Authority of assertion in repository contributions to the PID graph

George Macgregor
University of Glasgow
2023-09-07
https://purl.org/g3om4c
https://orcid.org/0000-0002-8482-3973

(w/ Barbara S. Lancho-Barrantes & Diane Rasmussen McAdie)

Slides available: `https://doi.org/10.17868/strath.00085167`

Overview

This paper: to outline an emerging challenge in repository metadata and users' discovery of repository content

Exploring issues when levels of 'authority of assertion' are lacking or are uncertain

- Emergence of increasingly fluid, distributed, and fragmented scholarly objects
- Emergence of persistent identification as a (partial) solution
- The growing scholarly graph, e.g. 'PID graph'
- Implications for repositories and the question of 'authority of assertion'
- Scholars' 'PID literacy' as an impediment to satisfactory scholarly objective description (and ergo discovery) [1]

Raising a manifesto for action?


Interpreten-Katalog
Interpreten-Katalog
Ensembles

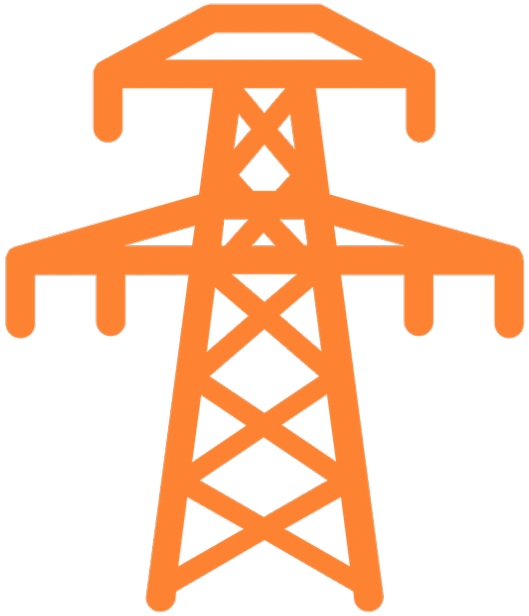
Metadata conventions...

- Generating statements of fact about (often) tangible things...
- Even in formal digital libraries, objects (often) static things with predictable descriptive metadata attributes (e.g. digitized objects, etc)
    - Linked Open Data (RDF/URIs)
- "...change is slow" [2]

😆

[Photo, previous slide ⤴️] Dresden-Neustadt. Sächsische Landesbibliothek - Staats- und Universitätsbibliothek Dresden, Europeana (CC-BY-SA)

Open scholarly infrastructure



[3]

Distribution, fragmentation, fluidity

Disorientation grows for metadata creators with research domain, but also for users in discovery for research content...

- Increasing levels of distribution, fragmentation, and fluidity
  - Emergence of open, distributed, multi object (often 'non-standard') transient scholarly objects
- Multi-object exists elsewhere (e.g. METS package) but all components under direct custodianship where *authority of assertion* exists
- PIDs increasingly central to linking, description, disambiguation, discovery

What are PIDs?

PID = *persistent identifier* 😄

- PIDs come in many flavours, e.g. DOI, ORCID, RAiD, ROR, etc. -- *and beyond!*

  `https://doi.org/10.17868/strath.00085975`

  `https://orcid.org/0000-0002-8482-3973`

  `https://ror.org/05j0ve876`
- Enables long-term identification but also 'actionability' because they are (*generally*) formed as URIs
- PIDs usually underpinned by some form of registry which registers new PIDs and resolves them (and contains *metadata*!) (e.g. DataCite, CrossRef)

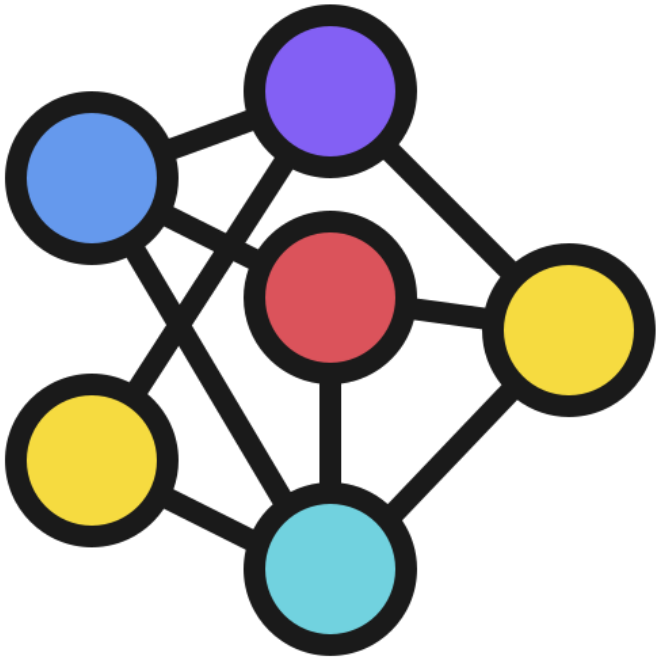
Why have PIDs?

PID = *persistent identification* 😉

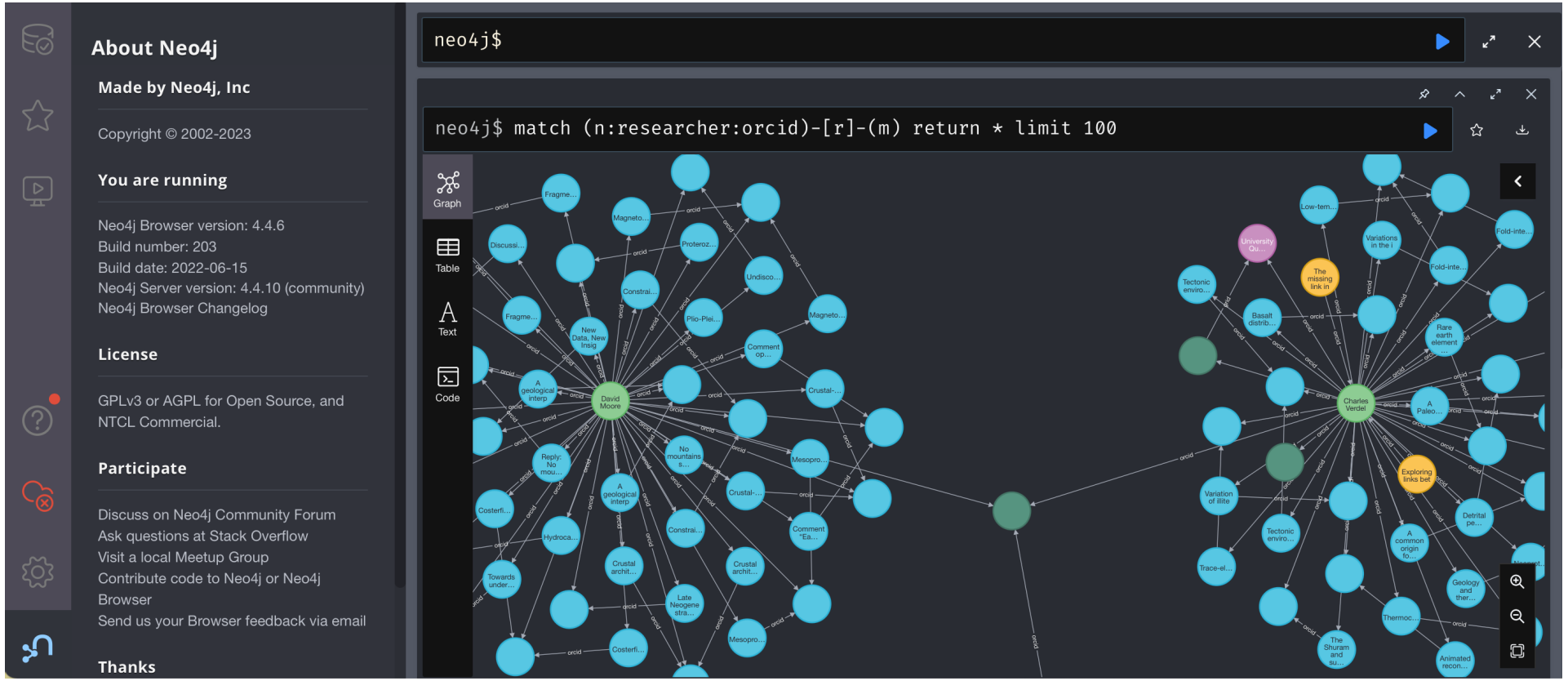
- Provides unique and persistent reference to an *entity* normally accessible over the Internet
- Maintenance of the scholarly record - citability, verification, reproducibility, replicability [5], [6]
  - '*Reference rot*' / '*link rot*' - e.g *Klein, Van de Sompel et al.* [6], [7]
- Uniquely identify entities on the web; enables discovery these entities -- happy times!
- *The quest to encode PIDs within metadata to link and grow the utility of data*

