RESEARCH ARTICLE

Challenges in building scholarly knowledge graphs for research assessment in open science

Paolo Manghi[1,2]

[1]National Research Council Institute of Information Science and Technologies Alessandro Faedo, Pisa, Italy
[2]OpenAIRE A. M. K. E., Marousi Athens, Greece

Keywords: open science, research assessment, research data, research impact, research software, scholarly knowledge graphs

ABSTRACT

Open science has revolutionized scholarly communication and research assessment by introducing research data and software as first-class citizens. Scholarly knowledge graphs (SKGs) are expected to play a crucial role in generating research assessment indicators being able to aggregate bibliographic metadata records and semantic relationships describing all research products and their links (e.g., citations, affiliations, funding). However, the rapid advance of open science has led to publication workflows that do not adequately support and guarantee the authenticity of products and metadata quality required for research assessment. Additionally, the heterogeneity of research communities and the multitude of data sources and exchange formats complicate the provision of consistent and stable SKGs. This work builds upon the experience gained from pioneering and addressing these challenges in the OpenAIRE Graph SKG. The aim is twofold and broader. First, we identify obstacles to the creation of SKGs for research assessment caused by the state-of-the-art publishing workflows for publications, software, and data. Second, we describe repurposing SKGs as tools to monitor such workflows to identify and heal their shortcomings, taking advantage of tools, techniques, and practices that support the actors involved, namely research communities, scientists, organizations, data source providers, and SKG providers, to improve the Open Science scholarly publishing ecosystem.

1. INTRODUCTION

Research assessment is the evaluation process that serves to determine the significance of scientific endeavors and understand their impact on the academic community and society. By assessing research, organizations, funders, and governments can make informed decisions about resource allocation, recognition of scientific contributions, and support for future research. In addition, research assessment helps determine researchers' reputation and recognition within the scientific community, playing a crucial role in shaping their careers.

In traditional *research assessment*, the focus is primarily on evaluating the impact of research in scientific, societal, and economic spheres, with the evidence based on *research publications* of specific *resource types*, like articles, chapters, monographs, books, and preprints. These publications, particularly those disseminated through "recognized" *publishing venues*, such as journals, conferences, and preprint servers, are considered representative of scientific output. Assessment is performed via metrics and indicators representing different facets of impact (e.g., citation counts, impact factors, social media mention counts, number

an open access journal

Citation: Manghi, P. (2024). Challenges in building scholarly knowledge graphs for research assessment in open science. *Quantitative Science Studies*, *5*(4), 991–1021. https://doi.org/10.1162/qss_a_00322

DOI: https://doi.org/10.1162/qss_a_00322

Peer Review: https://www.webofscience.com/api/gateway/wos/peer-review/10.1162/qss_a_00322

Received: 3 January 2024
Accepted: 4 June 2024

Corresponding Author:
Paolo Manghi
paolo.manghi@isti.cnr.it

Handling Editor:
Rodrigo Costas

The MIT Press

of patents, and visualization). The production and reliability of such numbers derives from the authoritative bibliographic information gathered into *scholarly knowledge graphs* (SKGs) (Aryani, Fenner et al., 2020; Manghi, Mannocci et al., 2021; Peng, Xia et al., 2023). SKGs are data structures modeling scholarly knowledge, representing metadata of entities, concepts, and relationships between them, for scientific information across disciplines and publishing venues. Many kinds of SKGs exist (Peng et al., 2023), targeting applications such as discovery, scientific reasoning, reproducibility, and research analysis; for example, PubGraph in Ahrabian, Du et al. (2023); OKRG in Brack, Hoppe et al. (2020); and SciKGraph in Tosi and Dos Reis (2021).

In this paper we focus on SKGs specifically conceived to address research assessment, such as Dimensions, Scopus, Web of Science, OpenAlex, OpenCitations, and the OpenAIRE Graph. Such SKGs are graph-shaped databases obtained by aggregating, enriching (e.g., AI, FTM, web crawling), and deduplicating metadata information from data sources and from research entity registries like ORCID.org, Crossref.org, and ROR.org. In this context, data sources are intended as services that store and provide access (via APIs and/or web applications) to the metadata and files of research publications submitted to a given publishing venue. As illustrated in Figure 1, researchers submit a publication to a publishing venue, which preserves its metadata and files in a data source of reference—note that, the same data source can serve multiple (even thousands of) venues (e.g., journals and publisher archives).

In the last decade, as a reaction to the holistic approach of open science, pushing for FAIR (Findability, Accessibility, Interoperability, and Reuse) and open access, and the open science mandates from funders worldwide, the scientific community showed growing interest in publishing research products beyond the traditional publications and considering open access publishing venues, such as institutional repositories, catch-all repositories, and data repositories. Scientists extended the range of outcomes to other kinds of research publications, such as preprints (the Green Road), reviews, presentations, deliverables, and reports, and to other classes of products, like research data and research software. The growing magnitude of the

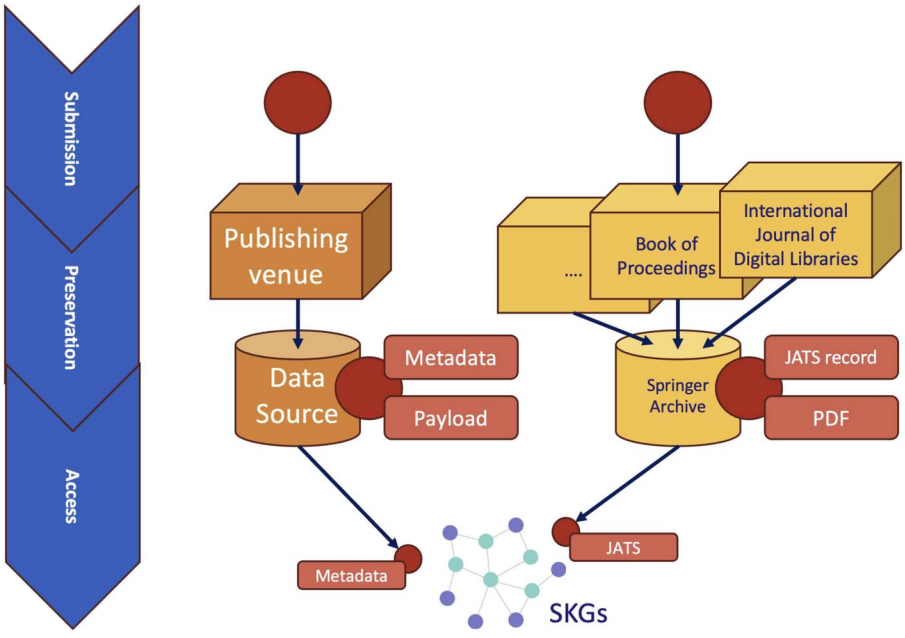

Figure 1. Publishing venues: the traditional article publishing workflow.

change is supported by evidence, with more than 600,000 ORCID researcher profiles reporting at least one research data object in their CV. These facts, combined with the huge investments from funders and companies (€400 million from the European Commission in 2020), led to a natural process of reform of research assessment (CoARA agreement on reforming research assessment[1], Evaluation of research proposals: The why and what of the ERC's recent changes[2], Towards global principles for Open Science monitoring[3], Open Science Monitoring Initiative[4]). On the one hand, new indicators[5] are required to assess scientific CVs, which include extra facets of scientific work, such as data contributions, software contributions, scientific dissemination (e.g., education and presentations), collaboration activities (e.g., reviews and reports of working groups), and industrial innovation (e.g., patents). On the other hand, for open science to be effective and worth its cost, open science research practices must be monitored and assessed, to "establish a baseline, provide punctual support to scientists, monitor compliance to policies, and understand their effects"[6].

In this rapidly evolving scenario, research assessment entails a holistic approach to support the evaluation of multifaceted scientific CVs and the evaluation of open science practices, from open access, open source, and reproducibility of science. Scientists, funders, and organizations are interested in monitoring and assessing the return of investments of the human efforts and investments underpinning open science. New questions (and combinations of those) arise, such as "How many research data sets and software have been produced by researchers of the CNRS research center?," "What is the impact of the dissemination activities of Amir Subramanian?," "What is the impact of the research data published by Laura Rossi?," "What is the rate of open access publications and data funded by the Italian Ministry of Research under open access mandates?," and "What are the trends of linking scientific articles to research data and research software in cultural heritage?".

In response to such questions, some SKGs have started to extend their data models to represent the aforementioned research products and to consider as eligible new typologies of publishing venues. Examples are the OpenAIRE Graph in Europe, the ResearchGraph.org initiative in Australia, and research.fi in Finland. Novel metrics and indicators have been conceived to assess open science practices and to assess scientific quality in this richer domain (Schomberg, Britt Holbrook et al., 2019). Expectations are high because policy directives mandating data and software publishing and adherence to FAIR principles have positively influenced the scientific community, and scientists can rely on a variety of tools, data and software repositories, archives, and databases, operating at the institutional, thematic, and cross-community level, to share, find, and reuse such products.

Despite its seemingly smooth façade, this scenario poses two main questions and challenges for SKG providers and consumers. The first is "Which products should be regarded for research assessment?" Determining the scope of research assessment is no longer an obvious task and highly depends on the context of the evaluation and on the ability to identify such products. Publications, research data, and software are available through thousands of venues

[1] https://coara.eu/agreement/the-agreement-full-text/.
[2] https://erc.europa.eu/sites/default/files/2024-02/Evaluation_of_research_proposals-.pdf.
[3] https://www.ouvrirlascience.fr/building-an-open-science-monitoring-framework-with-open-technologies-unesco-workshop-19-12-23/.
[4] https://open-science-monitoring.org/.
[5] https://handbook.pathos-project.eu/.
[6] Iain Hrynaszkiewicz, Open Research Solutions, presentation at Year of Open Science Culminating Conference, March 2024

and data sources that feature diverse, community-specific resource type vocabularies. Questions like "Which products qualify as research data?" may have different, sometimes conflicting, yet reasonable answers in terms of their significance to science.

The second question is: "Is this product worth assessment?" For many publishing venues, the practice of *gatekeeping* metadata accuracy and scientific trust of the product is not always supported. For example, peer review, a widely accepted practice for evaluating publication quality, is not applied in Green Open Access and may not be suitable for data (Parsons & Fox, 2013), being replaced by community endorsement or, in some cases, omitted altogether.

Such discrepancies from the past complicate the construction of comprehensive SKGs in open science. This work exposes some of the common and context-related challenges that arise in the construction of such SKGs. As such, it is not intended to offer a survey of SKGs for research assessment in open science, but rather to identify the issues in the current open science publishing workflows that may undermine their construction; hence the establishment of an open and transparent open science research assessment. Specifically, first, the paper identifies the shortcomings in existing scholarly publishing workflows for publications, software, and data that impede the creation of reliable SKGs for research assessment. Second, it repurposes SKGs as means to monitor such workflows, by enabling quality tracking of the data, to identify and heal their shortcomings. SKGs, combined with tools and practices, can establish a virtuous cycle engaging scientists, their organizations, publishing venues, data sources, and scholarly communication infrastructure providers, where SKGs play a pivotal role in open science research assessment but also in the identification and adjustment of anomalies in scholarly publishing practices.

The OpenAIRE Graph[7] (Manghi, Atzori et al., 2023) represents one of the early endeavors in constructing an interdisciplinary and cross-border SKG that encompasses today the largest range of scholarly products, including publications, data, software, and other scholarly outputs. Its data provisioning chain, whose steps are transparently documented at https://graph.openaire.eu/docs, is confronted with various challenges, which we shall refer to as sources of motivation and evidence for our context analysis. The numbers reported in this paper refer to February 2024 and can be verified via the web portals OpenAIRE Explore[8] and OpenAIRE MONITOR[9], via the OpenAIRE Graph APIs[10], or via the Graph data sets in Zenodo.org[11].

1.1. Outline

Section 2 presents research assessment in the traditional setting of research publications, introducing the concept of gatekeeping in terms of metadata trust and scientific trust; Section 3 highlights how open science demands are making this scenario outdated and how gatekeeping has changed when research data and software are involved. Section 4 details how such changes have affected the construction of SKGs for research assessment. Section 5 recommends possible solutions that, using SKGs as the pivot of information exchange between researchers, venues, SKG providers, and their consumers, can facilitate the development of a virtuous and synergic research assessment infrastructure.

[7] https://graph.openaire.eu.
[8] https://explore.openaire.eu.
[9] https://monitor.openaire.eu.
[10] https://graph.openaire.eu/docs/apis/home.
[11] https://graph.openaire.eu/docs/category/downloads.

