September 27, 2023

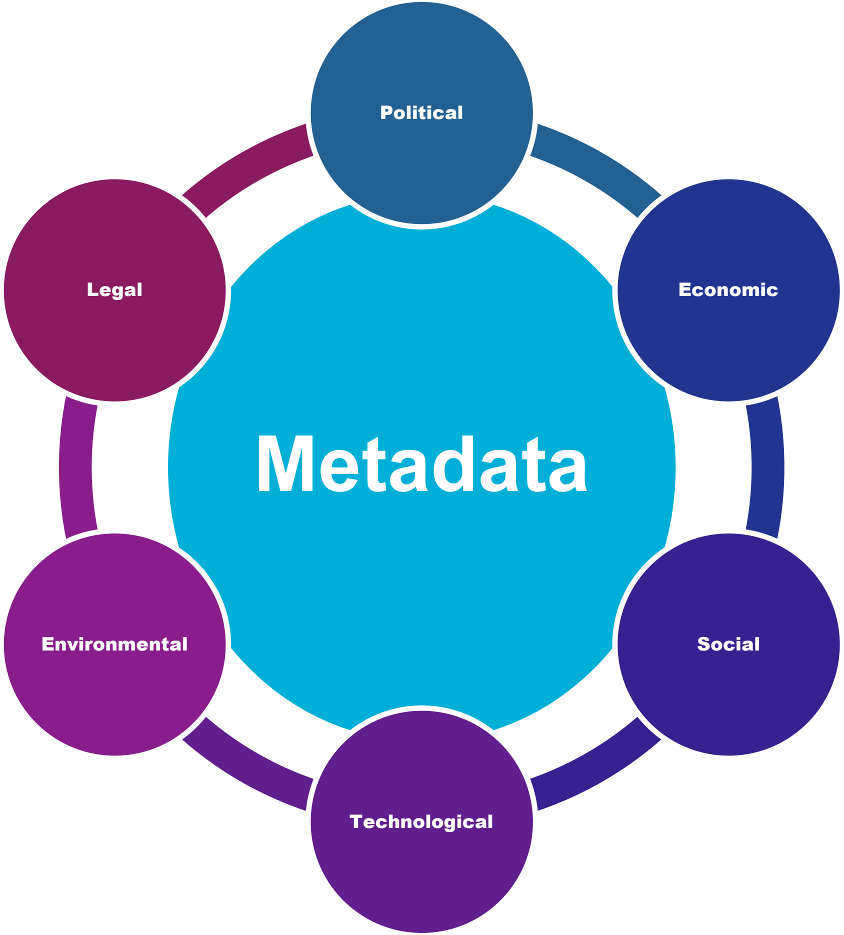
Transforming Metadata: Valuing metadata in changing environments

OCLC

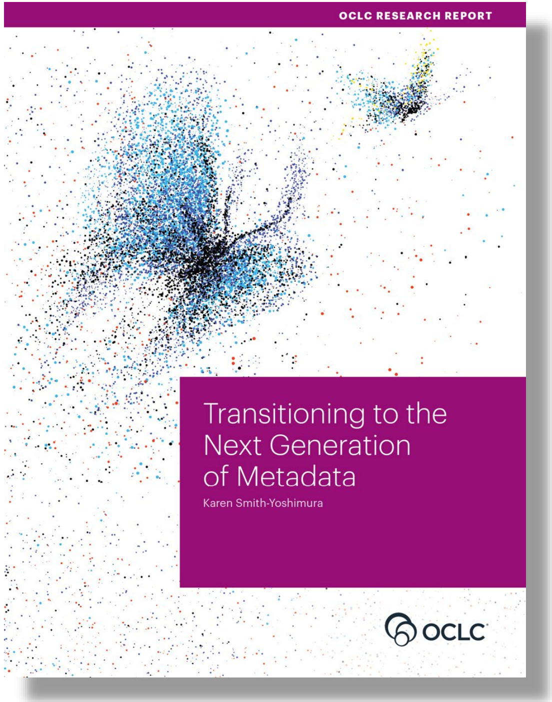
We acknowledge and celebrate the Indigenous peoples on whose traditional lands and airways we meet, and pay our respect to the Elders past and present.

If you are unsure whose land you are currently residing upon, we encourage you to visit native-land.ca

Transforming



Next-generation metadata



OCLC RESEARCH REPORT

Transitioning to the Next Generation of Metadata

Karen Smith-Yoshimura

oc.lc/nextgen-metadata-report

Transition to linked-data and identifiers

- Expanded reliance on PIDs for entity management.

