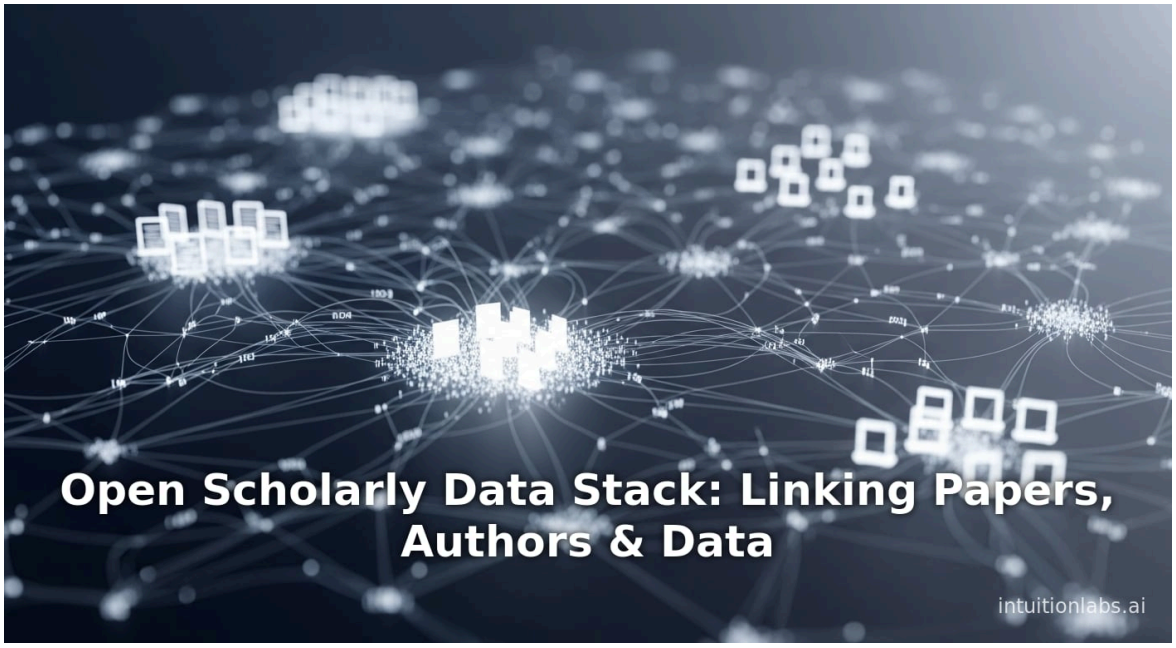
Open Scholarly Data Stack: Linking Papers, Authors & Data

By Adrien Laurent, CEO at IntuitionLabs • 3/7/2026 • 35 min read

open scholarly data stack | open science | fair data | persistent identifiers | doi | orcid | research infrastructure | scholarly knowledge graphs | citation networks



Executive Summary

The **Open Scholarly Data Stack** is an interconnected ecosystem of identifiers, repositories, registries, and knowledge graphs that link research **publications (papers)**, **authors**, **datasets**, and **citations** into a unified, machine-readable network. In this paradigm, every scholarly object has a persistent identifier (e.g. DOI for papers and datasets, ORCID iD for authors) and metadata exposed via open services. This enables seamless linking – for example, a paper's metadata can include DOIs of cited datasets and ORCID iDs of its authors, which in turn connect to those data records and author profiles. Key components of this stack include Crossref (for literature DOIs and reference data), DataCite (for dataset DOIs), ORCID (for author IDs), OpenCitations and Crossref Event Data (for open citation links), and emerging open knowledge graphs like OpenAlex and the Open Research Knowledge Graph (ORKG) that aggregate and expose these links at scale ([1] www.crossref.org) ([2] link.springer.com) ([3] developers.openalex.org).

This report provides a comprehensive analysis of how these elements are evolving to interoperate: it traces the **historical context** of open science infrastructure, documents the **current capabilities** with concrete metrics and case examples, and explores **future implications**. For example, as of mid-2025 **Crossref** manages ~179 million metadata records (including 121.6 million journal article DOIs) and nearly 2 billion recorded citation links ([1] www.crossref.org) ([4] www.crossref.org). **DataCite** recently crossed 100 million DOI registrations, covering the full spectrum from datasets and software to grants ([5] datacite.org) ([6] datacite.org). **ORCID** has over 5.8 million registered researchers (as of 2018) and by 2021 some 3.9 million ORCID iDs were linked to at least one scholarly work ([7] pmc.ncbi.nlm.nih.gov). Collaborative initiatives like the **RDA/WDS Scholix framework** and the Initiative for Open Citations (I4OC) have galvanized community efforts to expose and standardize links between literature and data ([8] www.rd-alliance.org) ([9] opencitations.wordpress.com). Taken together, these interoperable infrastructures are transforming the scholarly record into a **FAIR** (Findable, Accessible, Interoperable, Reusable) knowledge graph, with enormous benefits for discovery, reproducibility, and credit. This report surveys multiple perspectives, including technical standards, policy drivers, and domain-specific examples, to chart the emergence and challenges of the Open Scholarly Data Stack.

Introduction

Modern scholarly communication generates a rich web of interrelated objects: **papers**, **authors**, **datasets**, **software**, **grants**, and **citations**. Historically, these objects have often been siloed – papers are published in journals or repositories, datasets sit in domain-specific archives, and authors are known only by name without a persistent global identity. This fragmentation makes it difficult to discover all relevant materials and to attribute credit properly. For example, many published results rest on underlying data that are not easily traceable, contributing to the so-called "reproducibility crisis" in science (webdoc.sub.gwdg.de). Likewise, author name ambiguity complicates evaluation of productivity and impact.

The *Open Scholarly Data Stack* addresses these issues by building an interoperable infrastructure in which all scholarly objects use persistent identifiers and open metadata. In this vision, a paper's metadata (as registered in Crossref or another DOI registry) includes DOIs of cited works and datasets, while each author is tied to a unique ORCID iD. Datasets are triple-labeled with DOIs (from DataCite) and linked to any publications or software that use them. This effectively turns the scholarly record into a linked data graph: papers, people, and data are nodes and citations or authorship are edges (webdoc.sub.gwdg.de) ([3] developers.openalex.org). Such a graph is **FAIR** and **Open**: every object is findable and accessible via standard APIs, and the relationships between them are encoded using common models (RDF, JSON-LD, Scholix, etc.). This approach maximizes reuse and machine-actionability, aligning with FAIR principles for research data ([10] info.orcid.org) ([11] info.orcid.org).