PID graph

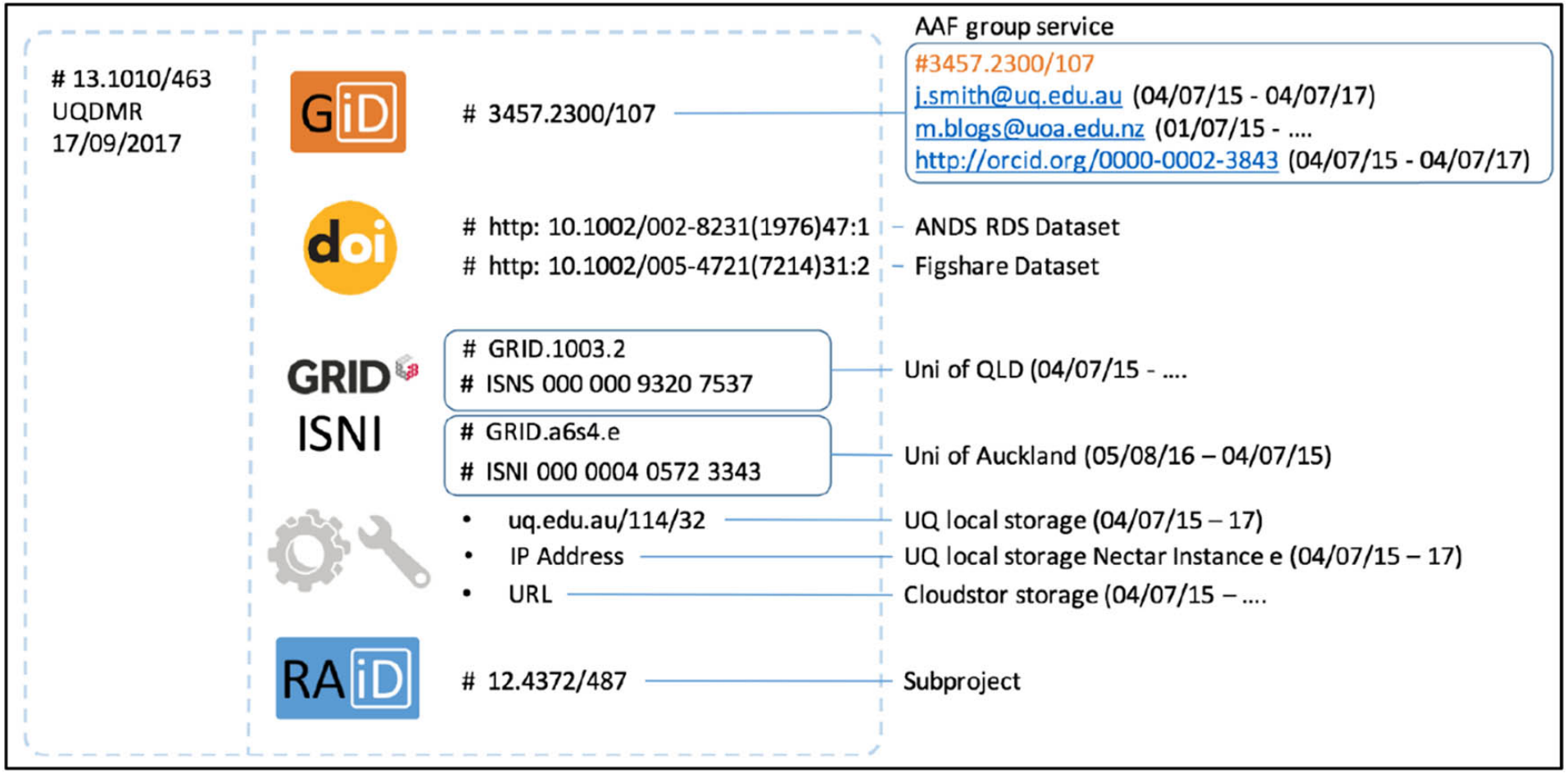
.... or Linked Open Data déjà vu... ? [4]



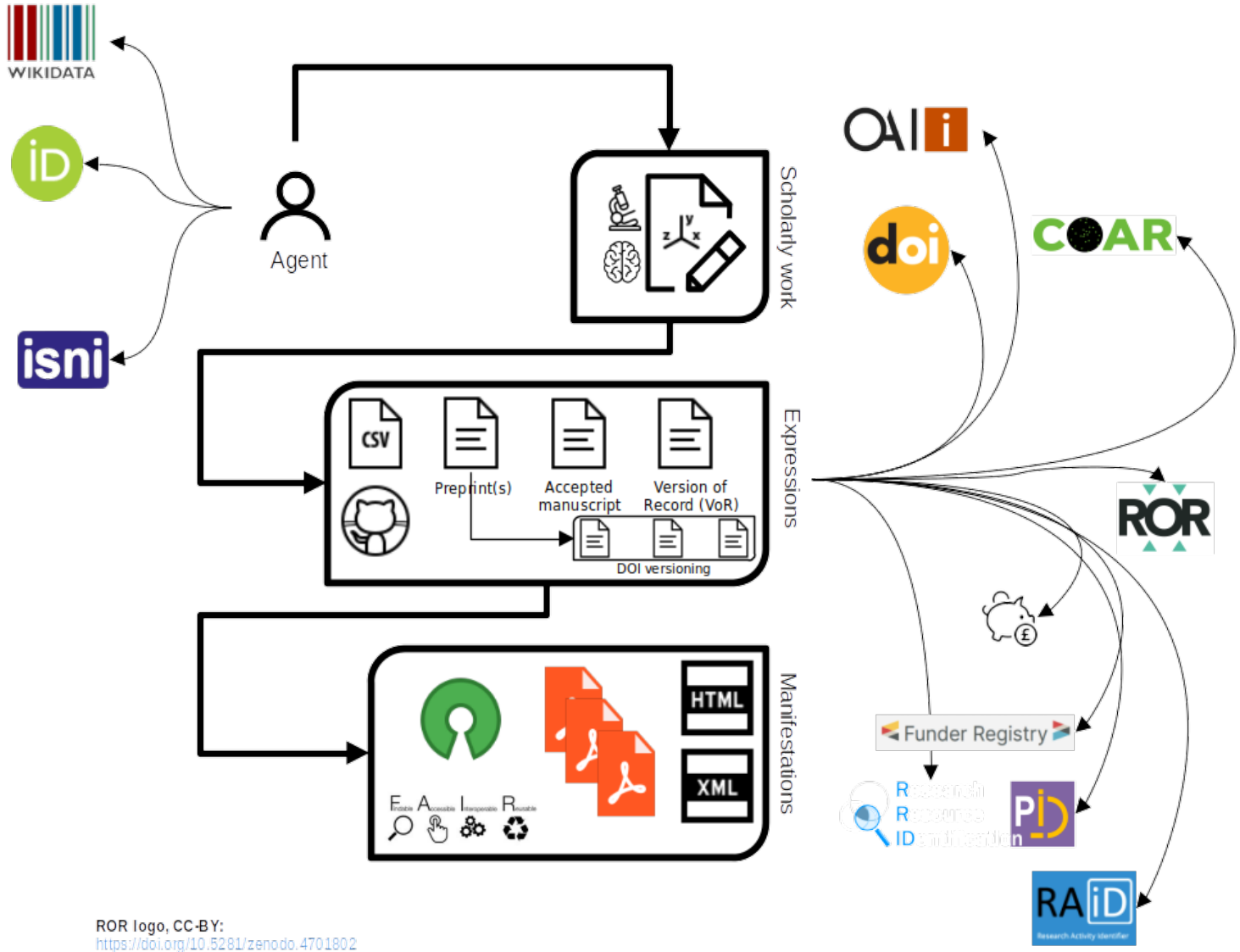
TIB – Leibniz Information Centre for Science and Technology - PID Service (CC-BY)



Exploring the graph with Neo4j...