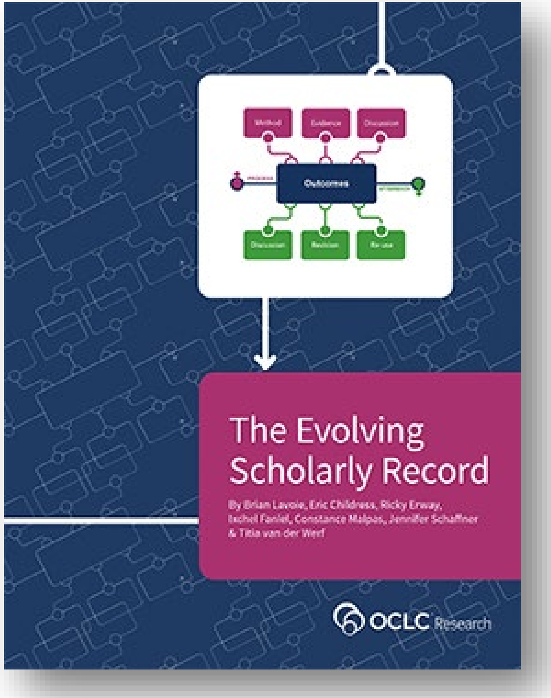
Describing inside-out and facilitated collections

- Increased expectations for research information management.

Future staffing requirements

- Shifting learning towards new tools and skills.

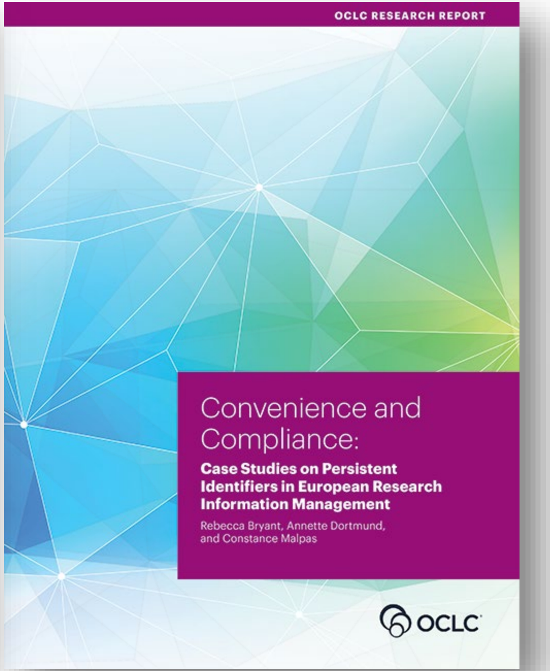
Transforming social & technical needs



The Evolving Scholarly Record

By Brian Lavoie, Eric Childress, Ricky Erway, Ixchel Faniel, Constance Malpas, Jennifer Schaffner & Titia van der Werf

OCLC Research



OCLC RESEARCH REPORT

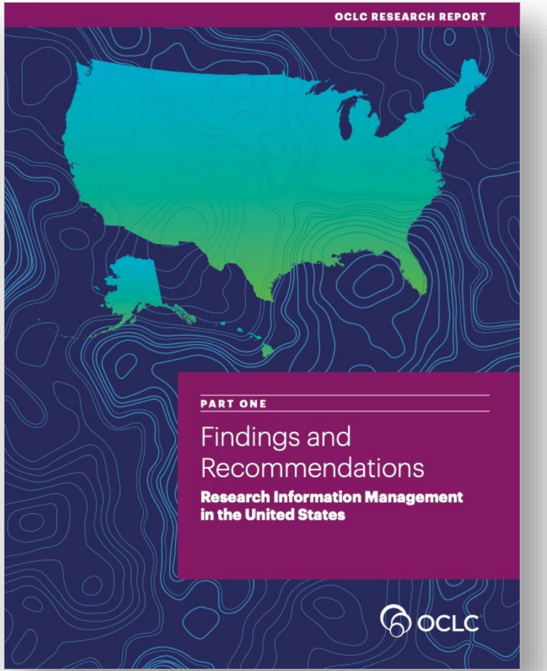
Convenience and Compliance:

Case Studies on Persistent Identifiers in European Research Information Management

