



Richer metadata. Together.



# Building a Community-Driven Future for Collaborative Metadata Enrichment

Force2026 | Panel Discussion | 3 June 2025  
Adam Buttrick and Dione Mentis



# About COMET

# The COMET community

- Stewards of quality scholarly metadata.
- Practitioners of an evidence-based community enrichment model.
- Collaborators building shared workflows to propagate community-sourced enrichments into the systems that maintain and disseminate scholarly metadata—at scale.

# COMET's story so far



## Reignite momentum

Discussions at Force2024 and the Paris Conference on Open Research sparked renewed interest in shared approaches to metadata enrichment.

## Convene community

The community formed a task force to discuss needs, approaches, and potential solutions, culminating in a call to action.

## Pilot validation

Pilot projects with community partners tested approaches and produced enriched metadata to round-trip into PID infrastructures.

## Foundations for scale

Pilot work is being operationalised into standards, guidelines, and integration pathways to extend community capacity and infrastructure.



Why we are here

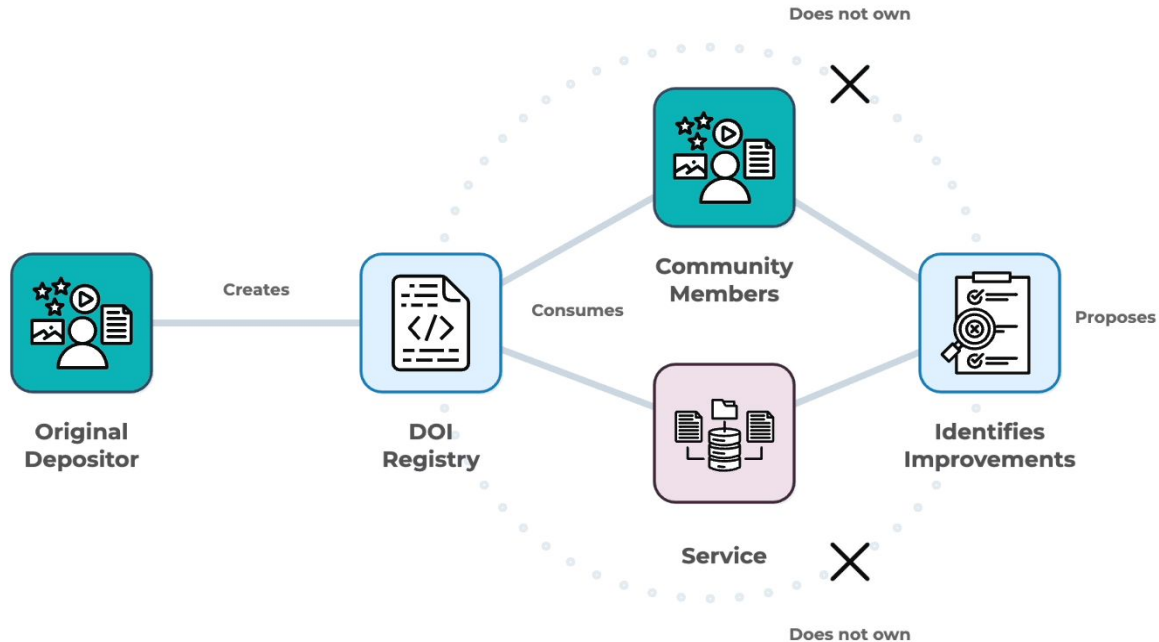
“ Our current workflows for maintaining and improving the scholarly record are disconnected and inefficient.

