How to Track Progress towards **Open Data** in **Poland**

Dimas Putra
Senior Product Manager - Elsevier
d.putra@elsevier.com

Nov 2022

Contents

Trends in research data management

Research Data has become a high priority

- **Funders policies**
  Most funders around the world have a policy on Research Data sharing

“No opting out of RDM”

- **Publishers policies**
  All the largest publishers (incl. Elsevier) have a policy on Research Data sharing

- **Government strategic agendas**
  Govts and national bodies around the world

National Science Centre (NCN) - Poland

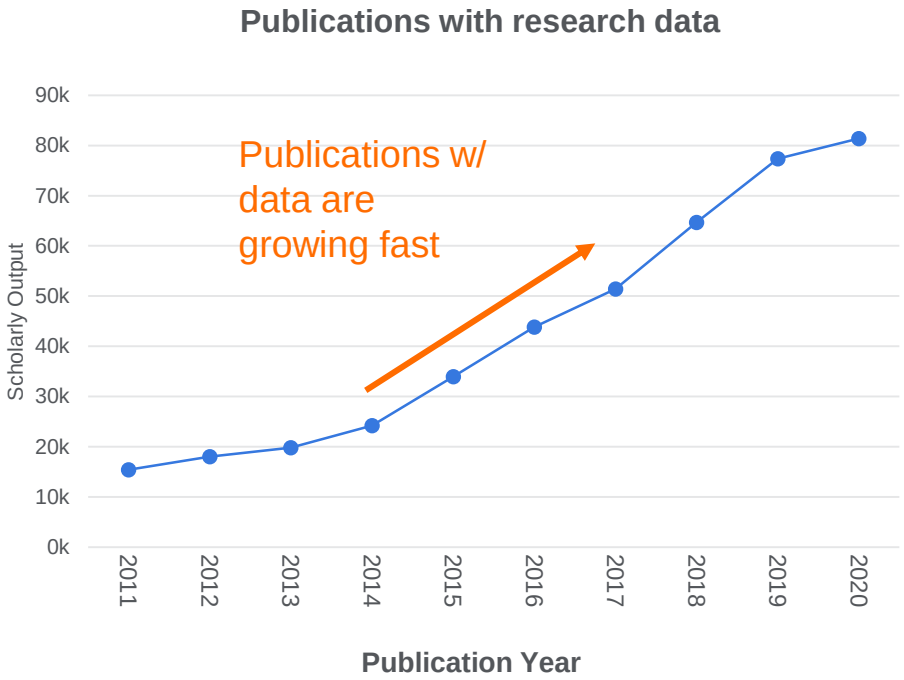
"We have set out to draw up a set of joint European guidelines on research data management and open-access publication methods. A practical guide to uniform European research data management practices was published in January 2019."

National Science Centre
Signed by Zbigniew Błocki, Director

NATIONAL SCIENCE CENTRE
POLAND

Source: https://www.ncn.gov.pl/sites/default/files/pliki/2019_04_03_pismo_dyrektora_NCN_zarzadzanie_danymi_naukowymi_ang.pdf

Publishers mandates on RDM are growing



CellPress

THE LANCET

Taylor & Francis Group
an informa business

F1000Research

Institutions are ramping up RDM efforts

- Hiring data stewards / data librarians
- Data Management Plans
- Data Repository
- Reporting (compliance, assessment, etc) on Data
- Showcasing Data

} Highly strategic

Institutions are ramping up RDM efforts

- Hiring data stewards / data librarians
- Data Management Plans
- Data Repository
- Reporting (compliance, assessment, etc) on Data
- Showcasing Data

} Highly strategic

Challenge: Where is the institutions' research data?

Where is the Research Data

The 90% challenge

Where is the institutions' Research Data?

Up to 90%

- is deposited in any of the 1000s domain-specific / generalist repositories
- This is in accordance to community best practices and funders recommendations (e.g. NIH)

Sources:
1: E. Zudilova-Seinstra; A. Zigoni; W. Haak (2020), "Analysis of research data for 11 Institutions – Data Monitor", doi: 10.17632/k5p45z33kb.3

Finding it manually is an impossible task

"With all the copy and pasting, checking and improving on the metadata, it probably took us around 20-30 minutes a dataset"

