Young Information Scientist 7 (2022), 13–27
DOI:10.25365/yis-2022-7-2

Quality of metadata describing research data and the influence of repository characteristics

Dorothea Strecker*

Objective — This article captures the status quo of metadata for research data, and identifies factors at the repository level that influence metadata quality.
Methods — Based on a joint analysis of DataCite metadata records and re3data repository descriptions, this paper evaluates the quality of metadata records describing research data and analyzes differences in metadata quality between repositories of different types and between repositories with or without formal certification to determine if these factors correlate with high metadata quality.
Results — Of individual metadata elements, mandatory elements are used most frequently, followed by recommended and optional elements. More than half of all metadata elements are used in less than 5 % of metadata records. With the exception of related identifiers, persistent identifiers are rarely used. The average descriptions has 487.3 characters. On average, 18.7 elements are used in metadata records, which corresponds to 24.7 % of the elements available. The homogeneity of metadata records varies considerably between repositories, on average, 50.9 % of metadata records use the same common set of metadata elements. The analysis revealed statistically significant differences across repositories of varying type and certification status in the use of individual metadata elements, the comprehensiveness of descriptions, and the completeness of metadata records.
Conclusion — This paper presents a first systematic analysis of metadata quality for research data and the influence of repository characteristics on metadata quality. It discusses difficulties of using a generic metadata schema for describing diverse research data. The results show that some repositories appear to have established successful metadata practices and workflows, but some metadata elements remain underused. There is evidence of repository type and certification status affecting metadata quality, but more research is needed to identify specific factors.

Keywords — research data, research data repository, metadata quality

Qualität von Metadaten zur Beschreibung von Forschungsdaten und Einflussfaktoren auf der Ebene von Repositorien

Zielsetzung — In diesem Beitrag werden der Status quo von Metadaten für Forschungsdaten erfasst und Faktoren auf der Ebene der Repositorien ermittelt, die die Qualität der Metadaten beeinflussen.
Forschungsmethoden — Auf der Grundlage einer gemeinsamen Auswertung von DataCite-Metadatensätzen und re3data-Einträgen wird die Qualität von Metadatensätzen bewertet, die Forschungsdaten beschreiben, und Unterschiede in der Metadatenqualität zwischen Repositorien verschiedener Typen und zwischen Repositorien mit oder ohne formale Zertifizierung analysiert, um festzustellen, ob diese Faktoren mit einer hohen Metadatenqualität korrelieren.
Ergebnisse — Von den einzelnen Metadatenelementen werden obligatorische Elemente am häufigsten verwendet, gefolgt von empfohlenen und optionalen Elementen. Mehr als die Hälfte aller Metadatenelemente wird in weniger als 5 % der Metadatensätze verwendet. Mit Ausnahme von related Identifiern werden persistente Identifier nur selten verwendet. Beschreibungen umfassen durchschnittlich 487,3 Zeichen. Im Durchschnitt werden 18,7 Elemente pro Metadatensatz verwendet, was 24,7 % aller verfügbaren Elemente entspricht. Die Homogenität der Metadatensätze variiert beträchtlich zwischen den Repositorien, im Durch-

* Dorothea Strecker | Berlin School of Library and Information Science, Humboldt-Universität zu Berlin | dorothea.strecker@hu-berlin.de | ORCiD: 0000-0002-9754-3807

Young Information Scientist (YIS) wird vom Verein zur Förderung der Informationswissenschaft (VFI), Wien, herausgegeben. Alle Beiträge unterliegen einem Peer Review. ISSN: 2518-6892

schnitt verwenden 50,9 % der Metadatensätze dieselbe gemeinsame Menge von Metadatenelementen. Die Analyse ergab statistisch signifikante Unterschiede zwischen Repositorien verschiedener Art und verschiedenem Zertifizierungsstatus hinsichtlich der Verwendung einzelner Metadatenelemente, des Umfangs der Beschreibungen und der Vollständigkeit der Metadatensätze.
Schlussfolgerungen — In diesem Beitrag wird eine erste systematische Analyse der Qualität von Metadaten für Forschungsdaten und des Einflusses von Repositoriumseigenschaften auf die Metadatenqualität vorgestellt. Schwierigkeiten, die aus der Verwendung eines generischen Metadatenschemas für die Beschreibung diverser Forschungsdaten resultieren, werden diskutiert. Die Ergebnisse zeigen, dass einige Repositorien offenbar erfolgreiche Metadatenpraktiken und Workflows etabliert haben, dass aber einige Metadatenelemente nach wie vor nicht ausreichend genutzt werden. Es gibt Hinweise darauf, dass die Art des Repositoriums und der Zertifizierungsstatus die Qualität der Metadaten beeinflussen können, aber es sind weitere Untersuchungen erforderlich, um spezifische Einflussfaktoren zu ermitteln.

Schlagwörter — Forschungsdaten, Forschungsdatenrepositorien, Metadatenqualität

Diesem Beitrag liegt die folgende Abschlussarbeit zugrunde / This article is based on the following thesis:
Strecker, Dorothea: *Quantitative assessment of metadata collections of research data repositories*. Masterarbeit (M.A.), Humboldt-Universität zu Berlin, 2021. DOI: 10.18452/22916

1. Introduction

In the context of the Open Science movement, research data are increasingly regarded as distinct and valuable research outputs. By their nature, research data can be packaged and moved across disciplinary or other boundaries, for example to scrutinize results or to be used in an entirely new setting (Leonelli 2020). Currently, a broad political and cultural shift towards making data sharing the norm is unfolding. This is demonstrated by the proliferation of data sharing policies and the FAIR Principles, a set of principles intended to make research data more useful for machines and humans (Wilkinson et al. 2016). However, the success of data sharing and the fulfillment of the FAIR Principles depend on capturing extensive and appropriate metadata (Musen 2022).

While ambitious policies are put in place and the number of infrastructures supporting data sharing increases, little is known about the status quo and results of data curation activities (York et al. 2018). One area where understanding is significantly lacking is metadata creation and maintenance, especially differences across repositories specializing on data sharing (Gregg et al. 2019).

This paper describes results of a quantitative analysis that combines the evaluation of metadata quality with information on research data repositories. First, the quality of metadata records describing research data is evaluated. The second step analyses differences in metadata quality between repositories of different types and between repositories with or without formal certification to determine if these factors correlate with metadata quality.