# Characteristics of metadata production workflows

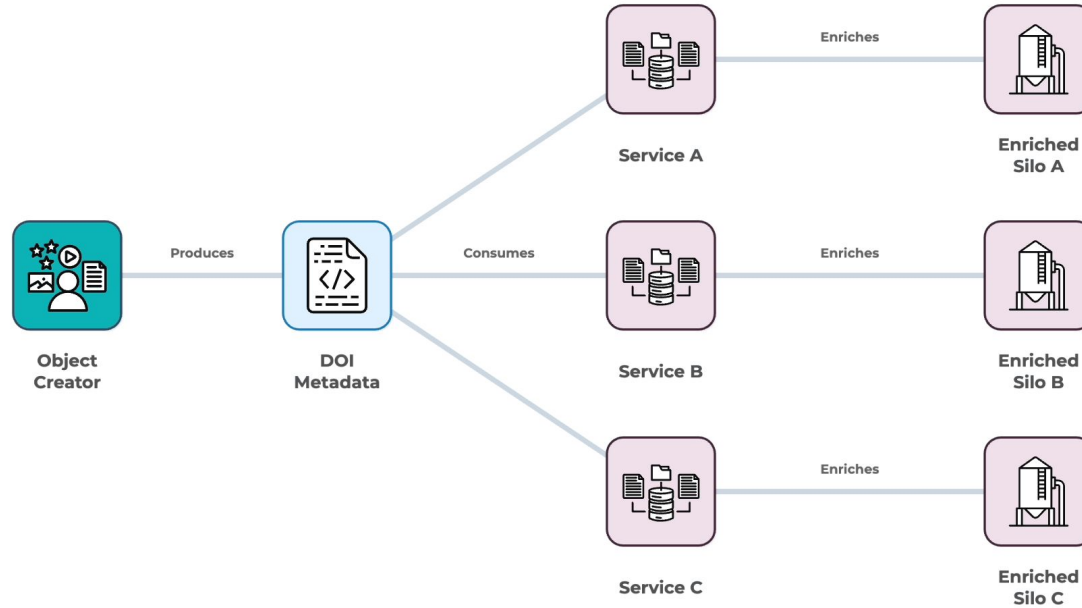
- **Complexity:** many systems, different schemas, multiple versions
- **Non-transparency:** undocumented, siloed in proprietary systems
- **Duplicative:** improvements rarely shared between systems
- **Resource intensive:** high curation effort, high infrastructure costs



