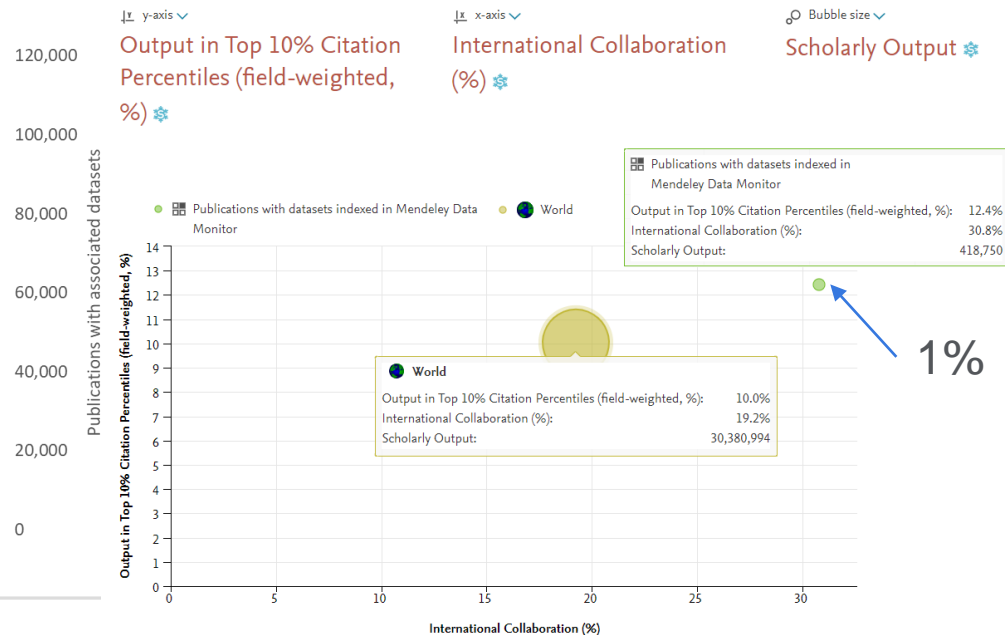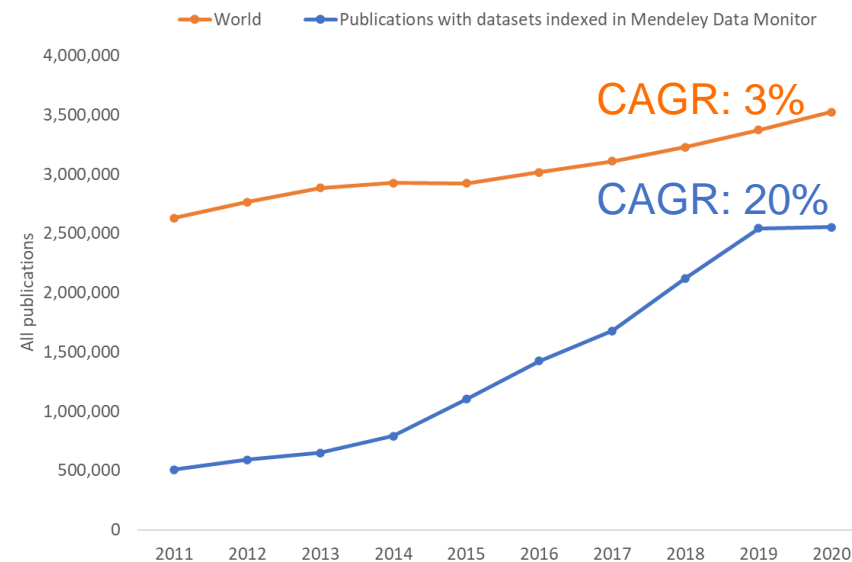# Session II: Exploring the current state of data citation methods and tools

Changing the Culture of Data Management and Sharing

A Workshop

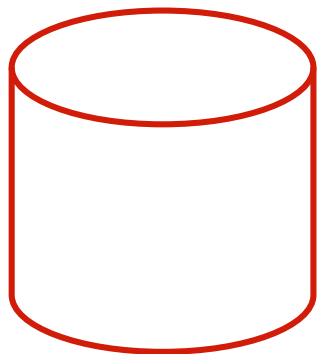28 – 29 April 2021

Alberto Zigoni
Market Development Director, Research Data Management Solutions, Elsevier BV