2. "TRADITIONAL" RESEARCH ASSESSMENT

The traditional scholarly communication infrastructure has long regarded scientific publications, which document experimental findings, as the primary evidence of research impact. Over the years, the scope of research assessment has been shaped by consolidating lists of "reputable" publishing venues, such as the journal directories maintained by Scopus, Web of Science, and SJR, which policymakers and scientists can refer to for assessment and for publishing, respectively. Lately, such listings have admitted thematic and curated thematic preprint and postprint servers (Chaleplioglou & Koulouris, 2023), such as PubMed[12] and ArXiv .org[13], which are increasingly being trusted by scientists as sources of relevant knowledge.

Although several concerns have rightly been raised regarding their governance and their consequence in terms of "publish or perish" (Tennant, 2020), such venue listings establish a stable and widely recognized scope of research assessment, which is ideal for SKG construction. Indeed, they restrict the classes of resource types to an agreed-on and limited set (e.g., articles in journals, conferences, workshops, chapters of books, or monographs); and guarantee that such publications are subject to gatekeeping controls, and hence feature degrees of scientific trust and metadata trust. Scientific trust guarantees that all publications aggregated by the SKG are to be regarded for assessment because the research community trusts the underlying scientific endeavor; this is often established through processes such as "peer review" or community endorsement, seen in platforms like ArXiv.org ("Guild publishing model" by Kling, Spector, & McKim, 2002); note that scientific trust should not be conflated with scientific quality (Haucap, Thomas, & Wohlrabe, 2018), including when peer review is involved (Siler, Lee, & Bero, 2015). Metadata trust guarantees instead that the SKGs can rely on the authoritativeness of the publication metadata, meaning that experts (librarians, publisher operators) and tools ensure completeness and precision of the metadata as provided by the authors under submission, such as correctness of authors, affiliation templates and spelling, appropriateness of the resource type, and usage of PIDs (e.g., ORCID for authors, ROR.org, ISNI for organizations).

Figure 2 illustrates the workflow of SKG construction when relying on "reputable" venues: the process entails an aggregation phase (i.e., metadata harvesting and harmonization), followed by enrichment and deduplication phases (Verma, Bhatia et al., 2023).

As shown in Table 2 (see Section 4.1), harvesting and harmonization tasks are generally affordable for publication-oriented SKGs due to the limited number of data sources involved, uniform resource typing practices across them, and data sources support for bibliographic metadata standards.

The task of SKGs enrichment aims to enrich, complete, and adjust metadata records by extracting, safeguarded by the scientific trust, the full text of publications or related web pages to identify authoritative data, such as title, authors, topics, abstract, and citations. Humans or machines can, at the significant cost of crawling, storing, and mining full texts and related splash pages, extract metadata and references to generate richer SKGs. Citation and affiliation relationships are regarded as particularly important for research assessment and are rarely provided manually during the submission process. For example, the OpenAIRE Graph extracts around 10 million citation relationships from 30 million full texts and feeds them to the OpenCitations index; as a result of this and many other similar activities (Crossref citation index, Heibi, Peroni, & Shotton, 2019), OpenCitations' index publishes an open access collection of 2 billion publication-publication citations. In the past decade, open

[12] https://pubmed.ncbi.nlm.nih.gov.
[13] https://arxiv.org.

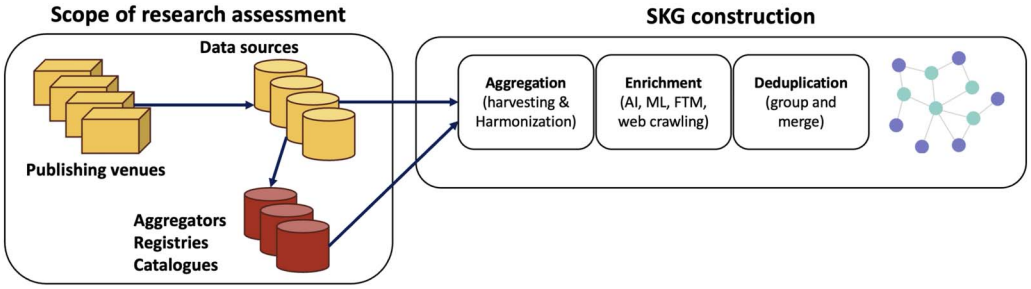

Figure 2. Scope of research assessment and SKG construction phases.

access has mandated facilitated access to full texts, diminishing the commercial edge of full-text availability for nonopen access publishers (Bologna, Iorio et al., 2021).

Given the rich and reliable metadata and limited range of resource types, identifying equivalent publication records is relatively straightforward. Deduplication of authors (Sanyal, Bhowmick, & Das, 2021) and organizations (Ancona, Cerqueti, & Vagnani, 2023) is instead a major challenge, requiring humans in the loop. Indeed, although several authorities exist for researcher and organization identifiers, scientists do not always use them, and many publishing venues still do not support their use in the submission process.

3. RESEARCH ASSESSMENT IN OPEN SCIENCE

Companies such as Google, Elsevier, and Clarivate rely on the selective research assessment scope described above to develop valuable SKGs and services for research evaluation, including Scholar, Scopus, SciVal, AMiner, and Web of Science. In recent years, the limitations of such an ecosystem have become more evident, above all its commercial orientation and toll-gated access to services and data, which hinder its ability to truly benefit the scientific community, and its bias towards the (journal) impact factor and citation counts, which have been criticized for their limitations.

Starting around 20 years ago, open access initiatives and mandates endorsed by policymakers, research communities, and scholarly societies aimed at gradually dismantling this ecosystem. Their recommendations and mandates brought into the picture other models of scientific publishing, such as Green, Bronze, Gold, and Diamond open access; as well as new publishing venues such as institutional, thematic, or catch-all repositories, Open Access journals, and overlay journals. Their recommendations and mandates brought into the picture other models of scientific publishing, including "preprint" Green open access and Brown and Gold open access, and brought into the picture new publishing venues such as institutional, thematic, or catch-all repositories, Open Access journals, and overlay journals, sometimes via transformation agreements.

More recently, open science and its principles of FAIRness, open data, and open software (Aksnes, Langfeldt, & Wouters, 2019) aimed to make scientific publishing even more effective, by fostering and improving reproducibility, collaboration, and transparency to reduce cost and facilitate its reuse and technology transfer. Research funders and institutions are increasingly looking to update their evaluation frameworks to deliver a more equitable assessment of individual researchers and to gain a holistic understanding of the reach of the activities by their grantees and faculty. To address such expectations of a holistic and reproducible science, both scholarly publishing and research assessment opened to "unconventional" products, such as research software and research data, as well as to a variety of unordinary products, traditionally disregarded for research assessment, such as scientific reviews, reports, technical documentation, and

presentations. Such products, key to reproducibility and collaboration, enable the evaluation of other facets of the scientific life cycle, in terms of impact and practices, and in terms of different perspectives of scientific contribution, beyond traditional research impact.

In this new ecosystem, research assessment becomes an overarching process, measuring scientific performance in terms of the more "traditional" research impact (publications, data, software), but also in terms of research support (dissemination, communication, reuse) and research practices (open science conduct, compliance with funders' mandates, FAIRness, transparency). Products are published in a variety of publishing venues and data sources, including institutional repositories, catch-all repositories, data repositories, and software repositories. Adequate multifaceted indicators and metrics that better reflect the complexity of contemporary research practices and diverse scientists' profiles are key to fostering open science principles, enhancing transparency, reproducibility, accessibility, and quality of research, and steering funding for research and innovation. For example, scientists' curricula can be more articulated (Mannocci, Irrera, & Manghi, 2022), highlighting technical contributions (software, services, tools), data manufacturing contributions (Research Data Management, Data Management Plans), beyond the traditional authorship roles (CrediT Ontology[14]). Analyzing the OpenAIRE Graph, around 600,000 researchers with an ORCID identifier (out of 6,518,007 ORCID with at least one product) reportedly produced research data, with almost 3 million data sets bearing at least one ORCID identifier, and another 4.5 million "unique" full names appearing without an ORCID (indicating that many more should be potentially reported to ORCID). For software, numbers are lower, but promising, as software publishing has only recently become a practice.

Broadening the spectrum of venues and products used in research assessment leverages the shift towards a holistic view of scholarly contributions today promoted by funders, ministries, research performing organizations, and research communities, and acknowledges that impactful and reliable research relies on a much broader ecosystem of outputs and practices. However, as we shall highlight in the following, this action disrupts the traditional conditions for research assessment described earlier and burdens the construction of SKGs. In the following sections, we will discuss the vulnerabilities and opportunities of this evolving landscape by describing the scope of research assessment and gatekeeping practices for each type of product (Table 1): publications, research software, and research data.

3.1. Research Publications

3.1.1. Scope of assessment

In the past decade, open science has greatly influenced scientific literature publishing, leading to the adoption of open access practices. The increased availability of open access versions of scientific articles has enhanced accessibility to science, boosted innovation (Bryan & Ozcan, 2021), and overall proved to bring a positive impact on article citations (Eysenbach, 2006). Initiatives like EC open access mandates[15] and cOAlition S (Schiltz, 2018) have successfully encouraged researchers to adhere to open access guidelines, such as Green, Gold, and Platinum open access. As a result, the range of publishing venues eligible for research assessment has expanded beyond traditional journals, conferences, and thematic repositories for preprints and postprints. New entities, like institutional repositories (e.g., HAL.fr) and catch-all repositories (e.g., Zenodo.org, Dryad, Figshare), as well as national and thematic aggregators (e.g.,

[14] https://credit.niso.org/.

[15] https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/open-access_en.htm.

Table 1. Research product types: Scope of assessment and gatekeeping practices

	Scope of assessment		Scientific trust	Metadata trust	
Research product	*Publishing venues*	*Resource types*	*Diffusion*	*Diffusion*	*Inference*
Publications in traditional research assessment	Reputable venues (journals, conferences, thematic repos)	Cross-discipline: articles, chapters, books, monographs	Present for all products	Present for all products	Applicable to publications' full text
Publications in open science publishing	Reputable venues plus institutional/catch-all repositories	Cross-discipline plus other literature types	Absent/ unknown for nonreputable venues	Absent/ unknown for nonreputable venues	Applicable to products in reputable venues
Research software	Institutional/catch-all/ data repositories, few dedicated software venues	Cross-discipline: software, code, tool, workflow	Absent/ unknown for majority of sources	Absent/ unknown for majority of sources	Applicable to software repositories with CFF and codemeta files
Research data	Institutional/catch-all/ data repositories	Discipline-specific: different vocabularies per discipline	Absent/ unknown for majority of sources	Absent/ unknown for majority of sources	Rarely applicable: format diversity and missing biblio information

Recolecta in Spain, LaReferencia in Latin America), have become eligible publishing venues and data sources. Moreover, the push for a holistic research assessment extended the range of resource types to encompass other types of literature, such as technical reports, white papers, presentations, lectures, project deliverables, and theses.

In this context, the interoperability challenge is magnified compared to the traditional settings. There are thousands of data sources (the OpenDOAR.org registry of institutional and thematic repositories and aggregators alone has around 6,000 entries), each obeying potentially different resource type vocabularies, metadata models, and formats. As no common definition of scope of assessment is generally recognized, SKGs tend to differ from each other, typically carving out their own dedicated set of eligible publishing venues, resource types, and target SKG data model, maintaining an aggregation framework that copes with the heterogeneity.

3.1.2. Gatekeeping practices

In this context, the axiom of dealing with "reputable" publishing venues is lost. Venues may support various gatekeeping measures, from stringent requirements imposed by leading publishers to the complete absence of control and trust of catch-all repositories, leaving metadata trust to the researchers' common sense and scientific trust to their good intentions. In the following, we describe in more detail and with examples how these conditions affect SKG construction and, in turn, research assessment.

3.1.2.1. Metadata trust. The inclusion of publishing venues with inadequate metadata trust practices poses a risk to the reliability of metadata in SKGs, which is crucial for research assessment. Authors sometimes make errors when depositing descriptive metadata (e.g., wrong resource type, misspelled title) and attribution metadata (e.g., ORCID, ROR identifiers; Baglioni, Manghi et al., 2021); and metadata can be modified at any time, leaving room for malicious usage to manipulate research assessment (Besançon, Cabanac et al., 2023).

