Copyright Clearance Center | RightsDirect

Whitepaper

Meeting the Challenges and Opportunities in Scholarly Communications with FAIR Data

During a time of thorough transformation towards Open Access and, moreover, Open Science, the field of scholarly communications is facing new challenges. Some of the challenges are related to an aging and somewhat antiquated infrastructure, and others are related to new technologies and the evolving landscape of scientific dissemination, where the disclosure of data is becoming central to and an essential part of research and reproducibility.

In this white paper, we highlight three use cases in scholarly communications that need data, and particularly FAIR data considering these developments. The use cases are

- Artificial intelligence (AI).
- Research integrity.
- Researcher affiliation identification and disambiguation.

Artificial Intelligence

AI poses new questions with the emergence of generative AI, in particular, for text generation and other language processing tasks based on Large Language Models (LLMs). AI in other forms has been used for decades, albeit in more focused and specific environments. We are now at a point where we need to consider what these new technologies mean for publishers and stakeholders in scholarly communications.

One potential opportunity presented by LLMs is that researchers, not just those in data science, can utilize LLMs to help with their work, so they need to be sure these tools will work for them as intended.

Let's look at the applications of AI that will enable efficiencies, discovery, and content creation.

- Publishers are experimenting by feeding their own content into LLMs to automatically generate abstracts and summaries to make content more discoverable.
- Services are utilizing AI to power search and discovery, refining search results to get more relevant answers to questions, freeing up researcher time from sifting through an overwhelming amount of content.
- Researchers can, at least for some publishers, write manuscripts with AI assistance, provided the use of AI is disclosed within the manuscript. AI is also being used to assist researchers with language checking and technical manuscript compliance.
- Service providers are utilizing AI to write plain language summaries to assist researchers in cross-disciplinary areas and help authors promote their work more widely.

And what of the challenges?

- There is currently a lack of standards around the use of AI and the data it consumes.
- We need a better understanding of the source of AI-generated content, the provenance and trustworthiness of input data being critical, to avoid garbage in, garbage out (GIGO).
- We need to enable AI systems to be trained on curated, high-quality datasets, which adhere to copyright and licensing, to demonstrate the provenance of the data used by AI, and substantiate their corresponding outputs.
- We need to ensure that those doing the training of AI understand the datasets they are using, and the provenance of the data, requiring data scientists and new skill sets.

How FAIR data helps

So, what has this got to do with FAIR? Inherent to FAIR principles is the requirement that machines can Find, Access, Interoperate with and Reuse data. This applies to research datasets, the metadata attached to them, and of course the metadata attached to the content that they relate to.

Findable
The data and metadata are assigned globally unique and persistent IDs (PIDs) which machines can read to locate and differentiate data points. This should include PIDs for the content items themselves, the researchers, their organization affiliations, and the funder at minimum. The metadata should be rich so that machines can find information about the content, and it should be registered in a machine searchable resource.

Accessible
AI will require access to the data, it should not be encumbered by utilizing a non-standard or proprietary communications protocol. In other words, it needs to be retrievable via a standard and open technology such as HTTPS or FTPS. Where the data is protected for privacy or licensing reasons, it should be available through authentication methods in a way that machines can automatically execute the requirements for access.

Interoperable
The metadata needs to be understandable to the machine using a formal and broadly applicable language, self-describing schemas such as XML provide this. Data and metadata should utilize vocabularies and ontologies that clearly and consistently describe the information and there should be as many cross references to other metadata and PIDs as possible. This provides context and valuable information about the data and content being consumed by the AI.

Reusable
The data and metadata should provide a wide range of relevant attributes for the AI to determine the usefulness to the context. Any license or legal rights to reuse the data should be clearly available to the AI system to ensure compliance. Data should have clear provenance: where it came from, how it was developed, and who generated it, in a machine-readable form. Data and metadata should follow standards, ideally open standards, wherever possible.