RAiD envelope diagram. From Janke et al. [1], CC-BY.



ROR logo, CC-BY:
https://doi.org/10.5281/zenodo.4701802

Repositories on front line of PID data contributions

`distribution + fragmentation + fluidity`

Such metadata key to making scholarly works:

Interpretable...

- to contextualize; to provide provenance

Support reproducibility...

- to facilitate reproduciblity, verification, replication

Discoverable...

- to enable discovery through a variety of access points, some unconventional

What is 'authority of assertion'?

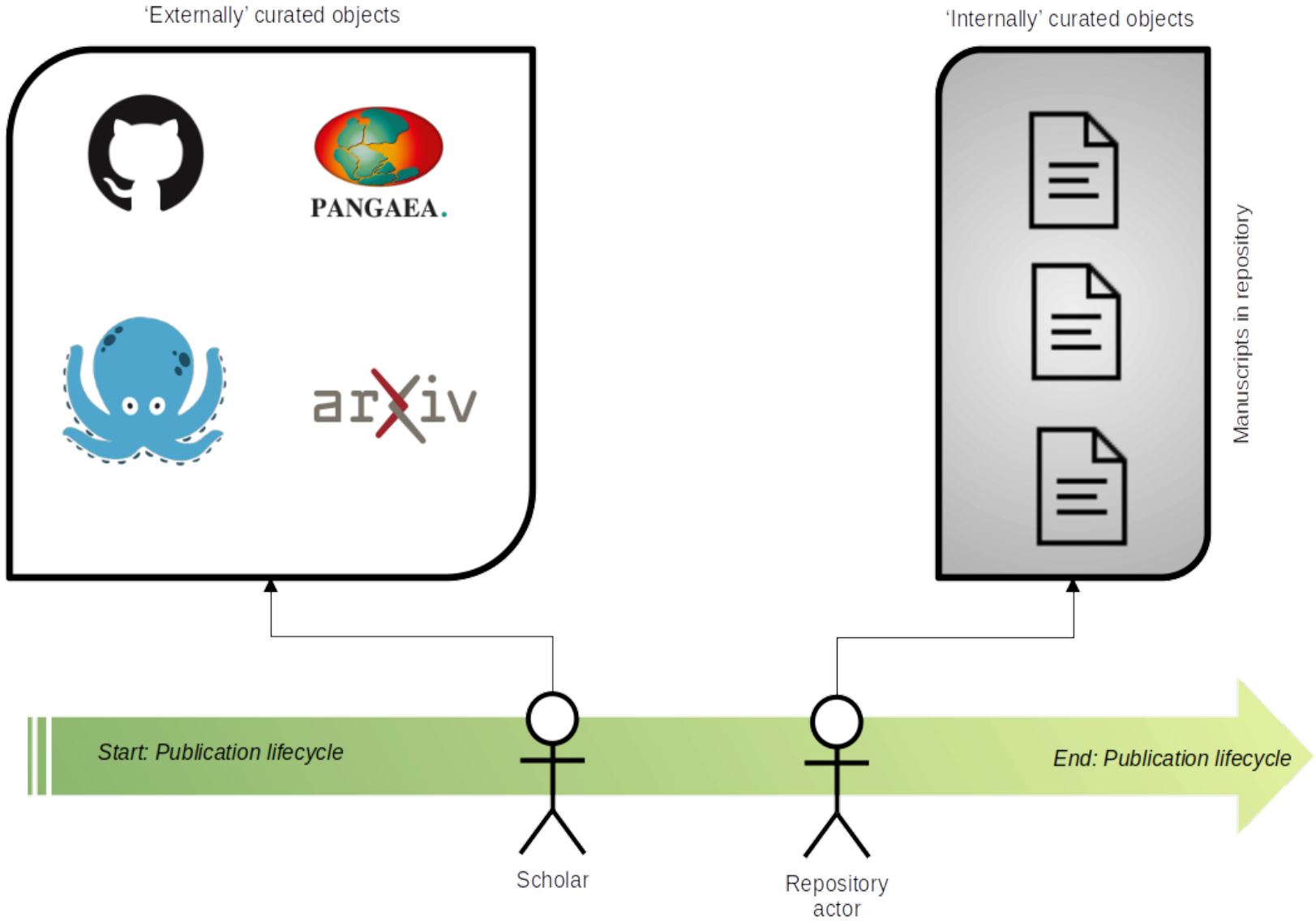
Concept of 'authority of assertion' exists but an under-research area; no definitions

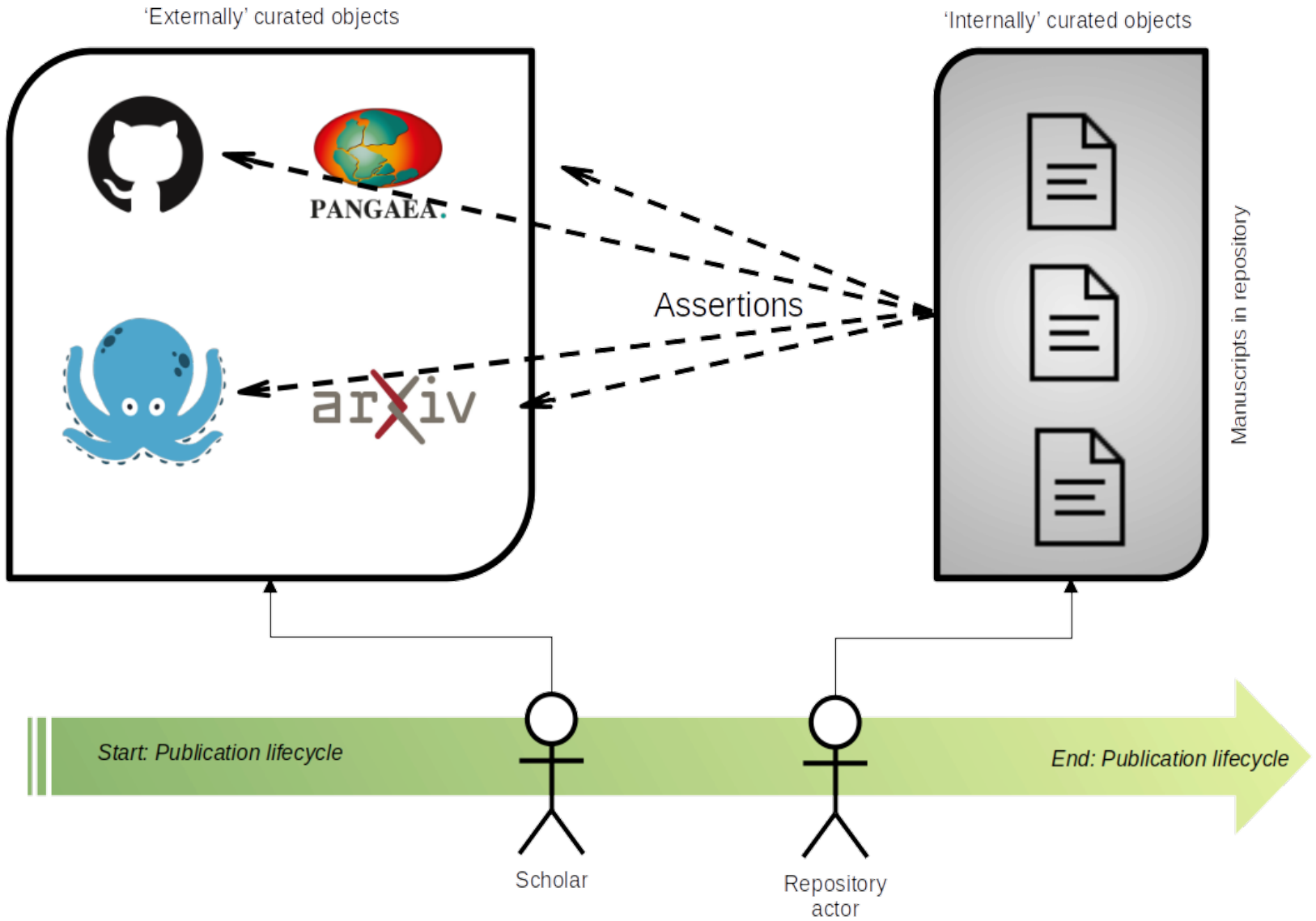
- **authority** = "*official permission or the legal right to do something*"- Cambridge Dictionary
- **assertion** = "*a statement that you strongly believe is true*" - Cambridge Dictionary