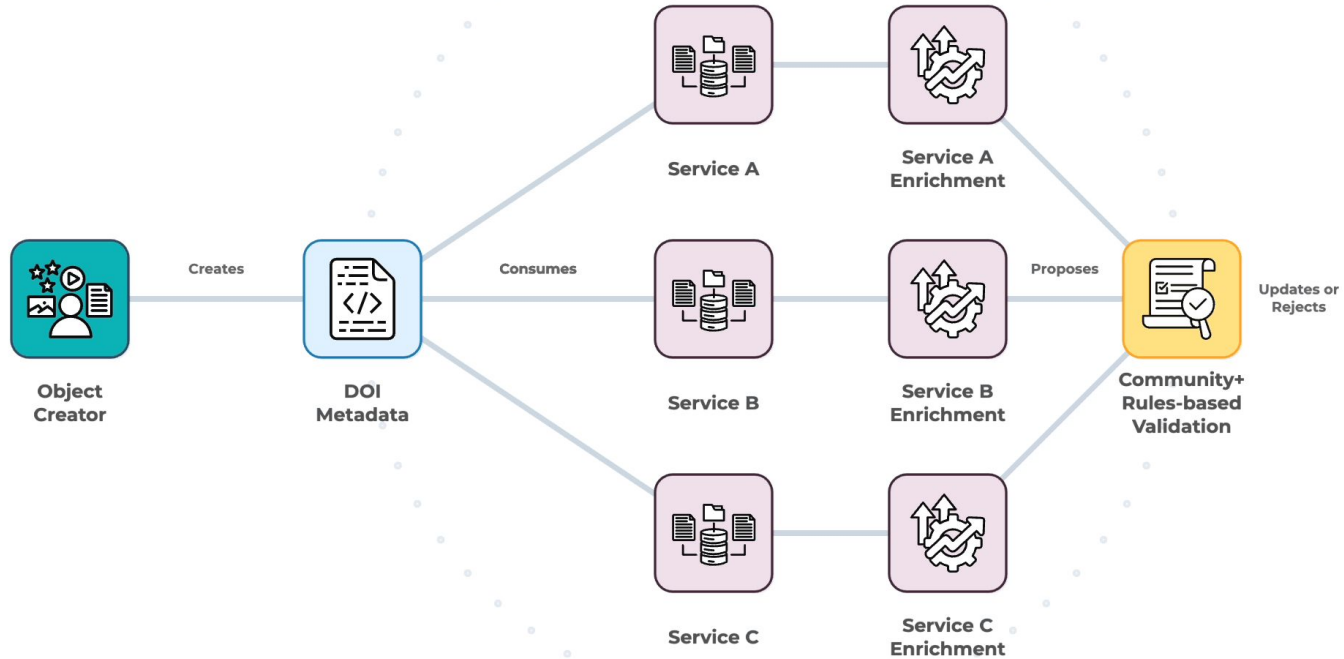
# Individual stewardship model



# Isolated benefits



# Shared stewardship and benefits





# The COMET Model



**Collective stewardship**

**Collective benefit**

**Trust through transparency**

## Foundational principles

1. Shift the responsibility of metadata quality from individual depositors to a distributed community partnership
2. Address the fragmented benefits that result when organisations tackle metadata problems in isolation
3. Establish trust through open processes

# Four focus areas

1. Unite quality metadata sources and enrichment methods
2. Target diverse content types and stakeholders
3. Provide multiple round-tripping pathways
4. Treat metadata fields as features

# Eight prioritised “fields as features”

1. Author/Creator and Contributor
2. Affiliation
3. Funder/Funding
4. References and Related Identifiers
5. License/Rights
6. Title
7. Work/Resource Type
8. Language



# COMET pilot projects



# Community enrichment is real!

- COMET has generated tens of millions of enrichment records, closing high-impact gaps in the scholarly record
- Built the infrastructure to do this work repeatedly
- COMET enrichments are already flowing into DOI metadata
- All work is fully open source. Anyone can reuse and extend the data, models, and methods
- Proven model for how to do enrichment, building capacity anyone to do the same work



How did we make it exist?



# Community

# Starting with community

- First step, solicit community expertise
- Highly-focused sessions, worked through all technical and social aspects of making metadata quality a collective responsibility
- Distilled those insights into a set of thematic principles (COMET Model) and technical requirements



# Principles to practice

# Principles to practice – pilots

- Organize work as a series of pilot projects, each addressing an important metadata gap
- COMET goes first, working in the open, so that real-world challenges and opportunities are exposed
- Build out infrastructure capacity so that others have both a model and the means to take up the same work themselves

## Example pilot - arXiv enrichment

- arXiv registers its DOIs with DataCite letting us test enrichment and integration with DOI metadata
- Metadata improvements are felt immediately and widely
- Large, well-known metadata gaps in affiliations and funding
- Corpus is large enough to confront hard problems at scale, but bounded enough to be tractable, whatever we solve generalizes to any full-text source

# Evaluation

- Empirical forms of evaluation tell us what success really means
- Evaluate with train and test datasets from random samples of the enrichment input
- Design evaluations for those datasets using standard measures like precision, recall, and f-scores





# Evaluation

- Train and test datasets become durable artifacts
- Empirical benchmarks make the best approach to enrichment more than a matter of opinion
- Community can build new solutions with these datasets and evals



# Constraints

- Every method has to be fast, efficient, cost-effective, fully open, and able to be iterated upon
- Scholarly record is a wildly diverse and constantly evolving target
- Methods that are slow, expensive, or closed will not find traction and cannot be maintained
- Constraints drive creativity and act as another form of evaluation, measuring cost, speed, and maintainability

## arXiv enrichment example - funding metadata

- End-to-end, deriving funding metadata from all of arXiv cost ~\$1,500 USD total, ~1 day to produce on cheap, widely-available compute
- Matches frontier model performance at ~66x cost reduction
- Artifacts the community fully owns, can extend, and run themselves, in perpetuity
- Enrichment once limited in closed bibliometric systems is now in reach of anyone willing to do the work and has a clear path into PID metadata

# arXiv enrichment methods

- Trained small LLMs to extract affiliation and funding metadata from all works, then enriched with Crossref's ROR ID matching
- Mined full text for dataset and software repository links, used community ML models to identify when a software link is the work's own implementation
- Derived and asserted back over 1 million missed citations from Crossref metadata
- Reused and extended what the community already built well
- Every component had to clear the same bar, state of art on benchmarks while being fast, efficient, cost-effective, and fully open