> “The principles emphasize machine-actionability (i.e., the capacity of computational systems to find, access, interoperate, and reuse data with none or minimal human intervention) because humans increasingly rely on computational support to deal with data as a result of the increase in volume, complexity, and creation speed of data.”
>
> https://www.go-fair.org/fair-principles/

All of these things make data and metadata AI ready, they are discoverable, consumable, and given context in a form that AI systems can reuse and understand. What FAIR doesn't address is the quality of the data and content being consumed by the AI – the garbage in, garbage out (GIGO) problem. However, the metadata attached to data and content provide context that enables those training the AI to understand the provenance of the information, the context of its creation, the terms under which it can be used (or not) in the training input for the AI system. FAIR will not remove the GIGO issue, but it does at least help us understand how to try and avoid it.

Scholarly communications sits on a wealth of data that is trusted, peer reviewed, and verified. For some use cases, this is the data that should be used to train AI under the appropriate licensing terms.

> "
>
> In order to establish trust in AI systems, their developers should provide provenance, license, security, and other useful information in a transparent manner through an open standard so that the consumers of these systems can ascertain what data have been used in building the model. An AI system could be trained on a set of data stemming from many sources. It is imperative that a method for establishing trust and security in AI systems should be developed and the FAIR-ification of the training data will be essential for that purpose."
>
> **Haralambos Marmanis**
>
> https://www.copyright.com/blog/unlocking-the-power-of-fair-data-building-trust-and-success-in-the-ai-era/

Research Integrity

For our second use case, we will look at maintaining the integrity of research, an area that is greatly concerning publishers at present. The rise of paper mills, image manipulation, peer-review rings and other fraudulent activity has the potential to erode trust in the scholarly record and scientific process itself while also negatively impacting publishers. There are clearly concerns that generative AI could exacerbate this situation:

> Fueled in part by AI's lightning-speed generation of fraudulent content, scholars, learners, and other stakeholders will expect providers in this sector to raise their game when it comes to maintaining research integrity and delivering quality publications.
>
> Kate Worlock
> Outsell, Inc.

How FAIR data helps

So how does FAIR data assist with discovering research integrity issues? FAIRifying data doesn't solve the problems with research integrity by itself, but it does provide the ability to track and find patterns in data and give context to the mass of information needed to sift through to address the problems publishers are presented with.

If we take the first principle of FAIR that a given piece of research including its data are given PIDs, we are not preventing unethical behavior. Attaching metadata such as a DOI, an ORCID, or an ISNI ID does not inhibit someone from favorably reviewing papers in a peer-review ring. It doesn't prevent someone from falsifying data. It also doesn't stop them from using generative AI to entirely create the research and the article, or indeed to falsify citations. The presence of PIDs in and of themselves, does not impart trustworthiness on the entity they are identifying.

However, if we consider that one of the key pieces to flagging possibly dubious research lies with machine detection, the FAIR data principles ensuring machine readability as described above (Findable, Accessible, Interoperable, Reusable) apply here also. FAIR data cannot prevent research fraud, but by mapping patterns and tracking connections, particularly against retracted works or expressions of concern, we are able to use the metadata and PIDs in particular, to locate potential problem areas to be investigated. PIDs provide a consistent reference point for people, places, and things. By providing the provenance of data and information we learn more about it en-masse, who works with whom, who reviews whom, who works where, what are they working on (or not), we see more in the mass of metadata than we do in the metadata for individual pieces of research.

We note that NISO recently released a draft version of the Communication of Retractions, Removals, and Expressions of Concern (CREC) Recommended Practice for public comment. Best practices and standards for reporting research integrity issues are extremely welcome and will work very well with FAIR data. We gain the ability to consistently look for that data in context, and patterns in behavior across the published corpus.

PIDs can help with building pattern matching. For example, if organization PIDs (such as the Ringgold ID) are applied at a granular level of the organizational hierarchy, and if researcher IDs are missing in the metadata, the granularity of the specific organizational context helps to more accurately disambiguate one researcher from the other. We can see specifically where people are based, who and which organizations are working together to help, alongside other metadata, in developing the patterns that publishers can look to as part of the information they need.

