



Affiliation Information in DataCite Dataset Metadata: a Flemish Case Study

RESEARCH PAPER

NIEK VAN WETTERE 

]u[ubiquity press

ABSTRACT

This article aims to evaluate how and to what extent metadata of datasets indexed in DataCite offer clear human- or machine-readable information that enables the research data to be linked to a particular research institution. Two main pathways are explored. First, researchers can encode their affiliation information at the moment of data submission. This can be done by means of free-text metadata fields or via the inclusion of identifiers such as GRID/ROR and ORCID. Second, affiliation information can be traced indirectly through linking between a dataset and associated publications, given that the metadata of publications is often more explicit about affiliation information than the metadata of datasets. Both pathways of affiliation information encoding are evaluated on the basis of metadata pertaining to datasets created at the five Flemish universities. It is shown that good practices such as encoding of affiliation information in a dedicated metadata field or inclusion of ORCID in the metadata are on the rise, but could be expanded further. Finally, the establishment of links between datasets and related publications is often lacking in dataset metadata, although there are important differences between data repositories, as is also demonstrated in a more data-intensive follow-up analysis based on random samples of metadata records. It is important that data repositories address this issue by providing a metadata field clearly dedicated to associated publications, prominently displayed on the landing page of the dataset.

CORRESPONDING AUTHOR:

Niek Van Wettere

Vrije Universiteit Brussel, BE

niek.van.wettere@vub.be

KEYWORDS:

DataCite; Scholix; research data; metadata; affiliation

TO CITE THIS ARTICLE:

Van Wettere N. 2021. Affiliation Information in DataCite Dataset Metadata: a Flemish Case Study. *Data Science Journal*, 20: 13, pp. 1–18. DOI: <https://doi.org/10.5334/dsj-2021-013>

In the broader context of funder requirements with regard to research data and open science, long-term preservation of research data is increasingly gaining in importance. As research data are more and more being archived by researchers in trustworthy data repositories, it is important that research institutions can keep track of the research data that were collected or created at their institution (Khan, Pink & Thelwall 2020). Essentially, this linking problem hinges upon the completeness and quality of the metadata associated with the research data and the degree to which links between research objects are available in metadata hubs. Unsurprisingly, the possibility to connect data with researchers, institutions and associated publications constitutes an important user requirement towards data repositories (Wu et al. 2019). In this article, different existing possibilities to link datasets to their respective institutions are explored and evaluated.

Concretely, this study aims to answer the following two research questions concerning affiliation linking:

1. In which ways is affiliation information included in the metadata pertaining to research data? If a researcher is able to record his/her affiliation to certain research institutions in the metadata of the archived research data, the research data can be more easily linked to the host institution(s) in question. Affiliation information can be encoded as free text, but also in a structured, machine-readable format.
2. To what extent are links between publications and datasets available in metadata pertaining to research data? Given that institutions often already have a good overview of publications created at their institution, linking between publications and datasets in external metadata hubs constitutes an interesting avenue to identify those research data that are associated with a particular institution, even in the absence of explicit affiliation information encoded in the metadata of the dataset.

The first research question will be addressed in Section 4.2, whereas the second research question is examined in Section 4.3. More specifically, these two issues will be analyzed on the basis of DataCite metadata of datasets collected or created at one of the Flemish universities. In Section 4.4, a more in-depth follow-up analysis on the basis of randomly collected DataCite metadata examines these two research questions more closely, in particular the encoding of links between a dataset and related research outputs.

The remainder of this article is structured as follows. Section 2 introduces DataCite as an important metadata hub for research data. Next, the methodology that was adopted in order to collect the relevant metadata from DataCite is explained in Section 3. The results drawn from the metadata that were harvested, are presented in Section 4. Finally, Section 5 establishes the main conclusions.

2 DATACITE METADATA HUB

This study draws on metadata collected from the DataCite metadata hub. DataCite is currently one of the most important metadata hubs and search engines for research data.¹ Moreover, DataCite is at the origin of the DataCite Metadata Schema, which is a widely used domain-agnostic metadata standard for research data.² The DataCite Metadata Schema captures basic information about research datasets such as data creator, publisher of the research data etc. [Table 1](#) gives an overview of the different DataCite metadata fields³ that are taken into account for this study.

In addition to DataCite, we also briefly introduce the Scholix framework because of its aptitude to detect links between datasets and associated literature, particularly relevant to the second research question (cf. Section 4.3). Scholix is a Data-Literature Interlinking (DLI) Service, based on a global standard for links between research data and literature (Burton

¹ Next to DataCite, OpenAire will probably be developed into the main European metadata hub for all types of research outputs. It is not entirely clear how these two main players will interact in the future.

² An alternative domain-agnostic metadata standard is Dublin Core.

³ Contributor fields are not included in this study.

DATA CITE METADATA FIELD	EXPLANATION
DOI (or other persistent identifier)	Persistent identifier that refers to the landing page containing the metadata of the dataset. If the data are open, download of the data is also available.
URL	URL that refers to the landing page containing the metadata of the dataset. If the data are open, download of the data is also available.
Publisher	The archiving organization/data repository that publishes the dataset.
Client ID	Unique identifier for a DataCite client. Since the publisher field sometimes contains different names for the same publisher, this Client ID seems useful to group together dataset records originating from the same archiving organization.
Publication year	The year when the data was or will be made publicly available. The DataCite documentation also specifies that, if there is no standard publication year value, the date that is preferred from a citation perspective should be used.
Description	Description of the dataset (free text), for example in the form of an abstract.
Name of data creators	First name and family name of the researchers who collected or created the data.
Identifier of data creators	ORCID that uniquely identifies each data creator.
Affiliation of data creators	Research institution to which each data creator contributing to the dataset is affiliated.
Identifiers of related output	Persistent identifiers that refer to research outputs related to the dataset in question. These outputs can be other datasets or associated publications (such as articles in journals).

Table 1 Overview of the different DataCite metadata fields included in the study.

et al. 2017; Cousijn et al. 2019). As such, their main objective is to provide a superordinate framework for the establishment of links between research publications and their associated research data. Scholix draws on existing sources such as DataCite and Crossref to aggregate links between publications and datasets, regardless of whether the link has been established in the dataset or publication metadata. In this way, Scholix links are in principle bidirectional: from the dataset to the associated datasets or literature and vice versa. Contrary to DataCite, Scholix indicates the category of the related research output (dataset/literature/other), so that dataset-to-dataset links can be easily distinguished from dataset-to-publication links. Since most links between datasets and associated literature made available via the REST API of Scholexplorer seem to originate from DataCite at this stage (Khan, Pink & Thelwall 2020), the current study chooses to primarily focus on what can be gleaned from the DataCite metadata. However, as will become apparent in Section 4.3, Scholix is able to contribute new links between data and publications that are not yet available in the DataCite metadata in a limited number of cases.

3 METHODOLOGY

In order to collect the data necessary to conduct this study, metadata pertaining to datasets collected or created at one of the five Flemish universities were harvested from DataCite in December 2020. The harvest was performed using the R package *rdatacite* (Chamberlain 2020). Metadata related to a unique dataset DOI and containing at least in one field one of the institutional names were extracted for subsequent analysis.

Table 2 below lists the queries that were used per university in order to detect the institution names of the Flemish universities in the metadata of the research datasets. Insofar as possible, the aim was to formulate for each university the official Dutch name, at least one common abbreviation of the Dutch name, the English equivalent of the Dutch institution name and, finally, the corresponding ROR identifier (Lammey 2020). Given that titles and abstracts of datasets are also part of the metadata that was taken into account, abbreviated institution names that do not contain the full name of the corresponding city where the university is located (e.g. KUL, VUB) were not included as to minimize the number of false positives.

If the institution name contains different words (for example, ‘Ghent University’ comprises ‘Ghent’ and ‘University’), a query was formulated that requires all the words of the institution

name to be present in the metadata, hence the AND-operator between different components of the institution name in [Table 2](#). Importantly, the structure of the search query does not entail that the words have to appear in that specific order, nor does it imply that the words have to concatenate. In other words, it is possible that other words (such as *of* etc.) intercalate between the component terms of the search query.