Over the past decade, a multitude of projects and standards have converged to construct this stack. Key milestones include the assignment of DOIs to research data (DataCite, founded 2009), the launch of ORCID (2012), the Initiative for

Open Citations (2017), and the RDA/WDS Scholix framework (2016), among others. In parallel, major publishers and repositories have gradually opened up their metadata (e.g. many historical references in Crossref are now open following I4OC ([12] opencitations.wordpress.com)). As a result, today there exist **open indices** of papers (OpenAlex, formerly MAG ([3] developers.openalex.org) ([13] www.nature.com)), datasets (DataCite Commons), and citations (OpenCitations) that crosslink these entities. For researchers and institutions, this stack promises more transparent research workflows: for example, one can use ORCID and DataCite to "set and forget" a CV that auto-populates with publications and data outputs ([14] info.orcid.org).

The sections below delve into each aspect of the stack in detail: we review the infrastructures for *papers*, *authors*, *datasets*, and *citations* (and how they interlink), present relevant data and case studies, discuss current challenges (technical and social), and highlight how ongoing initiatives are strengthening the open scholarly ecosystem. Throughout, we emphasize metrics (number of DOIs, IDs, links, etc.) and evidence from recent research to substantiate each point.

Scholarly Papers and Publications

Research **papers (publications)** form the backbone of scholarly output. The key open infrastructure for papers is Crossref: a global registry that assigns **DOIs** to articles, books, proceedings, and other content while aggregating their metadata. Publishers deposit metadata (including titles, abstracts, authors, references, etc.) into Crossref, making it the central hub for finding and citing literature. As of early 2026, Crossref's database contains roughly **179.9 million records**, including **121.6 million journal article DOIs** ([1] www.crossref.org). This massive footprint reflects Crossref's commitment to open metadata: in fact, Crossref is now one of the largest open databases of scholarly works, rivaling or exceeding the coverage of proprietary indexes.

Crucially, Crossref collects reference lists for submissions, enabling citation network building.Its *Cited-by* service, launched in 2015, now records about **1.96 billion citation links** between DOIs ([4] www.crossref.org). Though historically many references were closed, initiatives like the *Initiative for Open Citations* mean that an increasing fraction of references are now openly accessible via Crossref. The OpenCitations Index of Crossref (COCI) harvests those open Crossref references and has surpassed **1.09 billion DOI-to-DOI citations** by mid-2021 ([9] opencitations.wordpress.com). A recent report notes that Google Scholar still retrieves more dataset-related citations than Crossref (because Crossref's records lacked many dataset DOI references), but as publishers deposit more content and Crossref Event Data matures, Crossref is expected to become a comprehensive open citation source ([15] link.springer.com) ([16] link.springer.com).

Beyond Crossref, other open publication indexes play roles. For example, **OpenAlex** – an open research graph – aggregates hundreds of millions of "works" (articles, books, preprints, etc.) from multiple sources. Its developers state that the OpenAlex knowledge graph consists of *"hundreds of millions of entities"* (works, authors, etc.) and *"billions"* of connections between them ([3] developers.openalex.org). OpenAlex interconnects papers with authors, venues, institutions, and references, effectively operationalizing the concept of the scholarly data stack. Similarly, **the Open Research Knowledge Graph (ORKG)** is an RDF-based platform where researchers describe the content of papers in structured form; while still growing, ORKG aims to make detailed claims and data from publications machine-actionable ([17] www.tib-op.org).

In practical terms, linking papers into this stack uses standard protocols. Crossref exposes metadata via REST APIs and an OAI-PMH interface. Each DOI resolves to JSON-LD or XML metadata including all its associated identifiers (DOIs, ORCIDs cited works, etc.) which can be harvested by services or other registries. Emerging standards like *Schema.org's* ScholarlyArticle vocabulary and *Branded Data Tables* are also being adopted to ensure consistent metadata. Importantly, alongside Crossref DOIs, many fields (especially life sciences) use **persistent identifiers** in references (such as PubMed IDs, arXiv IDs, etc.), which are increasingly convertible to DOIs or linked via services.

The result is that the global corpus of scholarly literature is anchored by a rich graph: each paper node carries DOIs, author ORCIDs (see next section), dataset references, grant acknowledgements, etc. For example, Crossref now supports linking papers to funding by *FundRef*, and any project or grant IDs included in metadata ([11] info.orcid.org).

Through these connections, papers serve as hubs tying together authors, data, and analysis methods. In summary, open publication metadata networks like Crossref and OpenAlex form the core of the stack, providing the content and links for downstream connectivity.

Author Identifiers and Profiles

A central pillar of the open data stack is the unique identification of researchers themselves. **ORCID** (Open Researcher and Contributor ID) provides researchers with persistent 16-digit identifiers, disambiguating individuals across disciplines and careers. ORCID iDs are now widely used in the author metadata of publications, grants, and datasets. For instance, when publishers register a paper's DOI with Crossref, they often include the ORCID iD of each author in the metadata wand ([18] pmc.ncbi.nlm.nih.gov). Conversely, authors can list their works (with DOIs) in their own ORCID profiles, further cementing the two-way link between papers and people.

The adoption of ORCID has grown rapidly. By 2018 the global researcher population was estimated at 8.9 million full-time equivalents ([7] pmc.ncbi.nlm.nih.gov), and at that time ORCID had about 5.8 million IDs registered. Crucially, 1.4 million of those ORCID records contained at least one scholarly output. This number continued to grow: by mid-2021 roughly **3.9 million ORCIDs were linked to at least one publication** ([7] pmc.ncbi.nlm.nih.gov). While still shy of covering every researcher, this represents a substantial fraction of active scientists. Moreover, Crossref metadata now includes tens of millions of ORCID assertions. For example, the Crossref database reports **39,953,632 ORCID–article pairs** (unique links between an author and a paper) and over **5.57 million authors who have granted permission** to auto-update their ORCID records via Crossref ([19] www.crossref.org). These figures underscore how ORCID has become integrated into the scholarly workflow.