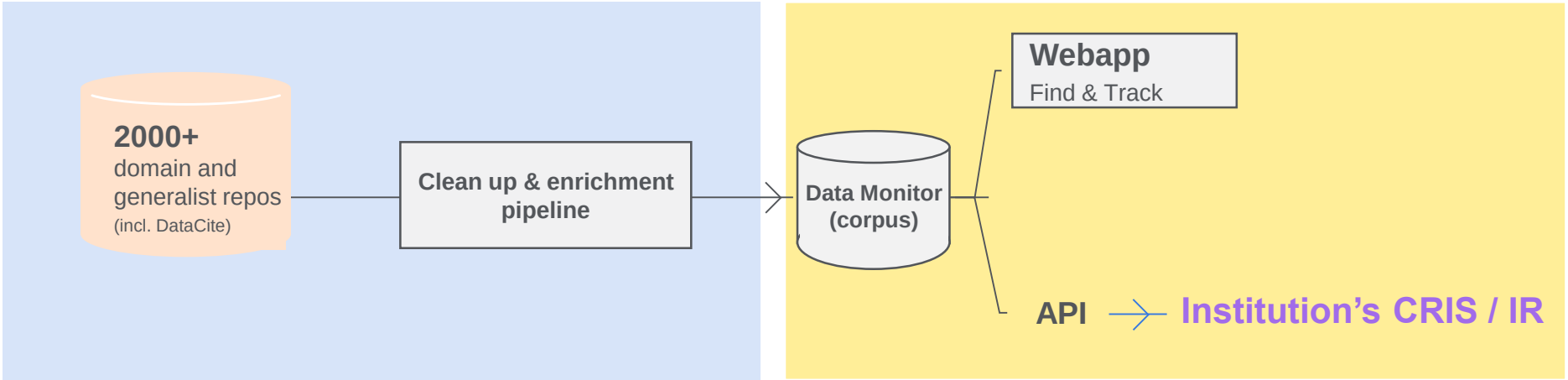
"We knew we had to find an automated solution"

Christina Elsenga
University of Groningen Library

Sources: University of Groningen Case Study

How institutions can start tracking their research data

Data Monitor helps to overcome 90% challenge



How we track the Research Data

- **Harvest** metadata from 2000+ repos
- **Normalize** metadata (OpenAIRE schema)
- **Clean-up** i.e. remove duplicates, non-research data e.g. articles
  - How: ML techniques, integration w/ Scopus, experts
- **Enrich** metadata with publications/ author / institution
  - How: NLP techniques; integration w/ Scopus, Scholix, DataCite

What users see and get:

- Webapp
- API
  - Integration
  - Batch

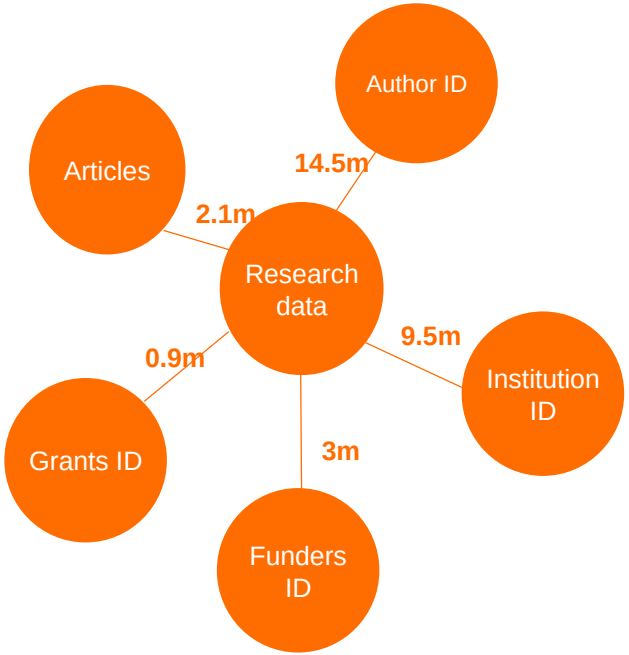
Data Monitor: Ever Improving Coverage, Enrichment, and Quality

Data Enrichment

- **Institutional** affiliations
- **Authors** Disambiguation
- **Metrics**: Citations, Usage... *and more*
- **Publications, Funders, Grants**.... and more
- Related Links...

