Data-Driven Community Building: Measuring and Improving Connectivity in Domain Repositories

Erin Robinson[1] and Ted Habermann[2]

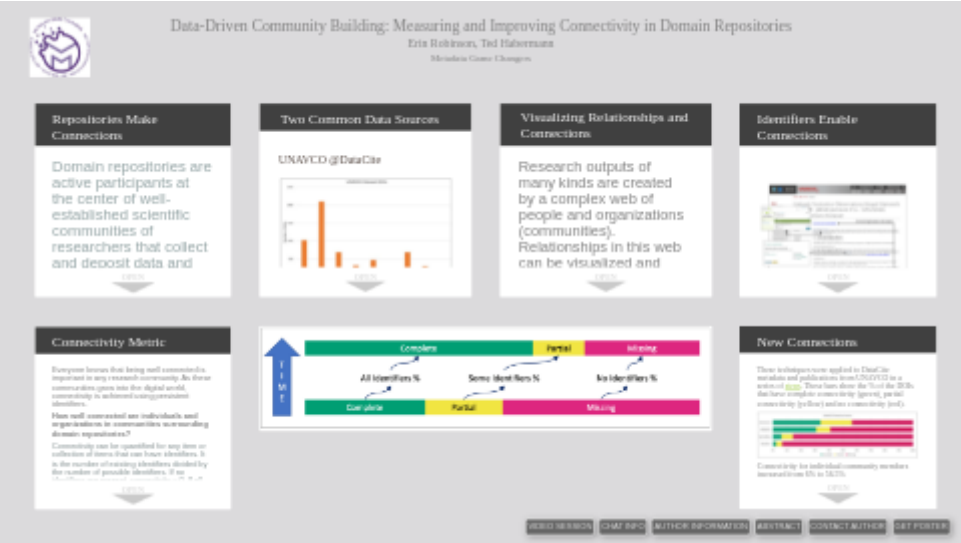
[1]Self Employed
[2]Metadata Game Changers

November 24, 2022

Abstract

Domain repositories can be an integral part of extensive community support systems that extend 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. These long-term relationships are reflected in multiple contributions (data, software, results, papers, . . .) by community members and recognizing these contributions should be an important community-building best-practice for these repositories. Identifiers for people and organizations are critical for recognizing community members and, equally important, for making connections between them and all of the various objects in the research ecosystem. This Figure demonstrates connections that can made once identifiers are integrated into the research ecosystem. Most domain repositories provide DOIs for datasets in the repository. The metadata for those DOIs can include identifiers for some authors (ORCIDs) along with names of organizations they are affiliated with (affiliations). In practice, most authors in these metadata records do not have ORCIDs but, if they have an ORCID once, that ORCID can be spread across all of the datasets they have contributed to, increasing connectivity across the repository. Affiliations can also be spread across multiple contributions, with some caveats. If identifiers (i.e. RORs) exist and can be found for the affiliated organizations, they can be inserted into the metadata, again increasing connectivity. Many domain repositories maintain lists of research papers that have used data from the archive. Metadata for these papers also provide a potential source for identifiers and affiliations. These can also be harvested and spread across the repository, again improving connectivity. These ideas and techniques were applied to UNAVCO, a repository for data related to geodesy with a well-developed community with over 5000 archived datasets. The connectivity for the repository is below 10% for dataset contributors and 0% for RORs. Applying these techniques can increase the connectivity to 56% for contributors and 49% for RORs.

Data-Driven Community Building: Measuring and Improving Connectivity in Domain Repositories



Erin Robinson, Ted Habermann

Metadata Game Changers

PRESENTED AT:

REPOSITORIES MAKE CONNECTIONS

Domain repositories are active participants at the center of well-established scientific communities of researchers that collect and deposit data and publish results based on those data in a wide variety of forms.

These repositories serve the important role of facilitating connections across their communities through active websites, trainings and workshops, town halls, advisory groups, and other means.

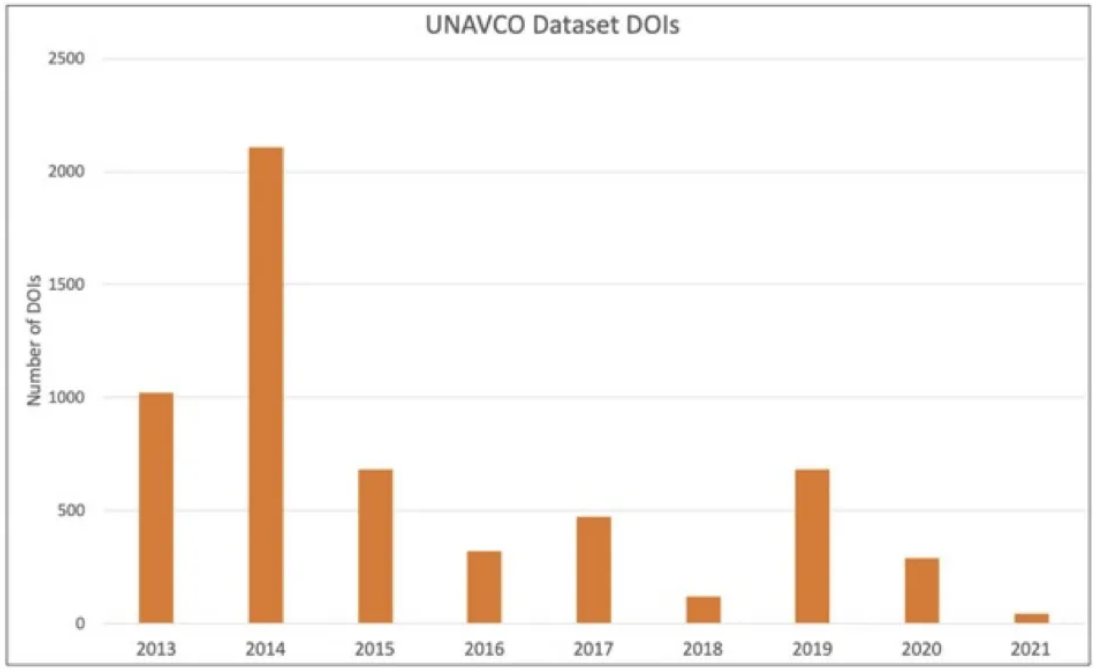
Making these connections in the ever-growing universe of scientific communications requires persistent and unambiguous identifiers for all people, organizations, and resources in the community.