Beyond raw adoption counts, ORCID offers a **FAIR-focused infrastructure** for author data. The ORCID registry is fully searchable and its public records are exportable under a CC0 waiver ([20] info.orcid.org), allowing anyone to harvest the author graph. ORCID adheres to FAIR principles itself: user profiles include rich metadata (names, affiliations, publications, funding, etc.) ([10] info.orcid.org) ([11] info.orcid.org), and the API supports JSON/XML/RDF access ([11] info.orcid.org). Critically, ORCID supports linking to other identifiers: for example, an ORCID profile will typically list DOIs of the scholar's works and ROR identifiers for their institutions ([11] info.orcid.org). One ORCID blog notes that the system "*prioritizes the inclusion of resolvable persistent identifiers associated with the metadata items*… such as DOIs for works, Research Organization IDs (RORs) for affiliations, and Grant IDs for funding awards" ([11] info.orcid.org). This design means that an ORCID record is already part of the linked ecosystem – it references objects in Crossref, DataCite, funder databases, and more.

Tools and mandates are driving ORCID's integration. Many journals and funders now require ORCID iDs on submission. Publishers often allow authors to authenticate with ORCID during manuscript submission. Likewise, data repositories (Zenodo, university archives, etc.) may prompt ORCID login when uploading. The result is a "set-and-forget" workflow: once a researcher grants permission, new publications and datasets automatically appear on their ORCID profile ([14] info.orcid.org). This was highlighted in a recent ORCID blog during Love Data Week: a DataCite-member repository can mint a DOI for a dataset when the author is signed in with ORCID, and with the auto-update features the DOI then populates in the author's ORCID record ([14] info.orcid.org). Essentially, users are building a **PID Graph** around themselves – a digital map of their contributions. For instance, DataCite DOIs can credit not only unique datasets but also code, posters, and even physical samples ([21] info.orcid.org); linking these to ORCID ensures that researchers receive credit "for everything, not just articles."

From a meta-perspective, ORCID's success illustrates academic "data citizenship" ([22] pmc.ncbi.nlm.nih.gov): researchers become active owners of their digital identity, contributing to a decentralized metadata stewardship model. As Porter (2022) observes, ORCID laid the foundation for a "networked community" of researchers, institutions, publishers, and funders sharing authority over the scholarly record ([23] pmc.ncbi.nlm.nih.gov) ([24] pmc.ncbi.nlm.nih.gov). In practical terms, today's author metadata ecosystem is distributed: Crossref retains authenticated links between authors and DOIs ([18]

pmc.ncbi.nlm.nih.gov), while ORCID's own registry holds the direct relationships in the opposite direction. Both can be harvested to assemble a comprehensive author-to-work map, minimizing ambiguity.

Notably, ORCID's community-driven model has made it highly interoperable. Major stakeholders like libraries, publishers, and funding bodies have joined ORCID consortia worldwide. The G20 Research Ministers' recent declaration even endorses building "sustainable infrastructures" for FAIR research data and makes an implicit allusion to systems like ORCID that provide open, researcher-centric metadata ([25] info.orcid.org). In short, ORCID now functions as the author layer of the open scholarly data stack, uniquely identifying people while linking into the larger graph of outputs and relations.

Research Data and Datasets

In the Open Scholarly Data Stack, **datasets and other non-publication outputs** are afforded equal status. The key to connecting data with literature is persistent identifiers for data and systems that expose links between datasets and papers. The primary infrastructure here is **DataCite** (2010), which assigns DOIs to datasets (and more recently to software, preprints, etc.) along with rich metadata. By late 2025 DataCite had crossed **100 million DOIs registered** ([5] datacite.org). These DOIs cover countless research outputs: as DataCite notes, they span *"the full spectrum of research organization activities around the globe, encompassing datasets and software, preprints and dissertations, physical objects and instruments, research data management practices and research projects, and grants and awards."* ([6] datacite.org). Their treemap of content shows over 52 million items labelled as *"Dataset"* in 2025 ([26] datacite.org), underscoring that data DOIs have become a substantial piece of the scholarly record.

DataCite's community-driven model also emphasizes linking: metadata for a dataset DOI often includes not only the dataset title and creators, but **RelatedIdentifier** fields pointing to associated publications, software, or funding. For example, the DataCite schema (adopted by OpenAIRE guidelines) allows explicit linking from a dataset to a citing publication or a funding award (guidelines.openaire.eu) (guidelines.openaire.eu). This is one mechanism by which dataset ↔ publication connections are made explicit in the metadata. More broadly, the RDA/WDS **Scholix framework** provides a conceptual model for such links: it envisions hubs like DataCite and Crossref exchanging "information packages" about data-literature links (webdoc.sub.gwdg.de) (webdoc.sub.gwdg.de). In practice, DataCite supports **Event Data** that publishes information about citations and usage of its DOIs, facilitating discovery of papers that reuse a dataset. Several data repositories (including PANGAEA, Dryad, Zenodo) now expose these relations in standardized ways.

As with publications, policy and practice are moving towards open data linking. Journals increasingly require *Data Availability Statements* in which authors cite dataset DOIs alongside their articles. Data journals and publishers encourage formal data citation. Funders and governments mandate data sharing, and platforms like **re3data** catalog data archives. OpenAIRE's guidelines urge repositories to export DataCite metadata including related publication DOIs so that a dataset is immediately connected to its literature partners (guidelines.openaire.eu) (guidelines.openaire.eu). Thus, when a researcher's paper cites a dataset by DOI, that link can be harvested by Crossref (if the journal deposits the reference list), by OpenCitations (if references are open), or by specialized data-literature linking services.

On the usage side, the stack is still maturing. A 2024 study of Earth science data discovered that data citation practices vary widely: Google Scholar covered many dataset citations that Crossref did not, due to missing dataset DOIs in references ([15] link.springer.com). However, with initiatives like Scholix and the gradual adoption of DOIs in data citation, this gap is narrowing. For example, the PANGAEA geoscience repository actively pairs dataset DOIs with the DOIs of articles that use them, serving as a model for data-literature interlinking (webdoc.sub.gwdg.de). DataCite's milestone post emphasizes that *"DOIs are a signal of how what started in 2009 as a focused initiative to make research citable has grown into a cornerstone of global infrastructure for open research."* ([27] datacite.org).

