Open Science Graphs Must Interoperate!

Amir Aryani[1][0000−0002−4259−9774], Martin Fenner[2][0000−0003−1419−2405], Paolo Manghi[3][0000−0001−7291−3210], Andrea Mannocci[3][0000−0002−5193−7851], and Markus Stocker[4,5][0000−0001−5492−3212]

[1] Swinburne University of Technology: Hawthorn, VIC, Australia
[2] DataCite e.V., Germany
[3] Institute of Information Science and Technology – CNR, Pisa, Italy
[4] TIB Leibniz Information Centre for Science and Technology, Hannover, Germany
[5] MARUM Center for Marine Environmental Sciences, PANGAEA Data Publisher for Earth & Environmental Science, Bremen, Germany

Abstract. Open Science Graphs (OSGs) are Scientific Knowledge Graphs whose intent is to improve the overall FAIRness of science, by enabling open access to graph representations of metadata about people, artefacts, institutions involved in the research lifecycle, as well as the relationships between these entities, in order to support stakeholder needs, such as discovery, reuse, reproducibility, statistics, trends, monitoring, impact, validation, and assessment. The represented information may span across entities such as research artefacts (e.g. publications, data, software, samples, instruments) and items of their content (e.g. statistical hypothesis tests reported in publications), research organisations, researchers, services, projects, and funders. OSGs include relationships between such entities and sometimes formalised (semantic) concepts characterising them, such as machine-readable concept descriptions for advanced discoverability, interoperability, and reuse. OSGs are generally valuable individually, but would greatly benefit from information exchange across their collections, thereby improving their efficacy to serve stakeholder needs. They could, therefore, reuse and exploit the data aggregation and added value that characterise each OSG, decentralising the effort and capitalising on synergies, as no one-size-fits-all solution exists. The RDA IG on Open Science Graphs for FAIR Data is investigating the motivation and challenges underpinning the realisation of an Interoperability Framework for OSGs. This work describes the key motivations for *i)* the definition of a classification for OSGs to compare their features, identify commonalities and differences, and added value and for *ii)* the definition of an Interoperability Framework, specifically an information model and APIs that enable a seamless exchange of information across graphs.

Keywords: Open Science · Research · Knowledge Graph · Interoperability

1 Introduction

The Open Science movement is urging scientists, communities, institutions, and policymakers to define and adopt methodologies, practices, and tools for open

publishing research artefacts beyond the scientific article, including research data, software, and digital experiments. As a consequence of this trend, researchers are increasingly depositing these artefacts and metadata about them, together with relationships among artefacts and other relevant contextual entities such as those described, in metadata registries about authors (e.g. ORCID[6]), organisations (e.g. ROR[7], GRID[8]), and data repositories (e.g. re3data[9]). De facto, Open Science publishing practices materialise a global and decentralised Open Science Graph.

Naturally, there is great interest to contribute to and/or consume such a graph for discovering and reusing artefacts as well as monitoring Open Science. To address this interest, several initiatives are building specialised Open Science Graphs (OSG), capable of serving specific user needs: Google Scholar, Scopus [3], Web of Science [4], Microsoft Academic Graph [17], FREYA PID Graph [7], Research Graph Foundation [1], OpenAIRE Research Graph [12], Open Research Knowledge Graph [11], Scholexplorer [5], Human Brain Project Knowledge Graph[10], Open Citations [14], Crossref [9], SciGraph[11], Semantic Scholar [8], Dimensions [10], as well as the CERIF[12] graphs built via Current Research Information System (CRIS) are just a few existing OSGs.

The fragmentation of these specialised OSGs motivates our interest to provide OSGs with an Interoperability Framework, whose drivers are manifold. First, interoperability would reduce duplication of effort and capitalise on synergies and complementarity. Second, interoperability enables information to circulate and thus ensures the enrichment and quality of individual OSGs as well as more redundancy to safeguard information availability and persistence. Third, interoperability will elevate OSGs as the backbone of Open Science scholarly communication.