Table 2. SKG construction challenges by research product types

Research product	Aggregation	Enrichment	Deduplication
Publications in traditional research assessment	*Harvesting*: tens of data sources. *Harmonization*: lightweight	Collecting trusted full texts/web pages (via agreements with publishers or open access crawling); inference of metadata and links	By DOI and rich metadata among few resource types
Publications in open science publishing	*Harvesting*: thousands of data sources. *Harmonization*: heterogeneity of resource types, formats, and unreliability of PIDs	Collecting trusted and nongatekeept full texts/web pages (via agreements with publishers or open access crawling); inference of metadata and pub-pub links	Complicated by less reliable and incomplete metadata, and heterogeneous and unreliable resource types
Research software	*Harvesting*: thousands of data sources. *Harmonization*: heterogeneity of resource types, formats, and unreliability and unavailability of PIDs	Infrastructure as for publication in open science; inference of pub-software links into software repositories; collecting software code to infer software metadata from CFF and codemeta files	By software repository URLs or by SoftwareHeritage PIDs, rarely by metadata match due to unreliable or incomplete metadata records
Research data	*Harvesting*: thousands of data sources. *Harmonization*: heterogeneity of resource types, formats, unreliability of PIDs, and granularity	The variety of formats (e.g., tables, PDFs, database entries, GiHub files) requires dedicated solutions to infer metadata and links from payload	By PID or metadata, complicated by unreliable and incomplete metadata

3.1.2.2. Scientific trust. The absence of scientific trust blurs the borders of research assessment, introducing an issue of filtering at the data source level. For example, Figure 3 shows that resource types are not enough to detect "trusted" products: The product is deposited in Zenodo.org as an "article" but it was instead a mere test of the deposition process left undetected because of the lack of support for scientific trust from Zenodo.org[16]. On the other hand, Zenodo.org is also a publishing venue for publications subject to gatekeeping. Figure 4 shows how Zenodo.org may act as a data source for multiple publishing venues. The episciences.org publisher platform for Open Access overlay journals, which ensures both scientific trust and metadata trust, uses Zenodo.org (and HAL.fr) data source to store peer-reviewed articles. Accordingly, when collecting from nongatekeept data sources, SKGs may end up including untrusted products.

3.2. Research Software

3.2.1. Scope of assessment

Recent and ongoing activities in Europe (e.g., SoftwareHeritage.org, FAIRCORE4EOSC project[17], FAIR-IMPACT project[18], Research Data Alliance Interest Group on Software Source Code[19]) aim to shed some light on research software publishing (Gruenpeter et al., 2021) and provide research software archiving services for the European Open Science Cloud[20] in

[16] Zenodo.org supports spam filters, blacklisting unconventional or malicious usage of the repository, such as SEO boost and advertising.
[17] https://faircore4eosc.eu.
[18] https://fair-impact.eu.
[19] https://www.rd-alliance.org/groups/software-source-code-ig.
[20] https://eosc.eu.

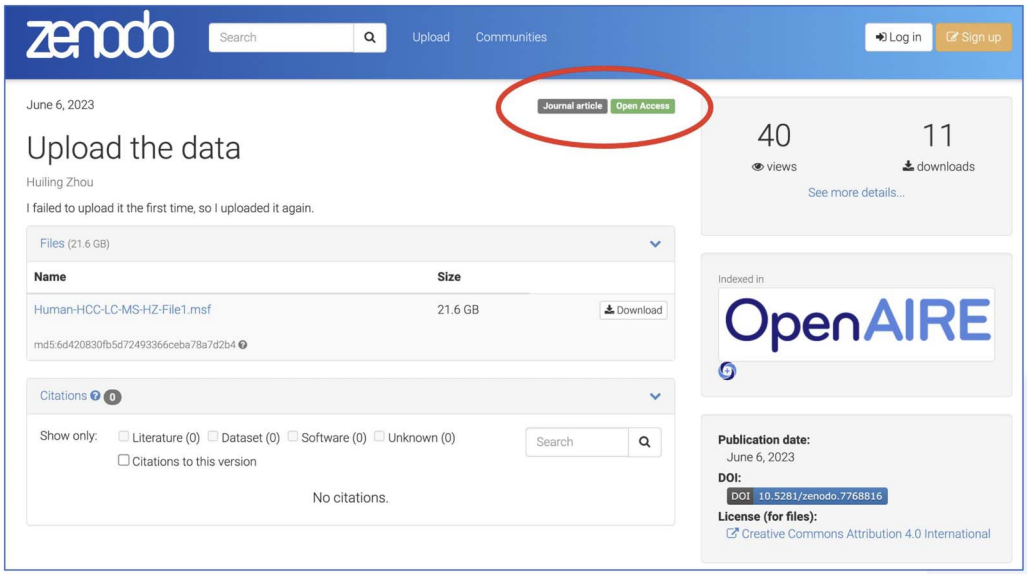

Figure 3. Zenodo.org: Example of research publication escaping gatekeeping practices.

support of it. The focus is on software as "code" and on establishing a common understanding of research software publishing workflows (Barker, Hong et al., 2022; Hong, Katz et al., 2022; Tennant, Agarwal et al., 2020) at global and community levels: research software vs. software for research (Gruenpeter et al., 2021), attribution metadata for software, PIDs for software, software archiving, etc.

As of today, there is no rigorous and widely adopted notion of publishing workflow for research software nor of publishing venue for software products. Indeed, many scientists reckon software versioning platforms to be an ideal publishing venue. Unfortunately, such platforms do not distinguish between "software" and "research software," do not collect

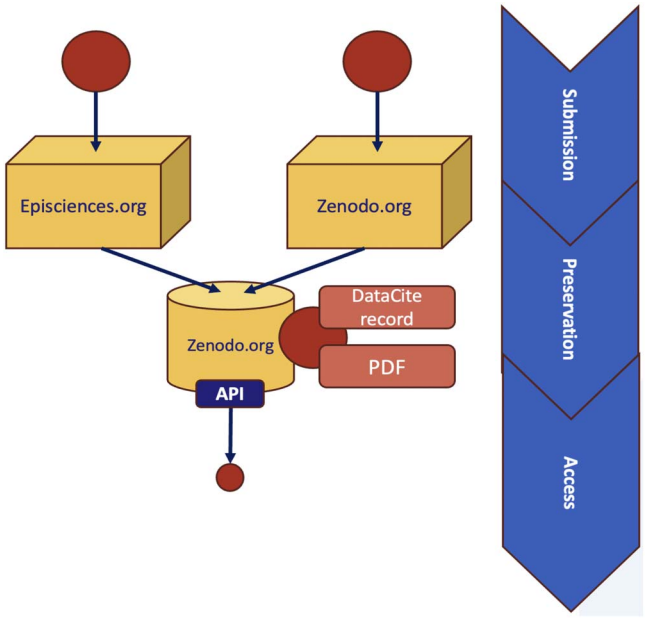

Figure 4. Data source supporting publishing venues with diverse gatekeeping degrees.

bibliographic metadata, and do not ensure preservation, and cannot therefore be considered valuable metadata data sources by SKGs. Exceptions exist, such as ESCAPE's OSSR Repository (Vuillaume, Al-Turany et al., 2023), ELIXIR's bio.tools, CodeOcean.com, DOE Code[21], and the Research Software Directory[22], where research software follows a dedicated publishing workflow, but the majority of research software is published in venues such as data repositories, institutional repositories, and catch-all repositories. The challenge with such sources is to identify research software products. Depending on their discipline-specific or cross-discipline nature, the resource types used for software are not uniform (e.g., "software," "code," "method," "algorithm") and SKGs must cope with the interoperability barrier and its scale, due to the large number of sources.

The lack of well-established publishing practices and the inability to frame a realistic research assessment scope undermines the ability of SKGs to identify and include the full pool of research software scientists produce.

3.2.2. Gatekeeping practices

Gatekeeping practices for research software are not the norm. From a tentative analysis, in the OpenAIRE Graph, only around 20,000 software products, out of 390,000, can be tracked back to gatekeeped venues that offer both metadata trust and scientific trust. In the following, we depict the current maturity of metadata trust and scientific trust.

3.2.2.1. Metadata trust. Software metadata quality, completeness, and trust highly depends on the user's commitment and understanding of scholarly communication. This holds for record metadata, but also for relationships from/to publications that cite the software or for which the software is supplementary material (i.e., the paper describing the experiment for which the software was produced). Scientists do not excel today in this practice, as indicated by the 45,000 citations between software and articles in the OpenAIRE Graph, of which 23,000 are inferred by full-text mining, and thus not provided at publication time. In the following, we bring examples of common sense-based practices, which may lead to rich and complete records, very poor ones, or to hiding software records from SKGs.

Figure 5 shows a software product deposited in the ESCAPE OSSR repository, a venue that provides for degrees of metadata trust and scientific trust and utilizes Zenodo.org as a data source.

Figure 6 shows an example of software deposited into the Zenodo.org catch-all repository to get both a PID and bibliographic metadata. At the user's request, the software payload is fetched and preserved by Zenodo.org from GitHub[23]. The user has provided complete and accurate metadata; still, no metadata trust means that accuracy cannot be guaranteed.

For many scientists "data set" is the type intended for "everything that is not a publication." Figure 7 shows an example of software deposited into the Figshare.org catch-all repository but "erroneously" typed as a "data set" due to lack of metadata trust.

3.2.2.2. Scientific trust. Only in the minority of cases is research software FAIR-published under scientific trust control, for example, as a supplement to a scientific article, with mandatory bibliographic data, as per request from a journal or a conference (e.g., *Journal of Open*

[21] https://www.osti.gov/doecode.
[22] https://research-software-directory.org.
[23] https://docs.github.com/en/repositories/archiving-a-github-repository/referencing-and-citing-content.

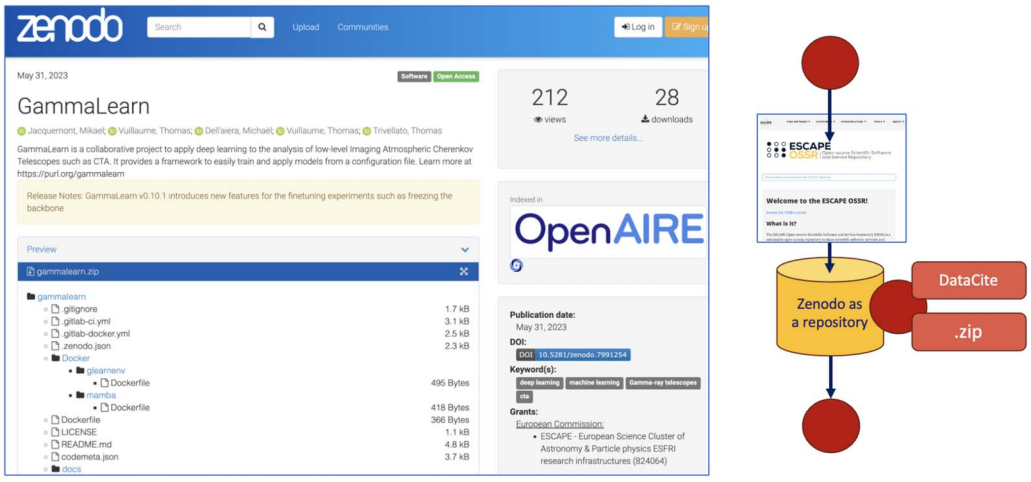

Figure 5. Example of research software published in the ESCAPE OSSR repository, which ensures metadata trust and scientific trust.

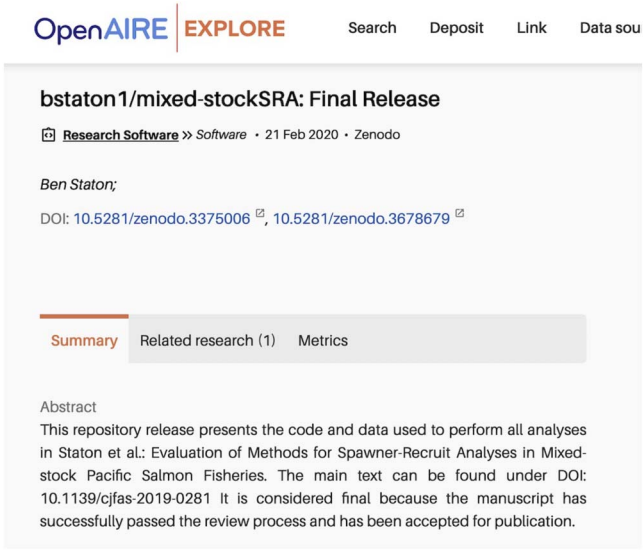

Figure 6. Examples of research software published in Zenodo to get a DOI but without metadata trust and scientific trust.