Rebecca Bryant, Annette Dortmund, and Constance Malpas

OCLC

oc.lc/rim-europe



OCLC RESEARCH REPORT

PART ONE

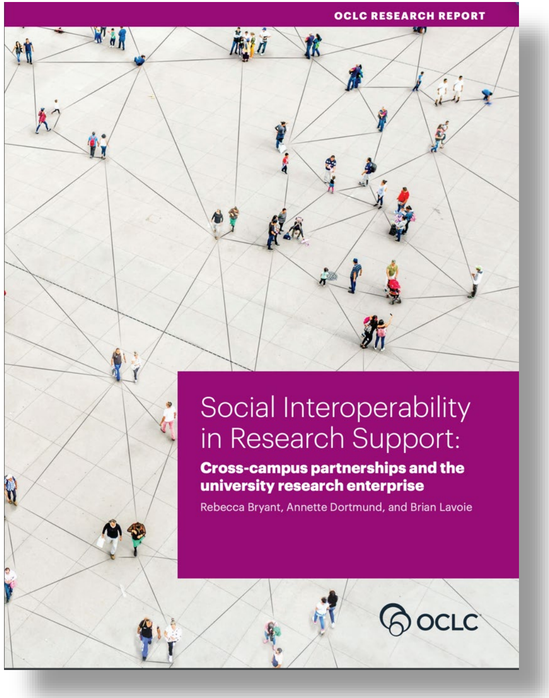
Findings and Recommendations

Research Information Management in the United States

OCLC

oc.lc/us-rim-report

Social interoperability & collaboration



OCLC RESEARCH REPORT

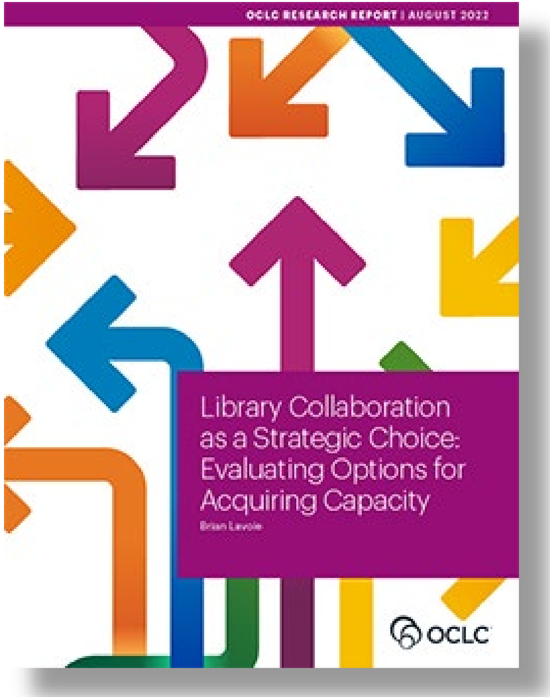
Social Interoperability in Research Support:

Cross-campus partnerships and the university research enterprise

Rebecca Bryant, Annette Dortmund, and Brian Lavoie

OCLC

oc.lc/social-interoperability



OCLC RESEARCH REPORT | AUGUST 2022

Library Collaboration as a Strategic Choice: Evaluating Options for Acquiring Capacity

Brian Lavoie

OCLC

oc.lc/strategic-collaboration-report

Transforming the workforce

- **Metadata Managers Focus Group participants face continuing challenges to hire for *attitude* and *aptitude* in changing environments.**

Library Futures series

Open everything, everywhere, all at once

OCLC’s role in the Open Access ecosystem

Sustainable stewardship: WorldCat as a membership good

https://hangingtogether.org/tag/OPEN/

Ioana Hulbert

Researcher, Ithaka S+R

Tracy Bergstrom

Program Manager, Collections and Infrastructure, Ithaka S+R

ITHAKA S+R

Transforming Metadata

September 27, 2023

ITHAKA S+R

We provide research and strategic guidance to help the academic and cultural communities serve the public good and navigate economic, technological, and demographic change.

Findings from the Library Deans and Directors Survey 2022

Ioana G. Hulbert, PhD | ioana.hulbert@ithaka.org
Researcher, Ithaka S+R

Thanks to the sponsors of the US Library Survey 2022

NON SOLUS

ELSEVIER

JSTOR

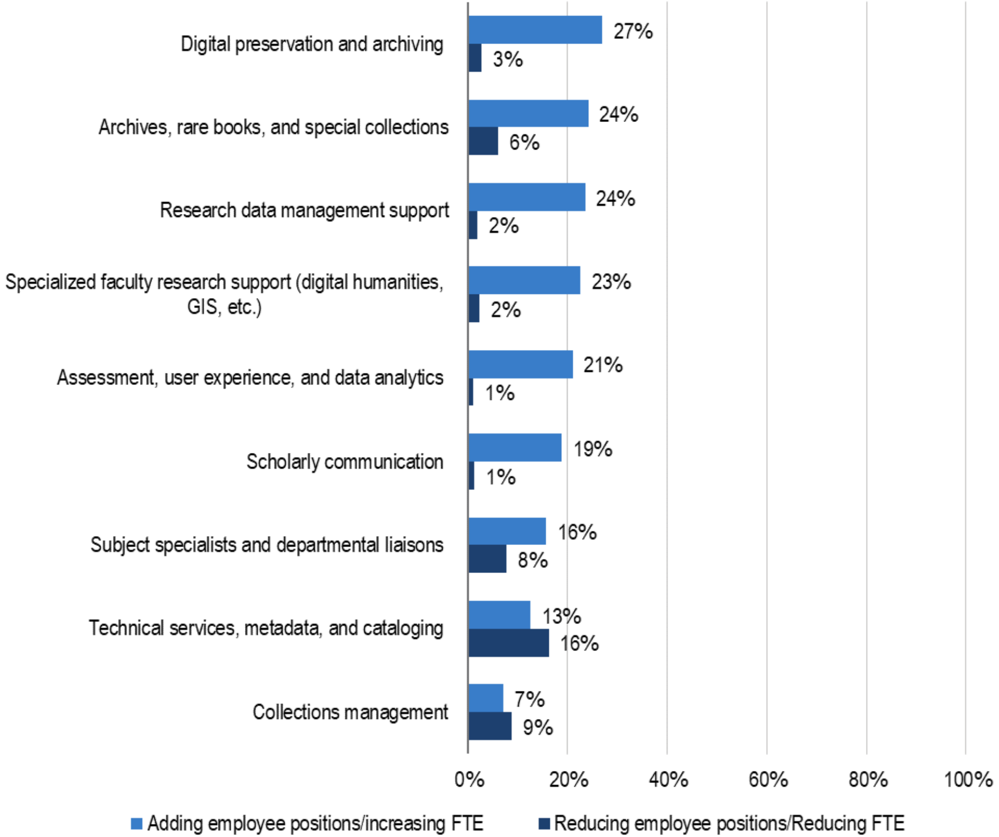
- 12% are struggling to **retain** talent skilled in cataloging & metadata.
- 17% are struggling to **recruit** talent skilled in cataloging & metadata.

