Improving Domain Repository Connectivity

Ted Habermann, Metadata Game Changers

This work appeared originally as a series of Metadata Game Changer blogs published during early 2021.

Table of Contents

Establishing a Baseline

As I have been working with domain repositories to understand and describe their practices and apply for Core Trust Seal certification, I have been struck by the close, long-term relationships that these repositories form with their communities. In some cases, like UNAVCO, the repository is an integral part of an extensive community support system that 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 work together to create and extend our understanding of the shape of the earth and how it changes (the science of Geodesy).

The UNAVCO Community described the responsibilities of players in open science communities during 2012 (https://doi.org/10.1029/2012EO260006) and developed an open data policy based on those responsibilities. These responsibilities included identifying datasets with PIDs and connecting data to papers with citations, that is, establishing an important element of the PID Graph: connections between papers and data.

I introduced the concept of Connectivity last month and have been thinking about it ever since. 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. I am initially focusing on people, identified by ORCIDs, and organizations, identified by RORs.

As the breadth of identifiers and connections continues to expand, I made the leap from the strong connections between real people and organizations in the UNAVCO Community and connections between these entities in the PID Graph. Specifically, I wondered if the multitudinous real-world connections could help us populate identifiers in the metadata and related connections in the PID Graph. I begin the exploration of this question here with UNAVCO datasets described in DataCite.

UNAVCO Datasets in DataCite

UNAVCO has minted over 5000 dataset DOIs with DataCite since 2013 (Figure 1). UNAVCO maintains an 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.

erin@metadatagamechangers.com
https://orcid.org/0000-0001-9998-0114
ted@metadatagamechangers.com
https://orcid.org/0000-0003-3585-6733

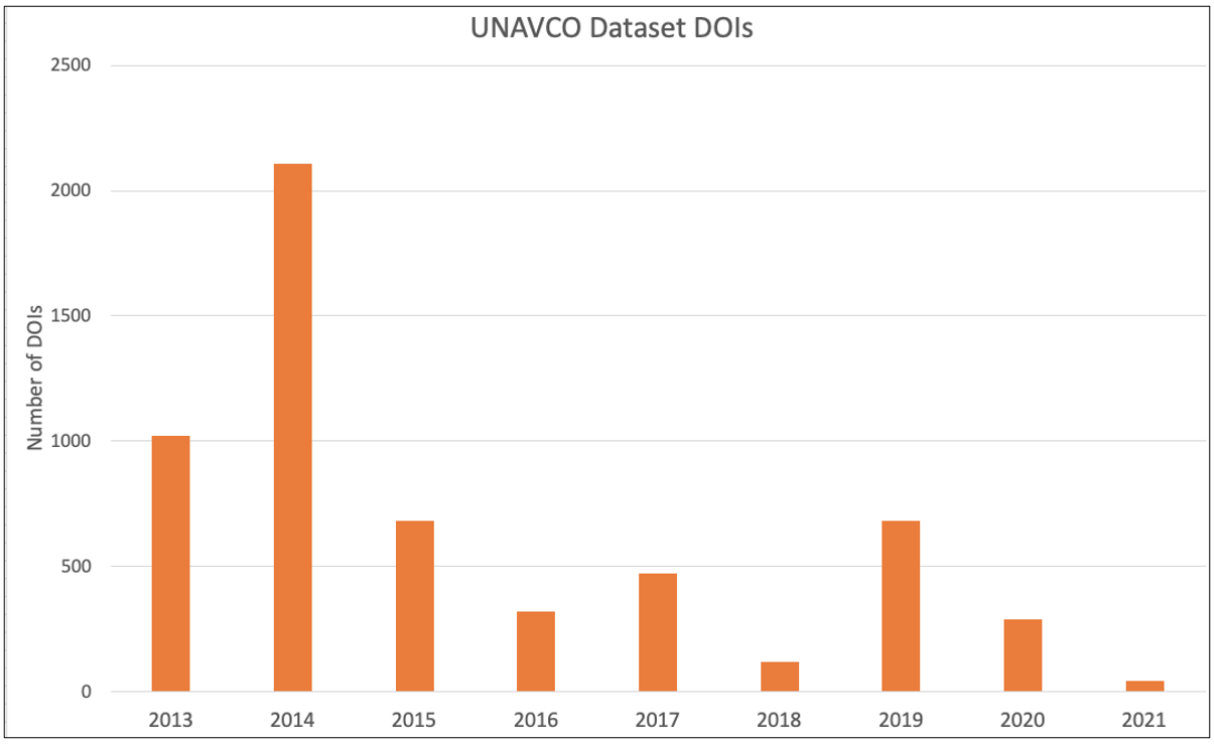


Figure 1. The number of UNAVCO datasets registered in DataCite per year.

DataCite is also the place where the UNAVCO Community connects their data to the broader scientific world through identifiers included in the metadata. As mentioned above, I am particularly interested in connectivity through ORCIDs and RORs. Characterizing the current state of these identifiers in the collection is the first step in understanding the collection and measuring improvements in the connectivity that might be achieved through time.

Visualizing Connectivity

Our goal is to understand how to improve connectivity in domain repositories and to use connectivity as a metric for measuring progress as connectivity improves. In order to do this, we must be able to express connectivity as numbers and pictures. I do this using a horizontal bar which represents the entire collection and color sections of the bar green for items that have complete connectivity, yellow for items that have partial connectivity, and red for items that have no connectivity (Figure 2).

erin@metadatagamechangers.com
https://orcid.org/0000-0001-9998-0114
ted@metadatagamechangers.com
https://orcid.org/0000-0003-3585-6733

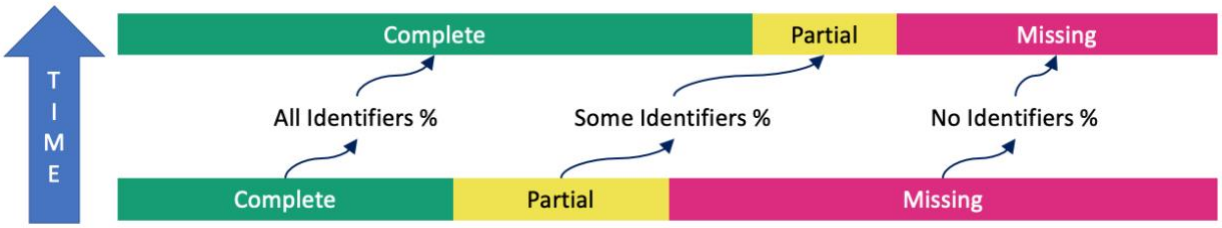


Figure 2. Visualizing connectivity as the % of items with all identifiers (complete, green), with some identifiers (partial, yellow), and with no identifiers (missing, red).

