(meta)data, distortion, and dashboards

Emma Wilson
Edinburgh Research Office
Edinburgh Open Research Conference
18 June 2026

Background

Research metadata can help us navigate research outputs:

- Aids discoverability in online systems
- Tells us something about the work
- Is data in its own right

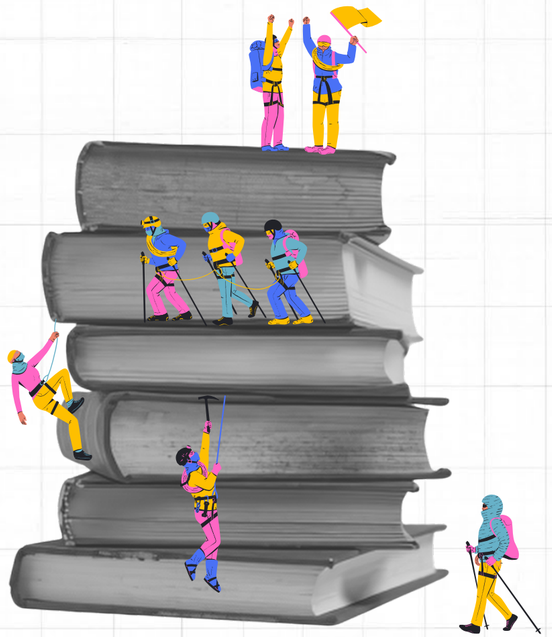
Growing interest in using research (meta)data to **build dashboards** for:

- Research assessment (funder and institutional monitoring)
- Open research monitoring
- Large-scale research summaries as part of meta-research / research synthesis activities

Facilitated by **large, open catalogues** of research information, e.g., OpenAlex.

However, easily accessible (CCBY or CC0) data does not mean easy to analyse and interpret or straightforward use.

Focusing on my PhD experiences, hoping to contribute to a dialogue around **inclusive and ethical use of scholarly metadata** that can support research activity.



(meta)data?

Type of data	What it might include	Examples of where data might be used
Bibliographic	Titles, abstracts, references, DOI, author and affiliation, publication venue	Using DOIs to retrieve further data about the output Examining rates of open access based on publication venue
Research	Metadata on software, research data, samples, instruments	Examining how often data and code are shared Using RRIDs to track the use of samples and instruments
Funding	Funder and grant	Tracking who is funding what
Contributor	Organisations and contributors	Identifying outputs from organisations using ROR ID Identifying authors using ORCID ID

Barcelona Declaration on Open Research Information, 2024

Data accuracy

Missing data

Data transparency

Missing data

No database is fully comprehensive in terms of what gets to be included, and **missing data occurs unevenly** at various levels:

1. **Focus on written text documents** (e.g. journal articles and books) excludes non-traditional outputs (particularly AHSS) and indigenous knowledge.

The majority of works in OpenAlex are classified with the document type “article” (Alperin et al., 2024).

2. **Focus on English language** and larger (often commercial) publishers means excluding outputs from the Global Souths and smaller publishers.

Journals from low income countries and those published in languages other than English are still underrepresented (Chavarro et al., 2025).

3. **Database (meta)data is mainly provided by publishers**, and this is patchy.

For example, a lack of standardisation, disconnected systems, and cost and capacity issues have limited the quality and quantity of funding metadata (Barcelona Declaration et al., 2026).

Data accuracy

Where data are missing, **metadata enrichment** efforts aim to fill in gaps. This often involves rule -based (e.g., ifelse , pattern matching) or generative AI (e.g., LLM) methods.

Many of these methods will look at performance metrics to assess how well these methods might be working, based on True / False Positives & Negatives.

- **Deployed at scale** (think millions of research outputs) even seemingly well performing methods can result in many data inaccuracies (Wilson, 2025 [Thesis]).
- **Generative AI methods are problematic** (non-transparent, encode bias, stochastic text prediction) (Guest et al., 2025a, Guest et al., 2025b).

Open access type

OpenAlex uses rules to determine open access type:

"*For each work, OpenAlex decides whether the work meets this definition using several techniques, such as looking for it in the Directory of Open Access Journals, or trying to find a publicly available PDF on the internet. If a work is Open Access, OpenAlex is able to provide you with a link to the full text of the work.*" (Open Acess Support Page, OpenAlex)

"***Diamond OA*** *: Published in a fully OA journal —one that is indexed by the DOAJ or that we have determined to be OA —with no article processing charges (i.e., free for both readers and authors).* "

Not all diamond OA journals are listed in DOAJ and unclear (potentially ad hoc) process for what "we have determined" means. What about diamond OA books?

Could underestimate the quantity of diamond open access outputs.

ORCID IDs

ORCID is a free, unique persistent identifer that can be used by researchers to highlight their contributions. It's especially useful where researchers have a common name or change their name during their career, or there are culturual differences in name order.

ORCID launched in 2012 and even now not every output contains an ORCID ID for all authors, so there is a lot of missing data.

"As part of [a data science] course, I developed a simple programme where a user could input an ORCID number[...] the programme would search OpenAlex for that author and output a basic, structured biography detailing the author's name, affiliation, and key research interests.

"Through testing this programme using my own ORCID ID, I discovered OpenAlex had misattributed publications by other researchers sharing my name to my ORCID ID. This was not a problem with ORCID but rather a problem with OpenAlex attempting to fill in missing data. " (Wilson, 2025 [Thesis])

ORCID ID becomes unusable because we cannot be sure the ORCID was used in the output or was "guessed" by the database provider.

Transparency

The source of data (i.e. publisher -provided or AI-enriched) is not always apparent to users.

In the case of OpenAlex:

The information on how open access status is determined was found on a help page through internet searching and not in the API documentation.

When downloading data, there is no way to differentiate between ORCID IDs that have been provided by publication venues versus added using metadata enrichment.

Information on the provenance of data should be accessible and transparently documented. Is a lot of data really useful if we can't trust it?

Possible solutions

Invest in open

Initiatives like Barcelona Declaration on Open Research Information

Open infrastructures for publishing and disseminating data about research

Accuracy & transparency

Make sure documentation is easy to find and comprehensive

Consider the impact of metadata enrichment beyond % performance values

Take care with data

Ask questions about who is missing and who is represented in the data

Consider data ownership and biases

Use quant metrics to supplement not replace qual analysis

Thank You

If you're interested in continuing this conversation,
please get in touch!

Emma Wilson
Email: emma.wilson@ed.ac.uk
BlueSky: emmawilson.eurosky.social

