RDA WG Case Statement "Scientific Knowledge Graphs - Interoperability Framework" (SKG-IF)

Table of content

Overview

The Scientific Knowledge Graph - Interoperability Framework (SKG-IF) Working Group (WG) will target the definition of a framework to enable a seamless exchange of information among diverse initiatives regarding Scientific Knowledge Graphs, intended as knowledge bases of scholarly knowledge content (e.g. repositories, databases, catalogues, knowledge graphs, LOD collections). The WG stems from and incorporates the activities of the pre-existing RDA Interest Group (IG) on "Open Science Graphs for FAIR Data"[1] and tightly liaises with it. As motivated in the case statement of the IG, the topic is highly relevant, and a priority globally, with this specific RDA IG featuring in several calls for project proposals by the European Commission in 2022.[2] The change of name, from “Open Science Graphs” in the IG to “Scientific Knowledge Graphs” in the WG, stems from the specific name convention used by the European Commission in the EOSC-related working programme.

Charter

The Open Science movement is encouraging scientists, communities, research organisations, and policymakers to define and adopt methodologies, practices, and tools for

[1] https://rd-alliance.org/groups/open-science-graphs-fair-data-ig

[2] https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/wp-call/2021-2022/wp-3-research-infrastructures_horizon-2021-2022_en.pdf

the open publishing of research artefacts beyond the scientific article, thus including research data, software, and in silico experiments.

As a consequence of this trend, policymakers and funders have adopted Open Science policies so that researchers are increasingly depositing these artefacts and their metadata descriptions, together with relations among artefacts and other relevant contextual entities, such as those described, in repositories available online, i.e., catch-all, general purpose repositories (e.g., Zenodo, Figshare), institutional and thematic (e.g., arXiv, ChEMBL), methodology repositories (e.g., Protocols.io), research software repositories (e.g., Software Heritage reproducibility and data management suites (e.g., WholeTale, Gigantum, Globus). Similarly, services born for causes other than research have been adopted and re-purposed to benefit science and are currently used as means for the deposition of research products (e.g., GitHub, Bitbucket, Google Colaboratory, Docker). At a higher level of the hierarchy, scholarly registries keep information on literature and data repositories (e.g., FAIRsharing, OpenDOAR, ROAR, re3data), authors' details (e.g. ORCID), organisations (e.g. ROR), projects and funding (e.g., CORDIS).

De facto, Open Science publishing practices unfold the very fabric of scientific progress by digitally materialising a global and decentralised research record and the current status of the scientific forefront.

Naturally, there is a great interest in contributing to and/or consuming such information in order to serve a broad range of applications and studies. To this end, several initiatives are, more or less independently, building specialised Scientific Knowledge Graphs (SKGs), providing ad-hoc views of such research global record capable of serving specific user needs - the new name convention, “SKG vs OSG”, was introduced by the European Commission work programme; the WG adopted it for the sake of clarity. Google Scholar, Scopus, Web of Science, Microsoft Academic Graph (alas, phased out at the end of 2021), OpenAlex, Dimensions, the OpenAIRE Graph, OpenCitations, Crossref, the Research Graph Foundation, the Open Research Knowledge Graph (ORKG), the Human Brain Project Knowledge Graph, SciGraph, Semantic Scholar, as well as several Current Research Information Systems (CRIS) are just a few existing SKGs.

The high fragmentation and heterogeneity of information within these SKG motivate our interest in providing them with an Interoperability Framework (SKG-IF) whose drivers are manifold. Firstly, interoperability would reduce duplication of effort and capitalise on synergies and complementarity. Secondly, interoperability would enable information to circulate and thus ensures the enrichment and enhancement of individual SKGs both quantitatively and qualitatively, as well as more redundancy to safeguard information availability and persistence. Thirdly, interoperability would elevate SKGs as the backbone of modern Open Science scholarly communication.

