WHY METADATA MATTERS FOR RESEARCH INTEGRITY AND HOW TO CONTRIBUTE

A GUIDE TO USING CROSSREF AND DATACITE

Version 1.0 | April 2026

https://doi.org/10.5281/zenodo.19695957

By:
Madhura Amdekar https://orcid.org/0000-0003-1741-6646
Xiaoli Chen https://orcid.org/0000-0003-0207-2705
Helena Cousijn https://orcid.org/0000-0001-6660-6214
Sara El-Gebali https://orcid.org/0000-0003-1378-5495
Patricia Feeney https://orcid.org/0000-0002-4011-3590
Ginny Hendricks https://orcid.org/0000-0002-0353-2702
Kelly Stathis https://orcid.org/0000-0001-6133-4045

Cite as: Amdekar M., Chen X., Cousijn H., El-Gebali S., Feeney P., Hendricks G., Stathis K. (2026, April). Why metadata matters for research integrity and how to contribute. Zenodo. https://doi.org/10.5281/zenodo.19695957

DataCite Crossref

Table of Contents

Introduction: The Relationship Between Metadata and Research Integrity

The size of the research domain has grown in recent years, with more parties sharing more scientific information online than ever before and in numerous formats. The process of communicating one's research has become digital-first and highly programmatic. With that, there is increased attention from all sides on defining and establishing the trustworthiness of research, and metadata is playing a critical role in denoting, for example, how the research was planned, what methods were followed, or who funded the activities.

About this guide

Both Crossref and DataCite recognise their growing roles in helping all those involved in assessing research to create and use metadata as part of the research evidence trail. We have each grown our respective operations to help establish and preserve the rigour and integrity of the scholarly record in several ways. This joint guide aims to provide practical help to all stakeholders who are responsible for research integrity—that is, researchers, institutions, funding bodies, and publishers—to think of metadata as a tool that helps tell part of the story and to act in two ways: as members and creators to create and enrich their metadata records with Crossref and DataCite; and also as users and assessors to look to our open APIs and incorporate this information in their analyses and systems.

What is research integrity?

Responsible research demands independence, accuracy, efficiency, and impartiality. Research integrity concerns the soundness of the scientific process and can include the ethical treatment of humans and animals as well as consideration of conflicts of interest, honesty, and equality.

The part of research integrity relevant to open scholarly infrastructure organisations like DataCite and Crossref is about the reporting and acknowledgement of input, process, output, and evidence, including the clear identification of the individuals and organisations accountable for upholding research integrity and addressing any issues that arise. We help our member organisations assert facts about the research objects that they steward, such as methods and protocols, data management plans, registered reports, and research funding grants.

Both organisations have a mission to support the rigour and integrity of the scholarly record. Crossref's Integrity of the Scholarly Record (ISR) program can be viewed here, and DataCite also emphasizes the importance of rich metadata in various ways, for example, in its blog post "Connecting the Dots with DataCite DOI Metadata", by stating that

assigning DOIs with comprehensive metadata enhances discoverability, contextualization, and the integrity of research outputs.

What is metadata?

Metadata is detailed information that communicates nuance and context about something. In scientific research, these 'somethings' are objects (such as articles, datasets, reports, grants, projects, etc.) and entities (such as contributors like authors, funders, institutions, publishers, as well as readers/reusers). Together, this metadata makes up 'the scholarly record'. The detailed information can include what went into the work, such as funding, methods, data management plans, and what is known about the work, such as titles, descriptions, and dates. Metadata can also denote how the object or entity relates multilaterally to numerous others, and how it is being adapted, updated, or used. This kind of metadata is called relationship metadata or connection metadata. The image below shows an example of a Crossref record for a journal article and some of the metadata that can be included to enrich the record and add context and nuance to the story of the research, to help downstream users assess the work.