Adoption of these identifiers provides an exciting opportunity for domain repositories to continue to carry their connector role into the future.

We demonstrate how existing metadata and publication lists can help kick-start the adoption process and propose a metric for measuring progress.

TWO COMMON DATA SOURCES

UNAVCO @DataCite



UNAVCO has minted over 5000 dataset DOIs with DataCite (https://datacite.org/) between 2013 and 2021. UNAVCO maintains an in-house archive of these datasets with extensive metadata for discovery, access, and understanding, so the primary roles of the DataCite repository are minting DOIs for identification and citation of the datasets and facilitating connections to people and organizations.

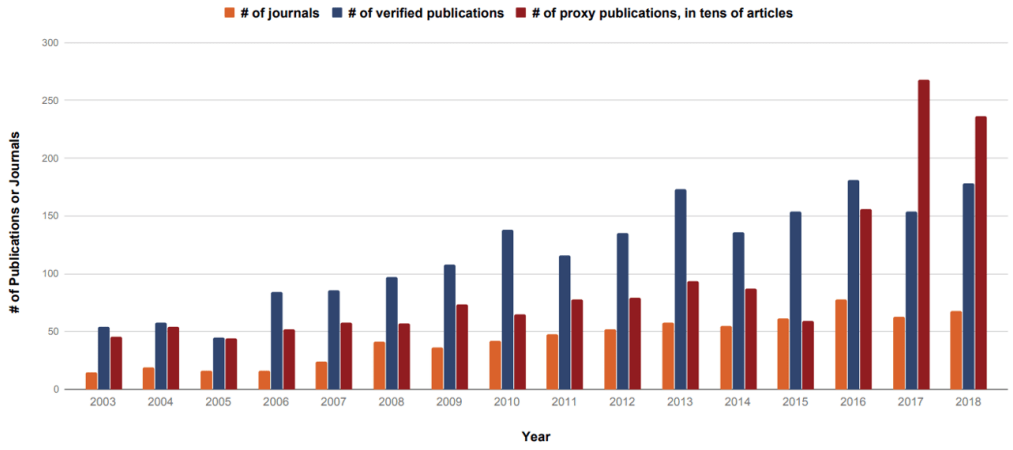
Community Publications

GAGE
National Science Foundation's Geodetic Facility for the Advancement of Geoscience

UNAVCO

Peer-reviewed publications using UNAVCO- or GAGE-related facilities and data 2003-2018

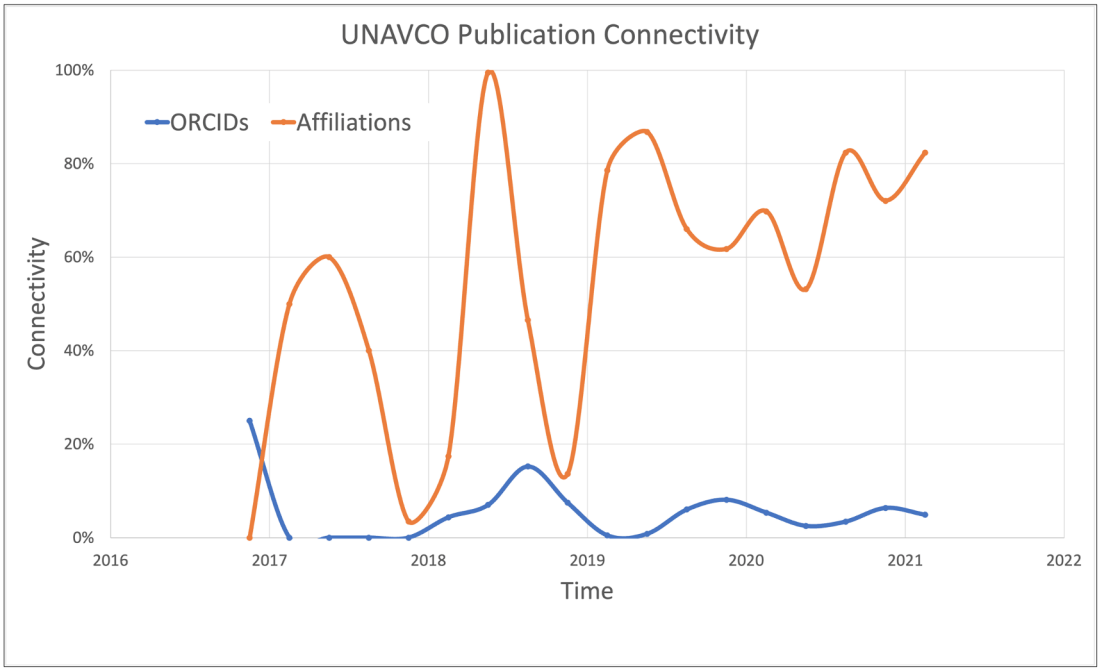
Broad impact of UNAVCO and GAGE resources as reflected by peer-reviewed publications



UNAVCO maintains a list of over 1500 papers that have been published using UNAVCO data. Metadata for these papers from Crossref (https://www.crossref.org/) also include authors with identifiers and

affiliations.

