the COVID-19 pandemic.** *Eur J Philos Sci.* 2021; **11**(4): 99.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**
2. Tolley EE, Severy LJ: **Integrating behavioral and social science research into microbicide clinical trials: challenges and opportunities.** *Am J Public Health.* 2006; **96**(1): 79–83.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**
3. Valente PK, Wu Y, Cohen YZ, *et al.*: **Behavioral and social science research to support development of educational materials for clinical trials of broadly neutralizing antibodies for HIV treatment and prevention.** *Clin Trials.* 2021; **18**(1): 17–27.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**
4. Lees S, Sariola S, Schmidt-Sane M, *et al.*: **Key social science priorities for long-term COVID-19 response.** *BMJ Glob Health.* 2021; **6**(7): e006741.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**
5. de Meijere G, Valdano E, Castellano C, *et al.*: **Attitudes towards booster, testing and isolation, and their impact on COVID-19 response in winter 2022/2023 in France, Belgium, and Italy.** *medRxiv.* 2022.12.30.22283726.
Publisher Full Text
6. Pearson H: **How COVID broke the evidence pipeline. The pandemic stress-tested the way the world produces evidence — and revealed all the flaws.** *Nature - NEWS FEATURE.* 2021.
Publisher Full Text
7. Mabry PL, Olster DH, Morgan GD, *et al.*: **Interdisciplinarity and Systems Science to Improve Population Health: A View from the NIH Office of Behavioral and Social Sciences Research.** *Am J Prev Med.* 2008; **35**(2 Suppl): S211–S224.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**
8. Lindström M: **A commentary on "The trouble with trust: Time-series analysis of social capital, income inequality, and COVID-19 deaths in 84 countries".** *Soc Sci Med.* 2020; **263**: 113386.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**
9. Van de Casteele M, Waterschoot J, Anthierens S, *et al.*: **Saliva testing among teachers during the COVID-19 pandemic: Effects on health concerns, well-being, and precautionary behavior.** *Soc Sci Med.* 2022; **311**: 115295.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**
10. Ohmann C, Canham S, Panagiotopoulou M, *et al.*: **Bridging scientific domains with metadata: CESSDA and ECRIN**. EDDI2022: The 14th Annual European DDI User Conference (EDDI2022), Sciences Po, Paris, France, *Zenodo.* 2022.
Publisher Full Text
11. Riley J: **Understanding metadata. What is metadata, and what is it for?** Primer publication of the American National Information Standards Organization (NISO), 2017.
Reference Source
12. Mortelmans D, Pasteels I: **Using register data in the social sciences**. In: *Sage Research Methods Cases Part 1*. SAGE Publications, Ltd, 2014.
Publisher Full Text
13. Doorn P, Tjalsma H: **Introduction: archiving research data.** *Arch Sci.* 2007; **7**(1): 1–20.
Publisher Full Text
14. Canham S: **ECRIN Metadata Schema for Clinical Research Data Objects Version 6.0 (August 2021) (6.0).** *Zenodo.* 2021.
http://www.doi.org/10.5281/zenodo.5554961
15. Akdeniz E, Borschewski K, Moilanen K, *et al.*: **CMM CESSDA Metadata Model (2.0).** *Zenodo.* 2021.
Publisher Full Text
16. Ohmann C, Panagiotopoulou M, Canham S, *et al.*: **An assessment of the informative value of data sharing statements in clinical trial registries.** *BMC Medical Research Methodology.* (under resubmission).
17. Moilanen K, Canham S, Kleemola M, *et al.*: **Crosswalk between CESSDA Data Catalogue (CDC) Metadata Profile and ECRIN Metadata Schema.** (Version 1) [Data set]. Zenodo. 2023.
http://www.doi.org/10.5281/zenodo.8129621
18. Akdeniz E, Jakobsen M, Storviken S: **Mapping CDC to OpenAIRE, B2find, schema.org and Dublin Core (1.2).** *Zenodo.* 2021.
Publisher Full Text
19. Sim I, Carini S, Tu S, *et al.*: **The human studies database project: federating human studies design data using the ontology of clinical research.** *Summit Transl Bioinform.* 2010; **2010**: 51–5.
PubMed Abstract | **Free Full Text**
20. Ohmann C, Panagiotopoulou M, Moilanen K, *et al.*: **EOSC-Future Test Science Project "META-COVID": Final report for CESSDA – ECRIN use case (Version 1).** *Zenodo.* 2023.
Publisher Full Text