Read the full report here:

Please indicate whether your library is considering outsourcing these skills to a third-party provider/another department (top five).

Scenario	Baccalaureate	Master’s	Doctoral
Considering Outsourcing			
Cataloging and Metadata	13%	17%	20%
Technology and programming skills	8%	6%	16%
Acquisitions, collections management and procurement	4%	4%	9%
User data analysis and assessment	6%	5%	5%
Data management and research	1%	2%	5%
Currently Outsourcing			
Facilities management	14%	13%	11%
Technology and programming skills	16%	10%	10%
Cataloging and metadata	11%	8%	10%
Marketing and fundraising	13%	8%	5%
Data management services	2%	3%	3%

To the best of your knowledge, will your library add or reduce employee positions in any of the following areas over the next five years?



ITHAKA S+R

In the next five years, do you anticipate the share of overall resource expenditure (including direct expenditures and staffing) to increase, remain the same, or decrease for each of the following?

Budget Scenario	Baccalaureate	Master's	Doctoral
General Collections			
Increase	24%	13%	23%
Remain the same	45%	52%	50%
Decrease	32%	35%	27%
Rare, special, and other distinctive collections			
Increase	22%	15%	37%
Remain the same	62%	56%	53%
Decrease	16%	29%	10%
Services to support research data management			
Increase	39%	28%	37%
Remain the same	53%	60%	56%
Decrease	8%	12%	7%

Findings from the research project “The Second Digital Transformation of Scholarly Publishing”

Tracy Bergstrom | tracy.bergstrom@ithaka.org
Program Manager, Collections and Infrastructure, Ithaka S+R

Project Scope and Starting Points

- The first digital transformation in scholarly communication saw a shift of content from paper to digital → the current second digital transformation is focused on the transformation of infrastructure.
- Within this context, we defined shared infrastructure as broadly as possible to include a variety of shared tools and frameworks that underpin scholarly communication processes.
- A robust and nimble infrastructure is imperative to support the vital work of scholarly communication to meet the emerging service needs of different stakeholders.
- This is complicated by the dynamics— including divergent communities, priorities and financing— that differentiate STEM, humanities, and social sciences publishing.

RESEARCH REPORT

April 24, 2023

Common Scholarly Communication Infrastructure Landscape Review

Oya Y. Rieger
Roger C. Schonfeld

ITHAKA S+R

Landscape Review

- In April 2023, the research team published a Scholarly Communications Infrastructure Landscape Review.

- The Landscape Review is intended to provide illustrations of representative elements by category of the shared infrastructure.

Landscape Review: Confirmation of Metadata Principles

- Metadata continues to play a crucial role in enabling discovery and access as well as facilitating information encoding and exchange through interoperability.
- Semantic technologies such as natural language processing, data mining, artificial intelligence (AI), category tagging, and semantic search are being increasingly used to improve metadata and lead to better discovery and access.
- Metadata quality is key to the functionality of persistent identifiers (PIDs), which provide unique and long-lasting reference to digital objects, contributors, and organizations to facilitate discovery, access, linking, and assessment of scholarly content.
- In addition to descriptive metadata, administrative and preservation metadata includes technical information to support long-term management and preservation of digital collections.

White Paper: Project Overview

- In spring and early summer 2023, we conducted interviews with 49 infrastructure service providers, publishers, librarians, advocates, analysts, funders, and policy makers.

- Metadata was one of the topics we inquired about, but it also underpins many of the infrastructure components including:
    - Discovery, Syndication, Persistent Identifiers, Preservation, Publishing Platforms and Repositories, Researcher Identity, Research Data Curation

- Draft report was released for comment in July 2023; we're currently revising the report based on this feedback. The project's final report will be published in conjunction with the Frankfurt Book Fair in mid-October 2023.

- STM Solutions provided sponsorship support that made this project possible, for which we express our gratitude.

Discussion: Metadata as a Key Component within the Evolution of Scholarly Publishing

Preliminary Findings: Syndication and Aggregation

- Shared infrastructure, and the interconnectedness of research content through a network of syndication and aggregation, has tremendous benefits for researchers. It also introduces complexity.

- If error exists in descriptive metadata, for instance, it is difficult to rectify once the error has proliferated across systems.

- It is also unclear whose responsibility it is to enhance or correct metadata once it is in various syndication pipelines, or where might be the point to do this.

- This issue has both practical implications, but also larger implications for perception of the trustworthiness of content.

- Research integrity was one of the foremost issues of concern voiced by interviewees.