Demo

# Meeting community expectations

- Once we've produced the enrichment data and the open methods, make sure the work meets the community's standards in real time
- Run structured feedback sessions that walk community partners through every aspect of the project
- Hold public community meetings to interrogate the work with wider audience
- Focused outreach for those doing similar work to find opportunities to extend, refine, and collaborate

# Pilot project partners



# What comes next

- Fully-documented pilot projects, open training and benchmark datasets, endorsed by the community
- Provides others with a working model to do the same work
- Co-design an additive enrichment layer with DataCite that integrates COMET's enrichment alongside the original DOI metadata





# Round-tripping

# What do we mean by round-tripping?

- **Successful round-tripping:** enrichments flowing back into the systems that maintain and disseminate scholarly metadata
- **Many pathways:** publisher platforms, repositories, CRIS systems, and, importantly, PID infrastructures.
- **PIDs are crucial:** offering a shared distribution pathway that keeps improvements from fragmenting across the ecosystem.



## Round-tripping pathway: integration with DataCite

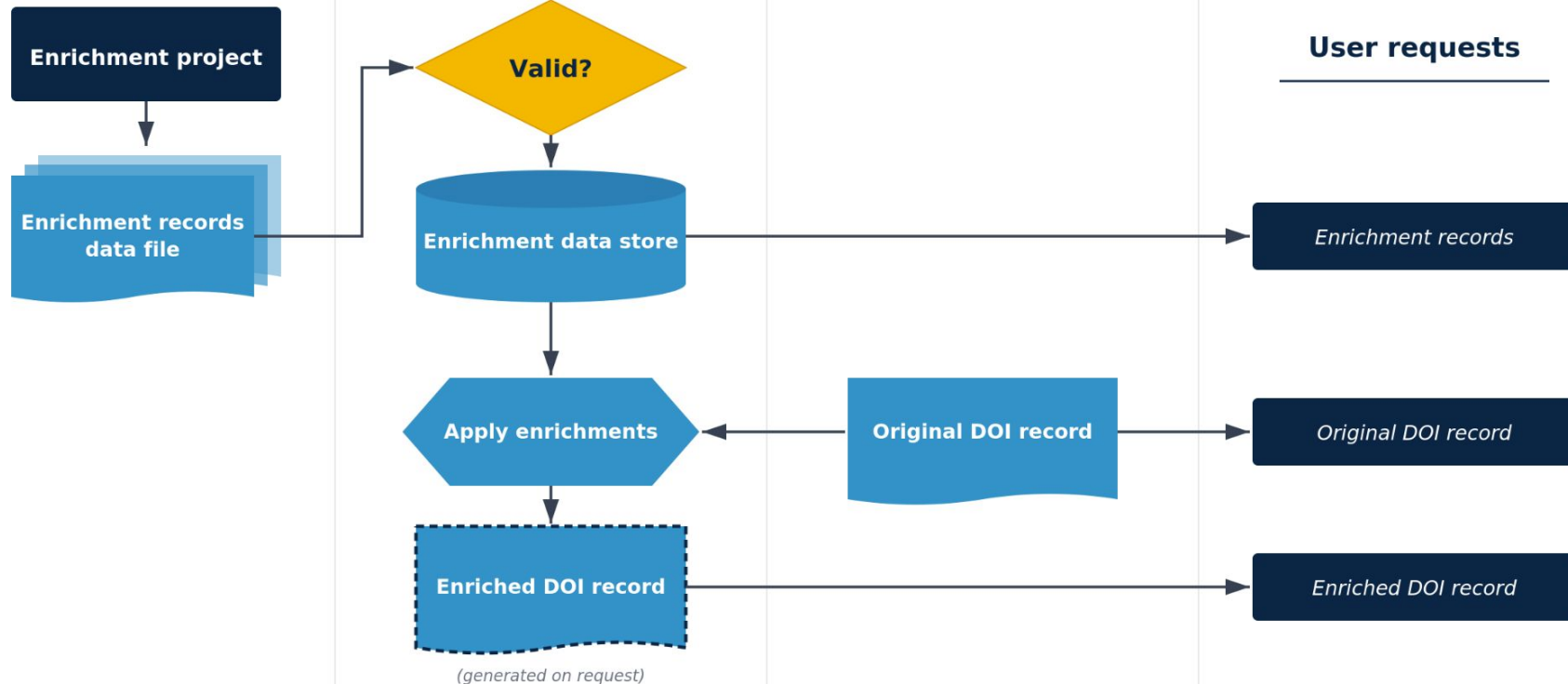
## COMET

## DataCite enrichments pipeline

## DataCite metadata store

## DataCite REST API

### User requests



# DataCite enrichment record data model

	Field	Description
Target DOI	doi	The target DOI of the enrichment record.
Provenance	contributors	The source entities of the enrichment represented as an array of contributors.
	resources	Resources about the enrichment method, such as project documentation and related datasets, represented as an array of relatedIdentifiers.
Action	field	The top-level field to enrich. For example, creators, relatedIdentifiers, or types.