The practice of linking articles and datasets is growing and is associated with higher collaboration and citation impact

CAGR: 3%

CAGR: 20%

Output in Top 10% Citation Percentiles (field-weighted, %)

International Collaboration (%)

Scholarly Output

Publications with datasets indexed in Mendeley Data Monitor

Output in Top 10% Citation Percentiles (field-weighted, %): 12.4%
International Collaboration (%): 30.8%
Scholarly Output: 418,750

World

Output in Top 10% Citation Percentiles (field-weighted, %): 10.0%
International Collaboration (%): 19.2%
Scholarly Output: 30,380,994

1%

Source: SciVal, Data Monitor. Data extracted on April 21, 2021 - CAGR: Compound Annual Growth Rate

2

But where is your data?

~90%
of the public research data is hosted on external general subject or domain repositories

~10%
of public research data is hosted on institutional repositories

Dr. Andrew White
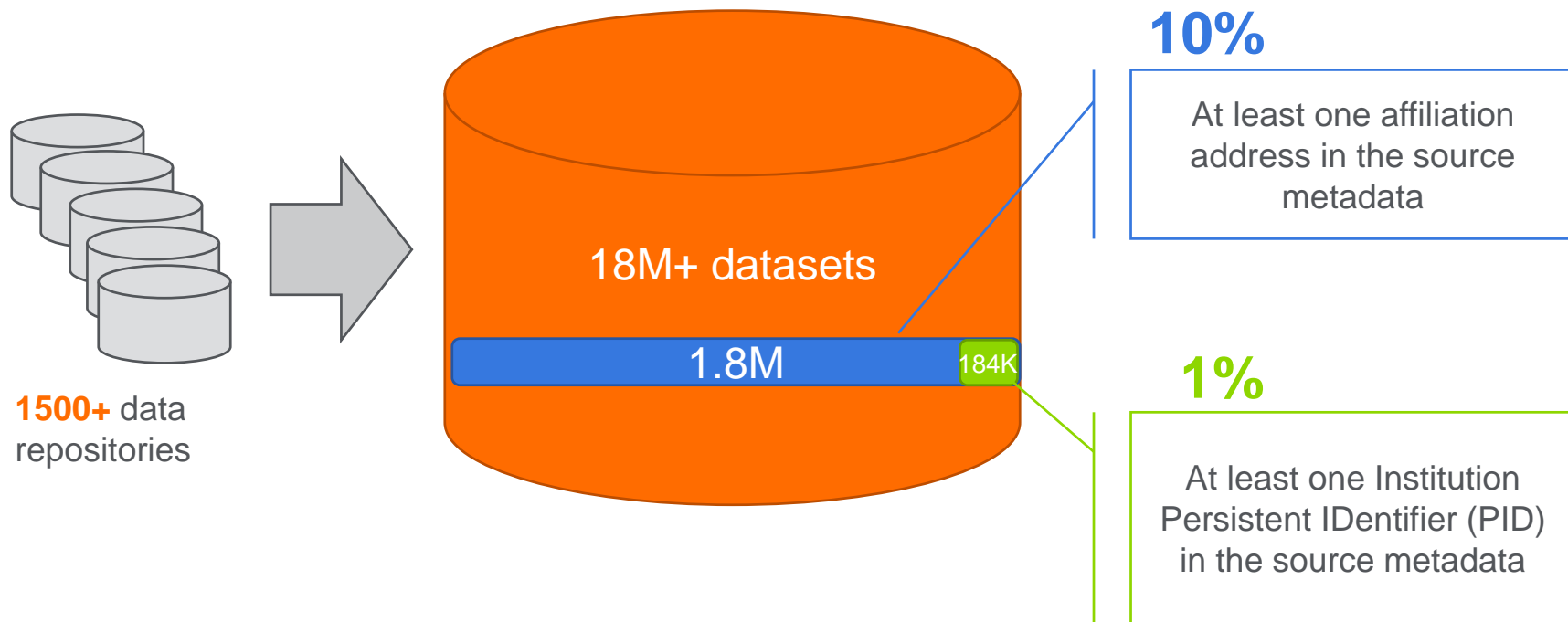Director of Library Information Services at Rensselaer Polytechnic Institute

"There is a broad spectrum of RDM practices used by our researchers, often due to a multitude of data types and a range of discipline-specific practices.

