International FAIR Digital Objects Implementation Summit 2024

Full Papers

https://doi.org/10.52825/ocp.v5i.1418

Published: 18 Mar. 2025

Towards Machine-Actionable Scientific Knowledge as FAIR Digital Objects

Claudia Biniossek[1,2,3], Dirk Betz[1,2,3], Lars Vogt[1], and Markus Stocker[1,4,*]

[1]TIB - Leibniz Information Centre for Science and Technology, Hannover, Germany

[2]Center for Empirical Research in Economics and Behavioral Sciences, University of Erfurt, Germany

[3]Bioinformatics Group, Institute for Computer Science, Albert-Ludwigs-University of Freiburg, Germany

[4]Knowledge-based Systems Department, Institute of Data Science, Leibniz University Hannover, Germany

*Correspondence: Markus Stocker, markus.stocker@tib.eu

Abstract. Building on the Open Research Knowledge Graph as an infrastructure for the production, curation, and publication of FAIR scientific knowledge, we present a concept that models original articles and the corresponding expression in the ORKG as independent and interlinked FDOs by organizing the content describing an article into semantic units.

Keywords: Scientific Knowledge, ORKG, Digital Scholarship, FAIR Principles, FAIR Digital Objects (FDO), RO-Crates, Nanopublications, Semantic Unit

1. Overview

Published research articles are the main source of scientific knowledge. Empirical articles typically include data, analysis, and theory-based explanations, usually presented in the form of text, tables, and figures. To date, independent (primary or final) data publications are not common in many disciplines, so empirical-analytical articles are the main source of data. The multilayered content (e.g., data, analysis, and interpretations) of articles offers a wide range of possibilities for documentation, extraction, and reuse, and makes it particularly fruitful not only to express scientific knowledge according to the FAIR Guiding Principles [1] but also to produce knowledge from the outset as FAIR Digital Objects (FDO) [2], in order to make it human- and machine-actionable for knowledge processing tools and services. With a focus on the Open Research Knowledge Graph (ORKG) [3], we present how the ORKG supports the structured description of research findings published in articles and describe how the combination of FDO, ORKG, and semantic units can maximize the reusability of knowledge.

An article is an identified Digital Object (DO) as a bit sequence (mostly in PDF format) described by a metadata record that conforms to some metadata schema. DOI is the default identifier type and the metadata record conforms to the Crossref metadata schema. However, the scientific knowledge published in research articles, expressed as a bit sequence (DO) in PDF (or HTML) format(s), is not machine-actionable. As a result, machine support for the reuse of scientific knowledge is insufficient. Nanopublications [4] and the ORKG are emerging infrastructures that enable the production, publication, and reuse of a machine-actionable expression of scientific knowledge classically expressed as narrative text, tables, figures, diagrams, etc.

A classic example is an article that reports a Student's t-test by plotting the data and a sentence stating the p-value and statistical significance for the dependent variable under investigation. As illustrated in more details in [2], such a t-test can be described in structured form in terms of input data, dependent variable, and output data, in this case a p-value. To make the description machine-actionable, the tabular input data can follow a canonical syntax (e.g., CSWV [5]) and the dependent variable can be linked to a term of some terminology.

2. Open Research Knowledge Graph (ORKG)

2.1 Maximize reusability by interlinking articles and ORKG papers as FDOs

ORKG supports the specification of schemas for arbitrary data types by means of ORKG templates. The approach is similar to CNRI's Type Registry [6], Schema.org or SHACL [7], with the major difference that ORKG templates utilize ORKG terminology (for properties, data types, classes) and thus facilitate the production of ORKG-compatible data.

We argue that structured descriptions of research findings in ORKG can be easily transformed into FDOs: The structured description is a DO as a bit sequence with a machine-actionable syntax (e.g., JSON-LD). This DO conforms to one or more schemas (ORKG templates) and can be identified using a DOI, specifically a DataCite DOI of resource type *dataset*, and is thus described with an additional (bibliographic) metadata record that conforms to the DataCite metadata schema. This implementation enables an interesting possibility, namely the persistent interlinking of the original research article and the corresponding ORKG paper with structured descriptions including the underlying data of research findings published in the original research article. This amounts to DOI-based (bidirectional) interlinking of FDOs using "related identifiers" in DataCite and Crossref metadata. Given a Crossref DOI identifying an original research article, machines can discover the corresponding ORKG paper and thus machine-actionable descriptions of the findings published in the context of the research work. Such a mechanism could improve machine support for scientific knowledge reuse, especially in knowledge synthesis. Figure 1 graphically visualizes the relationship between these two FDOs.