FLEMISH UNIVERSITY	SEARCH QUERY
Katholieke Universiteit Leuven	Katholieke AND Universiteit AND Leuven
	KU AND Leuven
	KULeuven
	Catholic AND University AND Leuven
	https://ror.org/05f950310
Universiteit Antwerpen	Universiteit AND Antwerpen
	UAntwerpen
	University AND Antwerp
	https://ror.org/008x57b05
Universiteit Gent	Universiteit AND Gent
	UGent
	Ghent AND University AND NOT (Global AND Biodiversity AND Information AND Facility) ⁴
	https://ror.org/00cv9y106
Universiteit Hasselt	Universiteit AND Hasselt
	UHasselt
	Hasselt AND University AND NOT (Global AND Biodiversity AND Information AND Facility)
	https://ror.org/04nbhqj75
Vrije Universiteit Brussel	Vrije AND Universiteit AND Brussel ⁵
	https://ror.org/006e5kg04

Table 2 Flemish universities and name variants.

In total, the metadata harvested from the DataCite API consists of 1050 DOIs. Note that the search terms ‘UAntwerpen’ and ‘UHasselt’ did not yield any results, nor did the ROR identifiers.⁶ However, these raw metadata still contain false positives and redundant DOIs that exclusively refer to a particular version or part of a dataset. In order to filter out the DOIs that do not refer to the main dataset, but rather versions or parts of the dataset, different computational strategies had to be used. First of all, clusters of potentially redundant DOIs were detected on the basis of an overlapping combination of the three metadata fields Client ID, publication year and names of data creators. These clusters can then be examined more closely. When the ‘IsVersionOf’ or ‘IsPartOf’-relation is available in the metadata associated with a certain DOI, this can be considered as an indicator to remove the DOI,

⁴ Since downloads of (recombined) subsets of datasets are also registered as separate DOIs for the Global Biodiversity Information Facility (GBIF), the number of DOIs associated with GBIF is artificially high. Therefore, the combination of the search terms Global, Biodiversity, Information and Facility is excluded in the search queries ‘Ghent University’ and ‘Hasselt University’, for which the number of GBIF results is important and complicates the metadata harvesting process.

⁵ The (incorrect) name variant ‘Free University of Brussels’ caused an internal error at the DataCite server. Therefore, this variant was not included. Moreover, the abbreviation ‘VUB’ was not used to avoid noisy data with numerous false positives: ‘VUB’ could potentially refer to many other things than the Vrije Universiteit Brussel, especially in titles and abstracts of datasets.

⁶ However, it seems that the DataCite Commons (<https://commons.datacite.org/>) that was officially launched in October 2020, is able to generate supplementary results (via <https://commons.datacite.org/ror.org?query=>) that were not obtained via our methodology. Since these results seem for the moment restricted to DOIs from the Dryad repository, this is probably a consequence of the ROR implementation at Dryad, facilitating the discovery of DOIs related to a particular institution (Lammey 2020).

on the condition that at least one other DOI in the cluster is not endowed with this relation type.⁷ Unfortunately, not all cases of versioning or part-to-whole relationship could be revealed by means of ‘IsVersionOf’- or ‘IsPartOf’-relations readily available in the metadata. Therefore, two indicators in the URL were also taken into account: an URL ending in ‘.v1’ or an URL referring to one particular file within the broader data package (e.g. ‘https://dataverse.harvard.edu/file’). Most cases could be resolved automatically in this way, although some remaining cases had to be sorted out manually. Next, false positives without an actual link with a Flemish university (for example a historical dataset about the city Ghent compiled by a different university than Ghent University) were removed after manual scrutiny. After removal of all the false positives and redundant DOIs, 257 unique DOIs remain to be analyzed. Given that only metadata records that contain an explicit reference to one of the Flemish universities are taken into consideration, it seems reasonable to expect that this collection of DOIs is still non exhaustive, but it is difficult to assess how (in)complete this set of DOIs exactly is.⁸

Following the DataCite harvest, the API of Scholexplorer (cf. also <http://api.scholexplorer.openaire.eu/v2/ui/>) was interrogated in order to check for each of the 257 DOIs whether Scholix establishes a link to a related dataset or publication. More specifically, each harvested DataCite DOI was fed into the Scholix query field ‘sourcePid’ (source persistent identifiers) to filter all the Scholix relationships according to these input DataCite DOIs. If, for a given DataCite DOI, Scholix links to multiple datasets or articles, only one of the linked DOIs is included in our dataset. In any case, there do not seem to be any datasets that link to both associated datasets and publications for the subset of metadata that was collected. Consequently, the share of DOIs that link to related publications can be reliably measured. In sum, this second Scholix harvest gives rise to the addition of two new variables to our dataset, namely the category of the research output to which the dataset in question is linked and the corresponding persistent identifier of the related research output.

Finally, the data repositories used by Flemish researchers according to the harvested DataCite metadata were also subjected to a more in-depth follow-up analysis. Concretely, a random sample of 1000 DOIs belonging to data published in 2019 or 2020 was drawn for each data repository, by means of the `dc_dois` command from the `rdatacite` package. The most recent publication years 2019 and 2020 were chosen in order to reflect the most recent situation. Subsequently, potentially redundant DOIs, redundancy being defined as an overlapping combination of the three metadata fields Client ID, publication year and names of data creators, were removed from each sample. Archiving organizations with relatively low sample sizes (the maximum sample size was not obtained following the `dc_dois` command and/or too many redundant DOIs) were left out of the analysis. Next, the highest common sample size shared between the samples of the remaining repositories was determined and all samples were aligned on this sample size (= 450 DOIs). These samples were then analyzed via a heatmap and a mosaic plot, as reported in Section 4.4.

4 RESULTS

The results section addresses four main topics, namely (1) an overview of the metadata extracted from DataCite for this study, (2) an analysis of the different ways in which data creators encode affiliation information in the metadata of datasets, (3) an evaluation of how DataCite performs with regard to the detection of research output related to archived datasets and (4) a follow-up analysis that examines more closely the different data repositories on the basis of larger random samples.

⁷ If all DOIs within a cluster refer to specific versions or parts, only one DOI is retained in the final set of metadata.

⁸ A possible avenue to tackle this problem is to screen the DataCite metadata records in a different way. Instead of searching directly for mentions of institution names in the metadata of the datasets, a list of publication DOIs extracted from institution’s CRIS databases could be used to check whether metadata hubs such as DataCite contain metadata records about datasets that contain these publication DOIs, for instance in the metadata fields pertaining to related identifiers. Of course, this approach will not give rise to an exhaustive overview either, because it relies on the fact that researchers themselves take the time to encode machine-readable links to associated publications in the metadata of datasets, but could nonetheless yield more complete results than the approach followed in this study.

4.1 OVERVIEW OF THE METADATA EXTRACTED FROM DATACITE

The overall aim of this first subsection is to concisely summarize the content of the metadata that were harvested from DataCite. [Table 3](#) offers a view on the number of distinct datasets⁹ (operationalized as distinct DOIs) per publication year¹⁰ in our data.

PUBLICATION YEAR	NUMBER OF DOIS
1989	1
2002	1
2006	1
2007	15
2009	1
2011	2
2012	1
2013	3
2014	8
2015	16
2016	61
2017	31
2018	29
2019	40
2020	47
Total	257

Table 3 Number of datasets, per publication year (1989–2020).

[Table 3](#) suggests that digital data archiving in Flanders is more prevalent in the period 2015–2020 than in the previous period. This reflects the general tendency towards increasing awareness about Open Science and RDM in Europe over the past years.

Furthermore, [Table 4](#) adds an extra dimension to the picture, namely the archiving organization where the dataset is archived for long-term preservation. Based on [Table 4](#), we can evaluate which archiving organizations are the most popular ones among researchers affiliated with Flemish research institutions. For each archiving organization, the number of distinct DOIs is determined per publication year.

Clearly, the discipline-specific repository ‘Marine Data Archive’ (VLIZ in Dutch) is an important actor in the Flemish archiving landscape. The rest of the list is dominated by domain-generic data repositories, such as Harvard Dataverse, Zenodo and DANS. Commercial domain-generic repositories such as Figshare and Mendeley form a minority, relatively speaking.