Open Peer Review

Current Peer Review Status: ✔ ? ?

Version 1

Reviewer Report 13 November 2023

https://doi.org/10.21956/openreseurope.17579.r35761

© **2023 Braukmann R.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Ricarda Braukmann
[1] DANS-KNAW, Nijmegen, The Netherlands
[2] DANS-KNAW, Nijmegen, The Netherlands

Summary of the article

The paper "ECRIN – CESSDA strategies for cross metadata mappings in selected areas between life sciences and social sciences and humanities" by Ohmann and colleagues describes the comparison of metadata standards and vocabularies used in the Metadata Repository of ECRIN (MDR) and the Data Catalogue of CESSDA (CDC). With the background of the COVID-19 pandemic, the authors pose the question whether researchers in the clinical domain could have benefited from insights and data from the social sciences, and vice versa. A role is seen for retrieval systems that operate across different disciplines and the authors want to assess the resemblance between the metadata and vocabularies used for resource discovery in two exemplary domain-specific systems.

The authors performed a literature search and they compared the metadata from the MDR and CDC. They assessed the search system, as well as creating a crosswalk between the two metadata schema's and the authors systematically compare how a selection of 12 discovery metadata elements are implemented in each system.

The authors observe that with respect to names, identifiers, and descriptions the metadata schemas of CESSDA and ECRIN appear broadly similar, whereas some information related to the methodologies in the disciplines is harder to map to each other.
In the discussion section, proposals are made for how existing elements and Controlled Vocabularies (CVs) could be used to align the metadata to improve the interoperability. The authors finish their assessment by presenting four possibilities how an integrated search mechanism could be realized (i.e. including metadata in each other's system, creating a new system or including both in an existing overarching system) and addressing the possibilities and challenges of each option.

Overall the authors conclude that there is a resemblance in the discovery metadata used in the ECRIN MDR and the CDC which is promising when assessing the possibilities to have a common search mechanism for clinical and social sciences data, yet the four options for implementation which are presented would need to be further assessed.

General impression

The paper assesses the resemblance of metadata used in two domains which could benefit from more integration. I find this paper therefore a very useful exercise and a good initiative to assess how metadata from different existing domain-specific systems could be combined. The selection of clinical data and social sciences data seems very relevant when referring to the COVID-pandemic in the introduction, although this exercise could be replicated with other example catalogues from different domains as well.

The authors assessed the used metadata profiles in a lot of detail and created a crosswalk as well as making concrete suggestions on how existing fields and CVs could be used by the other system to enhance the interoperability.

I have some recommendations and requests for clarification which I will briefly introduce here and outline below in more detail.
Firstly, one aspect that I was missing in the paper that I believe the paper would benefit from is including more information about the **user perspective** and considering what end user is envisaged in the common search system for which the metadata comparison is made. It would be great if the authors would elaborate on this end user perspective when comparing the two systems as well as when evaluating the four options that are presented for implementing a common search system in the discussion.

Next to elaborating on the user perspective, I have some concerns about the **comparability of the two systems** and how the search results that are presented can be interpreted. I would like to propose some adjustments that could provide clarification.
I also recommend to present the **literature study** in a different way.

I have some questions regarding the **assessment of the discovery metadata elements**, in particular the PIDs, and the Data Access Terms.

In the discussion, I would recommend a more extensive **elaboration on the fourth option** that is presented.