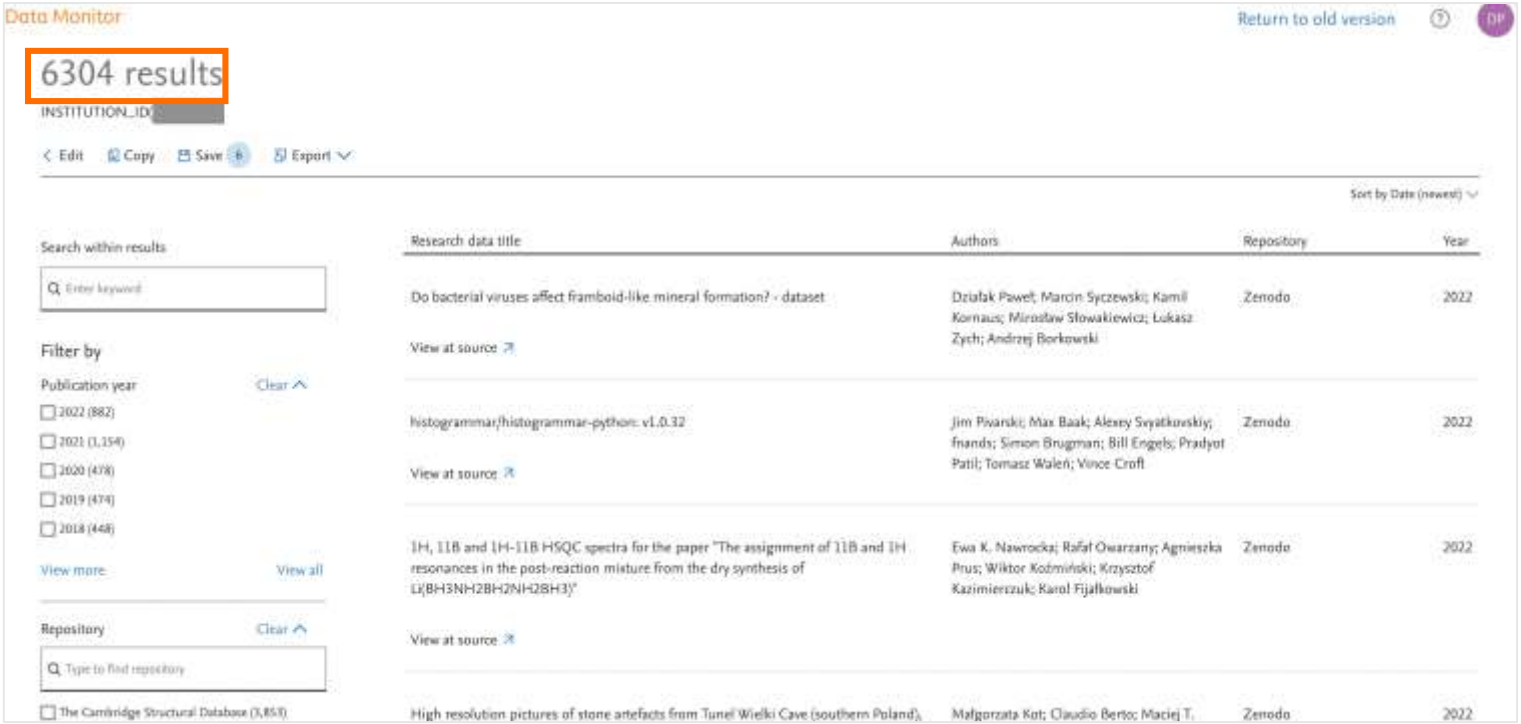
Data Clean-Up:

- Non-research data clean up
- Spam & Dead-links clean up
- De-duplication
- Versioning management



Source: Data Monitor Q2 2022

Example: research data for a leading Polish University



Source: rdmonitor.com

Examples: Pure – Data Monitor Integration

Source: https://research.vu.nl/

Source: https://tmu.pure.elsevier.com/

Case study examples: helping leading universities track research data

	Research data deposited in …	
	… the institutional repository	…all repositories (domain-specific, generalist, institutional)
UK institution (Figshare)	234	16724
NL institution (Figshare)	534	16372
US institution (Dataverse)	46	3951
Chinese institution (home-grown)	1328	71760
Australian institution (DC Data)	11	785

30x- 70x

Source: Data Monitor and institutional repositories public portal (data extracted on June 21, 2022)

Benefits: Measuring and Showcasing Research Data

Measure **progress towards data sharing and Open Data**

> We can make datasets visible. I think that visibility is one of the most important objectives we have, and we can report on the number of datasets grouped by the faculties or research institutes."
>
> - CRIS/RIMS Administrator at a leading university in the Netherlands

Showcase institutional research data

> Why are we doing all this Open Science/Open Data if we cannot showcase it and show the world that we are involved? What's the purpose of all this if nobody uses the data?"
>
> – Open Science leader at a leading institution in Netherlands

Benefits: Compliance and Improving Administration

Track **compliance** with RDM (funders) **policies**

> We can’t force researchers to deposit in our repository because many opt for domain-specific repositories. Instead of "forcing them" having the overview will help us to monitor compliance.”
>
> - Research Admin Leader at a leading university in the United States

Reduce **administrative burden**

> Having the data linked to an output, then having your data come into staff profile via CRIS system and being able to link up to your outputs as well is really nice, that integration is lovely!
>
> – CRIS Research Data Admin at a leading university in the United Kingdom