⁹ Note that the term ‘dataset’ can also mean in a broader sense a ‘data package’. Strictly speaking, the data associated with a particular DOI can comprise more than just ‘a dataset’ (in the sense of a coherent set of individual, non-aggregated data values). For example, a single DOI can refer to both a dataset containing variables measured on people and associated computational scripts that formalize all the analysis steps applied to the dataset in question. This second component is of course also very important from a reproducibility perspective. Consequently, it can be debated whether the term ‘data package’ is not more appropriate to capture the possible heterogeneity of the data bundled together. Moreover, the term ‘data package’ also emphasizes the need to avoid excessive ‘salami slicing’, in which case even the smallest piece of the same overarching data package is archived independently, with separate DOI attribution. However, since the term ‘dataset’ is most commonly used, this term will also be adopted in the remainder of the text.

¹⁰ As mentioned in [Table 2](#), the metadata field ‘publication year’ can also refer to ‘any date that is relevant from a data citation perspective’. Consequently, it is possible that certain dates do not reflect the date that the data were uploaded in a public repository.

ARCHIVING ORGANIZATION	NUMBER OF DOIS
delft.vliz (Marine Data Archive)	72
gdcc.harvard-dv (Harvard Dataverse)	46
cern.zenodo (Zenodo)	40
dans.archive (DANS - Data Archiving and Networked Services)	26
gbif.gbif (Global Biodiversity Information Facility) ¹¹	24
delft.rbins (RBINS - Royal Belgian Institute for Natural Sciences, OD Nature - Directorate Natural Environment, BMDC - Belgian Marine Data Centre)	10
figshare.ars (Figshare)	10
pangaea.repository (PANGAEA - Data Publisher for Earth & Environmental Science)	9
bl.mendeley (Mendeley)	7
delft.data4tu (4TU.Centre for Research Data)	3
doe.lbnl	2
dryad.dryad (Dryad)	2
bl.oxdb (FAIRsharing)	1
europ.odin (European Commission JRC)	1
geis.gesis (GESIS Data Archive)	1
geis.icpsr (ICPSR - Interuniversity Consortium for Political and Social Research)	1
ieee.dataport (IEEE DataPort)	1
tib.ideo (IEDA - Interdisciplinary Earth Data Alliance)	1
Total	257

Table 4 Number of DOIs, per archiving organization.

4.2 ENCODING OF AFFILIATION INFORMATION IN DATACITE METADATA

In order for research institutions to be able to retrieve the metadata of datasets that were collected/created by a researcher affiliated with their institution, it is important that researchers are able to encode affiliation information in the metadata associated with their archived data. In other words, encoding of affiliation information should be an obligatory component of data archiving as it is implemented by research data repositories. Currently, the affiliation sub-property of the data creator element is optional in the DataCite Metadata Schema (but recommended in the OpenAire guidelines, cf. https://guidelines.openaire.eu/en/latest/data/field_creator.html). Consequently, the affiliation metadata field is often incomplete in metadata records (Habermann 2019). In the metadata that were extracted from DataCite for the purpose of this study, two main strategies to encode affiliation information were discerned. First, affiliation information can be encoded as free text (i.e. the name of the research institution as it is submitted in written form by the researcher) in certain metadata fields such as name(s) of the data creator(s) and/or affiliation. This practice is analyzed in Section 4.2.1.

Second, affiliation information can also be deduced from the researcher's ORCID. This affiliation information can in principle be directly retrieved in the ORCID profile of the researcher in question, provided that this information is kept up-to-date by the researcher and is open to the public. An alternative is that institutions screen metadata hubs for ORCIDs that match the ORCIDs of researchers with an active appointment at their institution (as registered in their CRIS-system). In order to optimize the relevance of the matches, only metadata of datasets archived within the period of the active appointment are to be harvested. Section 4.2.2 will explore in more detail the usage of ORCIDs in DataCite metadata.

In addition to these two strategies, other possibilities can be contemplated to relate datasets to the research institutions where they were collected or created. The most straightforward way to accomplish this is the encoding of an organizational identifier in the metadata of

¹¹ Even though the Global Biodiversity Information Facility was maximally excluded in the search queries (cf. supra), a relatively low number of GBIF DOIs for which publisher names such as 'Ghent University' and 'Marine Biology Section Ugent' are registered, were harvested and thus included in our data. It goes without saying that the number of GBIF DOIs reported in **Table 4** is an underestimation of the actual number of DOIs associated with GBIF.

the dataset. The DataCite metadata schema has been modified in 2019 to include a field for organizational identifiers such as a GRID (Hook, Porter & Herzog 2018) or ROR ID, which ‘will enable more efficient discovery and tracking of publications by institutions and is making unambiguous affiliation information widely and freely available’ (DataCite Metadata Schema v.-4.3). Once this innovation is also implemented by the different data repositories¹², which have to adapt their metadata ingestion process accordingly, this would constitute a promising and, what is more, machine-readable third way to effectively encode affiliation information in metadata. For the moment, we do not see this practice in our collected data.

Finally, since publication output is an important factor in public funding allocated to research institutions, links between publications (articles in journals etc.) and research institutions are often already more generally and more reliably established than is currently¹³ the case for research datasets. If there exists an efficient linking between publications and their corresponding datasets, affiliation information pertaining to the datasets could be retrieved via the link of the dataset with the publication (cf. also the methodology reported in Khan, Pink & Thelwall 2020), provided that the metadata of the publication contains affiliation information¹⁴ and/or is registered in the appropriate institutional CRIS-system(s). This also presupposes that the link between publications and their corresponding datasets differentiates between new data that were specifically collected/created for the linked publication and re-use of already existing data. Of course, in the latter case, affiliation information concerning the publication cannot be directly attributed to the datasets that were analyzed to conduct the research.¹⁵ Linking between publications and associated datasets will be discussed in Section 4.3.

4.2.1 Free text metadata fields

Our data show that affiliation information is encoded in a wide variety of free text metadata fields.¹⁶ In addition to the metadata field specifically dedicated to affiliation information, numerous researchers also add affiliation information to the field describing the names of the data creator(s) (mostly name + research institution between parentheses) or to the description field. Frequently, affiliation information is encoded redundantly in two or more fields. In general, the affiliation information included in these fields is in line with each other, although the exact wording can differ to some extent. However, in some cases, one field is more complete than the other field. For example, one of both fields sometimes lacks one or more of the institution names associated with a particular author. Moreover, one of both fields can contain more in-depth information, such as affiliation information at the department level, whereas the other field only mentions the affiliation information at the university level. Next to the affiliation and name metadata fields, other fields are also used to encode affiliation information, such as the publisher¹⁷ field.

Figure 1 shows which strategies researchers in our data use to encode affiliation information for the period 2006–2020. The x-axis corresponds to the year in which the dataset was published, whereas the y-axis relates to the number of unique DOIs. The points on the figure are jittered so that overlapping points remain visible. In order to guarantee the readability of the plot, combinations of metadata fields that do not occur in more than two publication years (such as the combination ‘Affiliations_data_creators + Description + Identifiers_related_outputs + Names_data_creators’), are left out of the plot. It goes without saying that the trends that can be deduced from this figure are very tentative.

In the early period 2006–2012, researchers encode affiliation information exclusively in the name field (cf. ‘Names_data_creators’) or the description field (cf. ‘Description’). The former strategy, making use of the name field, continues to be important between 2012 and 2017, but starts to fall into disuse from 2017 onwards, after the affiliation field was introduced in the DataCite

¹² Note that retrospective assignment of RORs applied to existing metadata can prove to be problematic (Habermann & Lowerberg 2019).

¹³ It is to be expected that this will change in the coming years.

¹⁴ Note that information on the author’s institution is often missing from the article metadata at Crossref (Lammey 2020).

¹⁵ There might be exceptions to the former case as well, but this will probably be a minor subset.

¹⁶ Since our data were harvested based on the encoding of affiliation information somewhere in the metadata, all collected DOIs contain some form of encoded affiliation information.

¹⁷ In other words, the archiving organization, cf. [Table 2](#).