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 2 by the change between the lower and upper bars.

ORCID Connectivity Baseline

Connectivity can be measured for many kinds of identifiers and for any collection of research objects or other entities in the PID Graph. For a single paper, ORCID connectivity is the % of authors that have ORCIDs (see Connectivity). This calculation can be easily extended to a collection with the ORCID connectivity being the % of authors in all collection items that have ORCIDs.

The baseline ORCID connectivity for the UNAVCO DataCite collection is shown in Figure 3. The largest part of the datasets in the collection have no ORCIDS (93%, 5005/5356 DOIs) while 234 (4%) have some ORCIDS, and 117 (2%) have all ORCIDS.

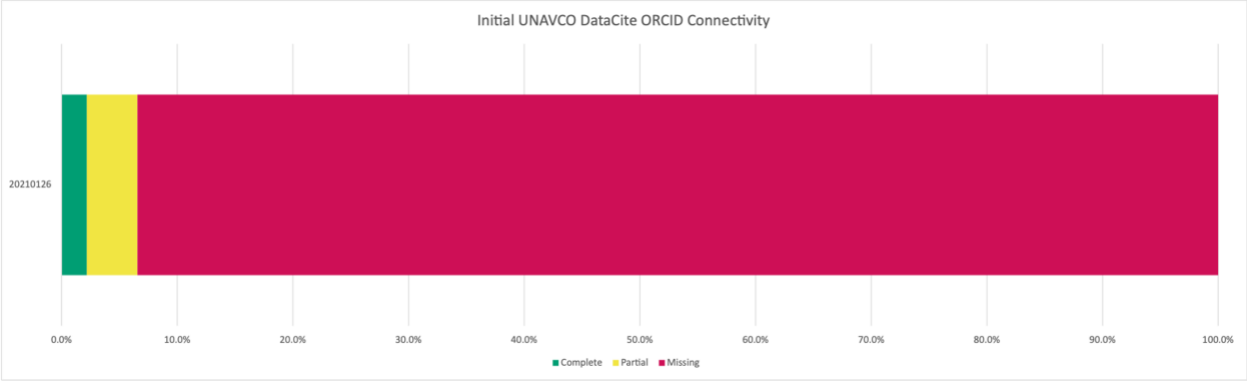


Figure 3. Initial (baseline) connectivity for ORCIDs in UNAVCO DataCite metadata.

The lack of ORCIDs in the UNAVCO metadata is not unusual. An assessment of 144 DataCite repositories in the TIB Consortium showed that, on average, less

erin@metadatagamechangers.com
https://orcid.org/0000-0001-9998-0114
ted@metadatagamechangers.com
https://orcid.org/0000-0003-3585-6733

than 15% of the records in these repositories have identifiers and a similar assessment of all Crossref metadata during 2019 showed that the average portion of Crossref records with ORCIDs was less than 10% and it was only during mid-2020 that the average number of ORCID's per article in Crossref passed 2.0. We are clearly at the beginning of the ORCID adoption process across the scientific publishing world and we have a lot of room for increased adoption.

On the positive side, fifty-three authors have ORCIDs in this metadata that occur in a total of 499 datasets and just over 350 datasets had complete or partial ORCID coverage. The most common ORCID belongs to Marianne Okal, an engineer at UNAVCO. Her ORCID occurs 130 times in these data. Severn other ORCIDS occur more than ten times.

Affiliation Connectivity Baseline

The UNAVCO DataCite metadata do not currently include any organizational identifiers but the metadata do include affiliation names that can give us an idea of the maximum organizational connectivity that we can achieve if we can find identifiers for all of the affiliations.

Figure 4 shows the initial affiliation connectivity for UNAVCO metadata at DataCite which is very similar to the data for ORCIDs in Figure 3. In fact, the numbers are slightly better, with 382 records having complete or partial connectivity. Unfortunately, 93% of the records are still missing affiliation information.

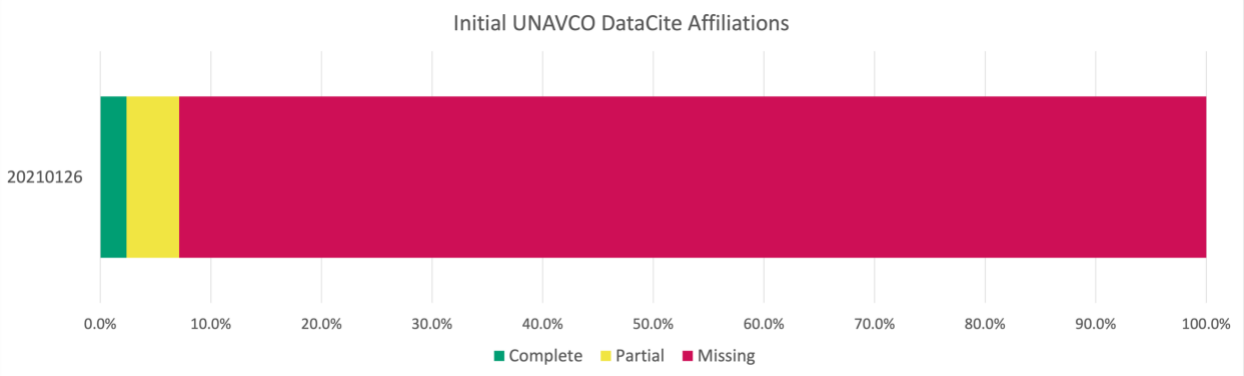


Figure 4. Initial (baseline) connectivity for Affiliations in UNAVCO DataCite metadata.

Identifier Awareness

Closer examination of the data shows that many of the records that have ORCIDs also have affiliations, while records without ORCIDs also lack affiliations. This observation is reflected in Figure 5 which shows the average connectivity per year for ORCIDs (orange) and affiliations (blue). Note the similarity of the

erin@metadatagamechangers.com
https://orcid.org/0000-0001-9998-0114
ted@metadatagamechangers.com
https://orcid.org/0000-0003-3585-6733

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.