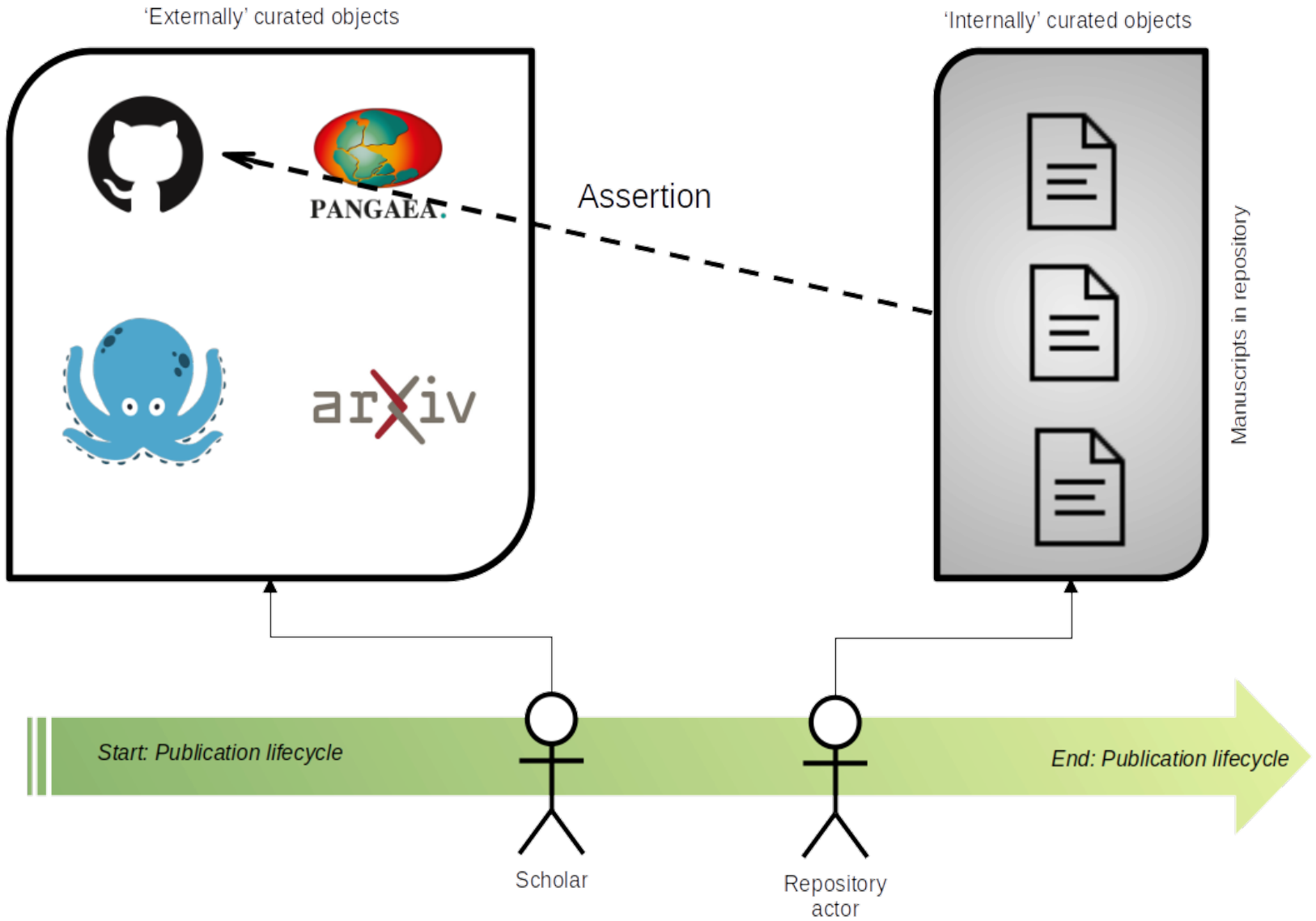
Ergo, could we propose the following:

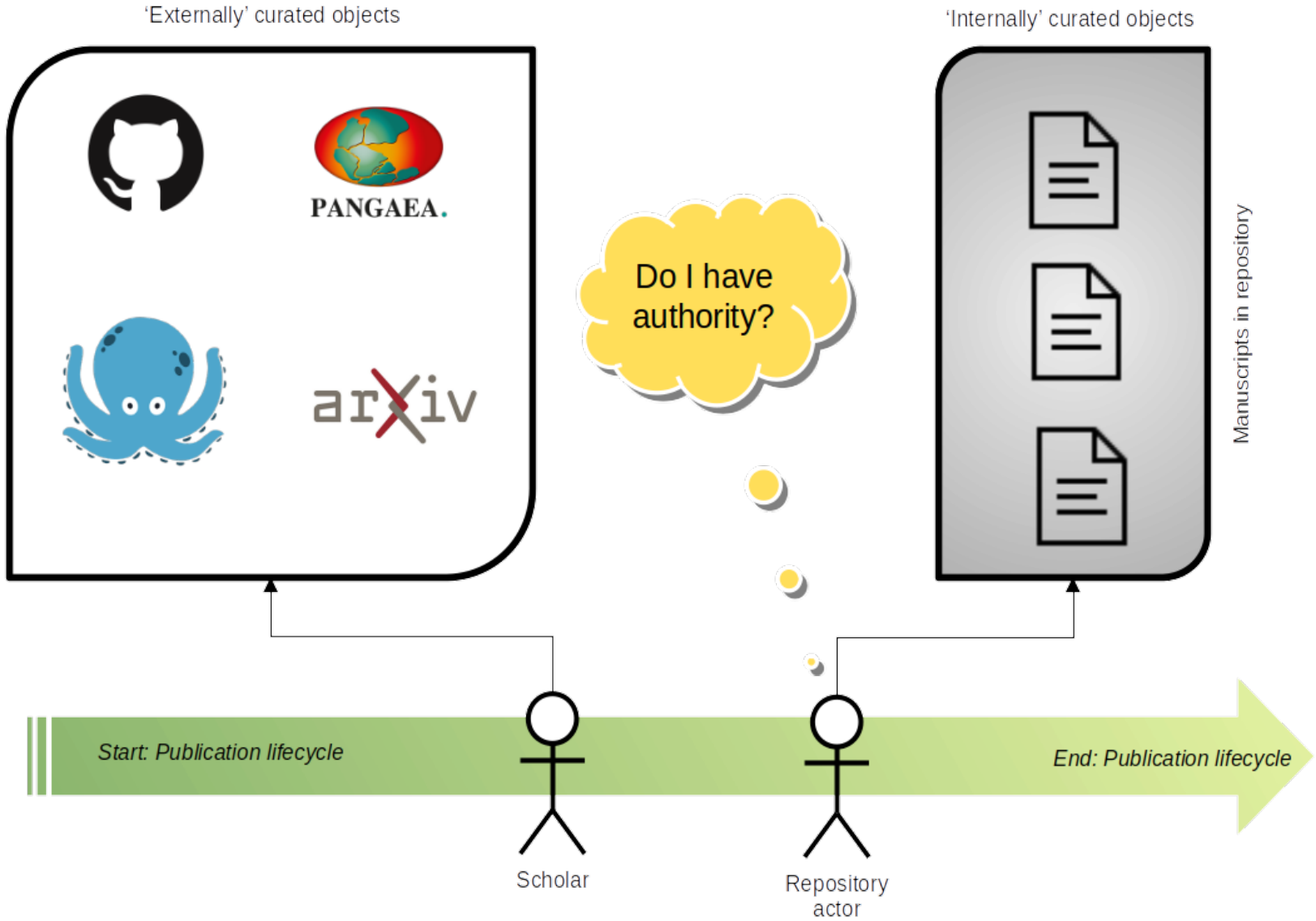
"*The extent to which there exists a right to make metadata statements about objects and their relations, such that these statements are reliable, verifiable, and true*"