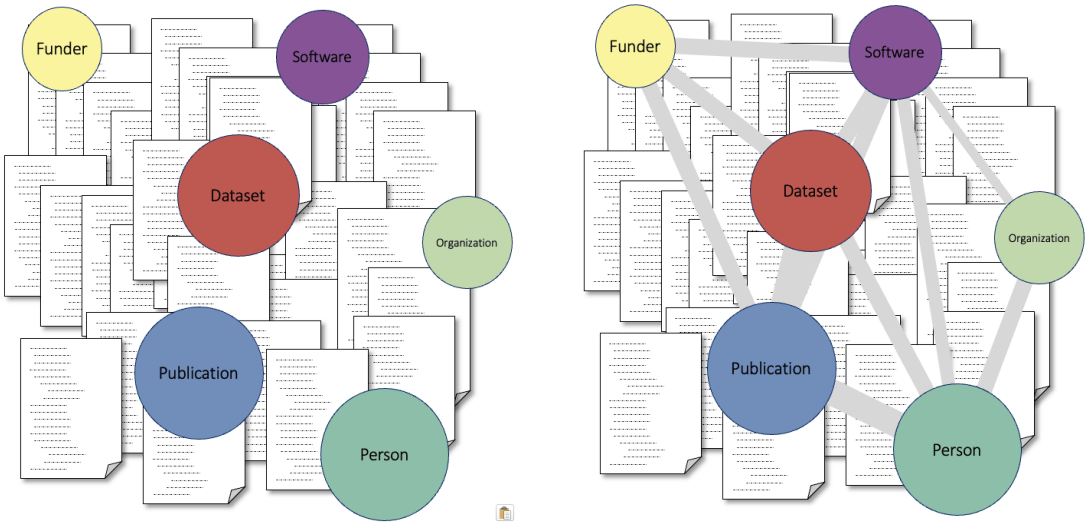
Publication Connectivity



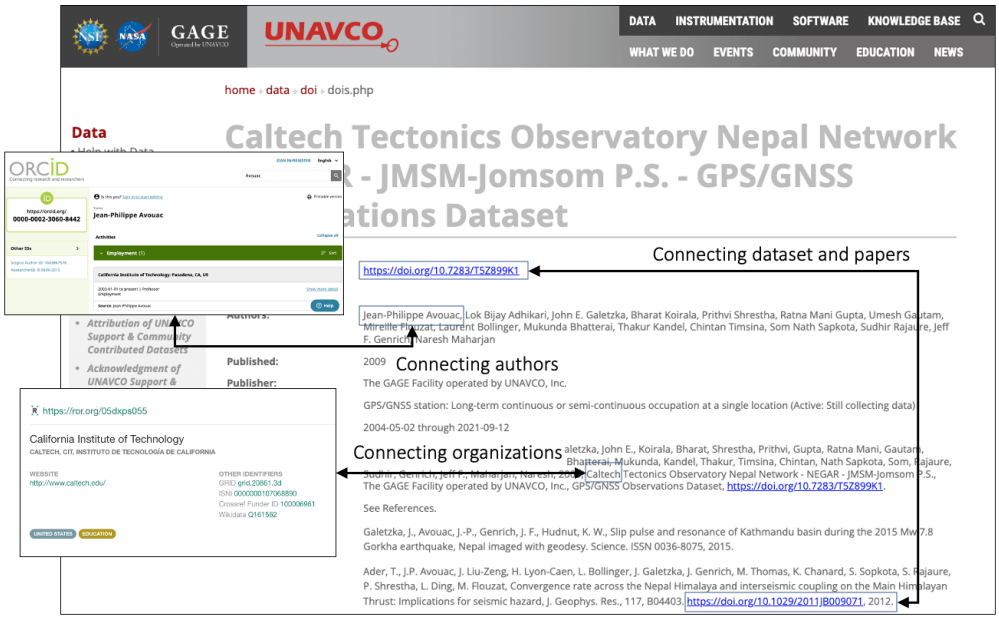
Papers have affiliation information for all authors and identifiers for only one (the corresponding author), limiting connectivity for authors.

VISUALIZING RELATIONSHIPS AND CONNECTIONS

Research outputs of many kinds are created by a complex web of people and organizations (communities). Relationships in this web can be visualized and analyzed using the PID Graph (https://blog.datacite.org/using-jupyter-notebooks-with-graphql-and-the-pid-graph/).



IDENTIFIERS ENABLE CONNECTIONS



Persistent Identifiers (PIDs) are the glue that makes unambiguous connections across the web of research objects possible.

We focus on three types of identifiers: DOIs, ORCIDs, and RORs.