The Research Data Alliance (RDA) Interest Group (IG) on Open Science Graphs for FAIR Data[13] is currently investigating the motivation and challenges underpinning the realisation of an Interoperability Framework for OSGs. The work presented here describes the motivations and challenges underlying the goal of an OSG Interoperability Framework, identified as:

– Need to define a classification for OSGs that supports assessing their value, compare their features, and identify differences. To this end, the presented preliminary analysis of the FREYA PID Graph, OpenAIRE Research Graph, Open Knowledge Research Graph, Research Graph, and Scholexplorer paves the way for a classification of OSGs.

[6] ORCID, `https://orcid.org`
[7] ROR, `https://ror.org`
[8] GRID, `https://grid.ac`
[9] Re3data, `http://re3data.org`
[10] The Human Brain Project, `https://www.humanbrainproject.eu`
[11] Springer Nature SciGraph, `https://scigraph.springernature.com`
[12] CERIF, `https://www.eurocris.org/cerif/main-features-cerif`
[13] RDA Interest Group on Open Science Graphs for FAIR Data, `https://rd-alliance.org/groups/open-science-graphs-fair-data-ig`

– Need to define an agreed-upon information model and APIs that enable the seamless exchange of information across OSGs.

The results of our preliminary investigation suggest that there is a need for a community-driven initiative that ensures common terminology (i.e. classification) and interoperability-enabled added value scholarly communication services that exploit the full potential of OSGs.

2 A Classification for Open Science Graphs

The fabric required to enact Open Science is a digital infrastructure based on an Interoperability Framework that captures research artefacts (in particular articles, datasets, software, services, workflows), metadata about artefacts, people and institutions as well as their relationships, as they evolve over time. This infrastructure relies on the adoption of Persistent Identifiers (PIDs) and metadata standards for the persistent identification and description of such entities across data sources (e.g. repositories, archives), thematic services (e.g. research infrastructures), and research communities.

Open Science Graphs (OSGs) are use case driven specialisations of Scientific Knowledge Graphs that build on the fabric of PIDs, metadata, and relationships. Figure 1 depicts OSGs in their context. Their scope differs according to served stakeholders, whose needs range from discovery, access, and reuse of research artefacts to monitoring and evaluating funding efforts, and identifying research trends. Stakeholder needs also drive the selection of data sources a particular OSG ought to integrate and the required data processing and enrichment capabilities.

The increasingly diverse and complex OSG landscape fuels an urgent demand for a classification to *i)* facilitate service providers in building needed added value services on OSGs, *ii)* assist consumers in selecting the services that meet their needs, and *iii)* facilitate OSG providers in identifying and communicating the characteristics of their service, and in understanding how to benefit from other OSGs.

In a first attempt to develop a classification framework that supports the systematic description of OSG characteristics and identification of their commonalities and differences, in the following, we introduce some existing OSGs that have been developed in recent years, namely: FREYA PID Graph, OpenAIRE Research Graph, Open Knowledge Research Graph, Research Graph, and Scholexplorer. This selection is by no means exhaustive. Indeed, additional initiatives do exist, e.g. Microsoft Academic Graph, SciGraph, Crossref, Dimensions, Semantic Scholar, Open Citations. Still, we argue that the selected OSGs reasonably represent the broader landscape.

2.1 Existing OSGs

FREYA PID Graph. The PID Graph is a scholarly infrastructure built by the partners in the EC-funded FREYA project, with the core infrastructure hosted

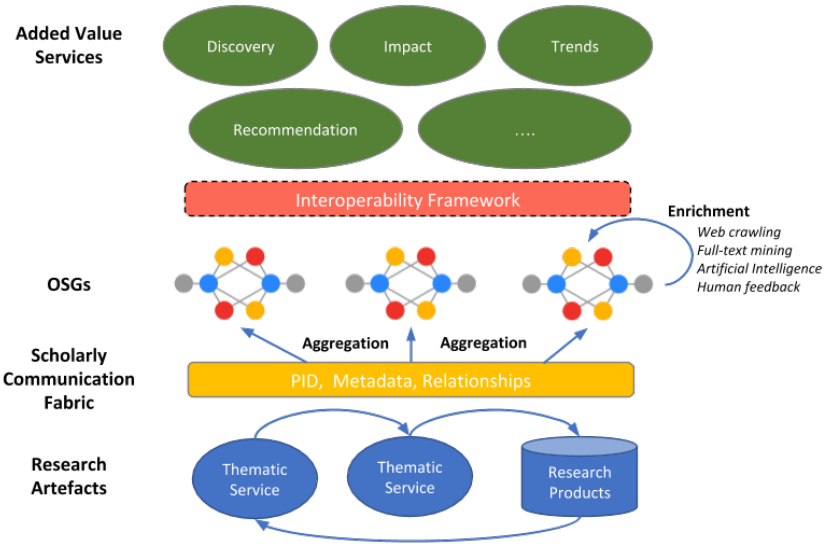

Fig. 1. OSGs and scholarly communication architecture)