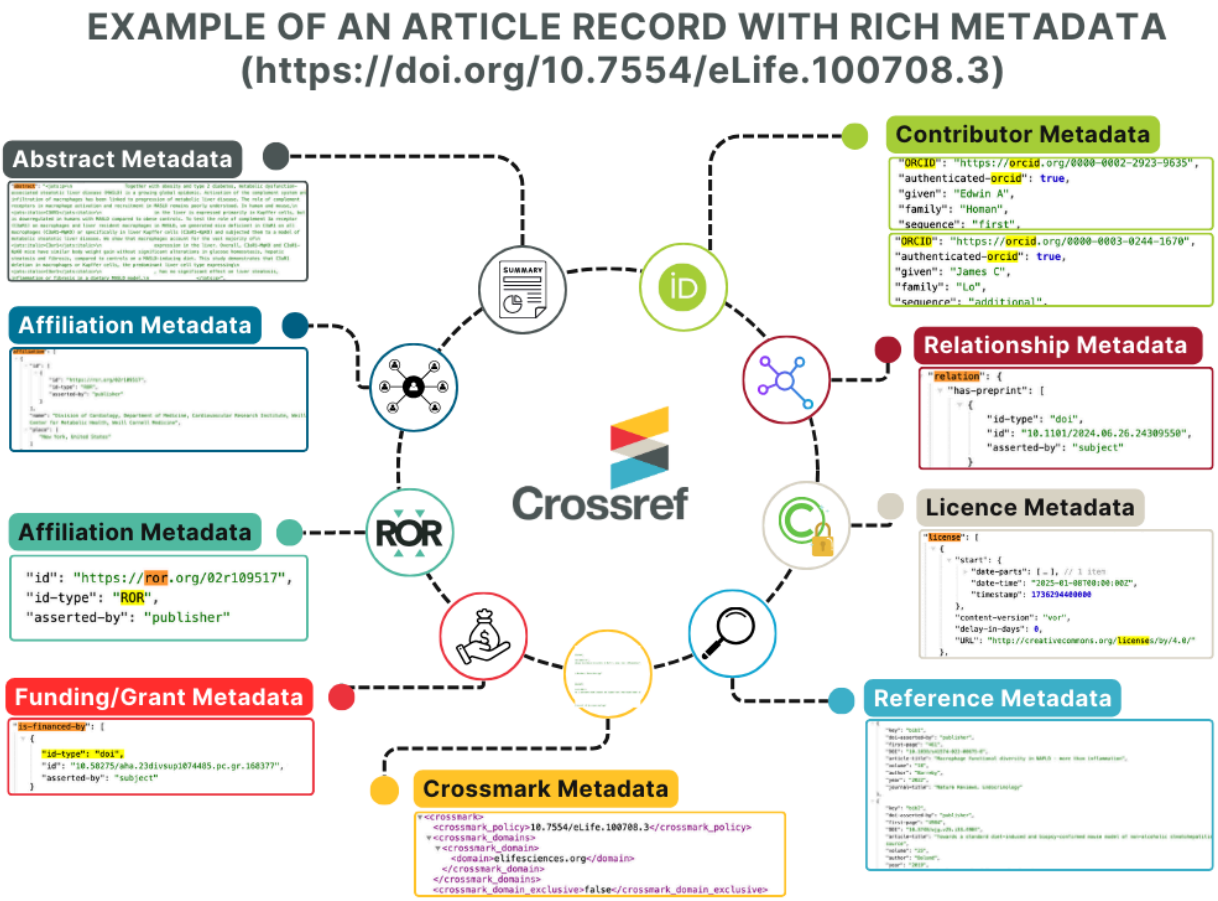

Figure 1. *Case example of an article record in Crossref with rich metadata (Credit: E Atoni).*

Who is responsible for the metadata?

Metadata is maintained and added to over time by an authority on that object, such as a repository, institution, publisher, funder, or research group. These 'stewards' join open scholarly infrastructure organisations like DataCite and Crossref as members, in order to create and maintain the metadata, and they commit to stewardship over the long term, ideally with preservation approaches in place too. They founded and continue to support DataCite and Crossref in distributing such metadata about the scholarly record throughout the research ecosystem to thousands of services such as discovery engines, library systems, or analytics platforms, so anyone can discover and assess information about the research and make more informed decisions about trustworthiness and reuse. Such users often add to and enrich and fill gaps in the metadata and both organisations are working on ways to accommodate multi-party community assertions back into the records they run.

The metadata elements that can help assess the integrity of the record

Metadata	Crossref	DataCite
Contributors and their roles	✓	✓
Affiliations	✓	✓
Dates	✓	✓
Publication year	✓	✓
Funder, funding, and grant information	✓	✓
Versioning	✓	✓
Retractions, corrections, and updates	✓	
Abstract/description	✓	✓
Clinical trials	✓	
References	✓	✓
Peer reviews	✓	✓
Publisher/steward	✓	✓
Record types/resource types	✓	✓
Alternate identifier		✓
Relationships/related identifier	✓	✓
Rights, licenses, and access indicators	✓	✓

Contributors and their roles

Contributor metadata plays a vital role in maintaining research integrity by capturing detailed information about individuals or organisations involved in a piece of work, including their names, roles, affiliations, and identifiers like ORCID.

The use of ORCID IDs helps in reliably associating research outputs with the people responsible for authoring them. While names can be used to distinguish between researchers, this method is not always reliable because names are not unique, may change over time, vary in spelling, and can be hard to disambiguate. While Crossref and DataCite accept other identifiers for contributors, they both prefer ORCID as a persistent and open infrastructure, into which both organisations push updates to works programmatically.

