RESEARCH PAPER

Improving Domain Repository Connectivity

Ted Habermann[†]

Metadata Game Changers

Keywords: Persistent Identifier; Domain Repository; ORCID; ROR; Connectivity; PID Graph

Citation: Habermann, T. Improving Domain Repository Connectivity. Data Intelligence 5(1), 6-26 (2023). doi: 10.1162/dint_a_00120

Received: October 7, 2021; Revised: February 5, 2022; Accepted: February 7, 2022

ABSTRACT

Domain repositories, i.e. repositories that store, manage, and persist data pertaining to a specific scientific domain, are common and growing in the research landscape. Many of these repositories develop close, long-term communities made up of individuals and organizations that collect, analyze, and publish results based on the data in the repositories. Connections between these datasets, papers, people, and organizations are an important part of the knowledge infrastructure surrounding the repository.

All these research objects, people, and organizations can now be identified using various unique and persistent identifiers (PIDs) and it is possible for domain repositories to build on their existing communities to facilitate and accelerate the identifier adoption process. As community members contribute to multiple datasets and articles, identifiers for them, once found, can be used multiple times.

We explore this idea by defining a connectivity metric and applying it to datasets collected and papers published by members of the UNAVCO community. Finding identifiers in DataCite and Crossref metadata and spreading those identifiers through the UNAVCO DataCite metadata can increase connectivity from less than 10% to close to 50% for people and organizations.

1. INTRODUCTION

For many years repositories have focused on discovery systems based on a small number of "discovery metadata" elements, mostly text fields (title, author, abstract, keywords), and extents (spatial and temporal) [1, 2, 3]. The emergence and widescale adoption of identifiers for people, organizations, and research objects is reviving a powerful tool from the traditional discovery arsenal: connections. Persistent

[†] Corresponding author: Ted Habermann (E-mail: ted.habermann@gmail.com; ORCID: 0000-0003-3585-6733).

identifiers (PIDs) are the keys that enable these connections and the PID Graph [4] is a visualization and search tool built using them. To take advantage of this development, repositories must create identifiers for resource objects, find identifiers for the people and organizations that produce these objects, and add these identifiers to metadata records. The community is currently in the nascent stages of this adoption process.

UNAVCO [5] is an excellent example of a domain repository that has developed close, long-term relationships with their community. UNAVCO supports instruments, data, and enginering for terrestrial and satellite geodetic technologies; GPS networks for Earth, atmospheric, and polar science applications; and the Global Navigation Satellite System (GNSS). Datasets and products provided or enabled by UNAVCO span the fields of seismology, hydrology, glaciology, geomorphology, geology, atmospheric sciences, data science, and others. UNAVCO has always been an integral part of the geodetic community support system. Their role extends from proposal planning and writing, through project initiation and implementation, data collection, management, and archive, to publication of results and access to data by other community members. Scientists, engineers, logistics specialists, data managers, software developers, and educators are all part of this community.

The UNAVCO Community described the responsibilities of participants in open science communities during 2012 [6] and developed an open data policy [7] based on those responsibilities. These responsibilities included assigning PIDs to datasets and using these PIDs to cite those data from papers, that is, establishing an important element of the PID Graph: connections between papers and data.

As the breadth of identifiers and connections continues to expand, there are clear parallels between the strong connections between real people and organizations in the UNAVCO Community and connections between these entities in the PID Graph. Can the multitudinous real-world connections help us populate identifiers in the metadata and related connections in the PID Graph? This paper addresses this question using methods for mining existing metadata in DataCite and Crossref for Open Researcher Contributor IDs (ORCIDs) [8] and Research Organization Registry Ids (RORs) [9] and the application of the results to UNAVCO metadata in DataCite. The results indicate that community connections can be very helpful in the identifier adoption process.

2. DATA

This work is based on two datasets: 1) metadata for datasets registered by UNAVCO in DataCite [10], 2) Crossref metadata for a collection of papers that used UNAVCO data provided on the UNAVCO website [11], retrieved using DataCite [12] and Crossref APIs [13].

2.1 UNAVCO Datasets in DataCite

UNAVCO has minted over 5000 dataset DOIs with DataCite between 2013 and 2021 (Figure 1). UNAVCO maintains an in-house archive of these datasets with extensive metadata for discovery, access,

and understanding, so the primary role of the DataCite repository is minting DOIs for identification and citation of the datasets.

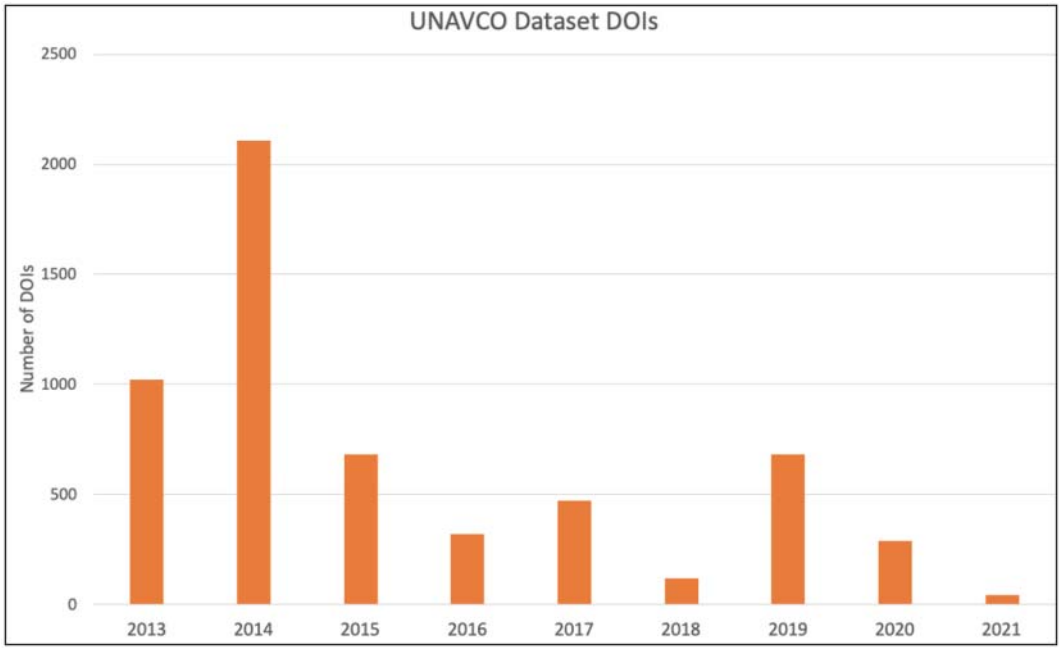

Figure 1. The number of UNAVCO datasets registered in DataCite per year, as of November 2020.

2.2 Papers Using UNAVCO Datasets

Like many domain repositories, UNAVCO keeps track of papers that are published based on repository data. Dataset DOIs and clear citation guidelines [14] both make it easier to do this tracking. The list of UNAVCO community publications available on the website includes 1569 articles published between 2003 and 2018. This is a rich source of identifiers (ORCIDs) for community members and affiliations (leading to RORs) that were not included in the original DataCite metadata.

3. METHODOLOGY

3.1 Measuring Connectivity

Connectivity measures how well research objects or collections of research objects are connected to the global research web, represented by the PID Graph. These connections depend on identifiers for all kinds of research objects. Here I focus on people, identified by ORCIDs [8], and organizations, identified by RORs [9]. All people and organizations in the UNAVCO DataCite metadata are listed as creators rather than contributors, so connectivity for people and organizations in other roles cannot be examined.