Source Software[24]) or by request of a thematic publishing venue, whose publishing workflows are designed to publish research software under quality controls, review, or the endorsement of trusted communities (e.g., biotools.org). Accordingly, SKGs tracking software product publishing must carefully select the publishing venues and data sources or cope with the possibility of unwelcome products being included in the collection.

3.3. Research Data

3.3.1. Scope of assessment

Publishing venues for research data range from disciplinary ones, with highly specialized scientific databases (e.g., proteins in ProtBD), to cross-disciplinary ones like national data

[24] https://joss.theoj.org.

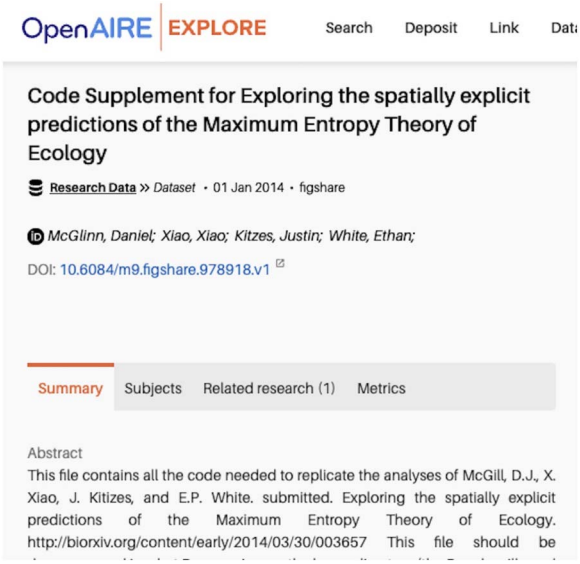

Figure 7. Example of research software published with resource type "data set."

archives or catch-all repositories. The re3data.org registry of data repositories counts around 3,000 sources, and FAIRSharing.org around 2,000. As to resource types, unlike research publications and research software, research data are rather disciplinary animals and will always be. Starting from the general definition, different communities use the term "data set" to intend rather different products: "packages" of product files, which may include software, data, etc.; "collections" of pointers to distinct data files; results of queries over a database (Pröll & Rauber, 2013); or entire databases; in other cases, "data set" and "research data" are used interchangeably. When called to classify their products as "research data," communities (or individual scientists in the "long tail of science") show somewhat different interpretations and often feature dedicated ontologies (Chawinga & Zinn, 2019; Gomez-Diaz & Recio, 2022a). De facto, defining the scope of research assessment for research data requires an in-depth dive into the community practices, facing the following main challenges:

- *Establishing eligibility of publishing venues for research assessment.* Which publishing venues, and hence which underlying data sources, should be regarded to support research assessment is a bottom-up, community-driven, decision. Organized communities refer to thematic data repositories to publish domain-specific research data described by domain-specific metadata (e.g., Pangaea.de). Long-tail data scientists may instead rely on general-purpose institutional and national data repositories (e.g., dataverseNL) or catch-all repositories (e.g., Zenodo.org, Figshare.org). Venues in the first class tend to support gatekeeping policies, which give quality reassurance to SKG providers. Still, not all venues contain products relevant to research assessment. For example, the Integrated Interaction Database[25] is an online database of predicted (inferred) protein-protein interactions (PPI) in 18 species, including humans and 12 domesticated species. The products therein are relevant for science but not for research assessment, being generated by machinery and not human thinking or skills. Venues in the second class are instead "black holes" where disciplinary metadata is missing and

[25] https://iid.ophid.utoronto.ca.

gatekeeping can be absent or ensured only to some extent; there are exceptions though, for example some "collections" in catch-all repositories (e.g., OpenPlantNLR in Zenodo .org[26]) are curated (gatekeeped) by communities.

- *Establishing eligibility of research data for research assessment*. Which entities should be mapped as research data onto SKGs and which should be regarded for research assessment is a community decision, in turn subject to revisions over time. When research data comes from thematic data sources, resource types and metadata formats radically differ from discipline to discipline and mapping them onto SKG models is not straightforward. For example, the life sciences produce bioentities, such as proteins, sequences, genes, etc., classify them as "research data" (the OpenAIRE Graph counts 32 million[27]), publish them into scientific databases, assign accession numbers as PIDs, and share bioschema.org as common metadata model and format. Other disciplines feature different PID schemes (handle, DOIs, etc.), resource types (e.g., crystals, clinical trials, drugs, design, maps), formats (e.g., CIDOC[28]), and payload shapes (e.g., sets of files, queries, time series, tables).
- *Granularity of research data*. Granularity is often a discipline-specific choice. For example, Pangaea.de organizes research data into collections of data sets, where the authors of the collection are also the authors of the data sets; in SKGs authors are therefore assigned a "plus one" due to the proxy collection product. When the publishing venue is not thematic and structured, researchers follow their common sense and adapt to the technical requirements of the platforms; for example, data sets stored into smaller fragments whose size and number depend on repository limits (e.g., 50 Gbyte per deposition in Zenodo.org, 3 Gbyte per file in DataVerse).

The definition of SKGs for research data research assessment is challenging, complex, and hard to maintain in terms of evolution, especially when dealing with cross-disciplinary scenarios, implying multiple specific choices for the three challenges above and feedback with community experts.

3.3.2. Gatekeeping practices

Research data publishing has become a practice, with scientists sharing data autonomously or via journal platforms supporting the data sharing as supplementary material. Still, the maturity of practices varies from discipline to discipline, ranging from rigorous gatekeeping approaches to an absolute lack of rules. Some publishing venues apply scientific trust controls to research data and ensure metadata quality (e.g., the Pangaea.de repository in the Earth Sciences[29]). In other cases, publishing venues support metadata trust but do not ensure scientific trust controls; for example, dryad.org verifies that a minimal level of metadata quality is reached[30] but does not support peer review workflows nor ensure that only trusted community actors can deposit products. Finally, many publishing venues do not support gatekeeping at all, as in the case of Zenodo.org, Figshare.org, and some institutional repositories, dangerously leaving metadata quality to the user's common sense (Quarati & Raffaghelli, 2022). In the following, we illustrate the implications of such a variegated universe.

[26] https://zenodo.org/communities/openplantnlr.
[27] https://explore.openaire.eu/search/find/research-outcomes?type=datasets&instancetypename=Bioentity.
[28] https://www.cidoc-crm.org.
[29] https://wiki.pangaea.de/wiki/Curation_levels.
[30] https://datadryad.org/stash/mission#our-curation-and-publication-process.

3.3.2.1. Metadata trust. FAIR principles encourage data publishing and citations from publications to research data, showing an encouraging trend. The OpenAIRE Graph counts around 60 million products classified as research data and around 3 billion publication-data relationships. Data citation increases visibility and article citations (Colavizza, Hrynaszkiewicz et al., 2020; Piwowar, Day, & Fridsma, 2007). On the other hand, cross-discipline citation practices are missing, raising issues of semantics, and also of identification, completeness, and fixity (Silvello, 2018), at the level of SKGs.

3.3.2.2. Scientific trust. In some disciplines, the quality and added value of the payload are reviewed before publishing by scientists, following the "publication metaphor" (e.g., peer review) or via validation tools (Schmidt, Struckmann et al., 2021; Wagner & Henzen, 2022). In other cases, scientific trust can be guaranteed by the publishing venue's submission workflows (Callaghan, 2019; Parsons & Fox, 2013), which support community-trusted practices, tools (Agosti, Buccio et al., 2012), or endorsement strategies, ensuring that research data payloads (e.g., files, records in a database) are produced by known instruments, communities, or services (Gomez-Diaz & Recio, 2022b; Sengupta, Bachman et al., 2019). Finally, there are scenarios where the publishing venue is not supporting scientific trust at all. The case of catch-all repositories and institutional repositories for data is exemplary. Their role is crucial for the community, enabling open access publishing of all kinds of products, but their lack of scientific trust leaves room for uncontrolled usage by users and platforms, which can harm research assessment in a number of ways. Here are some concrete examples:

- *Usage of catch-all repositories as file repositories*, to store results of digitalization, photos, videos, etc. For example, the Belgium Herbarium of Meise Botanic Garden[31] is a Zenodo.org collection of book page scans all associated with the project grant that funded their digitalization. Each scan is counted as a research data product (but is it one?), valuable for assessment of the grant, the researchers, and the organizations involved;
- *Supplementary material*: a peculiar and frequent subcase of the former point, is that of "supplementary material," such as figures and tables, deposited as "research data" in catch-all repositories by publisher's submission platforms as a result of the article's editorial process. For example, the OpenAIRE Graph counts 450,000 research data whose title includes the term "figure"[32]. Such platforms deposit supplementary material as research data products with links (semantics "supplementOf") to the related article DOI; an example from the OpenAIRE Graph of such an article is shown in Figure 8. Another form of pollution arises when scientists, to satisfy publishers' demands to deposit supplementary research data, publish a copy of their peer-reviewed article as "research data" because the article contains a "data" table.
- *Catch-all repository spamming*. Spamming is a well-known issue in publishing involving the mass uploading of records to open repository APIs for SEO purposes. Despite constant efforts to combat spam, thousands of DOIs are minted, sometimes finding their way into SKGs.

4. SKGS IN OPEN SCIENCE

In this section we describe how the gatekeeping scenario depicted above affects the construction of SKGs for research assessment. Table 2 gives an overview of the issues raised by the

[31] https://zenodo.org/communities/belgiumherbarium.

[32] https://explore.openaire.eu/search/find/research-outcomes?f0=q&fv0=figure&type=datasets.

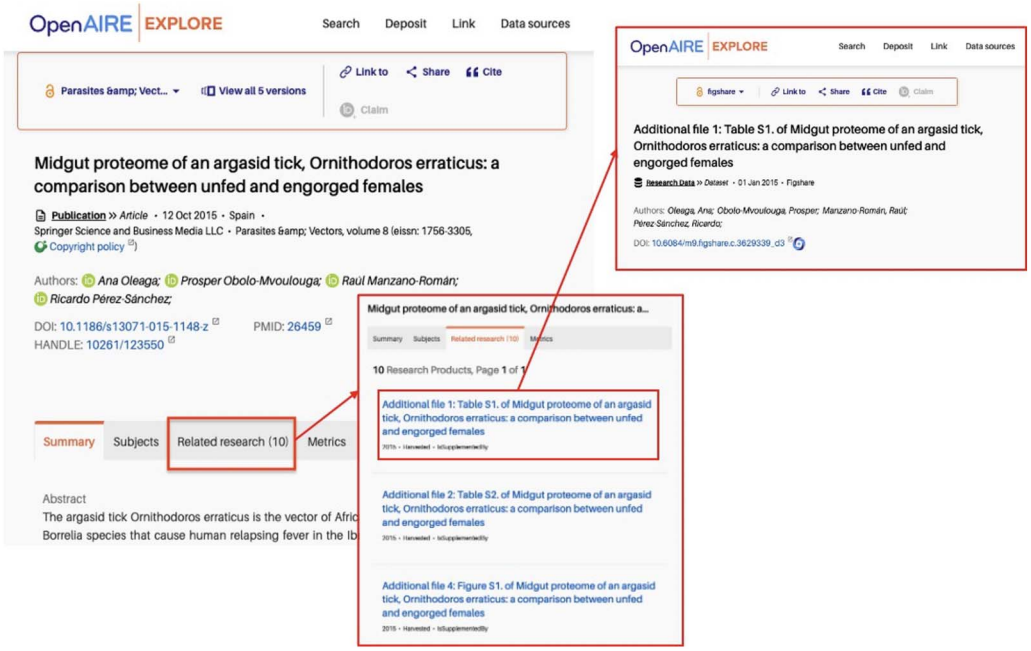

Figure 8. OpenAIRE Graph: example of a scientific article with supplementary material.

different kinds of products in the construction phases of aggregation, enrichment, and deduplication.

4.1. Aggregation

The number of publishing venues eligible for research products poses a serious scalability and heterogeneity challenge. Each source may export metadata records based on different metadata formats, thereby requiring specific crosswalks from the data source format to the SKG format and data model (e.g., resource types, dates, author names, PIDs). For example, the OpenAIRE Graph collects directly from 2,100+ data sources around 500 million metadata records and maintains more than 250 mapping rules. Defining proper crosswalks is an onerous and continuous activity to be monitored and maintained over time, sometimes hampered by challenging technical and methodological barriers. The most notable ones are caused by metadata aggregators resource types, resource type "others," persistent identifier's reliability, and affiliations.