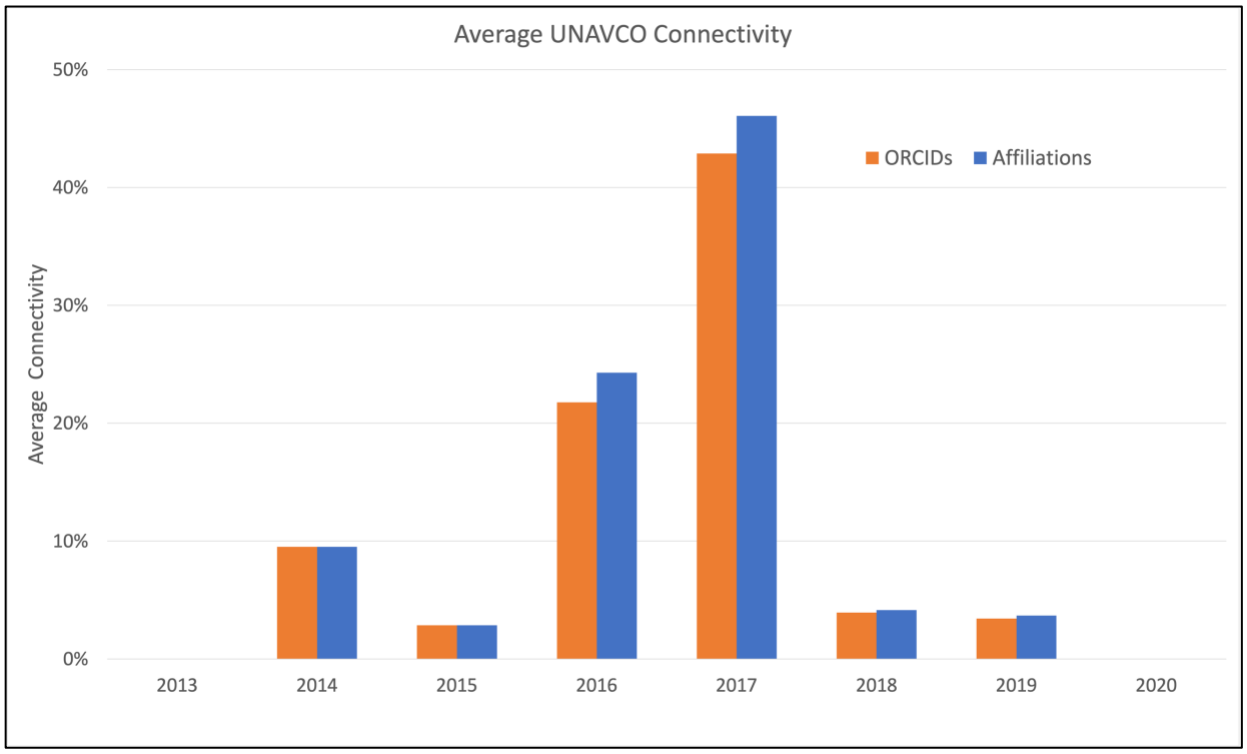


Figure 5. Average connectivity for ORCIDs (orange) and affiliations (blue) per year.

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.

For example, we know that Marianne Okal, an engineer at UNAVCO, is the author with the most common ORCID in these data, it occurs in 130 datasets.

erin@metadatagamechangers.com
https://orcid.org/0000-0001-9998-0114
ted@metadatagamechangers.com
https://orcid.org/0000-0003-3585-6733

She is an author on five other datasets that do not include here ORCID in the metadata, so, she is an author 135 times and has 130 ORCIDs and her connectivity is 130/135 = 96%. We can increase the number of records with ORCIDs by five by adding her ORCID to five records that are currently missing it. This also increases her connectivity to 1, i.e., complete.

Figure 6 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%.

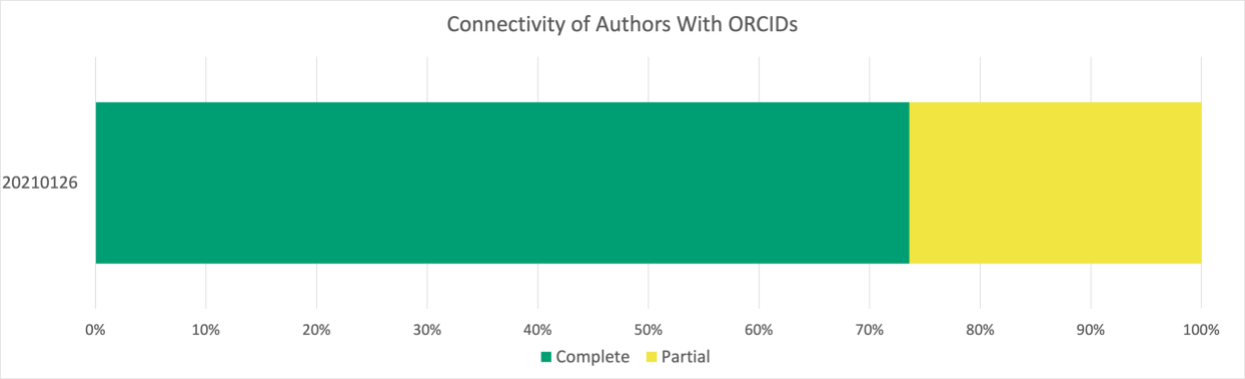


Figure 6. Author connectivity in UNAVCO DataCite metadata.

Figure 7 compares the ORCID connectivity before and after the addition of known ORCIDs. As expected on the basis of 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%. As mentioned earlier, this improvement was achieved with the completely safe assertions that ORCIDs for authors do not change.

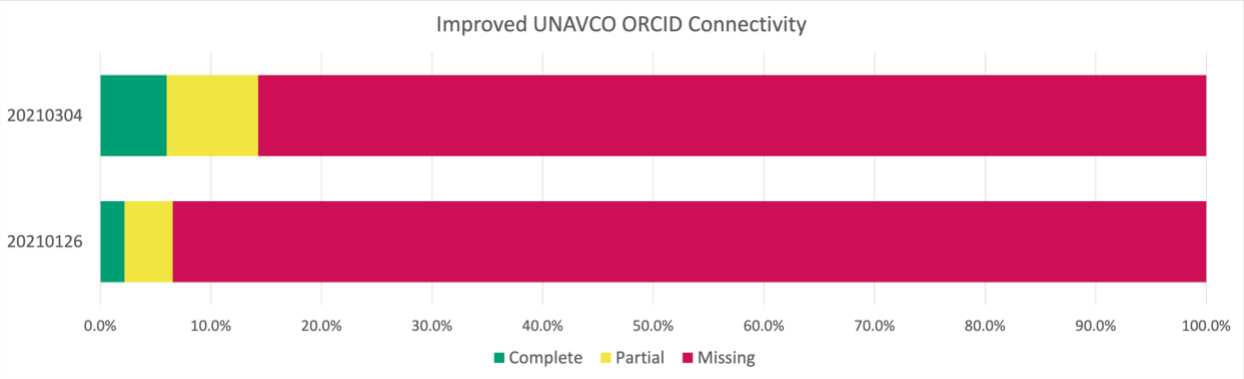


Figure 7. Improvement in ORCID connectivity associated with spreading known ORCIDS to metadata without ORCIDs.

erin@metadatagamechangers.com
https://orcid.org/0000-0001-9998-0114
ted@metadatagamechangers.com
https://orcid.org/0000-0003-3585-6733

Conclusion

UNAVCO is a domain repository with very strong real-world connections to the Geodetic community in the United States and across the world. In order for these connections to be reflected in the PID Graph, UNAVCO datasets and the papers that use them must have unique and persistent identifiers and metadata for those research objects must include identifiers for people, organizations, and other related research entities.

