
RESEARCH ARTICLE

Semantic Labour and the Politics of Metadata: Scholarly Infrastructure in the AI-Driven Academy

Elena Battaner Moro

Universidad Rey Juan Carlos

The structured, machine-readable metadata that researchers and librarians generate to make scholarly outputs retrievable has become, in the age of artificial intelligence (AI), the decisive input shaping what knowledge can be found, ranked, and generated. This article argues that this metadata is produced through a cycle of semantic labour—the cognitive, administrative, and technical work of encoding research outputs in structured, machine-readable form—that is largely unremunerated by its commercial beneficiaries and that this labour is the mechanism through which a broader process of epistemic capture operates. As enriched metadata is enclosed within proprietary knowledge graphs and AI products, especially the retrieval-augmented generation (RAG) systems now central to scholarly search, the coverage gaps, classification biases, and ranking logics embedded in those systems are operationalised at scale, transferring authority over what counts as legible knowledge from scholarly communities to commercial infrastructural providers. This article traces this dynamic through a six-stage model of the metadata production cycle, analyses the mechanisms of commercial enclosure and their epistemic consequences, and assesses the conditions under which open metadata infrastructures could constitute a genuine counterweight. The conclusion argues that metadata sovereignty, understood as the capacity of scholarly communities to maintain collective authority over the semantic conditions of their epistemic legibility, is a central challenge for scholarly communication in the AI era.

Keywords: metadata sovereignty; semantic labour; epistemic capture; retrieval-augmented generation; knowledge infrastructures; scholarly communication

1. Introduction: The Politics of Metadata in the AI-Driven Academy

For most of the twentieth century, scholarly metadata occupied a peripheral role in academic publishing.¹ Maintained by librarians and indexers, it consisted of the bibliographic information required for cataloguing and retrieval—for example, author names, titles, journal designations, and subject headings. It was mainly a descriptive supplement to the scholarly record rather than a constitutive element of it.

That description no longer holds. In the current wave of artificial intelligence (AI) integration into scholarly workflows, enriched metadata has moved from the margin to the centre of value creation. It is now the decisive input shaping what scholarship can be retrieved, how scholarship is ranked, and what AI tools can generate about the scholarly record. Matthew Mayernik (2020) argues that metadata should be understood not as a flat definitional shortcut but as both process and product, created through human effort and inseparable from the infrastructures that give it meaning. In the AI era, that infrastructural significance has become explicitly political.

¹ For the concept of metadata, see Matthew Mayernik (2019, 2020), who understands metadata as both process and product and, more specifically, as the processes and products that enable entities to become accountable as evidence. Kyle Parry (2023) rejects the reductive definition of metadata as “data about data,” emphasising instead that metadata are situated and interpretive structures through which objects are organised, related, and made intelligible.

Infrastructure studies (Bowker and Star 1999) has long established that infrastructures are never neutral. Social and political choices are embedded within technical systems, shaping what is visible, connectable, and governable, and scholarly metadata infrastructures are no exception. The question is no longer simply whether metadata is accurate but who controls it, on what terms, and to whose benefit.

This article addresses that question through two arguments that are structurally connected. The first concerns labour. The enriched, machine-readable metadata that now underpins AI tools in scholarly publishing is not produced by publishers alone: It is generated through what can be called semantic labour, understood as the cognitive, administrative, and technical work of encoding research outputs in structured, machine-readable form. Researchers select controlled keywords, enter persistent identifiers, link datasets, and structure abstracts for machine parsing. Librarians normalise affiliations, disambiguate author names, and curate repository records. This work is typically unremunerated by its commercial beneficiaries, uncredited in formal reward structures, and rarely acknowledged in workload models. Understood in these terms, semantic labour forms part of the broader field of digital work analysed by Ursula Huws (2014) and Tiziana Terranova (2000), in which value is extracted from activities that remain marginal to formal systems of compensation and recognition.

The second argument concerns epistemic power. When enriched metadata is captured within proprietary knowledge graphs, enclosed by a small number of commercial publishers and analytics firms, the consequences extend beyond labour exploitation into the epistemic structure of scholarship itself. Coverage gaps and algorithmically encoded hierarchies embedded in proprietary indexes become the operative parameters of AI-generated knowledge. Miranda Fricker (2007) has established that epistemic harm can be structural—that is, embedded in the interpretive resources available to an epistemic community rather than in any particular act of exclusion. The dynamic examined here operates at a different register—infrastructural rather than intersubjective—but it produces analogous effects: Certain bodies of knowledge, certain research traditions, and certain scholarly communities become systematically less legible, not through deliberate exclusion but through the accumulated logic of proprietary classification. This is epistemic capture: the transfer of authority over what counts as legible knowledge from scholarly communities to proprietary infrastructural providers.