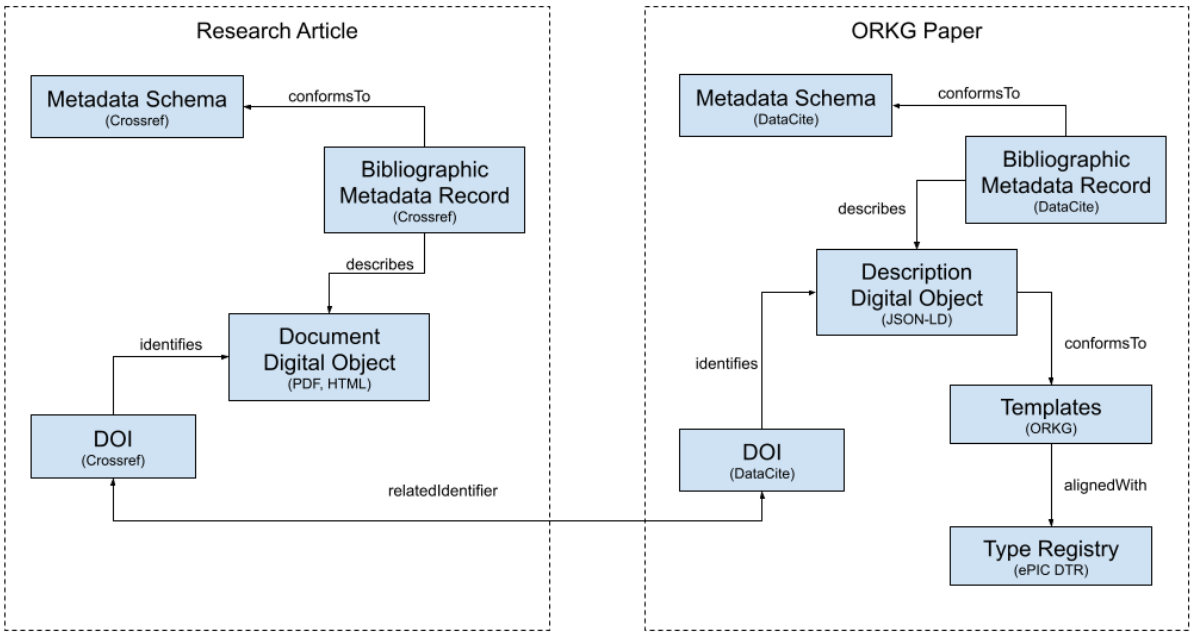

Figure 1. *Conceptual model relating published research articles and ORKG Papers as FDOs.*

An important avenue for future work is the alignment of ORKG templates with a type registry such as the ePIC Persistent Identifier Consortium's Information Type Registries. This would effectively align system-specific terminology and schemas with those maintained in an external registry, thereby increasing the ability of systems to produce data that are consumable by other systems in a distributed system of systems.

2.2 Producing FDOs with ORKG

An example of such a distributed system of systems is the production of machine-actionable scientific knowledge in (statistical) computing environments such as R or Python [2]. The Student's t-test mentioned above can be implemented in Python using Jupyter [8]. A trivial extension to the Python script provides a structured description of the t-test, e.g. in JSON-LD, so that the data produced conforms to a schema of a type registry. The JSON-LD data is interlinked as supplementary data to the research article and published in a distributed manner as an FDO. Assuming an alignment with the type registry, systems such as the ORKG can now harvest machine-actionable scientific knowledge and support integrated reuse in added value services for, e.g., knowledge synthesis.

3. FDO scientific knowledge model and its impact in research

3.1 Semantic units as knowledge model building blocks

Decomposing the content of a research article, i.e. scientific knowledge, into a set of different, partly nested, partly overlapping semantically meaningful units, i.e. semantic units [9], which can be documented using different types of FDOs, is a key building block of a scientific knowledge model for FDOs. Semantic units are representational entities that can contain information at different levels of granularity. The smallest level would be represented by statement units (statement not in the RDF sense of a single triple, but in the sense of assertions and thus simple sentences that researchers would build to communicate a piece of information). By considering statements as minimum units of information and communication, we can understand research articles as collections of statement units. Depending on the statement, such statement units can be modelled in the ORKG by one or more triples, each resulting in a small graph. Such a statement unit can be documented as an FDO in the form of a nanopublication.