2. Literature review

2.1. Metadata for research data

Metadata encapsulate information describing information-bearing objects, often in structured form, following specifications of a metadata schema (Zeng and Qin 2022, p. 11). There are a number of metadata schemas available for describing research data specifically that vary in their degree of specialization and disciplinary focus.[1] Metadata schemas define individual metadata elements that fulfill different functions. For example, metadata elements can generally serve finding and understanding a

1 The Research Data Alliance maintains a list of metadata schemas used for describing research data: https://rdamsc.bath.ac.uk/scope

resource (*descriptive* metadata), decoding and rendering files (*technical* metadata), long-term management of a resource (*preservation* metadata), defining intellectual property rights (*rights* metadata), and specifying relationships with other resources (*structural* metadata) (Riley 2017, p. 6). A metadata record is the sum of metadata elements describing an information object based on a metadata schema (Pomerantz 2015).

In the context of research data, metadata creation is an essential curation activity that is often undertaken jointly by data providers and repository staff (Lee and Stvilia 2017). Data providers have detailed knowledge of datasets, whereas data curators can assist in standardizing metadata descriptions (Rodrigues et al. 2019).

2.2. Metadata quality

Metadata sustain core functionalities of digital repositories, and it is often via metadata that users first come in contact with research data. Therefore, ensuring high metadata quality is an important factor for successfully operating a research data repository.

Metadata quality refers to "the degree to which the metadata in question perform the core bibliographic functions of discovery, use, provenance, currency, authentication, and administration." (Park 2009, p. 215) When evaluating metadata quality, the conformity to a set of requirements is determined. Bruce and Hillmann (2004) list core criteria used for assessing metadata quality: completeness, accuracy, provenance, conformance to expectations, logical consistency and coherence, timeliness, and accessibility (ibid., 5 ff). Metadata quality criteria can either be applied to individual metadata elements, to metadata records, or to entire metadata collections (Zeng and Qin 2022).

Metadata quality for research data has been discussed for several years (Rousidis et al. 2014). However, research on evaluating metadata quality for research data specifically is still limited. Existing literature is often driven by the intention to establish metrics measuring the impact of datasets (Cousijn, Feeney, et al. 2019; Robinson-Garcia et al. 2017), and improving data retrieval (Chapman et al. 2019).

Metadata completeness specifically is a challenge for research data service providers. For example, metadata availability at the data discovery service Google Dataset Search is very limited beyond the two mandatory metadata elements *title* and *description*. In an analysis of their metadata collection, service operators reported that license information is available for 34.80 % of all datasets, and only 11 % are assigned a DOI (Benjelloun et al. 2020). Currently, the PID service provider DataCite is one of the most comprehensive sources for metadata on research data. Still, metadata records are not complete. For example, less than half of the metadata records provided subject information about the resource (Robinson-Garcia et al. 2017). Both services acknowledges that metadata completeness presents a major challenge and are taking steps to improve it.

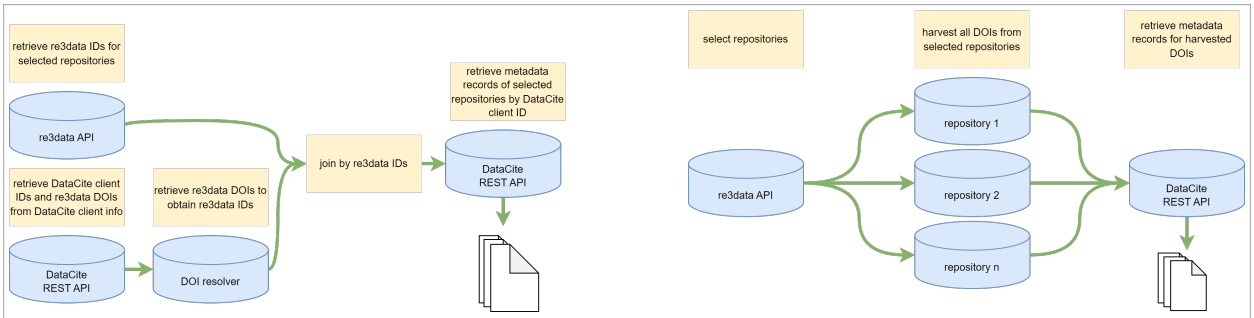
2.3. Research data repositories

Although research data repositories share similar objectives, they vary in specific characteristics and the practices they adopt to support data sharing in their communities (Kindling et al. 2017).

Research shows that metadata practices at repositories are heterogenous. For example, an analysis of data deposit guidelines of 20 repositories revealed significant differences in metadata requirements within and across disciplines (Kim et al. 2019). Metadata schemas implemented by generalist repositories differ in the number, use of controlled vocabularies and obligation levels of metadata elements related to aspects of data sharing (Assante et al. 2016). Institutional repositories also use different metadata elements to describe research data (Manninen 2018).

Repositories contribute essential resources and services to data curation, but little is known about results of their efforts (York et al. 2018). Even within one research area, metadata quality at discipline-specific repositories varies (Gonçalves and Musen 2019). It is also unclear whether high-quality metadata contributes to increased data use - at the generalist research data repository Figshare, no correlation was found between the quality of metadata records and views or downloads of datasets (Quarati and Raffaghelli 2020).

In summary, little is currently known about factors contributing to varying metadata quality across repositories. This paper adds to the current body of research by examining the influence of repository characteristics on metadata quality.



(a) Matching by client IDs

(b) Matching via OAI-PMH

Figure 1: Matching procedures

3. Methodology

3.1. Data sources and sampling

The analysis is based on combining two data sources: DataCite for metadata records describing research data, and re3data for repository information. Therefore, the sampling is centered around repositories that are represented in both data sources, and the ability to successfully match metadata records in DataCite to a repository entry in re3data. Academic social networks and repositories that only publish text publications were excluded from the analysis.

Two procedures were used to match the two data sources. The first used re3data identifiers listed in DataCite client information (see figure 1a), the second was based on harvesting dataset DOIs from repository APIs and retrieving DOI metadata from DataCite, if available (see figure 1b). The first method added 41 repositories to the sample, the second an additional 6.

3.2. Data collection and processing

Metadata records of all 47 repositories in the sample were harvested from the DataCite OAI-PMH interface between August 3rd and August 10th, 2020. An upper time limit was defined to restrict results to metadata records registered with DataCite up to and including July 31st, 2020. At the time of data collection, the most current version of the DataCite Metadata Schema was version 4.3 (DataCite Metadata Working Group 2019). Therefore, data collection, processing and analysis are based on this version (see table 1 for an overview of the DataCite Metadata Schema, Version 4.3).