CONNECTIVITY METRIC

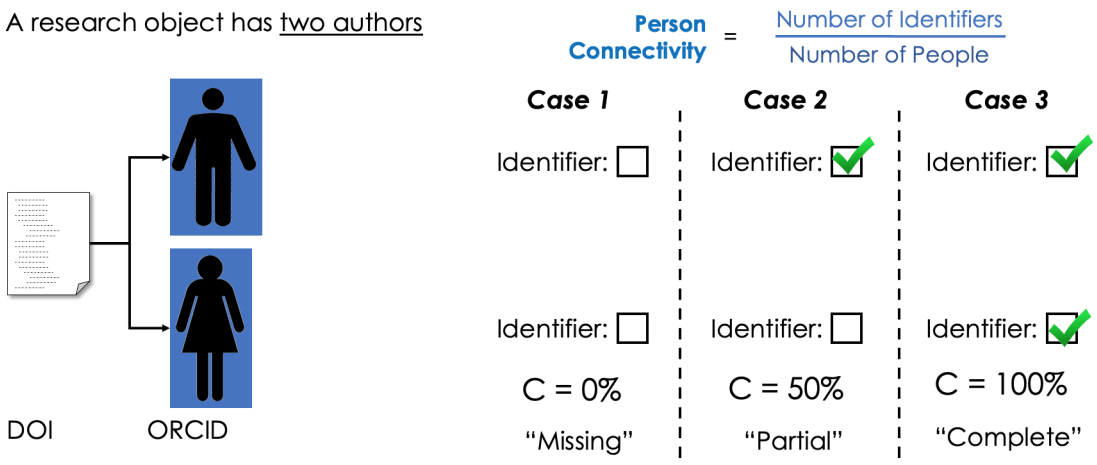
Everyone knows that being well connected is important in any research community. As these communities grow into the digital world, connectivity is achieved using persistent identifiers.

How well connected are individuals and organizations in communities surrounding domain repositories?

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.

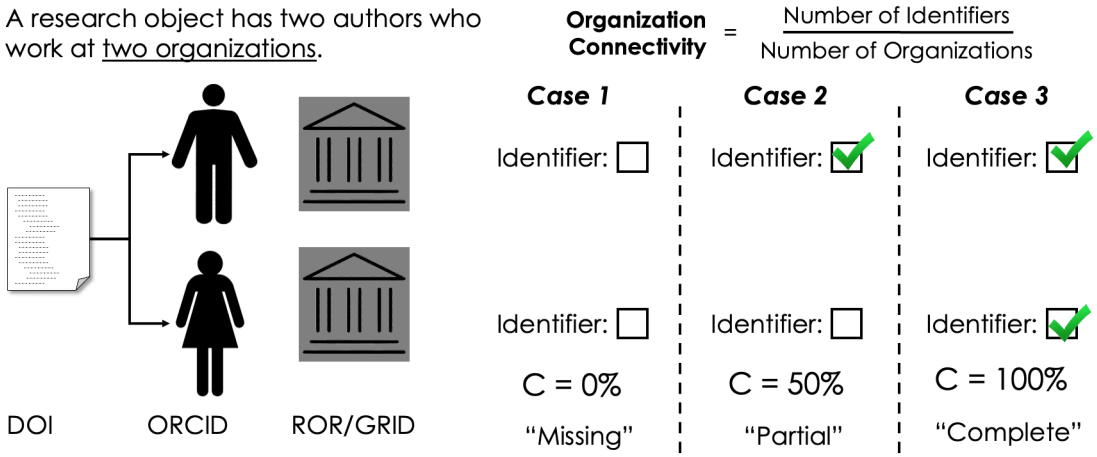
Connectivity for people:

ORCIDs are identifiers for people and this resource has two authors. Connectivity can be 0 (no ORCIDs), 0.5 (1 ORCID), or 1 (2 ORCIDs).



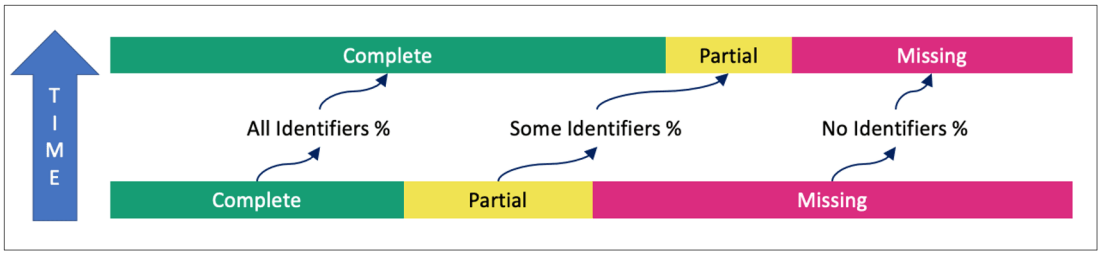
For Organinzations:

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

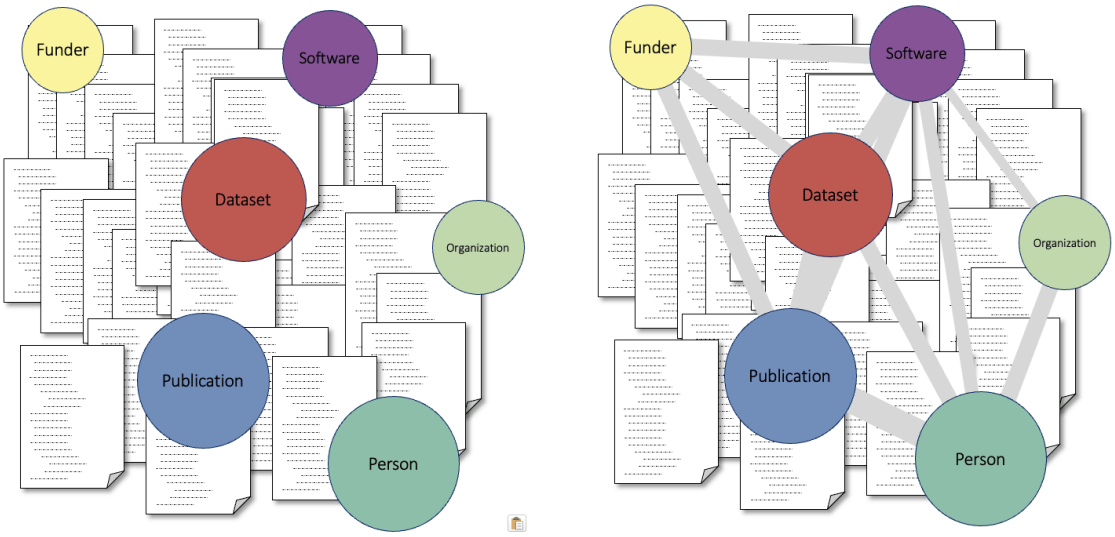


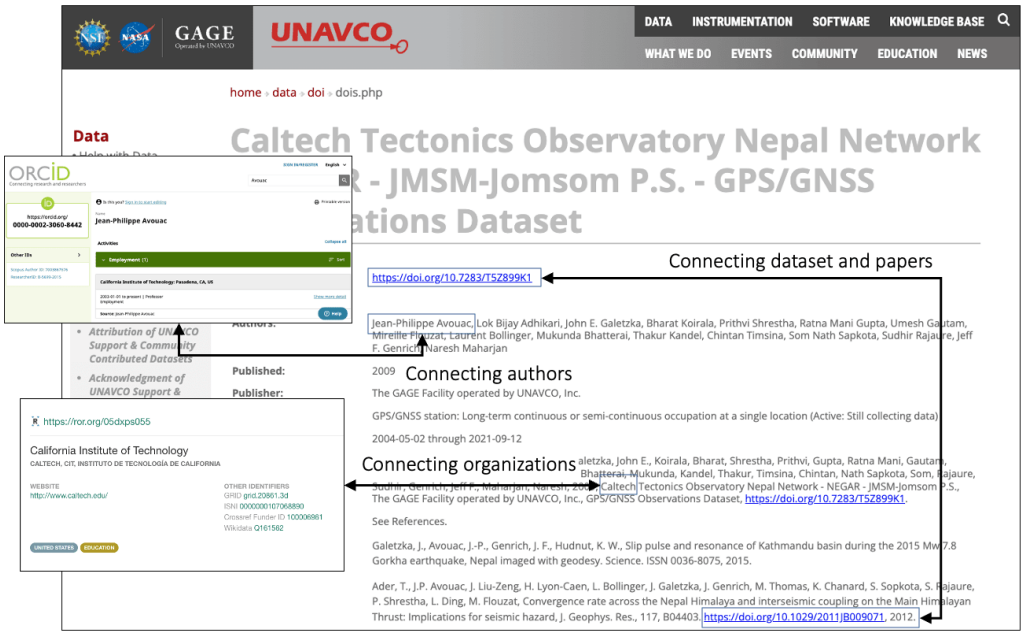
Visualizing Connectivity:

Tracking improvements in connectivity is easier if we can make it a picture. This can be done using a horizontal bar which represents the entire collection and color, with green sections of the bar for items that have complete connectivity, yellow for items that have partial connectivity, and red for items that have no connectivity.

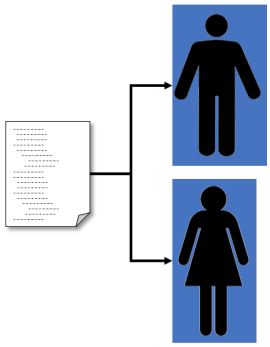


As connectivity improves, the red part of the bar gets smaller!





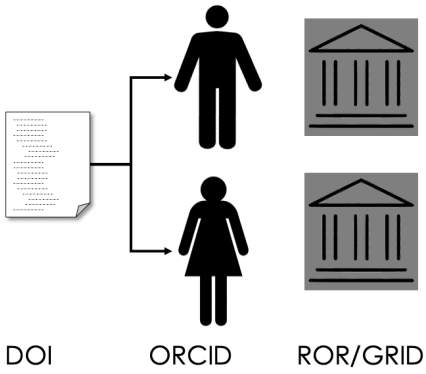
A research object has <u>two authors</u>



Person Connectivity = Number of Identifiers / Number of People

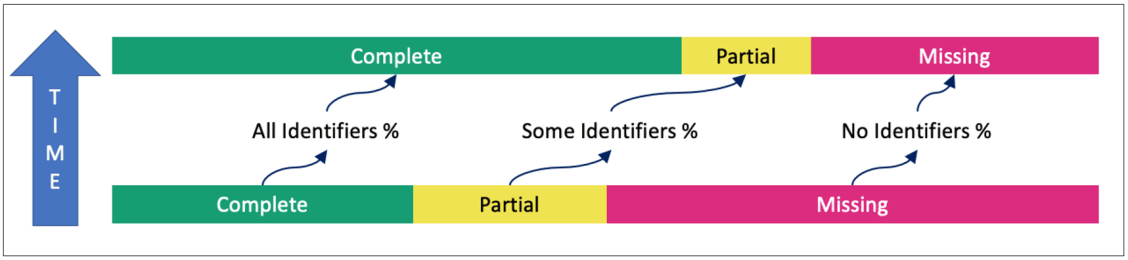
Case 1	Case 2	Case 3
Identifier: ☐	Identifier: ☑	Identifier: ☑
Identifier: ☐	Identifier: ☐	Identifier: ☑
C = 0%	C = 50%	C = 100%
"Missing"	"Partial"	"Complete"