Preliminary Findings: Atomization of the Scholarly Record

- As the scholarly record becomes more heterogeneous, variable, dynamic, and distributed, keeping track of and archiving various versions and components (article, preprint, underlying data, external components, etc.) will get even more complex.
- Ideally, all these individual components will be connected together to represent the complete set of outputs of a given research project.
- As the scholarly record becomes more fragmented, the strength of the metadata connecting these components will be all the more critical.
- The atomization of the scholarly record in these component parts raises questions about assessing the quality, integrity, and impact of research at various levels.

Preliminary Findings: The Critical Role of Persistent Identifiers

- Object identifiers encompass a broad range of resources including books, articles, white papers, chapters, datasets, tables, figures, and videos. A single resource may have multiple identifiers associated with its different components.
- If properly structured, PIDs have the potential to reduce administrative burden and redundancy by eliminating the creation of redundant metadata.
- This makes it all the more important that the PID metadata is trustworthy, defined as that it can be trusted and understood as valid by a variety of partners
- PIDs also play an important role in enabling AI systems to access, integrate, and analyze data from multiple sources.

Thank you!

Alice Meadows

Cofounder, MoreBrains Cooperative

Incentives to invest in persistent identifiers

Two cost-benefit analyses of PIDs in national research systems

Alice Meadows, Cofounder, MoreBrains Cooperative

MOREBRAINS

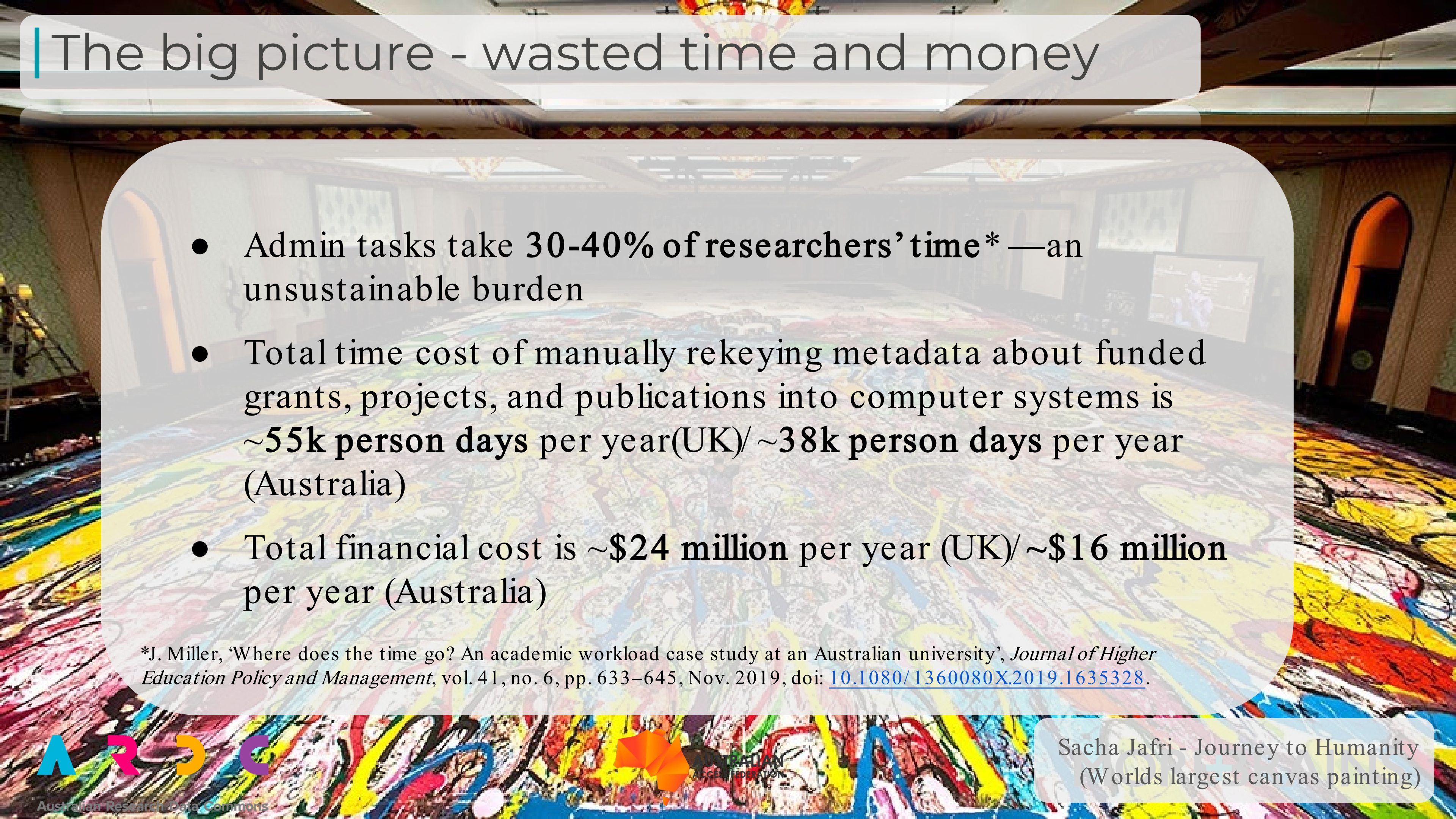
Picture courtesy of Josh Brown
All other images Wikipedia
unless otherwise indicated