Harvested metadata records were processed to extract information relevant to the analysis: the occurrence of metadata elements and the combined length of descriptions for each metadata record. In addition, the content of the element *resourceTypeGeneral* was extracted. The element was used to remove metadata records describing text publications from the sample, as suggested by Robinson-Garcia et al. (2017).

Since the publication of version 2.0, the DataCite Metadata Schema has been adapted several times, and new elements were added. These revisions result in varying sizes of element sets across schema versions. For a more precise estimate of the total number of metadata elements available, the *datestamp* provided by DataCite for each metadata record was used to determine the latest schema version available when that record was registered. Release dates of the schema versions were retrieved from the DataCite website (*DataCite Metadata Schema* n.d.) (see table 2).

Information on repository type and certification was retrieved via the re3data API on August 3rd, 2020. Repository type (*r3d:type*) categorizes repositories based on the extent of services offered. There are three values (*disciplinary*, *institutional*, *other*), and a repository can be assigned more than one value. Certification status is derived from *r3d:certificate* and indicates whether a repository has acquired any type of formal certification.

element name	obligation level	metadata type	# of child elements and attributes
identifier	mandatory	descriptive	1
creator	mandatory	descriptive	10
title	mandatory	descriptive	1
publisher	mandatory	descriptive	0
publicationYear	mandatory	descriptive	0
resourceType	mandatory	technical	1
subject	recommended	descriptive	3
contributor	recommended	descriptive	11
date	recommended	descriptive	2
relatedIdentifier	recommended	structural	6
description	recommended	descriptive	1
geoLocation	recommended	descriptive	16
language	optional	descriptive	0
alternateIdentifier	optional	structural	1
size	optional	technical	0
format	optional	technical	0
version	optional	structural	0
rights	optional	rights	4
fundingReference	optional	descriptive	7

Table 1: Description of the main elements in the DataCite Metadata Schema, Version 4.3

schema version	release date	size of the element set
4.3	2019-08-16	83
4.2	2019-03-20	76
4.1	2017-10-23	72
4.0	2016-09-19	66
3.1	2014-10-16	44
3.0	2013-07-24	42
2.2	2011-07-01	31
2.1	2011-03-28	31
2.0	2011-01-24	31

Table 2: Version history of the DataCite Metadata Schema

3.3. Metrics and statistical tests

In order to provide a first systematic overview of metadata quality for research data and study the influence of repository characteristics, this paper focuses on the metadata quality dimension *completeness* and includes aspects of *logical consistency and coherence* (Bruce and Hillmann 2004). The indicators used in this paper describe metadata quality at various levels (individual metadata elements, metadata records, metadata collections):

Completeness

use of individual metadata elements the number of metadata records using a metadata element

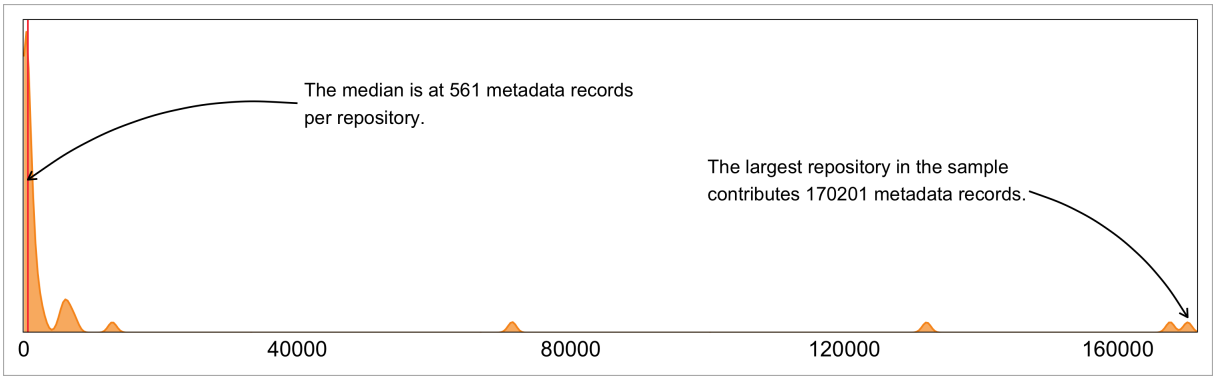


Figure 2: Frequency distribution of the number of metadata records per repository (n = 47 repositories)

use of persistent identifiers the number of metadata records using a persistent identifier
comprehensiveness of descriptions the combined character length of descriptions in a metadata record
completeness of metadata records the number of metadata elements used in a metadata record in relation to the number of metadata elements available

Logical consistency and coherence

homogeneity of metadata collections the number of metadata records using the most common combination of metadata elements within a metadata collection

The second part of the study is based on a between-group analysis that evaluates differences in metadata quality between repositories with certain characteristics. The statistical tests used require independent groups; therefore, repositories with overlapping characteristics were excluded from the analysis (this concerns 6 repositories with overlapping type).

After conducting an Anderson-Darling normality test, the assumption of a normal distribution for all dependent variables was rejected. As a result, non-parametric methods were chosen over parametric methods for investigating differences between groups. Since there are three levels (disciplinary, institutional and other) of the independent variable *repository type*, the Kruskal-Wallis test was selected (effect sizes are reported in η^2). In the case of the independent variable *certification status*, there are two levels (true and false), therefore, the Mann-Whitney U-test was used (effect sizes are reported in r).

4. Findings

4.1. Sample characteristics

In total, 606091 metadata records of 47 repositories were included in the analysis. Repositories in the sample are listed in Appendix A. Figure 2 shows a frequency distribution of the number of metadata records per repository in the sample. Repositories contain between 11 and 170201 metadata records, with a median of 561 and an average of 12895 records. Overall, the metadata sample is skewed towards smaller metadata collections of less than 600 records.