The UNAVCO DataCite metadata currently includes some or all ORCIDs for 6% of the datasets, reflecting only a small portion of the existing connections. This blog post describes a quantitative measure of the repository connectivity and uses that metric to demonstrate a significant increase in connectivity accomplished by including known ORCIDs across all records. This first step suggests that connectivity can be increased using existing community information resources. Further improvements will be described in future blogs.

erin@metadatagamechangers.com
https://orcid.org/0000-0001-9998-0114
ted@metadatagamechangers.com
https://orcid.org/0000-0003-3585-6733

Identifying Organizations

The UNAVCO DataCite Repository has over 5000 records that describe datasets created by researchers from many organizations, all of which are members of the tight-knit and well-established UNAVCO community. In the first blog of this series, I proposed that connecting these organizations to the PID Graph depends on having unique identifiers, i.e., RORs, for these organizations. Most of these organizations have contributed multiple datasets to the community, so they occur multiple times in the metadata. This characteristic of domain communities, i.e. multiple contributions from the same people and organizations, simplify the process of populating organizational identifiers in the repository because each identifier is used many times.

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. 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%). It is also important to note that RORs were identified for all organizations in the metadata.

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

erin@metadatagamechangers.com
https://orcid.org/0000-0001-9998-0114
ted@metadatagamechangers.com
https://orcid.org/0000-0003-3585-6733

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:

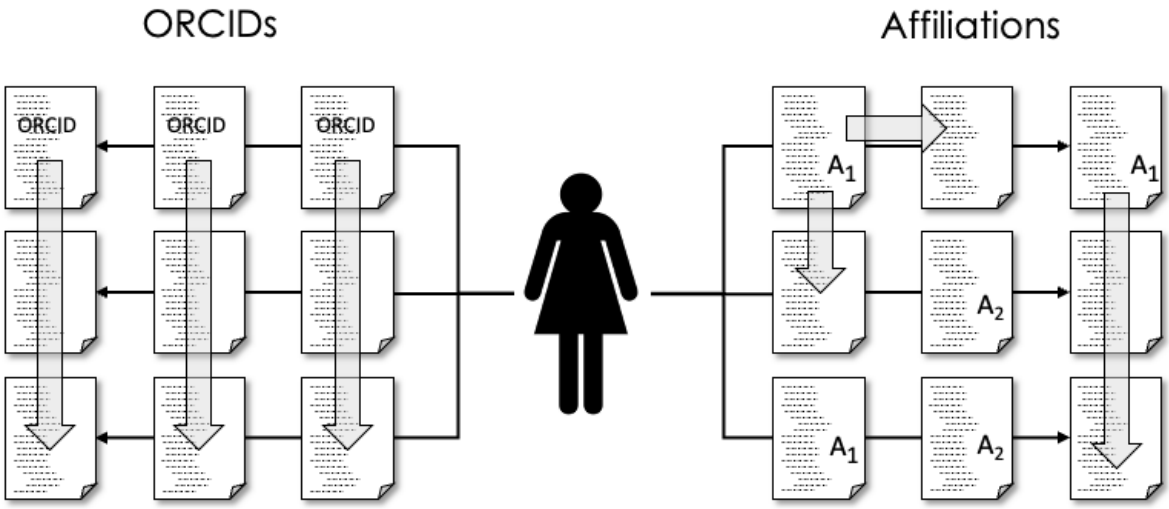
```
{
   "name": "Doe, Jane",
   "nameType": "Personal",
   "affiliation": [
      "UNAVCO, Inc."
   ]
}
```

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. Is it possible to spread the known affiliations across the datasets which do not have affiliations in the metadata?

In the first blog in this series, I described the process of spreading ORCIDs for dataset authors across occurrences of the authors in the metadata that did not originally include ORCIDS. This increased the number of ORCIDs in the metadata significantly and, because the association between a person and their ORCID is one-to-one, there is high confidence in the assertion of the connection between the person and the ORCID. In the affiliation case, the confidence is not so high, as authors can readily switch organizations.

erin@metadatagamechangers.com https://orcid.org/0000-0001-9998-0114
ted@metadatagamechangers.com https://orcid.org/0000-0003-3585-6733

This situation is illustrated in Figure 1. This author has authored nine datasets and ORCIDs are included in three of them. In this case, the ORCID connectivity for this author is 33%. The connectivity is increased to 100% by adding the ORCID for this author to the six datasets that originally had no ORCIDs, indicated by the grey arrows.



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? Of course, there is no answer here that has 100% confidence. The rules used here were as follows to spread affiliations as shown by the grey arrows:

1. if only one affiliation exists, use it
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.

In addition, affiliations identified this way are 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 corrected.

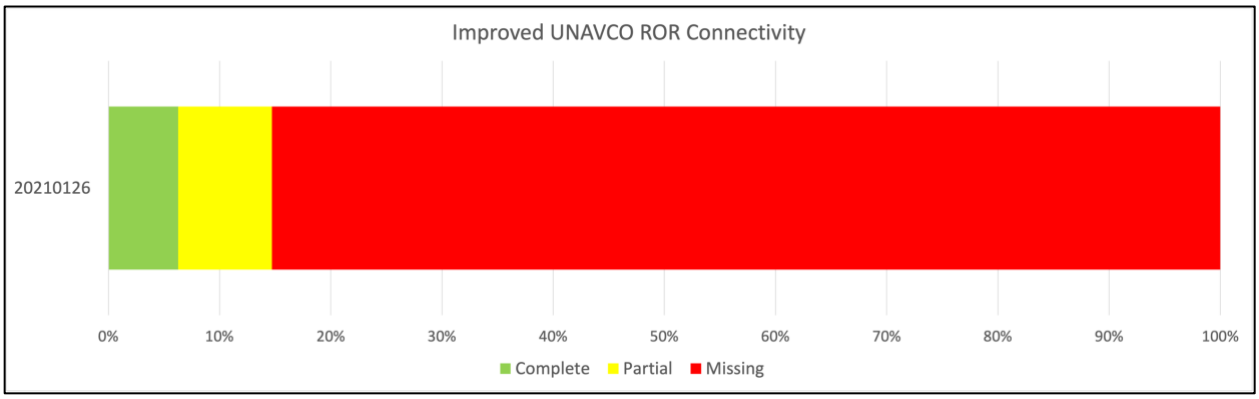
Fortunately, this ambiguity only occurred in four out of fifty-four cases. In the others the authors only had one associated affiliation.