Connectivity can be quantified for any item or collection of items that can have identifiers. It is the number of existing identifiers divided by the number of possible identifiers. If no identifiers are present, connectivity = 0. If all potential identifiers are present, connectivity = 1.

The example Figure 2 shows a resource that has two authors. In this case the identifiers are ORCIDs, and connectivity can be 0 (no ORCIDs), 0.5 (1 ORCID), or 1 (2 ORCIDs).

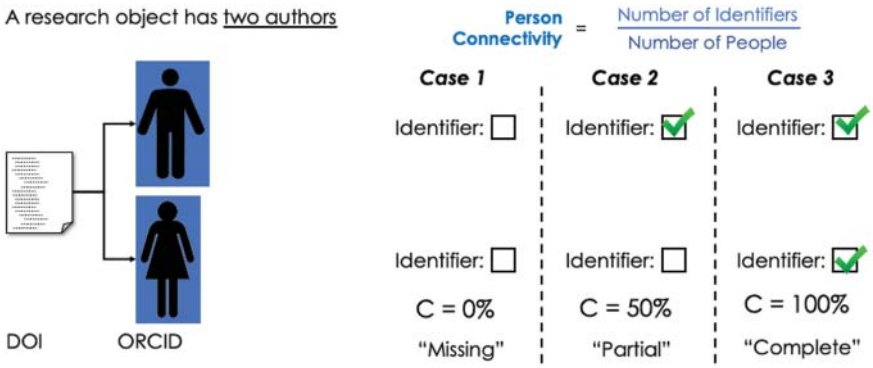

Figure 2. Quantifying connectivity for a resource with two authors.

The calculation is similar for a resource that has two affiliations (Figure 3). In this case, the identifiers are RORs and the connectivity can be 0 (no RORs), 0.5 (1 ROR), or 1 (2 RORs)

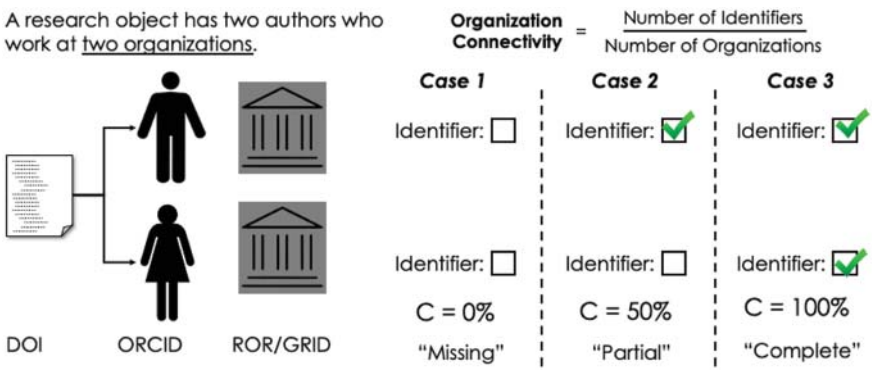

Figure 3. Quantifying connectivity for a resource with two organizations.

These calculations scale up easily to complete collections of resources. In those cases, the total number of possible person identifiers is generally the total number of authors across all resources and the total number of possible organizational identifiers is the total number of affiliations of those authors which is typically greater than the number of authors because of multiple affiliations per author.

3.2 Visualizing Connectivity

The goal is to understand how to improve connectivity in domain repositories and to measure connectivity as a metric for showing progress as connectivity improves. Describing connectivity in pictures can facilitate this process. This can be done using a horizontal bar which represents the entire collection and color, with green sections on the left side of the bar for items that have complete connectivity, yellow sections in the middle of the bar for items that have partial connectivity, and red sections on the right side of the bar for items that have no connectivity (Figure 4).

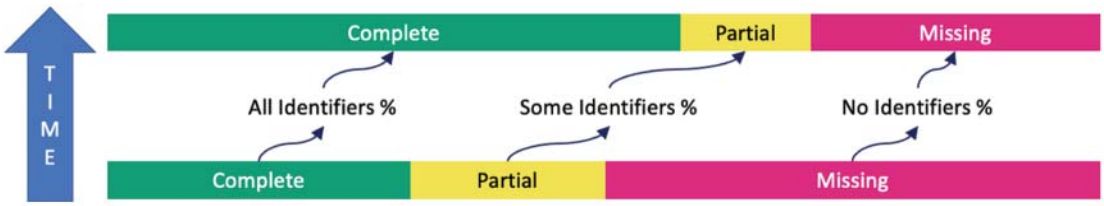

Figure 4. Visualizing connectivity as the % of items with all identifiers ("complete", green), with some identifiers ("partial", yellow), and with no identifiers ("missing", red).

The desired end state for connectivity is maximizing the % of the collection that has complete connectivity, so improvements make the green part of the bar larger and the yellow and red parts of the bar smaller, illustrated in Figure 4 by the change between the lower and upper bars.

3.3 Spreading Identifiers

The proposal underlying this work is that people and organizations in domain repository communities make multiple contributions to the community in the form of data and results published based on them. If this is true and, if identifiers can be found for one contribution, those identifiers can be added to other contributions in the repository, i.e. "spread" across the repository.

The process of spreading ORCIDs across a repository can be done with confidence as the association between a person and their ORCID is one-to-one. In other words, there is high confidence in the assertion of the connection between the person and the ORCID. In the affiliation case, the confidence is not as high, as authors can readily switch organizations.

This situation is illustrated in Figure 5. This author has authored nine datasets and ORCIDs are included in three of them. In this case, the initial ORCID connectivity for this author is 33%. The connectivity is increased to 100% by spreading the author's ORCID to the six datasets that originally had no ORCIDs, indicated by the grey arrows.

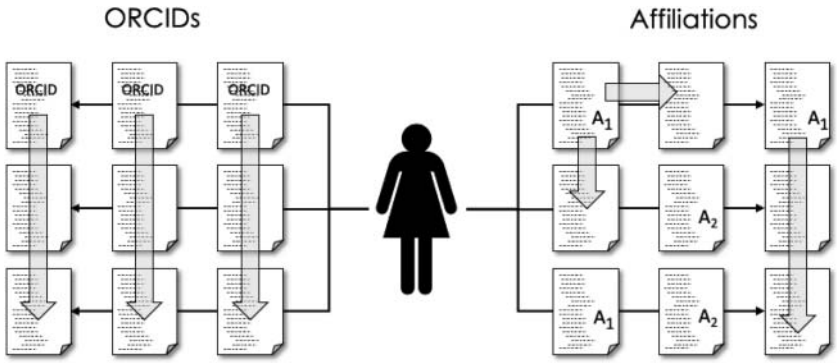

Figure 5. Spreading ORCIDs and affiliations.

The same datasets are shown again on the right side of the Figure along with affiliations. In this case affiliation A_1 occurs three times and A_2 occurs two times and there are four papers without affiliations. Can either affiliation be added to the other four datasets? There is no solution here that has 100% confidence. The rules used here to spread affiliations as shown by the grey arrows were as follows:

1. if only one affiliation exists, use it in all papers
2. if more than one affiliation exists, use the most common one
3. if two affiliations exist and occur an equal number of times, use both.

Affiliations identified using these rules can flagged for evaluation by the author, or by a community member that is familiar with their affiliation history. So, if the spreading introduces errors, they can be identified and corrected.

3.4 Finding Identifiers (Metadata Archeology)

Extracting identifiers from DataCite metadata for a particular DOI is a straightforward process of retrieving the metadata using the DataCite API [12] and searching the appropriate metadata properties for identifiers. Finding identifiers in a set of citations is more difficult, particularly if, as in the UNAVCO case, the citations do not include identifiers (DOIs).

