# The Indexing Mandate: Infrastructure as Legitimacy

## The Infrastructure of Legitimacy: A Critical Re-evaluation of Indexing Protocols and Metadata Stewardship in the Diamond Open Access Ecosystem

zeba.academy

Zeba Academy

**First Published by Zeba Academy and Zeba Books.**

**Publication Year:** 2026

**Document Series:** Zeba Academy Blueprints -- Sovereign Systems Technical Directives

**Purpose:** This series of blueprint directives is authored to combat the "enshittification" and unnecessary bloat of modern software. Our goal is to reclaim sovereign control over our systems by bridging the gap between deep academic theory and high-stakes industrial implementation. We believe that software should be fast, permanent, and most importantly, understandable to the person who owns and uses it.

**Principal Architect: Sufyan bin Uzayr,** Google Cloud-Certified Professional DevOps Engineer.

**Core Stack:** Linux, Rust, Zig, C++, Flutter, and PHP.



**Email:** hello@zeba.academy

# The Indexing Mandate: Infrastructure as Legitimacy

## *The Infrastructure of Legitimacy: A Critical Re-evaluation of Indexing Protocols and Metadata Stewardship in the Diamond Open Access Ecosystem*

## Introduction

Indexing is frequently misunderstood as a sign of prestige and achievement, or as a sign that a journal has "made it." However, this misunderstanding ignores the fact that indexing is the circulatory system that determines whether knowledge flows through the global research body or remains trapped on an island of invisibility. The gap between Diamond Open Access (OA) journals, which do not charge author or reader fees, is existential. Because they work outside the commercial sector and rely on improvised technology stacks, their study does not qualify as "technical citizenship" in the global research community.

If we understand the purpose of indexing as infrastructure, we can move our focus from status and achievement to movement. Like the electrical grid, the system of indexing aims to facilitate the transfer of citations. Without such systems of movement, research is trapped in a vacuum, inaccessible to the search engines and aggregators that have become the usual tools of discovery.

The challenge for Diamond OA is interoperability: the ability of the research to be discovered and linked within the fragmented digital ecosystem.[1]

This paradigm is based on the assumption that survival in Diamond OA is dependent on investment in infrastructure and not prestige-based behavior. There is a need to alter our journals from static PDF-based research artifacts to dynamic research nodes that can be understood by machines. The argument is based on four important technical junctures:

- **Liquidation of 'Metadata Debt':** Closing technical gaps, such as the absence of consistency in XML, which renders journals invisible to discovery tools.
- **Technical Compliance as a Beginning Point:** Using structural best practices, such as the DOAJ, as a model for institutional quality.
- **The Active Network of Metadata:** Using Crossref's relational strength to ensure research is networked rather than simply named.
- **Strategic Alignment:** Entails coordinating local initiatives with global platforms and specialized niche registries.

However, reframing our understanding of the function of indexing in this manner is more than just a technical exercise; it is, in fact, a fundamental shift in how we understand indexing as a start, as opposed to a finish, in our academic lives.

[1] Marching with Mashup: Application of Information Mashup for Developing Open Library System - https://www.academia.edu/27006477/Marching_with_Mashup_Application_of_Information_Mashup_for_Developing_Open_Library_System - Accessed: 19 March 2026

# The Metadata Debt: Why Niche Journals Remain Invisible

The transition from static PDF hosting to digital interoperability is frequently hampered by a backlog of technical errors. This backlog not only causes administrative issues, but it also fundamentally separates the journal from the global citation graph.

## Defining Metadata Debt

"Metadata debt" is an expansion of the notion of "technical debt" in software development. It refers to the long-term repercussions of choices made for short-term convenience and ease of implementation. In the context of Diamond OA, metadata debt refers to the cumulative effect of errors, omissions, and non-standard data standards that render journal content inaccessible to discovery engines.