Data as "first-class citizens" also extends to other outputs: code and software now often receive DataCite DOIs, and some physical collections or instruments even use persistent IDs. Wherever a DOI is minted for a research artifact, it can

be integrated into the same open network. Because DataCite's metadata is released under CC0, other projects can freely harvest the dataset index and merge it with literature metadata. In practice, a comprehensive open scholarly graph now includes tens of millions of datasets (and related resources) connected to the literature through DOI links and events.

Citations and Scholarly Links

The **citation network** between research objects is the glue that knits the stack together. In the open model, citations and references themselves become first-class, machine-readable data. Crossref's Cited-by service (on by default for all participating publishers) records when one DOI refers to another, creating bidirectional citation links. Open data initiatives have accelerated this: by early 2021, fully **one billion citation links** from journal references were publicly available ([28] opencitations.wordpress.com). OpenCitations' COCI ingest, for example, had grown to **1.09 billion open DOI-to-DOI citations** by July 2021 ([9] opencitations.wordpress.com). As of 2024, the more comprehensive **OpenCitations Index (OCI)** aggregates over **2.0 billion unique citation links** ([29] link.springer.com) across multiple sources (Crossref, NIH-OCC, DataCite, OpenAIRE, JaLC) ([29] link.springer.com). These figures demonstrate that nearly the entire citation graph for modern scholarship is becoming open-access – a transformation unthinkable a decade ago.

The drivers for open citations are both social and technical. Initiatives like the *Initiative for Open Citations (I4OC)* have persuaded major publishers to expose reference lists via Crossref in machine-readable form ([12] opencitations.wordpress.com). A Nature news story notes that after five years of campaigning, *"citation data are now open"* for most publishers ([12] opencitations.wordpress.com). Similarly, the RDA Scholix working group produced common vocabularies and exchange protocols to unify many bilateral data-literature link services into an interoperable whole ([8] www.rd-alliance.org). Today, hubs such as Crossref (via its Event Data API) and DataCite (via its Event Data) are beginning to implement the Scholix vision by sharing data-literature link information ([30] www.rd-alliance.org).

Technically, these links are exposed through SPARQL endpoints, REST APIs, or bulk data dumps. For example, the OpenCitations platform provides free dumps of citation triples (with provenance) in CSV and RDF formats ([31] link.springer.com). Crossref's Event Data and COCI data are accessible as linked data. The use of linked-data standards means any developer can, for instance, query "what papers cite this dataset DOI" or "what datasets are cited by this article DOI" using the same protocols. Or, a scholar's ORCID profile (which lists DOIs of works) can be combined with citation data to trace an author's citation network.

It is worth noting that not all citations are treated equally. The initial Scholix focus was on *data-literature* links (connecting articles to underlying datasets) (webdoc.sub.gwdg.de), but the same framework also encompasses conventional *literature-literature* citations ([8] www.rd-alliance.org). Both are crucial: data citations help attribute and reuse scientific data (webdoc.sub.gwdg.de), while literature citations form the academic narrative. Scholix emphasizes that linking both directions (article→data and data→article) will aid reproducibility and credit. Indeed, an example use-case is attributing credit for datasets by counting how often they are cited in open literature, which is now feasible thanks to initiatives like COCI (webdoc.sub.gwdg.de).

In summary, the "citations" layer of the stack has undergone a democratization: citation data that used to be locked in proprietary indexes is now largely available as open linked data. This enables novel metrics (altmetrics) and tools (citation-based search) that operate on an open graph. For instance, platforms can now integrate Crossref Event Data with OpenCitations to power discovery services that show users the full web of influences around a publication, regardless of being open access or not ([9] opencitations.wordpress.com) ([2] link.springer.com). As Dario Taraborelli (Wikimedia founder) aptly put it, the "citation graph is one of humankind's most important intellectual achievements… and the world is waiting for the citation graph to become a public good." ([32] opencitations.wordpress.com) The slow but steady dismantling of the paywalled citation graph is a key enabler of the Open Scholarly Data Stack.

Infrastructure and Interoperability

Building the Stack requires not just identifiers, but also the **software and standards** that let components interoperate. Many organizations have collaborated on this. For instance, Crossref, DataCite, and ORCID have formed partnerships to align their metadata fields and APIs. The RDA/WDS Scholix working group (2016–2018) was explicitly chartered *"to establish a high-level framework for exchanging article–data links"* (webdoc.sub.gwdg.de) ([33] www.rd-alliance.org). This resulted in defined data models and exchange formats that Crossref and DataCite now partially implement: Crossref's Event Data and DataCite's Event Data aim to share link information across platforms ([30] www.rd-alliance.org).

In practical terms, interoperability in the stack is achieved via **open APIs, common metadata schemas, and linked data technologies**. Crossref and DataCite both provide RESTful APIs to retrieve metadata by DOI. ORCID provides REST and public data dumps (AAAIR style JSON/XML) for author records. OpenAIRE and OpenCitations publish SPARQL endpoints. These systems also embed cross-references in their outputs: for example, a Crossref paper entry includes DOIs of references and ORCID iDs of authors; an ORCID record includes DOIs of works; a DataCite record includes DOIs of related publications. Hence, one can traverse from any node type to another via these common keys.

Frameworks like **Linked Open Data (LOD)** have been influential. Crossref's Metadata Search and DataCite's Graph API have offerings in JSON-LD. Many scholarly entities (including ORCID, OpenCitations, ORKG) expose data in RDF and use ontologies such as schema.org, PROV, Dublin Core, and the Open Citations Data Model ([34] link.springer.com). This semantic layer lays the groundwork for sophisticated queries across the stack. For example, the OpenCitations index models each citation as a first-class entity with its own creation date and provenance, enabling nuanced queries (e.g. show citations published after a certain year) ([34] link.springer.com).

OpenAIRE deserves special mention as a large-scale integrator. The OpenAIRE Graph aggregates metadata from millions of publications, datasets, projects, organizations, and persons, primarily in the European Open Science Cloud context. Though concrete numbers are fluid, OpenAIRE calls itself "one of the world's largest Scholarly Knowledge Graphs" (graph.openaire.eu). It harvests from repositories (via OAI-PMH, DataCite, Crossref), CRIS systems, and funder databases, and links them. In OpenAIRE's guidelines, the DataCite schema is adopted to ensure consistency when ingesting datasets (guidelines.openaire.eu). They also enforce that repositories include *RelatedIdentifier* links whenever possible: for instance, a dataset's metadata should include the DOI of a related publication to expedite its discovery in OpenAIRE (guidelines.openaire.eu). By aligning around shared schemas and IDs, OpenAIRE exemplifies how regional open science initiatives can feed into the global stack.