4.2. Use of individual metadata elements

The use of individual metadata elements indicates how frequently a metadata element is used in the sample of metadata records. Figure 3 shows the use of schema elements present in more than 5 % of all metadata records by obligation level (see table 1 for an overview of the main elements in the DataCite Metadata Schema). Of the 8 mandatory schema elements, only *resourceType* is not available for all metadata records, because it was not mandatory in previous schema versions. In comparison, recommended elements are used less frequently. Of the 35 recommended elements, 8 are present in more than

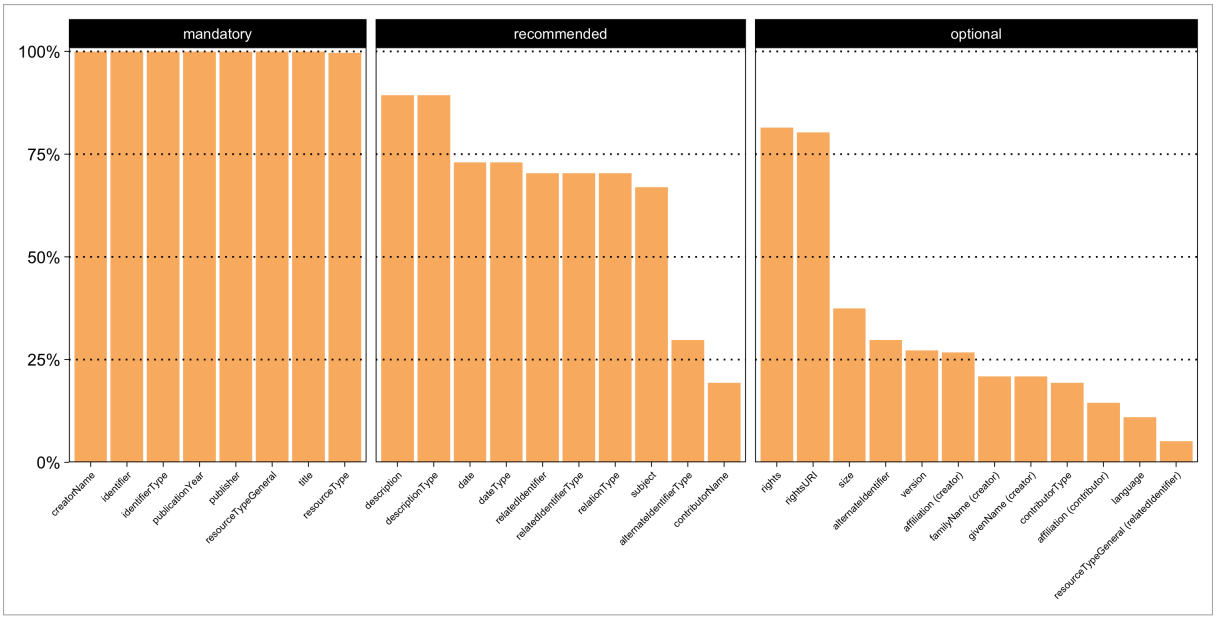


Figure 3: Use of schema elements present in more than 5 % of all metadata records by obligation level (n = 606091 metadata records)

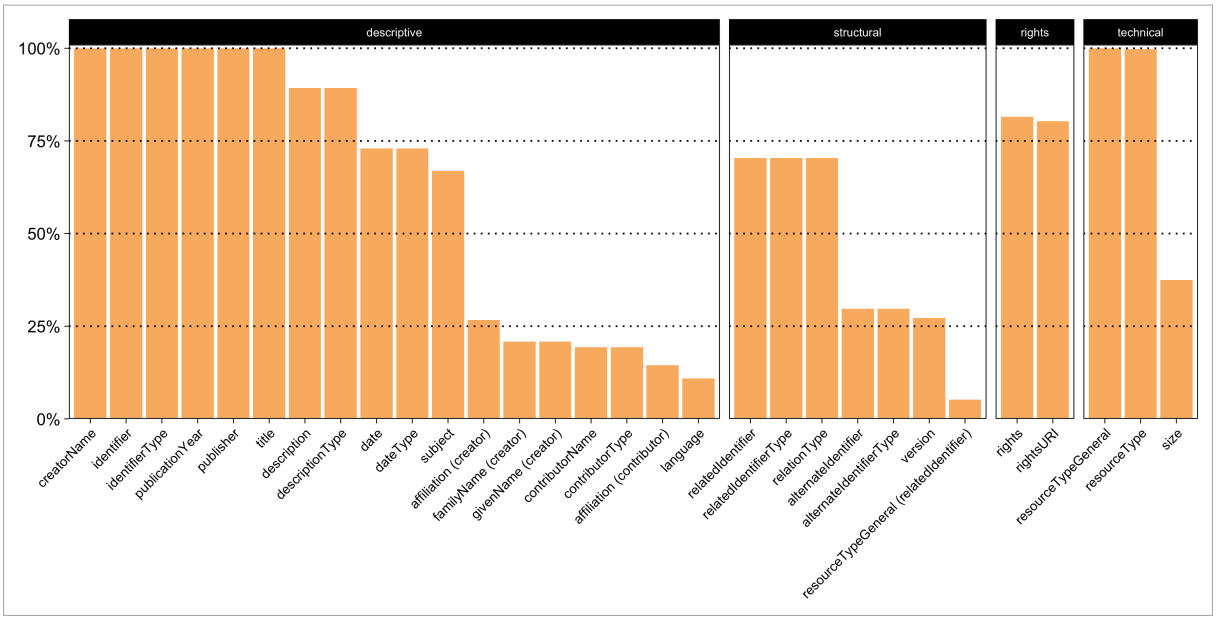


Figure 4: Use of schema elements present in more than 5 % of all metadata records by metadata type (n = 606091 metadata records)

50 % of metadata records. The elements *description* and *descriptionType* are most common (89.3 %), followed by *date* and *dateType* (73.0 %), *relatedIdentifier*, *relatedIdentifierType* and *relationType* (70.4 %), and *subject* (66.9 %). Overall, optional elements in the DataCite metadata schema are used least frequently. The most common optional elements are *rights* and *rightsURI*, which are present in more than 80 % of metadata records. Other optional elements are used significantly less.