The activities coordinated by the RDA Interest Group (IG) on "Open Science Graphs for FAIR Data" are currently investigating the motivation and challenges underpinning the realisation of an Interoperability Framework for Scientific Knowledge Graphs (SKGs). Believing that SKGs could (and should!) interoperate more and therefore exchange the locally available wealth of information and added value to benefit the ecosystem globally operating in Open Science and Scholarly Communication, the Scientific Knowledge Graph - Interoperability Framework (SKG-IF) Working Group (WG) will foster the definition of guidelines towards devising a framework to achieve information exchange across the diverse SKGs willing to collaborate.

By bridging academia and industry and bringing together practitioners and stakeholders with diverse backgrounds and interests, the WG will constitute a fertile ground to spark the discussion and converge on a pragmatic approach towards interoperability across SKGs.

Description of Work

The SKG-IF WG will facilitate the work towards the definitions of guidelines for the implementation of an Interoperability Framework for Scientific Knowledge Graphs. In doing so, the WG will promote and include the following principles:

- The FAIR Data Principles[3]: Findable, Accessible, Interoperable, and Reusable;
- The RDA Guiding Principles[4]: Openness, Consensus, Inclusive, Harmonization, Community-driven, Non-profit and technology-neutral;
- The RDA Code of Conduct[5].

Furthermore, the advancements in terms of interoperability will enhance the machine-actionability of information and semantic relations contained in the various graphs.
The SKG-IF WG concretely responds to the demand of the research community, policymakers, and industry to achieve consensus on the way scholarly communication and research ecosystem data are exchanged. The Work Programs of the European Commission funded grants (and will fund others) to implement the results of this working group, which are explicitly mentioned in the Calls.

Objectives

The primary objective of the SKG-IF WG is to outline the guidelines towards a framework for interoperability among SKGs.

The following specific objectives will be pursued:

1. Analysis of an extensive selection of Scientific Knowledge Graphs, standards and metadata formats and how they model the relevant information;
2. Study of the main entities represented (e.g., literature, research data, software, venues, authors, funding) and relation among such entities;
3. Analysis of the main attributes of such entities and specification of crosswalk mappings across different models;
4. Study of the main standards and protocols available for the development of an Interoperability Framework.

Organization and Activities

The SKG-IF WG will meet monthly (tentatively), taking over the actions initiated within the SKG for FAIR Data IG from June 2022. We foresee the WG will be running until Dec 2023. The WG is organised into four distinct task forces (TFs) with different objectives and life cycles. These are defined as follows:

- **Task Force 1 - SKG-IF Core Information Model** (Jun - Nov 2022)

[3] https://www.nature.com/articles/sdata201618
[4] https://www.rd-alliance.org/get-involved.html
[5] https://rd-alliance.org/rda-code-conduct-and-how-report-breach

    - Identification of the entities and relevant relations
    - Identification of common real-scenario use cases (e.g., OpenAlex, SemanticScholar, Scopus, OpenAIRE Graph, OpenCitations, PID Graph, ResearchGraph, Open Research Knowledge Graph)
    - Survey on (de facto) standard information models, PIDs, for the entities (e.g., CERIF, DataCite, JATS, Dublin Core, ROR.org)
    - Definition of the Core Information Model (incremental, introducing entities one by one)
- **Task Force 2 - SKG-IF Data Exchange Commons** (Nov 2022 - Feb 2023)
    - Survey on (de facto) Standard metadata formats for the entities (e.g. CERIF, DataCite, JATS, Dublin Core, ROR.org, OpenCitations Data Model) - performed as part of the Core Information Model
    - Definition/identification of Core Metadata Format
    - Crosswalks from known standards to Core Information Model (incremental, follows Information Model approach)
    - Identification of exchange formats and definition of the schema for the Core Information Model
- **Task Force 3 - SKG-IF Access Protocol Commons** (Feb - Mar 2023)
    - Analysis of (de facto) standard protocols and approaches for exchanging SKG data (e.g. RESTful API, GraphQL, SPARQL, etc.)
    - Identification of a set of protocols to be adopted by SKGs to exchange data according to the Core Information Model
- **Task Force 4 - SKG Profiles** (Apr - Jun 2023)
    - Definition of a high-level SKG profile, describing the main features of an SKG at hand, such as main entities represented, schema, ID/PID system, access protocols, frequency of update, adopted vocabularies, locally added value, algorithms applied, level of compliance towards the Interoperability Framework, and so on. Such profiles will serve the purpose of easing the side-by-side comparison of different SKGs and understanding which is the most suitable one for the application at hand.
- **Task Force 5 - SKG-IF Data Exchange Commons - 2nd round** (Apr - Sep 2023)
    - Extension of the Core Information Model with information regarding trust, provenance (sources, algorithms), or annotations of the SKG entities and relationships