The relationship between these two arguments is the central claim of this article: Semantic labour is the mechanism through which epistemic capture operates. The work of metadata creation, externalised from the cost structures of those who profit from it, feeds the proprietary systems that then govern scholarly visibility. This process generates a recursive loop: Labour exploitation and epistemic dependency are not separate problems but a single integrated dynamic. Researchers and librarians produce the structured semantic layer that makes AI tools functional; those tools, built on enclosed versions of that layer, then reproduce and amplify coverage gaps and hierarchical biases embedded in the indexes that trained them. The semantic labour of knowledge production is thus not merely an economic grievance but an operative link between commercial enclosure and the narrowing of the epistemic conditions of scholarship.

Recently, Jefferson Pooley (2024) charted how oligopolistic publishers have extended surveillance publishing into a new phase of data extraction and platform monetisation, and Lai Ma (2024) identified the risks that generative AI poses to epistemic diversity in scholarly publishing. The present article does not dispute either diagnosis; rather, its contribution is in identifying metadata as the mechanism connecting commercial enclosure to epistemic harm, to show that it is precisely through the semantic layer, and through the appropriation of the labour that produces it, that platform capitalism restructures the epistemic conditions of scholarship. Where Pooley maps the political economy of the publishing oligopoly and Ma interrogates the epistemic implications of AI-generated content, this article focuses on the infrastructural substrate that makes both dynamics possible and on the specific form of labour that substrate appropriates.

This analysis proceeds as a theoretically grounded intervention, drawing on critical infrastructure studies, digital labour theory, and platform capitalism scholarship; empirical cases and examples are introduced as illustrative rather than as the basis of a systematic empirical study. Section 2 traces the historical enrichment of metadata and its integration into AI systems; section 3 examines the distributed semantic labour that produces metadata, tracing a six-stage cycle from research activity to governance feedback; section 4 analyses commercial enclosure and its epistemic consequences; and section 5 assesses open infrastructure alternatives and the governance conditions required for metadata to function as a genuine public resource, closing with the argument for metadata sovereignty.

2. From Bibliographic Record to AI Infrastructure

For most of the twentieth century, the metadata attached to scholarly publications was rather modest in both scope and ambition. The print-based record required only what was necessary for citation and retrieval: author names, article titles, journal names, volume and issue numbers, page ranges, and a handful of

subject headings assigned by specialist indexers. This skeletal descriptive layer served primarily human use, facilitating the production of printed indexes, bibliographies, and library catalogues, and its creation was centralised within library and publishing institutions.

The shift to digital databases from the late 1960s onwards marked the first step towards metadata as a computational input. Early online databases relied on structured metadata for keyword search and Boolean retrieval, but they still operated with relatively limited fields: title, author, abstract, and subject terms. The central innovation was speed and remote access; the fundamental scope of metadata remained constrained.

The launch of Crossref in 2000 was a more decisive turning point. By introducing the digital object identifier (DOI) system for scholarly publishing, Crossref normalised the principle that every article should carry a persistent, machine-resolvable identifier linked to a centralised metadata record. DOIs enabled direct linking of references, transforming citation metadata from a static list into a networked structure. Publishers now had an incentive to enrich metadata beyond the minimum since richer linking meant higher discoverability and citation counts.

From the 2010s onwards, metadata entered an enrichment era, driven by the convergence of several policy and technical developments. The introduction of persistent identifiers—like the ORCID system for authors, the Research Organization Registry (ROR) for institutions, and grant identifiers added by Crossref and DataCite—enabled reliable disambiguation of entities across platforms. Funding information became a standard metadata component, driven by open access mandates and funder compliance requirements. The Contributor Roles Taxonomy (CRediT) embedded granular contribution statements within article records. The FAIR (Findability, Accessibility, Interoperability, and Reuse) principles (Wilkinson et al. 2016) consolidated enriched, machine-readable metadata as a cornerstone of open science policy. At each stage, these developments responded to genuine needs for interoperability and discoverability; they also, collectively, constructed a semantic layer whose potential for commercial exploitation was not yet fully apparent.

In the age of AI, the value of this enriched layer derives precisely from its structured nature. While natural language processing can extract entities and relationships from full text, such processes are resource intensive and error prone. Enriched metadata functions as pre-processing at scale: Persistent identifiers resolve entity ambiguities that raw text cannot; controlled vocabularies enable concept disambiguation essential to topic modelling; explicit citation links create machine-readable edges in citation graphs; and funding metadata connects outputs to projects and institutions. For machine learning systems—whether scholarly recommendation engines, impact forecasting models, or large language models fine-tuned for academic domains—these structured metadata features are cleaner and more interoperable than full text, and their production has required decades of cumulative human labour that remains largely externalised from the cost structures of those who benefit commercially from its products. As Stacy Allison-Cassin and Dean Seeman (2022) argue, metadata should not be regarded as an ancillary record but as a form of knowledge in itself, shaping how research is produced, organised, and encountered. In the AI era, that claim has acquired a commercial dimension that the authors did not anticipate.