The first problem is raised by SKGs whose construction process (rightly) delegates part of the aggregation task to metadata aggregators (or registries and catalogues), which define and operate metadata crosswalks across different data sources. Such intermediate services unburden SKGs from point-to-point metadata collection but hide the original data source resource types from the end-stream SKGs, thereby hindering SKGs' capacity to control or revisit the mapping. Figure 9 shows an example of a data set published by the Atlas of Living Australia via DataCite typed as "text." This type is used by DataCite members/providers with different meanings, as a data set in textual form or as a publication. The OpenAIRE Graph aggregation chain assumes the second interpretation and *ad hoc* corrective actions are taken when this is not the case. The "indirection" problem raised for resource types may be extended to all record fields affected by the crosswalk. For example, metadata records collected from Zenodo.org are richer than Zenodo's metadata collected from the DataCite registry.

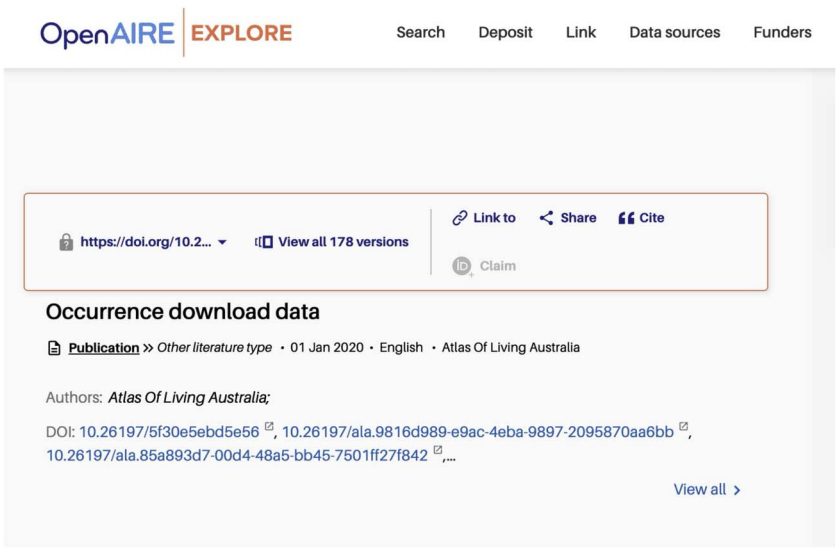

Figure 9. OpenAIRE Graph: example of a data set typed as a publication via the DataCite.org registry.

The second problem is a consequence of the resource type "others," exposed by repositories' user interfaces, and sometimes optioned by undecided or hasty authors; the OpenAIRE Graph search for "publications of type: other type of literature"[33] returns 14.5 million results. Considering such products for research assessment is not obvious, as the type may indeed conceal relevant products, such as preprints, but also second-class products, such as bulletins.

The third problem is caused by the erroneous usage of persistent identifiers, violating their implicit guarantee of unicity and identity (Juty, Wimalaratne et al., 2020). For instance, some noncurated data sources may export product records that bear DOIs in the identifier field (e.g., dc:identifier for Dublin Core), but they are instead references to supplementary material (e.g., data sets). In other cases, ORCID identifiers may be incorrectly assigned to the wrong author name string (Baglioni et al., 2021). SKG providers must adopt adequate countermeasures when data sources are not subject to metadata trust.

Finally, author affiliation information is vital for research assessment, but it is often missing in research data and software metadata; even when present, it may not be linked to PIDs but specified in the form of strings. Solutions to this problem range from complex and costly data source-specific metadata crosswalks to semiautomated deduplication of organization strings with organization metadata information collected by SKGs from registries, such as ROR.org or the European Commission's CORDIS database (Pavone & Atzori, 2021).

In the following sections, aggregation barriers specific to research product typologies are reported.

4.1.1. Research publications

Publishing venues for academic publications often adhere to standard bibliographic metadata and resource type vocabularies. However, harmonizing local records into an SKG data model

[33] https://explore.openaire.eu/search/find/research-outcomes?type=%22publications%22&instancetypename=%22Other%2520literature%2520type%22.

Figure 10. OpenAIRE Graph: example of scientific article typed as conference object.

presents significant challenges. For instance, the diversity in resource type vocabularies is extensive (e.g., COAR[34], CERIF[35]), and their interpretation can vary widely among individuals during the deposition process. A notable issue is that specific, well-defined types, such as "article," may be obscured by more generic categories like "conference object."[36] This generic category is commonly used by institutional repositories to classify a range of conference-related outputs, including articles, presentations, and demonstrations, as illustrated in Figure 10. The OpenAIRE Graph, for example, includes 11.7 million such entries[37], highlighting the potential impact on research assessment.

Another complicating factor is the presence of metadata export errors from data sources. Due to varying interpretations and attempts to conform to standard export metadata formats, data sources may incorrectly map internal resource types, resulting in the publication of misclassified records within SKGs. An illustration of this issue is provided in Figure 11, where a "newsletter" is mapped by the export process of the data source onto an "article" resource type.

These discrepancies underscore the complexity of achieving accurate and consistent aggregation of bibliographic metadata across thousands of publishing venues and data sources.

4.1.2. *Research software*

Scientists publish research software via an extensive range of venues, encompassing thousands of data, software, institutional, and catch-all repositories. SGKs must therefore identify software products by mapping various local resource types onto the SKG's resource types vocabulary and, more generally, address the heterogeneity of metadata formats.

[34] https://vocabularies.coar-repositories.org/resource_types.
[35] https://cerif.eurocris.org/vocab/html/OutputTypes.html.
[36] https://vocabularies.coar-repositories.org/resource_types/c_c94f.
[37] https://explore.openaire.eu/search/find/research-outcomes?type=%22publications%22&instancetypename=%22Conference%2520object%22.

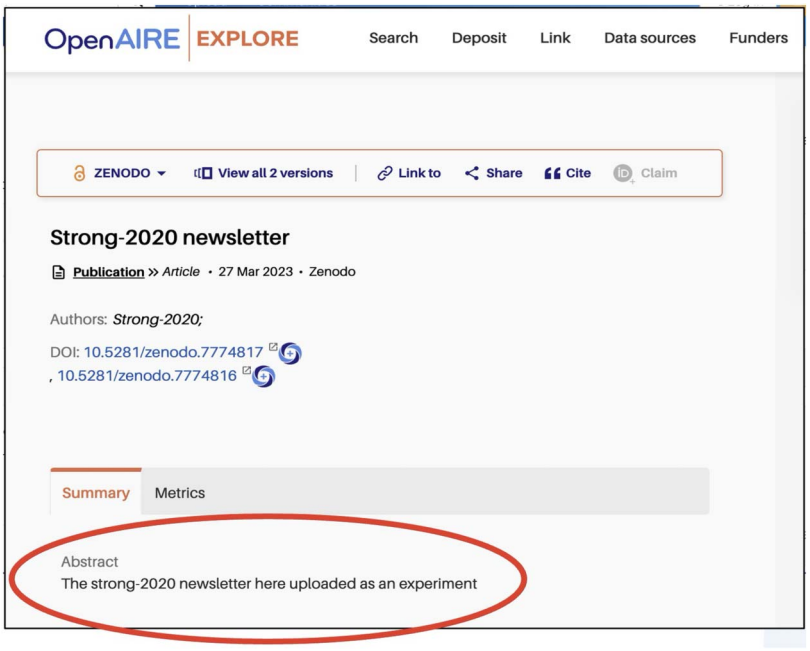

Figure 11. Data source exporting a "newsletter" as an "article."

It should be noticed, however, that a presumably large (and unmeasurable) portion of research software remains hidden from SKGs. An examination of the OpenAIRE Graph reveals that, as of today, out of 390,000 software products within the OpenAIRE Graph, approximately 218,000 originate solely from Zenodo.org. The analysis reports 20,000 products in 2018, 25,000 in 2019, 34,000 in 2020, 46,000 in 2021, 50,000 in 2022, and 51,000 in November 2023. The trend indicates promising growth but compared to the numbers of publications and research data in the Graph, it confirms a lack of awareness among scientists regarding the necessary steps to publish their software.

4.1.3. Research data

The aggregation of research data metadata and links in SKGs is complicated by the heterogeneity of metadata formats, the variety of resource types, granularity of information, and the general immaturity of community-established publishing practices, which leaves scientific trust and metadata curation in the hands of the end-users publishing their data. The process is further complicated by the problem of selecting which data sources are eligible for the SKG scope of assessment, out of the thousands available for data. To achieve optimal results, SKG providers must sustain and operate costly, heuristic-based, data source-specific crosswalks from local models onto the SKG data model. For example, the OpenAIRE Graph, after consultation with community experts, has removed the European Nucleotide Database[38] from the set of eligible research data publishing venues. The reason is that citations between articles and ENA sequences could be in the order of millions for each article, as the database supports a fine-grained data set granularity; there is no consolidated concept of a "collection of sequences associated to a given research"). Although useful for reproducibility, such granularity and links could be misleading for research assessment as they do not indicate scientific value, but relatedness.

[38] https://www.ebi.ac.uk/ena.

4.2. Enrichment

SKG enrichment techniques rely on obtaining the payloads associated with product metadata records, such as web "splash pages" (e.g., a publisher's detail page of an article) and the actual files of publications, data, or software, in order to extract metadata and complete the original records. These activities face a variety of obstacles, including legal and technical challenges, as well as concerns regarding the authoritativeness and reliability of the extracted information. The following sections will discuss these challenges in relation to different types of products.

4.2.1. Research publications

Techniques such as text mining and deep learning (Salatino, Mannocci, & Osborne, 2021) are often used to infer information from the full text of publications and enrich the related metadata records. Known inference tasks are publication citations, author affiliations, and topics (e.g., Sustainable Development Goals, Fields of Science). This approach is also applied to product web pages, crawled from the Web, which often contain structured bibliographic metadata, such as author affiliations, ORCID identifiers, and corresponding authors, to be extracted with custom scripts.

Downloading full texts and web pages is not straightforward, as it requires smart web crawling techniques and, when legally required, agreements with publishing venue providers (e.g., publishers, research organizations), and demands large storage and processing infrastructures. For example, the OpenAIRE Graph has signed agreements with Springer Nature, ACM, Wiley, and Elsevier, as well as agreements with data source providers, and empowers a storage (replicated) infrastructure of ~69 Tbyte for around 30 million PDFs and full texts extracted from the former via CERMINE[39].

The availability of open access publishing venues, such as preprint servers, facilitates this acquisition process. However, the lack of scientific trust underlying some open access publishing venues may affect the reliability of the outcomes. This holds for any attribution metadata extraction but especially for citation and affiliation extraction, foundational elements of research assessment. For instance, Google Scholar's SKG started to include some selected institutional and catch-all repositories in its collection. Consequently, research products with unverified scientific trust and metadata trust find a way into Scholar's machinery and are text-mined to extract citations: The search in Scholar for "A user journey in OpenAIRE services" results in a deck of PDF slides, from which references will be extracted. Such a "hack of the system" introduces undesirable distortions in the citation network and adds one more potential malicious practice to an already long list (Delgado López-Cózar, Robinson-García, & Torres-Salinas, 2014; Fong & Wilhite, 2017).

4.2.2. Research software

Software repositories obey common formats and file structures and may be utilized for mining to derive metadata. For example, codemeta or CFF files in software repositories are recommended to include software attribution metadata. Given one software repository platform, e.g., GitHub, or a software archive such as SoftwareHeritage.org, one approach could be to collect such files from repositories and, if available, extract metadata information and links to the reference publications. Unfortunately, such practices are not the norm (e.g., CFF is followed by less than 2% of software projects).

[39] https://github.com/CeON/CERMINE.

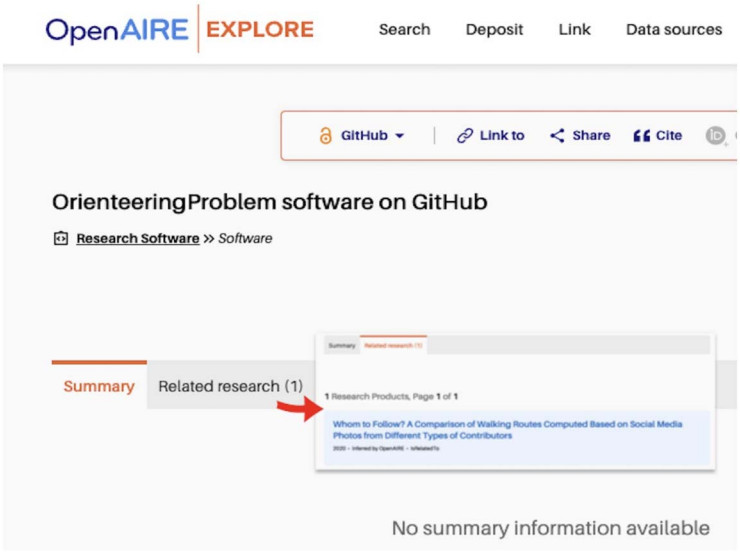

Figure 12. Examples of research software record inferred by text mining a scientific article's PDF and linked to the article's record.