There are four main ways that metadata debt occurs:

- **Identification gaps:** The lack of ORCID IDs makes it impossible to distinguish between two researchers with the same name.
- **Name Inconsistency:** Differences in author names and institutions between journal issues prevent them from being merged into a single publication history.[2]

[2] ORCID iD gives visibility to research outputs - https://www.uzh.ch/blog/ub/2023/11/15/orcid-id-gives-visibility-to-research-outputs/?lang=en - Accessed: 19 March 2026

- **The PDF Dead End:** The absence of machine-readable PDFs and associated JATS XML and metadata headers prohibits search engines from "understanding" links between citations, publication dates, and funding data.

## Structural Causes

This debt is seldom attributable to careless editing but rather to the resource-constrained world in which Diamond OA functions. As there is no technical assistance and financial commitment, volunteers are used, and they may not necessarily be knowledgeable about data standards.

The primary reason is the extensive usage of "fragmented and manual" systems. When articles are submitted via email and data is manually entered into a website, there is a significant margin for error. "Quick fixes" allow a journal to publish today but leave it with a massive debt that must be "repaid" with compound interest when the publication eventually wants to apply to huge indexes like Scopus or Web of Science.

## Technical Causes

The lack of a data framework provides the technical basis for the concept of metadata debt. The journals are, in fact, a digital filing system. This is best exemplified by the lack of JATS XML, which makes it difficult for search engines to differentiate between abstracts, references, and footnotes. Furthermore, the data lacks Persistent Identifiers, such as DOIs for articles, ORCIDs for

researchers, and RORs for institutional affiliations, preventing it from having a permanent location.

Instead, there is an overreliance on static PDFs and website-only hosting. A PDF can be read by humans, but it is essentially a 'dark' format for machines. Without the underlying metadata, the knowledge is trapped within the page and inaccessible to the bots that collect information for library catalogs and citation indexes.

## Implications of Metadata Debt

The first and most noticeable result is academic marginalization. If this type of metadata is absent or in a non-standard format, the visibility of the item in search engines such as Google Scholar or Dimensions is significantly compromised. This immediately compounds the problem of lack of presence on other indexing sites, such as Scopus or Web of Science, which require stringent metadata exports as a prerequisite.

From the author’s perspective, this translates into a scarcity of references. If an item does not lend itself easily to being linked or exported into a reference manager, it will not be referenced, regardless of its quality. This finally translates into a loss of institutional trust. The authors will be loath to contribute their most valuable work to a journal that does not guarantee its long-term visibility.

## The Compounding Effect

Metadata debt is a notion based on negative interest, meaning that its value increases over time. The cost of retroactively correcting thousands of pieces increases with the age of a journal. Older archives are effectively a "technical black hole" in the history of the past.

This forms a vicious circle. The new articles will most likely be no better than the old ones, adding only to the culture of technical simplicity. As visibility remains low, the magazine will receive fewer high-quality articles, reducing its prestige. This is known as the 'death spiral,' because it demonstrates that, in the digital age, a journal's existence is as much determined by its metadata as by its editorial board.

# The DOAJ Readiness Framework

DOAJ is not just a directory; it is the first step in the process of establishing the integrity of the Open Access publication process. To achieve a state of 'sovereign' status, the journal must be designed in a manner that meets the Best Practice guidelines set out by the DOAJ, which translates the editorial intention into a machine-readable format.[3]

Indexing is a technology gateway that links local hosting to global discovery. Publishers must be careful not to fall into the 'predatory' trap and should follow the transparency guidelines set out by the DOAJ, which include everything from

[3] COPE Code of Conduct - https://publicationethics.org/membership/code-of-conduct - Accessed: 19 March 2026

ethics statements set out by the Committee on Publication Ethics (COPE) to licensing statements. The metadata is then disseminated worldwide, building citation potential.

## Structural Requirements

### Editorial Transparency

To be open, a journal's content must be clearly licensed. In other words, the open license must be explicit. Creative Commons is the finest open license, according to the DOAJ. In this manner, the user of the content has complete control over what he or she can do with it without seeking permission. The license must be machine-readable from the article's metadata, such as HTML or PDF. In other words, the license must be clearly visible to search engines. It should be underlined that publishers must avoid any ambiguity, such as saying "All Rights Reserved" alongside "Open Access," as this creates legal "grey zones."