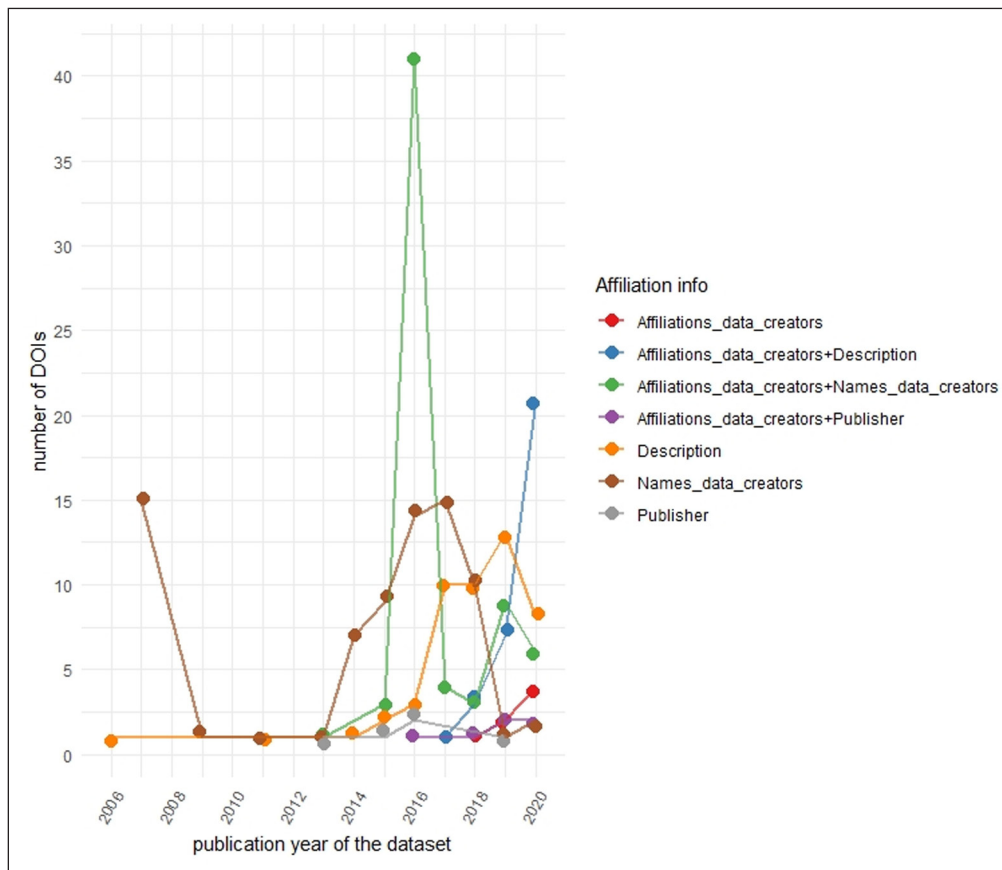


Figure 1 In which metadata fields do researchers encode affiliation information? An overview for the period 2006–2020.

Metadata Schema. The latter strategy, making use of the description field, remains successful beyond 2012, although there is a small drop in 2020. It is clear that many researchers revert to a general-purpose field such as description instead of the more specialized fields, possibly because they are not yet familiar enough with the global metadata structure. Crucially, the practice to encode affiliation information exclusively in the affiliation field is not very widespread, but gains in importance from 2017 to 2020. Since the dedicated affiliation field was only added in version 3.1 of the DataCite metadata schema (~ 2015), this increasing uptake is of course fairly recent. From a metadata management perspective, this last practice is what should be encouraged.

Let us now turn to the cases where affiliation information is encoded in two fields simultaneously. The practice to encode affiliation information in both the name and the affiliation field displays an upward¹⁸ and ensuing downward trend between 2012 and 2020, besides a late revival in 2019. Interestingly, researchers start to encode affiliation information in both the description and the affiliation fields in 2017. In the following years, this approach is clearly on the rise, even becoming the top strategy in 2020. It might be the case that researchers feel the need to highlight certain important elements redundantly in the description, so that they receive more emphasis in the global display of the repository landing page.

4.2.2 ORCID

In this section, it is examined to which degree researchers add ORCID information to the metadata of datasets. *Figure 2* shows how many DOIs have ORCID information associated with them per publication year. To be clear, it suffices that at least one ORCID is registered in the metadata, regardless of the number of data creators. Our analysis does not include an evaluation of the completeness of the ORCID information that is available (i.e. whether an ORCID is encoded for every data creator associated with a particular DOI).

ORCID was launched in 2012. Clearly, ORCID encoding in metadata is steadily on the rise from 2017 onwards in our data, to the point that the encoding of ORCID becomes even more prevalent than the absence thereof in 2020. In principle, ORCID is mandatory for Flemish researchers involved in projects funded by BOF/IOF or FWO since 2019. Presumably, this explains the steep

¹⁸ The peak observed in 2016 is exclusively due to datasets archived at the Marine Data Archive. Most data creators work at the Marine Biology section of Ghent University.

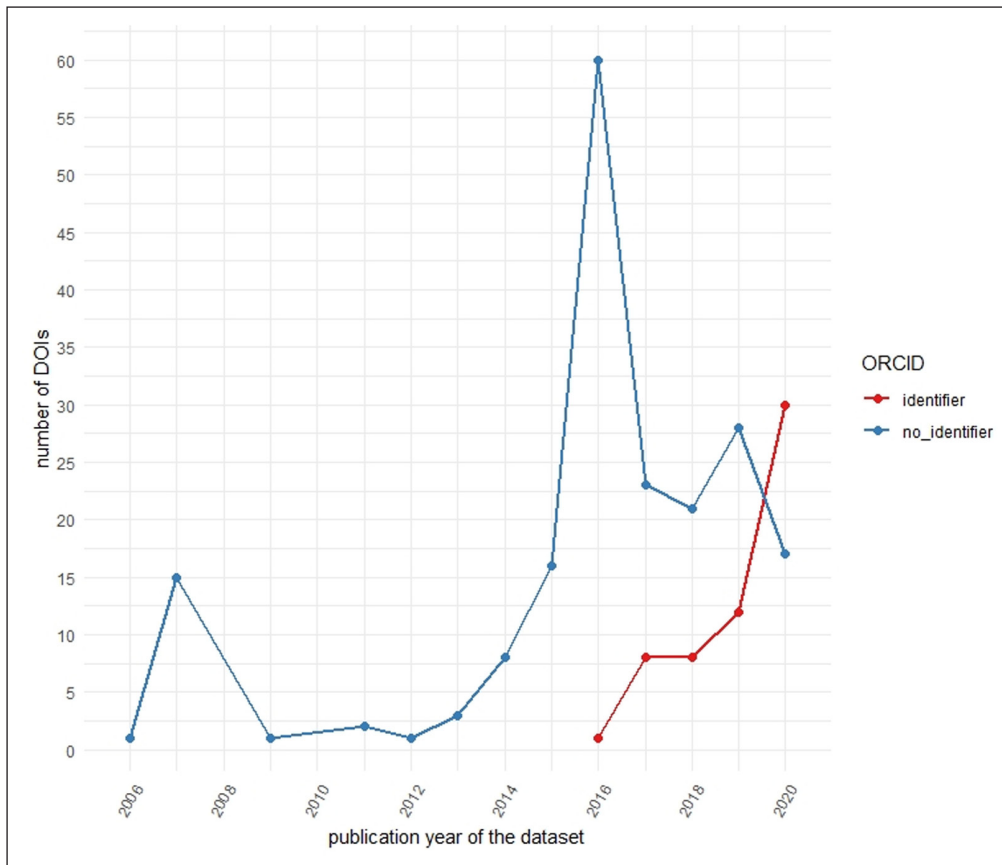


Figure 2 Do researchers add their ORCID at the moment of data archiving? An overview for the period 2006–2020.

increase in ORCID uptake observed for the publication year 2020 in [Figure 2](#). Recently, a growth trajectory has been established to enhance ORCID coverage among Flemish researchers even further (cf. FOSB 2020).

It can be hypothesized that some repositories offer more facilities to include ORCID in the metadata than others. Consequently, it is interesting to check uptake of ORCID encoding per archiving organization ('data publisher'), as visualized in [Figure 3](#).