In Crossref, each contributor is assigned a single role, such as author, editor, chair, reviewer, review assistant, stats reviewer, external reviewer, reader, or translator and in the case of grants, multiple co-investigators or one lead investigator can be included. This will be expanded in their forthcoming schema version 5.5 to support multiple contributors and CRediT for both individual and organisational contributors.

DataCite distinguishes between creators—the main researchers—and contributors, who may include institutions and individuals in various supportive roles, recognizing up to 20 different roles.

By documenting this variety of roles, contributor metadata allows for a nuanced assessment of individual contributions, helping to identify the specific roles that drive research forward and informing strategies to support these roles more effectively. Recognising a broad spectrum of roles promotes inclusivity and equity in research by valuing the input of all participants, not just the primary authors. This practice also elevates the visibility of underrepresented groups who play critical yet often overlooked roles in research. The Crossref schema also supports contributor information to be captured for peer reviews, and the availability of this information can be used to identify undesirable review practices such as authors conducting peer reviews of their own work.

Proper attribution and detailed information add valuable context to the research. This helps readers understand the background and expertise of the researchers involved, enhancing the work's credibility. It ensures that all contributors receive due recognition for their work, which is particularly important in multidisciplinary research where multiple authors may contribute in various ways. This fosters a culture of transparency and fairness, motivating researchers by making their contributions visible and credited. This bibliographic information allows readers to explore other works authored by the same contributors, providing valuable context about their overall contributions to the scholarly record.

Affiliations

Affiliation metadata, which captures the institution or organisation employing the contributors of the work, plays a pivotal role in maintaining and enhancing research integrity by ensuring a clear and traceable link between the research and the institution. This metadata enables research organisations to track the work their researchers have produced, also allowing funders to monitor the research they have financed, and assisting journals in identifying the affiliations of their authors.

Journals and publishers can leverage affiliation metadata to identify potential conflicts of interest by linking authors to their respective institutions, thus maintaining ethical standards in research publication by transparently disclosing any affiliations that might influence research outcomes.

For this system to work effectively, the metadata needs to include an identifier for the organisation. Both Crossref and DataCite support the use of standardised affiliation identifiers and prefer the Research Organization Registry (ROR). ROR is a collaborative initiative whose operations are supported by DataCite, Crossref, and the California Digital Library. ROR can also denote funding organisations, and this is supported in both schemas. In Crossref, ROR IDs and affiliation metadata for publication records can be nested within the contributor metadata, and for grant records, can be included in both investigator and member metadata, denoting both affiliations and funding organisations. In DataCite, affiliation metadata and ROR IDs can be part of both the creator and contributor properties across all records.

Dates

Dates are a key metadata element that offer a timeline that outlines the lifecycle of research inputs and outputs, and improve research transparency and traceability.

Dates that are supported in the Crossref schema include publication_date (date of publication of an article/preprint, etc.), acceptance_date (date on which the manuscript was accepted for publication), and update_date (date on which an article was updated). Publication_date allows capturing multiple dates to account for online and print versions of the work. Crossref requires dates for most record types (for example, peer review report records require a review_date, which captures the date when a review was published or posted).

In the DataCite schema, a number of controlled attributes, including "Created," "Submitted," "Issued," and "Updated," that define different dates associated with research outputs are included under the "Date" metadata element. For example, the "Created" date provides users with a starting point by indicating the precise moment the research output was created. The date "Submitted" indicates when the data was first uploaded to a repository, marking the beginning of its publication journey and peer review. "Issued" indicates the data's official publication date, and "Updated" offers a timeline of revisions, indicating an ongoing commitment to accuracy and relevance.

Date metadata provides a layer of reliability and accuracy as these dates can be used to assess the timeliness of the data and publication creation and the length of time between changes, and so users can know if they are looking at the most recent version. Date granularity can therefore enhance the trustworthiness of the research object.

Publication year

This indicates the year a research output was officially made available to the public. This critical piece of information has a direct impact on the reliability and validity of research findings because it ensures that the data is presented in the appropriate historical and temporal context. Knowing when the data was published allows users to verify its relevance and applicability to current issues or historical analysis, as well as assess its significance and contributions to the field, especially in disciplines where new discoveries rapidly shift the landscape of knowledge. In other words, accurate publication dates may prevent misinterpretation and the spread of outdated or incorrect findings.

In Crossref, years are included in the date element above, and their API allows for filtering by year. A "Publication Year" element is distinct in the DataCite schema and is also searchable.

Funder, funding, and grant information

Transparent funding references are important for maintaining ethical standards in research. By offering a clear account of the financial and other influences behind a study, it is easier to identify potential biases or conflicts of interest. This openness promotes integrity in the research process.