Data sharing in Poland

About this analysis

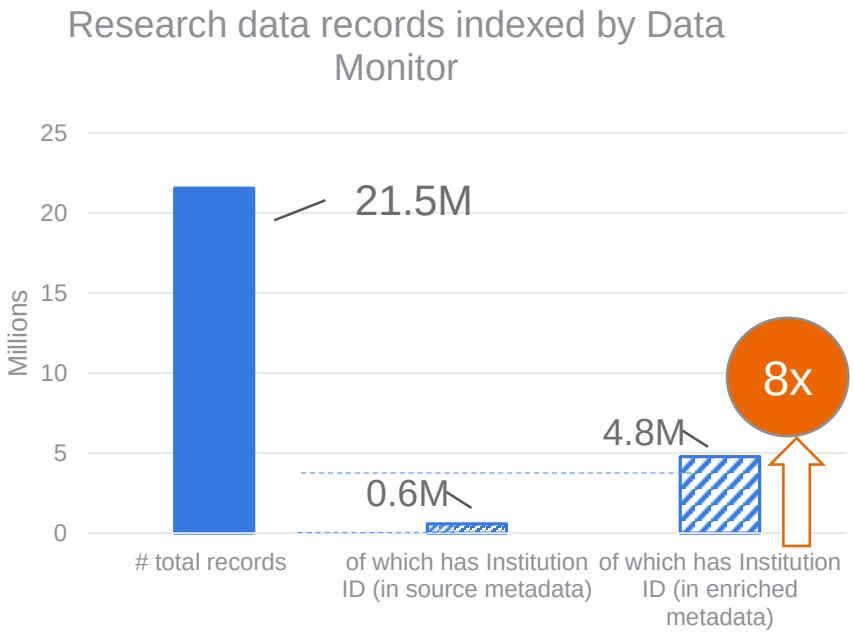
- **Disclaimer**: this is an **exploration** on current data sharing practices

- **Based on Data Monitor**, which indexes research data from **2000+** generalist, domain-specific and institutional repositories (incl. DataCite)
  - https://www.elsevier.com/solutions/data-monitor

- Definition of **research data**: primary/secondary data, raw/processed data, or discipline specific
  - Content types include, for example: dataset, tabular data, images, videos, …

About this analysis

- Data sharing practices not fully mature → **metadata annotation is often incomplete**
  - e.g. often lacking institutional affiliation
- **Data Monitor** aims to address this by **enriching metadata**, thereby providing a more realistic view about data sharing practices
- Reaching complete metadata coverage remains a challenge to tackle, by:
  - **Advocating best practices** among researchers
  - Continue **improving metadata enrichment processes**