Another approach is to mine publications' full text to find the URLs that authors write to refer to software repositories (Haris, Stocker, & Auer, 2022). OpenAIRE Graph exploits this strategy to identify software URLs (23,000 as of today) and implements scripts, one per software repository platform, through which metadata information is parsed out of the web pages and a bibliographic record is generated for the software. The record is added to the SKG together with a semantic relationship to the publication record that contained the software repository URL. Coverage depends on the number of full texts and the range of software repository systems handled by the process, each requiring dedicated APIs and web page scrapers. As exemplified in Figure 12, although this operation is undoubtedly relevant for research reproducibility, the resulting software metadata tends to be poor: Attribution metadata is incomplete and the relationship to the publication does not imply that the software was a supplement of the article or the result of research investigations.

A complementary approach is the one adopted by the French Open Science Monitor (via the software-mentions software[40]) to identify mentions of software within articles using AI models, in an attempt to classify research software produced in the context of an experiment. In this case, the existence of software is not implied by a URL to a repository, but rather by the in-text description within the publication. The software record is therefore reporting the existence of software, but potentially missing the actual URL to the software product.

4.2.3. Research data

Exploiting research data payloads to enrich SKGs is not straightforward due to the diversity of research data payloads across disciplines and scientists practices (e.g., CSV tables, text files, PDFs, database entries), which requires highly customized and context dependent interpretation of data, and the fact that research data files do not in general embed bibliographic metadata, such as attribution or linking to other products and entities, useful for research assessment.

[40] https://github.com/softcite/software-mentions.

4.3. Deduplication

The sheer number of data sources from which metadata can be collected increases considerably the likelihood of duplicate records (i.e., records provided via different publishing venues and describing the same research product or versions of it). Deduplication is an essential stage of SKG population to ensure a nonredundant science research assessment. The challenge is computational (e.g., the OpenAIRE Graph harvests 500 million records); semantic, as the similarity match of records suffers from low quality or lack of trust of the data; and of stability of the SKG, as the deduplication business logic may evolve over time, complicating the incremental construction of an SKG, returning in such cases misaligned releases of the same graph, and complicating the maintenance of SKGs with stable identifiers. Other technical challenges are involved, such as the merge policies of equivalent records to build a richer "representative" record while preserving provenance of contributing data sources and avoiding loss or redundancy of relationships after the merge (Manghi, Atzori et al., 2020).

4.3.1. Research publications

Publication duplicates are common across data sources, as coauthors publish their articles (or related preprints) in institutional repositories, which may be in turn part of aggregators, and in publisher's archives. For research assessment it is therefore crucial to identify duplicates and be able to distinguish their nature (e.g., preprint, postprint, published version).

Deduplication policies may differ for SKGs but are in general complicated by the unbalanced richness of metadata records from different sources, by their unmatching resource types, and by data source metadata export policies. For example, "articles" should be matched against "conference objects," as the latter may include conference articles; however, because they may also refer to presentation slides, this approach risks deduplicating and merging an article with its presentation slides at a conference. Another common example is that of records with the title "Preface," often with no authors, with the same publishing year, and collected from multiple sources, which require adequate heuristics so as not to bring false equivalent records. Other problems may derive from the author's name formats, which may obey different templates and exceptions (e.g., "et al.") and require dedicated business logic, in some case data source-specific ones.

The OpenAIRE Graph collects from repositories, publishers, national/thematic aggregators, and other SKGs, resulting in a 52% duplication rate for publications. Its process considers preprints and published versions as different versions of the same scientific effort, to be merged and counted as a single outcome.

4.3.2. Research software

The deduplication of research software records is limited by the poverty of metadata. For example, in the OpenAIRE Graph we cannot assume that the title of two matching records is identical or that the list of authors is exhaustive, as in many cases metadata records are inferred and may lack part of the information. One matching criterion is using the URL to the software repository, which in many cases acts as a stable identifier. Software Heritage identifiers are also an option, although not yet spread across repositories. In the OpenAIRE Graph, which counts a software deduplication rate of 60%, one of the most common deduplication scenarios derives from Zenodo.org (today counting 85% of software with DOIs), which exposes multiple and identical metadata records, one for each version of the software product.

4.3.3. Research data

Duplication of research data records across publishing venues is less common than for publications. Research data tend to be published once and to appear multiple times in SKGs due to metadata collection from both data sources and aggregators. Confirming this trend is the OpenAIRE Graph 12% duplication rate for research data. For such reasons, deduplication operates at the level of PIDs rather than matching metadata attributes. In some cases, however, the duplication strategy of the SKG considers multiple versions of the same data set, with different PIDs, as duplicates (e.g., for statistical reasons). The solution of applying general-purpose criteria across all research data products may, however, introduce false positives, and exceptions must be taken care of. Figure 9 shows how different research data products can be erroneously considered duplicates because of the same title and author metadata they bear; in this specific case, the publishing workflow supported by the venue Atlas of Living Australia automatically assigns the same title "Occurrence Download Data" and author "Atlas of Living Australia" to all depositions.

5. A GLANCE TO THE FUTURE

The sections above have shown how the urgency and jump-start of open science led to publishing workflows that, with degrees that depend on the maturity of the context, bring complexity to the task of building SKGs for research assessment. The traditional assumptions of gatekeeping are relaxed, thereby losing the conditions to first identify a clear research assessment scope and, second, a reliable one. The discipline-specific and data source-specific bias of such workflows introduces a degree of heterogeneity that is rather hard to capture and tackle without the involvement of researchers, communities, and publishing venues. On the technical side, harmonization, enrichment, and deduplication techniques require a variety of skills and an amount of resources that are not always easily available and sustainable for all SKG providers.

This scenario should not discourage our vision and intentions. We cannot expect the research ecosystem to spontaneously enact conditions of gatekeeping and alignment similar to the ones we have today for scholarly articles. While we need to acknowledge the challenges and make adjustments to current publishing workflows and infrastructure, we can track, measure, and understand what scientists produce today to organize and monitor such a transition in synergy with the research communities. To this end, we envision a synergic approach, involving SKG providers and actors in publishing workflows. First, we propose the use of SKGs as tools to track and monitor publishing workflows, so that research assessment can be performed with full awareness of the gatekeeping practices underlying each research product. Second, we highlight infrastructure tools and practices that we believe can be crucial to promote collaboration among research communities, organizations, policymakers, and SKG providers to address the challenges outlined in this paper.

5.1. Tracking Publishing Practices

SKGs can and should introduce methods to annotate metadata records to qualify and quantify the inherent trust and accuracy of publishing practices behind a given research product. SKGs can become maps of such practices, an observatory, and a source of inspiration for the evolution of open science publishing workflows. Such graphs can support research assessment applications in generating indicators that consider the intrinsic metadata trust and scientific credibility of research products, and publishing venues and researchers in collecting feedback on how to improve their publishing workflows and publishing practices. Indeed, observing such metadata annotations from the point of view of a discipline or a set of publishing venues

may help identify problematic practices common to a network of scientists. SKGs would become essential tools for research organizations, funders, and researchers to enhance and refine their publishing workflows and educate scientists to converge to common cross-disciplinary behaviors.

5.1.1. Gatekeeping annotations: The publishing venue perspective

SKGs should find ways to distinguish which products have undergone scientific trust and metadata trust controls. This is not always straightforward, as SKGs collect metadata records from data sources that may serve as storage and access support for publishing venues with rather diverse gatekeeping modalities. For example, as shown in Figure 13, Zenodo.org serves as data source for the venue ESCAPE OSSR open access repository, which supports metadata trust and scientific trust by community endorsement and for the venue Zenodo.org itself, acting as a free deposition catch-all repository without any gatekeeping. Accordingly, because SKGs harvest records for both venues from the same data source, the aggregation process should find ways to annotate records to highlight their distinct provenance.

In an ideal scenario, SKGs should rely on a public registry of publishing venues that describe their gatekeeping nature. They should also rely on metadata models that include attributes to explicitly refer to the publishing venue and describe the related gatekeeping practices. Such "gatekeeping annotations" would greatly benefit SKG consumers in their research assessment applications but, unfortunately, such registries and metadata standards do not exist.

However, SKGs can infer some of these annotations from data source registries or product records. For example, journals listed in Crossref and DOAJ and articles in ArXiv and PubMed undergo known forms of scientific trust, while content from catch-all repositories lacks any form of gatekeeping. During the aggregation phase, SKGs could annotate records from such data sources with a label or value measuring the metadata trust and scientific trust of the record. More generally, any attempt to reveal, measure or expose the degree of metadata trust and scientific trust underlying research products would highly benefit consumers and advance investigations in this domain.

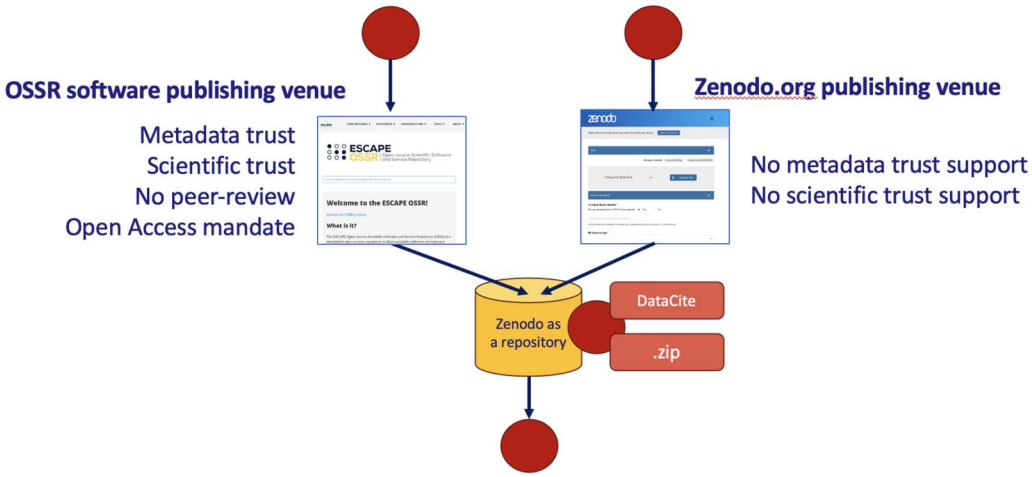

Figure 13. Publishing venues with different gatekeeping degrees on top of the same data source.

5.1.2. Publishing practices annotations: The researchers perspective

Regardless of the gatekeeping level of publishing venues, SKGs have the potential to identify researchers' open science "good practices" in providing metadata information. SKGs can enhance research product records with annotations that capture the extent to which they adhere to metadata quality expectations in terms of completeness or reproducibility, whether specific to publications, data, or software. Expectations could be mandatory or recommended, drawing a clear line between research products that are accountable for research assessment and those that are, but still lack crucial information that would improve open science. For instance, commonly expected mandatory criteria include the presence of a title, for all products, or at least one author for scientific articles, while "no authors" could be admitted for research data. Examples of recommended criteria are the presence of PIDs and abstracts or, for data sets, the presence of a link to the article they are supplementing. For software products, a good practice of reproducibility could entail recommending the inclusion of a URL to a software repository (e.g., GitHub reference) or referencing a publication product that documents the software, such as an article or a technical report. Such annotations could be inferred by analyzing the data in the SKG or be calculated via tools, such as F-UJI[41] for FAIR quality metrics.

When enhanced with such annotations, SKGs can be used to identify research products that suffer from shortcomings that may compromise their accountability for specific research assessment use cases; or recommend to researchers, scientific communities, or publishing venues how to improve their practices for the sake of accurate research assessment; or to produce metrics of Open Science practice adherence for data sources, venues, communities, etc.

5.2. Towards Open Science Scholarly Communication Commons

In the previous section, we proposed methods for how SKGs could address the shortcomings of current open science publishing workflows. In this section, we present and envision Open Science Commons, such as practices, guidelines, and tools that we believe could enhance the existing scholarly communication infrastructure, enabling more robust publishing workflows for open science.