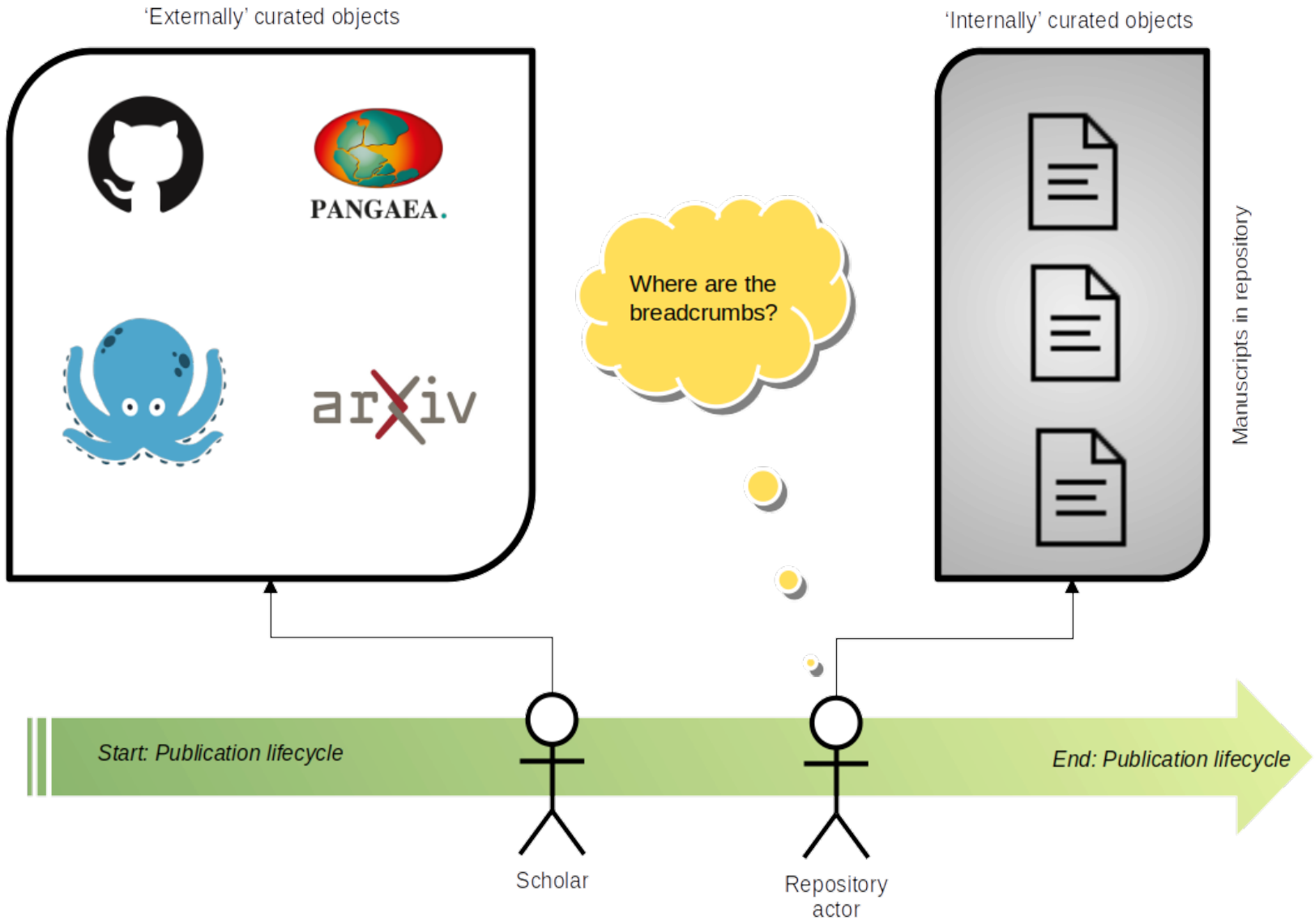
The 'bread crumbs' and authority of assertion