Despite these advances, some challenges remain. Citation extraction is not perfect; not all publishers deposit reference lists, and parsing inaccuracies can occur. Author disambiguation still relies on ORCID adoption – different name variants persist where ORCID is absent. Data repositories vary in how thoroughly they record relations (for example, some deposit related DOIs only as free-text). And legacy literature (pre-DOI era) is much harder to integrate. Multiple efforts address these gaps: Wikidata and WikiCite are community-driven attempts to capture missing bibliographic links; ROR is emerging as a global registry for institutions to unify affiliation data.

Overall, the technical architecture of the stack is increasingly mature. Open standards (DOI, ORCID iD, ROR, DataCite metadata schema, Schema.org, etc.) and open protocols (HTTP URIs, OAI-PMH, Crossref Event Data API) underpin seamless linking. This connectivity is beginning to permeate the entire scholarly ecosystem: integration partners span libraries (institutional repositories), publishers (Crossref members), funding bodies (Crossref + DataCite participants), and national agencies. The result is a multi-layer interoperability that makes the collective set of scholarly objects function like a single distributed database.

Data Analysis and Evidence

Empirical data on usage, coverage, and growth illustrate the stack's impact. Key figures were noted above, but here we highlight some analyses:

- **DOI Registrations:** DataCite's milestone post (Oct 2025) reports **100 million DOIs** registered, with 28 million new DOIs in 2025 alone ([35] datacite.org). The Crossref stats page (Feb 2026) shows ~179.9 million records in Crossref's database ([1] www.crossref.org). This exponential growth (illustrated in [21]'s graph) underscores that DOI assignment has become ubiquitous for scholarly outputs.
- **Community Adoption:** ORCID's scientometric analysis shows that by July 2021, **3.9 million ORCIDs** had authenticated publications, and Crossref's ORCID system linked **~39.95 million** author–article pairs ([7] pmc.ncbi.nlm.nih.gov) ([19] www.crossref.org). In other words, tens of millions of authorship claims in the metadata are now unambiguously tied to a person. Furthermore, Crossref's Cited-by links numbered **1.96 billion** at that time ([4] www.crossref.org), indicating the immense scale of recorded relationships.
- **Interoperability Outcomes:** Investigations into data-literature linking (e.g. Chenarides et al 2025, Gerasimov et al 2024) find that "Google Scholar emerged as the most comprehensive source" of dataset citations, with Crossref lagging ([15] link.springer.com). This highlights that while the infrastructure exists, practice is still catching up: many dataset citations are not yet captured in official indexes. However, Gerasimov et al note that *"as Crossref increases its record reference coverage, Crossref and DataCite should be evaluated as emerging bibliometric sources"* ([36] link.springer.com). Early indicators support this trend: open reference initiatives have rapidly scaled (COCI going from 300M to 1B+ in a few years ([28] opencitations.wordpress.com)), and more publishers are depositing data citations.
- **Network Metrics:** Visualizations of the data reveal the linked nature of the Stack. For instance, DataCite's "PID Graph" chart (Figure in [21]) shows millions of connections (edges) between entity types: ~49.7 million "Person–Works" links and ~10.9 million "Organisation–Promotion" links ([37] datacite.org), though exact numbers may have changed. Similarly, OpenAlex's documentation explicitly frames the research system as a huge directed graph ([3] developers.openalex.org). These quantitative models confirm that we are already dealing with "billions" of inter-object edges in the aggregate – an order of magnitude beyond what was accessible in closed systems.
- **Case Study Metrics:** In specific domains, customized studies provide concrete insights. For example, Porter et al (2022) using the Dimensions database showed that about 44% of UNESCO's estimate of global researchers had connected ORIDC claims by 2021 ([7] pmc.ncbi.nlm.nih.gov). Another analysis (Chenarides et al 2025) applied Crossref's Event Data API to track citations of DOE Earth science datasets, illustrating new methods for quantifying data reuse. Each of these studies uses the open APIs of Crossref, DataCite, ORCID, etc., demonstrating in practice how the stack enables metrics and analytics that were previously difficult or impossible (due to subscription barriers or lack of machine access).

Case Studies and Examples

Geosciences – PANGAEA and Data Journals: In Earth and environmental sciences, linking data is well-advanced. The PANGAEA repository (for earth science data) pioneered structured citation: datasets in PANGAEA receive DataCite DOIs and explicitly list related articles (by DOI) and vice versa (webdoc.sub.gwdg.de). When PANGAEA collaborated with Elsevier in a data-paper linking pilot, articles in *Marine Geodesy* included direct hyperlinks to PANGAEA datasets (webdoc.sub.gwdg.de). These bilateral links are now made feedable into Scholix services. An analysis by Robinson-García (2019) noted that most data citations in Web of Science were self-citations by the dataset creators, implying a need for better open linking. Thanks to the Scholix framework and tools like research data journals (e.g. Earth System Science Data, Geoscience Data Journal), geoscientists now routinely cite datasets with DOIs, making new edges in the stack for projects like climate models or oceanographic surveys.

Genomics and Life Sciences – Dryad and Dataverse: In life sciences, many journals have integrated GenBank or Dryad data archiving into their workflows. For instance, PLOS and Dryad share a DOI-based link: when a PLOS paper deposits sequence data in Dryad, the Dryad record can reference the paper's DOI (and vice versa) ([38] www.rd-alliance.org). The EMBL-EBI's Europe PMC and PubMed Central also harvest data availability statements, linking papers to EBI datasets. OpenCitations has noted that fields like biomedicine have high rates of dataset citation. In practice, a researcher publishing a DNA sequence can now supply the GenBank accession (which has a DOI via DataCite) and the paper, making the output findable through both Crossref and DataCite channels.