Multiple statement units can form semantically meaningful collections of statements at higher levels of representational granularity. We call such semantic units compound units. Each compound unit can be documented as its own FDO using RO-Crates [10]. It would be a container for a collection of associated statement unit FDOs and other compound unit FDOs. Several types of compound units can be distinguished. Item units are compound units that are collections of statement units that share the same subject resource. Granularity tree units are collections of statement units that are based on the same partial order relation, such as *hasPart*, *developsFrom*, or *before*, and that form a connected graph that describes a hierarchical tree (i.e., a granularity tree such as a taxonomy or partonomy). Additional types of compound units exist, such as context units, argument units, dataset units, etc. [9]. Following this approach, the FDO of a research article in ORKG would be a container of a collection of associated different types of semantic unit FDOs.

By organizing and structuring the information contained in a research article into semantic unit FDOs and thus into various, partly nested, partly overlapping, semantically meaningful units, the content of an article becomes easier to explore and browse. When searching in a knowledge graph for a specific concept, information can be provided for the corresponding resource in the graph, indicating which semantic units contain this resource. This provides contextual information about the resource that can be used by user interfaces to support users in further exploring the graph [11, 12].

Moreover, by documenting the statements of a research article as statement unit FDOs and semantically meaningful collections of statements as compound unit FDOs, researchers will be able to reference them individually. This could have far-reaching consequences and ultimately change the way we publish research. In particular, it would allow for targeted referencing rather than referencing the article as a whole. Users could make statements about the content of a given semantic unit by using the identifier of its corresponding FDO, resulting in a statement unit that can be documented and published as a statement unit FDO nanopublication. For example, users could indicate that observation A in paper X contradicts hypothesis B in paper Y and publish the statement in the ORKG. However, such cross-document referencing tasks form a significant part of reading and writing activities in science [13], and associating information within the same and across different articles seems to be challenging without the aid of digital tools [14].

3.2 Statement units metadata and its implications to research assessment

The statement unit FDOs play a central role in this approach, as they carry the actual content - compound units function only as containers. Therefore, they need to cover the following metadata in addition to the typical FDO metadata:

1. Specification of who created the FDO (i.e., creator) and who authored its content (i.e., author);
2. Specification of the schema identifier (i.e., shape, table structure, etc.) used to model the statement to support schema interoperability [15];
3. Specification of the formal logical framework, if any, used to model the statement, indicating whether the content supports reasoning and which logical framework must be used to do so;
4. Specification of the statement category:
    a. Assertional statement, such as “This swan is white”;
    b. Contingent statement, such as “Swans can be white”;
    c. Prototypical statement, such as “Swans are typically white”;
    d. Universal statement, such as “All swans are white”;
5. A human-readable representation of the statement to make the FDO more human-friendly (see also *cognitive interoperability* [11]).

If each statement unit FDO identifies its creator/author, researchers can choose from a variety of formats for publishing their findings in a knowledge graph, ranging from single assertions (i.e., statement unit FDOs) to larger collections of assertions comparable to articles and books (i.e., compound unit FDOs). As FDOs, each publication has its own identifier and can be referenced by other researchers. This approach would make each researcher’s contributions to a larger work more transparent. Applying this approach to research articles would contribute to a fairer reflection of the actual work that went into an article by each of the listed authors, as authorship could be differentiated at the level of individual assertions. In addition, citations could reference specific FDOs instead of citing the entire article, resulting in targeted citations. By documenting such citations as statement unit FDOs with additional information (e.g., supporting, contradicting, etc.), citations would become qualified, targeted references that would allow the development of new ways to quantitatively assess a researcher’s contributions that does not necessarily depend on the impact factor of the journals in which they are published. If, instead of considering h-index, journal impact factors, and number of citations by published article and book, a researcher were evaluated based on the number of different types of citations that come from their statement unit FDOs and any FDOs that refer to them, weighted by the quality (appraisal systems, batches, etc.) and the nature of the citations (supporting, contradicting, etc.), it could substantially change the way researchers build their careers. Researchers might be able to build a career without having a single publication in a well-established journal because they have made statement FDOs that have been referenced by

other researchers. As a result, high-impact journals may lose some of their appeal, and we may finally be able to emancipate ourselves a bit from the publishing industry towards a multi-dimensional measurement of quality, relevance and impact maturity indicators for research and researchers.

4. Conclusion

FDOs and their flavours such as RO-Crates and nanopublications are a disruptive technology that will not only unleash its power as a container within and between (research) data spaces, but also allow the creation of modular research pipelines in combination with workflow technologies. This will have a significant impact on the production and consumption of FAIR (research) data as FDOs. In research, the application of FDO technology to scientific knowledge has the potential to enrich classical article publications in PDF format with machine-actionable expressions of scientific data. The Open Research Knowledge Graph (ORKG) infrastructure for digitalized scholarship presented in this article has and will continue to support and accelerate the adoption of FDOs in research and contribute to unlocking the potential of FDO technology.