by DataCite[14]. PID Graph identifies all nodes in the graph using persistent identifiers (PIDs) and describes these nodes, as well as the edges between nodes using the metadata associated with these PIDs. The graph is a federated graph, with PIDs and associated metadata provided by a number of PID providers who store this information in their respective services.

The development of the PID Graph is driven by user stories that the FREYA project partners have initially identified and that are continuously expanded. A distinguishing feature of these user stories is that they cannot be easily resolved with existing scholarly infrastructure, as they assume an underlying graph. Many of these user stories are around the discovery of connected resources, and the tracking of reuse.

The research entities supported by the PID Graph currently include publications, datasets, software, physical samples, instruments, services, people, research organisations, funders, and research data repository registries from the PID providers Crossref, DataCite, ORCID and ROR, for a total of currently about 35 million resources.

The PID Graph uses GraphQL[15] to query the PID Graph[16], a widely used open source technology that aims to make it easy to build client applications for the PID Graph. The fields that describe resources have been harmonised across resource types to simplify working with the PID Graph and to enhance connections between resources. Many Jupyter notebooks have been written to explore the PID Graph, and they are openly available for reuse. All information

[14] DataCite, `https://datacite.org`
[15] GraphQL, `https://graphql.org`
[16] DataCite API, `https://api.datacite.org/graphql`

in the PID Graph is available for reuse without restrictions, the software stack powering the PID Graph is available as open source software.

OpenAIRE Research Graph. The mission of the OpenAIRE initiative[17], one of the foundations of the European Open Science Cloud (EOSC)[18], is to provide training, dissemination, and technical services to seed (and support) Open Science publishing practices into the research lifecycle. To this end, one key activity of OpenAIRE aims at the construction of the OpenAIRE Open Research Graph by aggregating and integrating metadata records relative to digital research products (literature, dataset, software, and others) from more than 13,000 scholarly data sources world-wide (scientific repositories, archives, registries, databases, publishers), for a current total of more than 114 million publications, and 10 million research data. The graph is also algorithmically enhanced so to *i)* find and merge metadata records that describe the same entity (literature, and organisations), and *ii)* apply inference techniques on the metadata records and mine full-texts of Open Access publications to add new properties and new semantic relationships. End-user claims provided via the Web portal are also fed in the loop, so to drive the processing of raw metadata.

The OpenAIRE Research Graph data model is described in detail in [13] and its modelled entities are: literature, datasets, software, funders, funding streams, grants, organisations, researchers, data sources. Its content supports a number of analytics and applications such as discovery, research impact assessment, Open Science monitoring, brokering, reporting to funders, and statistics.

The graph is redistributed free of charge for everyone to use[19], both in bulk access mode (snapshot dump [12], OAI-PMH), and in selective access mode via APIs (REST Search API, LOD) under CC-BY licence, due to the fact the graph integrates sources with licences stricter than CC0 (e.g. Microsoft Academic Graph, Springer Nature).

Open Knowledge Research Graph. The Open Research Knowledge Graph[20] (ORKG) [11] is a scholarly infrastructure and open project led by the TIB Leibniz Information Centre for Science and Technology that aims to publish scholarly knowledge communicated in the literature in structured and semantic form.

The entity of primary interest to ORKG is therefore the research article (paper) and, importantly, article content. ORKG models article contents as "research contribution", an abstract concept that, in general terms, relates the problem addressed by a contribution with the materials and methods used, and the obtained results.