Research Data Published by Institutions Indexed by Data Monitor

41.000+

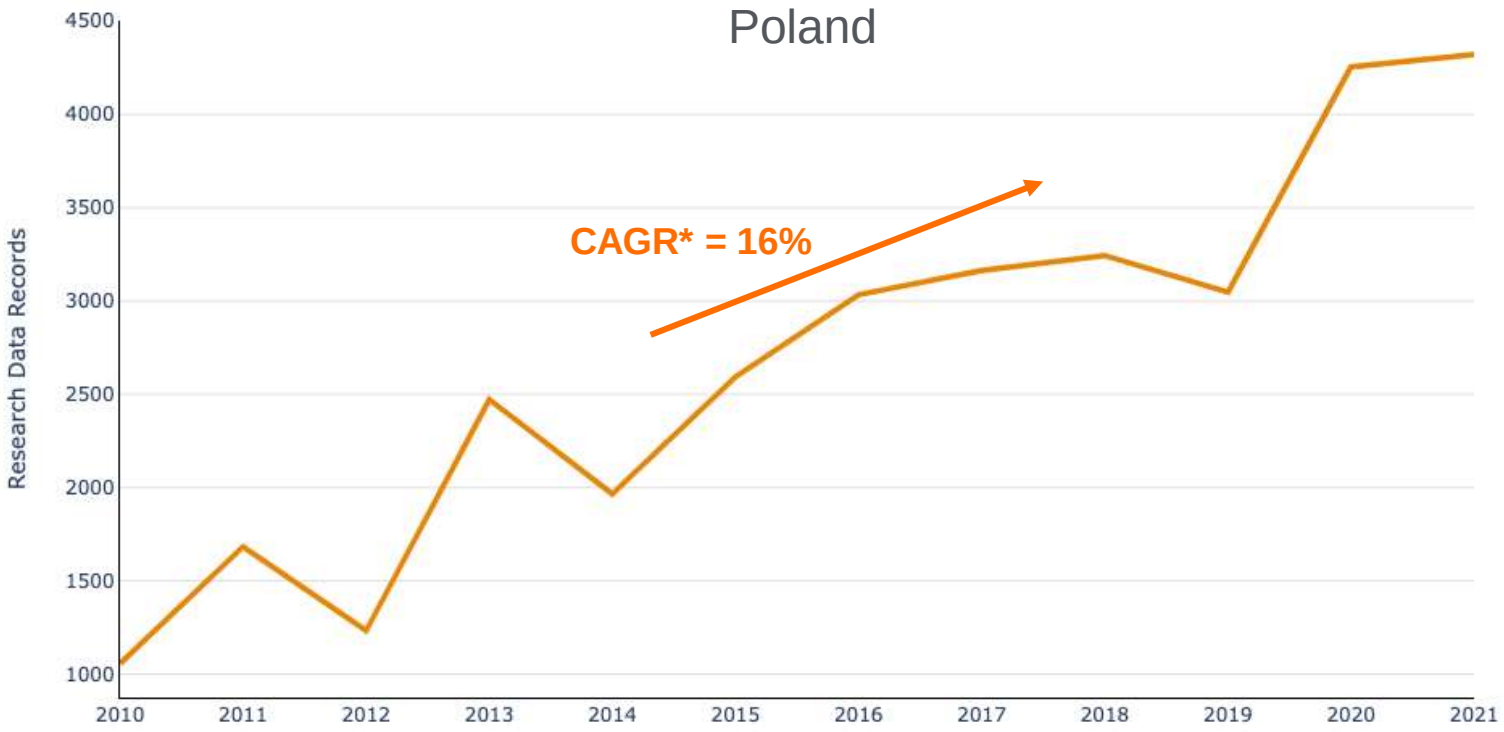
Poland

1.370.000+

EU

Source: **Data Monitor** Corpus
https://www.elsevier.com/solutions/data-monitor

Research Data by year

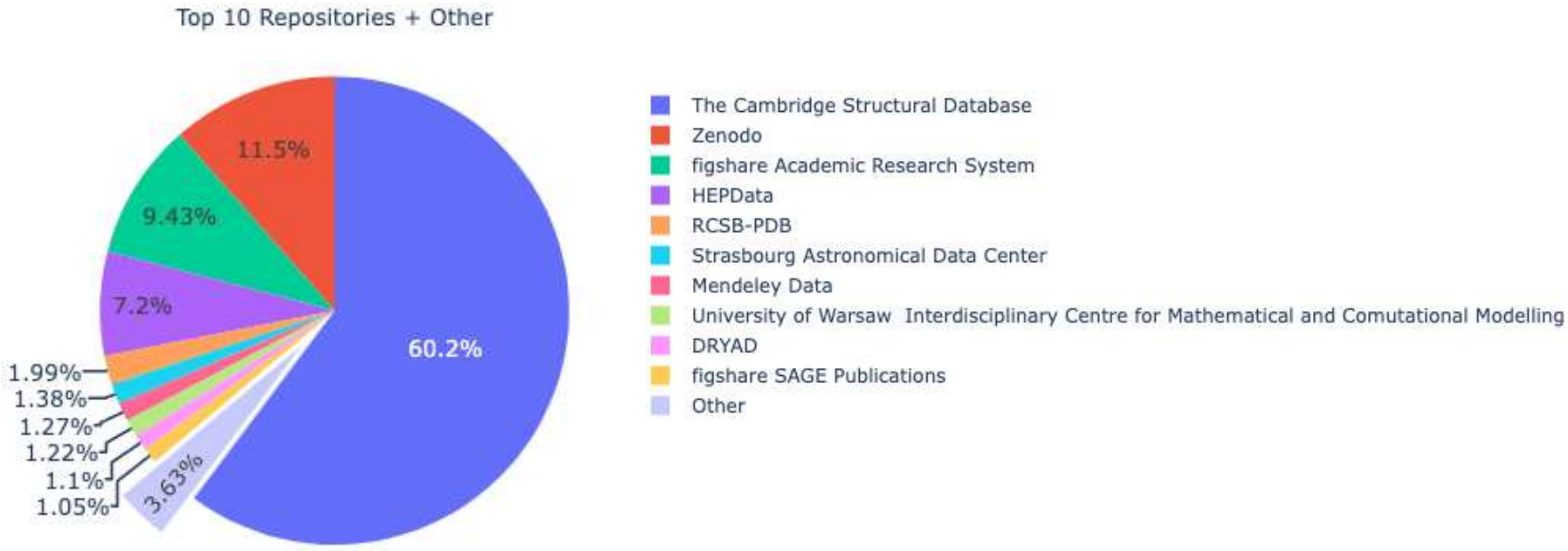


*CAGR (compound annual growth rate) calculated in the period 2012 – 2021

Source: **Data Monitor** Corpus
https://www.elsevier.com/solutions/data-monitor

