

European Infrastructure Integration and Research Knowledge Graphs

Christian Hauschke¹, Benjamin Kampe¹, Ivan Mršulja², Miloš Popović² and Tatiana Walther¹

¹ TIB – Leibniz Information Centre for Science and Technology, Hannover, Germany

² Faculty of Technical Sciences, University of Novi Sad, Serbia

christian.hauschke@tib.eu, benjamin.kampe@tib.eu,
ivan.mrsulja@uns.ac.rs, milos.popovic@uns.ac.rs,
tatiana.walther@tib.eu

Abstract

European research infrastructures increasingly emphasize cross-institutional knowledge graphs, linked open data, and interoperability across national and disciplinary boundaries. This paper highlights gaps in and contributions to the research knowledge graph (RKG) integration in Europe, focusing on the role of institutional knowledge graphs (IRKG) as central hubs in a vertically integrated ecosystem. The findings underscore the need for widespread OpenAIRE compliance and open infrastructure to build a trustworthy, reusable research information commons across Europe, and the role that knowledge-graph-based open-source research information systems like VIVO can play in this.

1 Introduction

A knowledge graph is “a graph of data intended to accumulate and convey knowledge of the real world, whose nodes represent entities of interest and whose edges represent potentially different relations between these entities.” (Hogan et al. 2021), where the conveyed knowledge is explicit knowledge in the sense of Nonaka (2009): “formal and systematic” knowledge, which “can be easily communicated and shared”. Thus, a Research Knowledge Graph (RKG) is a knowledge graph which represents explicit knowledge about research.

The knowledge represented in a RKG is derived from research information, which is metadata about research activities, actors and results that constitute the research ecosystem. This includes, for example, descriptions of projects, researchers, organisations, publications, software or conferences. RKGs are making this research information discoverable, comparable and reusable across systems and institutions.

Research information in Europe is increasingly organized around shared, open, and interoperable infrastructures, which are complementing institutional research information systems. This development

is driven by several initiatives with varying focus. The European Open Science Cloud (EOSC) is a federated infrastructure for open science across Europe. The European Commission's research funding programme Horizon Europe mandates open access and FAIR data as conditions of funding. The European Research Area (ERA) framework promotes cross-border collaboration and knowledge sharing. The European Commission (2025) analyzed feedback from their stakeholders, which points at "[t]echnical barriers [...] hindering the free circulation of knowledge", a "lack of standardized metadata and interoperable data infrastructures", "the lack of standardized metadata, as noted in the Barcelona Declaration" decreasing the discoverability, and "valuable data remain[ing] siloed across institutional domains". Furthermore, "[t]hese technical challenges are compounded by the dominance of proprietary systems, which restrict access and interoperability."

From the perspective of research assessment and evaluation, the Coalition for Advancing Research Assessment (CoARA) works towards ensuring fair and transparent research evaluation. It aims at incorporating a wider range of research activities into the evaluation process, and to respond to the dynamic developments set out in the Barcelona Declaration on Open Research Information. This makes it clear that both the coverage and quality of open research information must increase in the future. It also promotes richer, more transparent research information for research assessment. One approach to achieving this is the vertical integration of RKGs to facilitate the exchange of information across all levels - from the individual to the institutional and international - thereby creating a metadata commons that benefits all.

2 Types of Research Knowledge Graphs

RKGs can be differentiated from diverse perspectives. One of them is the granularity, or level of aggregation, of the subjects in its centre. Here, we distinguish between different levels.

At the most granular level, there are personal research knowledge graphs (PRKGs). They can be understood as a RKG that contains "information in the form of entities and their relationships that are of importance to the researcher" (Chakraborty et al. 2022). Usually, it contains the knowledge a researcher has about their own activities and environment too, making it a personal knowledge graph (Balog & Kenter, 2019; Skjæveland et al., 2024) that supports personal information management (Jones 2007). PRKGs typically contain information that is not intended or suitable for publication. This includes private notes, information about individuals, or interim results that are not (yet) suitable for publication.