ORKG enables a range of new applications, including automated comparisons. As a classic example, it is possible to automatically compare the characteristics of sorting algorithms, e.g. their best and worst-case complexity. Given

[17] The OpenAIRE project, `https://www.openaire.eu`
[18] EOSC, `https://www.eosc-portal.eu`
[19] Access to the OpenAIRE Research Graph, `http://develop.openaire.eu`
[20] ORKG, `http://orkg.org`

precision, recall and F1 score of classification algorithms across the literature on a specific problem, say road-vehicle detection, it is possible to create leaderboards automatically, showing the trend of classification performance over time and the currently leading approach.

The primary data sources for ORKG are peer-reviewed research articles. In case data published in the literature (e.g. as a plot) is deposited in a research data repository, such infrastructures are an additional important data source. Furthermore, ORKG relies on third-party terminologies to align resources and thus ensure data interoperability and reusability.

ORKG adds value by making scholarly knowledge published in the literature better accessible to and processable by machines. As a multimodal infrastructure, ORKG integrates diverse data (i.e. scholarly knowledge) acquisition forms, specifically manual crowdsourcing, automated text mining, and scholarly knowledge exchange among research infrastructure, services and tools, e.g. data analysis environments such as Jupyter.

ORKG data export and provision is primarily via its REST API and SPARQL endpoint. Supported data formats are JSON and various RDF serialisations. As an open project and infrastructure, ORKG publishes software and data under open licenses (MIT and CC BY-SA, respectively).

Research Graph. Research Graph is a distributed network of scholarly works including data from data repositories such as NCI in Australia [15], academic and grey literature (e.g. GESIS, ICPSR), grants and funders (e.g. Australian Research Council, NIH) and researchers and research organisation information. Research Graph initially formed as a Graph Database by participants in the DDRI Working Group of Research Data Alliance [2] to connect datasets and metadata about data collections across repositories and data infrastructures. This graph later extended to a distributed network of graphs connecting via graph augmentation functionality running on a hybrid (national, private and commercial) cloud. At the time of writing this article, the graph holds close to 250 million nodes, including metadata about 180 million publications, 51 million datasets, 55 thousand grants, 1.4 million organisations, and 8.6 million researchers.

Research Graph is accessible to the partner organisations via Augment API, that is a cloud-hosted capability which creates graphs from bibliographic records, and extends this graph using information available in Research Graph clusters. The schema used for this transformation is based on the minimum required fields for identifying a research object, a trade-off between completeness and practicality, in favour of practicality. The graph schema [1] supports both XML, RDF XML and JSON-LD [16], and the endpoint supports Cloud Hosted Services, REST API and GraphQL.

Research Graph is mainly used by data infrastructures, repositories, and research systems for discovery of related scholarly works such as related datasets, and connections between grants and research output. Metadata about Research Graph is available on researchgraph.org, and github.com/researchgraph/schema,

the input API supports RDF, DDI, RIF-CS, Dublin Core, Scholix, DataCite, Crossref, and many other metadata formats, and the output includes Research Graph Schema (JSON, XML), JSON-LD and RDF XML. Research Graph includes a subgraph reusable under CC-By licence while some other parts are accessible for limited use only under NC-ND-SA-CreativeCommons.

Scholexplorer. Scholexplorer[21] [5] populates and provides access to a graph of Scholix [6] links between dataset and literature objects, and between dataset and dataset objects. Links (and objects) are provided by data sources managed by publishers, data centres, or other organisations providing services to store and manage links between data sets and publications such as CrossRef, DataCite, PubMed, EMBL-EBI data sources, Pangaea, and OpenAIRE. Scholexplorer aggregates links metadata harvested from these data sources as Scholix records and out of these builds a harmonised and de-duplicated graph of scholarly objects counting today over 21 million publications, 53 million datasets, and over 269 million bi-directional semantic links between them. The graph is openly accessible under CC-BY licence via REST search APIs that return links in Scholix format, and via periodic dumps on Zenodo[22].

2.2 Classification

Based on a comparison of the five initiatives described above (Figure 2), we propose a first classification across seven main features, regarded as relevant to both OSG consumers and OSG providers.