Funding metadata contains several attributes, providing detailed information about the financial support for a research project as well as other support, such as prizes and the use of equipment or facilities. The metadata can include the names of funding agencies, award numbers/Grant IDs, and a funder identifier that captures the funding agency. The award_number element captures the internal grant number or other public persistent identifier, such as a grant DOI.

Crossref's Grant Linking System (GLS) and DataCite's award DOI support (awardURI) help grants to be persistently linked to research outputs, researchers and institutions. A wide range of types of financial support is supported by the Crossref GLS, such as awards, fellowships, grants, APCs, prizes, use of facilities, or salary awards, which means that the exact source and type of support are made transparent, as well as the monetary value and currency of the award (if relevant), and the projects, languages, and contributors. The unique identifier for a grant registered by the funding body is also used by the publisher in research outputs such as publications, preprints, and other metadata, which establishes a clear connection between funding and its outputs and outcomes.

As mentioned in the Affiliations section, ROR IDs can be included as funder identifiers in both Crossref and DataCite schemas, alongside the established Open Funder Registry

(OFR) IDs. Over time, the OFR, run since 2013 by Crossref with data curation by Elsevier, will merge into ROR as the preferred open identifier and registry of funding organisations.

Collectively, funding information can be used by funders to track the magnitude and impact of the work that they have funded. Researchers obtain important context about who funded the work that they are reading or citing. Publishers are also able to acknowledge funding in their publication metadata with the Grant IDs, and they can also trace who is funding their authors. Funders can see and monitor open acknowledgements of the outcomes of the work they have supported. Overall, funding metadata is critical for the transparency of scholarly funding and all of its outcomes.

Versioning

Versions play an important role in maintaining the integrity and reliability of scholarly records by capturing the progression and transformation of research outputs throughout their lifecycle. A work may undergo several changes and evolve over the course of its lifetime. Different versions of an article include draft, preprint, pending publication, author accepted manuscript, version of record, and updates, corrections or retractions. It is important to show the connection between accepted versions, such as the article version of record and its preprint, to ensure traceability. Version control is important as it enforces clarity in the scholarly record by clearly designating multiple versions of a document and allowing anyone to see the full discourse about a research project over time. DataCite provides Versioning Guidelines, and the NISO group is soon updating its Journal Article Version (JAV) guidance.

Preprint is the version of a paper before it has been peer-reviewed. Preprints have grown to be a critical part of science communications, especially growing during and since the 2020 global COVID-19 outbreak, where speedy communication of research, such as for vaccine studies, was essential. Preprints can be registered with both DataCite and Crossref. In Crossref, some of the relevant elements that are encompassed within this record type are: contributors; posted_date and acceptance_date; funding; access indicators to capture license terms, version numbers and version descriptions (added in their March 2025 schema 5.4 release); and, of course, a list of references.

To establish the connection between a preprint and its published version, the Crossref member is notified of the publication of a subsequent journal article, and the preprint record must be updated using the isPreprintOf relationship type. A clear and explicit link between the two, as well as versions of the preprint pre-journal, enables clarity in the scholarly record and allows anyone to see the relationship between a given work and its preprint. Persistent identifiers for preprints guarantee that the preprints will be a part of the scholarly discourse over time. Their correct linking to the published versions ensures that the scholarly record is complete and well-connected.

Retractions, corrections, and updates

Retractions, corrections, errata, and withdrawals are, in general, a good sign of a healthy scientific and academic ecosystem, iterating and adding to a body of evidence, and correcting and perfecting over time. In Crossref, this information is captured in the Crossmark metadata. The types of updates that can be submitted include retraction, addendum, clarification, correction, corrigendum, erratum, expression of concern, new edition, new version, partial retraction, removal, and withdrawal. Research is not stagnant and is continuously updated as new information comes to light. Existing research is used to inform current practices and is also the basis on which new research is planned. If the underlying research cannot be replicated or is deemed not trustworthy, any future research based on it is likely to be unreliable too. Such unreliable research, if not corrected in the scholarly record, can not only eliminate trust in science but can also lead to unfortunate consequences for the public, for example, in the case of biomedical research. Therefore, it is important to track any updates that happen to an article after its publication, so that the readers reading or citing the article are aware of these updates.

By registering updates, publishers can ensure that the scholarly record stays up to date. Crossref also incorporated the Retraction Watch dataset and partnered with them for its ongoing upkeep. This data, now available with Crossref member metadata, is also used by downstream services and percolates through the research ecosystem. A Crossmark button can be embedded on any website or PDF document to allow readers to see, with a green/amber/red traffic light system, whether the article is current, corrected, or retracted—even if that retraction or correction takes place many years after a user has downloaded and saved the PDF.