Bridging the personal and institutional levels is ORCID, which functions as a standardized, shared PRKG. While a PRKG is often private and local, ORCID provides a persistent identifier (PID) that connects researchers to their contributions and affiliations in a machine-readable format (Haak et al. 2012). In the context of vertical integration, ORCID acts both as a data source and as a data sink. As a source, it allows harvesting employment history or connections of an ORCID iD to other PIDs. As a sink, it can receive metadata from individual researchers or institutions, such as employment records or funding acknowledgments. This bidirectional flow can make ORCID a critical interoperability layer between the private PRKG and the IRKG.

An institutional knowledge graph represents the actors, processes, and output of an institution's research activities. It models and connects entities such as researchers, departments, projects, publications, datasets, grants, and facilities. This enables unified discovery and reporting across the organization. Institutional Research Information Systems (RIS) can be implemented as knowledge graphs, where bibliographic records and other research information are represented as nodes and edges to support flexible querying, aggregation, and semantic enrichment. Such a graph contains the knowledge that an institution has about itself. By leveraging persistent identifiers and open ontologies, an institutional knowledge graph facilitates interoperability with external sources by other actors in

academia such as publishers, funders, and research infrastructure. The adoption of semantic technologies, such as RDF/OWL, by RIS results in their qualification as IRKGs.

The next level comprises over-institutional research knowledge graphs. These can be disciplinary, regional, or international. Two prominent examples in the European context are OpenAIRE and CORDIS, which serve different functions:

- OpenAIRE is a European Research Knowledge Graph whose purpose is "to facilitate, foster, support, and monitor Open Science scholarly communication in Europe" (Manghi et al. 2019). The OpenAIRE knowledge graph relies on aggregating metadata from diverse information sources, including RIS, with a primary focus on research outputs such as publications and datasets.
- CORDIS, maintained by the European Commission, functions as an over-institutional graph focused on "comprehensive information about EU Research & Development projects", including information about funding, involved institutions, and grant-related outputs.

OpenAIRE has the potential to be a true research information commons in the sense of "information and knowledge resources that are collectively created and owned or shared between or among a community and that tend to be [non-exclusive], that is, be (generally freely) available to third parties. Thus, they are oriented to favor use and reuse, rather than to exchange as a commodity" (Fuster Morell 2010, p. 5).

Complementing these regional infrastructures are global research knowledge graphs such as OpenAlex (Priem et al. 2022). OpenAlex aims to provide comprehensive coverage of research outputs regardless of geography or funding source. Here, it serves as an example of a knowledge graph acting as a data source for IRKGs seeking to enrich their local records with global bibliographic context.

When integrating RKGs, it is essential to ensure that the provenance of the content is traceable, to verify the information, and to include only trustworthy information. An important role that IRKGs can play here in vertical integration is that of epistemic trust markers. If information is transferred from a PRKG via an IRKG to, for example, OpenAIRE, the intermediary between the researcher and OpenAIRE must guarantee the quality and accuracy of the metadata.

3 PIDs: The Connective Tissue of Research Graphs

PIDs serve as a connective tissue between entities in graphs, facilitating the integration of different knowledge graphs. They are globally unique, resolvable, and persistent references to entities in the research landscape. They are the foundation of interoperability in knowledge graphs and (open) research information systems. In RKGs, PIDs act as anchors that enable different data sources to be joined, deduplicated, and enriched. Some of the main PIDs for RKG integration are:

- ORCID (Open Researcher and Contributor ID) for persons. ORCID is the de facto global standard for researcher identification. It enables disambiguation of researchers across institutions, tracking the research of persons with name changes, and career moves.
- ROR (Research Organization Registry) for organizations. ROR is increasingly replacing proprietary organization identifiers in European research systems. Organization IDs are critical for disambiguation of affiliations.
- DOI (Digital Object Identifier) for research outputs such as publications, datasets, and software. The most prevalent DOI registries are CrossRef and DataCite. There are several other non-publication use cases for DOIs such as PIDINST (Stocker et al. 2020), a PID schema specifically developed for scientific instruments and equipment, DOIs for academic events

(Franken et al. 2022), RAID (Research Activity ID, Suominen et al. 2024) for projects, or Funder IDs in the Open Funder Registry (Hendricks et al. 2020).