Lastly, I a few **minor comments** and requests for clarification.

One technical thing I want to note is that I was not able to view the **Appendix** that is referred to in the paper. It was unclear to me where I can download this document.

Detailed Recommendations

Methods - Literature review

The literature search is presented in the abstract and introduction as the first core element of the

methodological part of the study. The literature study is, however, not described in much detail in terms of the methodology that was used to execute the study and the Results section lists examples of studies that were found rather than presenting a comprehensive list of results. With how the literature search is presented in the abstract and introduction, I would have expected more information about the search engine and search terms used, the selection criteria, the number of results, and the selection for further assessment. It appears that the literature study was conducted mainly to strengthen the existing assumption that it would be useful to compare the metadata from clinical and social science portals. The authors themselves state: "This work was seen as necessary to motivate the approach of mapping metadata schemas between ECRIN and CESSDA."

In my view, the focus of the paper would be improved if the literature study was included as part of the introduction rather than being described in the methods and results. It seems that this was in fact part of the background research that was done to justify the main goal of the paper - namely assessing the resemblance between the metadata used in the two domain portals.

Methods - Comparability of the Catalogues - end user perspective

While I do understand the reasoning behind the selection of the ECRIN MDR and the CDC for comparison, I do wonder a bit about their comparability. MDR is a registry of different resources associated with a given study, including data but also other aspects, whereas the CDC is a data catalogue.

I wonder if both systems are used for the same purpose by an end user. In the CDC an end user can find datasets - often datasets that are available for reuse - but no other information is available, e.g. on social science publications or preregistered studies. The MDR contains more studies to browse through, but - as far as I can evaluate - many of the records to not contain a link to the underlying data. This distinction could be stressed more in the paper and should be considered when discussing the merits of combining information from the two systems. Considering different end users that such a combined system would address would be useful. I can imagine for instance that a policy officer is interested in published results whereas a researcher might be interested in the underlying data. Discussing who a combined system would be targeted towards seems a useful contextualization.

Results - Comparability of the Catalogues

Due to the difference in the catalogues (see above), I also find it difficult to evaluate the comparison that is presented in the results in terms of the search results (i.e. presented in Table 1 and Table 2). A resource discovered in the CDC will always be a dataset, while a resource in the MDR will be study information that may or may not have a link to a dataset. I think it would be useful to present the results including the resource type for the MDR. I would be interested to see how many results contain links to datasets as this would make it better comparable to the CDC results presented.

Results - Assessment of the Discovery Metadata

PIDs

I wonder about the evaluation of Persistent Identifiers in the MDR. An important aspect for the FAIRness of the metadata would be the inclusion of an identifier that is not only persistent and unique, but is also actionable so that one can connect with the original resource. It would be great if this aspect of PIDs could be included in the evaluation.

Data Access terms

I find this aspect of the discovery metadata difficult to compare between the catalogues as the records in the CDC refer to datasets and include licences and data access statements, whereas the records in the MDR refer to studies where – in my understanding – not always a link to a dataset is included. I also noticed when looking at some records in the MDR that although a data sharing statement is available it is often not filled out ("NaN"). It would be great if the authors could elaborate on this aspect a bit more and also provide insights in whether certain fields in the Discovery Metadata are obligatory and typically provided in the two systems.

For evaluating the usability of the metadata and for an assessment by the end user, I suppose this would be an important aspect.
In addition to the access terms, I am missing an evaluation about the licence and whether a resource is openly available. This seems to me to be an important aspect of the access information for a particular resource.

Discussion - Generic versus topic-specific catalogue

I wonder if the fourth option that is presented should not make a distinction between an overarching system that is specific to a certain subdomain or topic, like the COVID-19 portal compared to an overarching system that is domain-agnostic like the OpenAIRE graph.
In the discussion of these different options, I also miss the user perspective. The options are outlined in terms of technical feasibility, but I would like to hear the authors perspective about the envisoned end user and in what way that perspective influences the choice of system in which the combined metadata would be presented?