@alicejmeadows | @alicemeadows@scicomm.xyz | @alicemeadows.bsky.social

The big picture - wasted time and money

- Admin tasks take **30-40% of researchers' time*** —an unsustainable burden
- Total time cost of manually rekeying metadata about funded grants, projects, and publications into computer systems is ~**55k person days** per year(UK)/ ~**38k person days** per year (Australia)
- Total financial cost is ~**$24 million** per year (UK)/ ~**$16 million** per year (Australia)

*J. Miller, 'Where does the time go? An academic workload case study at an Australian university', *Journal of Higher Education Policy and Management*, vol. 41, no. 6, pp. 633–645, Nov. 2019, doi: 10.1080/1360080X.2019.1635328.



Sacha Jafri - Journey to Humanity
(Worlds largest canvas painting)

Australian Research Data Commons

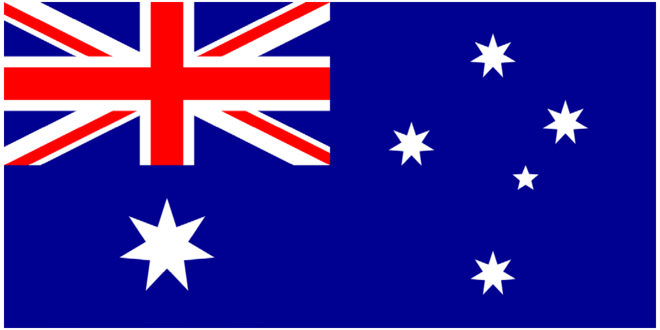
Calls to action

"Jisc to lead on selecting and promoting a range of unique identifiers, including ORCID, in collaboration with sector leaders with relevant partner organisations."

Prof Adam Tickell, Open access to research: independent advice

"I note the higher education sector's concern regarding the workload required for the current mode of delivery of the ERA assessment … Streamlining the processes undertaken during National Competitive Grant Program funding rounds must be a high priority for the ARC …I ask that the ARC identify ways to minimise administrative burden on researchers."

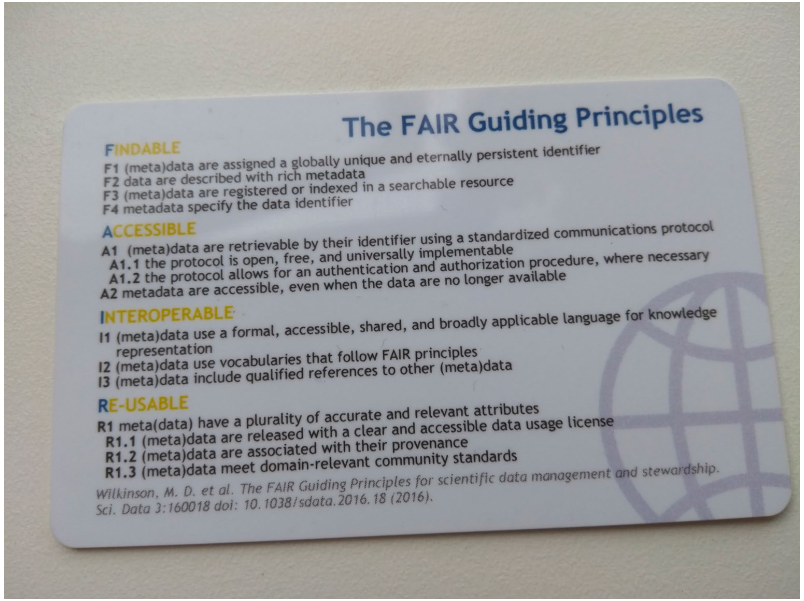
Hon Jason Clare MP, Minister for Education, Statement of Expectations 2022



How can PIDs help?

Persistent identifiers…

- Are unique
- Don't change
- Are connected to each other and to other metadata
- Can be read by computers/enable interoperability
- Facilitate tracking over time
- Save effort and time when finding and compiling information
- Support openness and transparency



The FAIR Guiding Principles

FINDABLE

F1 (meta)data are assigned a globally unique and eternally persistent identifier
F2 data are described with rich metadata
F3 (meta)data are registered or indexed in a searchable resource
F4 metadata specify the data identifier

ACCESSIBLE

A1 (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2 metadata are accessible, even when the data are no longer available

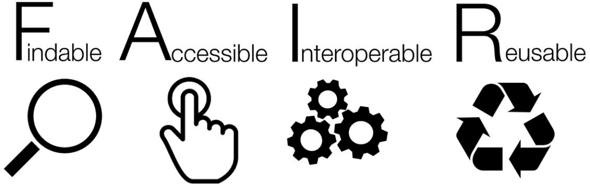
INTEROPERABLE

I1 (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation
I2 (meta)data use vocabularies that follow FAIR principles
I3 (meta)data include qualified references to other (meta)data

RE-USABLE

R1 meta(data) have a plurality of accurate and relevant attributes
R1.1 (meta)data are released with a clear and accessible data usage license
R1.2 (meta)data are associated with their provenance
R1.3 (meta)data meet domain-relevant community standards

Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3:160018 doi: 10.1038/sdata.2016.18 (2016).