5.2.1. Practices and guidelines

Practices, models, and guidelines are key to establishing common behavior, data representation, interpretation, disambiguation, and exchange. The landscape analysis in this paper highlighted a lack of important commons, such as standard SKG data models, transparent sharing of crosswalks from data source models to SKGs, and lack of practices that ensure authoritativeness of metadata for research data and research software.

5.2.1.1. Interoperability framework for SKG data exchange. The European Open Science Cloud (Corcho, Eriksson et al., 2021) has recently urged the definition of Interoperability Frameworks to be shared globally to ensure the interoperability of services and data. The EOSC interoperability framework for SKGs is the goal of the RDA Interest Group "Open Science Graphs for FAIR Data," engaging with stakeholders such as OpenAIRE.eu, OpenCitations.net, DataCite.org, Crossref.org, OpenAlex.org, ORKG.org, ResearchGraph.org, eurocris.org, and more than a hundred scientists working on the topics of SKGs. The framework[42] enables the exchange of SKG data by agreeing on a standard data model and export format, which represent the core

[41] https://www.f-uji.net.

[42] https://skg-if.readthedocs.io.

research entities (publications, data, software, researchers, organizations, grants, etc.), their properties and semantic relationships. The framework operates as a tool to enable metadata exchange between SKGs and standardize the definition of coherent metadata mappings from data source metadata formats onto a common data model in support of open science research assessment.

5.2.1.2. Data sources (and communities) responsible of cross-walking onto SKG data model. As highlighted in the first sections of this work, the construction of SKGs requires the definition of crosswalks from the data model of data sources to be aggregated onto the SKG data model. This process is complex and expensive as it requires disciplinary knowledge and must be followed up by prompt updates following local evolutions. In an ideal scenario, given the standard data model for SKGs mentioned in the previous section, data source providers must be responsible for adopting and using a common set of assets (PID, vocabularies, metadata models) recommended by the Global Open Science Commons and publishing and maintaining in dedicated services (e.g., the Metadata Schema and Crosswalk Registry[43]) a structural and semantic crosswalk from their local data model onto the global Open Science Commons models (e.g., SKG common data model). Data source registries (e.g., OpenDOAR.org, re3data.org, FAIRSharing.org) should refer from the data source profile description to the data source's crosswalk. SKG providers could therefore align (and refer to) such crosswalks to produce content according to consistent and community-approved terms.

5.2.1.3. Metadata freezing. We have highlighted how metadata records for data and software products represent the only source of attribution or citation information for a product (e.g., authors, organizations, funding, references to other products), as the related files do not contain attribution data, as in the case of scientific publications. As metadata can be updated any time after submission in most publishing venues (e.g., data repositories), the scenario is open to malicious behavior (e.g., adding authors to highly cited research data, adding incoming citations to a product). As a standard practice, publishing venues should, at publishing time, require authors to submit files with metadata together with the data or software files, freezing them under the same PID both product and attribution and citation descriptions. Something similar happens in software repositories, where codemeta and CFF files can be published with code files in the repository.

5.2.2. Tools and services

For publishing venues and scientists to implement and adhere to coherent open science publishing workflows, we perceive two main urgent issues. On the one hand, we have the operation of trusted entity registries to enable uniform representations of provenance, quality, attribution, reproducibility elements, etc., while on the other hand, there is a need to compensate for the lack of scientific trust without renouncing to the freedom of open science publishing workflows that can generate it.

5.2.2.1. Entity registries. The adoption of research entity registries for data sources, authors, organizations, etc. is critical for the population of a consistent scholarly record, and hence for the construction of SKGs.

Some registries are still missing, such as those for PID schemes, metadata schemas, disciplinary communities, gatekeeping practices, facilities, instruments, and publishing venues.

[43] https://faircore4eosc.eu/eosc-core-components/metadata-schema-and-crosswalk-registry-mscr.

Data source registries exist (e.g., re3data.org, OpenDOAR.org, FAIRSharing.org, ROAR) but they do not differentiate between data sources and publishing venues. As highlighted in previous sections, the establishment of publishing venue registries, listing venues, capturing their gatekeeping capabilities, and recording their data source of reference would be critical to support the construction of transparent SKGs. The European Commission funds some of these activities via projects like EOSC-Future[44], SciLake[45], and FAIRCORE4EOSC[46], which aim to deliver registries for services, metadata schemas and crosswalks, and SKGs, but none of these is production-ready yet. SKGs tend either to build their own disambiguated collections, or simply do not consider such entities as identifiable objects.

Finally, while some entities still lack dedicated registries, others showcase multiple registries, featuring distinct inclusion policies and data models (e.g., OpenDOAR, re3data.org, and FAIRSharing.org for data sources). SKGs must address the challenges posed by the disambiguation problem arising from the presence of multiple registries for the same entity. Coordinating interoperability and ensuring entity profile alignment between these registries is crucial to facilitate SKG construction and therefore research assessment (Baglioni, Mannocci et al., 2023).

5.2.2.2. Overlay tools for scientific trust controls. Scientific trust requires disciplinary skills and community recognition and, in general, qualified personnel that not all publishing venues or data source providers can afford. An area of investigation in this domain is the realization of overlay tools for "post-publishing scientific trust"; examples are the overlay open peer review systems (Ross-Hellauer, 2017) or annotation systems for data. Such systems operate on top of the publishing venues or SKGs, rely on product PIDs, and can become sources for SKGs to track gatekeeping practices.

We envision two classes of overlay tools, supporting automated review and human review, which can be integrated into publishing venues' workflows to compensate, when this is the case, for their lack of scientific trust support. In "automated review," tools allow for verifying structural and semantic compliance of research data payloads to ensure alignment with standards and, therefore, reuse. Services process the payload and return a score of expected quality (e.g., compliance of a data table with columns and data types). In "human review," in the spirit of a trusted network of scientists, tools allow authors to proactively request their colleagues to endorse their product payloads. The endorsement can be requested and/or provided at different degrees of inspection: from a lightweight "I trust my colleague on this product" (i.e., *à la* ArXiv.org), to a milder "I verified, and product files are good for me," or a stronger "I downloaded the files and personally verified their novelty, quality, and reproducibility" (e.g., article peer review).

6. CONCLUSIONS

This work has highlighted the vulnerabilities of today's publishing workflows for open science and how these currently affect the generation of SKGs in support of research assessment. The solutions sketched above foster the realization of a virtuous information cycle between authors and communities, publishing venues, and policymakers who need SKGs for their research assessment purposes. As illustrated in Figure 14, SKGs play a pivotal role in this process by harvesting and harmonizing metadata relying on community knowledge (e.g., crosswalks) and

[44] https://eoscfuture.eu.
[45] https://scilake.eu.
[46] https://faircore4eosc.eu.

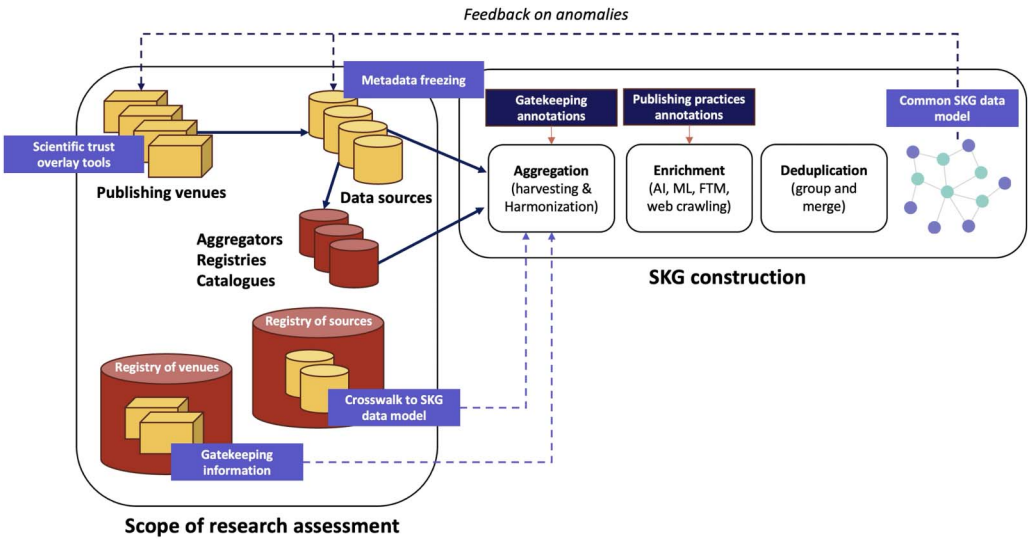

Figure 14. SKGs as pivots of a virtuous cycle of scholarly communication.

infrastructure services (e.g., registries), labeling metadata to track their provenance and gatekeeping support, and adapting their data model over time to match the emerging standards to align providers and consumers of data. Via SKGs, researchers and publishing venues can learn and improve their practices and tools toward global alignment while respecting disciplinary idiosyncrasies.

Still, for some time, we must accept that the instability of SKGs "is a feature, not a bug." Providers and consumers will likely observe changes in the final SKGs, which will not be limited to additions but include "restructuring" of the SKGs, in terms of identifiers and resource types, and hence in terms of deduplication (which depends on resource types), and in terms of the annotations mentioned above. Changes can be the result of many factors: improving the ways SKGs can interpret the community's intents by refining the crosswalks from data sources to SKG models; refinements of deduplication and enrichment processes; the publishing venues enhancing their gatekeeping capabilities; and authors consolidating practices to accurately metadata-ize their products.

The OpenAIRE Graph, used in this work as a reference for numbers and statistics, implements the full data provision workflow described in the previous sections. Its team contributes to the aforementioned recommendations in multiple ways: cochairing or participating in several RDA Interest and Working Groups for the definition interoperability frameworks and practices, providing requirements to the realization of a crosswalk and schema registry in the FAIRCORE4EOSC project, designing an SKG registry in the SciLake project, operating OpenOrgs[47], a service for bridging organization registries worldwide (ror.org, EC PIC numbers, ISNI, etc.), etc.

Finally, the OpenAIRE Graph team is engineering annotation techniques for labeling metadata records of scientific products with metadata trust and scientific trust features, to be derived via the establishment and operation of an open directory of publishing venues. As a result, expected in early 2025, we will investigate and confront publishing workflows patterns in disciplines, organizations, and funders, with the intent of collaborating with policymakers to feedback the actors and, where possible, undertake corrective actions.

[47] https://openorgs.openaire.eu.

ACKNOWLEDGMENTS

I extend my sincere gratitude to my colleagues Michele Artini, Claudio Atzori, Miriam Baglioni, Alessia Bardi, Michele De Bonis, Sandro La Bruzzo, Andrea Mannocci, Giambattista Bloisi, and Silvio Peroni for their constructive feedback, which significantly enhanced the quality of this work.

AUTHOR CONTRIBUTIONS

Paolo Manghi: Conceptualization; Funding acquisition; Investigation; Methodology; Supervision; Writing—Original draft; Writing—Review & editing.

COMPETING INTERESTS

The author has no competing interests.

FUNDING INFORMATION

This work was cofunded by the European Commission H2020 projects OpenAIRE Nexus (Grant agreement ID: 101017452) and EOSC-Future (Grant agreement ID: 101017536).

REFERENCES

Ahrabian, K., Du, X., Myloth, R. D., Ananthan, A. B. S., & Pujara, J. (2023). Pubgraph: A large-scale scholarly knowledge graph. *arXiv*. https://doi.org/10.48550/arXiv.2302.02231

Agosti, M., Buccio, E. D., Ferro, N., Masiero, I., Peruzzo, S., & Silvello, G. (2012). DIRECTions: Design and specification of an IR evaluation infrastructure. In *Information access evaluation. Multilinguality, multimodality, and visual analytics*. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-33247-0_11

Aksnes, D. W., Langfeldt, L., & Wouters, P. (2019). Citations, citation indicators, and research quality: An overview of basic concepts and theories. *SAGE Open*, *9*(1), 2158244019829575. https://doi.org/10.1177/2158244019829575

Ancona, A., Cerqueti, R., & Vagnani, G. (2023). A novel methodology to disambiguate organization names: An application to EU framework programmes data. *Scientometrics*, *128*(8), 4447–4474. https://doi.org/10.1007/s11192-023-04746-x

Aryani, A., Fenner, M., Manghi, P., Mannocci, A., & Stocker, M. (2020). Open science graphs must interoperate! In L. Bellatreche, M. Bieliková, O. Boussaïd, B. Catania, J. Darmont, ... M. Žumer (Eds.). *ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium* (pp. 195–206), Cham: Springer. https://doi.org/10.1007/978-3-030-55814-7_16

Baglioni, M., Manghi, P., Mannocci, A., & Bardi, A. (2021). We can make a better use of ORCID: Five observed misapplications. *Data Science Journal*, *20*, 38. https://doi.org/10.5334/dsj-2021-038

Baglioni, M., Mannocci, A., Pavone, G., De Bonis, M., & Manghi, P. (2023). (Semi)automated disambiguation of scholarly repositories. In *Proceedings of the 19th IRCDL (The Conference on Information and Research Science Connecting to Digital and Library Science)*. CEUR-WS.