The first step in finding identifiers from these citations is to find DOIs for the referenced papers. This was done using Google searches for the titles of the papers and searching results for pages with titles matching the title of the papers using Beautiful Soup [15]. If these matches exist, the metadata of the page can be scraped for a <meta tag> with the name "citation_doi" and content which is the DOI for the paper.

An example of this approach is illustrated in Figure 6 for the paper titled "A revised dislocation model of interseismic deformation of the Cascadia subduction zone". In this case, as in many examples, all goes well and the DOI is easily determined from the first link in the Google results.

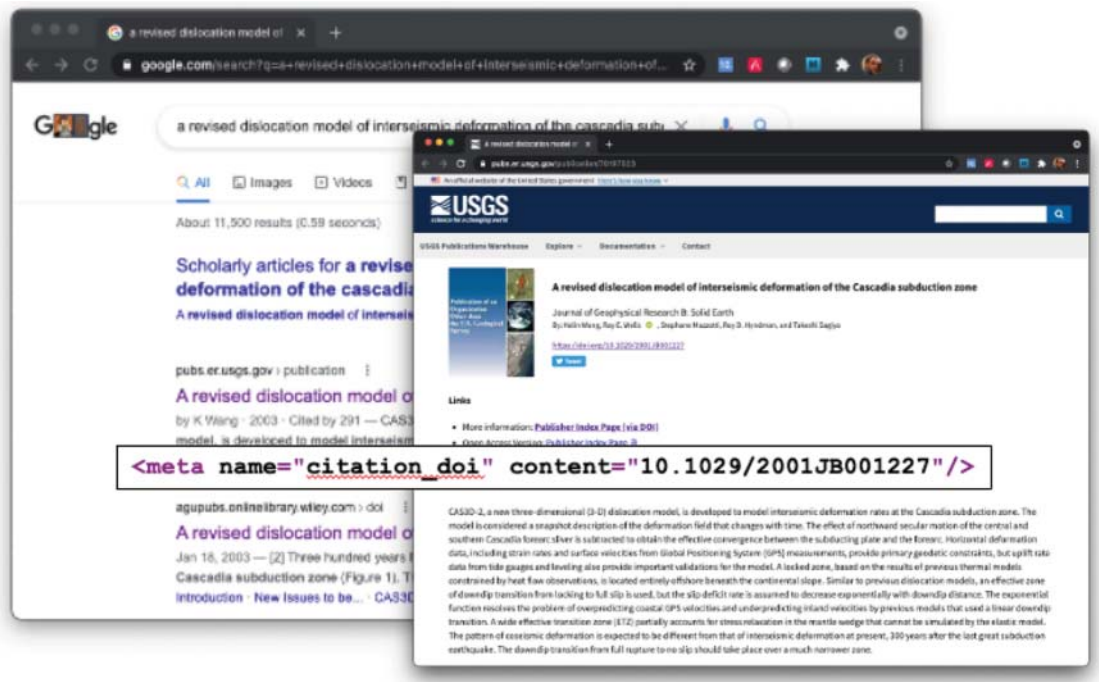

Figure 6. Finding DOIs in journal article web pages.

Once the DOIs are known, two approaches can be used to find ORCIDs and affiliations:

1. search Crossref metadata
2. search and scrape journal web pages.

The first approach is preferred because the Crossref metadata are in a standard, structured representation so retrieving ORCIDs and affiliations from the metadata is straightforward. These standard metadata are an invaluable resource for aspiring metadata archeologists. In contrast, scraping journal web pages is remarkably inconsistent. Affiliations (without identifiers) are many times available in meta tags (citation_author and citation_author_institution). Unfortunately, no citation_author_identifier or citation_author_institution_identifier tags exist. ORCIDs, if available, are many times hidden in non-standard mouseovers or popups or other exotic and non-standard approaches that may work for humans viewing pages but can be difficult to find automatically.

4. RESULTS

The methodology described above was applied to the UNAVCO datasets in a series of steps.

4.1 Establishing a Baseline

DataCite is where the UNAVCO Community connects their data to the broader scientific world through identifiers included in the metadata. Characterizing the current state of ORCIDs and RORs in the collection (the baseline) is the first step in understanding the collection and measuring improvements in the connectivity that might be achieved through time.

4.1.1 ORCID Connectivity Baseline

As described above, connectivity can be measured for any kind of identifiers and for any collection of research objects or other entities. The UNAVCO DataCite metadata include over 5,000 creators and the baseline ORCID connectivity for the UNAVCO DataCite collection is shown in Figure 7. The largest part of the datasets in the collection has no ORCIDs (93%, 5005/5356 DOIs) while 234 (4%) have some ORCIDs, and 117 (2%) have all ORCIDs.

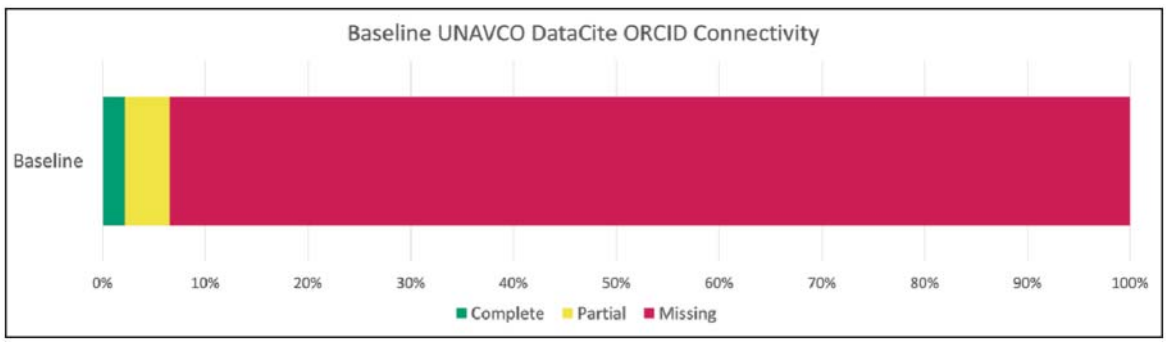

Figure 7. Initial (baseline) connectivity for ORCIDs in UNAVCO DataCite metadata.

While it is important to note that just over 350 datasets have complete or partial ORCID coverage, the overall paucity of ORCIDs in the UNAVCO metadata is not unusual. An assessment of 144 DataCite repositories in the TIB Consortium showed that, on average, less than 15% of the records in these repositories have identifiers [16] and a similar assessment of all Crossref metadata during 2019 showed that the average portion of Crossref records with ORCIDs was less than 10% [17] and it was only during mid-2020 that the average number of ORCID's per article in Crossref passed 2.0 [18]. We are clearly at the beginning of the ORCID adoption process across the scientific publishing world.

The proposal that we are testing is that community members make multiple contributions to these datasets. If that is true, single individuals should occur many times in these metadata. In fact, fifty-three authors have ORCIDs in these metadata that occur in a total of 499 datasets, i.e. an average of 9.4 times. The most common ORCID belongs to a system engineer at UNAVCO whose ORCID occurs 130 times in these data. Seven other ORCIDs occur more than ten times.

4.1.2 Affiliation Connectivity Baseline

The UNAVCO DataCite metadata do not currently include any organizational identifiers, but the metadata do include affiliations (organization names) that can give an idea of the maximum organizational connectivity that can achieve if identifiers can be found for all the affiliations.

Figure 8 shows the baseline affiliation connectivity for UNAVCO metadata at DataCite which is very similar to the data for ORCIDs in Figure 7. In fact, the numbers are slightly better, with 382 records having complete or partial connectivity. Unfortunately, 93% of the records are missing affiliation information.