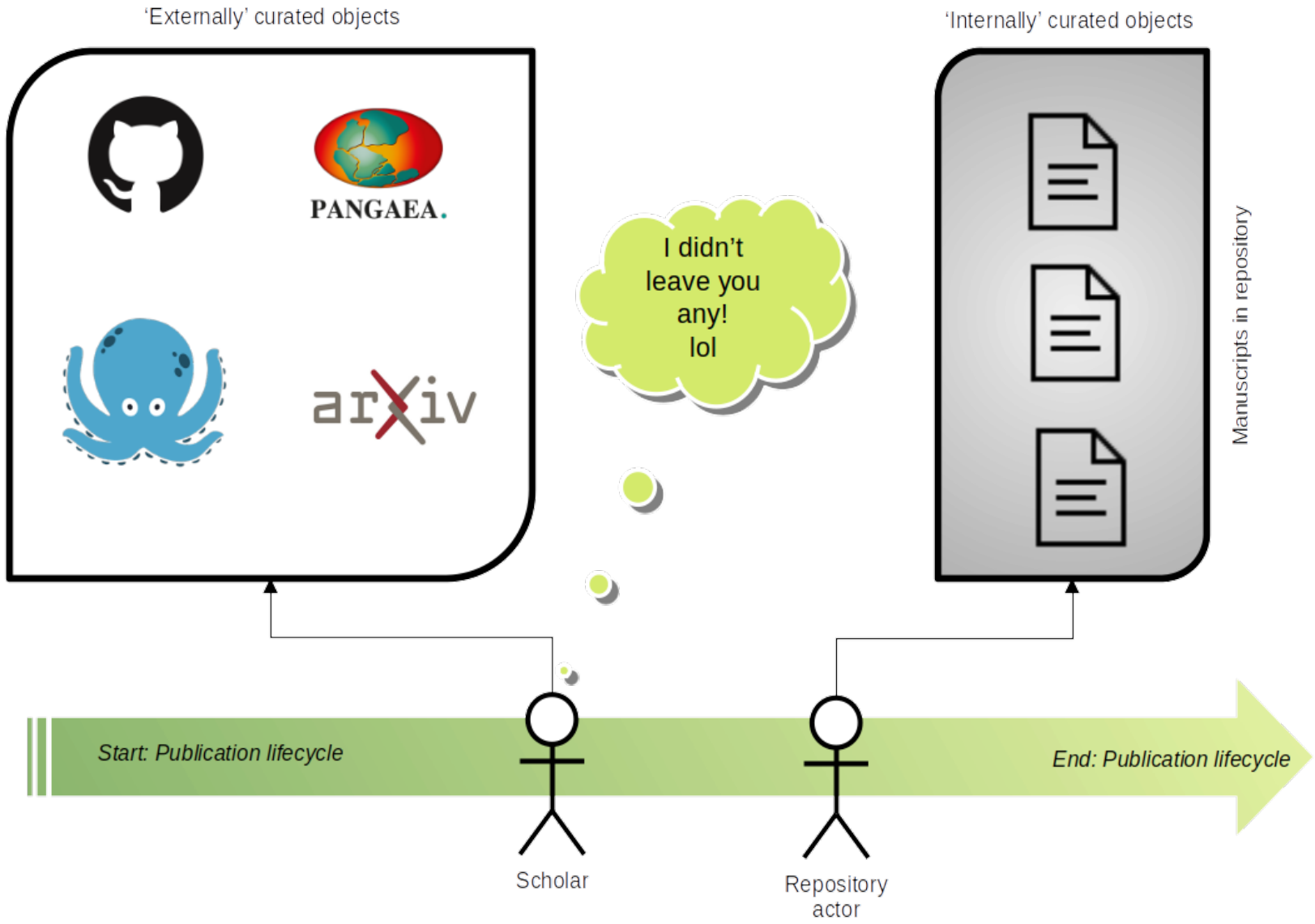

'Externally' curated objects
'Internally' curated objects
PANGAEA.
arXiv
Manuscripts in repository
Start: Publication lifecycle
End: Publication lifecycle
Scholar
Repository actor

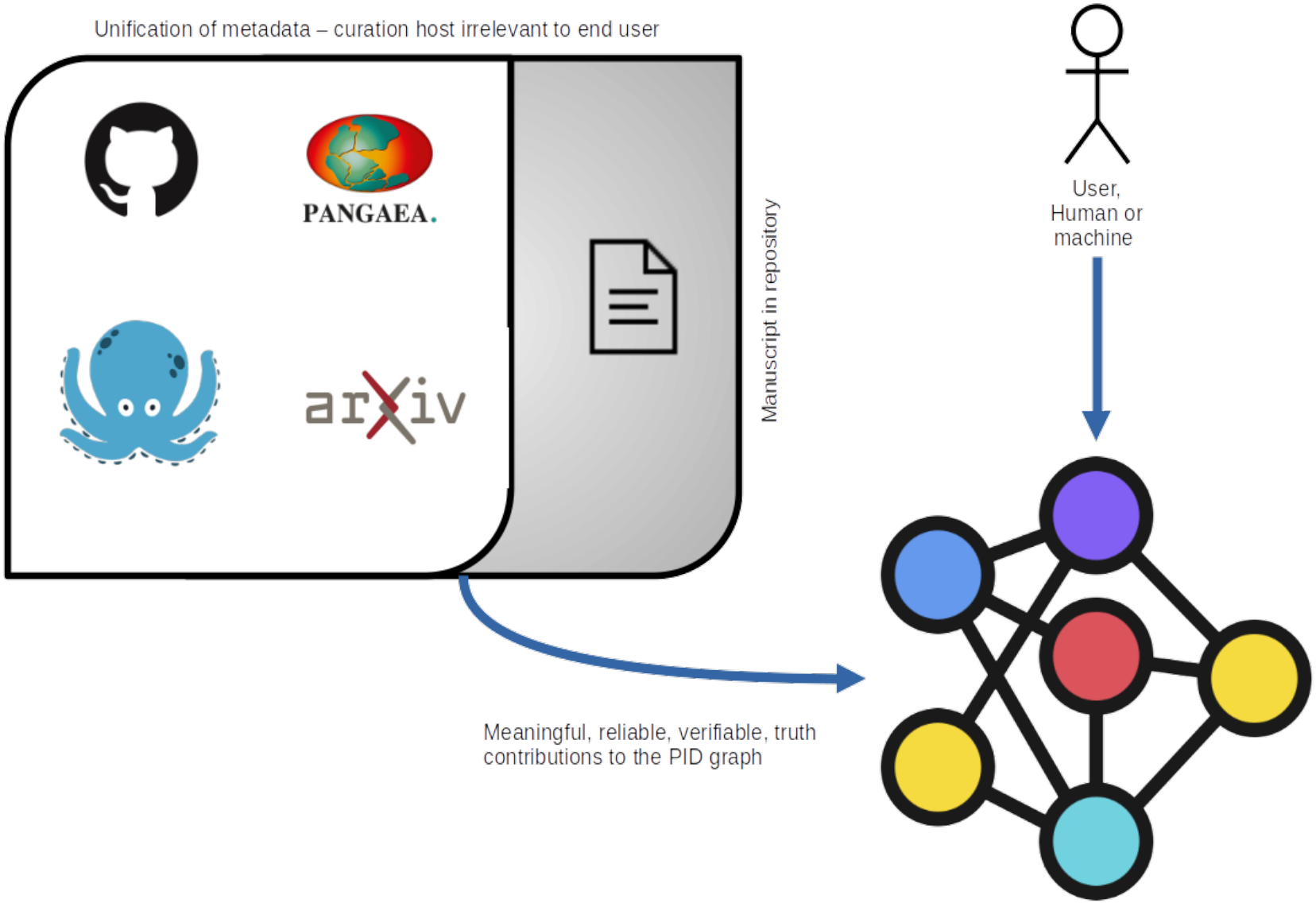

'Externally' curated objects
'Internally' curated objects
PANGAEA.
Assertions
arXiv
Manuscripts in repository
Start: Publication lifecycle
End: Publication lifecycle
Scholar
Repository actor


'Externally' curated objects
'Internally' curated objects
PANGAEA.
Assertion
arXiv
Manuscripts in repository
Start: Publication lifecycle
End: Publication lifecycle
Scholar
Repository actor


'Externally' curated objects
'Internally' curated objects
PANGAEA.
arXiv
Do I have authority?
Manuscripts in repository
Start: Publication lifecycle
End: Publication lifecycle
Scholar
Repository actor


'Externally' curated objects
'Internally' curated objects
PANGAEA.
arXiv
Where are the breadcrumbs?
Manuscripts in repository
Start: Publication lifecycle
End: Publication lifecycle
Scholar
Repository actor


'Externally' curated objects
'Internally' curated objects
PANGAEA.
arXiv
I didn't leave you any! lol
Manuscripts in repository
Start: Publication lifecycle
End: Publication lifecycle
Scholar
Repository actor


Unification of metadata – curation host irrelevant to end user
PANGAEA.
arXiv
Manuscript in repository
User,
Human or
machine
Meaningful, reliable, verifiable, truth
contributions to the PID graph

Assertion and 'PID literacy'

- Components of multi-object, non-standard scholarly objects can often be inferred (by human & *maybe* machine)
- Lack of authority to assert can make *reliable, verifiable, and true* statements difficult or impossible
  - We want to assert the predicate / attribute of that PID, e.g. `IsVersionOf`, `HasPart`, etc.
  - We want relational understanding

Analogy - cataloguing a book and discovering that half the title page is missing...

...

All of which leads to a lack of authority to assert:

- **Undermining** the multi-part scholarly object
- **Undermining** end-user discovery and establishment of provenance
- **Undermining** open research, esp. verification, reproducibility, etc.
- **Undermining** ability of repositories or other scholarly services to contribute meaningful data to the wider scholarly PID graph

Er, what the heck is 'PID literacy'?!