The use of metadata elements present in more than 5 % of all metadata records by metadata type is displayed in figure 4, based on the NISO metadata typology (Riley 2017, p. 6). Overall, many metadata elements in the DataCite Metadata Schema are used infrequently: of the 83 metadata elements, more than half (63.9 %, n = 53) are present in less than 5 %

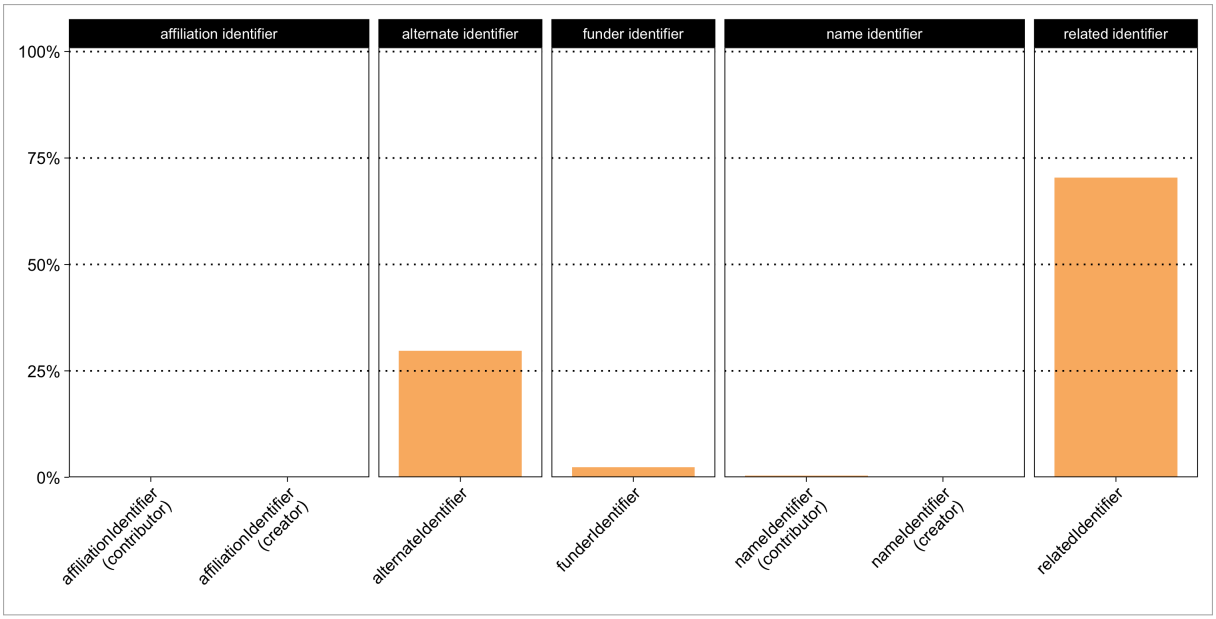


Figure 5: Use of persistent identifiers (n = 606091 metadata records)

of all metadata records. Of these elements, most (46) are categorized as descriptive metadata, including all 17 child elements and attributes of the main element *geolocation*. A smaller number of structural (3), rights (3) and technical (1) metadata elements do not surpass the 5 % threshold.

4.3. Use of persistent identifiers

An important subgroup of metadata elements are persistent identifiers. Related identifiers referring to a related resource are used most frequently (70.4 %), as figure 5 shows. 29.7 % of metadata records specify at least one alternate identifier, other persistent identifiers are used rarely.

In total, 35 repositories use related identifiers. 21 use alternate identifiers, 16 use funder identifiers, 14 use name identifiers for contributors, 4 use name identifiers for creators, and one uses affiliation identifiers.

4.4. Comprehensiveness of descriptions

The comprehensiveness of metadata descriptions refers to the use of the element *description*, particularly the number and length of descriptions provided.

Even though not all metadata records provide a description for the dataset they refer to, some offer up to 6 distinct description texts. The combined character length for all descriptions of a metadata record varies between 0 and 54,468 characters, with an average of 487.3 characters.

4.5. Completeness of metadata records

At the level of metadata records, completeness describes how many of the available metadata elements were used to describe an information object.

Metadata records in the sample use between 8 and 52 metadata elements. On average, 18.7 elements are used per record, which corresponds to 24.7 % of the available metadata elements.

4.6. Homogeneity of metadata collections

The homogeneity of metadata collections indicates how consistent metadata records within a collection are in terms of the metadata elements used. To determine the homogeneity of a metadata collection, the most common combination of metadata elements used is identified, as well as the number of metadata records using this combination.

Between 9.9 % and 100 % of metadata records in a repository's metadata collection use the same

parent element	element	η^2	p-Value
geoLocation	geoLocationPolygon	0.126	0.006
geolocation	polygonPoint	0.126	0.012
language	language	0.173	0.042

Table 3: Results of the Kruskal-Wallis test (repository type) for the completeness of individual metadata elements

parent element	element	r	p-Value
contributor	contributorType	0.348	0.018
contributor	contributorName	0.348	0.018
contributor	affiliationIdentifier (contributor)	0.349	0.02
contributor	affiliationIdentifierScheme (contributor)	0.349	0.02
contributor	schemeURI (contributor affiliation identifier)	0.349	0.02
date	dateInformation	0.349	0.02
relatedIdentifier	relatedIdentifier	0.29	0.048
relatedIdentifier	relatedIdentifierType	0.29	0.048
relatedIdentifier	relationType	0.29	0.048
format	format	0.316	0.032
geoLocation	geoLocation	0.51	< 0.001
geoLocation	geoLocationBox	0.493	< 0.001
geoLocation	geoLocationPlace	0.428	0.004

Table 4: Results of the Mann-Whitney U-test (certification status) for the completeness of individual metadata elements

combination of elements to describe datasets. On average, 50.9 % metadata records within a metadata collection share a common set of metadata elements. The size of this set varies between 9 and 39 elements, with an average of 19.6 elements.

4.7. Differences of metadata quality between repository groups

The quality indicators at the level of individual metadata elements, metadata records and metadata collections described in the previous section were analyzed to determine if there are significant differences between repository groups. The groups reflect the type and certification status of repositories.

4.7.1. Use of individual elements

Differences in the use of individual metadata elements across repository types are significant for three metadata elements providing information on geolocation and language, with a moderate effect size in all three cases (see table 3). In contrast, the number of elements that significantly vary in use at repositories with different certification status is larger (see table 4). Effect sizes are moderate for elements displaying information on contributors, date, format, and geolocation. The effect size for the parent element *geoLocation* is large.

4.7.2. Comprehensiveness of descriptions

Descriptions are on average most detailed for repositories of the type *other* (556.68 characters), and shorter for institutional (468.5 characters) and disciplinary (466.94 characters) repositories. Results of a Kruskal-Wallis test show that the difference is significant ($p < 0.001$) across repository types, but the effect size is small ($\eta^2 = 0.012$). At 549.31 characters on average, descriptions are longer at repositories without formal certification than at repositories with formal certification (185.69 characters). The Mann-Whitney U-test shows that the difference is significant at a 5 % significance level, with a moderate effect size ($r = 0.322$).