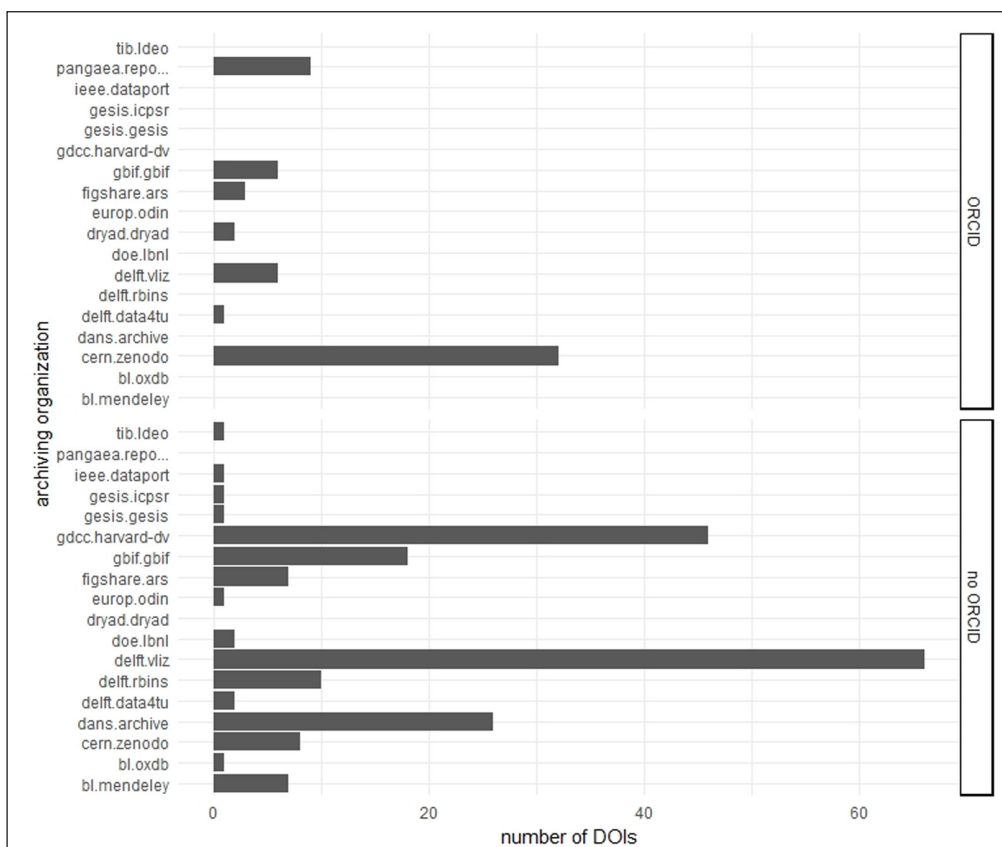


Figure 3 Overview of ORCID registration per archiving organization (operationalized as Client IDs).

Only a few repositories such as Zenodo, PANGAEA, GBIF, Marine Data Archive and Figshare account for most of the DOIs for which ORCID is available in the metadata. Of these repositories, only three archiving organizations, i.e. Zenodo, PANGAEA and Dryad, have more DOIs with ORCID registration than cases for which ORCID registration is absent.

4.3 DETECTION OF RELATED RESEARCH OUTPUTS VIA DATACITE DATASET METADATA

This section aims to assess in more detail to what extent dataset DOIs are linked to related research outputs in the metadata submitted at the different data repositories. Of course, our data only allow for an assessment that takes into consideration the links that are actually established. On the basis of the attested links, it is not possible to determine how many links between datasets and other research outputs are currently lacking in the big metadata hubs: it goes without saying that their absence is nowhere registered. As such, this type of analysis is beyond the scope of our inquiry.

Figure 4 below gives an overview of the number of dataset records that have and do not have related research outputs registered in their metadata, per data repository. Importantly, these related research outputs are not limited to associated publications, but can also correspond to other DataCite relation types, establishing for example relationships with other versions of the dataset. The top three repositories for which the DOIs in our data have links to related research outputs, are Zenodo, Marine Data Archive and PANGAEA. Unsurprisingly, these repositories also performed relatively well on the level of ORCID registration in Section 4.2.2.

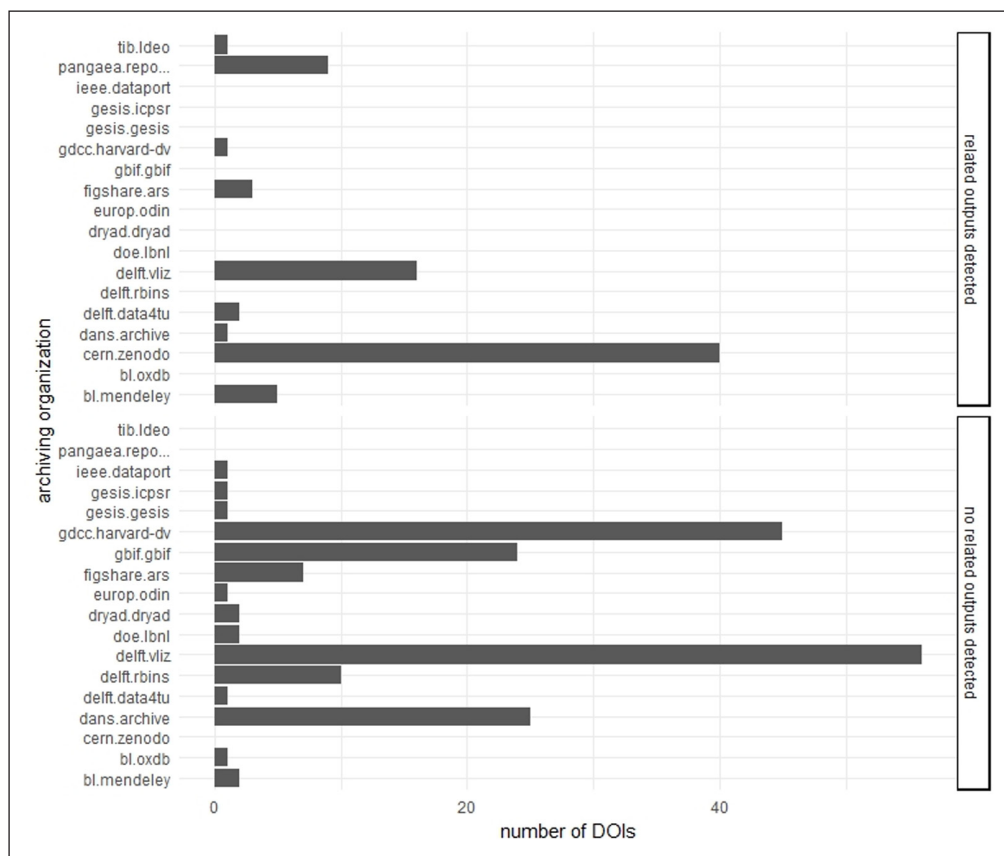


Figure 4 Detection of related outputs for different archiving organizations.

Thanks to the complementary information obtained via the Scholexplorer API, it is also possible to determine the number of DOIs in our data for which associated literature is registered in the Scholix framework. In this way, it becomes straightforward to isolate the related outputs that specifically correspond to literature, excluding other types of research output, which is not readily possible via DataCite. In total, Scholexplorer detected related literature for only 14 dataset DOIs (archived in Harvard Dataverse, Marine Data Archive, PANGAEA, Zenodo, Figshare, 4TU.Centre for Research Data and the Coherent X-ray Imaging Data Bank). Interestingly, for half of these DOIs, the DataCite metadata did not include the link to the related literature.

In this way, Scholexplorer can definitely complement the information found in the DataCite metadata. Moreover, all the related literature found by Scholexplorer for datasets archived in Harvard Dataverse are not available in the DataCite metadata. Since the 'related publication'-field is readily available in Harvard Dataverse, as demonstrated in [Figure 5](#), it seems surprising that this information is not registered with DataCite.