1. **Research entities**: Each OSG operates at a specific level of abstraction, and consequently models (including PIDs and metadata descriptions) specific research entities of interest to the scholarly communication domain, such as research artefacts (e.g. publications, data, software, samples, instruments) and concepts therein expressed (e.g. hypothesis, methods, algorithms, protocols), research organisations, researchers, services, projects, and funders.
2. **Applications:** OSGs are designed to serve specific use cases, which may range from the discoverability of research artefacts to the assessment of research impact, from quantitative science studies (e.g. science of science) to the computation of usage statistics, and monitoring, etc.
3. **Data sources**: OSGs are constructed by collecting and integrating information from different types of data sources (e.g. journals, repositories, archives, registries, other OSGs, etc.), or selections of such sources (e.g. thematic or discipline-driven, from a geographical area, related to organisations), to specific individual sources.
4. **Added value**: OSGs are often characterised by integration and enrichment techniques, which manipulate the aggregated metadata in several ways; for example, by *i)* harmonising metadata to map it onto the OSG information

[21] Scholexplorer, `http://scholexplorer.openaire.eu`
[22] Scholexplorer dump, `https://zenodo.org/record/3541646`

model (e.g. metadata structural and semantic transformations), and *ii)* by enhancing the metadata via web crawling, interlinking, inference, full-text mining, AI, user annotations and feedback, etc.

5. **Data export and provision:** OSGs offer access to their content via APIs (e.g. OAI-PMH, SPARQL, GraphQL, ad-hoc REST APIs, etc.) and standard exchange formats (e.g. XML, JSON, RDF) that implement standard metadata formats (e.g. DataCite, Scholix.org, Dublin Core, ORCID profile, CERIF) or proprietary formats.
6. **FAIRness:** FAIRness of OSGs regards their nature as dataset in regard to being Findable, Accessible, Interoperable, and Reusable. Practices vary, but in general, OSGs are available via standard exchange formats (e.g. XML, JSON, RDF) and accessible via standard protocols, from simple download to GraphQL, OAI-PMH, or proprietary search REST APIs. In some cases, accessibility is facilitated by minting a DOI to the OSG collection, and, in some cases, complicated by the fact consumers need to go through toll-gated cloud services to access the graph. OSG schemata give life to the hardest interoperability and reusability challenge, as they follow application-driven interpretations of research entities, which complicate OSG reuse and integration.
7. **Openness:** Different OSGs are released and redistributed under different licences (CC0, CC-BY, CC-SA, etc.). In general, the licence applies to the whole graph, but, in some cases, different parts of the graph can be released under different licences, be accessible only to a limited number of stakeholders, or be behind a paywall. In other cases, for example for Microsoft Academic Graph, the graph is released openly with ODC-BY licence, but a small fee is needed to sustain the provisioning platform (i.e. Azure).

While the table already provides evidence for the value of a classification, it also highlights the need for common agreements on classification criteria. For example, aspects such as coverage of the data sources aggregated by the OSG may be of interest, as a graph may focus on a geographical region, be cross-community or community-specific (e.g. computer science and neuroscience in the early Semantic Scholar), or be able to capture geospatial descriptions (e.g. INSPIRE in Research Graph).

3 A Framework for Open Science Graphs Interoperability

We advocate for the establishment of a community-driven Interoperability Framework in order to mediate the diverse data models and technologies used by existing OSGs. The drivers for conceiving an Interoperability Framework for OSGs are manifold.

Firstly, as we have seen in Section 2, the various OSGs differ in scope, extent and technological details as they strive to capture various aspects of scholarly communication from diverse perspectives and different abstraction and granularity. Thus, the information pertaining to different OSGs can be overlapping or

can be complementary. With overlap we gain plurality, e.g. different identifiers for same papers, authors, organisations, etc. while with complementarity we gain completeness and coverage, e.g. integrate information of various granularity as published by various OSGs.

Secondly, despite building on data sources with clear sustainability plans, some OSGs have unclear directions, lack a viable business model, and thus might cease to exist. Given this risk, OSG content should be federated, shared, and possibly fed back to original data sources where it can be managed sustainably for the common good of both the scientific community and, more broadly, society.

Thirdly, OSGs and more generally Scientific Knowledge Graphs should act as the backbone of modern Open Science scholarly communication, embody its core principles, and foster its adoption along several dimensions such as discoverability, monitoring, and FAIRness. This is especially relevant for the non-commercial OSGs and their leading role in open innovation with best-in-class, cutting-edge services, free at the point of use.