The RKG ecosystem in Europe is converging around open, community-governed registries. It is deliberately moving away from proprietary identifier systems. For example, an EOSC PID policy paper suggests using PIDs in open infrastructure with metadata that is “community property and openly available” (Buys et al. 2024).

4 RIS Participation in OpenAIRE

The foundations for the interlinking of knowledge graphs in Europe have now been established. But what does this look like in practice? How many IRKGS contribute content to OpenAIRE? To answer these questions, we have investigated how many European RIS are OpenAIRE-compliant, how much content they contribute to OpenAIRE, and to what extent the various European countries are involved in this. For this purpose, we used the OpenAIRE dataset provided by Sesame Open Science and last updated on 24 February 2026 via Orion-DBS (Neylon et al. 2026).

Of the 162,657 data sources in OpenAIRE, 1,430 were of the type “CRIS System”. Of these, 1,347 were of the type “Institutional CRIS”, 26 “National multidisciplinary CRIS”, 22 “CRIS System”, 12 “Funder CRIS” and 2 “CRIS Aggregator”. 1,380 of those 1430 “CRIS System” (96.5%) were not registered as OpenAIRE compliant data sources. Only very few are explicitly OpenAIRE compliant: OpenAIRE CRIS v1.1 contributes 43 records (3.0%), while other variants are rare. This includes OpenAIRE 3.0 (OA/funding) which contributes 4 records (0.3%), and OpenAIRE 2.0 (EC funding), OpenAIRE PubRepos v4.0, and proprietary systems, each with one record. Overall, the distribution shows a strong concentration of unregistered or not-yet-classified CRIS, with minimal uptake of standardised OpenAIRE compliant versions (see Fig. 1). For this investigation, we focus on institutional CRIS.

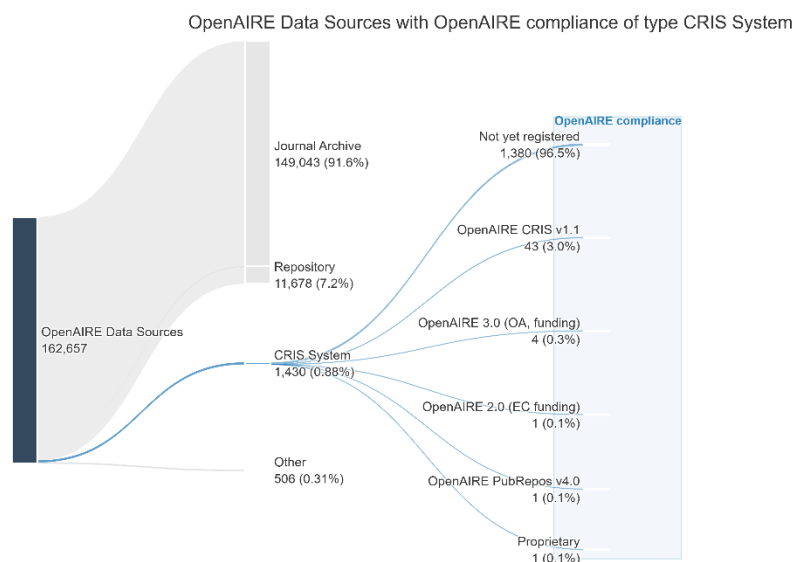


Figure 1: OpenAIRE Data Sources with OpenAIRE compliance of type CRIS System

The mapping of institutional CRIS data sources to countries is almost entirely dominated by single-country data sources (1,344 of 1,347 data sources). Two of them are attributed to two countries: one to India and Switzerland and another one to Lithuania and Spain. One has no country affiliation. These three are subsequently excluded from this investigation.