Research Data by repository

Poland

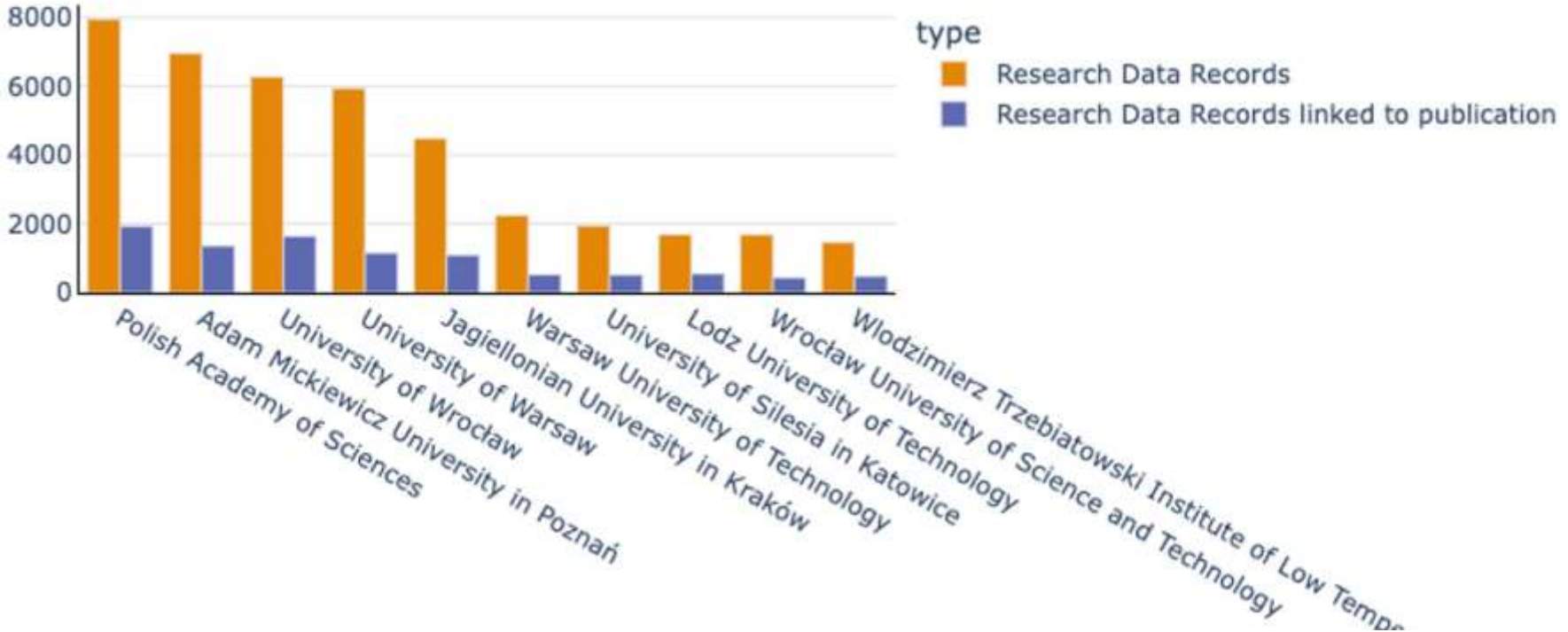


Source: **Data Monitor** Corpus
https://www.elsevier.com/solutions/data-monitor

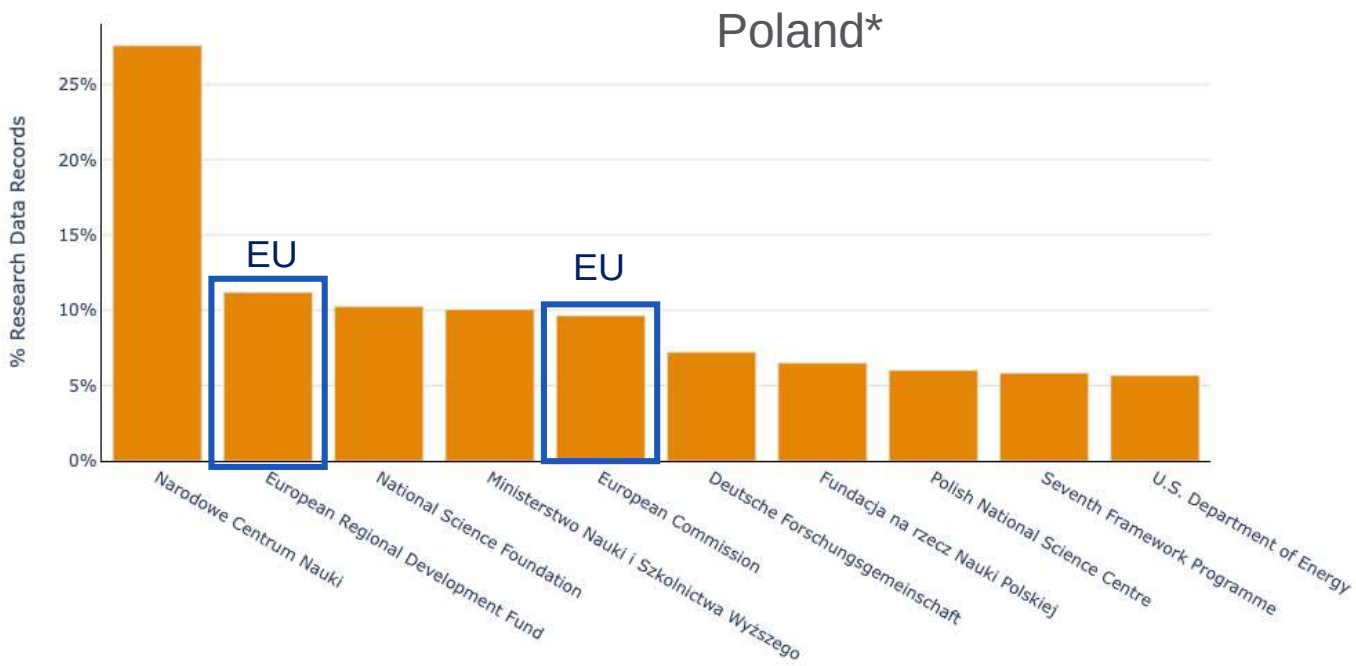
Research Data Top 10 Institutions

Poland

Source: **Data Monitor** Corpus
https://www.elsevier.com/solutions/data-monitor



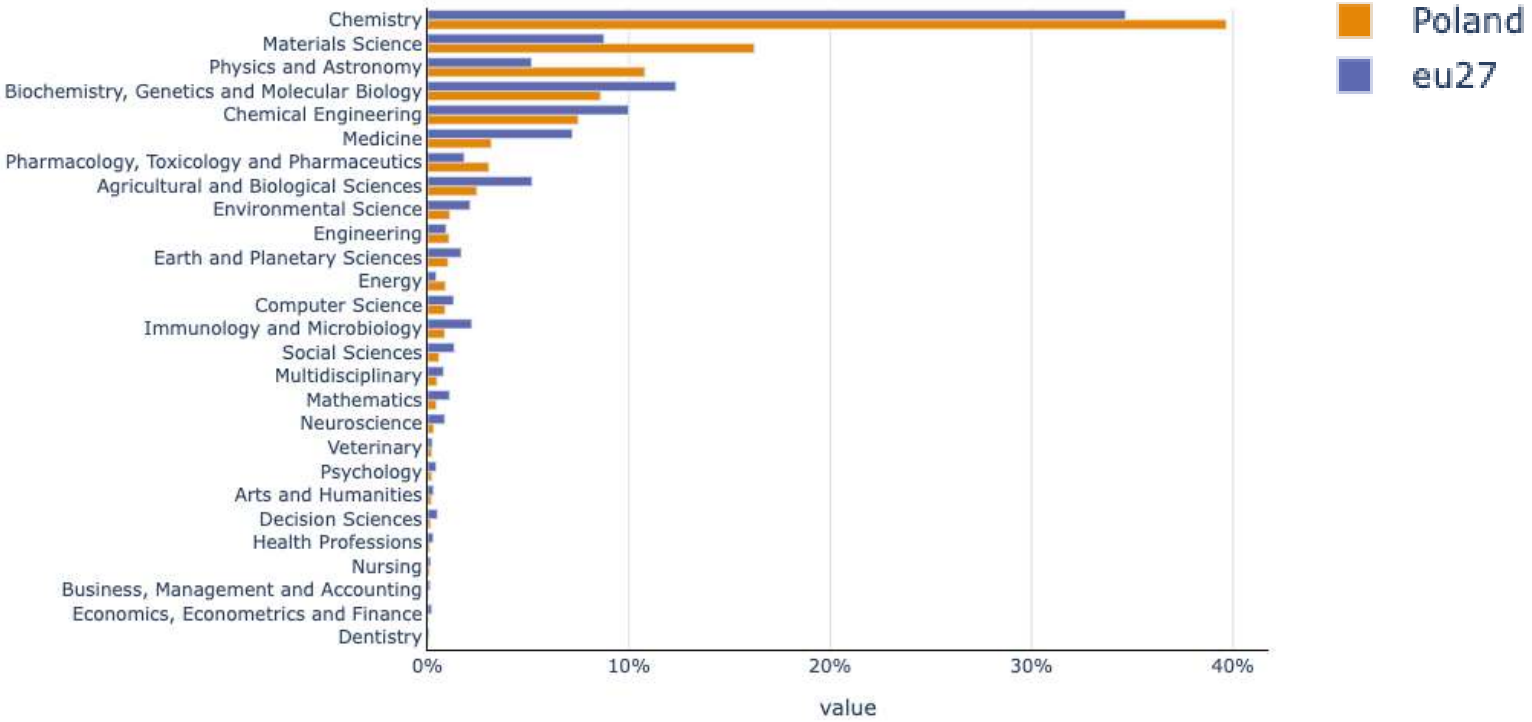
Research Data Top 10 Funders