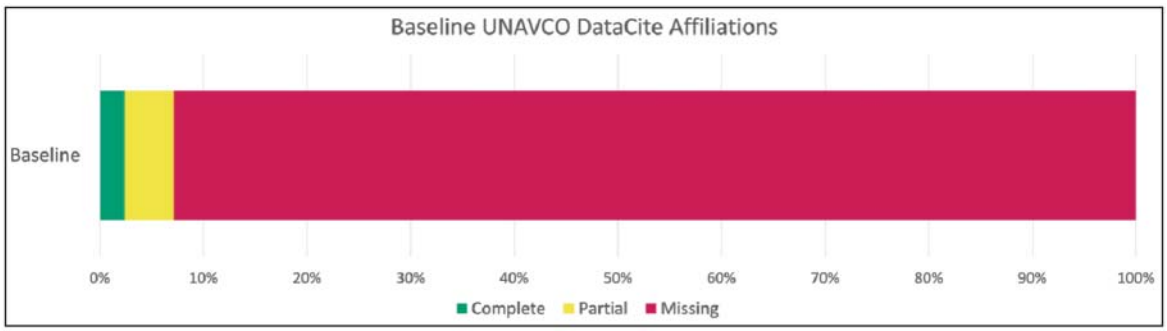

Figure 8. Initial (baseline) connectivity for Affiliations in UNAVCO DataCite metadata.

4.2 Author Connectivity

The observations so far have focused on connectivity for UNAVCO datasets represented by DOIs. As mentioned earlier, connectivity can be calculated for any entity in the PID Graph. The UNAVCO DataCite metadata provide an opportunity to calculate ORCID connectivity for authors in the metadata that have ORCIDs. In this context, authors with complete connectivity have associated ORCIDs in all metadata records where they appear, i.e. they are completely connected. Authors with partial connectivity have ORCIDs only in some of the records where they appear. The records that include these authors but do not have ORCIDs provide an easy and completely safe opportunity to improve connectivity in the metadata by adding know ORCIDs for these authors in records currently missing them, i.e. spreading identifiers as discussed above.

For example, we know that the author with the most common ORCID in these data occurs in 130 datasets. Five other datasets include this author but do not include the ORCID in the metadata, so, this author occurs 135 times and has 130 ORCIDs so connectivity is 130/135 = 96%. We can increase the number of records with ORCIDs by adding the ORCID to five records that are currently missing it. This also increases connectivity to 1, i.e., complete.

Figure 9 shows that 26% of the authors with known ORCIDs have partial connectivity. The total number of datasets authored by these authors without ORCIDs is 182. Adding these to the 499 records with ORCIDs increases the number of ORCIDs in the metadata by 36%.

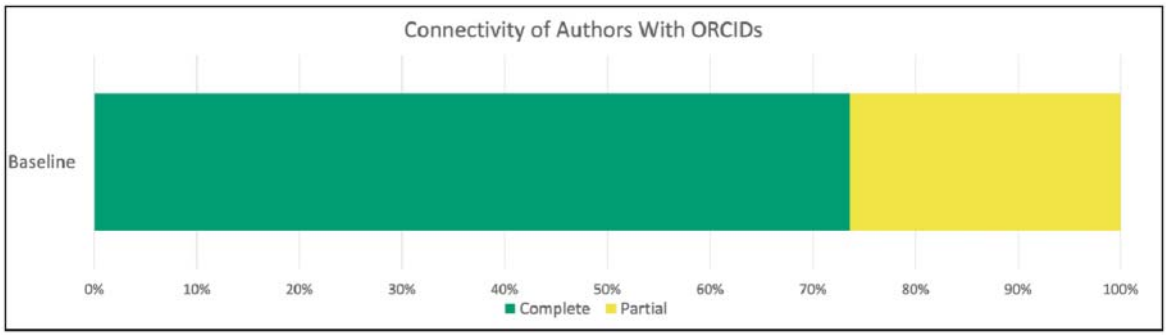

Figure 9. Author connectivity in UNAVCO DataCite metadata.

Figure 10 compares the ORCID connectivity before and after the spreading of ORCIDs. As expected, based on the discussion above, there is a significant improvement. The partial and complete DOIs now make up 14% of the collection as compared to 6% in the initial baseline and the number of records missing ORCIDs decreased by 8%.

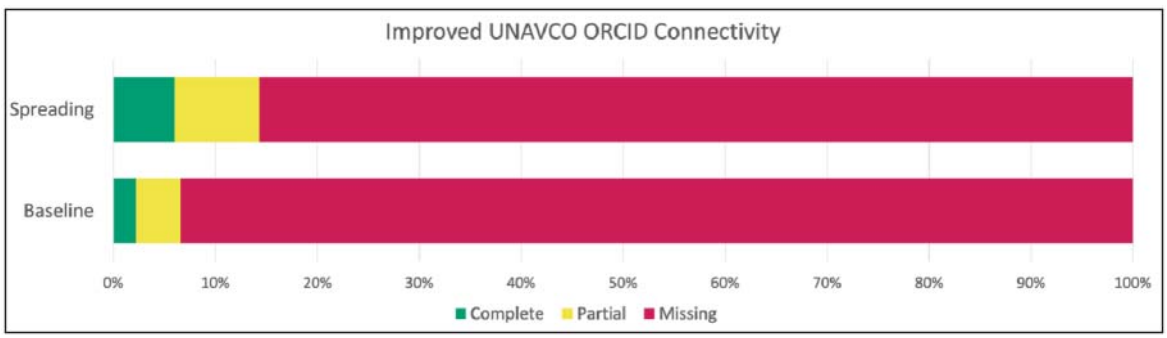

Figure 10. Improvement in ORCID connectivity associated with spreading known ORCIDs to metadata without ORCIDs.

4.3 Identifying Organizations

The UNAVCO DataCite Repository includes datasets created by researchers from many organizations, most of which are members of the tight-knit and well-established UNAVCO community. Most of these organizations have contributed multiple datasets to the community, so they occur multiple times in the metadata. Connecting these organizations to the PID Graph depends on having unique identifiers, i.e., RORs, for these organizations.

Table 1 shows the organizations that occur in the UNAVCO metadata along with RORs found for these organizations and the number of times they occur. Fortunately, RORs were identified for all organizations in the metadata. As expected, a small number of organizations (35) occur many times (2596) in these metadata. The most common organization is UNAVCO itself, which occurs in 1288 records (50% of all occurrences).

Table 1. Organizations found in UNAVCO metadata with RORs and number of occurrences.

