High-Quality Metadata: A Collective Responsibility and Opportunity

Adam Buttrick

August 13, 2024 . 8:32 PM

7 min read

https://doi.org/10.54900/j2q51-0jj98

Our community and tools rely on high-quality DOI metadata for building connections and obtaining efficiencies. However, the current model - where improvements to this metadata are limited to its creators or done within service-level silos - perpetuates a system of large-scale gaps, inefficiency, and disconnection. It doesn't have to be this way. By collaboratively building open, robust, and scalable systems for enriching DOI metadata, we can leverage the work of our community to break down these barriers and improve the state and interconnectedness of research information.

On August 3, 2024, at the FORCE11 conference in Los Angeles, the University of California Curation Center (UC3) hosted the first in what will be a series of discussions about community enrichment of DOI metadata: why we need it, how to do it, and who would like to be involved. As part of the California Digital Library (CDL), UC3 is an established leader in collaborative, open infrastructure and persistent identifier (PID) projects. This effort builds on our existing work to enhance scholarly communication and research data management using the collective expertise of our community. To broaden the scope of this engagement and include those who could not attend, we'd like to share the initial thoughts that guided this discussion, organized as a series of observations, principles, and goals.

Collaborative Infrastructure as a Shared Source of Truth

Building the corpus of DOI metadata over many years has taught our community an important lesson: when we work together to define how infrastructure should exist, how we want to build and improve upon it, we arrive at better outcomes than when we do this work alone. Collective stewardship of our shared sources of truth is what allows us to make the right decisions for as many people as we can.

It is thus unsurprising that Crossref and DataCite, two of the primary organizations responsible for this work, have grown in scope and impact alongside the systems they have brought into existence. In pursuing the immense value and network effects that result from solving the same problems in the same places, they have demonstrated that it is possible to align a diverse set of actors around the goals of open infrastructure and open research information. In their embrace of the Principles of Open Scholarly Infrastructure (POSI), in the example and sponsorship they have provided to new services, through their constant advocacy, they have steered our efforts toward greater openness and continued improvement.

The fruits of this labor extend beyond their own services to everything that is derived therefrom. From bibliometric analysis to research evaluation, from discovery services to funder impact tracking, this rich web of scholarly metadata is the basis for so much of our work. It bridges open and closed systems, fosters interoperability, and helps to guarantee the integrity of the scholarly record.

Enriching Metadata in Service-level Silos Creates Inefficiencies and Disconnects

Despite these successes, a persistent problem has arisen from the maintenance model of DOI metadata. In its current conception, making corrections or improvements to records are the almost exclusive remit of their depositors. As a result, much of the work to improve records is done in services that consume DOI metadata, as opposed to at their sources.

To a great extent, these efforts are admirable. They demonstrate the ingenuity and resilience of our community to route around any obstacles we encounter. This service-level enrichment, however, also leads to the duplication of work and a more fragmentary, isolated view of research information that our shared efforts seek to avoid. When the collaboration and observability derived from the "one place, one thing" model of DOI metadata is removed, changes to records occur multiple times in many different places. Each change introduces the potential for discrepancies that have to then be reconciled. Since DOI metadata relies heavily on the accurate linking between authors, institutions, and other works, these individual discrepancies can quickly compound into aggregate views that are wildly divergent from their sources and each other.

This is, again, contrary to our building and maintenance of DOI infrastructure as our source of truth. We derive value from DOIs by having a persistent reference to an object, a description of that object, and by being able to perform some basic validation that the object exists. Reliance on service-level enrichment leads to a more unstable arrangement, where to either provide or discern a more complete description, we have to stitch together different views of an object in multiple services that have no corresponding guarantee of stability, provenance, or persistence. As a result, organizations can invest a great deal of time into service-level workflows, only to lose access to them, for the services to change or degrade, and for all of their efforts to become non-transferable or lost.

A More Comprehensive Form of Research Information Can Be Achieved Through Diverse and Consensus-Based Descriptions

While it is important to acknowledge the complications that result from service-level enrichment, the history of this work has also shown that it is necessary to synthesize many forms of improvement to achieve complete and accurate descriptions. The investment made by users in these services results from them being permitted to make changes that cannot be made at the source and because it is simply not true that a depositor of DOI metadata always has the time, resources, or ability to produce a better form of it. Perhaps more importantly, the depositor can also not anticipate in advance what every user will require from their records. Instead, it is the diverse feedback from all users that captures their corresponding needs from this metadata, correcting for gaps, errors and biases that may be present when we rely on the depositor as the sole source of truth.