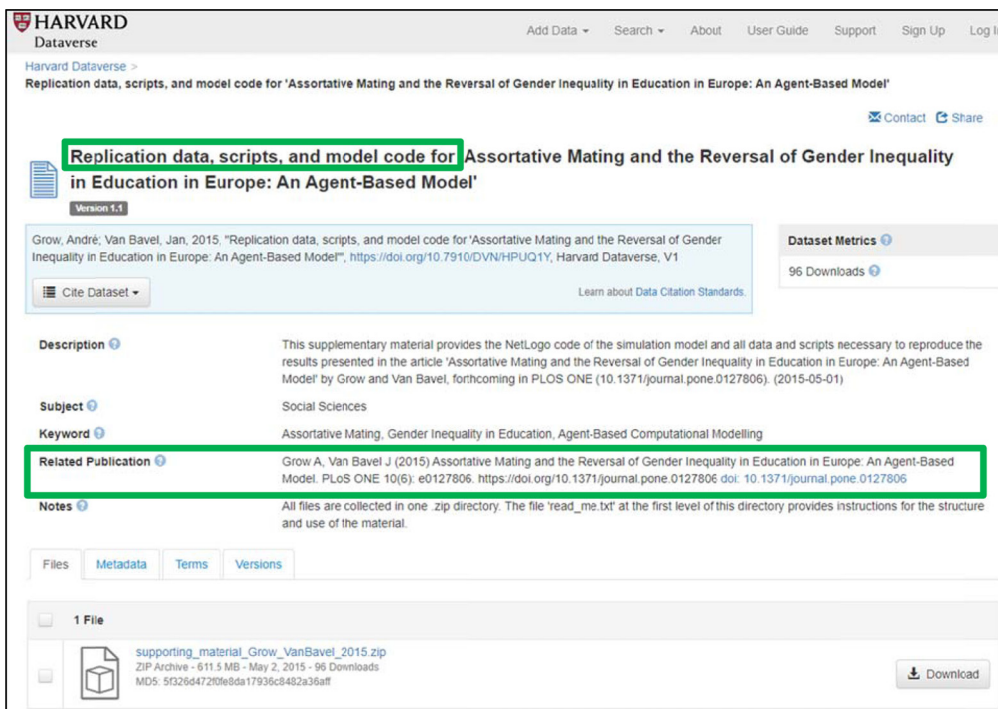


Figure 5 Example dataset on the Harvard Dataverse portal.

Furthermore, in the case of Harvard Dataverse, the (free text) title field of the metadata pertaining to the archived dataset (cf. 'Replication data, scripts and model code for:') also often incorporates the title of the corresponding publication, which further enhances the findability of 'publication - associated data' pairings. In sum, it is important that data repositories offer a built-in, machine-readable solution to link to related publications. If this possibility exists at the level of the data repository, the link can be registered by Scholix.

4.4 FOLLOW-UP ANALYSIS BASED ON A RANDOM SAMPLE OF DATACITE DATASET METADATA

The previous sections established the current state of affiliation information completeness in DataCite dataset metadata for the five Flemish universities. However, these insights are based on a relatively small sample size per archiving organization. The goal of this final results section is to offer a more general outlook on the issue of affiliation information completeness in DataCite metadata, beyond the Flemish research data landscape. This will enable us to draw conclusions about the performance of the different data repositories that are better generalizable, although it has to be stressed that this analysis is still very exploratory and remains to be confirmed in more large-scale studies.

Evidently, if certain archiving organizations underperform with regard to others, this can be due to their technical infrastructure and the way they capture metadata, or it can be a random side effect caused by individual data depositors who happen to provide metadata of poor quality. Of course, only the former is of interest here. In spite of the possible confounding factor of random data depositor negligence, relative underperformance of certain data repositories in comparison with others can be a useful indicator of a more structural problem, to be examined more closely in future research.

As stated in the methodology section above (cf. Section 3), the analysis developed here is based on a random sample of 450 DOIs per data repository, operationalized as Client IDs. The list of repositories is limited to those for which the minimum sample size of 450 DOIs was attained. This cut-off point of 450 DOIs is arbitrarily chosen and allows to examine and compare the

following repositories more closely: ‘dryad.dryad’, ‘figshare.ars’, ‘bl.mendeley’, ‘cern.zenodo’, ‘gesis.icpsr’, ‘delft.data4tu’, ‘ieee.dataport’, ‘pangaea.repository’ and ‘gdcc.harvard-dv’ (the repositories are named according to their Client IDs).¹⁹

In order to effectively summarize the profiles of the different repositories, a heatmap is used (Kolde 2019). The heatmap is visualized below in **Figure 6** and contains values between 0 and 1 for different variables (= columns), per data repository (= rows). These values between 0 and 1 indicate the proportion that a certain variable represents for each data repository. As such, the values sum up to 1 within each row (= data repository). The lower the numerical value, the more the associated colour evolves in the direction of dark blue. Conversely, the higher the numerical value, the more the associated colour evolves in the direction of dark red. Each variable (= columns) is a combination of a yes- or no-value for the following three parameters, in the specified order:

1. Is there an ORCID available in the metadata field ‘identifier of the data creator’?
2. Is there affiliation information available in the dedicated metadata field ‘affiliation of the data creator’?
3. Is there an identifier referring to a related research output available in the metadata, for which the type of relationship between the dataset and the related output involves one of the following strings: ‘supplement’, ‘cite’, ‘reference’ or ‘document’ (e.g. the DataCite relation types ‘IsSupplementTo’, ‘IsDocumentedBy’ etc.)? These four strings are chosen because the relation types that contain these strings seem to have a high likelihood of being used in cases of dataset-publication pairings, which is our main topic of interest here.²⁰ This also implies that both the absence of an identifier referring to a related research output as the presence of a linked identifier for which the relation type does not involve one of the aforementioned strings entail a no-value for this parameter. In this way, relation types such as ‘IsVersionOf’ or ‘IsPartOf’ do not lead to a positive value for this parameter.

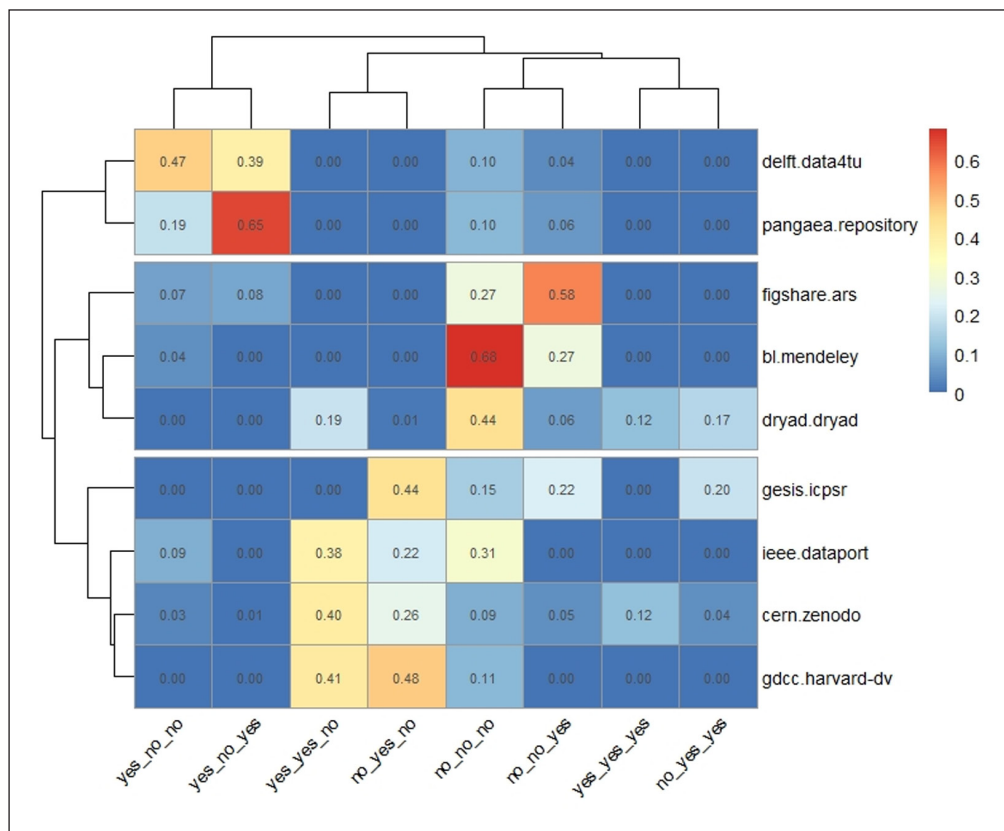


Figure 6 Heatmap of data repositories.