Organizational Connectivity

As mentioned in the last blog, connectivity is the % of items in a collection that have identifiers, ORCIDs and RORs in this case. The UNAVCO metadata do not include RORs, so affiliations were used as a proxy for the RORs and connectivity was calculated for ORCIDs and affiliations. Now that RORs have been identified,

erin@metadatagamechangers.com
https://orcid.org/0000-0001-9998-0114
ted@metadatagamechangers.com
https://orcid.org/0000-0003-3585-6733

we can calculate the organizational connectivity directly. Figure 2 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).



Conclusions

Improving the number of identifiers of all kinds (connectivity) across repositories is important as the role of connections increases in the discovery processes. Domain repositories can benefit from well-developed communities in making these improvements because community members, both individuals and organizations, make multiple contributions through time.

These benefits were demonstrated using the UNAVCO DataCite repository as an example. In this case, the three most common ORCIDs make up 50% of all ORCIDs in the repository. Similarly, as shown here, the three most common RORs make up almost 70% of the RORs in the repository.

Taking advantage of this characteristic, the connectivity of the UNAVCO repository was increased using information already in the repository. The portion of DOIs with all ORCIDs (complete connectivity) increased by a factor of three, from 2 to 6% while the portion of organizations with all RORs increased from zero to 6% (see Table 1).

	Party			Organizations		
Connectivity	Missing	Partial	Complete	Missing	Partial	Complete
Initial	93%	4%	2%	100%	0%	0%
Improved	86%	8%	6%	85%	8%	6%

erin@metadatagamechangers.com
https://orcid.org/0000-0001-9998-0114
ted@metadatagamechangers.com
https://orcid.org/0000-0003-3585-6733

Unfortunately, like in many other repositories, the portion of DOIs without ORCIDs or RORs remains very high. Fortunately, like many domain repositories, UNAVCO maintains a list of papers that have been written by community members using data from UNAVCO. Metadata for these papers is another source of information that can be brought to bear on the problem of increasing connectivity.

erin@metadatagamechangers.com
https://orcid.org/0000-0001-9998-0114
ted@metadatagamechangers.com
https://orcid.org/0000-0003-3585-6733

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. This dichotomy was managed with the introduction of the concept of 'party' which could be a person, organization, or position in the ISO TC211 metadata standards for geographic data. Each of the different types of parties have different properties, for examples, organizations can include people or positions, but people and positions cannot include organizations.

In the DataCite metadata schema, 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. This works quite well if this property is provided, but, if not, the default value of 'Personal' is used.

When humans are reading metadata this default value is not a problem as humans can tell the difference between Metadata Game Changers and Ted Habermann regardless of the value of 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 or GRID are more appropriate.

The UNAVCO Community

This series of blog posts explores the hypothesis that domain repositories are great places to work on improving connectivity because they build strong communities of people and organizations that contribute and use data from the repository many times in their science. In the UNAVCO case, the importance of the community is reflected in the observation that 'UNAVCO Community' and 'Community, UNAVCO' are by far the most common creator names in the UNAVCO DataCite metadata, occurring1471 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.

The UNAVCO Community is clearly an important contributor to the repository and their role needs to be reflected in the connections that make up the PID Graph. 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 1 shows the progression of connectivity through the various stages of this work with green being complete connectivity (identifiers for all creators), red being missing (some identifiers), and yellow being partial (some identifiers). The top bar shows the situation after identifiers were added for the UNAVCO community. Adding

erin@metadatagamechangers.com
https://orcid.org/0000-0001-9998-0114
ted@metadatagamechangers.com
https://orcid.org/0000-0003-3585-6733

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 went from missing to complete.

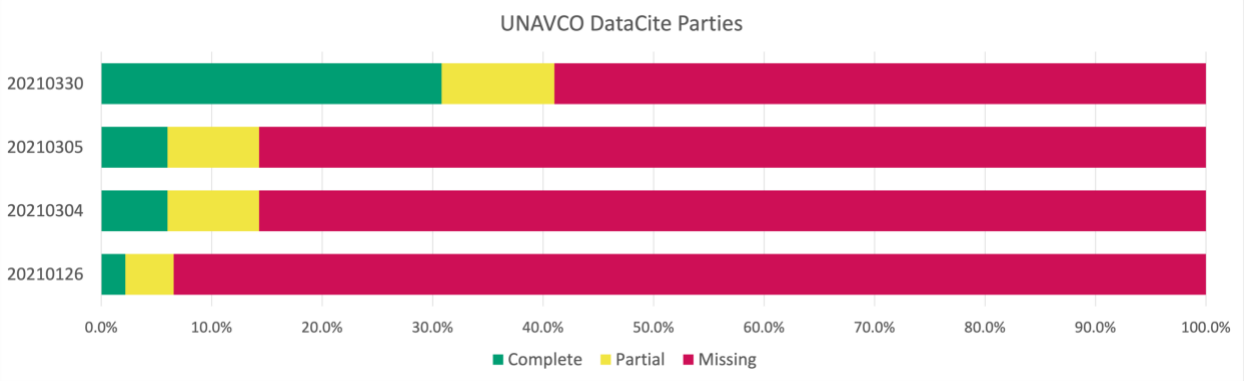


Figure 8. Connectivity for parties in UNAVCO DataCite metadata through time (increasing upward). Note the large improvement resulting in adding the identifier for the UNAVCO Community.

Defensive Metadata

This blog began discussing how people and organizations are differentiated in DataCite metadata using the nameType property and pointed out that '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:

erin@metadatagamechangers.com
https://orcid.org/0000-0001-9998-0114
ted@metadatagamechangers.com
https://orcid.org/0000-0003-3585-6733

```
{
   "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"
      }
   ]
}
```


I think of this approach 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 redundance seems like a small price to pay for making life easier for a variety of users.