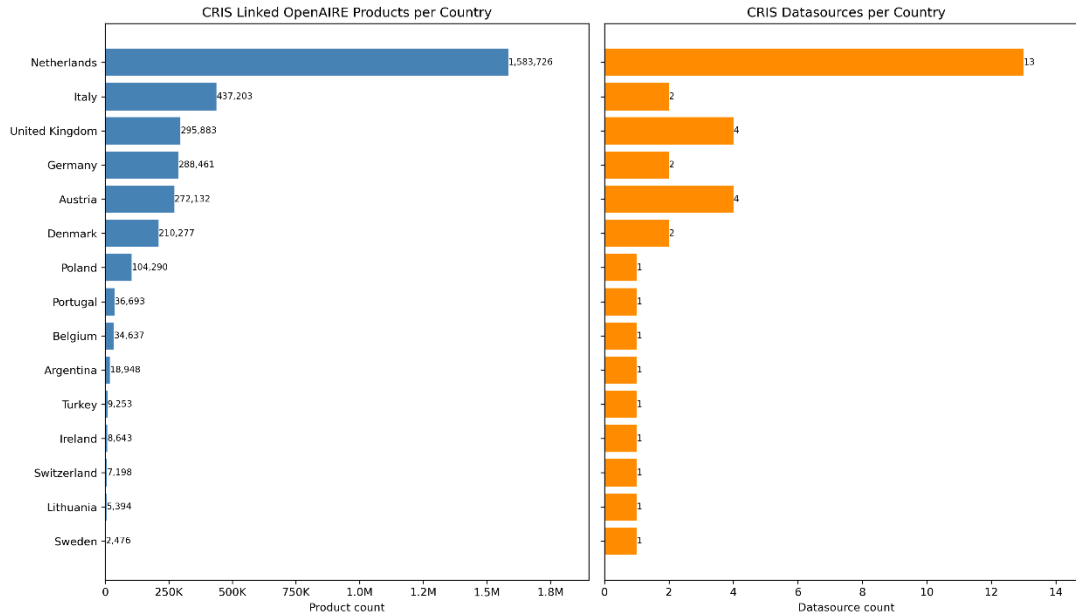


Figure 2: CRIS-linked OpenAIRE products per country and CRIS data sources per country

As can be seen in Fig 2, the Netherlands stands out not only in the number of CRIS data-providing systems, but also in absolute OpenAIRE product counts. This could be related to a structural transition in the Dutch research information infrastructure. Following the discontinuation of the national NARCIS system, efforts were made to strengthen distributed, OpenAIRE-compliant institutional harvesting. This included establishing the Netherlands Research Portal (<https://netherlands.openaire.eu/>) as a coordination layer for visibility and aggregation (De Jong et al. 2024). In contrast, across all other countries, CRIS-linked OpenAIRE product flows are clearly much lower and generally spread over only a small number of data sources. Unlike the Netherlands, which has a highly integrated and high-volume setup, most countries have a long-tail pattern with a comparatively weak CRIS presence and noticeably lower harvesting intensity within OpenAIRE.

5 Connecting institutional research knowledge graphs

As explained above, IRKGs play a significant role as epistemic trust markers in the exchange of research information between RKGs. They therefore require a wide range of integration options and mechanisms for acquiring and providing metadata. The open-source research information system VIVO (Conlon et al. 2019) is built on semantic technologies such as RDF and OWL. This enables VIVO to act as an IRKG. In this capacity, it represents the institutional research ecosystems as knowledge graphs, supporting interoperability through shared ontologies and an increasing number of PIDs (VIVO 2025).

5.1 OpenAlex to VIVO

OpenAlex provides a REST API for exporting metadata. The API allows filtering by various entity types, such as works, authors, and institutions. The VIVO Harvester is a suitable tool for automatically importing this data into VIVO and keeping it up to date, as it includes workflows for retrieving and mapping various data formats (<https://github.com/vivo-community/VIVO-Harvester>, see Fig. 3). The JSON format used by the OpenAlex API is also supported and can be processed without difficulty. Thus, by combining VIVO and OpenAlex, a knowledge graph for an institute, a group of institutes, or a specific community can be created with manageable effort. This approach was used for the VIVO based Research Atlas (Auhunas et al. 2024). There, the OpenAlex query for the Research Atlas filters publications to ensure they are relevant to the architecture and civil engineering community. Along with the publications, information regarding authors, their affiliations, publishers, and journals is also retrieved. The harvester maps all this information to the VIVO ontology, thereby constructing a comprehensive knowledge graph. To make things even simpler, a preconfigured package of all necessary components could be provided that offers the functionality to retrieve data from OpenAlex, requiring minimal configuration to get it up and running.