Abstract/description

Abstracts refer to the summary of a research object. They are primarily used by researchers to quickly identify the crux of the work and to decide whether it is relevant to their project. Abstracts can be registered as a metadata element for almost all record types (article, preprint, book chapter, conference paper, reports, dissertation, etc.). Openly available abstracts have the advantage of being amenable to text mining, which opens up a whole new avenue for how they can be leveraged as trust markers. Abstracts can be used to find articles that have similarity in their titles and abstracts, for extracting keywords to obtain annotations, and for finding tortured phrases (use of uncommon terms instead of well-established scientific terms). The ability to analyse abstracts at scale can unearth trends: for example, it can help in discovering whether certain phrases have been repeatedly used across articles.

DataCite's description property allows for a structured abstract that outlines the research context, goals, and results. Additionally, it can provide methodological details necessary for replication and validation. Clear and thorough documentation supports scientific rigor and ethical research practices, ensuring the integrity of findings.

Open abstracts and descriptions are very important for the research community to analyse and assess trends and trustworthiness. If they are not available through Crossref and DataCite, i.e. in a structured and scalable way, research integrity sleuths and others resort to scraping and harvesting the web, which is less reliable.

Clinical trials

Crossref members can include in their metadata records clinical trial numbers and the clinical trial registry where the trial is registered. This information is a crucial trust marker for biomedical literature. Having an identifier associated with every clinical trial not only makes it transparent for any conflicts of interest to be identified, but also provides decision-makers in the healthcare sector with information at the level of every individual study. Some journals have also been using Crossmark to link all the publications that reference a single clinical trial registration number. By connecting all the publications related to a particular clinical trial, all the results, methods, protocols, and reviews that arise from any clinical trial can be accessed and are visible to the reader.

References

Crossref members can register references when registering their content. By including Crossref DOIs in the reference lists, members can also link their references and those referenced objects can include the work in their cited-by counts. This allows readers to reach the referenced work by following the links from the reference list. With their schema 5.4 release, members can now also include a record type label for each reference, denoting whether it is a dataset, a blog post, an article, or another type of object.

In DataCite, references are captured by using relationType values such as “Cites” and “References,” which denote that the current resource is citing or referring to another work.

Registering references means that the list of references included in a research output is submitted as part of the metadata. References direct readers to the source of the content in the work, enabling them to know the influences and provenance of the work. References also provide wider context around how a work fits with and is related to (or adds to) the work that already exists in a given field of research. References can be used to evaluate whether published works are being cited correctly in the appropriate context, and to measure self-citations, mis-citations, and citation manipulation.

Peer reviews

Peer reviews, including referee reports, decision letters, author responses, and post-publication reviews, provide a key integrity check and can be registered with Crossref and DataCite as separate entities which are then linked through relationship metadata. In Crossref, the key metadata elements related to peer review that are supported in their schema are: contributor (captures the name and role of the reviewer), review date (captures the date when a review was published or posted), institution (organisation submitting the peer review), competing_interest_statement (competing interest statement provided by the review author during the review process), and

relationships (captures the link between the review and the work that is being reviewed using the relation isReviewOf*)*. Other key attributes related to peer review that are captured are the stage (whether pre- or post-publication), type, recommendation (minor or major revision, reject, reject with resubmit, accept), revision-round that denotes the revision round number, and language (of the review).

In DataCite, the relationship to the reviewed work is noted in the RelatedIdentifier property with the relation type “Reviews”.

This metadata provides visibility into the peer review process: by linking reviews to the published works, readers can get insights into the editorial and peer review process of the work. This data can also uncover any potential conflict of interest between the reviewers as well as between the reviewers and the authors. Very importantly, the availability of this information can be used to identify undesirable review practices such as authors conducting peer review for their own work, superficial or templated review comments, and duplicate review reports.

Publisher/steward

Stewardship of a metadata record is critical information so that downstream users can see which organisation(s) have taken responsibility for asserting the accuracy of the metadata as facts, and therefore assess the trustworthiness of the work.

The "Publisher" metadata element in the DataCite schema is critical for ensuring research integrity by verifying the origin and authenticity of research outputs. This component is used for determining the provenance of a research output, ensuring that each piece of research can be traced back to a credible source. Including publisher information in metadata improves the visibility of research outputs and encourages proper citation practices. Moreover, accurate publisher information indicates compliance with industry standards and archival practices, which is essential for the preservation and long-term accessibility of research outputs.

In Crossref, the publisher-name can be asserted and gleaned from metadata at the work, prefix, and member levels. Since the schema was initially focused on journal publisher content, there was a presumed 1:1 correlation between member name and publisher name, but as things have evolved to include many more types of members, the publisher name is not always accurate (it could actually show a funder as ‘publisher’ of a grant record, for example). It’s limiting, so a more general representation of the stewardship of a Crossref metadata record is planned in the future; the prefix will equal ‘steward’, and a separate field for publisher name will be added to all record types that don’t currently support it.

