

Tracking data reuse:  
following  
one thousand datasets  
from repositories  
into the published literature

Have the potential benefits  
been realized?

Are the data sets reused?

Have they saved money?

Enabled new science?

# Research Questions

- How often is data from repositories used in the published literature?  
What is the distribution of use across datasets and time?
- Who reuses data?  
Are investigators who reuse repository datasets similar to investigators who deposit data?
- What is data reused for?  
How similar are studies that reuse data to studies that deposit data?

$$10 \times 100 = 1000$$

# Choosing repositories

- navigable in English,
- publicly available, and
- established prior to 2004.
- support querying datasets by deposit date
- allow link to data collecting publication
- Nice to have:
  - offer accession numbers or unique IDs that have an easily queriable prefix (dois, hdls, or an alpha-numeric string)
  - diversity across domain
  - diversity across repository type

# Choosing repositories

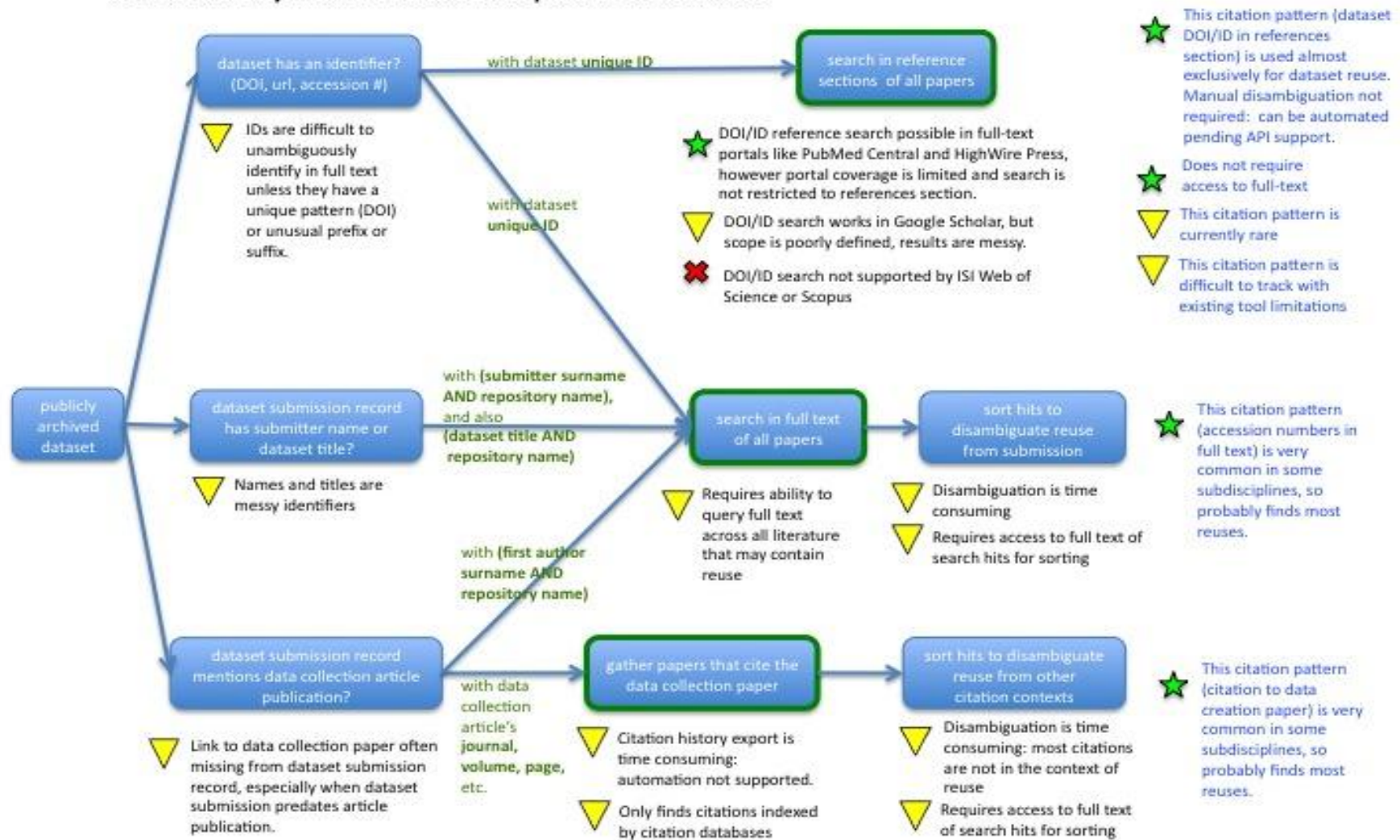
- IQSS Dataverse network
- Oak Ridge National Laboratory Distributed Active Archive Center (ORNL DAAC)
- PANGAEA® – Publishing Network for Geoscientific and Environmental Data
- NCBI's Gene Expression Omnibus
- EBI's ArrayExpress