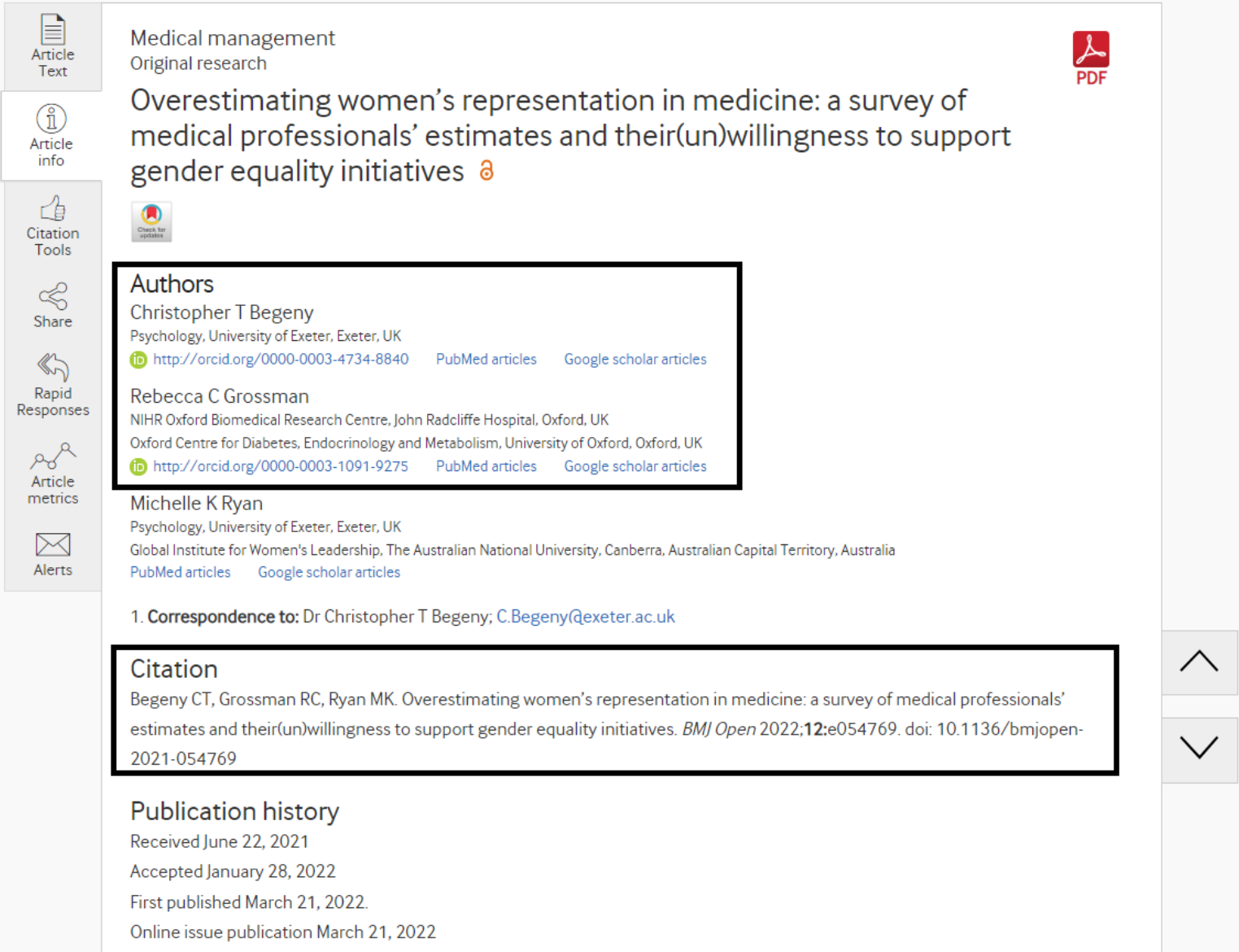
What does a PID literate scholar look like?

"PID literate scholar might display the following competencies:

1. An understanding of persistent identification in scholarship, when it should be used, and its importance to the scholarly record and the wider PID graph.
2. An ability to accurately identify, reproduce, and cite PIDs in scholarship activities.
3. Cognizance of adjacent PID types relevant to scholars' community of practice, such as those devised to identify scholarly 'things' other than academic papers." [1]

Where art thou, PID literacy?

- Challenges within the open research culture space generally
  - e.g. open data, data management planning (DMP), etc.
- Our research reveals [1] many scholars (~35%) demonstrate limited recognition of even common PIDs, e.g. DOIs & ORCIDs
  - Even less understanding of how they should be (re)used - widespread confusion
  - Concerning levels of failure in the PID tests we engaged our participant scholars in



Article Text
Article info
Citation Tools
Share
Rapid Responses
Article metrics
Alerts

Medical management
Original research

PDF

Overestimating women's representation in medicine: a survey of medical professionals' estimates and their(un)willingness to support gender equality initiatives

Authors

Christopher T Begeny
Psychology, University of Exeter, Exeter, UK
http://orcid.org/0000-0003-4734-8840 PubMed articles Google scholar articles

Rebecca C Grossman
NIHR Oxford Biomedical Research Centre, John Radcliffe Hospital, Oxford, UK
Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Oxford, UK
http://orcid.org/0000-0003-1091-9275 PubMed articles Google scholar articles

Michelle K Ryan
Psychology, University of Exeter, Exeter, UK
Global Institute for Women's Leadership, The Australian National University, Canberra, Australian Capital Territory, Australia
PubMed articles Google scholar articles

1. **Correspondence to:** Dr Christopher T Begeny; C.Begeny@exeter.ac.uk

Citation

Begeny CT, Grossman RC, Ryan MK. Overestimating women's representation in medicine: a survey of medical professionals' estimates and their(un)willingness to support gender equality initiatives. *BMJ Open* 2022;**12:**e054769. doi: 10.1136/bmjopen-2021-054769

Publication history

Received June 22, 2021
Accepted January 28, 2022
First published March 21, 2022.
Online issue publication March 21, 2022

RESEARCH ARTICLE

The effect of data sources on the measurement of open access: A comparison of Dimensions and the Web of Science

Isabel Basson[1,2]*, **Marc-André Simard**[1], **Zoé Aubierge Ouangré**[1], **Cassidy R. Sugimoto**[3], **Vincent Larivière**[1]

1 École de bibliothéconomie et des sciences de l'information, Université de Montréal, Montréal, Québec, Canada, **2** DST-NRF Centre of Excellence in Scientometrics and STI Policy; and Centre for Research on Evaluation, Science and Technology, Stellenbosch University, Stellenbosch, Western Cape, South Africa, **3** School of Public Policy, Georgia Institute of Technology, Atlanta, Georgia, United States of America

* isabel.basson@umontreal.ca

OPEN ACCESS

Citation: Basson I, Simard M-A, Ouangré ZA, Sugimoto CR, Larivière V (2022) The effect of data sources on the measurement of open access: A comparison of Dimensions and the Web of Science. PLoS ONE 17(3): e0265545. https://doi.org/10.1371/journal.pone.0265545

Editor: Frank Havemann, Humboldt-Universität zu Berlin, GERMANY

Received: November 16, 2021

Accepted: March 3, 2022

Published: March 31, 2022

Data Availability Statement: Restrictions apply to both datasets used in this paper. Aggregated data is available on Figshare (https://doi.org/10.6084/m9.figshare.18319238). The Web of Science data is owned by Clarivate Analytics. To obtain the bibliometric data in the same manner as authors (i.e. by purchasing them), readers can contact Clarivate Analytics at the following URL: https://clarivate.com/webofsciencegroup/solutions/web-of-science/contact-us/. The Dimensions data is owned by Digital Science, which has a programme

Abstract

With the growing number of open access (OA) mandates, the accurate measurement of OA publishing is an important policy issue. Existing studies have provided estimates of the prevalence of OA publications ranging from 27.9% to 53.7%, depending on the data source and period of investigation. This paper aims at providing a comparison of the proportion of OA publishing as represented in two major bibliometric databases, Web of Science (WoS) and Dimensions, and assesses how the choice of database affects the measurement of OA across different countries. Results show that a higher proportion of publications indexed in Dimensions are OA than those indexed by WoS, and that this is particularly true for publications originating from outside North America and Europe. The paper concludes with a discussion of the cause and consequences of these differences, motivating the use of more inclusive databases when examining OA, especially for publications originating beyond North America and Europe.