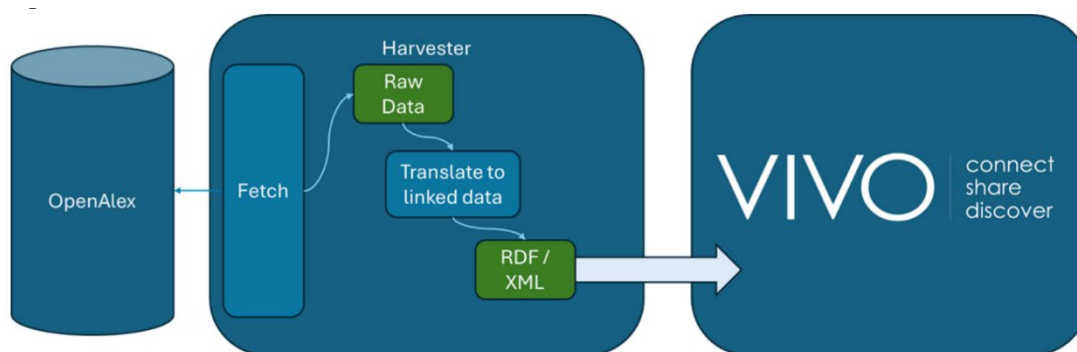


Figure 3: VIVO Harvester workflow for OpenAlex harvesting

5.2 CORDIS to VIVO

LORIS Projects (Leibniz Open Research Information System) is a VIVO prototype that aims at combining metadata about research projects funded by the Leibniz Association, German Research Foundation (DFG) and European Commission into one RKG. It gives an overview on what projects are administered by a particular member of the Leibniz Association, which institutes collaborate with each other, which institutes are doing research on a particular topic, what projects were conducted in a particular German state, how many projects were awarded by one of the three funders in a particular year.

CORDIS offers a convenient option to harvest project information via a SPARQL endpoint using the European Information Research Ontology (EURIO, Publications Office of The European Union 2025). To be able to import CORDIS data into LORIS Projects automatically and continuously, a mapping of EURIO to the VIVO Ontology was conducted. As a result, the VIVO Harvester could transform the metadata and incorporate it into its KG. One notable challenge encountered during the process was integrating organizations in LORIS Projects with those in CORDIS. The institutions had to be merged manually by means of CORDIS' local organization IDs (Participant Identification Code, PIC) and ROR ID in a prior step.

5.3 VIVO to ORCID

The integration extends the existing VIVO ORCID functionality while maintaining full backward compatibility. Researchers can explicitly authorize metadata synchronization by initiating an OAuth 2.0 workflow from their VIVO profile page. Upon successful authorization, the researcher's ORCID iD is automatically associated with the corresponding VIVO profile and access credentials are securely stored. The user retains full control over synchronization and may disable metadata pushing at any time.

A custom synchronization service periodically exports research information from VIVO to ORCID. The export process uses SPARQL queries against the institutional knowledge graph to retrieve employment and education affiliations as well as scholarly works. The harvested information is transformed into ORCID-compliant metadata and synchronized with the researcher's ORCID record using the ORCID Member API (v3.0). Both creation of new records and updates of previously exported records are supported, enabling continuous synchronization between the institutional and personal research knowledge graphs. To ensure compliance with ORCID metadata requirements and best practices, the synchronization process implements several validation and filtering mechanisms. For example, affiliations are exported only when organizations can be linked to persistent organizational identifiers such as ROR, while works must contain suitable identifiers before publication. Organizational hierarchies are traversed to identify the closest parent organization with a valid ROR identifier, while lower-level organizational units are preserved as department information in accordance with ORCID affiliation guidelines.

The synchronization infrastructure is implemented through a lightweight scheduling framework integrated into VIVO. Scheduled background tasks periodically synchronize research outputs and affiliations, refresh expiring OAuth access tokens using ORCID refresh tokens and remove expired credentials when necessary. Since authorization tokens are refreshed automatically and securely stored, researchers are not required to repeat the OAuth authorization workflow after the initial consent. This approach minimizes user effort while ensuring uninterrupted synchronization.