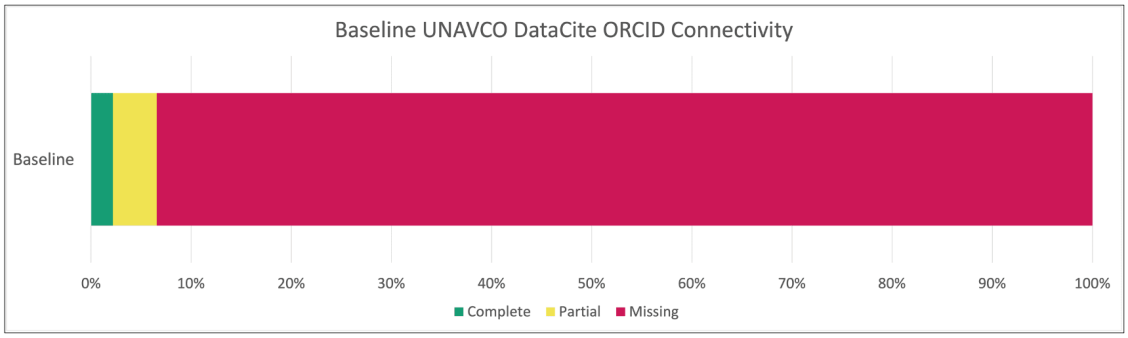
A research object has two authors who work at <u>two organizations</u>.

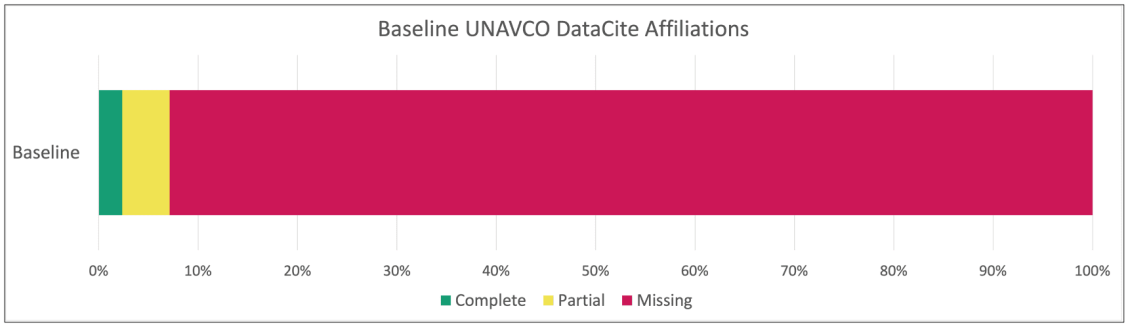


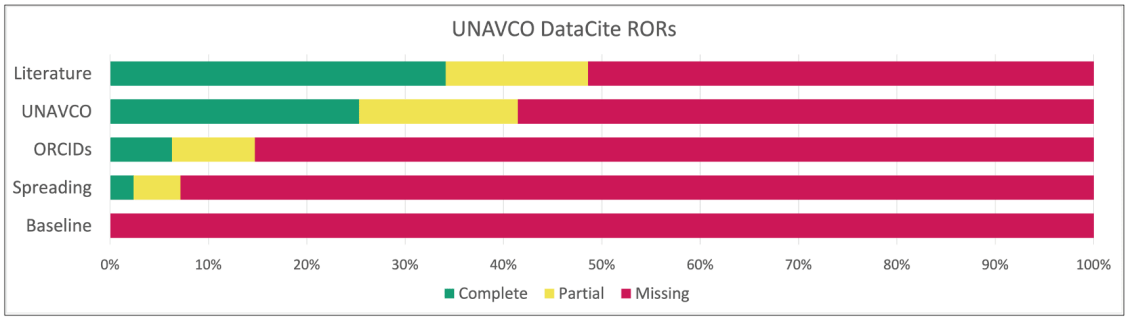
Organization Connectivity = Number of Identifiers / Number of Organizations

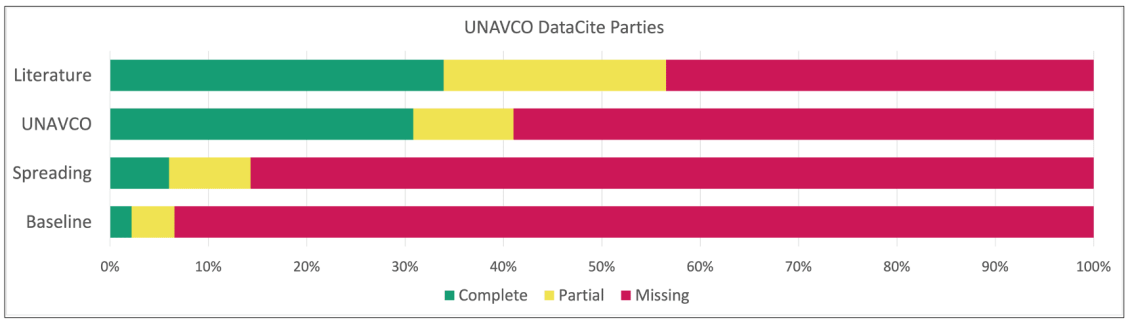
Case 1	Case 2	Case 3
Identifier: ☐	Identifier: ☑	Identifier: ☑
Identifier: ☐	Identifier: ☐	Identifier: ☑
C = 0%	C = 50%	C = 100%
"Missing"	"Partial"	"Complete"