The AI architecture most prominently adopted in commercial scholarly publishing tools is retrieval-augmented generation (RAG), an approach that has attracted significant attention in library and information science for its potential to transform academic search and retrieval (Bevara et al. 2025). Unlike language models that rely solely on patterns encoded during training, RAG systems combine a generative model with an external retrieval layer: When a user submits a query, the system first searches a curated corpus for relevant documents and then feeds excerpted versions of those documents into the language model to guide its response. In the scholarly domain, the retrieval corpus is typically a proprietary index such as Scopus, Web of Science, or Dimensions, and the quality of the generative output depends directly on the richness and precision of the metadata describing each document in that corpus. Metadata enables structured filtering before retrieval by, for example, document type, publication date, funding agency, and subject classification. Metadata also determines how documents are ranked before being passed to the generative model, and it links publications to associated datasets, grant records, and related works, enabling richer generative contexts. The consequence is that metadata is not merely a descriptive layer in RAG architectures: It has become the operative mechanism that sets the boundaries of what the system can retrieve and therefore of what it can generate.

Publishers have been quick to capitalise on this dependency: Elsevier's [Scopus with AI](#) retrieves from the Scopus index, using controlled vocabularies and disambiguated author profiles to surface high-relevance records for generative processing, and Digital Science's [Dimensions AI Assistant](#) draws from a knowledge graph linking publications, grants, and patents. All these connections are made possible through persistent identifiers and curated funding metadata. Clarivate's [Web of Science](#) applies citation-derived rankings and journal classifications before passing retrieved content to its generative model. Beyond traditional

publishing, platforms such as Consensus, ResearchRabbit, Elicit, and Perplexity deploy RAG architectures combining public or licensed corpora with metadata-based filtering to produce AI-generated scholarly answers. In each case, the retrieval stage is metadata-dependent, and the proprietary index that supplies that metadata sets the epistemic frame within which AI outputs are constructed.

A particularly instructive illustration of how metadata enrichment is transformed into commercial product is Elsevier's [Fingerprint Engine](#). The system applies natural language processing to publication abstracts, grant descriptions, and other scientific texts, mapping them to thesauri covering more than five hundred thousand concepts across disciplines. These semantic profiles are then aggregated into researcher, institutional, and regional "fingerprints" embedded in commercial platforms such as [SciVal](#) (Elsevier), [Pure](#) (Clarivate), or [Expert Lookup](#) (Elsevier too but recently discontinued). What begins as metadata enrichment—much of it produced through the labour of researchers and librarians—becomes a marketable knowledge graph that institutions must license (again) in order to analyse their own research outputs. The Fingerprint Engine thus demonstrates, in miniature, the dynamic that this article examines at scale: Enriched metadata, produced through scholarly labour, is transformed into proprietary infrastructure and sold back to the community that produced it.

3. Semantic Labour and the Metadata Production Cycle

The enriched metadata that now underpins AI-driven scholarly publishing is the product of a complex, multi-actor, and multi-stage process. Understanding this process as a cycle of semantic labour, moving from the initial production of research through successive layers of encoding, infrastructural ingestion, commercial capture, and governance feedback, is essential to grasping both the economic and epistemic stakes of its enclosure. Each stage in this cycle demands human work, much of it unremunerated by its commercial beneficiaries or unrecognised in academic reward structures and all of it indispensable to the computational apparatus of contemporary scholarly communication. This pattern of labour extraction resonates with Terranova's (2000) analysis of free labour in digital economies, where value is systematically produced through activities that fall outside formal compensation structures, and with Huws's (2014) broader account of digital labour as a site of invisible value extraction in which the products of unremunerated work are appropriated within commercial information systems.

Stage 1. Research Activity

The cycle begins with research itself: the conception of a question, the design of an investigation, the execution of experiments or analyses, and the writing of results. In the digital scholarly ecosystem, the outputs of this stage—articles, datasets, software, conference proceedings—are not only ends in themselves but containers for metadata. Where traditional print publishing assigned most metadata retrospectively through publisher and librarian intervention, researchers are now expected to anticipate machine-readable use cases from the outset by, for example, crafting structured abstracts, selecting keywords from controlled vocabularies, standardising author affiliations to match persistent identifiers, and linking datasets before submission. The metadata work begins before publication, embedded in the research process itself.

Stage 2. Semantic Encoding

This stage in the metadata production cycle is the core of what is here called semantic labour: the intellectual, administrative, and technical work of encoding research outputs in structured, machine-readable form. It includes, but is not limited to, titles and abstracts that balance human readability with algorithmic discoverability; keywords drawn from controlled vocabularies or disciplinary thesauri; author identifiers such as ORCIDs, ensuring disambiguation across platforms; funding information linked to funder IDs and grant numbers; persistent identifiers for the work itself and for linked datasets and references; and subject classifications such as Fields of Research codes, UNESCO descriptors, or journal-specific taxonomies. Much of this work is performed by researchers during manuscript submission and funder reporting. But librarians, data stewards, and repository managers are equally essential contributors: They normalise names, verify affiliations, ensure correct classifications, and link outputs to datasets and related records. As Amanda Belantara and Emily Drabinski (2022) demonstrate, cataloguers do not merely record technical details but interpret and narrate scholarly work, embedding intellectual judgements within seemingly administrative procedures.