Barker, M., Hong, N. P. C., Katz, D. S., Lamprecht, A.-L., Martinez-Ortiz, C., ... Martinez, P. A. (2022). Introducing the fair principles for research software. *arXiv*, arXiv:2307.02647. https://doi.org/10.48550/arxiv.2307.02647

Besançon, L., Cabanac, G., Labbé, C., & Magazinov, A. (2023). Sneaked references: Cooked reference metadata inflate citation counts. *arXiv*, arXiv:2310.02192. https://doi.org/10.48550/arxiv.2310.02192

Bologna, F., Iorio, A. D., Peroni, S., & Poggi, F. (2021). Can we assess research using open scholarly knowledge graphs? A case study within the Italian national scientific qualification. *arXiv*, arXiv:2105.08599. https://doi.org/10.48550/arxiv.2105.08599

Brack, A., Hoppe, A., Stocker, M., Auer, S., & Ewerth, R. (2020). Requirements analysis for an open research knowledge graph. In M. Hall, T. Merčun, T. Risse, & F. Duchateau (Eds.), *Digital libraries for open knowledge* (pp. 3–18). Cham: Springer. https://doi.org/10.1007/978-3-030-54956-5_1

Bryan, K. A., & Ozcan, Y. (2021). The impact of open access mandates on invention. *Review of Economics and Statistics*, *103*(5), 954–967. https://doi.org/10.1162/rest_a_00926

Callaghan, S. (2019). Research data publication: Moving beyond the metaphor. *Data Science Journal*, *18*, 39. https://doi.org/10.5334/dsj-2019-039

Chaleplioglou, A., & Koulouris, A. (2023). Preprint paper platforms in the academic scholarly communication environment. *Journal of Librarianship and Information Science*, *55*(1), 43–56. https://doi.org/10.1177/09610006211058908

Chawinga, W. D., & Zinn, S. (2019). Global perspectives of research data sharing: A systematic literature review. *Library and Information Science Research*, *41*(2), 109–122. https://doi.org/10.1016/j.lisr.2019.04.004

Colavizza, G., Hrynaszkiewicz, I., Staden, I., Whitaker, K., & McGillivray, B. (2020). The citation advantage of linking publications to research data. *PLOS ONE*, *15*(4), e0230416. https://doi.org/10.1371/journal.pone.0230416, PubMed: 32320428

Corcho, O., Eriksson, M., Kurowski, K., Ojsteršek, M., Choirat, C., ... Coppens, F. (2021). EOSC interoperability framework reference architecture. *Zenodo*. https://doi.org/10.5281/zenodo.4420095

Delgado López-Cózar, E., Robinson-García, N., & Torres-Salinas, D. (2014). The google scholar experiment: How to index false papers and manipulate bibliometric indicators. *Journal of the*

Association for Information Science and Technology, *65*(3), 446–454. https://doi.org/10.1002/asi.23056
Eysenbach, G. (2006). Citation advantage of open access articles. *PLoS Biology*, *4*(5), e157. https://doi.org/10.1371/journal.pbio.0040157, PubMed: 16683865
Fong, E. A., & Wilhite, A. W. (2017). Authorship and citation manipulation in academic research. *PLOS ONE*, *12*(12), e0187394. https://doi.org/10.1371/journal.pone.0187394, PubMed: 29211744
Gomez-Diaz, T., & Recio, T. (2022a). Research software vs. research data I: Towards a research data definition in the open science context. *F1000Research*, *11*, 118. https://doi.org/10.12688/f1000research.78195.1, PubMed: 36415208
Gomez-Diaz, T., & Recio, T. (2022b). Research software vs. research data II: Protocols for research data dissemination and evaluation in the open science context. *F1000Research*, *11*, 117. https://doi.org/10.12688/f1000research.78459.1, PubMed: 36483317
Gruenpeter, M., Katz, D. S., Lamprecht, A.-L., Honeyman, T., Garijo, D., ... Rabemanantsoa, T. (2021). Defining research software: A controversial discussion. *Zenodo*. https://doi.org/10.5281/zenodo.5504015
Haris, M., Stocker, M., & Auer, S. (2022). Scholarly knowledge extraction from published software packages. In Y.-H. Tseng, M. Katsurai, & H. N. Nguyen (Eds.), *From born-physical to born-virtual: Augmenting intelligence in digital libraries* (pp. 301–310), Cham: Springer. https://doi.org/10.1007/978-3-031-21756-2_24
Haucap, J., Thomas, T., & Wohlrabe, K. (2018). Publication performance vs. influence: On the questionable value of quality weighted publication rankings. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3126669
Heibi, I., Peroni, S., & Shotton, D. (2019). Software review: COCI, the OpenCitations index of Crossref open DOI-to-DOI citations. *Scientometrics*, *121*, 1213–1228. https://doi.org/10.1007/s11192-019-03217-6
Hong, N., Katz, D., Barker, M., Lamprecht, A.-L., Martinez, C., ... Yehudi, Y. (2022). *FAIR principles for research software released*. Research Data Alliance Europe. https://doi.org/10.59350/9qn73-phk11
Juty, N., Wimalaratne, S. M., Soiland-Reyes, S., Kunze, J., Goble, C. A., & Clark, T. (2020). Unique, persistent, resolvable: Identifiers as the foundation of FAIR. *Data Intelligence*, *2*(1–2), 30–39. https://doi.org/10.1162/dint_a_00025
Kling, R., Spector, L., & McKim, G. (2002). Locally controlled scholarly publishing via the internet: The guild model. *Proceedings of the American Society for Information Science and Technology*, *39*(1), 228–238. https://doi.org/10.3998/3336451.0008.101
Manghi, P., Atzori, C., Bardi, A., Baglioni, M., Schirrwagen, J., ... Pierrakos, D. (2023). OpenAIRE Graph dataset. *Zenodo*. https://doi.org/10.5281/zenodo.10037121
Manghi, P., Atzori, C., Bonis, M. D., & Bardi, A. (2020). Entity deduplication in big data graphs for scholarly communication. *Data Technologies and Applications*, *54*, 409–435. https://doi.org/10.1108/dta-09-2019-0163
Manghi, P., Mannocci, A., Osborne, F., Sacharidis, D., Salatino, A., & Vergoulis, T. (2021). New trends in scientific knowledge graphs and research impact assessment. *Quantitative Science Studies*, *2*(4), 1296–1300. https://doi.org/10.1162/qss_e_00160
Mannocci, A., Irrera, O., & Manghi, P. (2022). Will open science change authorship for good? In *Proceedings of the 18th Italian Research Conference on Digital Libraries*, Padua, Italy. *arXiv*, arXiv:2207.03121. https://doi.org/10.48550/arXiv.2207.03121
Parsons, M. A., & Fox, P. A. (2013). Is data publication the right metaphor? *Data Science Journal*, *12*, WDS32–WDS46. https://doi.org/10.2481/dsj.WDS-042
Pavone, G., & Atzori. C. (2021). OpenOrgs: Bridging registries of research organisations. Supporting disambiguation and improving the quality of data. *Zenodo*. https://doi.org/10.5281/zenodo.5101096
Peng, C., Xia, F., Naseriparsa, M., & Osborne, F. (2023). Knowledge graphs: Opportunities and challenges. *Artificial Intelligence Review*, *56*(11), 13071–13102. https://doi.org/10.1007/s10462-023-10465-9, PubMed: 37362886
Piwowar, H. A., Day, R. S., & Fridsma, D. B. (2007). Sharing detailed research data is associated with increased citation rate. *PLOS ONE*, *2*(3), e308. https://doi.org/10.1371/journal.pone.0000308, PubMed: 17375194
Pröll, S., & Rauber, A. (2013). Scalable data citation in dynamic, large databases: Model and reference implementation. In *2013 IEEE International Conference on Big Data* (pp. 307–312). https://doi.org/10.1109/bigdata.2013.6691588
Quarati, A., & Raffaghelli, J. E. (2022). Do researchers use open research data? Exploring the relationships between usage trends and metadata quality across scientific disciplines from the Figshare case. *Journal of Information Science*, *48*(4), 423–448. https://doi.org/10.1177/0165551520961048
Ross-Hellauer, T. (2017). What is open peer review? A systematic review. *F1000Research*, *6*, 588. https://doi.org/10.12688/f1000research.11369.2, PubMed: 28580134
Salatino, A. A., Mannocci, A., & Osborne, F. (2021). Detection, analysis, and prediction of research topics with scientific knowledge graphs. In Y. Manolopoulos & T. Vergoulis (Eds.), *Predicting the dynamics of research impact* (pp. 225–252). Cham: Springer. https://doi.org/10.1007/978-3-030-86668-6_11
Sanyal, D. K., Bhowmick, P. K., & Das, P. P. (2021). A review of author name disambiguation techniques for the PubMed bibliographic database. *Journal of Information Science*, *47*(2), 227–254. https://doi.org/10.1177/0165551519888605
Schiltz, M. (2018). Science without publication paywalls: cOAlition S for the realisation of full and immediate open access. *PLOS Medicine*, *15*(9), e1002663. https://doi.org/10.1371/journal.pmed.1002663, PubMed: 30178782
Schmidt, C. O., Struckmann, S., Enzenbach, C., Reineke, A., Stausberg, J., ... Richter, A. (2021). Facilitating harmonized data quality assessments. A data quality framework for observational health research data collections with software implementations in R. *BMC Medical Research Methodology*, *21*(1), 63. https://doi.org/10.1186/s12874-021-01252-7, PubMed: 33810787
Schomberg, R. v., Britt Holbrook, J., Oancea, A., Kamerlin, S. C. L., Ràfols, I., ... Wouters, P. (2019). *Indicator frameworks for fostering open knowledge practices in science and scholarship*. European Commission. https://doi.org/10.2777/445286
Sengupta, S., Bachman, D., Laws, R., Saylor, G., Staab, J., ... Bauck, A. (2019). Data quality assessment and multi-organizational reporting: Tools to enhance network knowledge. *eGEMs*, *7*(1), 8. https://doi.org/10.5334/egems.280, PubMed: 30972357
Siler, K., Lee, K., & Bero, L. (2015). Measuring the effectiveness of scientific gatekeeping. *Proceedings of the National Academy of Sciences*, *112*(2), 360–365. https://doi.org/10.1073/pnas.1418218112, PubMed: 25535380
Silvello, G. (2018). Theory and practice of data citation. *Journal of the Association for Information Science and Technology*, *69*(1), 6–20. https://doi.org/10.1002/asi.23917

Tosi, M. D. L., & Dos Reis, J. C. (2021). SciKGraph: A knowledge graph approach to structure a scientific field. *Journal of Informetrics*, *15*(1), 101109. https://doi.org/10.1016/j.joi.2020.101109

Tennant, J., Agarwal, R., Baždarić, K., Brassard, D., Crick, T., ... Yarkoni, T. (2020). *A tale of two 'opens': Intersections between free and open source software and open scholarship*. https://doi.org/10.31235/osf.io/2kxq8

Tennant, J. P. (2020). Web of Science and Scopus are not global databases of knowledge. *European Science Editing*, *46*, e51987. https://doi.org/10.3897/ese.2020.e51987

Verma, S., Bhatia, R., Harit, S., & Batish, S. (2023). Scholarly knowledge graphs through structuring scholarly communication: A review. *Complex & Intelligent Systems*, *9*, 1059–1095. https://doi.org/10.1007/s40747-022-00806-6, PubMed: 35965491

Vuillaume, T., Al-Turany, M., Füßling, M., Gal, T., Garcia, E., ... Verkouter, M. (2023). The ESCAPE open-source software and service repository. *Open Research Europe*. https://doi.org/10.12688/openreseurope.15692.2, PubMed: 38264265

Wagner, M., & Henzen, C. (2022). Quality assurance for spatial research data. *ISPRS International Journal of Geo-Information*, *11*(6), 334. https://doi.org/10.3390/ijgi11060334