The co-chairs will share responsibility for developing the meeting agendas, organizing meetings, leading discussions, keeping track of the work and deadlines, and arranging participation in plenary sessions. Specific task groups will be established to progress on the identified deliverables, with regular reporting. A mailing list will be used for asynchronous communications[6]. Other tools, such as shared worksheets and documents, may be used for collaborating on and tracking deliverables (e.g., on Google Suite).

[6] osg-if@googlegroups.com

Deliverables

The SKG-IF WG is planning to produce one deliverable by the end of its activities about the agreed guidelines towards the implementation of an Interoperability Framework across SKGs. The deliverable will serve as means of discussion and will be refined in a stepwise fashion, and it will cover the following aspects:

1. **SKG-IF Core Information Model and Interoperability Mappings** as resulting from the activities of TF 1 and TF 2. The deliverable will cover the description of the main entities represented by the Core Model, their attributes and crosswalks towards real-world SKGs.
2. **SKG Interoperability Framework v1.0:** as resulting from the activities of TF 2, the deliverable will describe the metadata exchange formats and related schema onto which SKGs metadata should be exchanged;
3. **SKG Interoperability Framework v2.0:** as resulting from the activities of TF 3, the deliverable will showcase the technical options that can be leveraged to access and exchange SKG metadata content
4. **SKG Profiles Handbook** as resulting from the activities of TF 4. The deliverable will outline the specifications to publish SKG profiles online.

Operational and Adoption Plan

Months 0-12 will be dedicated to the survey and implementation phase. Regular monthly virtual meetings will be held among co-chairs, members and the operational team, using web meetings functionality. Monthly virtual meetings, or more often as required, will be held between the co-chairs and TFs members. When possible, face-to-face meetings will be organized, particularly in conjunction with existing and relevant conferences, RDA plenaries and events. The co-chairs will be responsible for moderating the discussion and driving the development to meet the deliverables according to the timelines. Conflicts and any time adjustments on the timeline developments will be managed and addressed by the co-chairs, as appropriate.

Within **months 12-18**, we will initiate specific activities aiming toward the dissemination and adoption of the results produced, beyond the members of the group. The specific plan for encouraging adoption will include publications and presentations via RDA, EOSC, OpenAIRE, etc. meetings, and those of other partners. Early endorsement by the core members will also be used as adoption exemplars to other communities.

Background & Founding Members

The SKG-IF Working Group develops out of the Scientific Knowledge Graphs from FAIR Data Interest Group and continues the activities herein initiated towards the definition of an Interoperability Framework for SKGs.

The following members of that Interest Group now continue with the SKG-IF WG:

- **Andrea Mannocci** (he/his) is a Staff Research at CNR-ISTI in Italy. He currently works as a data scientist within the framework of the EU projects OpenAIRE Nexus, GraspOS, and SciLake. His research interests span from the analysis of enabling

services for Open Science to Science of Science, complex networks and the analysis of research as a global-scale phenomenon inserted in a delicate socioeconomic and geopolitical context. He obtained his PhD degree in Information Engineering from the University of Pisa (Italy). He co-organised the Scientific Knowledge Graph (SKG 2020) workshop and the Sci-K workshop series at the Web Conference from 2020 to 2023, and previously the Reframing Research (RefResh) workshop in 2018 and 2020 at the European Computational Social Science Symposium and SocInfo 2020 respectively. He also acted as a PC member in several international conferences and workshops.