I also wonder if these four options are mutually exclusive or if different options would be considered depending on the end user in mind.

With respect to the fourth option of integrating both into an existing system, I was wondering whether CDC and ECRIN metadata are already mapped to a common standard. It would be great if the authors could include information about existing crosswalks between the CDC, ECRIN metadata and other standards. I would particularly be interested in existing mappings to OpenAIRE and schema.org or DCAT. Including these in the supplementary materials or providing a persistent link to an existing mapping would be a useful addition to the paper.

As OpenAIRE is explicitly mentioned I would also be interested to know how the assessment of the two systems relates to what is available in OpenAIRE with respect to discovery metadata elements that were assessed. The authors mention that not all items can be mapped but it would be nice if more specific information was included, for instance in the mapping that is mentioned to be in the Appendix. I do have to note that I was not able to find the Appendix in the pdf of the paper or online so I was not able to evaluate what information is already included.

Other comments

- The article says to focus on the life sciences and social sciences *and humanities,* yet the data that is assessed is from the clinical research field and from the social sciences. Examples provided in the introduction seem to be mainly social sciences research (psychological behavioral data) and I miss the humanities perspective. Therefore I would recommend to rephrase the title and the reference to Social Sciences and Humanities to only include Social Sciences. I understand that SSH is often used as a term to describe the wider domain, but in my opinion referring to social sciences covers better what was assessed in this paper. Similarly, one might want to consider rephrasing life sciences to clinical research, yet as this is not my area of expertise I am not sure how this would be perceived by life science researchers.

- It would be useful to include a list of abbreviations at the end of the document.

- In table 3, it would be helpful if the letters the authors use for the discovery metadata in the Methods section is also added in the table as the letters are used in the text and this would be helpful for referencing to Table 3.

- I found a few hyperlinks that seem to link to other websites than described in the text:
    - "With the possible exception of the schema used internally by Vivli and a few other specialist clinical data repositories" yet the link is to a CESSDA vocabulary collection https://vocabularies.cessda.eu/vocabulary/ModeOfCollection

    - "is available as an enhanced DDI profile. CESSDA also offers tools for ensuring that the metadata is CDC valid" The link is https://cmv.cessda.eu/profiles/cdc/ddi-2.5/1.0.4/profile.xm where I believe an "l" is missing. https://cmv.cessda.eu/profiles/cdc/ddi-2.5/1.0.4/profile.xml does work however I believe a newer version of the CDC profile is available. If this is done explicitly as this profile was used for comparison it would be good to state this. I would find it even more useful if a link to an overview page is given rather than directly to the XML.

    - "The CDC is part of the EOSC (European Open Science Cloud) marketplace and has recently developed a search API (https://api.tech.cessda.eu/" although the text says something different, the hyperlink goes to https://bioportal.bioontology.org/ontologies/OCRE/

    - When referring to the crosswalk, I would mention the DOI (https://doi.org/10.5281/zenodo.8129621) instead of the Zenodo url (https://zenodo.org/record/8129621).
- As a lover of PIDs, I would love to encourage Steve Canham, Katja Moilanen and Maria Panagiotopoulou to include an ORCID ID that can be linked to their authorship on this paper.

- The text mentions that the article was created as part of the BY-COVID project and I was curious if the grant information is also included in the articles metadata.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and does the work have academic merit?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Not applicable

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: My organisation is involved in the BY-COVID project, although I personally am not financed by the project or actively involved in the work performed in the project. My organisation is a Service Provider of CESSDA and my organisation was involved in the SSHOC and FAIRsFAIR project in which some of the authors have been involved as well. I have in particular collaborated with Mari Kleemola in the past three years on different occasions within CESSDA and for project deliverables.