Organization	ROR	Count
UNAVCO or UNAVCO, Inc.	https://ror.org/02n9tn974	1288
University of Colorado Boulder	https://ror.org/02ttsq026	408
The Ohio State University	https://ror.org/00rs6vg23	248
United States Geological Survey	https://ror.org/035a68863	124
Pennsylvania State University	https://ror.org/04p491231	102
New Mexico Institute of Mining and Technology	https://ror.org/005p9kw61	69
Colorado State University	https://ror.org/03k1gpj17	32
University of Montana	https://ror.org/0078xmk34	32
University of Oregon	https://ror.org/0293rh119	24
Oregon State University	https://ror.org/00ysfqy60	24
Georgia Institute of Technology	https://ror.org/01zkghx44	23
San Diego State University	https://ror.org/0264fdx42	21
Idaho State University	https://ror.org/0162z8b04	16
George Washington University	https://ror.org/00y4zzh67	16
Boston University	https://ror.org/05qwgg493	16
Dartmouth College	https://ror.org/049s0rh22	12
University of Miami	https://ror.org/02dgjyy92	12
University of Chicago	https://ror.org/024mw5h28	12
Goddard Space Flight Center	https://ror.org/0171mag52	12
Office of Polar Programs	https://ror.org/05nwjp114	12
National Aeronautics and Space Administration	https://ror.org/027ka1x80	12
The University of Texas at San Antonio	https://ror.org/01kd65564	12
University of Washington	https://ror.org/00cvxb145	10
University of California, Davis	https://ror.org/05rrcem69	9
Gustavus Adolphus College	https://ror.org/007q4yk54	8
Harvard University	https://ror.org/03vek6s52	8
University of Tennessee at Knoxville	https://ror.org/020f3ap87	6
The University of Texas at El Paso	https://ror.org/04d5vba33	4
Woods Hole Oceanographic Institution	https://ror.org/03zbnzt98	4
National Park Service	https://ror.org/044zqqy65	4
Bates College	https://ror.org/003yn7c76	4
University of Minnesota	https://ror.org/017zqws13	4
Texas A&M University	https://ror.org/01f5ytq51	4
University of Michigan–Ann Arbor	https://ror.org/00jmfr291	3
University of Michigan, Ann Arbor	https://ror.org/00jmfr291	1

These Affiliations and RORs were found using two different techniques. In most cases (2184) the affiliations were included in the metadata along with the individual creator. For example:

```
{
  "name": "Doe, Jane",
  "nameType": "Personal",
  "affiliation": [
    "UNAVCO, Inc."
  ]
}
```

In this case, the association of the author and the organization is clear.

In some cases, however, an author appears on some datasets with affiliations and in others without. In these cases, the affiliations were spread across the repository using the rules described above (Figure 5). Fortunately, affiliation ambiguity only occurred in four out of fifty-four cases. In the others the authors only had one associated affiliation given in the repository.

4.4 Organizational Connectivity

As described above, connectivity is the % of items in a collection that have identifiers, ORCIDs and RORs in this case. Now that RORs have been identified, we can calculate the organizational connectivity directly. Figure 11 shows the ROR connectivity after RORs were identified. The results show that 6% of the DOIs have RORs for all authors (complete connectivity), 8% have RORs for some authors (partial connectivity), and 85% have no RORs (missing connectivity). This is a significant improvement over the initial state in which no DOIs had RORs (100% missing).

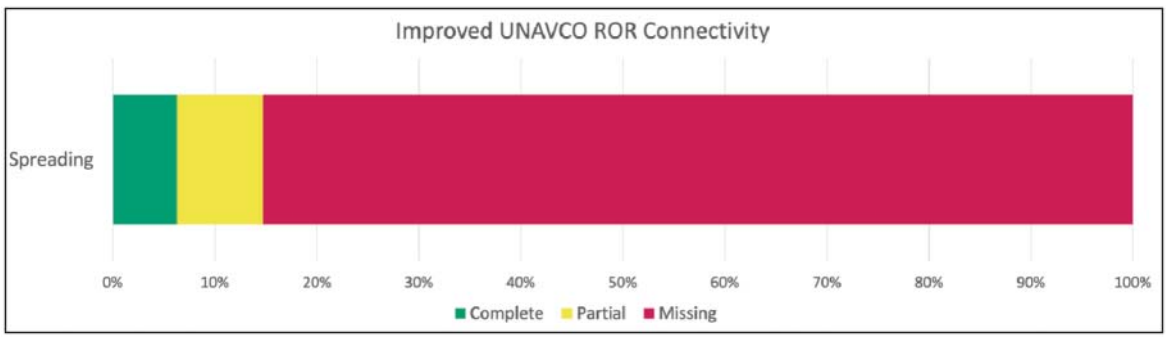

Figure 11. ROR connectivity after identification and spreading.

4.5 Person or Organization?

Most current metadata standards recognize that people and organizations can play similar roles in the creation and management of datasets and other research objects. In the DataCite metadata schema (since Version 4.1, [19]), the dichotomy is managed by soft-typing the creator and contributor objects, i.e. including the nameType property that is "Personal" for names of people and "Organizational" for names of organizations. If this property is provided the default value "Personal" is assumed.

When humans are reading metadata this default value is not a problem, as humans can usually tell the difference between organization and individuals names regardless of the nameType. When machines are reading the metadata, it can cause some problems, one being the identifier type appropriate for the party. If it is a person, ORCID is the first place to search, if it is an organization, ROR is more appropriate.

This difference is important in this case because the contributions of the UNAVCO community are reflected in the observation that "UNAVCO Community" and "Community, UNAVCO" are by far the most common creator names in the UNAVCO DataCite metadata, occurring 1471 times in the metadata collection. In other words, they occur in over 27% of the DOIs, outnumbering the other major contributors by over 1000 occurrences.

It seems reasonable to identify the community using the ROR for UNAVCO itself, as the community is an inseparable part of the organization. Of course, this has a major effect on the connectivity of the repository. Figure 12 shows the progression of connectivity through the various stages of this work. The top bar shows the situation after identifiers were added for the UNAVCO community. Adding these identifiers resulted in a five-fold increase in the number of DOIs with identifiers for all creators and a decrease of 27% in the number of DOIs with no identifiers. Note that the number of DOIs with partial connectivity did not change, indicating that the UNAVCO Community was the only creator on most of the datasets where it is listed as a creator. When we added the identifier, the connectivity for those datasets improved from missing to complete.

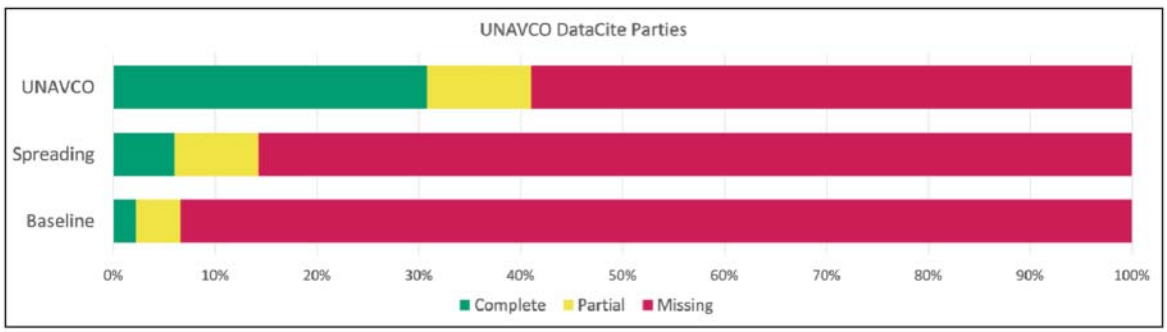

Figure 12. Connectivity for parties in UNAVCO DataCite metadata through time (increasing upward). Note the large improvement resulting in adding the identifier for the UNAVCO Community (middle to top row).

4.6 Metadata Archeology: Article Metadata

In previous steps we used DataCite metadata from UNAVCO to demonstrate how identifiers could be found and spread through the metadata collection to improve connectivity for people and organizations. The rest of the process relies on a list of community publications published on the UNAVCO website [11].

4.6.1 Finding Article DOIs

The process of finding DOIs described above involves Google searches and HTML processing. When using this approach on over 1500 papers, there are inevitable hiccups and challenges. In all, DOIs for 1222

(78%) of the papers were identified and those DOIs were searched using the approach described above. Most ORCIDs/Affiliations identified were from Crossref metadata rather than journal pages, one of the benefits of structured metadata.