### Archiving and Preservation

To be effective, open access must be permanent. Long-term access plans must be implemented to prevent the "digital disappearance" of published works if the publisher goes out of business or the servers fail. The journal should explore employing a distributed archiving system, such as LOCKSS (Lots of Copies Keep Stuff Safe) or CLOCKSS, to build a 'black archive' of the journal. Depositing metadata and full-text articles in institutional repositories or national libraries is

also a 'fail-safe.' Without this, scholarship records will be lost, and years of research will become inaccessible due to 'link rot.'

### Persistent Identifiers

Persistent Identifiers (PIDs) are digital fingerprints utilized for interoperability.

- **DOIs (Digital Object Identifiers)** for publications ensure that even if the journal's URL changes, the article's link remains intact.[4]
- **ORCID** assigns a unique identifier to each author, preventing name duplication and ensuring that each researcher is properly credited for their work.
- **ROR (Research Organization Registry)** identifies institutions, making it easy to monitor where research originates.

These are primarily used for interoperability, enabling various databases, citation indices, and financing platforms to "talk" to one another. When these PIDs are included in the metadata, the research becomes part of a larger network rather than a separate silo.

## Technical Requirements

### JATS XML is the official source of validation.

To achieve technological independence, ACS must transition from the old "PDF-first" strategy to a new rule-based structure centered on the JATS XML

[4] The Identifier What is a DOI? - https://www.doi.org/the-identifier/what-is-a-doi/ - Accessed: 19 March 2026

format. A PDF is a "flat" file format, meaning there is no structure or relationships between data types, whereas JATS (Journal Article Tag Suite) provides the "semantic" foundation required for machine actionability. The approach involves employing JATS high-fidelity tags (one-to-one mapping) for:

- **Front Matter:** Explicitly defining `<article-meta>` including `<contrib-group>` for author roles and `<kwd-group>` for taxonomics.
- **Body Content:** Use structured data tags (hierarchical and paginated) to allow for more granular text mining and screen reader accessibility.
- **Back Matter:** Documenting citations using DOI cross-referencing so that citation indexing can link citations automatically. With the above-mentioned strict structuring and tagging, the article's metadata will be 100% accurate, reducing the'metadata debt' caused by manually collecting metadata from visual layouts.

## Implementing OAI-PMH Harvesting

Universal discoverability is expected to be made available through OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting), rather than requiring a request. In this case, sovereignty is defined as your capacity to set a base URL for your local repository that may be accessed by aggregators. This is accomplished by configuring your server to respond to six separate verbs (ListRecords, GetRecord, etc.) via Dublin Core or MarcXML schema.[5] As a result, by making OAI-PMH endpoints publicly available, the journal may automate the

[5] Distribution Settings - https://docs.pkp.sfu.ca/learning-ojs/3.4/en/settings-distribution - Accessed: 19 March 2026

handshake process with worldwide discovery services, ensuring that all new issues are indexed within hours of publication, with no user intervention.

## Choosing Infrastructure: OJS vs. Custom Middleware

The choice of infrastructure will ultimately determine the journal's long-term interoperability. The chosen infrastructure is Open Journal Systems (OJS) or Janeway, both of which provide "compliance as a service" solutions by automating the complexities of generating XML and exposing OAI-PMH.

Publishers will need to design or integrate middleware (e.g., a REST API) to connect their custom user experience with international metadata standards. The purpose of this solution is to prevent technical isolation, which occurs when a visually pleasing website fails to meet the strict API standards imposed by Scopus or Web of Science.

## Automating Metadata Pipelines

The ability to scale up is tied to moving from manual data input to automated ingestion pipelines. This can be accomplished through a combination of technical methods.

**XML Validation -** automation tools, such as Schematron or DTD (Document Type Definition),[6] can be used to perform an XML "well-formedness" check prior to deploying into production.