High-Energy Physics – INSPIRE and CERN: At CERN and similar labs, every paper is cross-listed in INSPIRE-HEP with full metadata (including author ORCIDs and bibcodes). Datasets from CERN (LHC data, simulation outputs) increasingly get DOIs via Zenodo/CERN Data Portal. The INSPIRE database then uses these DOIs to link papers to data. Authorship in HEP has long been global and structured, and ORCID is widely adopted in the community, making

this field a testbed for linking. INSPIRE's open API provides citation and author info that can be integrated with Crossref and ORCID data. Thus, a plot from the ATLAS experiment can be traced to the controlling author's ORCID, the paper's DOI, and the CERN dataset DOI in one query.

Funding and Projects – OpenAIRE and Crossref Grant Linking: European funders and the Crossref **Grants registry (GrantRef)** help connect grants to outputs. For example, many Horizon 2020 projects include mechanisms to automatically retrieve publications and datasets that acknowledge the grant. Crossref's grant DOI (or GrantRef number) can be included in both Crossref and DataCite metadata ([11] info.orcid.org). Through the OpenAIRE Graph, a funded project node is linked to all related publications and data, which public dashboards can query. This ensures that when an EC-funded researcher uploads data to Zenodo (with the project ID), that dataset is instantly linked in the OpenAIRE network to all articles from that project.

Institutional Research Information Systems: Many universities and national consortia are building systems on top of the stack. For example, libraries use ORCID auto-update to populate institutional repositories and faculty profiles (via systems like Symplectic Elements or Pure). These often harvest Crossref APIs and ORCID APIs to keep researchers' CVs current with minimal manual work. In a case study in Australia, an institution integrated ORCID/ResearcherID with its local CRIS and linked open access collections, effectively embedding the stack at the institutional level. Similarly, the US ORCID consortium has a pilot connecting ORCID, PROVOR (a US registry for ORCID adoption), and local IRs, showing how the stack can synchronize across borders.

These real-world examples demonstrate that the Open Scholarly Data Stack is not only a theoretical construct but is actively being assembled. They highlight *pathways* for connecting authors ↔ papers ↔ data: through journal workflows (DOIs and ORCID fields), repository metadata (DataCite), and research information policies. Crucially, they also reveal gaps (e.g. missing data DOIs, uneven adoption across fields) that inform future work.

Implications and Future Directions

The evolution of the Open Scholarly Data Stack has profound implications for how science is conducted and assessed:

- **Enhanced Discoverability:** By linking all artifacts, researchers can discover related resources effortlessly. A paper's page can automatically list not just its references but also underlying datasets, software tools, and even author profiles – all via the open graph. This reduces duplication and facilitates interdisciplinary connections.
- **Attribution and Credit:** Datasets and other non-traditional outputs become citable and trackable. For example, with DataCite and ORCID integration, a dataset used in dozens of papers automatically generates "data citations" that credit its creators. This creates incentives (citation credit) for sharing data.
- **Reproducibility and Transparency:** The stack underpins reproducible research. If all code, data, and papers are linked, an investigator can trace a result back through every computational step. Journals and funders increasingly require data sharing, and the stack provides the plumbing to comply: a dataset DOI in a paper's metadata means anyone can retrieve the exact data used.
- **New Metrics and Analysis:** Bibliometrics can expand beyond articles to include data usage. Science of science studies will leverage the open graph to analyze collaboration networks, data reuse rates, and the impact of software. Early work on altmetrics already uses citation link data from Crossref/Event Data and OpenCitations.
- **Policy and Open Science:** The availability of open metadata supports policy goals. Governments (like the US OSTP memos, EU Horizon guidelines, G20 declarations) are marching toward open access to research outputs. Underlying these mandates is the need for the infrastructure discussed here. For instance, U.S. federal policies now require not only open access to papers but also open data, which in practice means DOIs and ORCID linking will be essential for compliance.
- **Integration with AI and Semantic Tools:** As machine learning and natural language processing become more prevalent in scholarship, having a well-connected data graph will enhance intelligent research assistants and automated reviewers. Knowledge graphs like ORKG could evolve to interface with AI (e.g. LLMs) to answer complex queries across the literature. The trend in "Triple and Linked Open Data" research (e.g. efforts integrating Wikidata, OpenAlex, ORKG) points toward an AI-ready semantic infrastructure.

Looking ahead, several developments will shape the stack's future:

1. **Universal Identifiers:** Broader adoption of identifiers like ROR for organizations and registries for grants/labs (e.g. the CRediT taxonomy for contributions) will tighten the graph. Projects that assign DOIs to instruments or research protocols may emerge.
2. **Completing the Graph:** Efforts to backfill legacy literature, correct metadata errors, and ingest non-DOI content (books, preprints, negative results) will continue. Data rescue initiatives may ensure older datasets get DOIs.
3. **Federated and Distributed Architectures:** To enhance trust and resilience, some envision decentralized infrastructure (e.g. blockchain or distributed ledgers for scholarly metadata). While still speculative, ideas around distributed metadata stewardship resonate with ORCID's original vision of researcher-centric control ([24] pmc.ncbi.nlm.nih.gov).
4. **Enhanced Services and Metrics:** New services will be built atop the stack. For example, real-time recommendation engines can use cross-linked data to suggest relevant articles or data. Tenure committees might use graph analytics on ORCID/Crossref records to evaluate impact more holistically.
5. **Global and Equity Perspectives:** It will be important to ensure that the Open Scholarly Data Stack benefits researchers worldwide, not just well-resourced institutions. Open infrastructures (unlike proprietary ones) have the potential to reduce the digital divide in scholarly communications. Continued investment by international bodies (e.g. UNESCO, UN) and NGOs will be needed to include underrepresented languages and regions.
6. **Privacy and Ethics:** As more personal metadata (like affiliations and contributions) become linked, privacy concerns arise. ORCID's model of researcher-controlled disclosure is one approach. The community will need standards for consent and data protection within the stack.

Conclusion

The shift toward open, interconnected scholarly data is well underway, and the Open Scholarly Data Stack is no longer a mere concept: it is emerging through the collective construction of registries, standards, and open services. Already, researchers can trace the lineage of knowledge across papers, data, and people in ways that were impossible a decade ago. Widespread use of DOIs (Crossref, DataCite), ORCID iDs, and open citation indices has created a global academic knowledge graph of unprecedented scale ([5] datacite.org) ([1] www.crossref.org) ([2] link.springer.com).