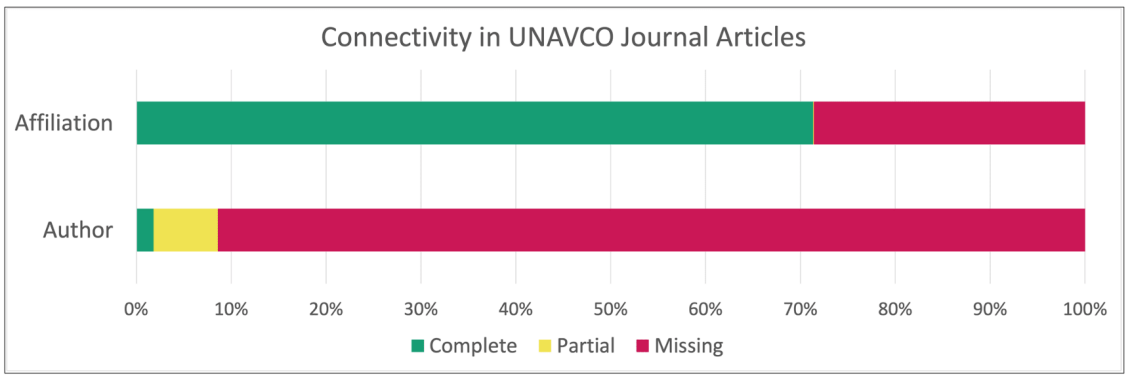


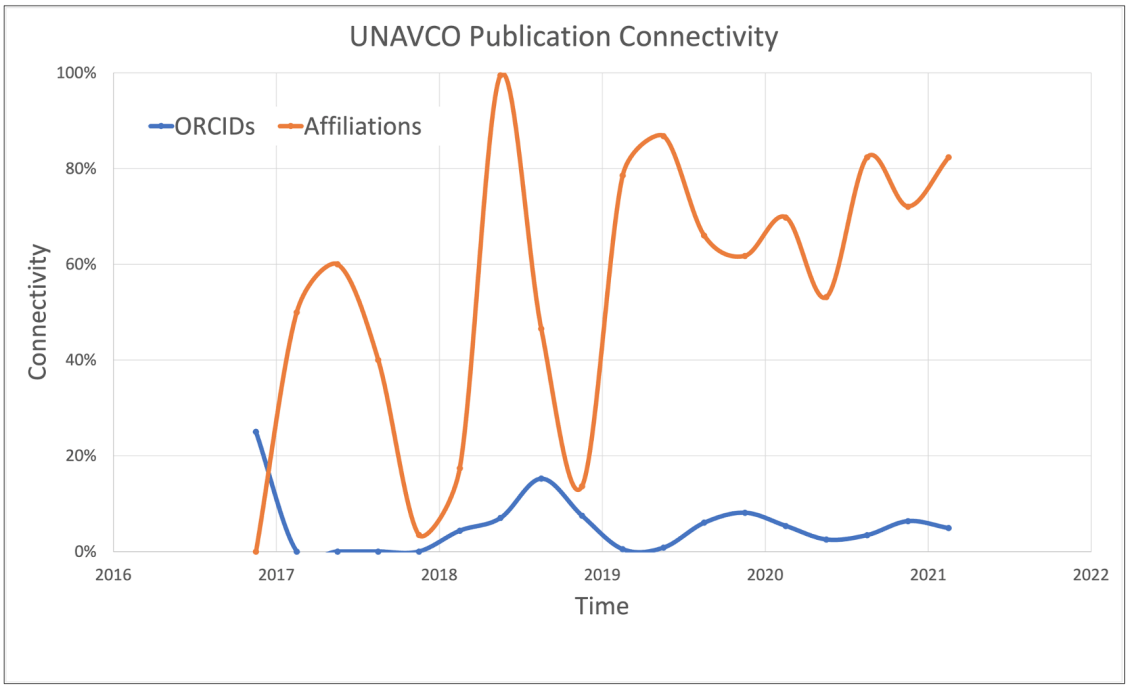





UNAVCO DataCite RORs
Literature
UNAVCO
ORCIDs
Spreading
Baseline
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
Complete
Partial
Missing

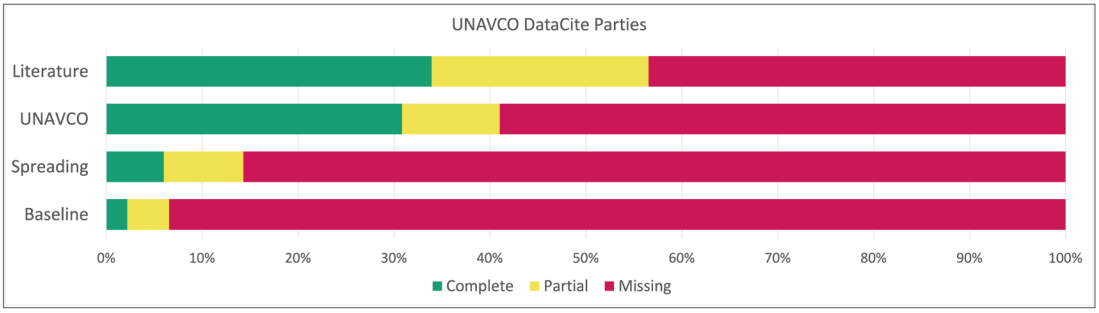

UNAVCO DataCite Parties
Literature
UNAVCO
Spreading
Baseline
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
Complete
Partial
Missing


Connectivity in UNAVCO Journal Articles
Affiliation
Author
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
Complete
Partial
Missing