Stage 3. Infrastructural Ingestion

Once generated, metadata enters the technical infrastructures that manage scholarly information: institutional current research information systems (CRIS), university repositories, library catalogues, and subject-based repositories such as arXiv or PubMed Central. At this stage, metadata is cleaned, normalised,

and further enriched through both manual curation and automated scripts. It is then shared with external aggregators and registries like Crossref, DataCite, OpenAIRE, or other disciplinary databases from which it can be harvested by a range of downstream platforms, both open and commercial. This is the point at which data pipelines begin and at which the trajectory from scholarly labour to commercial asset becomes technically feasible.

Stage 4. Platform Capture

Here the dynamic shifts from scholarly to commercial control. Metadata flows into the proprietary indexes and knowledge graphs of large publishing and analytics firms like Elsevier's Scopus, Clarivate's Web of Science, and Digital Science's Dimensions. Once captured, it is embedded in subscription-based dashboards and AI products. Algorithmic classification, topic modelling, and influence scoring are applied: Fields of research may be redefined according to platform-specific taxonomies, citation data is normalised for disciplinary difference, and metrics are calculated. At this point, metadata ceases to function merely as a description of research and becomes an input to evaluative and predictive algorithms that can shape statuses, careers, and institutional strategies.

Stage 5. Epistemic Automation

With the metadata structured, linked, and enriched, it becomes the substrate for a range of analytics like citation counts, field-normalised impact scores, co-authorship network analyses, trend forecasts, and, increasingly, for AI-driven tools like recommender systems, grant-matching services, and RAG platforms such as Scopus with AI or Dimensions AI Assistant. In these AI contexts, metadata serves simultaneously as retrieval index and relational context, guiding both the selection of documents and the frame within which the generative model produces its output. The richer and more standardised the metadata, the more precise the AI output and, hence, the more extensive the epistemic authority exercised by the proprietary metadata infrastructure that feeds it.

Stage 6. Governance Feedback

The final stage of the cycle is governance. Metadata-derived indicators feed into institutional rankings, departmental evaluations, and national research assessment exercises; funders use these indicators to assess the return on their investments, and universities use them to allocate resources and set hiring priorities. This evaluative use feeds back into researcher and librarian behaviour: Knowing that certain keywords, affiliations, or publication venues boost visibility in key indexes, scholars adapt their semantic labour accordingly, optimising not only for disciplinary peers but for algorithmic legibility. The cycle is therefore recursive: Metadata not only describes research but shapes what kind of research is pursued, how it is framed, and where it is published. Semantic labour is simultaneously a technical and a strategic act, constituted by the very infrastructures it sustains.

Taken together, these stages reveal the full scope of what is at stake in the governance of scholarly metadata. The labour involved is substantial, and its economic value, when made visible, is considerable. Yet this labour is structurally invisible: absent from workload models, unrewarded in promotion criteria, and rarely acknowledged in the contractual arrangements through which metadata flows into commercial platforms. The result is a structural asymmetry between the scholarly community that produces the semantic layer and the commercial actors that enclose and monetise it. This asymmetry follows what Armin Beverungen, Steffen Böhm, and Christopher Land (2012) identify as a double appropriation in academic publishing: The productive labour of the scholarly community is first extracted without remuneration from its commercial beneficiaries, then its products are sold back to the institutions that bore the original costs.

What distinguishes the account developed in this article from Pooley's (2024) mapping of surveillance publishing is its focus on the production end of the cycle: The argument is not only that metadata is appropriated once it enters commercial platforms but that the act of semantic encoding—understood as the work of making research machine-readable—is itself the founding moment of the political economy diagnosed here.

4. Commercial Enclosure and Epistemic Capture

The point at which enriched metadata leaves scholarly and institutional systems to enter commercial platforms marks a decisive shift in its economic and epistemic status. What began as a set of structured descriptions to aid discovery and interoperability becomes, at this stage, a proprietary asset controlled, packaged, and sold under terms set by a small number of firms. This section analyses that transformation in two movements: first, the mechanisms by which commercial enclosure operates; second, and more centrally, the epistemic consequences it produces when enclosed metadata infrastructures become the operative substrate of AI-generated scholarly knowledge.