[6] Schematron - https://schematron.com/ - Accessed: 19 March 2026

**API Interconnectivity -** CrossRef APIs allow automatic DOI minting when an article is published, creating an instantaneous connection between the citation network and the journal. As a result, manual workflows are not compatible with high-end indexing because they cannot maintain zero error for persistent identifiers.

### Continuous Metadata Audit

Sovereignty is maintained as long as the Metadata Quality Control (MQC) process is continuous and proactive; however, this will not be the last formal check, but will consist of a continuous "loop" of checks.

**Validation -** Use an NCBI JATS Validator to verify that there is schema compliance (e.g., to migrate from JATS 1.2 to 1.3).

**Audit Trail -** Conduct periodic link checking (i.e., checking for "link rot") on DOIs, and ensure the persistence of ORCIDs and RORs throughout the entire catalogue.

In the digital scholarly community, metadata is the most valuable resource. The integrity of metadata determines the ability for research to be included within the broader network of scholarship or to be lost in the "dark web" of unindexed archives.

## The Crossref Ecosystem

The Crossref ecosystem consists of a distributed, high-concurrency metadata and citation network that will provide infrastructure sovereignty for publishers

who will transition from passive DOI minting to an active and relational deposit strategy to transform individual, isolated PDFs into interconnected scholarly objects, as a global machine-readable "map" or node of scholarly work.

To accomplish this, publishers must build a means to integrate high-concurrency connection information into their present process utilizing the Crossref REST API, which provides them with:

- Real-time monitoring of research integrity.

- Flow of funds,

- Citation graphs.

## Advanced Crossref Infrastructure

### Operationalizing the Citation Graph (Cited-by)

To operationalize the citation graph (cited-by), publishers must submit the cited-by reference list of their articles as structured XML information during the DOI registration process. This enables Crossref to do recursive lookups against the whole Crossref database.

**Technical Execution**: Include the `<citation_list>` block in your XML export file and make sure that each entry contains a DOI whenever possible.

**Result:** This would enable the journal's front end to query the Crossref API to retrieve and display dynamically "citing works" as a measure of real-time verified article impact, allowing publications to stop relying on annual static measures of effect, such as the journal impact factor.

## Crossmark: The API-Driven Status Layer

In a digital setting, editorial independence is founded on the ability to modify academic records after they have been published. Crossmark enables this with a persistent status layer.[7]

**Technical Execution:** Include the Crossmark snippet in HTML or PDF files, and register a metadata change when corrections and retractions occur.

**Result:** when a reader clicks the Crossmark link, the Crossref database is searched in real time to verify that readers are accessing the most recent version of the document and that the journal's integrity is not jeopardized by the circulation of outdated or discredited materials.

## Funder Registry (FundRef) Integration

To improve the ability to demand grant acknowledgement standardization, take chaotic author-supplied strings and map them to a controlled vocabulary of 30,000 unique funder IDs.

**Technical Execution:** Include a lookup field in the submission workflow that searches the FundRef API.

**Result:** As a result of tagging articles with funder IDs and grant numbers, journals can automate their reporting for Open Access mandates, resulting in a direct, machine-verifiable link between grant money and research outcomes.

---

[7] Post-publication discussions and corrections - https://publicationethics.org/news-opinion/post-publication-discussions-and-corrections - Accessed: 19 March 2026

### Metadata Enrichment and Discovery SEO

A Metadata-conditioned DOI is a "dark" Identifier. To enhance Discovery SEO, publishers should enrich their XML Files by providing:

**Abstracts and References:** By providing "surface area" for semantic search engines (i.e., Dimensions, Lens.org) for crawling;

**ROR/ORCID Mappings:** Disambiguating institutional and author affiliations at the source; and Moving from passive labeling to a high-fidelity metadata deposit (i.e., not only is your content hosted, but also integrated into the global discovery "stack").

## Similarity Check and Institutional Trust

### Operationalizing the Similarity Check API

Publishers must operationalize Similarity Check (enabled by iThenticate) as a mandatory gatekeeper in order to implement an integrated integrity workflow, rather than using it passively as a checking tool. The Similarity Check service differs from consumer-grade plagiarism detection software in that it has access to a unique database of over 90 million paywalled scholarly artifacts, as well as 'black' archives of scholarship from all over the world.