6 Challenges

A major challenge for the knowledge graph integration is the very dynamic field of PIDs. There are multiple emerging PIDs, such as PIDINST or RAID, and it is uncertain if there will be significant uptake by researchers and research organizations. A related issue is the inclusion of these PIDs in major metadata schemas.

Another concrete issue is the non-use of established PIDs. One example is the use of the Participant Identification Code (PIC) in CORDIS as an organizational identifier. Not using an open and standardized PID such as ROR makes RKG integration unnecessarily difficult.

Several challenges remain regarding standards, governance, and infrastructure. The uptake of new or under-development metadata schemas, such as CERIF 2.0, and emerging PIDs, such as IGSN, PIDINST, and RAID remains uncertain.

It is clear that IRKGs play an important role in the integration of RKGs. Therefore, important steps must be taken to strengthen the open research information commons as a whole. Firstly, all relevant RIS software must be enabled to publish OpenAIRE-compliant metadata. For Vitro, on which VIVO is based, at least the OAI-PMH interface has already been implemented as a prototype for the SAH Digital project (Walther et al. 2025).

Furthermore, the institutions operating these RIS must become aware of their role and responsibility. To this end, it would be helpful to describe the benefits of such a commons and make this description accessible to decision-makers.

Funding

This work was partially funded by the German Federal Ministry of Research, Technology and Space (BMFTR), project number 16RBM1018B, and by the Lower Saxony Ministry of Science and Culture, project „Aufbau eines Netzwerkes und Services zu Forschungsinformationen (FIS)“.

Acknowledgments

We would like to thank Bianca Kramer and Sophie Biesenbender for their insights and for the productive discussions.

References

- Auhunas, S., Hagemann-Wilholt, S., Hauschke, C., Kampe, B., Rudakova, T., Shapovalov, V., & Shapovalov, Y. (2024). Persistent conference identifiers in CRIS – first implementations with VIVO and URIS. *Procedia Computer Science*, 249, 254–263. <https://doi.org/10.1016/j.procs.2024.11.071>
- Balog, K., & Kenter, T. (2019). Personal Knowledge Graphs. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval* (pp. 217–220). *ACM. ICTIR '19: The 2019 ACM SIGIR International Conference on the Theory of Information Retrieval*. <https://doi.org/10.1145/3341981.3344241>
- Conlon, M., Woods, A., Triggs, G. et al. (2019). VIVO: a system for research discovery. *Journal of Open Source Software*, 4(39), 1182. <https://doi.org/10.21105/joss.01182>
- Buys, M., Dreyer, B., Hellström, M., Kálmán, T., Mejias, G., Parkoła, T., Růžicka, M., & Suominen, T. (with Klemenčič, B.). (2024). *Community-specific (and stakeholder category-specific) perspectives on the EOSC PID architecture and the EOSC PID policy*. Zenodo. <https://doi.org/10.5281/ZENODO.11396803>
- Chakraborty et al. (2022). Personal Research Knowledge Graphs. *ACM. WWW '22*. <https://doi.org/10.1145/3487553.3524654>
- De Jong, R., Baars, C., Bardi, A., & Vanderfeesten, M. (2024). *Switching the national portal of The Netherlands to OpenAire within six months*. Zenodo. <https://doi.org/10.5281/ZENODO.12579311>
- European Commission (2025). *European Research Area Act. Analysis of Responses to the Call for Evidence*. 19 December 2025. Feedback from all Stakeholders. https://european-research-area.ec.europa.eu/sites/default/files/documents/2026-02/CFE_RTD_ERA_report_final_040226.pdf
- Franken, J., Birukou, A., Eckert, K., Fahl, W., Hauschke, C., & Lange, C. (2022). Persistent Identification for Conferences. *Data Science Journal*, 21. <https://doi.org/10.5334/dsj-2022-011>
- Fuster Morell, M. (2010). *Governance of online creation communities: Provision of infrastructure for the building of digital commons*. https://www.onlinecreation.info/wp-content/uploads/2007/12/final_final_mayofustermorell_spthesisaugust2010_final_verc280a6.pdf
- Haak, L. L., Fenner, M., Paglione, L., Pentz, E., & Ratner, H. (2012). ORCID: a system to uniquely identify researchers. *Learned Publishing*, 25(4), 259–264. <https://doi.org/10.1087/20120404>
- Hendricks, G., Tkaczyk, D., Lin, J., & Feeney, P. (2020). Crossref: The sustainable source of community-owned scholarly metadata. *Quantitative Science Studies*, 1(1), 414–427. https://doi.org/10.1162/qss_a_00022
- Hogan, A., Blomqvist, E., Cochez, M., D'amato, C., Melo, G. D., Gutierrez, C., Kirrane, S., Gayo, J. E. L., Navigli, R., Neumaier, S., Ngomo, A.-C. N., Polleres, A., Rashid, S. M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., & Zimmermann, A. (2021). Knowledge Graphs. *ACM Computing Surveys*, 54(4), 1–37. <https://doi.org/10.1145/3447772>