Reviewer Expertise: Social Sciences, Data Archiving Services

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 06 Dec 2023

Christian Ohmann

Reviewer: more information about the **user perspective** and considering what end user is envisaged in the common search system for which the metadata comparison is made - Discussion on the audience and user perspective has been added to the discussion. The methods and results of the literature search are moved to the introduction. Strengths and weaknesses of both systems (ECRIN MDR, CESSDA DC) are elaborated (see discussion). Reviewer: I wonder about the evaluation of Persistent Identifiers in the MDR. We agree with the comment from the reviewer and added a statement:- The following statement is added to the text: **The MDR collects study metadata from the study registries covered by the system. So, fields in the MDR are available when the metadata fields in the original registries are mandatory. This is, for example, the case for CliniocalTrials.gov, were most of the metadata elements are mandatory but not all (e.g., IPD data sharing statement). Even if fields are mandatory, quality checks are limited. Nevertheless, some fields are almost always present (e.g., registry ID, title. Information for the MDR has been added, overing the managing organisation, the access type (e.g., open, controlled), the access details and resources.** Option 4 has been specified with respect to domain-specific and generic systems. The user perspective in general was added in the

discussion. With respect to the interaction between user and the choice of the system, a statement was added: In the revised version we have considered and compared the 4 options for the researcher as end user. From the viewpoint of the authors the 4 options are mutually exclusive. At the end of the results section, information about crosswalks between metadata schemas is included. This covers also discussion of mapping to OpenAIRE. Throughout the paper “social sciences” was used, except for the title (social scineces & humanities) List of abbreviations added to the text. Letters have been added to table 3. The links have been corrected.

Competing Interests: No competing interests were disclosed.

Reviewer Report 13 November 2023

https://doi.org/10.21956/openreseurope.17579.r35765

© 2023 Van den Eynden V. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Veerle Van den Eynden
[1] KU Leuven, Leuven, Flanders, Belgium
[2] KU Leuven, Leuven, Flanders, Belgium

This paper provides a valuable comparison of metadata schemas and discovery metadata characteristics for two data catalogues to investigate the feasibility of cross-disciplinary exploration of data resources to facilitate cross-disciplinary research. The paper takes the case of two data catalogues, CESSDA for the social sciences & humanities, and ECRIN for the health sciences and clinical trials, to examine comparability and divergence, and to formulate recommendations for potential future cohesion.

Whilst the paper presents clear explanations and evidence, there are suggestions for improvement:

1. I miss a description of the aims and objectives of this piece of work.

2. It would be good to say something about the role of EOSC to facilitate such discovery of resources across disciplines in the introduction;

3. It may be worth pointing out that whilst disciplinary metadata schemas may develop to meet discipline-specific needs, they often have historically developed starting from a common basis like DataCite and Dublin Core. This is worth mentioning.

4. For the literature search (first step in methods), indicate which databases / portals were searched and which search terms were used (for reproducibility of the research)

5. Provide explanations of abbreviations when first used, e.g. GDPR, API.

6. I'm not sure the statements on page 4-5 that study-originated resources may not be easily identified outside the context of the study and may lack PIDs and unique titles (whilst for study independent resources it is the opposite) are accurate. If you have evidence for this, then provide it, or reference it. The CESSDA DC has over 15000 datasets with 'study' in the title. Which datasets resulting from studies are not easy to find?

7. Organise the results in such way that it is clear what corresponds with steps 1, 2 etc. described in the methods.

8. Is the statement on page 6 that it is currently difficult to discover similar research types & designs across domains correct? What about portals like Web of Science and Google Scholar for cross-domain discovery?

9. For DDI specify that the CDC profile is based on DDI Codebook.

10. Should URL links in the text be replaced by references, e.g. for DDI, the Zenodo link for the metadata mapping, etc.?