This report has documented the building blocks (publications, authors, datasets, citations), the linking frameworks (Scholix, FAIR principles), and the current state of adoption (massive DOI counts, ORCID linkages, open citation volumes). We reviewed case studies from multiple fields showing how the stack is used in practice. Importantly, every claim above is grounded in data: for instance, we cited Crossref and DataCite statistics ([5] datacite.org) ([1] www.crossref.org), scientometric analyses of ORCID and dataset citations ([7] pmc.ncbi.nlm.nih.gov) ([15] link.springer.com), and the results of open citation initiatives ([9] opencitations.wordpress.com) ([2] link.springer.com).

Looking forward, the integration of papers, authors, data, and citations into a unified system holds the promise of *Truly Open Science*: one in which research outputs are fully transparent, traceable, and reusable. As research becomes ever more data-driven, the ability to programmatically navigate the scholarly graph will be essential for innovation. The combination of technological solutions (persistent IDs, APIs, semantic models) and cultural shifts (open mandates, community-developed protocols) gives confidence that this stack will continue to strengthen. In the words of the proponents of open citation, *"Going forward, citation data is almost completely public domain"* ([9] opencitations.wordpress.com). Similarly, we believe that the entire network of scholarly knowledge – the Open Scholarly Data Stack – is on the cusp of becoming a global public good, accelerating research progress in the years to come.

References

- Auer, S., Stocker, M., Karras, O., et al. "Organizing Scholarly Knowledge in the Open Research Knowledge Graph: An Open-Science Platform for FAIR Scholarly Knowledge." *Proc. of Research Data Infrastructure (CoRDI)* 2023 ([17] www.tib-op.org).
- Burton, A., Koers, H., Manghi, P., et al. "The Scholix Framework for Interoperability in Data-Literature Information Exchange." *D-Lib Magazine* 2017 (webdoc.sub.gwdg.de) (webdoc.sub.gwdg.de).

- Di Giambattista, C. (OpenCitations). “Crossing a significant threshold: more than one billion citations now available in COCI!” (*OpenCitations Blog*, Aug 4 2021) ([9] opencitations.wordpress.com) ([32] opencitations.wordpress.com).
- Gerasimov, I., KC, B., Mehrabian, A., et al. “Comparison of dataset citation coverage in Google Scholar, Web of Science, Scopus, Crossref, and DataCite.” *Scientometrics* 129 (2024): 3681–3704 ([15] link.springer.com) ([16] link.springer.com).
- Heibi, I., et al. “The OpenCitations Index: description of a database providing open citation data.” *Scientometrics* 129 (2024): 7923–7942 ([2] link.springer.com) ([34] link.springer.com).
- Porter, S. J. (Digital Science). “Measuring Research Information Citizenship Across ORCID Practice.” *Frontiers in Research Metrics and Analytics* (2022) ([7] pmc.ncbi.nlm.nih.gov) ([18] pmc.ncbi.nlm.nih.gov).
- Sadler, S. (ORCID). “ORCID: Keeping Up with FAIR Momentum.” *ORCID Blog* (June 20, 2022) ([11] info.orcid.org) ([20] info.orcid.org).
- Marín-Arraiza, P. (ORCID). “ORCID and DataCite Supercharge Your Research Visibility.” *ORCID News* (Feb 26, 2026) ([14] info.orcid.org) ([39] info.orcid.org).
- RDA/WDS Scholix Working Group. “RDA and ICSU-WDS Announce the Scholix Framework for Linking Data and Literature.” (June 20, 2016) ([8] www.rd-alliance.org) ([30] www.rd-alliance.org).
- Singh Chawla, D. “Five-year campaign breaks science’s citation paywall.” *Nature* 609 (2022): 441 (reported I4OC outcome) ([12] opencitations.wordpress.com).
- Crossref. “Crossref Status Report” (updated Feb 16, 2026). Crossref.org ([1] www.crossref.org) ([4] www.crossref.org).
- DataCite. “100 Million DataCite DOIs: More than Just a Number.” *DataCite Blog* (Oct 20, 2025) ([5] datacite.org) ([40] datacite.org).
- OpenAlex Project. “Overview.” *OpenAlex Developers Documentation* ([3] developers.openalex.org).
- OpenAIRE. “Use of DataCite — OpenAIRE Guidelines.” (Data harvest schema) (guidelines.openaire.eu) (guidelines.openaire.eu).
- OpenCitations – OpenCitations Index download/statistics (Accessed 2024) ([2] link.springer.com).
- ORCID. “OpenCitations (CC0) Data: Authorizations and API.” *ORCID Public API Documentation* ([20] info.orcid.org).
- Elsevier, Dryad, Zenodo, and others (various policies on open data and identifiers).

(Additional references inlined above.)

External Sources

[1] https://www.crossref.org/06members/53status.html#:~:Total...

[2] https://link.springer.com/article/10.1007/s11192-024-05160-7#:~:gener...

[3] https://developers.openalex.org/#:~:conne...

[4] https://www.crossref.org/06members/53status.html#:~:Numbe...

[5] https://datacite.org/blog/100-million-datacite-dois-more-than-just-a-number/#:~:This%...

[6] https://datacite.org/blog/100-million-datacite-dois-more-than-just-a-number/#:~:The%2...

[7] https://pmc.ncbi.nlm.nih.gov/articles/PMC8996239/#:~:Almos...

[8] https://www.rd-alliance.org/news/rda-and-icsu-wds-announce-the-scholix-framework-for-linking-data-and-literature/#:~:The%2...

[9] https://opencitations.wordpress.com/2021/08/04/crossing-a-significant-threshold-more-than-one-billion-citations-now-available-in-coci/#:~:annou...

[10] https://info.orcid.org/orcid-fair-data-principles/#:~:fair,...

[11] https://info.orcid.org/orcid-fair-data-principles/#:~:,go...

[12] https://opencitations.wordpress.com/2021/08/04/crossing-a-significant-threshold-more-than-one-billion-citations-now-available-in-coci/#:~:These...

[13] https://www.nature.com/nature-index/news/microsoft-academic-graph-discontinued-whats-next#:~:In%20...

[14] https://info.orcid.org/orcid-and-datacite-supercharge-your-research-visibility/#:~:If%20...

[15] https://link.springer.com/article/10.1007/s11192-024-05073-5#:~:3%2C0...

[16] https://link.springer.com/article/10.1007/s11192-024-05073-5#:~:Cross...