**"How":** To apply technical sovereignty, each journal publisher's submission mechanism (OJS/custom) must be directly integrated with the iThenticate API.

**Implementation:** Upon submission, all papers are automatically submitted to the iThenticate server to obtain the manuscript's granular similarity score.

Before being reviewed by an editorial board, all submissions will be cross-referenced to iThenticate's exclusive corpus of worldwide research with surgical accuracy.

## Hard-Coding of Editorial Integrity

Editorial integrity is not a rule; it is a technological technique that can be validated. The screening mechanism must be hard-coded into the editorial timetable in order to meet COPE (Committee on Publication Ethics) standards.[8]

**Detecting "Salami Slicing":** Editors should use the system to identify patterns of duplicated publication, which occurs when text is reused across many submissions. This protects the journal's reputation from having to retract articles after publication.

**Transparency as Data:** By including the screening status in the article's internal metadata, you transform "integrity" from a concept to a machine-verifiable audit trail. This ensures that only original, high-impact contributions are included in the permanent scholarly record.

## Leveraging Screening as Institutional Currency

In the current funding environment, institutional trust acts as a proxy for financial sustainability. Universities and funding value publications that demonstrate rigorous, automated screening methods.

[8] Plagiarism in a submitted manuscript - https://publicationethics.org/guidance/flowchart/plagiarism-submitted-manuscript - Accessed: 19 March 2026

**APC Whitelisting:** Research offices mostly utilize active plagiarism detection to determine who receives Article Processing Charge (APC) funding and who has access to the library.

**Strategic Signaling:** A journal employs similarity to show the world that it is a "safe" location to publish. Check and provide clear data on how many contributions are rejected for being non-original. This technological commitment is the major currency for current open-access publications seeking institutional sponsorship and high-impact submissions.

# Strategic Indexing: Aligning with Scopus, WoS, and Specialized Databases

## Transitioning from Validation to Citation Impact

By seeing DOAJ as more than just an end goal and instead as a standard base for tiered indexing, you can achieve indexing sovereignty. While DOAJ aims to validate the transparency of the Open Access process, accessing higher-tiered databases necessitates abandoning "compliance" in favor of "influence." This is accomplished by using DOAJ metadata as a springboard to create a journal that meets the increasingly severe quantitative/qualitative requirements of the world's main citation databases.

## Major Indexing Systems: The Technical Gateways

### Scopus: Optimizing for Big Data Ingestion

Elsevier's Scopus is a citation database with a huge volume. To be considered for inclusion,[9] journals must meet the conditions for operationalizing citation consistency and citation performance.

**Operationalizing Citation Consistency:** The publication schedule is "zero-variance" for at least two years.

**Citation Engineering:** The journal's metadata (from the JATS XML) must be in a format that allows Scopus crawlers to correctly correlate citations to their sources, increasing the journal's CiteScore and exposure to the Elsevier ecosystem.

### Web of Science (WoS) Editorial Rigor: Data Point

Clarivate's WoS is in the top tier of selective inclusion. The concept of "sovereignty" in the context of WoS is based on "editorial rigor."

When we convert to a multi-stage review procedure, we cross-check all of the expert reviewers against their own WoS/Publons citation records.

**Technical Audit:** WoS will check the peer-review path for clarity. To do so, we must maintain an immutable digital log of reviewer comments, modifications, and editorial team decisions that fulfill the Science Citation Index Expanded (SCIE) requirements.

---

[9] Scopus - https://www.elsevier.com/products/scopus - Accessed: 19 March 2026

## Regional and Specialized Databases

Scopus has a limited global reach; a pure global approach may result in insufficient semantic indexing depth. Sovereignty might thus be defined as going beyond generic metadata to allow individual engines to successfully employ their high-fidelity taxonomies.