Introduction

Over the past 30 years, the democratization of the internet has made it possible for researchers, journals, and publishers to provide free online access to scholarly papers. This practice, also known as open access (OA), allows anyone with an internet connection to access, read, distribute, and download scientific publications for free with no legal or technical barriers [1]. OA publishing is no longer a marginal phenomenon, thanks to a massive rise in OA mandates [2], the introduction of several new OA publishers and OA options for legacy publishers [3], the creation of open-source software that facilitates the production of publications (such as the Public Knowledge Project), and the rise of OA mega-journals such as PLOS ONE and Scientific Reports [4].

The advantages of OA have been well-documented: increased global visibility [5], higher citation rates [6, 7], and a better use of taxpayers' money [8]. Several studies have attempted to

OPEN ACCESS

Citation: Basson I, Simard M-A, Ouangré ZA, Sugimoto CR, Larivière V (2022) The effect of data sources on the measurement of open access: A comparison of Dimensions and the Web of Science. PLoS ONE 17(3): e0265545. https://doi.org/10.1371/journal.pone.0265545

Editor: Frank Havemann, Humboldt-Universität zu Berlin, GERMANY

Received: November 16, 2021

Accepted: March 3, 2022

Published: March 31, 2022

Data Availability Statement: Restrictions apply to both datasets used in this paper. Aggregated data is available on Figshare (https://doi.org/10.6084/m9.figshare.18319238). The Web of Science data is owned by Clarivate Analytics. To obtain the bibliometric data in the same manner as authors (i.e. by purchasing them), readers can contact Clarivate Analytics at the following URL: https://clarivate.com/webofsciencegroup/solutions/web-of-science/contact-us/. The Dimensions data is owned by Digital Science, which has a programme

...

But enthusiasm for PIDs in scholarship; they are perceived positively even though many do not know what they are (exactly) or how to use them

😃

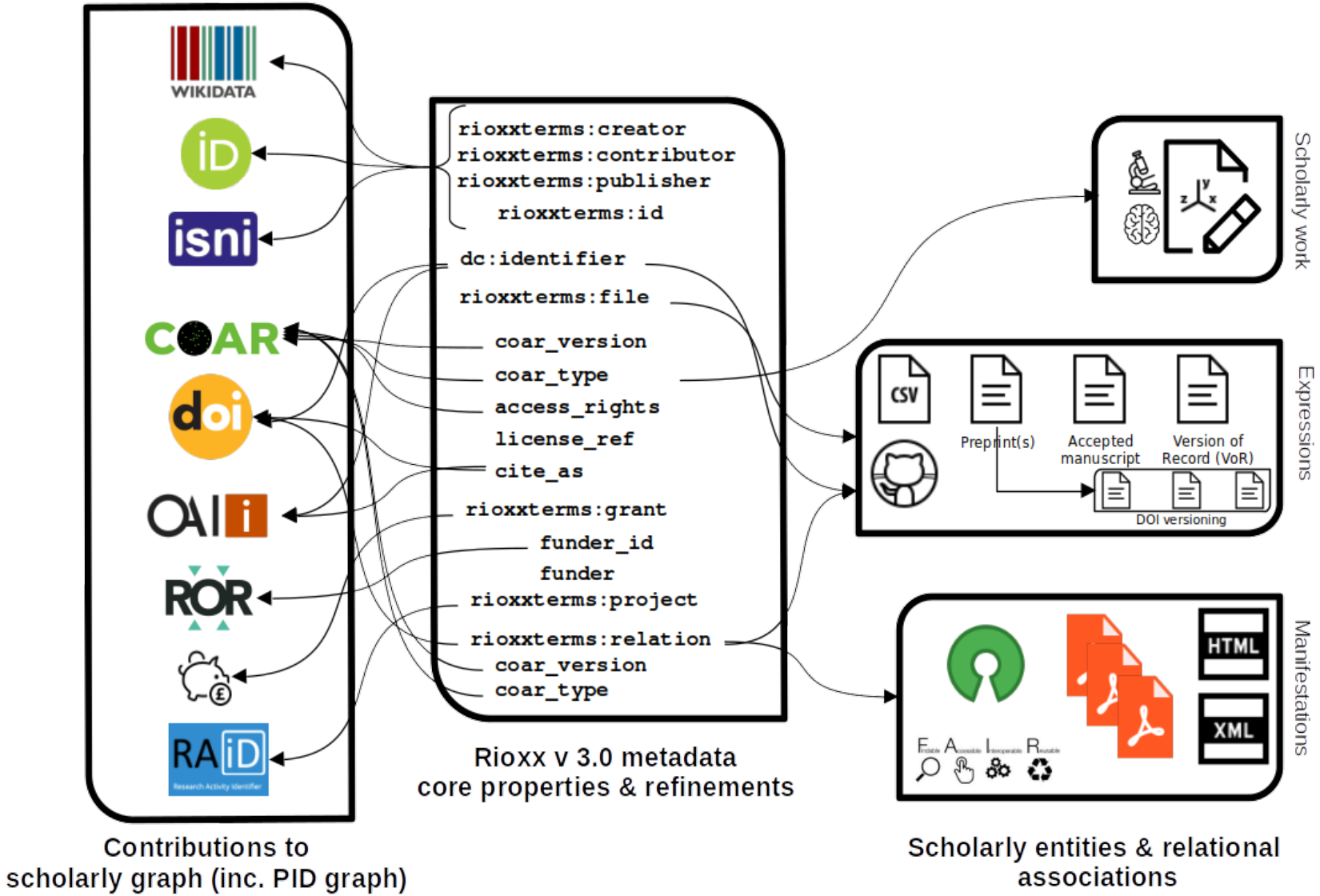
Centrality of scholars to making a better scholarly web

HELP NOW PLEASE! 😟

- To go beyond the academic paper we *need* help from PID literate scholars
- Radical improvement of 'digital scholarship' skills necessary
- 'To do' for open research / scholarly communications teams: advocate; communicate; educate

Better metadata modelling of the new nature of scholarly research objects increasingly reflected in schema

- Rioxx v 3.0 and DataCite Schema



Conclusion and way forward...

- Distributed, multi-part scholarly objects are here -- and will only increase
- Modelling reality and servicing users requires that we describe this complexity in a meaningful way, contributing to the wider scholarly graph of data
- But improving PID literacy as part of wider 'digital scholarship' engagement essential to guarantee sufficient authority to assert, as well as open research goals
- Notwithstanding PID illiteracy...

Questions?

References

[1] G. Macgregor, B. S. Lancho-Barrantes, and D. R. Pennington, 'Measuring the concept of PID literacy: user perceptions and understanding of PIDs in support of open scholarly infrastructure', *Open Information Science*, vol. 7, no. 1, 2023, available: `https://doi.org/10.1515/opis-2022-0142`

[2] G. Alemu. (2022). 'Resource Description and Access (RDA)'. In *The Future of Enriched, Linked, Open and Filtered Metadata: Making Sense of IFLA LRM, RDA, Linked Data and BIBFRAME* (pp. 197-212). Facet, London, available `https://doi.org/10.29085/9781783304943.010`

[3] G. Bilder, J. Lin, and C. Neylon, 'The Principles of Open Scholarly Infrastructure', 2020, available: `https://doi.org/10.24343/C34W2H`

[4] H. Cousijn et al., 'Connected Research: The Potential of the PID Graph', *Patterns*, vol. 2, no. 1, p. 100180, 2021, available: `https://doi.org/10.1016/j.patter.2020.100180`

[5] H. M. Sandy et al., 'Making a case for open research: implications for reproducibility and transparency', *Proceedings of the Association of Information Science & Technology*, vol. 54, no. 1, pp. 583–586, 2017, available: `https://doi.org/10.1002/pra2.2017.14505401079`

[6] S. M. Jones et al., 'Scholarly Context Adrift: Three out of Four URI References Lead to Changed Content', *PLoS ONE* vol. 11, no. 1, p. e0167475, 2021, available: `https://doi.org/10.1371/journal.pone.0167475`

[7] S. M. Jones, M. Klein, and H. Van de Sompel, 'Robustifying Links To Combat Reference Rot', *Code4Lib Journal*, no. 50, 2021, available: `https://journal.code4lib.org/articles/15509`