¹⁹ This means that the following archiving organizations are excluded from this analysis: ‘tib.ldeo’, ‘gesis.gesis’, ‘bl.oxdb’, ‘dans.archive’, ‘delft.vliz’, ‘europ.odin’, ‘gbif.gbif’, ‘delft.rbins’ and ‘doe.lbnl’.

²⁰ The ‘IsSupplementTo’ relation type seems to be the default option in these cases. However, it is to be expected that some variation exists in the type of link used, especially because researchers often have minimal experience with the differences between the various relation types.

For example, the variable 'yes_yes_no' means that the answer to the first two questions is affirmative, but the answer to the last question is negative.

Furthermore, the rows and columns are clustered according to a hierarchical agglomerative clustering procedure (Kaufman & Rousseeuw 1990). Consequently, data repositories and variables with similar values are grouped together on the plot. This clustering process is also made visible in the graph via the dendrograms that are added on the left and top side of the heat map. Of course, similarity is always relative, as the cluster analysis minimizes within-group variance and maximizes between-group variance. Cluster analysis requires to select a distance measure and a clustering method. The 'ward.D2' clustering method was chosen for the current analysis. The Euclidean distance was used for the distance between the rows, whereas the Canberra distance was used to measure the distance between the columns. On the basis of the average silhouette width criterion, it was determined that a three-cluster solution (with the highest average silhouette width, = 0.36) is optimal for the rows.

As a first observation, it is noteworthy that every data repository is characterized by a low proportion for the combination of three yes-values, which is arguably the best constellation from the perspective of metadata quality. Zenodo and Dryad obtain the highest proportion for 'yes_yes_yes' (0.12). At the other end of the spectrum, Mendeley and Dryad are characterized by the highest proportions for the combination of three no-values (0.68 and 0.44, respectively). Since Dryad has extreme values for both ends of the continuum, it can be hypothesized that the degree to which groups of researchers use (or do not use) the metadata facilities provided by the repository obviously has an impact on the reported values.

Let us now take a closer look at the different clusters. The cluster containing 'gesis.icpsr', 'ieee.dataport', 'cern.zenodo' and 'gdcc.harvard_dv' is characterized by relatively high values for 'yes_yes_no' (except 'gesis.icpsr') and 'no_yes_no'. Put differently, affiliation information is, comparatively speaking, often available in the dedicated metadata field, as well as ORCID information, but information about related publications seems rather incomplete.²¹ In the case of Harvard Dataverse, this might also be due to the fact that DataCite does not always seem to register related publication information available at the Harvard Dataverse repository (see Section 4.3). This problem might be affecting other repositories as well.

Another interesting cluster is the one comprising 'delft.data4tu' and 'pangaea.repository'. Both repositories score high on the 'yes_no_yes' variable: only affiliation information in the dedicated metadata field is lacking. The absence of affiliation information in the dedicated metadata field seems systemic for these two repositories, because no other combination with an affirmative value for the second parameter is attested. Moreover, 'delft.data4tu' also obtains a high value for the 'yes_no_no' combination, where related publication information is also unattested.

Next, the cluster containing 'figshare.ars', 'bl.mendeley' and 'dryad.dryad' has high values for the 'no_no_no' variable, as already stated above. However, the cluster is also characterized by relatively high values for the 'no_no_yes' variable, in particular 'figshare.ars' and 'bl.mendeley'. Clearly, these two repositories do rather well on the level of interlinking between data and publications,²² although data creator identifiers and affiliation information in a dedicated metadata field seem missing. Overall, it can be concluded that repositories can differ quite substantially when it comes to their characteristics.

Finally, [Figure 7](#) below zooms in on the third parameter about identifiers of related output, excluding the other two parameters from the equation. The mosaic plot (Friendly 2002) demonstrates how data repositories diverge with regard to the third parameter. This plot can be read as follows. The rectangles in the mosaic plot correspond to the proportions of the respective categories. Those that are coloured in red/blue correspond to categories that are overrepresented/underrepresented, respectively.

²¹ Since 'gesis.icpsr' has relatively high values for the variables 'no_no_yes' and 'no_yes_yes', which hints at the availability of information about related publications, the main shortcoming of this repository seems to correspond to lacking ORCID information.

²² Note that we accidentally discovered that, for this version DOI '10.17632/CRMFJDGYR6.1', a related publication is encoded in the DataCite metadata, but this is not the case for the overarching concept DOI '10.17632/CRMFJDGYR6' (Mendeley). More research is probably needed to unravel these peculiarities.

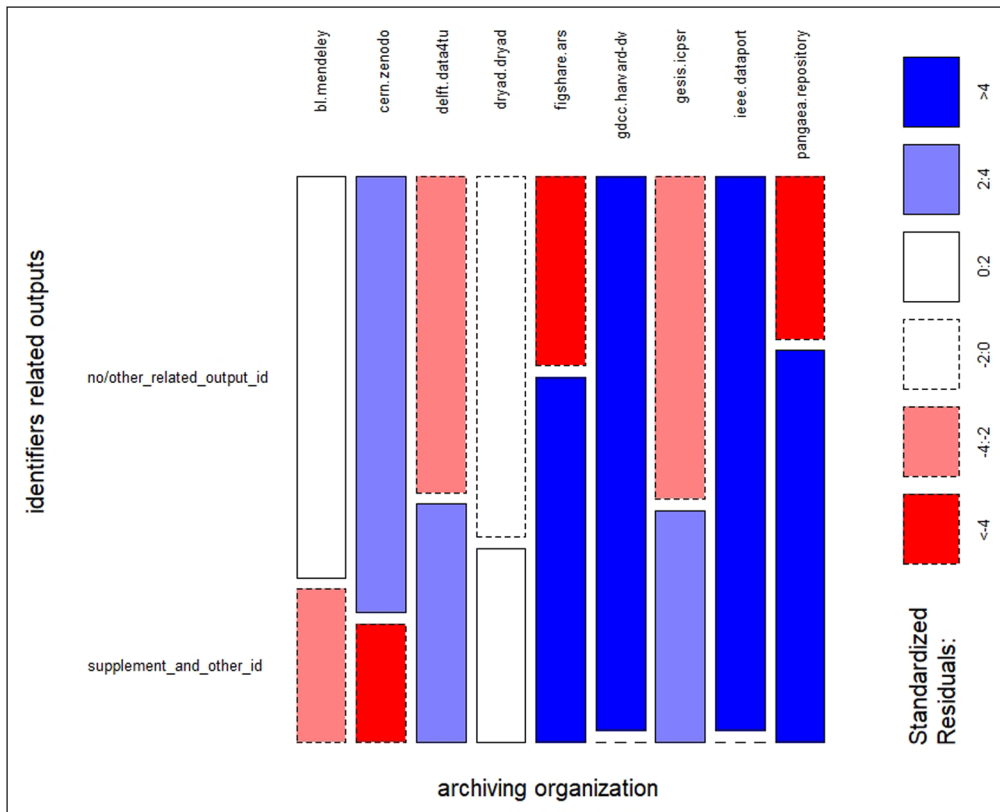


Figure 7 Mosaic plot cross-tabulating archiving organizations with the third parameter about identifiers of related output.

From this plot, it can be deduced that a repository such as PANGAEA has a high share of relation types involving the strings ‘supplement’, ‘cite’, ‘reference’ and ‘document’, whereas this category is clearly underrepresented in a repository such as Zenodo. These differences between repositories might be explained by the different ways they implement interlinking between data and publications. For example, PANGAEA adds the reference to the associated publication just below the full reference to the dataset situated at the top of the web page, as shown in [Figure 8](#).



Figure 8 Example dataset with associated publication on PANGAEA.

In contrast, the linked publication receives a less prominent place on Zenodo, where it is relegated to the right-most column of the web page, under the general headings ‘Published in’ and ‘Related identifiers: Supplement to’, as shown in [Figure 9](#). It is not inconceivable that this less visually noticeable place in the overall web page design causes researchers to be less aware of the need to include information about related publications. The impression that this information does not really stand out on Zenodo seems corroborated by the fact that the creator of the dataset referenced to in [Figure 9](#) apparently felt the need to repeat the full citation of the associated publication in the free-text field at the center of the Zenodo web page. Further research is needed.