Researcher affiliation identification and disambiguation

For our third use case, we look to the ongoing problem of researcher affiliations identification. Researcher affiliations are a central part to both topics above, again not in isolation, but of great importance.

The identification of researcher affiliations, through FAIR metadata and PIDs, provide us with one of the key points of the provenance of research — where it was undertaken. As we have seen above, the more granular the information held here, the more we understand provenance, which also enables a higher degree of disambiguation of the researchers themselves. While of course PIDs exist for researchers and are often used, a recent analysis of MEDLINE metadata found that only 3% of author data held an associated ORCID ID. This rises to just under 17% in 2021 (still a pretty low number)."[1] The granularity of affiliation data helps to disambiguate researchers within a field from one another where names are the same, and particularly when first names are abbreviated to letters.

How FAIR data helps

As with the above use cases, FAIR data for the identification and disambiguation of researchers also aids discoverability, not only of threads of research undertaken by certain people or teams, but affiliations also identify centers of research in a particular discipline.

Research affiliation data also enables the shift to Open Access, in that it becomes easier to know and communicate to authors that they are entitled to funding or discounts for publication charges. As publishers structure these OA agreements for their journals, the need for granularity in affiliations is a requirement. We know that publishers need to understand whether the author's affiliation, within a particular part of an institution is included in the agreement and communicate that efficiently to authors.

Accurately asserted affiliations using FAIR data enables publishers to run machine modeling and analytics against prior publication metadata to confidently structure and negotiate their OA agreements. It also enables them to provide both granular and collective reporting to institutions.

Funders using granular FAIR data on affiliations can provide research impact reports back to institutions, both at the institutional and departmental level, as is performed by the Portuguese Government body, Fundação para a Ciência e Tecnologia, using the Ringgold Data.

> Before adopting the Ringgold Identify Database, our team spent far too much time going back and forth with the institution curation and validation to make sure we had accurate data... We selected Ringgold above all other PID options because its comprehensive and curated data set is critical in reducing our administrative burden while improving service to our researchers."
>
> PLOS

Conclusion

In all three of the use cases discussed here, FAIR data and metadata are key enabling factors and are vital to facilitate the use of technologies to solve these problems and take advantage of opportunities.

As a key part of the FAIR metadata attached to research data and content, researcher affiliations serve not only the path to Open Access but provide a critical component of the FAIR data needed to support Artificial Intelligence use cases and the understanding of what is used as input to train AI. FAIR data including metadata about organizations, are vital in the tracking of patterns, identifying potential research integrity issues, and maintaining the scholarly record. Without assumptions on what metadata will be required by people and machines, and the more granular the FAIR data are, the more new use cases are enabled. Using FAIR data and PIDs also enables research reproducibility, it enables the continuation of a research theme. Where metadata are FAIR, and provide plenty of information about the research data made available, researchers can understand the context of the data, its creation and relation to other information. Where the research data itself is provided in a FAIR manner, it enables other researchers to further explore and reuse the data, extract additional information and build upon findings to create new knowledge.

While FAIR is the enabling factor in these use cases, it does not solve all the problems in and of itself. The FAIR Principles do not address data quality, they do not address privacy, security, or sustainability. Each of these require another set of principles and frameworks. Data quality frameworks about the dimensions of the data should also be addressed to understand the utility and validity of data. Compliance frameworks should be part of the data structure to understand the privacy risks and restrictions associated with the data. Economic frameworks ensure the sustainability of the data. With all of these things put together with the facilitating concept of FAIR data, new uses of data are enabled regardless of whether they have been devised yet.

References

[1] https://doi.org/10.1093/database/baad070

About CCC
A pioneer in voluntary collective licensing, CCC advances copyright, accelerates knowledge, and powers innovation. With expertise in copyright, data quality, data analytics, and FAIR data implementations, CCC and its subsidiary RightsDirect collaborate with stakeholders on innovative solutions to harness the power of data and AI.

CCC | RightsDirect

Learn more

Contact us at:

copyright.com/Ringgold

solutions@copyright.com