The connectivity of this collection of DOIs with identifiers for people and affiliations (no RORs yet) can now be determined. Figure 13 shows the baseline connectivity of this collection for affiliations and people using the same visualization used for the DataCite metadata. The pictures are very different. Over 70% of the papers have affiliations for all authors (green in top bar of Figure 13) while only 2% of the papers have ORCIDs for all authors. More importantly, over 90% of the papers have no ORCIDs. The average connectivity for ORCIDs is 4.2% while the average for affiliations is 71%. This reflects the common observation that it is typical for all authors of a paper to have affiliations while only a few, typically the corresponding author, have ORCIDs.

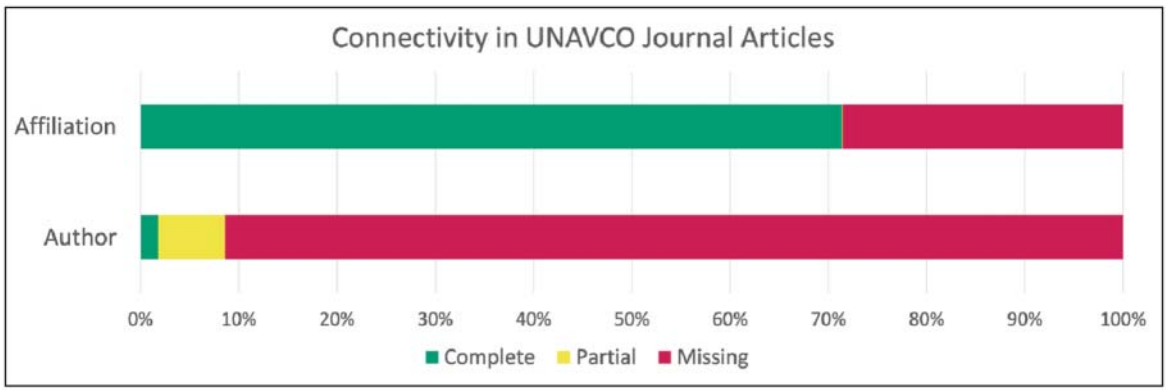

Figure 13. Baseline connectivity for UNAVCO journal articles.

It is important to note that these connectivity data are for papers rather than datasets, so the completeness of the identifiers is in the hands of the authors and the journals that publish the papers and provide the metadata for the papers to Crossref rather than the UNAVCO repository.

In addition to the general paucity of ORCIDs in publication metadata, author names for journal articles are entered into many different systems at different times and, as a result, they are very inconsistent. A person with two initials (F and M) and a family name can show up in different article author lists as F. M. Family, F M Family, F. Family, F Family, First M. Family, First M Family, First Family. Of course, this inconsistency is the primary motivator for using unambiguous and unique identifiers in the first place, so it is not surprising to observe it. Nevertheless, it makes matching names across a collection of papers more challenging.

In the 1222 UNAVCO articles with DOIs, there were 188 family names with identifiers that occurred 260 times. In most cases initials and names suggested that multiple combinations were the same person, but there were some cases that could not be resolved unambiguously.

The next step in the connectivity improvement process is to find article authors with identifiers that occur in the dataset metadata but do not have identifiers there. Thirty-five of the 188 article authors with ORCIDs did not have ORCIDs in the dataset metadata. These thirty-five authors occur 6316 times in the DataCite metadata, an average of 180 occurrences /author. These are, as expected, valuable identifiers.

It is interesting to note the rather small overlap between authors in the journal articles and authors of the datasets in the UNAVCO repository. This small overlap (19%) suggests significant re-use of the data in the UNAVCO repository by researchers that are not involved in the creation of the datasets, or at least not included in the metadata as creators. Of course, enabling and encouraging this re-use is the goal of the repository and they are doing it quite well by this measure.

As expected, based on the numbers above, adding ORCIDs and RORs into the repository significantly improved the connectivity. Figure 14 shows the evolution of the ORCID connectivity through all analysis stages. Before this work, the repository included ORCIDs for all (green) or some (yellow) authors for 6.6% of the datasets. Spreading those initial ORCIDs increased connectivity two times, adding identifiers for UNAVCO Community increased connectivity roughly six times, and adding in ORCIDs from the literature increased complete and partial connectivity over the initial state by a factor of 8.6, from 6.6% to 56.5%.

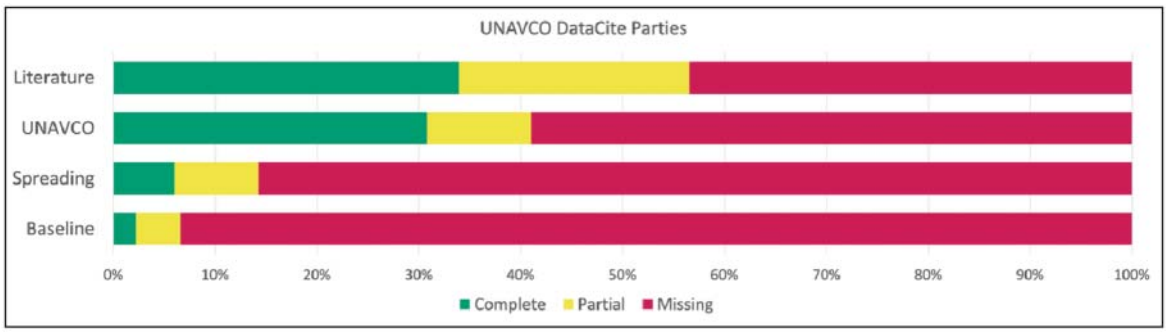

Figure 14. UNAVCO ORCID connectivity improvement.

Connectivity evolution for RORs is shown in Figure 15. In this case, the initial state included no RORs (0%) and complete and partial connectivity increased to 48.6%.

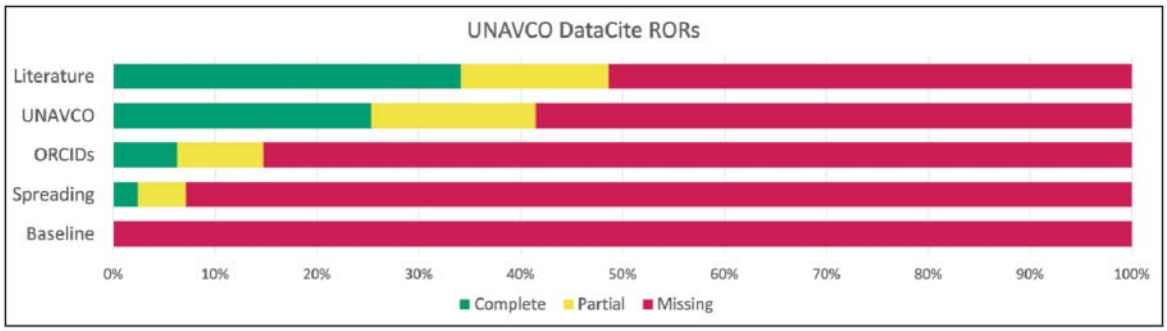

Figure 15. UNAVCO ROR connectivity improvement.

5. DISCUSSION

Several observations emerged during the analysis described here that may apply more broadly than in the UNAVCO case.

5.1 Identifier Awareness

When determining baseline connectivity for people and affiliations in UNAVCO, examination of the data indicated that many of the records that include ORCIDs also have affiliations, while records without ORCIDs also lack affiliations. This observation is reflected in Figure 16 which shows the average baseline connectivity per year for ORCIDs (orange) and affiliations (blue). Note the similarity of the time histories for both identifiers. This pattern may reflect increased awareness and attention to identifiers of several types in metadata workflows at UNAVCO during 2016 and 2017 compared to other time periods. It may also be related to tooling used for collecting and maintaining information about community members and getting it into metadata records. In any case, this observation suggests that repository practices and tool changes do impact metadata completeness and community connectivity.