	action	The action that the enrichment performs.
	originalValue	When the action is update, updateChild, or deleteChild, the original value of the field or child value. Otherwise, this field is empty.
	enrichedValue	When the action is update, updateChild, or insert, the enriched value of the field or child value. Otherwise, this field is empty.

# Transparency of provenance and process

## Contributors

Who produced the enrichment

name

**Collaborative Metadata (COMET)**

nameType

Organizational

contributorType

Producer

## Resources

Sources the enrichment links to

relationType

IsDocumentedBy

resourceTypeGeneral

Project

relatedIdentifierType

DOI

relatedIdentifier

10.82461/m8a8-m211

relationType

IsDerivedFrom

resourceTypeGeneral

Dataset

relatedIdentifierType

URL

relatedIdentifier

<https://huggingface.co/datasets/cometadata/arxiv-preprint-matching-results>



# Discussion



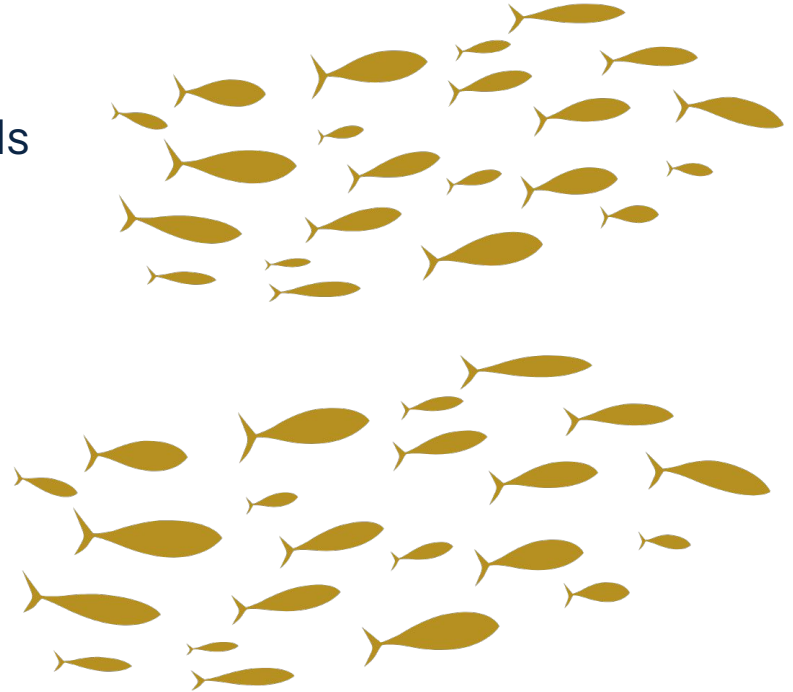
Join us



# Many ways to get involved

- Showcase your enrichment project
- Contribute towards enrichment standards
- Participate in project collaborations
- Offer in-kind or financial support
- Join our community meetings

[cometadata.org/join-us](https://cometadata.org/join-us)



# COMET community meetings

Join us at the next COMET community meeting on **14 July** at 3 PM UTC.



[cometadata.org/community](https://cometadata.org/community)

# Open calls for community feedback

- **DataCite enrichment layer:** Try out the DataCite enrichments API and give input into its next iteration → [datacite-suggestions/discussions/236](https://datacite-suggestions/discussions/236)
- **Posters.science project:** Improve PosterSentry, the scientific posters classifier in developed, with better training data → [survey.posters.science](https://survey.posters.science)



# Thank you!

[cometadata.org](https://cometadata.org)