At the same time, to guarantee that records remain usable for all, we need to build consensus mechanisms that define how and when changes should be applied, as well as when they are correct and appropriate. Here, we could have an extensive discussion about these specificities, but the point should never be to anticipate every possible scenario in advance. Instead, we should determine what structures are needed to navigate these issues as a community, from their most basic to their most complex. There are countless examples we can draw from: ROR's community curation model, the coopetition framework of the Generalist Repository Ecosystem Initiative (GREI), the rigorous analysis done by the Centre for Science and Technology Studies (CWTS), all of which demonstrate that collaborative, community-driven approaches are both effective and sustainable in guiding improvements to our sources of truth.

Empowering the Community to Validate and Improve Metadata

While unsurprising relative to the many systems we know to be doing this work, that records may be more frequently improved outside their sources than they are within them suggests the need for a change in approach. Past, successful efforts to improve DOI metadata have focused on lobbying depositors to contribute better and more complete records. Although still needed for certain aspects of records that can only be improved by their authors, the overall work should be refocused away from this advocacy model, relative to what we know can be accomplished from service-level improvements.

Specifically, we must allow for the same community enrichment of DOI metadata to occur at the source, meaning Crossref and DataCite, such that these records are maintained at a comparable level of quality and completeness. By doing so, we better reflect the existing reality where users are direct contributors to this metadata, further refining it from the baseline provided by depositors to be more comprehensive, correct, and aligned with their needs. Existing work being done at the service-level can then move upstream, and achieve the same visibility and collective stewardship that has been integral to the success of DOI infrastructure.

New Systems for Enrichment Should Be Open, Reproducible, Scalable, and Technically Sound

This visibility and stewardship requires open and reproducible enrichment processes. At the most basic level, openness and reproducibility are needed to validate both the quality and performance of any enrichment process. Without them, we have no way of accurately determining whether a given set of improvements meet the needs of the community or are useful to apply at scale. We likewise establish confidence in enrichment by allowing users to validate things like the representativeness of our benchmarks, the soundness of our designs, and the overall improvements that result from any work. This openness also allows users to immediately leverage and iterate upon any enrichment process, such that they can derive value from it separate from or in the absence of its implementation.

To succeed in this way, the work of enrichment must also be able to transition from any one system to another. Who has the resources, interest, and expertise to engage in these activities can and will shift over time. Openness and reproducibility ensure that we can adapt to these changes, transfer responsibilities, welcome new contributors, and accommodate attrition.

Enrichment Systems Require Shared Standards and Provenance Information

We know from past efforts that we need to bring together a diverse group of users and a disparate set of systems to improve DOI metadata. We likewise can gather from the success of DOI infrastructure and the enrichment found in service-level descriptions that this is an achievable outcome. However, to realize this aim also requires that community enrichment occur in consistent and actionable ways.

At a practical level, what this means is shared formats for describing enrichment that can be generated by any system and include provenance information linking the enrichment back to its source. Whether a user is submitting an individual correction or some matching process is updating to millions of records, we should indicate the source for these actions such that they can be evaluated, approved, or reverted, as needed. Enrichment must likewise be described in machine-actionable ways, meaning that if we establish consensus or thresholds for forms of improvements, these can be acted on automatically and occur at scale.

This approach has firm precedents in our existing systems. Both Crossref and DataCite's schemas have been refined through multiple iterations of planning and community feedback and are used in a constant stream of reference, creation and updates to existing works. We can thus use these as models to rationalize enrichment within their well-defined frameworks.

Moving Forward Together

Community enrichment of DOI metadata poses significant challenges, but not insurmountable ones. Our initial meeting in Los Angeles reaffirmed the community's interest in tackling this together, just as we have done with other successful infrastructure initiatives. Through collaboration and use of our shared expertise, we can build a better, more connected system of research information. UC3 will be continuing these critical discussions, and we encourage you to stay engaged with us. If you have any additional questions or would like to contribute further to these conversations, please feel free to reach out to me at adam.buttrick@ucop.edu. We hope you will join us in this work!