MORE+BRAINS

Five priority (open) PIDs

Grants Crossref (and soon DataCite) DOIs for grants allow unique identification and certainty of referencing

People ORCIDs represent a verifiable, reusable record of an individual's education, employment, funding, and outputs

Outputs DOIs (eg, Crossref and DataCite) are unique, permanent identifiers of publications and other outputs

Projects Research Activity Identifier (RAiD) is a new, portable container of research project activities connecting everyone, everywhere, everything involved

Organizations Unique ID for organizations to support accurate discovery of their activities, outputs, and impacts

PID-optimized research lifecycle

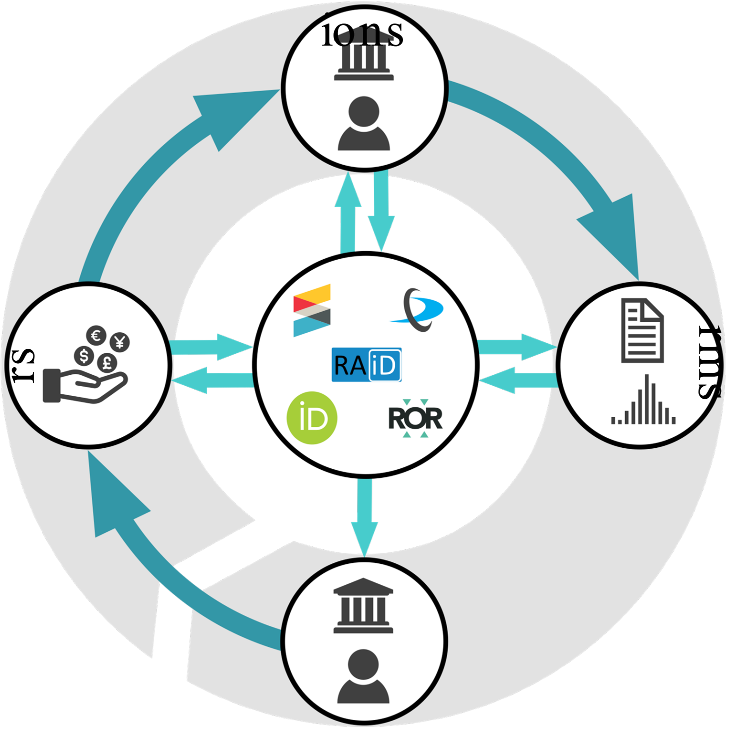
There are three major areas of benefit:

1. **Moving data between systems**
2. Efficiencies through automation
3. Better strategic decision-making

https://resources.morebrains.coop/pidcycle/

Brown, Josh, Jones, Phill, Meadows, Alice, & Murphy, Fiona. (2022, September 16). PID-optimised workflows: A vision of a more efficient future. Zenodo. https://doi.org/10.5281/zenodo.7085489

Researchers/ Institutions



Funders

Publishers/ Platforms

Researchers/ Institutions

Cost benefit analysis - our methodology (quantitative)

1. Identify and quantify entities in the research system for which PIDs are available (eg, FTE researchers, funding grants, etc)
2. Match these entities to the time taken for manual metadata re-keying of each one
3. Quantify the number of metadata re-key events

Multiplying these together generates the total time spent on manual metadata re-keying; multiplying that number by salary per minute provides the total cost.

Gathering the data (1)

Entities for which PIDs are available

- Number of researchers
  - Higher Education Research Data Collection (HERDC) - Australia
  - Higher Education Statistics Agency (HESA) - UK
- Average number of researchers per publication*
- Number of funded grants
  - Dimensions data and data from Australian Research Council (ARC) and Medical Research Future Fund (MRFF) - Australia; HESA - UK
- Number of publications
  - Dimensions data, Crossref data

* D. Fanelli and V. Larivière, 'Researchers' Individual Publication Rate Has Not Increased in a Century', PLoS ONE, vol. 11, no. 3, p. e0149504, Mar. 2016, doi: 10.1371/journal.pone.0149504.

Gathering the data (2 and 3)

Time taken for manual metadata rekeying and number of rekeying events

- Average time taken to rekey project / grant / publication information*
- Number of times information about grants and publications is entered into systems?
  - Australian repository managers' survey

Average salaries were based on official figures for average salaries for research administrators, junior and senior researchers

Note - some data aren't/aren't readily available…

* Research Consulting: 'Counting the Costs of Open Access', London Higher and SPARC Europe, and M. H. Klausen: 'Even Minor Integrations Can Deliver Great Value – A Case Study'