Once metadata enters open infrastructure registries such as Crossref or DataCite, it is available for harvest by both non-commercial and commercial platforms. This dynamic reflects the broader logic of platform capitalism, in which infrastructures designed for interoperability are systematically repurposed as engines of data extraction and value generation (Srnicsek 2017). Major publishers pull metadata into proprietary indexes and knowledge graphs, integrating it with algorithmically derived enhancements such as disambiguated author profiles, subject classifications, and citation normalisation, and then restrict access through subscription tiers, paid API agreements, and licensing arrangements. As Lucy Montgomery et al. (2021) argue, even open scholarly infrastructures are sites of political contestation: Metadata may remain enclosed within proprietary value-added layers even when the content it describes is openly accessible. The justification for enclosure is commercial protection of “value-added” processing, but the raw material for that processing is overwhelmingly produced outside the corporate payroll, through the semantic labour examined in the preceding section.

The revenue potential of these proprietary knowledge graphs has expanded dramatically in the AI era. Publishers now market their metadata holdings not only through analytics dashboards but as licensable training corpora for machine learning systems, and several developments in 2024 illustrate this trend. Taylor & Francis announced a \$10 million agreement granting Microsoft access to specialist content for AI model training; Wiley confirmed agreements with major technology firms supplying scholarly content and metadata for AI purposes; and Elsevier, rather than licensing outward, integrated Scopus metadata into proprietary AI products and new subscription tiers (Pooley 2024). High-quality structured metadata is a key differentiator in AI performance: It reduces pre-processing costs, improves retrieval accuracy, and increases confidence in entity disambiguation. For companies training domain-specific models, it is frequently the most commercially significant component of the licensed package.

The enclosure of metadata in the AI era operates, moreover, on two levels simultaneously. The metadata itself functions as a licensable asset, used for retrieval, training, and knowledge graph construction. But the behavioural data generated when users interact with AI tools via queries, clicks, dwell time, and selections is also harvested and fed back into the knowledge graph to refine relevance models, improve recommendation engines, and identify marketable user segments (Zuboff 2019). Following Pooley’s (2022) framing of surveillance publishing, this dual commodification transforms past and present research activity into data assets that support predictive analytics. The outputs are sold back to universities and funders as strategic intelligence, further embedding commercial platforms within the governance of research. The irony here is structural: Metadata labour performed without remuneration from its commercial beneficiaries, in compliance with funder mandates or institutional requirements, generates the training signal that enables commercial AI systems to predict and monetise future scholarly behaviour.

The epistemic consequences of this enclosure become apparent when enclosed metadata infrastructures function as the retrieval substrate of AI tools. What counts as visible, retrievable, or generatable within AI-assisted scholarly knowledge is increasingly determined by the classifications, coverage decisions, and ranking mechanisms embedded in commercial knowledge graphs, and these mechanisms carry structural biases that are operationalised at scale.

Coverage asymmetries are well documented. Commercial indexes privilege English-language, STEM-intensive, and Global North outputs, marginalising humanities, social sciences, and regionally produced research. The consequences for non-Anglophone scholarship are particularly acute: Systematic undercounting in major indexes translates directly into systematic underrepresentation in AI-generated outputs, since the boundaries of what any RAG system can retrieve are set by the coverage of its retrieval corpus (Céspedes et al. 2025). A literature search conducted through Scopus AI or Dimensions AI Assistant therefore reflects not the global body of scholarship but the contours of a commercially curated index, an effect that is structurally invisible to most users, who encounter the output of these tools as seemingly comprehensive knowledge.

Bias is further compounded through ranking mechanisms. Algorithms that weigh citation counts or journal prestige, often operationalised through metrics such as the journal impact factor, encode hierarchies of visibility into the retrieval process, producing recursive loops in which highly ranked work is surfaced more often, read more widely, and cited more frequently. This dynamic has long been identified in the sociology of science as the Matthew Effect: the tendency of prior advantage to generate cumulative advantage, independent of the intrinsic merit of individual contributions (Merton 1968). In metadata-dependent AI systems, the Matthew Effect is operationalised infrastructurally rather than socially, as it is not peer recognition but algorithmic ranking that amplifies existing hierarchies, and it does so at a scale and velocity that makes the amplification largely invisible (Merton 1988). Research that falls outside dominant metadata parameters, is published in regional journals, is written in languages underrepresented in major indexes, or is produced in disciplinary traditions excluded from commercial classification schemes risks a form of visibility

loss generated not by the quality or relevance of the work but by the infrastructural conditions of its legibility (Arboledas-Lérida 2024). These consequences extend directly to research assessment: Metadata-derived analytics, embedded in dashboards such as InCites, SciVal, and Dimensions Analytics,² shape hiring decisions, resource allocation, and strategic planning, encoding classification biases and algorithmic weightings as though they were objective measures of scholarly value.

At the micro level, the fragility of metadata quality compounds these structural asymmetries. Lonni Besançon, Guillaume Cabanac, Cyril Labbé, and Alexander Magazinov (2024) have shown that fabricated references and distorted metadata propagate through commercial databases, artificially inflating citation counts and corrupting bibliometric indicators. When such data underpins AI-driven retrieval and generation, the distortions are reproduced at scale and presented with the apparent authority of automated systems. The problem is not exceptional but endemic to infrastructures in which commercial incentives and metadata quality guarantees are only partially aligned.