Conclusion

At this point all possible information has been extracted from the existing UNAVCO DataCite metadata. We increased the portion of datasets with identifiers for all people from 2% to 31% and the portion of datasets with identifiers for organizations from 0% to 25% and the confidence in most of the assertions we made about connections to identifiers is very high. The next step involves searching the UNAVCO community publications for identifiers, affiliations, and RORs. Stay tuned!

erin@metadatagamechangers.com https://orcid.org/0000-0001-9998-0114
ted@metadatagamechangers.com https://orcid.org/0000-0003-3585-6733

Metadata Archeology: Article Metadata

In previous blogs we used DataCite metadata for UNAVCO to demonstrate how identifiers could be found and spread through the metadata collection to improve connectivity for people and organizations. The community built around UNAVCO over time was a critical part of this process as community members, both individuals and organizations, make many contributions over time. We also benefited from the fact that all contributing organizations have RORs and that the UNAVCO Community itself was recognized as an important contributor to the datasets described in the metadata.

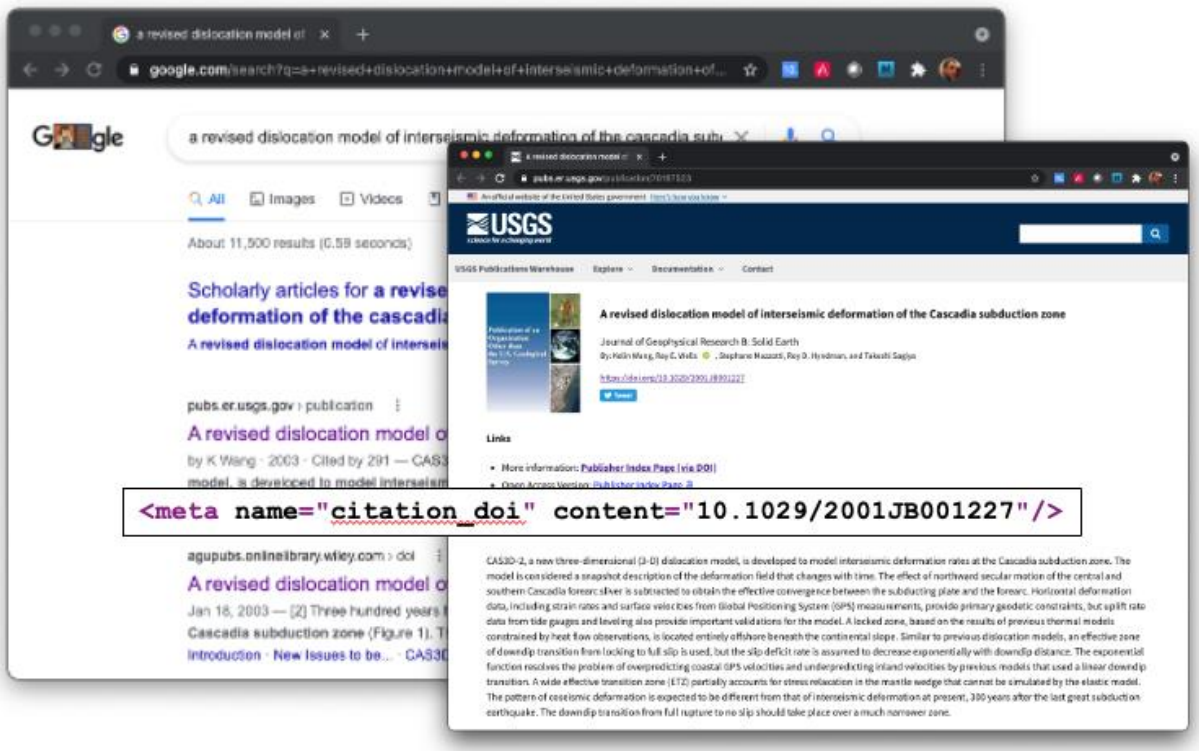
Like many domain repositories, UNAVCO keeps track of papers that are published using data that are in the repository. Dataset DOIs and clear citation guidelines both make it easier to do this tracking. The list of UNAVCO community publications available on the website includes 1569 articles published 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. Finding these identifiers in the papers is termed Metadata Archeology as it involves searching for exciting finds in a large information set.

Finding Article DOIs

The first step in the process of finding identifiers within these papers is to find DOIs for the papers themselves. This was done using google searches for the titles of the papers and searching results for pages with titles matching the titles of the papers. 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 1 for the paper titled "A revised dislocation model of interseismic deformation of the Cascadia subduction zone" (https://doi.org/10.1029/2001JB001227). In this case, as in most examples, all goes well and the DOI is easily determined from the first link in the google results. When using this approach on over 1500 papers, there are inevitable hiccups and challenges. I was able to retrieve DOIs for 1222 (78%) of the papers.

erin@metadatagamechangers.com
https://orcid.org/0000-0001-9998-0114
ted@metadatagamechangers.com
https://orcid.org/0000-0003-3585-6733



Once the DOIs were known, I used two approaches to finding 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 and retrieving ORCIDS and affiliations 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 mouseovers or popups or other exotic and clever approaches that may work for humans but are difficult for machines. In any case, the vast majority of ORCIDs/Affiliations identified were from Crossref.

Connectivity

Now we have a collection of DOIs with identifiers for people and affiliations (no RORs yet) which means we can determine the connectivity of the collection. Figure 2 shows the baseline connectivity of this collection for affiliations and

erin@metadatagamechangers.com
https://orcid.org/0000-0001-9998-0114
ted@metadatagamechangers.com
https://orcid.org/0000-0003-3585-6733

people using the same visualization used earlier for the DataCite metadata. The pictures are very different. Over 70% of the papers have affiliations for all authors (green in Figure 2) 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 observation that it is typical for all authors of a paper to have affiliations while only a few, typically the corresponding author, have ORCIDs.