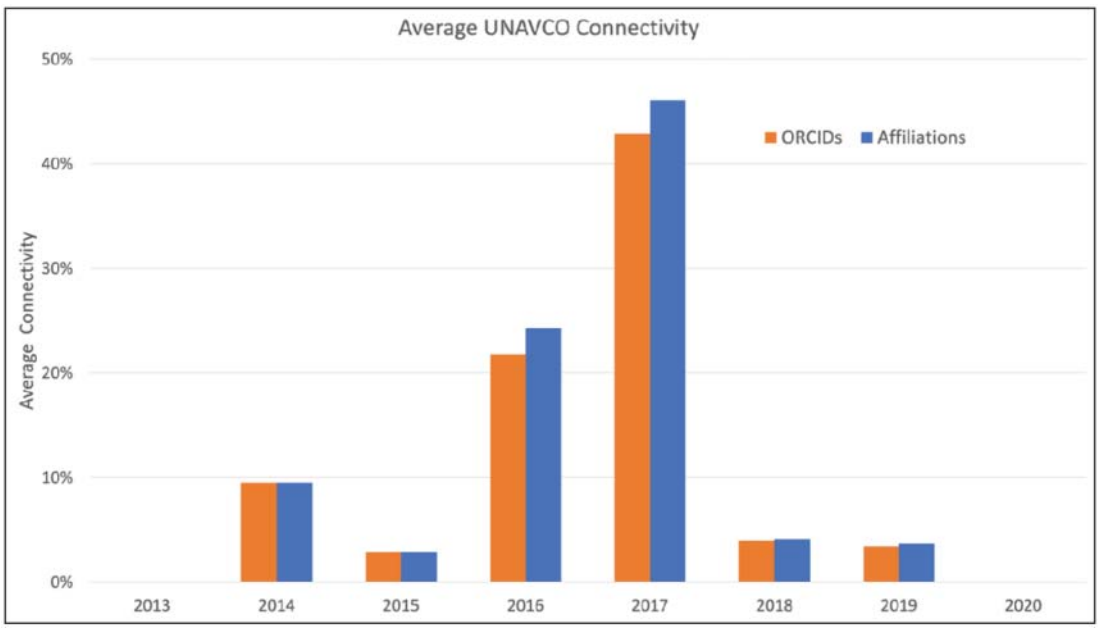

Figure 16. Average connectivity for ORCIDs (orange) and affiliations (blue) per year.

5.2 Affiliations and ORCIDs in Papers

This work included searching Crossref metadata for UNAVCO papers for ORCIDs and affiliations. The data indicate that this collection of papers is a much richer source for affiliation information than for ORCIDs. Over 70% of the papers have affiliations for all authors while only 2% of the papers have ORCIDs

for all authors (Figure 13), like the baseline numbers observed for ORCIDs in the DataCite metadata (Figure 7).

Figure 17 shows the average connectivity for ORCIDs and Affiliations in these papers over time. It confirms the general disparity between identifiers for people and affiliations. It also indicates that the connectivity for affiliations has increased over the last several years, after a more erratic pattern prior to 2019. Hopefully this reflects a general trend of increasing recognition by authors and journals of the importance of including affiliation information in the Crossref metadata.

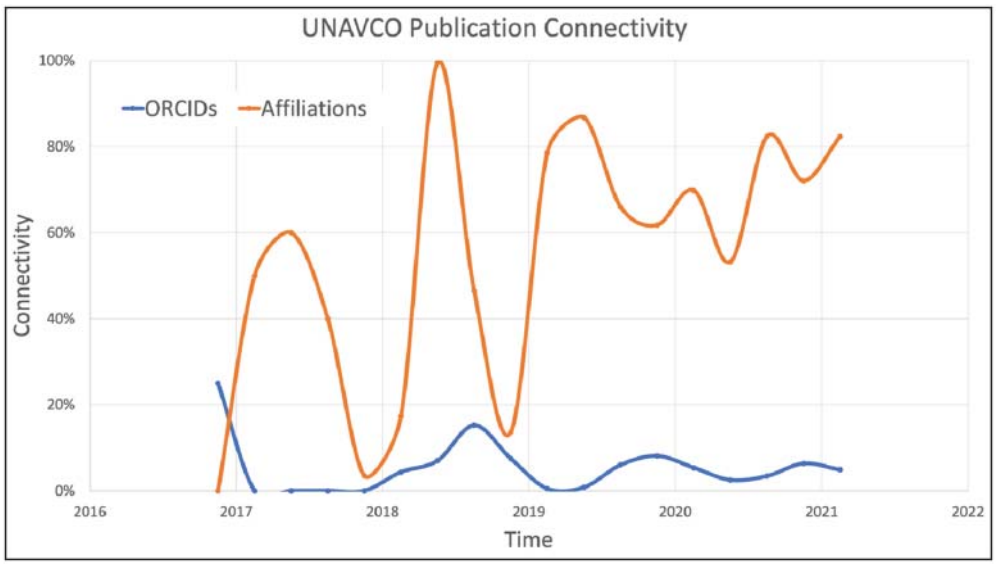

Figure 17. Average connectivity for UNAVCO papers versus time.

5.3 Defensive Metadata

People and organizations are differentiated in DataCite metadata using the nameType property and "Personal" is the default value for this property. This can result in organization names being misidentified as personal if metadata creation processes do not differentiate between people and organizations. In fact, this is the case in the UNAVCO metadata, i.e. UNAVCO Community is written without a nameType and, therefore, identified as a personal name:

```
"creators": [
      {
      "name": "UNAVCO Community",
    "affiliation": []
  }
]
```

As mentioned above, this can cause problems in searches for identifiers.

There are at least two ways to avoid these problems. First, the code that reads the metadata can search multiple identifier services, i.e. ORCID and ROR, for each name and record the type of the identifiers found. Second, the metadata creator can provide identifiers two ways: as a nameIdentifier and as an affiliationIdentifier. Using this approach users will find the identifier either way they search. In this case, the metadata looks like:

```
{
    "name": "UNAVCO Community",
    "affiliation": [
        {
            "affiliationIdentifier": "https://ror.org/02n9tn974",
            "affiliationIdentifierScheme": "ROR",
            "affiliation": "UNAVCO Community"
        }
    ],
    "contributorType": "creator",
    "nameType": "Organizational",
  "nameIdentifiers": [
        {
            "nameIdentifier": "https://ror.org/02n9tn974",
            "nameIdentifierScheme": "ROR"
        }
    ]
}
```

This approach might be considered as defensive metadata — accepting redundant information in the metadata to make sure users and tools find the information they are looking for regardless of where they look. The redundancy seems like a small price to pay for making life easier for a variety of users and, in this case, ensuring that the proper identifier is found in the metadata.

6. CONCLUSION

UNAVCO is a well-established domain repository with very strong real-world connections to the global Geodetic community. The UNAVCO DataCite repository, with over 5000 datasets, is much larger than the median DataCite repository size of ~180 resources and, because of the strong, long-term community, it provides an excellent sample for testing tools for increasing connectivity. The methods described here can be applied to repositories of any size, but they will be most effective for repositories with well-established and long-lived communities. In addition, these tools should be applicable to any domain-repository with a published metadata schema that includes metadata for people and organizations. The fact that UNAVCO

uses DataCite and DOIs for identifying datasets makes this work easier because the DataCite metadata are written using a well-known and described metadata schema [19].

The UNAVCO community recognized the importance of persistent identifiers for datasets during 2012 [6] and started using DataCite DOIs as unique and persistent dataset identifiers. Since that time, it has become clear that other kinds of identifiers, e.g. for people and organizations, are also important for recognizing and connecting members of the community.