Dataset Open Access

Ice sheet model initialisation Greenland: an ISMIP6 intercomparison

255 views 445 downloads
[See more details...](#)

Indexed in
OpenAIRE

Publication date:
May 25, 2018
DOI: [10.5281/zenodo.1173088](https://doi.org/10.5281/zenodo.1173088)

Keyword(s):
Model output ISMIP6 intercomparison
Ice sheet model Greenland ice sheet initMIP

Published in:
The Cryosphere: 12 pp. 1433-1460.
Related identifiers:
Supplement to
[10.5194/tc-12-1433-2018](https://doi.org/10.5194/tc-12-1433-2018)

License (for files):
[Creative Commons Attribution Non Commercial 4.0 International](https://creativecommons.org/licenses/by-nc/4.0/)

Output produced as part of the publication "Design and results Greenland: an ISMIP6 intercomparison", published in The Cryosphere

Authors: A. Aschwanden, A. Calov, R. Gagliardini, O. Gillet-Chaulet, F. Goussais, P. Kennedy, J. H. Larour, E. Lipscomb, W. H. Le clec'h, S. M. M. Rückamp, M. Saito, F. Schlegel, N. Seroussi, H. Shepherd, A. Shepherd, A. Siemen

Find it here:
ISMIP6
MIP-Greenland

Part of the data.
Ongoing support of CMIP, users are also obligated to join groups.

Figure 9 Example dataset with associated publication on Zenodo.

5 CONCLUSIONS AND RECOMMENDATIONS

At the end of our inquiry, the following conclusions and recommendations can be established:

- Affiliation information is currently encoded in different metadata fields by researchers. Researchers should be prompted to use the dedicated affiliation field in order to register affiliation information (and not the name field etc.). Our analysis shows that there is some progress with regard to this issue in recent years, because researchers are using the affiliation metadata field more and more to encode this information. At the same time, repositories should enable this practice by providing a dedicated affiliation field in their submission template, ideally via a dropdown menu from which researchers can choose their research institution. In order to improve machine-readability and consistency, organizational identifiers such as GRID or ROR could be implemented (Hahnel & Valen 2020), following the example of the Dryad repository (Lammey 2020).
- Data repositories more and more capture ORCID information from data depositors. However, there still seem to be important discrepancies between repositories. Ideally, ORCID registration at the moment of data submission is expanded in order to become the default for every repository. Ideally, every data creator is accompanied by an ORCID in the metadata.
- Certain links between data and associated publications go undetected in the DataCite metadata, but are established in the Scholix framework. This seems especially true for Harvard Dataverse repositories. Moreover, there seems to be considerable variation between data repositories concerning the availability of links between data and related publications. PANGAEA provides researchers with a specific and prominent field dedicated to related publications in their metadata structure. Unsurprisingly, this improves linking between publications and research data. Other repositories such as Zenodo seem more proficient at connecting related datasets and/or different versions of the same dataset.

In general, it is crucial that awareness is raised among researchers about the importance of interlinking, both between different research outputs as between research outputs and

research institutions. Recent popularization of the FAIR principles definitely contributes to this, but more targeted action may be needed. At the same time, researchers need to be informed about the technical limitations of repositories with regard to interlinking performance, so that they can make a well-founded choice concerning which archiving organization to choose for their research data. The recommendations for data repositories that are listed above could be integrated into certification criteria (cf. CoreTrustSeal) and minimum metadata requirements defined by metadata hubs (cf. DataCite and OpenAire) or EOSC (cf. the EDMI²³-initiative or the 'Descriptive Core Metadata'-metric as suggested by the FAIRsFAIR project), in order to increase their uptake among data repositories.

DATA ACCESSIBILITY STATEMENT

Van Wetteere, N. 2021. Replication Data for: "Affiliation Information in DataCite Dataset Metadata: a Flemish Case Study" (Version 01) [Data set]. Zenodo. DOI: <http://doi.org/10.5281/zenodo.4582681>.


ACKNOWLEDGEMENTS

I would like to thank the two anonymous reviewers for their suggestions and comments, as well as my colleagues at the VUB R&D department. Any remaining errors are my own responsibility.

COMPETING INTERESTS

The author has no competing interests to declare.

AUTHOR AFFILIATION

Niek Van Wetteere  orcid.org/0000-0002-9455-368X
Vrije Universiteit Brussel, BE

REFERENCES

- Burton, A, Koers, H, Manghi, P, La Bruzzo, S, Aryani, A, Diepenbroek, M and Schindler, U.** 2017. The Data-Literature Interlinking Service: towards a Common Infrastructure for Sharing Data-Article Links. *Program*, 51(1): 75–100. DOI: <https://doi.org/10.1108/PROG-06-2016-0048>
- Chamberlain, S.** 2020. rdatacite: Client for the 'DataCite' API. R package version 0.5.0. Available at <https://CRAN.R-project.org/package=rdatacite>.
- Cousijn, H, Feeney, P, Lowenberg, D, Presani, E and Simons, N.** 2019. Bringing Citations and Usage Metrics Together to Make Data Count. *Data Science Journal*, 18(9): 1–7. DOI: <https://doi.org/10.5334/dsj-2019-009>
- DataCite Metadata Working Group.** 2019. DataCite Metadata Schema Documentation for the Publication and Citation of Research Data. Version 4.3. DOI: <https://doi.org/10.14454/7xq3-zf69>
- Flemish Open Science Board.** 2020. KPIs for Open Science in Flanders. Available at <https://www.ewi-vlaanderen.be/sites/default/files/bestanden/5fc5f512b328e9000c0007f3.pdf>.
- Friendly, M.** 2002. A Brief History of the Mosaic Display. *Journal of Computational and Graphical Statistics*, 11(1): 89–107. DOI: <https://doi.org/10.1198/106186002317375631>
- Habermann, T.** 2019. MetaDIG recommendations for FAIR DataCite metadata. *DataCite blog*. DOI: <https://doi.org/10.5438/2chg-b074>
- Habermann, T and Lowerberg, D.** 2019. Missing Data for Data – Our Quest to Clean Up Institutional Affiliations in Dryad: Wrangling RORs. Presented at the csv,conf,V4, Portland, OR, USA. *Zenodo*. DOI: <http://doi.org/10.5281/zenodo.3254969>
- Hahnel, M and Valen, D.** 2020. How to (Easily) Extend the FAIRness of Existing Repositories. *Data Intelligence*, 2(1–2): 192–198. DOI: https://doi.org/10.1162/dint_a_00041
- Hook, DW, Porter, SJ and Herzog, C.** 2018. Dimensions: Building Context for Search and Evaluation. *Frontiers in Research Metrics and Analytics*, 3(23). DOI: <https://doi.org/10.3389/frma.2018.00023>
- Kaufman, L and Rousseeuw, PJ.** 1990. *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: John Wiley & Sons Inc. DOI: <https://doi.org/10.1002/9780470316801>

23 "EOSC Datasets Minimum Information".

- Khan, N, Pink, CJ and Thelwall, M.** 2020. Identifying Data Sharing and Reuse with Scholix: Potentials and Limitations. *Patterns*, 1(1). DOI: <https://doi.org/10.1016/j.patter.2020.100007>
- Kolde, R.** 2019. pheatmap: Pretty Heatmaps. R package version 1.0.12. Available at <https://CRAN.R-project.org/package=pheatmap>.
- Lammey, R.** 2020. Solutions for identification problems: a look at the Research Organization Registry. *Sci Ed*, 7(1): 65–69. DOI: <https://doi.org/10.6087/kcse.192>
- Wu, M, Psomopoulos, F, Khalsa, SJ and de Waard, A.** 2019. Data Discovery Paradigms: User Requirements and Recommendations for Data Repositories. *Data Science Journal*, 18(3): 1–13. DOI: <https://doi.org/10.5334/dsj-2019-003>

Van Wettere
Data Science Journal
DOI: 10.5334/dsj-2021-013

18

TO CITE THIS ARTICLE:

Van Wettere N. 2021. Affiliation Information in DataCite Dataset Metadata: a Flemish Case Study. *Data Science Journal*, 20: 13, pp. 1–18. DOI: <https://doi.org/10.5334/dsj-2021-013>

Submitted: 01 October 2020

Accepted: 24 February 2021

Published: 29 March 2021

COPYRIGHT:

© 2021 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

Data Science Journal is a peer-reviewed open access journal published by Ubiquity Press.