It is therefore of paramount importance to exchange OSG content and capitalise on the non-negligible acquisition, integration and enrichment efforts performed by the various OSGs. To facilitate information exchange between OSGs, the Interoperability Framework may rely on an agreed-upon *lingua franca*. This was already achieved with the specification of Scholix [6], an agreed-upon high-level interoperability framework for exchanging information about the links between scholarly literature and data. However, Scholix operated within a much narrower scope. Given the complexity of the modelled information and the ambition of the endeavour, for OSGs a set of "dialects" rather than a single lingua franca may be more viable while still efficiently catalyse interoperability.

OSG content exchange has to occur on at least two levels of abstraction: *information model* and *technological*. In regard to information modelling, there is an urge for the various OSGs to define bottom-up a common model that can maximise information exchange and has the flexibility to accommodate unforeseen extensions, use cases, and stakeholders. From a technological standpoint, we need a portfolio of operational frameworks supporting a seamless exchange of information across different OSGs by means of operators/primitives. Doing so implies supporting a plethora of exchange formats (and the relative mappings to the common model) such as CSV, XML, RDF, JSON-LD, Scholix, and OAI-ORE, as well as different APIs enabling the provisioning of OSG information such as REST, SPARQL, and GraphQL.

To this end, we envisage the European Open Science Cloud (EOSC) as one optimal channel through which such an Interoperability Framework for OSGs could be developed via consensus and for the benefit of Open Science, at least at a pan-European level. EOSC is being constructed having a System of Systems paradigm, where local autonomy and differences are fostered as they can be an added value, and where convergence is recommended and facilitated via common interoperability frameworks to optimise cost and maximise the efficiency of science. OSGs would, in such an ecosystem, become the mean for *i)* bridging research infrastructures, i.e. thematic and scholarly communication services, and

ii) offer to EOSC users, such as researchers, research communities, policymakers, and SMEs the tools to discover and monitor trends and impact of science. The RDA IG on Open Science Graphs for FAIR Data is and will be contributing to the definition of the EOSC interoperability frameworks to ensure that specific solutions will be sought after. Finally, another channel that is potentially conducive so to bring this discussion onto the global landscape and the long-term perspective of a broader "Global Open Science Cloud" could be the RDA IG on Global Open Research Commons[23].

4 Conclusions

In this paper, we targeted two challenges of working with Open Science Graphs (OSGs). On the one hand, OSGs would benefit from a classification framework that enables their inspection and comparison along key features. On the other hand, we argue that an Interoperability Framework is pivotal to enable a seamless exchange of information among OSGs with the resulting suggested benefits.

We proposed a preliminary classification framework by analysing a selection of representative OSGs, namely: FREYA PID graphs, OpenAIRE Research Graph, Open Knowledge Research Graph, Research Graph, and Scholexplorer. Moreover, we outlined the main drivers and *desiderata* of a possible Interoperability Framework.

Going forward, we see the RDA Interest Group on Open Science Graphs for FAIR Data as an important community to make further progress on aligning the various OSG initiatives, in particular concrete work on interoperability.

Acknowledgements

This work was co-funded by the EU projects OpenAIRE-Advance (Grant agreement ID: 777541), FREYA (Grant agreement ID: 777523), ScienceGRAPH (Grant agreement ID: 819536), and the TIB Leibniz Information Centre for Science and Technology.

References

1. Aryani, A., Poblet, M., Unsworth, K., Wang, J., Evans, B., Devaraju, A., Hausstein, B., Klas, C.P., Zapilko, B., Kaplun, S.: A Research Graph dataset for connecting research data repositories using RD-Switchboard. Scientific Data **5** (2018). https://doi.org/10.1038/sdata.2018.99
2. Aryani, A., Wang, J.: Research graph: building a distributed graph of scholarly works using research data switchboard. In: Open Repositories Conference (2017). https://doi.org/10.4225/03/58c696655af8a

[23] RDA IG on Global Open Research Commons, `https://www.rd-alliance.org/groups/global-open-research-commons-ig`