4.7.3. Completeness of metadata records

The average record completeness is highest for disciplinary repositories (26.1 %), followed by repositories of the type *other* (24.8 %) and institutional repositories (24.5 %). A Kruskal-Wallis test shows

that differences across repository types are are significant ($p < 0.001$), but effect sizes are small ($\eta^2 = 0.006$). On average, repositories without formal certification offer metadata records with a slightly higher degree of completeness (24.8 %) compared to repositories with formal certification (24.5 %). The Mann-Whitney U-test shows that the difference is significant ($p < 0.001$), and the effect size is small ($r = 0.145$).

4.7.4. *Homogeneity of metadata collections*

On average, metadata collections of disciplinary repositories are most consistent (61.1 %), followed by repositories of the type *other* (53.2 %) and institutional repositories (41.1 %). Repositories with formal certification have more homogenous metadata collections on average (67 %) compared to repositories without formal certification (48 %). Differences in collection homogeneity are neither significant across repository types nor across repositories with and without formal certification.

5. Discussion

5.1. Using a generic metadata schema for describing diverse research data

The analysis revealed that on average, metadata records use 24.7 % of the metadata elements available. More than half of all metadata elements are used in less than 5 % of metadata records, with most of these element being descriptive. These figures may appear low or suggest an insufficient level of description. However, this conclusion cannot be drawn without considering the objective and design of the DataCite Metadata Schema. The DataCite Metadata Schema is intentionally inclusive as a result of its application – registering DOIs as well as enabling retrieval and citation of research output (DataCite Metadata Working Group 2019).

Generally, it allows for making uniform statements about a large number of heterogenous research data, but not all schema elements are applicable to all datasets. Throughout revisions over time, the schema has also become more detailed in some areas compared to others. A good example for this is the main element *geoLocation*, which comprised 17 child elements and attributes in version 4.3 of the DataCite Metadata Schema. Despite the large number of elements related to geolocation information, not all datasets are associated with a particular region or place. Therefore, the observed variance in element use can at least in part be explained by the characteristics of repository collections: metadata completeness is not just an indicator of how well a dataset is described, but also of how suitable the metadata schema is for describing it. Because of their diversity, it is very difficult to compare descriptions of research data sets, even if they are based on the same metadata schema. For example, the analysis showed that the average completeness of the metadata collection of a given repository varies considerably between 13.47 % and 48.47 %. Future research should account for specific characteristics of data collections by examining individual disciplines separately. This is a challenge, however, because subject information is not available for all metadata records, as the analysis highlighted.

The analysis shows that some elements are used frequently, suggesting that they are applicable to a wide spectrum of datasets. If there were a core set of metadata elements for describing research data, it would likely include the mandatory elements *identifier*, *creator*, *title*, *publisher*, *publicationYear*, *resourceType*; the recommended elements *description*, *date*, *relatedIdentifier*, *subject*; and the optional element *rights*. In order to ensure a basic level of usefulness of datasets, repositories should focus on capturing this information at the minimum.

5.2. Underused metadata elements

Some elements of the DataCite Metadata Schema remain underused given their importance for reusing or contextualizing data, for example the elements *size*, *version*, and *format*. In the case of *format*, the element is used more frequently at repositories with formal certification. Certification may require repositories to reflect on the benefits of providing format information—for example, CoreTrustSeal mentions data formats explicitly in the requirements *Background Information & Context (RQ0)*, *Deposit & Appraisal (R8)*, *Preservation Plan (RQ9)*, *Quality Assur-*

ance (Q10), and *Reuse (R13)* (CoreTrustSeal Standards and Certification Board 2022). Repositories could improve the availability of these metadata elements by promoting their use among data providers, for example in data deposit guidelines or checklists. In addition, these elements (particularly *size* and *format*) could be candidates for retrospective, automated metadata enrichment, because they refer to technical specifications inherent to datasets.

Rights metadata should also be considered underused. Although some form of rights information is available for the majority of metadata records, the legal parameters for using some datasets remain unclear. Since missing licenses pose a significant barrier to data reuse, a change in the obligation level of the metadata element *rights* might be considered for future versions of the DataCite Metadata Schema.

Adding persistent identifiers to the metadata record contextualizes a dataset by connecting it to other entities including researchers, organizations or other documents. The result is a comprehensive representation of the research landscape that can serve novel use cases (Cousijn, Braukmann, et al. 2021). However, not all types of persistent identifiers are used with equal frequency. The analysis showed that the majority of datasets included at least one identifier referring to a related resource, for example to other versions of a dataset. Related identifiers can also refer to resources based on the dataset, such as journal articles. Journals increasingly publish guidelines that ask for references to the reported data. Authors also potentially benefit from establishing links between text and data publications in the form of increased citation rates (Colavizza et al. 2020). In contrast to related identifiers, identifiers for alternative resources, funders, researchers, and organizations were used less frequently. Repositories should make it a priority to add persistent identifiers to metadata records. In doing so, they can promote the use of persistent identifiers and connect their collections to the growing network of entities related to research and teaching.

A related issue is the adequate recognition of contributors. Information on contributors is only available for a small fraction of metadata records. Contributors are defined in the documentation of the DataCite Metadata Schema as "[t]he institution or person responsible for collecting, managing, distributing, or otherwise contributing to the development of the resource." (DataCite Metadata Working Group 2019, p. 18) Repositories and repository workers put a lot of resources and labor into managing datasets, but their contributions often remain invisible. Certified repositories are significantly more likely to provide contributor information. Other repositories should also seek recognition for contributors, thereby making their labor and contributions to data curation visible.

5.3. Influence of repository characteristics on metadata

The analysis revealed statistically significant differences across repositories of varying type and certification status in the use of individual metadata elements, the comprehensiveness of descriptions, and the completeness of metadata records.

The analysis clearly shows differences in the use of some individual elements between the groups, for example is contributor information more frequently available at repositories with formal certification. This could be interpreted as an indicator that certified repositories are more aware of the benefits of showcasing the effort behind data curation work and the contribution of diverse roles, including repository staff, to the usefulness of datasets and metadata records. Future research could focus the content of the element *contributorType* to establish who is being credited in metadata records.

Differences in the comprehensiveness of descriptions are also significant. The combined character length of descriptions is highest on average for repositories of type *other* and repositories without formal certification. The reason for this observation could be that repositories with long descriptions follow a self-deposit approach where data depositors provide metadata themselves, largely unsupervised by repository staff. Additional research on dataset descriptions would be useful to identify characteristics that make dataset descriptions useful from the perspective of data reusers.