This project explored the hypothesis that communities built around domain repositories provided fertile ground for adoption of identifiers because community members, both individuals and organizations, make many contributions to the repositories and published literature using data from the repositories. The hypothesis was tested through several phases:

Establishing a Baseline — we defined the concept of a connectivity metric, i.e. the % of datasets in the repository with researcher and organizational identifiers, proposed a visualization of that metric (Figure 2, Figure 3, and Figure 4), and established a baseline for that metric based on the UNAVCO DataCite repository. Roughly 6% of the datasets in the repository had some researcher identifiers (ORCIDs) and by spreading those known identifiers through the repository that number increased to ~14%.

Identifying Organizations — initially none of the organizations identified through affiliations in the repository had identifiers in the metadata. Fortunately, RORs could be identified for all those organizations. Spreading these RORs through the repository required an assumption that the most common affiliation was correct for researchers that had multiple affiliations but, given that assumption, the connectivity for organizations increased from 0% to ~7%.

Person or Organization — the most commonly occurring creator in the UNAVCO DataCite metadata is the UNAVCO Community itself, so giving that community the organizational identifier of UNAVCO provided recognition of the important role of the community and provided a jump in the connectivity to over 40% for individuals and organizations.

Article Metadata Archeology — UNAVCO compiles a list of papers that are published using data from the repository and adding DOIs for these papers provides access to metadata at Crossref and in journal pages. These metadata can be searched for author ORCIDs and affiliations. The connectivity for these papers is much better for affiliations (70% complete) than for authors (<10% complete or partial), reflecting the general paucity of ORCIDs in journal metadata.

Closing the Circle — in this final step new identifiers harvested from the literature were ingested back into the DataCite repository. Identifiers for less than 100 researchers and associated organizations were used over 9000 times in the repository for an average return of almost 100/identifier. The complete and partial connectivity in the repository increased to 56.5% for researchers and 48.6% for organizations.

These results demonstrate that the connectivity of the UNAVCO repository could be increased significantly using existing identifiers in the repository and related journal articles the connectivity, clearly confirming

the hypothesis in this case. This suggests that other domain repositories can also increase connectivity significantly using existing DataCite and Crossref resources.

AUTHOR CONTRIBUTION

Dr. Habermann was responsible for all aspects of this work including data access and analysis, graphics, and writing the manuscript.

REFERENCES

[1] NOAA Data One Stop. Available at: https://data.noaa.gov/onestop/. Accessed 9 October 2021
[2] NASA Earth Data Search. Available at: https://search.earthdata.nasa.gov/search. Accessed 9 October 2021
[3] Data.gov. Available at: https://catalog.data.gov/dataset. Accessed 9 October 2021
[4] Cousijn, H., Braukmann, R., Fenner, M., et al.: Connected Research: The Potential of the PID Graph. Patterns 2(1), 100180 (2021). Available at: https://doi.org/10.1016/j.patter.2020.100180
[5] UNAVCO. Available at: https://www.unavco.org/
[6] Pritchard, M., Owen, S., Anandakrishnan, S., et al.: Open Access to Geophysical Data Sets Requires Community Responsibility. Eos, Transactions American Geophysical Union 93(26), 243–243 (2012). Available at: https://doi.org/10.1029/2012EO260006
[7] UNAVCO Data Policy. Available at: https://www.unavco.org/community/policies_forms/data-policy/data-policy.html. Accessed 11 November 2020
[8] ORCID. Available at: www.orcid.org. Accessed 11 November 2020
[9] ROR. Available at: www.ror.org. Accessed 11 November 2020
[10] UNAVCO DataCite Metadata. Available at: https://api.datacite.org/dois?client-id=unavco.unavco. Accessed 11 November 2020
[11] UNAVCO, Community Publications. Available at: https://www.unavco.org/science/community-publications/community-publications.html. Accessed 12 Feb. 2021
[12] DataCite REST API Guide. Available at: https://support.datacite.org/docs/api. Accessed 11 November 2020
[13] Crossref Rest API. Available at: https://www.crossref.org/documentation/retrieve-metadata/rest-api. Accessed 11 November 2020
[14] UNAVCO Data Citation Guidelines. Available at: https://www.unavco.org/community/policies_forms/attribution/attribution.html. Accessed 11 November 2020
[15] Beautiful Soup. Available at: https://www.crummy.com/software/BeautifulSoup/bs4/doc/. Accessed 11 November 2020
[16] Habermann: A PID Feast for Research. Available at: https://metadatagamechangers.com/blog/2021/2/2/a-pid-feast-for-research-pidapalooza-2021 (2021). Accessed 11 October 2021
[17] Habermann: The Big Picture — Has Crossref metadata completeness improved? Available at: https://metadatagamechangers.com/blog/2019/3/25/the-big-picture-how-has-crossref-metadata-completeness-improved (2019). Accessed 11 October 2021
[18] Wynne, R.: Who uses ORCID IDs anyway? Available at: https://www.linkedin.com/pulse/who-uses-orcid-ids-anyway-richard-wynne (2020). Accessed 11 October 2021
[19] DataCite Metadata Working Group: DataCite Metadata Schema Documentation for the Publication and Citation of Research Data. Version 4.1. DataCite e.V. Available at: http://doi.org/10.5438/0014 (2017). Accessed 11 October 2021

AUTHOR BIOGRAPHY

Dr. Ted Habermann is currently the owner of Metadata Game Changers, a consulting company focused on helping organizations improve data discovery, access, and documentation. Projects include advising organizations like UNAVCO and NASA Langley Research Center on metadata systems, developing games as a means of building teams, and working with publishers to improve utilization of metadata for publications, software, and datasets. Ted was the Director of Earth Science at The HDF Group between 2013 and 2018. In addition to leading the group, he worked with the NASA Earth Science Data and Information System, the National Science Foundation, and many other U.S. and International groups to help facilitate the adoption ISO Standards for metadata and data quality. He led ISO Technical Committee 211 working groups that developed XML schema for the implementation of these standards. His vision is interoperable documentation for earth science datasets from many disciplines and locations. Many of his metadata presentations are available on SlideShare. He previously worked at NOAA's National Geophysical Data Center (NGDC) in Boulder, Colorado on a number of interoperability projects and international documentation standards. His group at NGDC developed and managed the NOAA Metadata Manager and Repository (NMMR) that included metadata for several thousand datasets produced and stewarded by NESDIS and the NOAA Data Centers. These records are provided in multiple standards (FGDC, DIF, NcML, THREDDS, and ISO) and views (FAQ, Text, HTML). More recently, Ted has focused on the challenges associated with evolution of organizational documentation systems and processes. Ted was a principle author on the recently approved NOAA Documentation Directive and participated in the Metadata Evolution for NASA Data Systems (MENDS) Project. Ted and his group have been active in the creation and sharing of ISO metadata for NESDIS, the NOAA Office of Climate Observations, the NOAA Observing System Architecture (NOSA), the GOES-R Project, the Integrated Ocean Observing System (IOOS), the Group for High-Resolution Sea Surface Temperature (GHRSST), Data.gov, and the World Meteorological Organization (WMO). This work has included translating several thousand metadata records from FGDC to ISO using several approaches, developing and sharing of ISO best practices, and participation in developing and implementing revisions to the ISO Standards. Recently Ted and his group have worked closely with NOAA's Unified Access Framework, Unidata and OPeNDAP to add ISO metadata capabilities to THREDDS and Hyrax Servers using a tool called ncISO.

ORCID: 0000-0003-3585-6733