- Jones, W. (2007). Personal Information Management. *Annual Review of Information Science and Technology*, 41(1), 453–504. <https://doi.org/10.1002/aris.2007.1440410117>
- Manghi, P., Bardi, A., Atzori, C., Baglioni, M., Manola, N., Schirrwagen, J., & Principe, P. (2019). *The OpenAIRE Research Graph Data Model (Version 1.3)*. Zenodo. <https://doi.org/10.5281/ZENODO.2643198>
- Neylon, C., Kramer, B., Fernandes Mazoni, A., Costas, R., Jahn, N., & van Eck, N. J. (2026). *Sharing the load: Building a collective to support open research information online*. Upstream Blog. <https://doi.org/10.54900/2pnyq-nhx95>
- Nonaka, I. (1991). The knowledge-creating company. *Harvard business review* (11). <https://hbr.org/1991/11/the-knowledge-creating-company-2>
- Priem, J., Piwowar, H., & Orr, R. (2022). *OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts (Version 2)*. arXiv. <https://doi.org/10.48550/ARXIV.2205.01833>
- Publications Office of The European Union (2025): *EUropean Research Information Ontology v2.6*. <http://publications.europa.eu/resource/dataset/curio>
- Richardson, R. A. K., Moon, J., Hong, S. S., & Amaral, L. A. N. (2025). Widespread misidentification of scanning electron microscope instruments in the peer-reviewed materials science and engineering literature. *PLOS One*, 20(7), e0326754. <https://doi.org/10.1371/journal.pone.0326754>
- Skjæveland, M. G., Balog, K., Bernard, N., Łajewska, W., & Linjordet, T. (2024). An ecosystem for personal knowledge graphs: A survey and research roadmap. *AI Open*, 5, 55–69. <https://doi.org/10.1016/j.aiopen.2024.01.003>
- Stocker, M., Darroch, L., Krah, R., Habermann, T., Devaraju, A., Schwarzmann, U., D’Onofrio, C., & Häggström, I. (2020). Persistent Identification of Instruments. *Data Science Journal*, 19. <https://doi.org/10.5334/dsj-2020-018>
- Suominen, T., Tatum, C., Sipponen, S., Rydman, W., & Ross, S. (2024). Piloting the use of RAiD in Research.fi. *Procedia Computer Science*, 249, 224–231. <https://doi.org/10.1016/j.procs.2024.11.068>
- Teixeira da Silva, J. A. (2025). Incomplete and mismatching descriptors of scanning electron microscopy (SEM) and SEM microscopes: a case study. *EXCLI Journal*, 24:Doc1019. <https://doi.org/10.17179/EXCLI2025-8605>
- VIVO (2025). *Survey Results: Identifier Use*. VIVO - Connect, Share, Discover. <https://vivoweb.org/2025/03/11/survey-results-identifier-use/>
- Walther, T., Rubach, B., Litvinov, G., & Hauschke, C. (2025). The GESAH Graphic Arts Ontology (GESAH GAO) and its reference application – concept and implementation. *Art Libraries Journal*, 50(1), 19–26. <https://doi.org/10.1017/alj.2025.6>