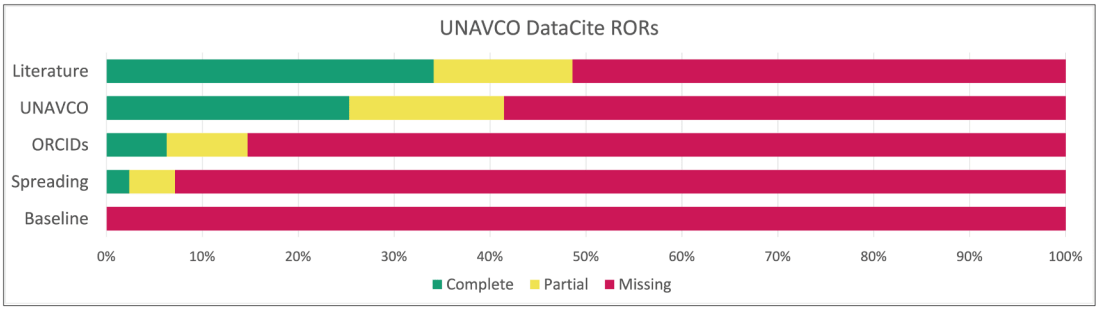

UNAVCO Publication Connectivity
ORCIDs
Affiliations
Connectivity
100%
80%
60%
40%
20%
0%
2016
2017
2018
2019
2020
2021
2022
Time

NEW CONNECTIONS

These techniques were applied to DataCite metadata and publications from UNAVCO in a series of steps (https://metadatagamechangers.com/blog/2021/5/14/improving-domain-repository-connectivity-closing-the-circle). These bars show the % of the DOIs that have complete connectivity (green), partial connectivity (yellow) and no connectivity (red).



Connectivity for individual community members increased from 6% to 56.5%



Connectivity for cummunity members that are organizations increased from 0% (no organization identifiers in initial repository) to 48.6%.

AUTHOR INFORMATION

Erin Robinson (0000-0001-9998-0114) and Ted Habermann (0000-0003-3585-6733), Metadata Game Changers (https://ror.org/05bp8ka05)

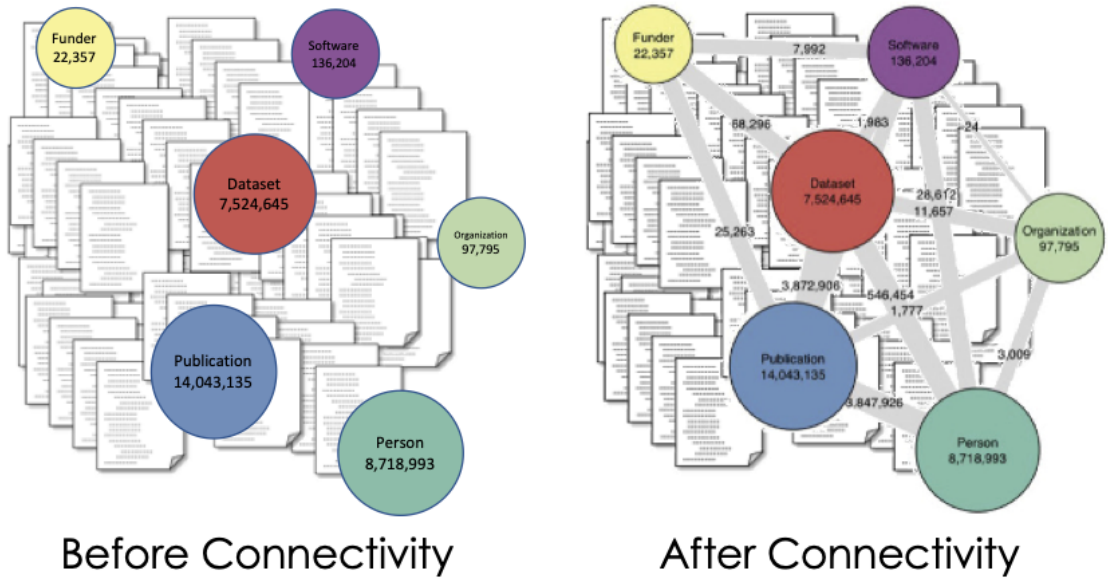
ABSTRACT

Domain repositories can be an integral part of extensive community support systems that extend 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. These long-term relationships are reflected in multiple contributions (data, software, results, papers, ...) by community members and recognizing these contributions should be an important community-building best-practice for these repositories. Identifiers for people and organizations are critical for recognizing community members and, equally important, for making connections between them and all of the various objects in the research ecosystem. This Figure demonstrates connections that can made once identifiers are integrated into the research ecosystem.

Most domain repositories provide DOIs for datasets in the repository. The metadata for those DOIs can include identifiers for some authors (ORCIDs) along with names of organizations they are affiliated with (affiliations). In practice, most authors in these metadata records do not have ORCIDs but, if they have an ORCID once, that ORCID can be spread across all of the datasets they have contributed to, increasing connectivity across the repository. Affiliations can also be spread across multiple contributions, with some caveats. If identifiers (i.e. RORs) exist and can be found for the affiliated organizations, they can be inserted into the metadata, again increasing connectivity.

Many domain repositories maintain lists of research papers that have used data from the archive. Metadata for these papers also provide a potential source for identifiers and affiliations. These can also be harvested and spread across the repository, again improving connectivity.

These ideas and techniques were applied to UNAVCO, a repository for data related to geodesy with a well-developed community with over 5000 archived datasets. The connectivity for the repository is below 10% for dataset contributors and 0% for RORs. Applying these techniques can increase the connectivity to 56% for contributors and 49% for RORs.



(https://agu.confex.com/data/abstract/agu/fm21/1/3/Paper_891031_abstract_822676_0.png)