This process constitutes epistemic capture: the transfer of authority over what counts as legible knowledge from scholarly communities to proprietary infrastructural providers. Fricker (2007) has established that epistemic harm is not always attributable to deliberate exclusion: It can be structural, embedded in the recognitional resources and interpretative frameworks available within an epistemic community. The dynamic examined here operates at a different register from the intersubjective harms Fricker analyses: It is infrastructural rather than interpersonal, and it is produced through market logic rather than individual credibility judgements. But it generates structurally analogous effects: Certain research traditions, languages, and scholarly communities become systematically less legible, not through deliberate censorship but through the accumulated logic of proprietary classification and algorithmic ranking. Amandine Catala (2024) has shown how epistemic oppression can be reproduced through knowledge infrastructures that present themselves as neutral; the argument developed here identifies the semantic layer as the specific site where that reproduction occurs, at the intersection of commercial enclosure and AI deployment.

Epistemic capture is not a problem parallel to the labour problem examined in the preceding section: The two are the same mechanism viewed from different positions in the cycle. The work of metadata production, externalised from the cost structures of those who profit from it, feeds the proprietary systems that govern scholarly visibility, which in turn shape the conditions under which future semantic labour is performed.

5. Towards Metadata Sovereignty: Open Infrastructures and Governance

If proprietary metadata holdings function as instruments of commercial enclosure and epistemic capture, open infrastructures represent a potential—but structurally constrained—counterweight. Initiatives such as Crossref, OpenAlex, Wikidata, OpenCitations, and OpenAIRE construct metadata pipelines that are openly licensed and, in principle, community governed. They embody the proposition that metadata, as the infrastructural condition of scholarly visibility, should be treated as a public good. The case for these alternatives is not merely normative: OpenAlex now indexes over 250 million scholarly works under a CCO licence, making its full metadata freely available for download and integration into AI systems; OpenCitations provides open citation data that commercial publishers had previously restricted; and OpenAIRE links European research outputs, funding records, and repositories within a single interoperable graph. These initiatives demonstrate, in practice, that enriched, structured metadata does not inherently require proprietary ownership to exist at scale.

Yet openness alone does not guarantee equity or sustainability. As Paolo Manghi (2024) documents in the case of the OpenAIRE Graph, even open scholarly knowledge graphs face significant challenges with metadata quality, heterogeneity, and governance. Open infrastructures harvest metadata from existing publication workflows and standards, which means that the coverage asymmetries and linguistic biases of the broader scholarly communication system are partially reproduced within them. A knowledge graph that is openly licensed but predominantly populated with English-language, Global North research does not resolve epistemic capture but relocates it within a more accessible infrastructure. The linguistic coverage of OpenAlex, while substantially broader than that of Scopus or Web of Science, remains uneven, with significant gaps in non-Anglophone and regionally published scholarship (Céspedes et al. 2025). As Geoffrey Bowker and Susan Leigh Star (1999) remind us, infrastructures always embody simultaneous inclusion and exclusion: What is standardised and recorded for some purposes erases other forms of knowledge. Open metadata infrastructures are not perfectly equitable; the question is whether they provide governance

² The names and features of these products are changing frequently. For further information, their commercial websites must be checked.

conditions under which inequities can be identified, contested, and progressively corrected. Proprietary systems structurally foreclose that possibility; open, community-governed systems do not.

Open infrastructures face at least three persistent structural challenges. The first is sustainability: Unlike commercial platforms, community-governed projects depend on unstable combinations of grants, institutional subscriptions, and volunteer labour, leaving them vulnerable to underfunding and, in some cases, to the commercial co-optation they were designed to resist. The second is governance: Decisions about schema design, classification systems, multilingual support, and quality standards are never neutral, and the communities currently represented in governance structures do not yet reflect the global diversity of scholarly production. The third is epistemic scope: Even the most inclusive open graphs depend on publication workflows and metadata standards developed primarily within Anglophone institutional contexts, which structurally disadvantage locally produced, non-standardised, or informally circulated scholarship.

These challenges do not disqualify open infrastructures as the appropriate institutional form for scholarly metadata. They identify what governance frameworks must address if open alternatives are to constitute a genuine counterweight to proprietary enclosure rather than merely a more accessible version of the same asymmetries. Several principles follow directly from the analysis developed in this article. First, metadata generated through publicly funded research should be required to remain openly licensed and auditable, a condition that could be embedded in funder mandates and institutional open access policies in most jurisdictions without new legislation. Second, open infrastructure projects require sustainable, publicly funded core support, insulated from the commercial pressures that have historically driven consolidation in scholarly publishing. Third, classification systems and controlled vocabularies embedded in open metadata infrastructure should be subject to transparent, pluralistic, and internationally representative governance, with specific provisions for multilingual and non-Anglophone scholarship. Fourth, where commercial platforms make use of openly licensed metadata—as they routinely do—collective licensing frameworks could redistribute a portion of the value generated back into the open infrastructure ecosystem, creating a modest structural counterweight to the current asymmetry of benefit. Fifth, and following directly from the analysis of semantic labour developed in section 3: The work of metadata production must be recognised within institutional workload models, academic credit frameworks, and, where possible, the contractual arrangements governing how that labour flows into commercial systems. Governance of metadata as infrastructure and recognition of the labour that sustains it are not separable demands because they address two dimensions of a single structural problem, and proposals that address one without the other will remain incomplete.