On average, metadata completeness is highest for disciplinary repositories and repositories without formal certification. This might be surprising, because certification encourages repositories to reflect on their metadata practices and highlights the usefulness of comprehensive metadata descriptions. However, this is might merely be a result of using

a generic metadata schema for describing diverse datasets (see above).

Collection homogeneity can be interpreted as an indicator of mature metadata practices – repositories having established consistent metadata curation workflows for repository workers, offering guidance to data providers et cetera. Homogeneity of metadata descriptions varies considerably across repositories in the sample, and future research could identify factors leading to more consistent metadata collections overall. On average, metadata collections are most uniform at disciplinary repositories and at repositories with formal certification, but these differences are not significant.

Although there is evidence of differences between repository groups, this should not conversely be interpreted as clear confirmation that repositories of a specific type or certification status always adopt particular metadata practices or achieve a certain degree of metadata quality. Each repository covers a niche, with its unique mission, collection, designated community and extent of services. In addition, the resources available at a repository for metadata curation vary. Repositories differ in these factors even within the groups analyzed here. This is only a first step in detangling the interrelation between repository characteristics and metadata quality, and more research is needed to study the effect of repository resources, expertise and workflows on metadata quality.

6. Conclusion

Generally, it can be observed that metadata schemas specialized on describing research data are widely used; pervasive infrastructures facilitating data publication, data retrieval and data citation are maturing; and the value of good metadata for research data is widely recognized. While this environment is conducive to good metadata quality, the burden for producing high-quality metadata currently rests predominantly on repositories. It would therefore be beneficial to explore approaches to support repositories in improving metadata quality, for example by testing automated enrichment of metadata elements that are underused in relation to their usefulness. There already are studies exploring metadata enrichment, such as the inference of missing subject information from titles, descriptions and keywords of datasets (Weber et al. 2020). The application of these models for enriching repository metadata should be investigated further.

The analysis showed that some repositories scored high in metrics used to measure metadata quality. Several repositories have already established very consistent metadata collections, which could point to mature metadata practices and workflows. Identifying factors that favor high-quality metadata could make it possible to transfer successful approaches to other repositories. The analysis revealed some significant differences across repositories of varying type and certification status, but overall, more research is required to identify specific factors contributing to metadata quality.

6.1. Limitations

This analysis does not include all criteria for evaluating metadata quality mentioned in the literature, but focuses on aspects of metadata completeness and logical consistency and coherence.

The process for matching the two data sources used in this paper requires repositories that are technologically mature. Therefore, the number of repositories included in the analysis is limited (47), and results should not be considered representative of all research data repositories.

The analysis has shown that due to the diversity of datasets, the metadata records describing them are difficult to compare, even if they are based on the same generic metadata schema. Future research should take discipline-specific features of metadata into account.

References

Assante, M.; Candela, L.; Castelli, D.; Tani, A. (2016). Are Scientific Data Repositories Coping with Research Data Publishing? In *Data Science Journal* 15(6). DOI: 10.5334/dsj-2016-006.

Benjelloun, O.; Chen, S.; Noy, N. (2020). Google Dataset Search by the Numbers. In *The Semantic Web – ISWC 2020*. Ed. by Pan, J. Z.; Tamma, V.; d'Amato, C., et al. Lecture Notes in Computer Science. Berlin: Springer, pp. 667–682. DOI: 10.1007/978-3-030-62466-8_41.

Bruce, T. R.; Hillmann, D. I. (2004). *The continuum of metadata quality: defining, expressing, exploiting*. HDL: 1813/7895.

Chapman, A.; Simperl, E.; Koesten, L., et al. (2019). Dataset search: a survey. In *The VLDB Journal* 29(1), pp. 251–272. DOI: 10.1007/s00778-019-00564-x.

Colavizza, G.; Hrynaszkiewicz, I.; Staden, I., et al. (2020). The citation advantage of linking publications to research data. In *PLOS ONE* 15(4), e0230416. DOI: 10.1371/journal.pone.0230416.

CoreTrustSeal Standards and Certification Board (2022). *CoreTrustSeal Requirements 2023-2025*. DOI: 10.5281/zenodo.7051012.

Cousijn, H.; Braukmann, R.; Fenner, M., et al. (2021). Connected Research: The Potential of the PID Graph. In *Patterns* 2(1), p. 100180. DOI: 10.1016/j.patter.2020.100180.

Cousijn, H.; Feeney, P.; Lowenberg, D., et al. (2019). Bringing Citations and Usage Metrics Together to Make Data Count. In *Data Science Journal* 18(1). DOI: 10.5334/dsj-2019-009.

DataCite Metadata Schema (n.d.). DataCite. https://schema.datacite.org/ visited on October 10, 2022.

DataCite Metadata Working Group (2019). *DataCite metadata schema documentation for the publication and citation of research data v4.3*. In collab. with Smaele, M. de; Dasler, R.; Ashton, J., et al. DOI: 10.14454/7XQ3-ZF69.

Gonçalves, R. S.; Musen, M. A. (2019). The variable quality of metadata about biological samples used in biomedical experiments. In *Scientific Data* 6(1), pp. 1–15. DOI: 10.1038/sdata.2019.21.

Gregg, W.; Erdmann, C.; Paglione, L., et al. (2019). A literature review of scholarly communications metadata. In *Research Ideas and Outcomes* 5, e38698. DOI: 10.3897/rio.5.e38698.

Kim, J.; Yakel, E.; Faniel, I. M. (2019). Exposing Standardization and Consistency Issues in Repository Metadata Requirements for Data Deposition. In *College & Research Libraries* 80(6), pp. 843–875. DOI: 10.5860/crl.80.6.843.

Kindling, M.; Pampel, H.; Sandt, S. van de, et al. (2017). The landscape of research data repositories in 2015: a re3data analysis. In *D-Lib Magazine* 23(3). DOI: 10.1045/march2017-kindling.

Lee, D. J.; Stvilia, B. (2017). Practices of research data curation in institutional repositories: A qualitative view from repository staff. In *PLOS ONE* 12(3), e0173987. DOI: 10.1371/journal.pone.0173987.

Leonelli, S. (2020). Learning from Data Journeys. In *Data Journeys in the Sciences*. Ed. by Leonelli, S.; Tempini, N. Berlin: Springer, pp. 1–24. DOI: 10.1007/978-3-030-37177-7_1.