We see the libraries role as one that is focused on tracking archival practices of raw research data sets, assigning appropriate ontologies/metadata, providing appropriate linking and checking that the data has Been stored in a secure location. "

Source: https://doi.org/10.1051/itmconf/20203301003

Why is it difficult to track datasets?

1500+ data repositories

18M+ datasets

1.8M   184K

10%

At least one affiliation address in the source metadata

1%

At least one Institution Persistent IDentifier (PID) in the source metadata

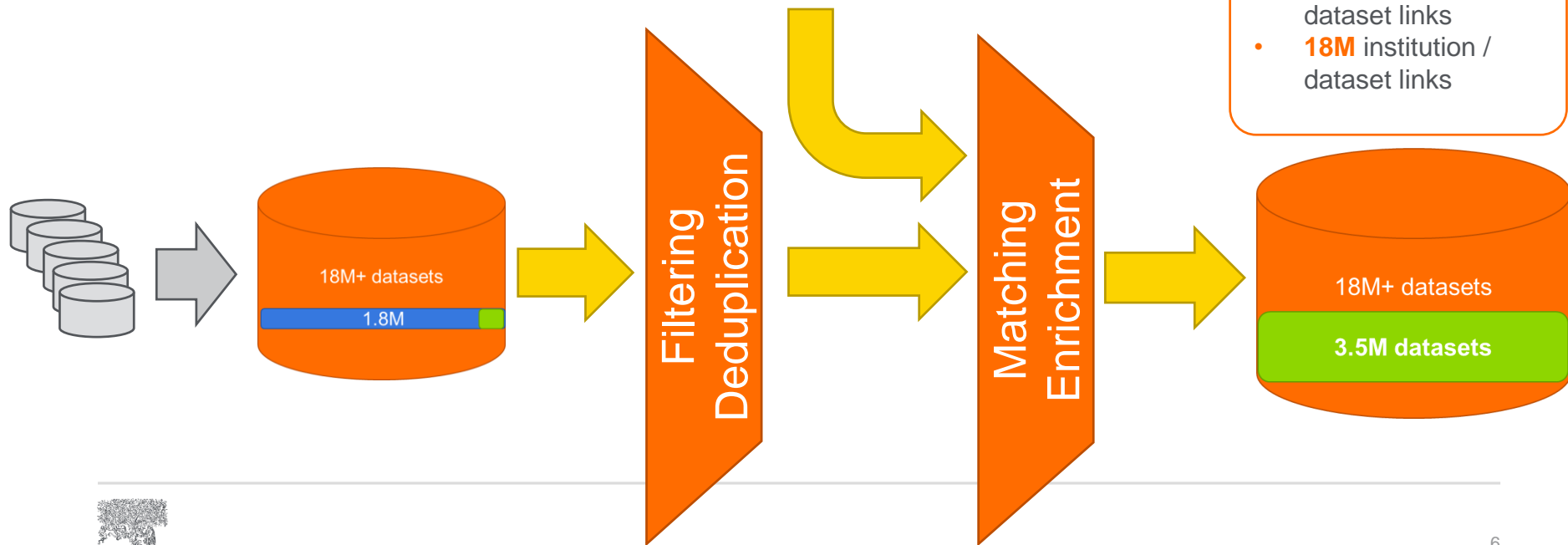# A great example of adoption of PIDs for institutions

- Dryad has consistently started to use ROR identifiers for institutions
- Affiliation metadata is significantly better than other repositories
- However, there are lot of repositories where researchers publish their data…

5

Using Data Science to enrich dataset metadata

Scopus® SciVal

- Publication metadata
- Author profiles
- Institution profiles

- **3.5M** datasets with at least one institution PID
- **14M** author / dataset links
- **18M** institution / dataset links

18M+ datasets

1.8M

Filtering Deduplication

Matching Enrichment

18M+ datasets

3.5M datasets

The results: 20x increase

Conclusions

- Data sharing is growing, and it contributes to collaboration and academic impact

- Metadata make data count! Adding metadata, including affiliation information, contributes to dataset discoverability and attribution

- Persistent Identifiers, particularly for authors and institutions, are the best approach to attribute datasets and be rewarded for sharing them

- When PIDs are not available, or metadata is scarce, analytical approaches can help improve metadata quality and enable tracking, reporting and showcasing

ELSEVIER

Thank you

Alberto Zigoni

a.zigoni@elsevier.com

Market Development Director

Research Data Management Solutions

Elsevier BV