These proposals follow from the logic of the problem: If epistemic capture operates through the enclosure of community-produced semantic labour within proprietary systems, then governance must address the conditions of both labour and enclosure as dimensions of the same structural problem.

Metadata Sovereignty

The concept organising these governance principles is metadata sovereignty: the capacity of scholarly communities to govern the structures that define their own epistemic legibility. The term is deliberately chosen to echo, without directly importing, the discourse of data sovereignty in knowledge governance contexts and in debates about national data infrastructures, fields in which the question of who governs the systems that make communities legible has been more extensively theorised than in scholarly communication (Bowker and Star 1999; Montgomery et al. 2021). In the scholarly context, metadata sovereignty does not require the elimination of commercial metadata services, nor does it demand that all metadata be produced and maintained by public institutions. It requires that the scholarly community retain meaningful collective authority over the standards, classifications, coverage criteria, and licensing conditions that determine what research is made visible, retrievable, and actionable within AI-mediated scholarly systems.

But metadata sovereignty will not be achieved through individual institutional decisions or through the voluntary adoption of open licences by publishers whose business models depend on enclosure as a revenue strategy. It requires coordinated policy intervention at the level of national research funders, international scholarly bodies, and, where appropriate, regulatory frameworks governing platform markets. The scholarly communication system is not alone in confronting this problem: Debates about algorithmic accountability, AI training data governance, and the political economy of data infrastructure are increasingly prominent across multiple sectors (Zuboff 2019). What distinguishes the scholarly case is the specific mechanism identified in this article: The labour of knowledge production itself—the cognitive and administrative work of encoding research in machine-readable form—is the raw material that proprietary AI systems appropriate and enclose. Addressing this asymmetry requires not only governance of metadata as infrastructure but recognition of semantic labour as a form of scholarly work with legitimate claims on the systems it sustains.

As AI tools become routine elements of research, the stakes of metadata governance will only intensify. The infrastructures of scholarly visibility are being redesigned, and the terms of that redesign—who controls the semantic layer, whose knowledge is retrievable, what counts as scholarship within an AI-mediated epistemic economy—will shape the conditions of knowledge production for years to come. If metadata is the substrate of epistemic automation, then its governance is a matter of scholarly sovereignty. Without coordinated intervention to recognise and protect metadata as public infrastructure, the epistemic commons of scholarship risks sustained enclosure by a small number of commercial actors, with the scholarly community cast indefinitely in the role of unremunerated provider of the semantic labour that makes that enclosure possible.

Competing Interests

The author declares that she has no competing interests.

Use of Generative AI

To write this article, the author used DeepL for Spanish–English translation and ChatGPT 5 to improve the English writing.