Manninen, L. (2018). Describing Data: A Review of Metadata for Datasets in the Digital Commons Institutional Repository Platform: Problems and Recommendations. In *Journal of Library Metadata* 18(1), pp. 1–11. DOI: 10.1080/19386389.2018.1454379.

Musen, M. A. (2022). Without appropriate metadata, data-sharing mandates are pointless. In *Nature* 609(7926), p. 222. DOI: 10.1038/d41586-022-02820-7.

Park, J.-R. (2009). Metadata Quality in Digital Repositories: A Survey of the Current State of the Art. In *Cataloging & Classification Quarterly* 47(3-4), pp. 213–228. DOI: 10.1080/01639370902737240.

Pomerantz, J. (2015). *Metadata*. Cambridge, MA and London: The MIT Press.

Quarati, A.; Raffaghelli, J. E. (2020). Do researchers use open research data? Exploring the relationships between usage trends and metadata quality across scientific disciplines from the Figshare case. In *Journal of Information Science* 48(4), pp. 423–448. DOI: 10.1177/0165551520961048.

Riley, J. (2017). *Understanding metadata: what is metadata, and what is it for?* National Information Standards Organization (U.S.) http://www.niso.org/publications/understanding-metadata-riley visited on October 10, 2022.

Robinson-Garcia, N.; Mongeon, P.; Jeng, W.; Costas, R. (2017). DataCite as a novel bibliometric source: Coverage, strengths and limitations. In *Journal of Informetrics* 11(3), pp. 841–854. DOI: 10.1016/j.joi.2017.07.003.

Rodrigues, J.; Castro, J. A.; Silva, J. R. da; Ribeiro, C. (2019). Hands-On Data Publishing with Researchers: Five Experiments with Metadata in Multiple Domains. In *Digital Libraries: Supporting Open Science*. Ed. by Manghi, P.; Candela, L.; Silvello, G. Communications in Computer and Information Science. Berlin: Springer, pp. 274–288. DOI: 10.1007/978-3-030-11226-4_22.

Rousidis, D.; Garoufallou, E.; Balatsoukas, P.; Sicilia, M.-A. (2014). Metadata for Big Data: A preliminary investigation of metadata quality issues in research data repositories. In *Information Services & Use* 34(3-4), pp. 279–286. DOI: 10.3233/ISU-140746.

Weber, T.; Kranzlmüller, D.; Fromm, M.; Sousa, N. T. de (2020). Using supervised learning to classify metadata of research data by field of study. In *Quantitative Science Studies* 1(2), pp. 525–550. DOI: 10.1162/qss_a_00049.

Wilkinson, M. D.; Dumontier, M.; Aalbersberg, I. J., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. In *Scientific Data* 3, p. 160018. DOI: 10.1038/sdata.2016.18.

York, J.; Gutmann, M.; Berman, F. (2018). What do we know about the stewardship gap. In *Data Science Journal* 17(19). DOI: 10.5334/dsj-2018-019.

Zeng, L.; Qin, J. (2022). *Metadata*. 3rd ed. London: Facet Publishing.

Appendices

A. Repositories in the sample

Table 5: Repositories in the sample

re3data ID	repository name	repository type	certification status
r3d100012587	EnviDat	disciplinary	FALSE
r3d100012825	Forschungsdaten-Repositorium der LUH	institutional	FALSE
r3d100012001	Illinois Data Bank	institutional	FALSE
r3d100012330	RADAR	other	FALSE
r3d100012646	Federated Research Data Repository	other	FALSE
r3d100012505	ORDaR	disciplinary	FALSE
r3d100012064	University of Reading Research Data Archive	institutional	FALSE
r3d100012927	Data Commons	institutional	FALSE
r3d100012140	Brunel figshare	institutional	FALSE
r3d100012190	ZBW Journal Data Archive	disciplinary	FALSE
r3d100012405	Research Data at Essex	institutional	FALSE
r3d100013062	Ifsttar research data	institutional	FALSE
r3d100012157	Fairdata IDA Research Data Storage Service	other	FALSE
r3d100011601	Structural Biology Data Grid	disciplinary | institutional	FALSE
r3d100012145	melbourne.figshare.com	institutional	FALSE
r3d100012633	ZivaHub	institutional	FALSE
r3d100011864	OpenKIM	disciplinary	FALSE

Continued on next page

Table 5: Repositories in the sample (Continued)

re3data ID	repository name	repository type	certification status
r3d100011890	Ag Data Commons	disciplinary	FALSE
r3d100011945	Research Data Leeds Repository	institutional	FALSE
r3d100012414	UEL Research Repository	institutional	FALSE
r3d100012147	Stockholm University repository for data	institutional	FALSE
r3d100011947	University of Bath Research Data Archive	institutional	FALSE
r3d100012417	UCL Discovery	institutional	FALSE
r3d100012384	CaltechDATA	institutional	FALSE
r3d100012369	Code Ocean	disciplinary	FALSE
r3d100012335	GFZ Data Services	disciplinary	FALSE
r3d100010216	4TU.ResearchData | science.engineering.design	disciplinary | institutional	TRUE
r3d100012564	ScholarBank@NUS	institutional	FALSE
r3d100011662	Landcare Research Data Repository	disciplinary | institutional	FALSE
r3d100010299	World Data Center for Climate	disciplinary	TRUE
r3d100010478	GigaDB	disciplinary	FALSE
r3d100012557	ETH Zürich Research Collection	institutional	FALSE
r3d100010731	Open Data LMU	institutional | other	FALSE
r3d100011038	Qualitative Data Repository	disciplinary	TRUE
r3d100012143	Loughborough Data Repository	institutional	FALSE
r3d100012965	IFREMER-SISMER Portail de données marines	disciplinary	TRUE
r3d100000044	DRYAD	other	FALSE
r3d100012538	DataverseNO	disciplinary | institutional | other	TRUE
r3d100000006	Archaeology Data Service	disciplinary	TRUE
r3d100010066	figshare	other	FALSE
r3d100010468	Zenodo	other	FALSE
r3d100010664	World Stress Map	disciplinary	TRUE
r3d100011108	heiDATA	institutional | other	FALSE
r3d100012757	RODARE	institutional	FALSE
r3d100013029	TUdatalib	institutional	FALSE
r3d100013084	SURF Data Repository	other	FALSE
r3d100013275	GRO.data	institutional	FALSE

Eingereicht: 10.10.2022; Revision 1: 14.11.2022; Revision 2: 23.11.2022; akzeptiert: 01.12.2022; publiziert: 14.12.2022