***focuses on the subset of research data linked to publications**

Source: **Data Monitor** Corpus
https://www.elsevier.com/solutions/data-monitor

Research Data ASJC27 Subject Breakdown



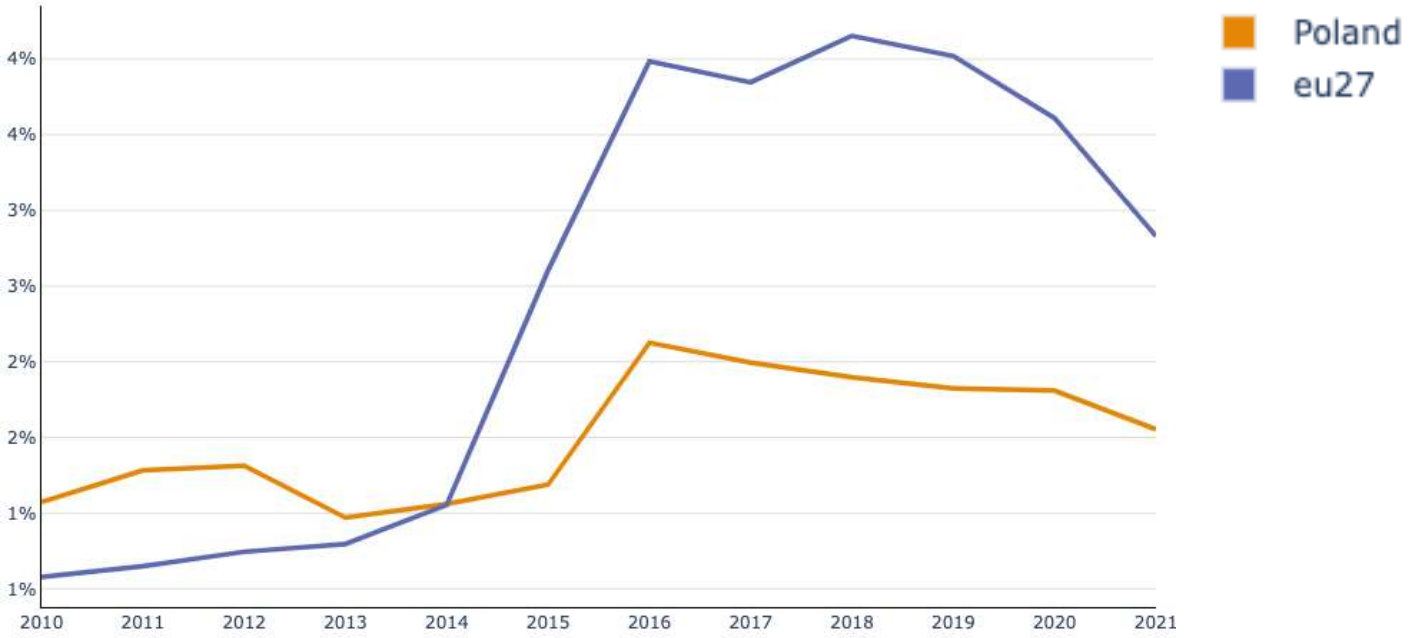
***focuses on the subset of research data linked to publications**

Source: **Data Monitor** Corpus
https://www.elsevier.com/solutions/data-monitor

Publications with Research Data / All Publications (trend)

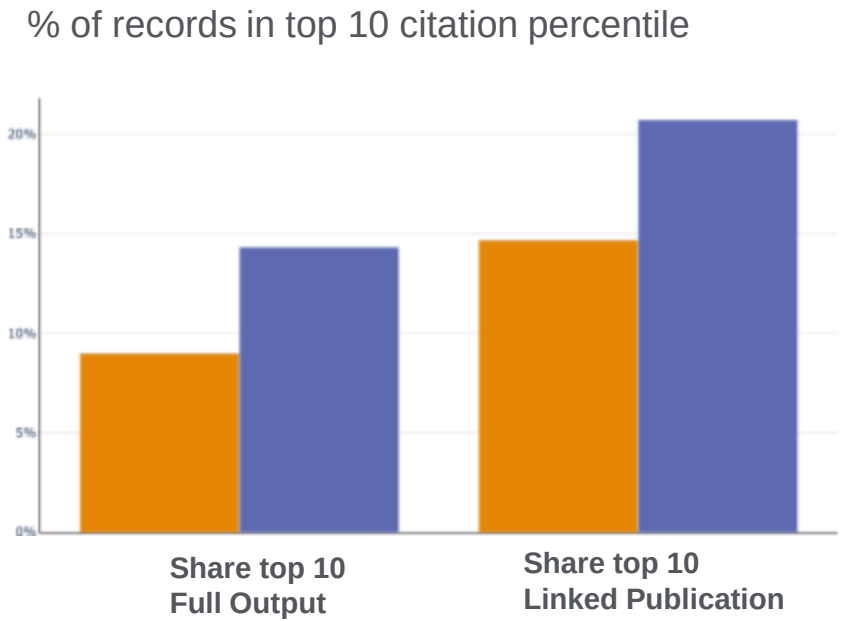
Source: **Data Monitor** Corpus
https://www.elsevier.com/solutions/data-monitor

OA Publications with Research Data / All OA Publications (trend)



Source: **Data Monitor** Corpus
https://www.elsevier.com/solutions/data-monitor

Academic impact of publications linked to data



Source: **Data Monitor** Corpus
https://www.elsevier.com/solutions/data-monitor

Thank you !

Dimas Putra
Senior Product Manager - Elsevier
d.putra@elsevier.com