# Tracking

1. Manually inspect citations to the data collection paper using Scopus or ISI Web of Science.
2. Query for the dataset accession number or unique identifier in the bibliography of published papers using Scopus or ISI Web of Science.
3. Query for the dataset accession number or unique identifier in the full text of published literature. Google Scholar plus one or more additional portals will be used, chosen based on the repository discipline (for example, PubMed Central and Highwire Press will be searched for biomedical repository IDs).
4. Query for both the repository name and the author's name in the full text of published literature, using the same portals as above.

# Tracking

## How to identify Dataset Reuse in the published literature



✗ This flow still misses attributions embedded in supplementary information, reuses attributed through a query description, etc.



# Timeline

IASSIST  
ASIS&T

# Budget

\$1000 SIGUSE

\$2-3k

UBC Work/study

# Hurdle:

## LOTS of citations

Querying by year of submission,  
Filtering to those with associated journal articles

Hurdle:

Citation tools don't support  
nontraditional products

# Hurdle:

## Querying repositories


Querying by year of submission,  
Filtering to those with associated journal articles


10.3334/ORNLDAAAC/\* -site:ornl.gov

### Field Tags


**TS**=Topic

**TI**=Title

**AU**=Author 

**GP**=Group Author 

**ED**=Editor

**SO**=Publication Name 

**PY**=Year Published

**CF**=Conference

**AD**=Address

**OG**=Organization

**SG**=Suborganization

**SA**=Street Address

**CI**=City

**PS**=Province/State

**CU**=Country

**ZP**=Zip/Postal Code

**FO**=Funding Agency

**FG**=Grant Number

**FT**=Funding Text

Quick Search

Scopus: 1

[More...](#)

[Web](#)

[Patents](#)

Your query: ALL(10.3334/ornlidaac\*)

[Edit](#)

[Save](#)

[Save as Alert](#)

[RSS](#)

[Search](#)

## Refine Results



Results: 1

[Show all abstracts](#)

Search within results



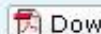
Output



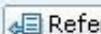
Citation tracker



Add to list



Download



References



Cited by

Select:



All



Page

Page 1 of 1

Document (sort by relevance)

Author(s)

▼ Date

Source Title

1. **Impacts of experimentally imposed drought on leaf respiration and morphology in an Amazon rain forest**

[Abstract + Refs](#)

[Check Article](#)

[Availability](#)

[Metcalf, D.B.](#), [Lobo-do-Vale, R.](#), [Chaves, M.M.](#), [Maroco, J.P.](#), [Aragão, L.E.O.C.](#), [Malhi, Y.](#), [Da Costa, A.L.](#), (...), [Meir, P.](#)

2010

[Functional Ecology](#) 24 (3), pp. 524-533



- Randel, W. J., Park, M., Emmons, L., Kinnison, D., Bernath, P., Walker, K. A., Boone, C., and Pumphrey, H.: Asian Monsoon Transport of Pollution to the Stratosphere, *Science*, doi:10.1126/science.1182274, 2010.
- Randerson, J. T., van der Werf, G., Giglio, L., Collatz, G., and Kasibhatla, P.: Global Fire Emissions Database, Version 2 (GFEDv2.1), doi:10.3334/ORNLDAAAC/849, 2007.
- Sanjeeva Rao, P.: Arabian Sea Monsoon Experiment: An Overview, *Mausam*, 56, 1–6, 2005.

- ✓ 54.   RANDEL WJ  
Asian Monsoon Transport of Pollution to the Stratosphere  
SCIENCE 328 : 611 DOI 10.1126/science.1182274 2010
- ✓ 55.   RANDERSON JT  
GLOBAL FIRE EMISSION : 2007
- ✓ 56.   RAO PS  
MAUSAM 56 : 1 2005

Thanks!