References

- Allison-Cassin, Stacy, and Dean Seeman. 2022. "Metadata as Knowledge." *KULA: Knowledge Creation, Dissemination, and Preservation Studies* 6 (3): 1–4. <https://doi.org/10.18357/kula.244>.
- Arboledas-Lérida, Luis. 2024. "A Marxist Analysis of the Metrification of Academic Labour: Research Impact Metrics and Socially Necessary Labour Time." *Science & Society* 88 (4): 573–600. <https://doi.org/10.1521/asiso.2024.88.4.573>.
- Belantara, Amanda, and Emily Drabinski. 2022. "Working Knowledge: Cataloguers and the Stories They Tell." *KULA: Knowledge Creation, Dissemination, and Preservation Studies* 6 (3): 1–10. <https://doi.org/10.18357/kula.233>.
- Besançon, Lonni, Guillaume Cabanac, Cyril Labbé, and Alexander Magazinov. 2024. "Sneaked References: Fabricated Reference Metadata Distort Citation Counts." *Journal of the Association for Information Science and Technology* 75 (12): 1368–1379. <https://doi.org/10.1002/asi.24896>.
- Bevara, Ravi Varma Kumar, Brady D. Lund, Nishith Reddy Mannuru, Sai Pranathi Karedla, Yara Mohammed, Sai Tulasi Kolapudi, and Aashrith Mannuru. 2025. "Prospects of Retrieval Augmented Generation (RAG) for Academic Library Search and Retrieval." *Information Technology and Libraries* 44 (2): 1–15. <https://doi.org/10.5860/ital.v44i2.17361>.
- Beverungen, Armin, Steffen Böhm, and Christopher Land. 2012. "The Poverty of Journal Publishing." *Organization* 19 (6): 929–938. <https://doi.org/10.1177/1350508412448858>.
- Bowker, Geoffrey C., and Susan Leigh Star. 1999. *Sorting Things Out: Classification and Its Consequences*. The MIT Press.
- Catala, Amandine. 2024. "Epistemic Injustice or Epistemic Oppression?" *KULA: Knowledge Creation, Dissemination, and Preservation Studies* 7 (1): 1–11. <https://doi.org/10.18357/kula.294>.
- Céspedes, Lucía, Diego Kozłowski, Carolina Pradier, Maxime Holmberg Sainte-Marie, Natsumi Solange Shokida, Pierre Benz, Constance Poitras, Anton Boudreau Ninkov, Saeideh Ebrahimi, Philips Ayeni, Sarra Filali, Bing Li, and Vincent Larivière. 2025. "Evaluating the Linguistic Coverage of OpenAlex: An Assessment of Metadata Accuracy and Completeness." *Journal of the Association for Information Science and Technology* 76 (6): 884–895. <https://doi.org/10.1002/asi.24979>.
- DOAJ (Directory of Open Access Journals). 2026. "DOAJ's New Premium Metadata Services." *DOAJ Blog*, March 3, 2026. <https://blog.doaj.org/2026/03/03/doajs-new-premium-metadata-services/>.
- Fricker, Miranda. 2007. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford University Press.
- Huws, Ursula. 2014. *Labor in the Global Digital Economy: The Cybertariat Comes of Age*. Monthly Review Press.
- Ma, Lai. 2024. "Generative AI for Academic Publishing? Some Thoughts About Epistemic Diversity and the Pursuit of Truth." *KULA: Knowledge Creation, Dissemination, and Preservation Studies* 7 (1): 1–5. <https://doi.org/10.18357/kula.287>.
- Manghi, Paolo. 2024. "Challenges in Building Scholarly Knowledge Graphs for Research Assessment in Open Science." *Quantitative Science Studies* 5 (4): 991–1021. https://doi.org/10.1162/qss_a_00322.
- Mayernik, Matthew S. 2019. "Metadata Accounts: Achieving Data and Evidence in Scientific Research." *Social Studies of Science* 49 (5): 732757. <https://doi.org/10.1177/0306312719863494>.

- Mayernik, Matthew S. 2020. "Metadata." *Knowledge Organization* 47 (8): 696–713. <https://doi.org/10.5771/0943-7444-2020-8-96>.
- Merton, Robert K. 1968. "The Matthew Effect in Science: The Reward and Communication Systems of Science are Considered." *Science* 159 (3810): 56–63. <https://doi.org/10.1126/science.159.3810.56>.
- Merton, Robert K. 1988. "The Matthew Effect in Science, II: Cumulative Advantage and the Symbolism of Intellectual Property." *Isis* 79 (4): 606–23. <https://doi.org/10.1086/354848>.
- Montgomery, Lucy, John Hartley, Cameron Neylon, Malcolm Gillies, Eve Gray, Carsten Herrmann-Pillath, Chun-Kai (Karl) Huang, Joan Leach, Jason Potts, Xiang Ren, Katherine Skinner, Cassidy R. Sugimoto, and Katie Wilson. 2021. *Open Knowledge Institutions: Reinventing Universities*. The MIT Press. <https://doi.org/10.7551/mitpress/13614.001.0001>.
- Parry, Kyle. 2023. "Metadata is Not Data About Data." In *Decolonizing Data: Algorithms and Society*, edited by Michael Filimowicz. Routledge.
- Pooley, Jefferson. 2022. "Surveillance Publishing." *The Journal of Electronic Publishing* 25 (1). <https://doi.org/10.3998/jep.1874>.
- Pooley, Jefferson. 2024. "Large Language Publishing: The Scholarly Publishing Oligopoly's Bet on AI." *KULA: Knowledge Creation, Dissemination, and Preservation Studies* 7 (1): 1–11. <https://doi.org/10.18357/kula.291>.
- Srnicek, Nick. 2017. *Platform Capitalism*. Polity Press.
- Terranova, Tiziana. 2000. "Free Labor: Producing Culture for the Digital Economy." *Social Text* 18 (2): 33–58.
- Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3: 160018. <https://doi.org/10.1038/sdata.2016.18>.
- Zuboff, Shoshana. 2019. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. PublicAffairs.

How to cite this article: Battaner Moro, Elena. 2026. Semantic Labour and the Politics of Metadata: Scholarly Infrastructure in the AI Driven Academy. *KULA: Knowledge Creation, Dissemination, and Preservation Studies* 9(2). <https://doi.org/10.18357/kula.330>

Submitted: 11 September 2025 **Accepted:** 15 May 2026 **Published:** 2 October 2026

Copyright: © 2026 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

KULA: Knowledge Creation, Dissemination, and Preservation Studies is a peer-reviewed open access journal published by University of Victoria Libraries.

OPEN ACCESS 