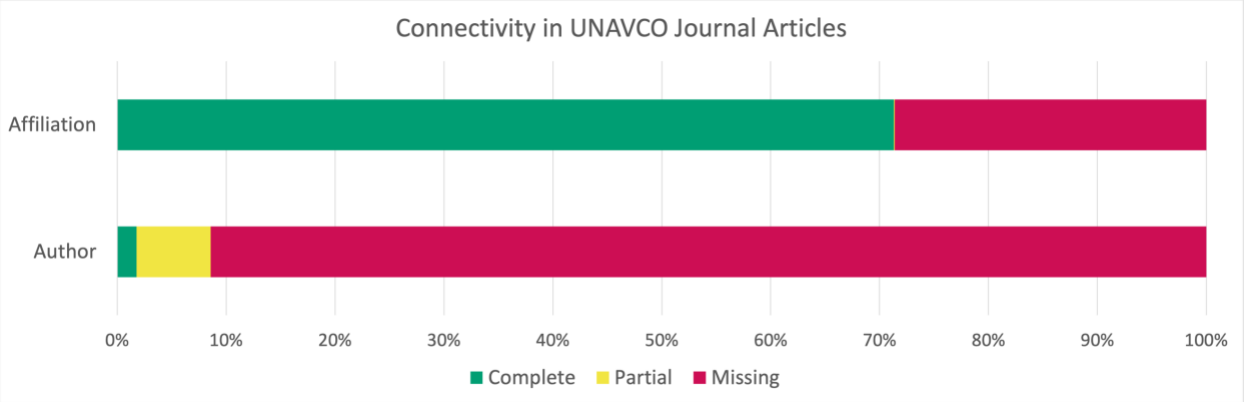
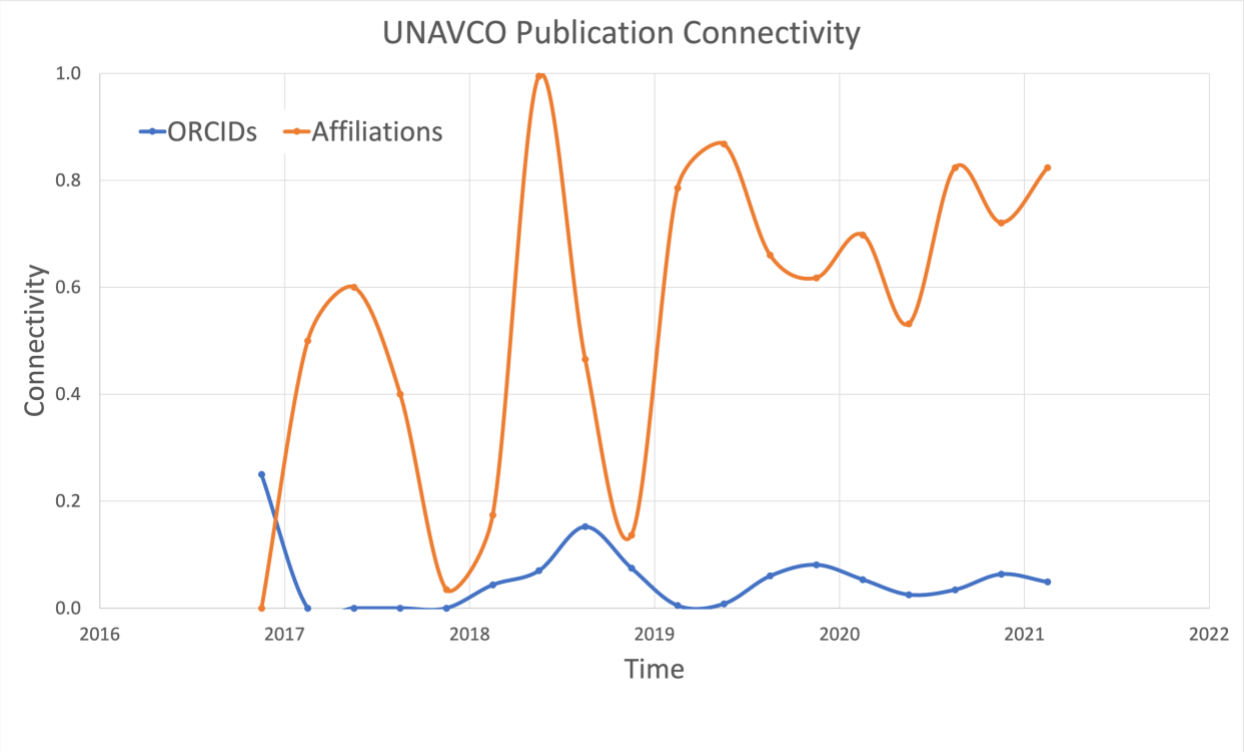


Figure 3 shows the average connectivity for ORCIDs and Affiliations over time. It confirms the general disparity between identifiers for people and affiliations. It also indicates that the connectivity for affiliations has increased over the last several years.

erin@metadatagamechangers.com
https://orcid.org/0000-0001-9998-0114
ted@metadatagamechangers.com
https://orcid.org/0000-0003-3585-6733



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 journals that publish the papers and provide the metadata for the papers to Crossref rather than the UNAVCO repository.

Conclusion

UNAVCO, like many domain repositories, tracks papers that use data from the repository. These papers are a potential source for author identifiers, affiliations, and organization identifiers. The search for these identifiers involves multiple steps, first finding DOIs for the papers using title searches, and then retrieving identifiers from Crossref or scraping journal web pages.

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, similar to the numbers observed for ORCIDs in the DataCite metadata.

The next step is to associate ORCIDs and affiliations with authors and to transfer discovered identifiers back to the dataset metadata in DataCite that are managed, and therefore controlled, by UNAVCO. This effort brings to the fore the ambiguity in author names, i.e. full names, initials, mixtures, as well as the

erin@metadatagamechangers.com
https://orcid.org/0000-0001-9998-0114
ted@metadatagamechangers.com
https://orcid.org/0000-0003-3585-6733

more significant ambiguity in affiliations for the same organization. This topic will be explored in the next blog.

erin@metadatagamechangers.com
https://orcid.org/0000-0001-9998-0114
ted@metadatagamechangers.com
https://orcid.org/0000-0003-3585-6733

Metadata Archeology 2: Closing the Circle

In the last blog in this series, we searched metadata for over 1500 journal articles written by members of the UNAVCO Community for ORCIDs and affiliations. The primary observation was that there were many more affiliations in the article metadata than ORCIDs, not surprising since most authors provide affiliations when articles are submitted while very few, typically just one, provide ORCIDs.

The fundamental hypothesis being explored here is that identifiers of any kind, even if they occur just once, can be useful in domain repository metadata because individuals and organizations make multiple contributions to the repository and the corpus of scientific literature based on it. The goal of the work is to increase the number of identifiers in DataCite metadata associated with datasets created at UNAVCO. The hypothesis was definitely true within the DataCite metadata as ORCIDs and affiliations, even if they occurred only once, could be spread throughout the repository to increase connectivity. The search for ORCIDs and identifiers was extended to include literature that had been identified by UNAVCO as being related to data in the repository.