3. Baas, J., Schotten, M., Plume, A., Côté, G., Karimi, R.: Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies. Quantitative Science Studies **1**(1), 377–386 (2020). https://doi.org/10.1162/qss_a_00019
4. Birkle, C., Pendlebury, D.A., Schnell, J., Adams, J.: Web of Science as a data source for research on scientific and scholarly activity. Quantitative Science Studies **1**(1), 363–376 (2020). https://doi.org/10.1162/qss_a_00018
5. Burton, A., Koers, H., Manghi, P., La Bruzzo, S., Aryani, A., Diepenbroek, M., Schindler, U.: The data-literature interlinking service: Towards a common infrastructure for sharing data-article links. Program **51**(1), 75–100 (2017). https://doi.org/10.1108/PROG-06-2016-0048
6. Burton, A., Koers, H., Manghi, P., Stocker, M., Fenner, M., Aryani, A., La Bruzzo, S., Diepenbroek, M., Schindler, U.: The scholix framework for interoperability in data-literature information exchange. D-Lib Magazine **23**(1/2) (2017). https://doi.org/10.1045/january2017-burton
7. Fenner, M., Aryani, A.: Introducing the pid graph (2019). https://doi.org/10.5438/jwvf-8a66
8. Fricke, S.: Semantic scholar. Journal of the Medical Library Association **106**(1), 145–147 (2018). https://doi.org/10.5195/jmla.2018.280
9. Hendricks, G., Tkaczyk, D., Lin, J., Feeney, P.: Crossref: The sustainable source of community-owned scholarly metadata. Quantitative Science Studies **1**(1), 414–427 (2020). https://doi.org/10.1162/qss_a_00022
10. Herzog, C., Hook, D., Konkiel, S.: Dimensions: Bringing down barriers between scientometricians and data. Quantitative Science Studies **1**(1), 387–395 (2020). https://doi.org/10.1162/qss_a_00020
11. Jaradeh, M.Y., Oelen, A., Farfar, K.E., Prinz, M., D'Souza, J., Kismihók, G., Stocker, M., Auer, S.: Open research knowledge graph: Next generation infrastructure for semantic scholarly knowledge. In: Proceedings of the 10th International Conference on Knowledge Capture. p. 243–246. K-CAP '19, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3360901.3364435
12. Manghi, P., Atzori, C., Bardi, A., Shirrwagen, J., Dimitropoulos, H., La Bruzzo, S., Foufoulas, I., Löhden, A., Bäcker, A., Mannocci, A., et al.: Openaire research graph dump (Dec 2019). https://doi.org/10.5281/zenodo.3516918
13. Manghi, P., Bardi, A., Atzori, C., Baglioni, M., Manola, N., Schirrwagen, J., Principe, P.: The openaire research graph data model (Apr 2019). https://doi.org/10.5281/zenodo.2643199
14. Peroni, S., Shotton, D.: OpenCitations, an infrastructure organization for open scholarship. Quantitative Science Studies **1**(1), 428–444 (2020). https://doi.org/10.1162/qss_a_00023
15. Wang, J., Aryani, A., Evans, B., Barlow, M., Wyborn, L.: Graph connections made by rd-switchboard using nci's metadata. D-Lib Magazine **23**(1/2) (2017). https://doi.org/10.1045/january2017-aryani
16. Wang, J., Aryani, A., Wyborn, L., Evans, B.: Providing research graph data in json-ld using schema.org. In: Proceedings of the 26th International Conference on World Wide Web Companion. pp. 1213–1218 (2017). https://doi.org/10.1145/3041021.3053052
17. Wang, K., Shen, Z., Huang, C., Wu, C.H., Dong, Y., Kanakia, A.: Microsoft Academic Graph: When experts are not enough. Quantitative Science Studies **1**(1), 396–413 (2020). https://doi.org/10.1162/qss_a_00021