**Operating Niche Indexing:** The 'how' needs the usage of domain-specific XML tagging. For example, in order for something to appear in PubMed, one must use PubMed Central (PMC) Tagging Guidelines rather than standard JATS with MeSH (Medical Subject Headings) explicitly coded into the metadata,[10] allowing the content to be discovered programmatically within the unique researcher workflows of that space.

**Regional Infrastructure as a Geopolitical Tool:** This infrastructure also serves as a vital tool to counterbalance traditional citation bias against research performed outside of the Western world. RTCS such as SciELO and AJOL will help alleviate citation bias by providing local Citation Bases.

- **Targeted Visibility -** To ensure that these infrastructures function properly, set their OAI-PMH Harvest Points to accommodate regionally generated metadata schemas. By doing so, all research conducted in a region should be visible and amplified throughout the surrounding geographic areas.

[10] PMC Tagging Guidelines: Elements - https://pmc.ncbi.nlm.nih.gov/tagging-guidelines/article/tags/ - Accessed: 19 March 2026

- **Community Data Loops -** These infrastructures provide a local Citation Base that creates a'sovereign' bibliometric record, which will eventually be recognized by bigger International Systems due to the cumulative volume of scholarly activity.

## Conclusion

Indexing sovereignty necessitates a shift away from editorial prestige and toward infrastructural autonomy in order to operationalize visibility as a primary product of high-fidelity engineering. To do so, Diamond Open Access must incorporate visibility-related aspects - interoperability via JATS XML, OAI-PMH endpoints, and multi-layered PIDs-within the publishing stack at a performance level sufficient for securing technical citizenship via hard-coded methods. Without the "plumbing" provided by this architecture,[11] research remains digitally invisible to the discovery algorithms that drive the global effect of scholarly work.

To fix the current systemic differences, we need to change our approach to sovereign technical execution:

**For Editors:** eliminated the existence of "metadata debt," which means that editors must inject as many ORCIDs, ROR Ids, structured citations, and so on into each JATS tag to build each article as a machine-readable node.

[11] Part III: Technical Guidance and Requirements - https://www.coalition-s.org/technical-guidance_and_requirements/ - Accessed: 19 March 2026

**For Institutions:** transitioned from respondents to potential grant recipients to fund a permanent infrastructure by providing DevOps assistance for OJS/Janeway instances and developing automated DOI/archival pipelines.

**For Policymakers:** must reclassify metadata protocols as a top financing priority in order to enable knowledge equity through shared discovery layers around the globe.

Prioritizing the development of these pipelines assures that the scholarly record is defined by scholarship quality rather than financial ability to purchase proprietary platforms.

**Zeba Academy** is a specialized technical research and training initiative dedicated to the principles of Sovereign Systems Engineering. Founded by Sufyan bin Uzayr - an author and university instructor as well as Google Cloud-Certified DevOps Engineer - Zeba Academy serves as a bridge between deep academic theory and high-stakes industrial implementation.

We reject the "enshittification" of modern software. Our core mission is the promotion of Anti-Bloat Architecture through the mastery of:

- **Systems Languages:** Using Rust, Zig, and C++ to build high-performance foundations that prioritize memory safety and deterministic execution.
- **SRE & DevOps:** Professional-grade automation via Google Cloud, Terraform, and Immutable Infrastructure to eliminate manual "toil" and operational fragility.
- **High-Performance Interfaces:** Utilizing Flutter for cross-platform development to deliver near-native mobile experiences without the lag of standard web-based wrappers.
- **Lean Web Publishing:** Reclaiming WordPress and PHP by stripping away the "slop", using Redis object caching and Unix sockets to transform standard platforms into high-speed, GEO-stable engines for modern publishing.
- **Legacy Modernization:** Applying memory-safe paradigms and modern build systems to century-old computational problems and aging C codebases.

Zeba Academy doesn't just teach code; we architect reliability. By merging the analytical rigor of Historical Research with the precision of Google-Certified Cloud Engineering, we provide our "Operatives" with the directives necessary to build systems that are safe, fast, and permanent.

Website:- https://zeba.academy

زیبا

Zeba Academy

zeba.academy