Record types/resource types

Documenting resource types ensures that all components of the research, from datasets to instruments, are accounted for and traceable to their origin. This traceability is crucial for verifying research results and ensuring the reliability of the findings. In DataCite, the “resourceType” field has two components: a required “resourceTypeGeneral” attribute with controlled list values and an optional free text description of the resource type. Accurate metadata for resource types allows future researchers to validate results by accessing the same resources, thus supporting the integrity of the research. For example, even if physical samples degrade or are lost over time, digital metadata preserves the essential information, enabling future research and verification. Sharing detailed metadata about resources also promotes transparency and openness in research, making it easier for other researchers to build on existing work.

Crossref calls resource types “record types”, and its schema covers types such as journal-article, grant, book, book chapter, conference proceeding, conference paper, report, dissertation, working paper, component, and some of the previously mentioned types like reviews and preprints. Members decide which record type or subtype they feel matches their type of objects best; e.g. blogs tend to be registered under posted-content->other, and the dataset type tends to be used for datasets, but also photographs and sometimes protocols. The record types should become more granular in the future and be mapped with other schemas in use from other initiatives.

Alternate identifier

The "alternateIdentifier" field in the DataCite Metadata Schema plays a central role. This field specifically addresses the need to manage multiple identifiers that a single research output might accumulate over its lifecycle. This linkage ensures that all manifestations of a work are traceable back to a common source, enhancing the reliability of data. Consider a scenario involving geological samples that are part of a larger environmental study. Each sample might be catalogued under a primary identifier when first collected, then receive additional identifiers as it is analysed by different labs or referenced in various publications. By using the "alternateIdentifiers" field, all these identifiers can be linked together in the metadata record. This ensures that any future researcher looking into this environmental study can access a comprehensive dataset, tracing each sample's journey through different analyses and references, thereby reinforcing the integrity and utility of the research data.

It also minimises the risk of duplicating records. The "alternateIdentifiers" field helps to mitigate this by providing a clear, unique reference system that distinguishes between different versions or records of the same work.

Relationships/related identifier

Metadata is not only useful to completely and accurately describe a specific research object, but it can also be used to establish links between different objects. If we take a research article as an example, an important way to assess the integrity of the article is to look at the other objects and entities that have been used to come to the conclusions described in the article.

The “relatedIdentifier” field in the DataCite Metadata Schema facilitates the creation of a network of linked research outputs, significantly benefiting the research ecosystem. This element fosters the continuation and expansion of research efforts by establishing connections between datasets, articles, supplemental materials, references and other resources. It allows researchers to validate the data and methodologies of studies they wish to replicate or verify, providing a solid foundation for research reproducibility. This approach promotes transparency, reproducibility, and accessibility in academia by ensuring that replication efforts are based on properly linked works and tools. For example, in a primary research article about the effects of climate change on marine biodiversity, this field could link to studies on the economic impacts of biodiversity loss, articles on marine species adaptations, software tools used for analysing datasets, and related datasets about ocean temperatures. Each link is properly labelled to clarify the nature of the relationship, which adds transparency to the process and increases trust. The result is an interconnected network of identifiers that enhances the article's value, providing researchers and policymakers with a comprehensive understanding of the relevant research landscape. The image below shows an example of an instrument record that is linked via DataCite’s relatedIdentifier element to other objects and entities:

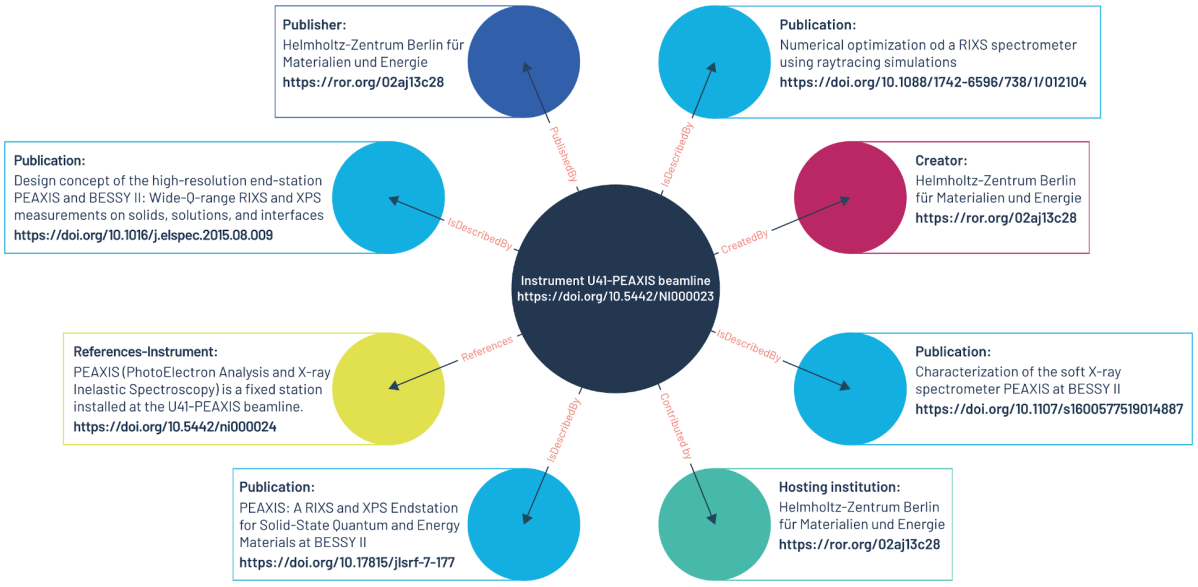

***Figure 2**: A DataCite record for an instrument, linked with related objects [Credit: S. El-Gebali]*

Relationships are also established within Crossref metadata, and metadata matching for enrichment is a key focus. Many members supply citations with their metadata records. A citation defines a relationship between the citing item and the item being cited. Crossref also connects items with corrections, updates, retractions, and clinical trials (using Crossmark), as well as grants and funding with funding metadata and funder identifiers. Registering component records for figures, tables, and supplemental materials automatically relates those items to the parent article.

Relationship-specific metadata can establish explicit relationships with objects that have a registered DOI. Other identifiers are also supported, including PMID, accession number, and URI. For example, a journal article can be linked to a dataset used in the research phase, or a preprint can be linked to a journal article. Relationships can link translated to original language content (isTranslationOf / hasTranslation), link versions to each other (isVersionOf / hasVersion), connect reviews to the works being reviewed (isReviewOf / hasReview), connect a research protocol to an article about the research (isRelatedMaterial / hasRelatedMaterial), and link grants to the outputs of their funding (finances/isFinancedBy).

When a relationship with a Crossref DOI is supplied, reciprocal relationship metadata will be inserted into the DOI record of the related item, noting that the relationship has been asserted by Crossref.

Rights, licences, and access indicators

Rights, license, and access information specify the legal permissions and restrictions associated with the reuse of a research output. Ensuring compliance with licensing terms is helpful for maintaining ethical standards in research. Detailed licensing information clarifies how the work can be used, shared, and cited, thereby supporting the ethical dissemination and reuse of research.

In Crossref, the license and access indicators can include a license-url, such as a Creative Commons or member-specific one, and also access indicators such as free-to-read with start and end dates, and for particular uses it applies to, such as tdm (text and data mining), vor (version of record), or am (accepted manuscript). Including license information (or access indicators) in your deposit is very helpful in letting readers know how they can access and use your content, for example, in text and data mining.

In DataCite, the Rights property similarly allows for a rightsURI to connect to a specific license, as well as a free-text statement for any rights information for the resource.

Limitations

Metadata quality and completeness

Thousands of organisations from hundreds of countries around the world are involved in making the open scholarly infrastructure work through Crossref and DataCite. Everyone has to play their part by stewarding the metadata for the objects and entities they are

responsible for and have the rights to. Metadata quality is a concern for all these organisations, and ensuring metadata creation and completeness is an ongoing effort, often involving multiple vendors for XML creation.

As the metadata flows throughout the many thousands of other users downstream (the aforementioned discovery engines and assessment tools), gaps and inaccuracies can be highlighted. Gaps in metadata records mean that the work isn't getting the same visibility and chance for being used as those records with more complete metadata. As well as metadata for research integrity, discoverability is a key purpose too, as is evidence and provenance for future work and generations to come.

No one metadata record at either DataCite or Crossref can be 100% complete, but these records are enriched and updated many times more frequently than new records are registered, so it's clear that, in general, the community has the intention to do as diligent a job as possible, knowing that improvements and adoption of new metadata standards can always be made.

Accessibility

Not everyone has the resources to join Crossref and DataCite, and they are missing out on the high visibility provided by including all the metadata elements described in this guide, such as citation networks, reference linking, and connections to related objects and identifiers. So, many high-quality, valuable resources, such as journals and datasets, are not yet included in the global open metadata network provided by the two infrastructure organisations.

Both organisations have programs to encourage as broad participation as possible and to lower barriers to entry, but adoption also depends heavily on local governments and university systems to support local infrastructure that connects with our global infrastructure.