11. When describing the results of Table 3, the text refers to topics a to l. These letters are not used in the table.

12. For suggestion 3 (source entities have different level of detail), what would / could be a solution?

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and does the work have academic merit?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Not applicable

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: research data management; data sharing: open science; researcher practices: researcher attitudes & motic-vations

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 06 Dec 2023

Christian Ohmann

The objective was included at the end of the introduction. EOSC mentioned in the introduction . In the text a statement was added about generic and domain-specific metadata. The literature research on studies combining data from social sciences and clinical research was targeted at identifying examples and therefore not systematic. Search started from reviews about this topic and was extended to the ECRIN MDR and the CESSDA Data Catalogue. According to another reviewer the search was moved to the introduction as background.. Abbreviations are spelled out in the text when first used. Reviewer: I'm not sure the statements on page 4-5 that study-originated resources may not be easily identified outside the context of the study ....:The reviewer is correct, the statement in the text maybe misleading and needs precision to be understandable. For that reason, the text has been improved. The headings of the result section are listed in the methods section. A statement concerning bibliographic databases is added to the text. The first versions of DDI Profiles for CDC are based on the DDI Codebook versions DDI 2.5 and DDI 1.2.2 has been added to the text. URL links to publications (e.g., ZENODO) are cited as references. URLs to other resources are included as hyperlinks. New column added to table 3. Reviewer: For suggestion 3 (source entities have different level of detail), what would / could be a solution? - This is difficult to handle for legacy data, but for future use one option could be to align this metadata discovery item between MDR and CDC.

Competing Interests: No competing interests were disclosed.

Reviewer Report 13 November 2023

https://doi.org/10.21956/openreseurope.17579.r35760

© 2023 Dusa A. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Adrian Dusa
[1] University of Bucharest, Bucharest, Romania
[2] University of Bucharest, Bucharest, Romania

This manuscript is a welcome and much awaited piece that attempts to promote the usage of secondary data through enhancing and stimulate the usage of metadata standard.

It does so by making a highly interesting comparison between two of the most important research infrastructures in Europe: ECRIN and CESSDA.

There are potentially unlimited possibilities of cross-domain research between the medical / clinical research and the social sciences, with a massive and deep impact over the whole European region.

I very much welcome this type of publications, and hope this will be circulated as widely as possible.

The manuscript is well written, using a clear and accessible language for any interested reader. I did not spot any obvious weakness in the text, and will therefore recommend it for publication.

If anything, I would only have a couple of suggestions, if not for the current form of the paper then for (perhaps) future similar ones.

First, sometimes the text seems to over employ the acronyms of the specific institutions, standards and procedures. They are of course presented earlier in the paper, but if the (especially) not specialised reader sometimes forgets what the acronym stands for, it is necessary to travel back to the original (first) mention of acronym. I would perhaps make an Annex with a list of all presented acronyms.

Second and most important, the manuscript is obviously written by professional staff with proven and long standing expertise in these fields. The readers, however, are not always experts. They can be part of different scientific areas, or even regular people from the general public. The text should, therefore, take this into consideration and perhaps be made readable for a more wider audience. I had no problem whatsoever understanding the text, but on the other hand I do have expertise in this field.

Third and perhaps more general, the context of this paper seems, yet again, addressed by professionals to professionals in the field. Science, however, has experienced an unprecedented opening, not only due to the open research but more because of involving the regular citizen, thus expanding far beyond the traditional research institution centred approach.

The tools mentioned in this paper are so important that (I believe) deserve a much wider impact.

Otherwise, the paper is well worthy of publication and I highly recommend it.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and does the work have academic merit?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Not applicable

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Social research methodology, statistics, data archiving, metadata standards

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 06 Dec 2023

Christian Ohmann

List of abbreviations included in the text. The intended audience is not the wider public but professionals that have prior knowledge about metadata and research data. Making the text easily understandable for a wider audience (e.g. citizens) would be a huge task and would require major restructuring and revision of the text. Nevertheless, the idea to make the paper more readable by a larger audience is a good one and will be followed up by the authors. Involving regular citizen: This is a good and relevant comment and should be kept in mind for future work as popularising scientific work and science communication are important tasks (see also the previous comment)

Competing Interests: No competing interests were disclosed.