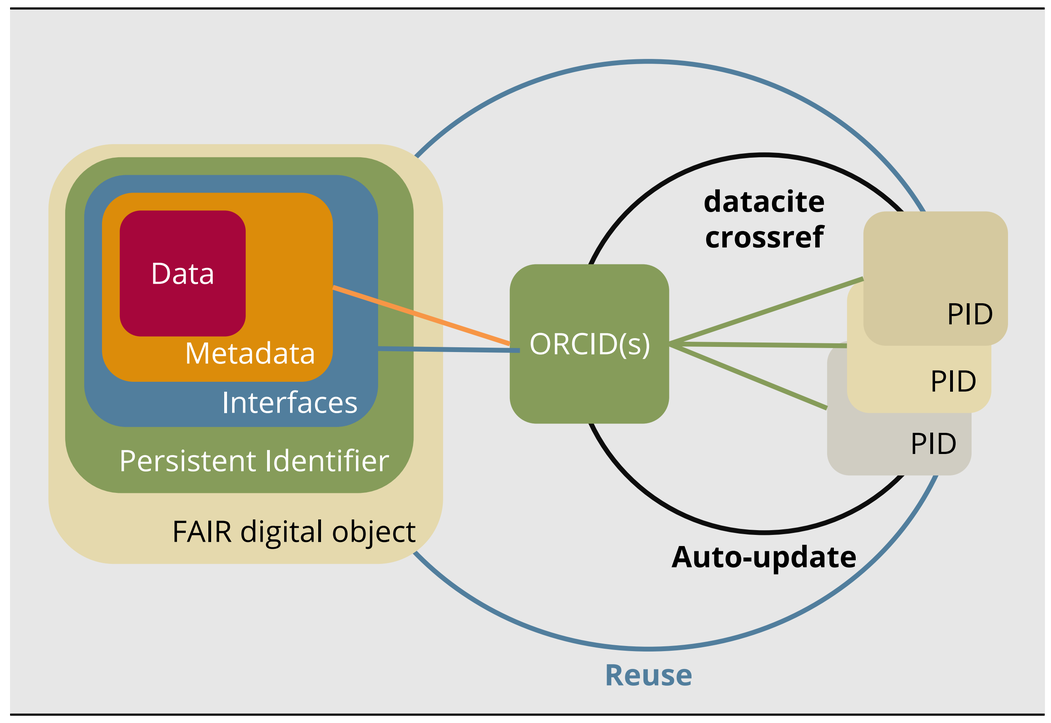
From Data Citations towards effective Valorization and Tracking of Data Reuse

Claire Jean-Quartier, Stefan Reichmann, Miguel Rey Mazon, and Ilire Hasani-Mavriqi
Graz University of Technology, Austria

Reuse of digital objects

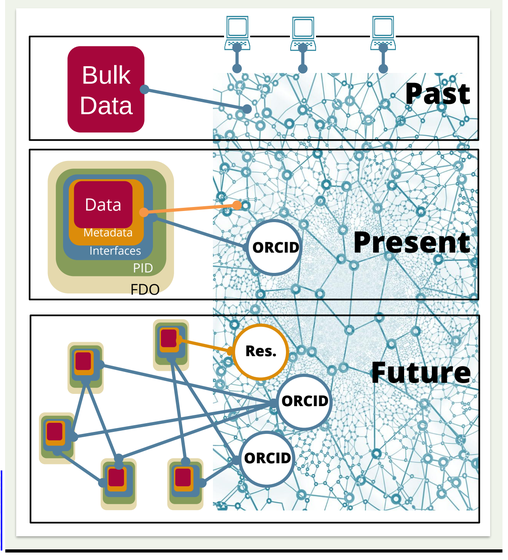
This poster presents a multilevel approach to valorize and track the reuse of digital objects, taking into account skewed adoption of Open Science practices (the “Matthew effect”), particularly in data citation. It introduces workflows to capture published digital records through various persistent identifiers (PIDs) and informal data citations, and extends these workflows to enable the automatic updating of researchers’ ORCID records when their research outputs are reused.

The research question is motivated by the observation that the principles underpinning the growth of the internet as a “web of documents”, based on the Resource Description Framework,[1] have not yet been fully applied to data.



Automation – updates of records upon FDO reuse based on linked PID

From data dumps towards reused FDOs



In the past, data was typically made available as raw dumps, without providing information (usually as linked Unified Resource Identifiers i.e. URIs) about relations to other documents or, more generally, Digital Objects, available online. At present, digital research objects are more typically published and assigned PIDs, but even minimal FAIR practices differ vastly across research contexts (as does data sharing and reuse).[2] Moreover, citation upon reuse is not a common practice,[3] and is at most done informally, either by reference to a publication, or by indicating the web resource where the data is hosted.

The FAIR Digital Object framework intends to solve this problem by providing Digital Objects with enough actionable information to automate its discovery, access, interoperation, and re-use.[4] Adding this information to all Digital Objects would facilitate the reuse of scholarly data and boost innovation by facilitating data discovery, in line with data re-use as one of the main motivations behind the FAIR principles for scientific data management.[5] Therefore, we argue that the FDO concept should be extended to include information in data records about who provided and who reuses data, to enable automated tracking of data reuse and increase the recognition of data providers in an automated manner.