Lack of integration

As this guide shows, Crossref and DataCite, whilst both DOI Registration Agencies (RAs), have separate systems and schemas, and handle almost all the same resource/record types in different ways. There is a risk of duplication because the two systems don't read from or write to each other, and other community services (e.g., DSpace, Open Journal Systems (OJS)) must build duplicate integrations across multiple RA systems. Both organisations recognise that this can be confusing and isn't ideal for our shared participants, and hope to explore collaboration in the future, such as building and matching across each other's metadata registries, with a mapping of relationship types planned for the next guide in this series.

Not everyone follows best practices

Whilst the overwhelming majority of Crossref and DataCite members act in good faith, some organisations register DOIs and metadata in ways that undermine the integrity of the scholarly record—whether through negligence, poor practice, or deliberate misuse. Crossref's ISR blog series, part 2, sets out how membership obligations address these cases: the board has authority to suspend or revoke membership where standards are not met, and in March 2025, voted to update the bylaws to clarify what 'for cause' revocation means in practice.

Suspension or revocation remains an absolute last resort—Crossref's first response is always to work directly with the member. A short-term Member Practices Working Group is meeting in the first half of 2026 to draft community-informed practices that guide member behaviour and revocation decisions. DataCite similarly emphasises rigour in metadata deposit and member conduct. These mechanisms aren't a guarantee that all metadata is reliable, but they reflect a shared commitment to trustworthy infrastructure—and the community is encouraged to report concerns via the Crossref forum's metadata quality channel.

The existence of a DOI does not guarantee trustworthiness

A DOI is a persistent identifier: it reliably resolves to a resource and anchors a metadata record. It is not, however, a quality mark. Any Crossref or DataCite member can register a DOI for any research object they steward, and the presence of a DOI says nothing about the soundness of the underlying research, the accuracy of the metadata, or the legitimacy of the research or publishing processes. The metadata elements described throughout this guide—contributors, affiliations, funding, peer review records, retractions, and relationships—are what provide the evidence trail the community needs to make those assessments.

Rich metadata supports research integrity; it does not confer it. A record with sparse or inaccurate metadata may still carry a DOI, as may one belonging to a publication of questionable standing. Downstream users should therefore treat DOI metadata as evidence to be weighed, not a stamp of approval—the judgement of trustworthiness ultimately remains with the community.

Call for actions

Learn more about open infrastructure and metadata

The open scholarly infrastructures, such as DataCite and Crossref, underpin these metadata services and operate discreetly, benefiting the research community without necessarily their explicit awareness. However, the significant role of DOI record metadata in contributing to and upholding the integrity of scholarship can be better recognized and utilized when the community is better informed about its organization and implementation. Both DataCite and Crossref provide comprehensive documentation and encourage exploration of the metadata schemas they use to uniquely identify scholarly works and ensure interoperability with various infrastructure systems.

Contribute more accurate metadata—enrich your records

The metadata record for research entities results from the combined efforts of researchers, service integrators, publishers, and curators. Researchers provide accurate, descriptive metadata when submitting outputs for sharing or publication, service integrators connect the metadata infrastructure with the tools and platforms researchers use, and publishers and curators incorporate as much metadata as possible when registering DOI records. Improving these processes enhances the quantity and quality of trustworthy metadata, leading to greater transparency and integrity in scholarly records.

Use existing metadata for investigation—API queries/notebooks

The existing metadata records hosted by the identifier infrastructure already provide a rich, persistent resource for insights into global scholarly activities, and scrutiny by the metascience community is welcome. Both Crossref and DataCite metadata are accessible via APIs and can be integrated into analytic environments, such as computational notebooks, for further analysis.

Report inconsistencies in the metadata

The open nature of DOI metadata allows for easy tracing and decoding of embedded information through the DOI string. Running a DOI prefix through the resolver on doi.org reveals information about the registration agency, and DataCite and Crossref's open APIs can be used to access the full metadata record, which contains provenance information, making it easy for reviewers to report inconsistencies to the metadata steward. Additionally, both Crossref and DataCite offer various support channels, such as the Research Integrity board on the Crossref forum, for reporting any inconsistencies in metadata.

Conclusion

To enhance the integrity of the scholarly record and safeguard scientific conduct and outputs, the continuous improvement and maintenance of the scholarly infrastructure supporting the creation and consumption of open metadata are essential. Metadata is a vital source of authoritative information on scholarly resources, containing rich contextual data that can act as trust signals for integrity assessment and investigation. Seamless integration across technologies and collaboration among organizations ensure the creation and oversight of metadata by stewards of sharing, publishing, and archiving processes. The infrastructure relies on community actors to unlock its potential: publishers, data stewards, and curators who oversee metadata quality; system integrators and tool providers who develop software solutions for metadata submission and harvesting; and meta-science researchers and investigators who utilize the rich insights provided by open metadata.