Article ORCIDs

As mentioned above, ORCIDs are generally rare in journal article metadata and the UNAVCO articles were typical: over 90% of the papers have no ORCIDs. In addition to this general paucity, 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. This inconsistency is the primary motivator for using unambiguous and unique identifiers in the first place, so it is not surprising to observe it.

In the UNAVCO articles, there were188 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.

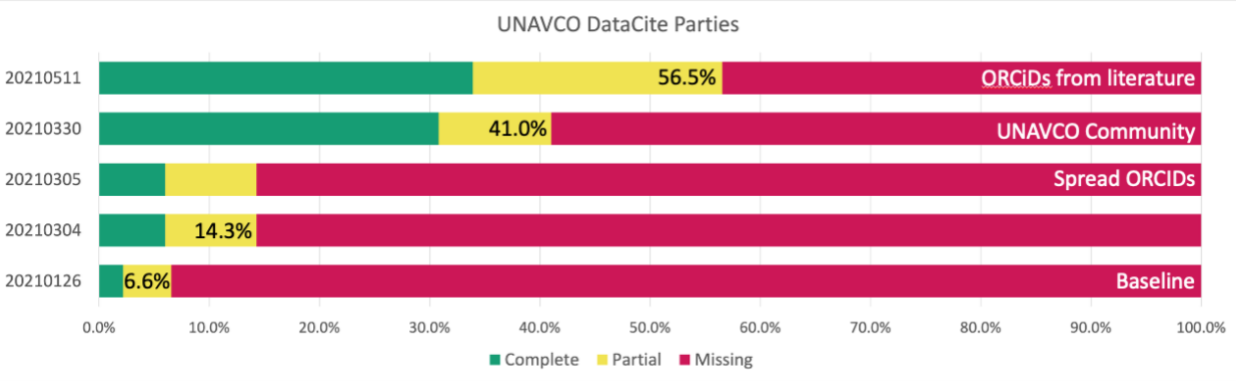
The next step in the connectivity improvement process is to find article authors with identifiers that do not have identifiers in the dataset metadata. Thirty-five of the 188 article authors with ORCIDS did not have ORCIDs in the dataset metadata. These authors support the hypothesis that we are testing by occurring 6316 times in the DataCite metadata, an average of 180/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

erin@metadatagamechangers.com
https://orcid.org/0000-0001-9998-0114
ted@metadatagamechangers.com
https://orcid.org/0000-0003-3585-6733

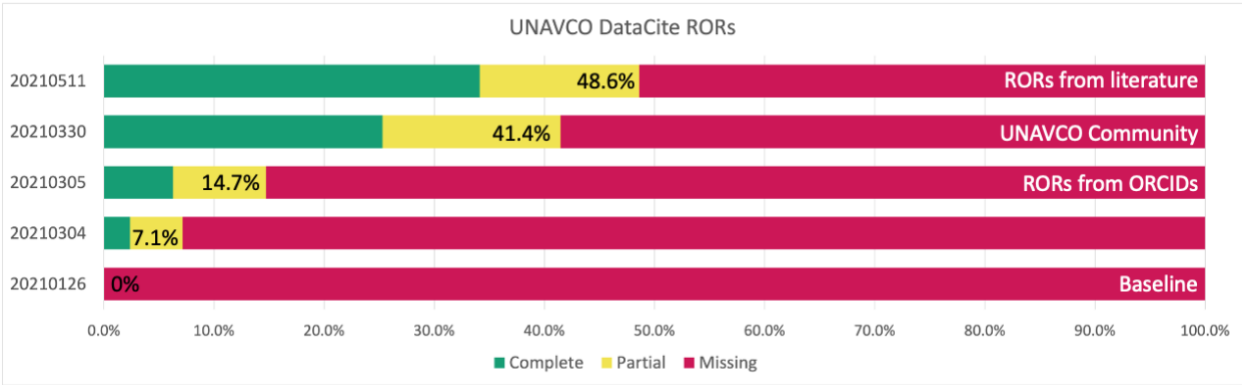
(19%) reflects significant re-use of the data in the UNAVCO repository by researchers that are not involved in the creation of the datasets. Of course, enabling and encouraging this re-use is the goal of the repository and they are doing it quite well by this measure.

Connectivity

As expected, based on the numbers above, adding new ORCIDs and RORs into the repository significantly improved the connectivity. Figure 1 shows the evolution of the ORCID connectivity through the analysis stages described in this blog series. 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, to 56.5%.



Connectivity evolution for RORs is shown in Figure 2. In this case, the initial state included no RORs (0%) and complete and partial connectivity increased to 48.6%.



erin@metadatagamechangers.com
https://orcid.org/0000-0001-9998-0114
ted@metadatagamechangers.com
https://orcid.org/0000-0003-3585-6733

Conclusion

This project began with the hypothesis that communities built around domain repositories provided fertile ground for adoption of identifiers across the repositories 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, each reported on in a blog post in this series:

1. 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 (Figures 1 and 2), 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%.
2. Identifying Organizations – initially none of the organization in the repository had identifiers in the metadata. Fortunately, RORs could be identified for all of 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 ~14%.
3. 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.
4. 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.
5. 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.

erin@metadatagamechangers.com
https://orcid.org/0000-0001-9998-0114
ted@metadatagamechangers.com
https://orcid.org/0000-0003-3585-6733

Other ORCIDs

The approach we have used builds on the connections between researchers and datasets in the UNAVCO repository or between researchers and published papers. These connections are created and curated by UNAVCO or by journals that publish the papers and this curation provides confidence in the assertions of relationships between the researchers, organizations, and the identifiers and, in some cases, multiple connections that also support these assertions.

Figures 1 and 2 show that roughly half of the datasets in the repository are still missing identifiers of any kind and we know that many individuals listed in the repository are still missing identifiers or affiliations that can be harvested for RORs or other organizational identifiers.

The ORCID registry can be searched for ORCIDs using researcher names either manually or using the ORCID API. The analysis described in this series identifies researchers that have made significant contributions to UNAVCO datasets or scientific results in papers. There are eleven researchers referenced over 100 times each in the repository for which ORCIDs were not found through the process described above and eight of those have ORCIDs that could be identified manually. Together these occurred 1490 times in the repository, raising the complete or partial ORCID connectivity to 84%. This suggests that automating this search and applying it to 293 remaining researchers that occur over 3300 times in the repository will add significantly to the overall connectivity.

erin@metadatagamechangers.com
https://orcid.org/0000-0001-9998-0114
ted@metadatagamechangers.com
https://orcid.org/0000-0003-3585-6733