Current initiatives to address this problem, like the OpenAIRE Graph, do not capture data reuse comprehensively.[6] In the near future, the emerging EOSC Federation might include a framework to allow unique identification of data, thus enabling to keep track of citation, avoid duplication, and impact tracking.[7] This should need to be further pursued to pave the way for valorizing data provision and create fair data reuse. More recently, a machine-readable “data reuse information” tag (DRI)[8] has been proposed to enable equitable data reuse for web resources by including information on the ORCID of data producers, to provide contact information to data consumers that can be indicated in the case of data reuse. This could be also realised by adding PIDs to data or other kind of Digital Object, as it is common for scholarly articles.

Multilevel approach to data reuse attribution

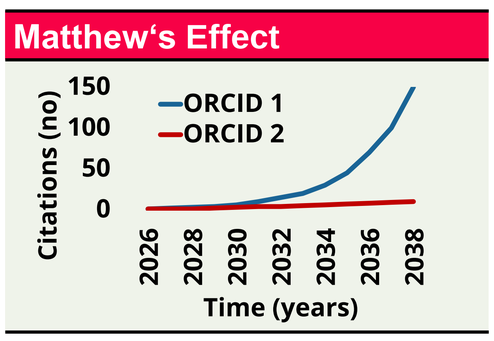
- Observing Matthew effects in the transition to open science practices
- Introducing workflows for capturing published digital records other than scholarly literature
  - based on various PIDs (DOI, HDL, ENA, EGA, W3ID, i.a.)
  - Tracking and integrating informally cited data (URL, i.a.)
- Extension to existing workflows enabling automatic updating of records associated with persistent researcher IDs upon published research objects‘ reuse.

The proposed approach to fostering attribution of reused digital objects builds on two complementary strategies. On the one hand, it extends interfaces of FDO services to automate cross-referencing of reused objects despite incomplete citation practices, complemented by web crawling and alternative methods to integrate digital objects lacking common persistent identifiers (PIDs); in both cases, attribution is relayed through linked persistent researcher identifiers. On the other hand, future research will examine the socio-dynamic effects of digital object citation, with the goal of counteracting inequitable attribution practices and improving the recognition of FDO value based on quality criteria, thereby mitigating biases related to the origin or prominence of citations.

Valorization – Recognition in academia

The technological change involved in data citation, however, must be accompanied by a cultural change. Among the ways suggested to achieve this, alternative metrics (such as the “data index”)[9] have been proposed to incentivise data sharing. However, as with other metrics (h-index, Impact Factor), the social dynamics of attention in academia results in highly skewed patterns of recognition (“Matthew Effect”)[9].

Matthew‘s Effect



Based on „Unpacking the Matthew effect in citations“, Journal of Informetrics 8(2), 2014, by Jian Wang.

Although the Matthew effect can have positive effects in terms of steering researchers’ attention, it also results in a stratified distribution of opportunities in academia, which has been shown to play a role in article citations, countries, institutions, departments, and journals. Similar dynamics can be expected for data citation.

References

1. Klyne, G., Carroll, J. (2004). Resource Description Framework (RDF): Concepts and Abstract Syntax - W3C Recommendation. https://www.w3.org/TR/rdf10-concepts/
2. Paul Groth, Helena Cousijn, Tim Clark, Carole Goble; FAIR Data Reuse – the Path through Data Citation. Data Intelligence 2020; 2 (1-2): 78–86. https://doi.org/10.1162/dint_a_00030
3. Bonino da Silva Santos, L. O., Sales, T. P., Fonseca, C. M., & Guizzardi, G. (2023). Towards a conceptual model for the FAIR Digital Object Framework. In Formal Ontology in Information Systems (pp. 227-241). IOS Press. https://doi.org/10.3233/FAIA231131
4. Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, (2016). https://doi.org/10.1038/sdata.2016.18
5. Paolo Manghi; Challenges in building scholarly knowledge graphs for research assessment in open science. Quantitative Science Studies 2024; 5 (4). https://doi.org/10.1162/qss_a_00322
6. EOSC architecture working group view on the minimum viable EOSC – Report from the EOSC Executive Board Working Group (WG) Architecture, Publications Office, 2021, https://data.europa.eu/doi/10.2777/492370
7. Hug, L.A., Hatzenpichler, R., Moraru, C. et al. A roadmap for equitable reuse of public microbiome data. Nat Microbiol 10, (2025). https://doi.org/10.1038/s41564-025-02116-2
8. Hood, A. S. C., & Sutherland, W. J. (2021). The data-index: An author-level metric that values impactful data and incentivizes data sharing. Ecology and Evolution, 11, 14344–14350. https://doi.org/10.1002/ece3.8126
9. Ross-Hellauer, Tony, Stefan Reichmann, Nicki Lisa Cole, Angela Fessl, Thomas Klebel, and Nancy Pontika. 2022. “Dynamics of Cumulative Advantage and Threats to Equity in Open Science: A Scoping Review.” Royal Society Open Science 9(1). https://doi.org/10.1098/rsos.211032

TU Graz – RDM Team
8010 Graz, Münzgrabenstraße 36/II
rdm.tugraz.at – mail: **rdmteam@tugraz.at**,