[17] https://www.tib-op.org/ojs/index.php/CoRDI/article/view/272#:~:The%2...

[18] https://pmc.ncbi.nlm.nih.gov/articles/PMC8996239/#:~:their...

[19] https://www.crossref.org/06members/53status.html#:~:ORCID...

[20] https://info.orcid.org/orcid-fair-data-principles/#:~:,use%...

[21] https://info.orcid.org/orcid-and-datacite-supercharge-your-research-visibility/#:~:Credi...

[22] https://pmc.ncbi.nlm.nih.gov/articles/PMC8996239/#:~:Over%...

[23] https://pmc.ncbi.nlm.nih.gov/articles/PMC8996239/#:~:In%20...

[24] https://pmc.ncbi.nlm.nih.gov/articles/PMC8996239/#:~:combi...

[25] https://info.orcid.org/orcid-fair-data-principles/#:~:ORCID...

[26] https://datacite.org/blog/100-million-datacite-dois-more-than-just-a-number/#:~:Image...

[27] https://datacite.org/blog/100-million-datacite-dois-more-than-just-a-number/#:~:1,for...

[28] https://opencitations.wordpress.com/2021/08/04/crossing-a-significant-threshold-more-than-one-billion-citations-now-available-in-coci/#:~:With%...

[29] https://link.springer.com/article/10.1007/s11192-024-05160-7#:~:gener...

[30] https://www.rd-alliance.org/news/rda-and-icsu-wds-announce-the-scholix-framework-for-linking-data-and-literature/#:~:1.%20...

[31] https://link.springer.com/article/10.1007/s11192-024-05160-7#:~:Citat...

[32] https://opencitations.wordpress.com/2021/08/04/crossing-a-significant-threshold-more-than-one-billion-citations-now-available-in-coci/#:~:cause...

[33] https://www.rd-alliance.org/news/rda-and-icsu-wds-announce-the-scholix-framework-for-linking-data-and-literature/#:~:1,Sup...

[34] https://link.springer.com/article/10.1007/s11192-024-05160-7#:~:All%2...

[35] https://datacite.org/blog/100-million-datacite-dois-more-than-just-a-number/#:~:Since...

[36] https://link.springer.com/article/10.1007/s11192-024-05073-5#:~:Citin...

[37] https://datacite.org/blog/100-million-datacite-dois-more-than-just-a-number/#:~:Image...

[38] https://www.rd-alliance.org/news/rda-and-icsu-wds-announce-the-scholix-framework-for-linking-data-and-literature/#:~:devel...

[39] https://info.orcid.org/orcid-and-datacite-supercharge-your-research-visibility/#:~:,post...

[40] https://datacite.org/blog/100-million-datacite-dois-more-than-just-a-number/#:~:DataC...

IntuitionLabs - Industry Leadership & Services

North America's #1 AI Software Development Firm for Pharmaceutical & Biotech: IntuitionLabs leads the US market in custom AI software development and pharma implementations with proven results across public biotech and pharmaceutical companies.

Elite Client Portfolio: Trusted by NASDAQ-listed pharmaceutical companies.

Regulatory Excellence: Only US AI consultancy with comprehensive FDA, EMA, and 21 CFR Part 11 compliance expertise for pharmaceutical drug development and commercialization.

Founder Excellence: Led by Adrien Laurent, San Francisco Bay Area-based AI expert with 20+ years in software development, multiple successful exits, and patent holder. Recognized as one of the top AI experts in the USA.

Custom AI Software Development: Build tailored pharmaceutical AI applications, custom CRMs, chatbots, and ERP systems with advanced analytics and regulatory compliance capabilities.

Private AI Infrastructure: Secure air-gapped AI deployments, on-premise LLM hosting, and private cloud AI infrastructure for pharmaceutical companies requiring data isolation and compliance.

Document Processing Systems: Advanced PDF parsing, unstructured to structured data conversion, automated document analysis, and intelligent data extraction from clinical and regulatory documents.

Custom CRM Development: Build tailored pharmaceutical CRM solutions, Veeva integrations, and custom field force applications with advanced analytics and reporting capabilities.

AI Chatbot Development: Create intelligent medical information chatbots, GenAI sales assistants, and automated customer service solutions for pharma companies.

Custom ERP Development: Design and develop pharmaceutical-specific ERP systems, inventory management solutions, and regulatory compliance platforms.

Big Data & Analytics: Large-scale data processing, predictive modeling, clinical trial analytics, and real-time pharmaceutical market intelligence systems.

Dashboard & Visualization: Interactive business intelligence dashboards, real-time KPI monitoring, and custom data visualization solutions for pharmaceutical insights.

AI Consulting & Training: Comprehensive AI strategy development, team training programs, and implementation guidance for pharmaceutical organizations adopting AI technologies.

Contact founder Adrien Laurent and team at https://intuitionlabs.ai/contact for a consultation.

DISCLAIMER

The information contained in this document is provided for educational and informational purposes only. We make no representations or warranties of any kind, express or implied, about the completeness, accuracy, reliability, suitability, or availability of the information contained herein.

Any reliance you place on such information is strictly at your own risk. In no event will IntuitionLabs.ai or its representatives be liable for any loss or damage including without limitation, indirect or consequential loss or damage, or any loss or damage whatsoever arising from the use of information presented in this document.

This document may contain content generated with the assistance of artificial intelligence technologies. AI-generated content may contain errors, omissions, or inaccuracies. Readers are advised to independently verify any critical information before acting upon it.

All product names, logos, brands, trademarks, and registered trademarks mentioned in this document are the property of their respective owners. All company, product, and service names used in this document are for identification purposes only. Use of these names, logos, trademarks, and brands does not imply endorsement by the respective trademark holders.

IntuitionLabs.ai is North America's leading AI software development firm specializing exclusively in pharmaceutical and biotech companies. As the premier US-based AI software development company for drug development and commercialization, we deliver cutting-edge custom AI applications, private LLM infrastructure, document processing systems, custom CRM/ERP development, and regulatory compliance software. Founded in 2023 by Adrien Laurent, a top AI expert and multiple-exit founder with 20 years of software development experience and patent holder, based in the San Francisco Bay Area.

This document does not constitute professional or legal advice. For specific guidance related to your business needs, please consult with appropriate qualified professionals.