	Research Entities	Applications	Data Sources	Added value	Data Export	FAIRness	Openness
PID graph	**PIDs:** DOI, ORCID, ROR, Crossref Funder ID **Entities:** publications, datasets, software, people, research organizations, funders, research data repository registries	Discovery, Research Impact, Open Science Monitor, Brokering, Reporting to funders, Statistics	PID providers	Standard GraphQL query interface with client libraries available in many languages. Strong support for Jupyter notebooks	**API**: GraphQL	**Findable:** Graphs are searchable by PIDs and metadata with GraphQL API **Accessible**: access described entities via PID **Interoperable**: Input: DataCite XML, Crossref XML, ORCID XML. Output: GraphQL JSON **Reusable**: Graphs are reusable	Redistributed free of charge
OpenAIRE Research Graph	**PIDs:** DOI, ORCID, accession numbers, PMCID, URLs, MAG IDs **Entities:** Literature, datasets, software, funders, grants, organizations, researchers, data sources **Relationships:** DataCite relationships, fundedBy, similarTo, hasAuthoraffiliatedWith	Discovery, Research Impact assessment, Open Science Monitoring, Brokering, Reporting to funders, Statistics	Any data source trusted by scientists: repositories, archives, registries, databases, publishers	Enrichment of metadata and relationships by full-text mining, User-feedback, Inference by context propagation, deduplication, Provenance tracking	**Format:** OpenAIRE XML format **APIs:** LOD, OAI-PMH, Dumps, REST Search APIs	**Findable:** searchable on Zenodo, accessible by DOI **Accessible**: every entity is openly accessible via HTTP API and dump **Interoperable**: OpenAIRE XML, documented online **Reusable**: Content can be reused with CC-BY licence	Redistributed free of charge under CC-BY licence
ORKG	**PIDs:** DOI, ORCID, URLs **Entities:** Literature and items of its content **Relationships:** addressed problems, utilized materials, employed methods, yielded results.	At the granularity of items of scholarly literature content, discovery, comparison, recommendation, visualization, reuse	Literature, research data repositories, terminologies, LIMS/ELN	Multimodal infrastructure for the acquisition, curation and publishing of machine-actionable scholarly knowledge.	**Format**: JSON and RDF serializations **APIs:** REST API, SPARQL	**Findable**: ORKG search, but currently lacking findability through 3rd party systems. Plans to assign DataCite DOIs to elements of ORKG content. All resources are URL-identified. Neo4j dumps or RDF exports can be deposited in a suitable data repository. **Accessible**: All ORKG resources are URL-retrievable, accessible and have their own descriptive landing page. Content can be accessed programmatically via REST API. **Interoperable**: Data use a graph-based data model, but currently lacking formal semantics. Alignment with external terminologies is technically possible but not broadly practised. **Reusable**: Content is reusable under CC BY-SA licence, provenance is tracked. Neo4j dumps or RDF exports can be made available for reuse.	Data and Software are released under CC BY-SA and MIT licences, respectively.
Research Graph	**PIDs:** DOI, ORCID, PURL, ISNI, GRID, PMCID, Scopus_ID **Entities:** academic articles and grey literature, datasets, funders, grants, organizations, researchers, datasets **Relationships:** authorship, funding, citation, usage, known_as, employment, custodian, management, etc.	Supporting repositories and research infrastructures	PID providers, data repositories, publishers, funders, discovery services and aggregators	Identity resolution, metadata enhancement, topic modelling, clustering, text mining and GIS mapping	**Format**: XML and JSON **APIs:** Cloud Hosted Services, REST API and GraphQL	**Findable:** Metadata available via researchgraph.org **Accessible**: Controlled Access available **Interoperable**: Input: RDF, DDI, RIF-CS, Dublin Core, Scholix, DataCite, Crossref, and many other metadata formats. Output: RDF and Research Graph Schema (JSON, XML) **Reusable**: Includes subgraph reusable under CC-BY licence while some other parts only accessible for limited use under NC-ND-SA-Creative Common	Controlled Access
Scholexplorer	**PIDs:** DOI, accession numbers, PMCID, URLs **Entities:** Literature, datasets **Relationships:** citedBy, supplementedBy, references	Discovery, Statistics, PID Resolution	DataCite, CrossRef Event Data, EMBL-EBI (EPMC), OpenAIRE, Scholix-compliant data sources	Deduplication by PID, provenance tracking	**Format:** Scholix [?] **APIs:** Dumps on Zenodo, REST Search APIs	**Findable:** searchable on Zenodo, accessible by DOI **Accessible**: every entity is openly accessible via REST API and dump **Interoperable**: Scholix links **Reusable**: Content is released under CC-BY licence	Redistributed free of charge under CC-BY licence

Fig. 2. Classification of OSGs initiatives