MORE+BRAINS

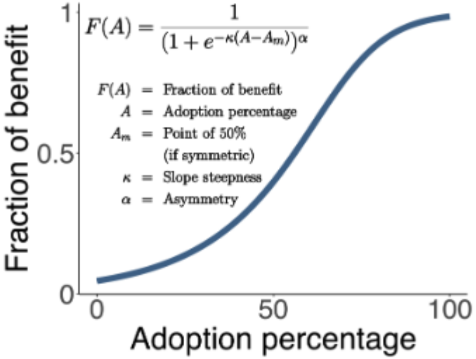
Cost benefit analysis - our methodology (qualitative)

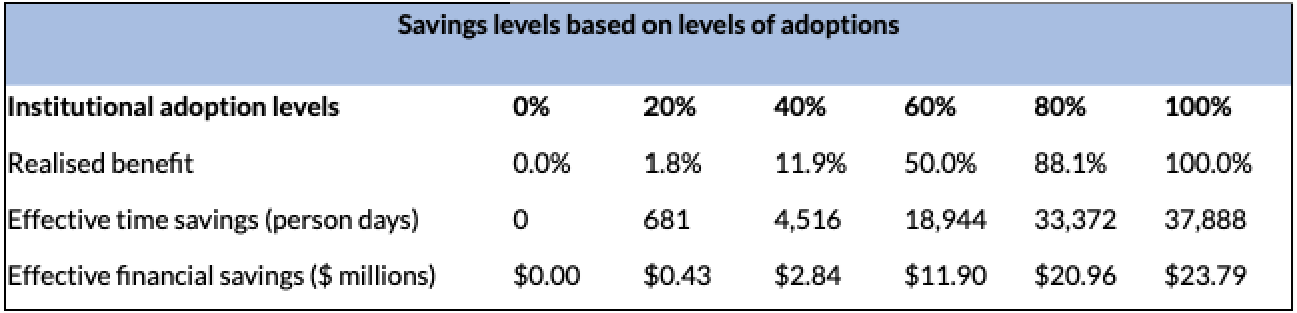
- UK - 11 semi-structured interviews with representatives from key stakeholder groups
- Australia - three case studies
  - ORCID integration at ARC*
  - The use of PIDs at the Terrestrial Ecosystem Research Network (TERN)
  - Central services provided by Australian Research Data Commons (ARDC) and the Australian Access Federation (AAF), such as the ORCID consortium, RAiD

*“*Before using ORCID, an average application used to take a few weeks; “formatting took time, getting the publications right took many days of work”. This took time away from the actual grant process.*”Joe Shapter, Pro-vice-Chancellor, University of Queensland

MORE+BRAINS

No one benefits till everyone benefits





Savings levels based on levels of adoptions

Institutional adoption levels	**0%**	**20%**	**40%**	**60%**	**80%**	**100%**
Realised benefit	0.0%	1.8%	11.9%	50.0%	88.1%	100.0%
Effective time savings (person days)	0	681	4,516	18,944	33,372	37,888
Effective financial savings ($ millions)	$0.00	$0.43	$2.84	$11.90	$20.96	$23.79

- The more organizations that adopt PID-enabled workflows, the more data is available, and the more everyone benefits, so…
- Develop a strategy to achieve a high level of adoption (80%+) for all five priority PIDs

The benefits of our cost benefit analysis

- Evidence of the value of PIDs and their metadata
  - For funders and policymakers
  - For institutions
  - For researchers
- Lays the foundation for developing a PID strategy
  - Benchmarking
  - Tracking progress
  - Demonstrating success
- Openly available methodology that can be replicated by others as needed

In case you want to learn more…

MORE+BRAINS

Incentives to Invest in Identifiers

A cost-benefit analysis of persistent identifiers in Australian research systems

AUSTRALIAN ACCESS FEDERATION

ARDC

Australian Research Data Commons

MORE+BRAINS

Revised cost-benefit analysis for the UK PID Support Network

18 November 2022

DOI: 10.5281/zenodo.7356219

Photo by Pixabay

cc BY

DOI: 10.5281/zenodo.7356219

DOI: 10.5281/zenodo.7100578

MORE+BRAINS

Thank you - any questions?

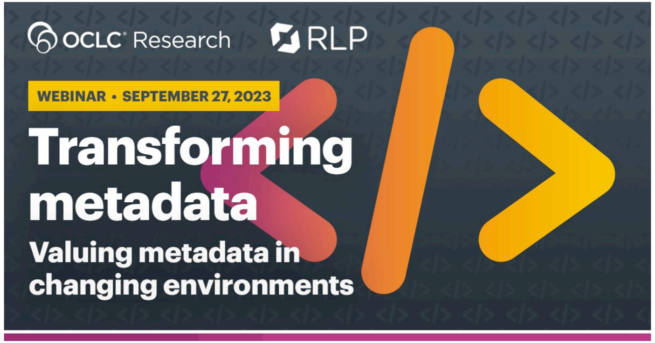
MORE+BRAINS

Thank you!

Richard J. Urban, Ph.D.

Senior Program Officer
Data science & next-generation metadata
OCLC Research Library Partnership

urbanr@oclc.org


OCLC Research
RLP
WEBINAR • SEPTEMBER 27, 2023
Transforming metadata
Valuing metadata in changing environments

Find recordings and information about future Transforming metadata events at:

https://www.oclc.org/research/events.html