Data availability statement

Not applicable.

Underlying and related material

Not applicable.

Author contributions

Conceptualization (CB, DB, LV, MS); Data curation (N/A); Formal Analysis (N/A); Funding acquisition (MS); Investigation (N/A); Methodology (N/A); Project administration (N/A); Resources (N/A); Software (N/A); Supervision (N/A); Validation (N/A); Visualization (N/A); Writing – original draft (CB, DB, LV, MS); Writing – review & editing (CB, DB, LV, MS).

Competing interests

The authors declare that they have no competing interests.

Funding

Parts of the work described in this article have been co-funded by the European Research Council (ERC) project ScienceGRAPH (GA: 819536) and the German Research Foundation (DFG) project NFDI4DS (PN: 460234259).

References

[1] Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. In Scientific Data (Vol. 3, Issue 1). Springer Science and Business Media LLC. https://doi.org/10.1038/sdata.2016.18

[2] Stocker, M., Snyder, L., Anfuso, M., Ludwig, O., Thießen, F., Farfar, K. E., Haris, M., Oelen, A., & Jaradeh, M. Y. (2024). Rethinking the production and publication of machine-reusable expressions of research findings (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2405.13129

[3] Stocker, M., Oelen, A., Jaradeh, M. Y., Haris, M., Oghli, O. A., Heidari, G., Hussein, H., Lorenz, A.-L., Kabenamualu, S., Farfar, K. E., Prinz, M., Karras, O., D'Souza, J., Vogt, L., & Auer, S. (2023). FAIR scientific information with the Open Research Knowledge Graph. In B. Magagna (Ed.), FAIR Connect (Vol. 1, Issue 1, pp. 19–21). IOS Press. https://doi.org/10.3233/fc-221513

[4] Kuhn, T., Banda, J. M., Willighagen, E., Ehrhart, F., Evelo, C., Malas, T. B., Dumontier, M., Merono-Penuela, A., Malic, A., Poelen, J. H., Hurlbert, A. H., Centeno Ortiz, E., Furlong, L. I., Queralt-Rosinach, N., & Chichester, C. (2018). Nanopublications: A Growing Resource of Provenance-Centric Scientific Linked Data. In 2018 IEEE 14th International Conference on e-Science (e-Science). 2018 IEEE 14th International Conference on e-Science (e-Science). IEEE. https://doi.org/10.1109/escience.2018.00024

[5] https://www.w3.org/TR/tabular-data-model/

[6] https://typeregistry.org/

[7] https://www.w3.org/TR/shacl/

[8] Granger, B. E., & Perez, F. (2021). Jupyter: Thinking and Storytelling With Code and Data. In Computing in Science & Engineering (Vol. 23, Issue 2, pp. 7–14). Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.1109/mcse.2021.3059263

[9] Vogt, L., Kuhn, T. & Hoehndorf, R. (2024). Semantic units: organizing knowledge graphs into semantically meaningful units of representation. J Biomed Semant 15, 7 (2024). https://doi.org/10.1186/s13326-024-00310-5

[10] https://www.researchobject.org/ro-crate/

[11] Vogt, L. (2023). Extending FAIR to FAIREr: Cognitive Interoperability and the Human Explorability of Data and Metadata. arXiv. https://doi.org/10.48550/arXiv.2301.04202

[12] Vogt, L. (2023). FAIR Knowledge Graphs with Semantic Units: a Prototype. arXiv. http://dx.doi.org/10.13140/RG.2.2.22809.49769

[13] Adler, A., Gujar, A., Harrison, B.L., O'Hara, K., Sellen, A. (1998). A diary study of work-related reading: design implications for digital reading devices. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '98, ACM Press, New York, New York, USA, pp. 241–248. https://doi.org/10.1145/274644.274679

[14] Tayeh, A.A.O., Signer, B. (2018). An Analysis of Cross-Document Linking Mechanisms. In: Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries, ACM, New York, NY, USA. pp. 69–78. https://doi.org/10.1145/3197026.3197053

[15] Vogt, L., Strömert, P., Matentzoglu, N., Karam, N. Konrad, M., Prinz, M., Baum, R. (2024). FAIR 2.0: Extending the FAIR Guiding Principles to Address Semantic Interoperability. arXiv. https://doi.org/10.48550/arXiv.2405.03345