- **Jason Priem** (he/his) is a co-founder of OurResearch (previously ImpactStory), a non-profit organisation building open software to promote the growth of Open Science. Since its foundation, Our Research has received funding from Arcadia, the US National Science Foundation, the US National Institutes of Health, the Alfred P. Sloan Foundation, and others. With OurResearch, Jason led the forefront development of successful scholarly services and platforms such as OpenAlex and Unpaywall. As a PhD student at the University of North Carolina at Chapel Hill, Jason led the creation of the field of altmetrics, coining the term, authoring the influential Altmetrics Manifesto, and leading the first altmetrics workshop.
- **Silvio Peroni** (he/his) is an Associate Professor at the University of Bologna. He is an expert in document markup and semantic descriptions of bibliographic entities using OWL ontologies, and he is one of the main developers of the SPAR (Semantic Publishing and Referencing) Ontologies, Director (with David Shotton) of OpenCitations, and founding member of the Initiative for Open Citations (I4OC) and the Initiative for Open Abstracts (I4OA). Among his research interests are Semantic Web technologies, markup languages for complex documents, design patterns for digital documents and ontology modelling, and automatic processes of analysis and segmentation of documents. His works concern theoretical studies and technical implementation of tools to foster semantic interoperability of Open Science services and infrastructures, the empirical analysis of the nature of scholarly citations, bibliometrics and scientometrics studies, visualisation and browsing interfaces for semantic data, and the development of ontologies to manage, integrate and query bibliographic information.
- **Sahar Vahadati** (she/her) is currently leading the research group named Nature Inspired Machine Intelligence at the Institute of Applied Informatics in Germany. She has done her PhD on the collaborative integration, publishing and analysis of distributed scholarly metadata. She developed OpenResearch, a semantic wiki-based platform for creating and curating scholarly knowledge graphs. Her work on this was the initiative of Open Research Knowledge Graph (ORKG). She was involved in the devolvement of several ontologies for scholarly domains, especially for review papers, and events. She was also a member of the technical team in the OpenAIRE project when working at the University of Bonn. She is a fellow scientist at Oxford University, where she worked on the unitisation of logical rules in knowledge graph embedding models, and on the example of scholarly graphs. Currently, she is focusing on providing representation learning approaches for the automated completion of scholarly knowledge graphs using predictions, as well as recommendation services for scientific communities.

Membership

Initially, former members of the RDA IG on "Scientific Knowledge Graphs for FAIR Data" will be surveyed for their preference and availability to take places in the work of the four constituted Task Forces. The SKG-IF WG will strive for an open and inclusive membership across countries, regions, and sectors and for fair coverage of different stakeholders and be representative of both academia and industry.

The SKG-IF WG will reach out across the RDA international community on its neutral social platform and assist in furthering Open Science and Scholarly communication common goals.

Timeline

The SKG-IF will complete its activities initiated in June 2022 within the RDA IG on "Scientific Knowledge Graphs for FAIR Data" and will conclude throughout an 18-month period, utilizing monthly web meetings (not reported for the sake of space in the timeline below), work streams, and focused discussions on its deliverables. The deliverables will all be due by the end of Dec 2023 in their latest revision.

Date	Activity
20 Jun 2022	Task Force kick-off @ RDA P19 in Seoul, South Korea
Nov 2022	Milestone 1: Task Force 1 wraps up and produces a first worksheet with the identified entities and relations of the Core Information Model. Task Force 2 starts off.
Feb 2023	Milestone 2: Task Force 2 wraps up and delivers a crosswalk from different models and standards to the identified Core Information Model. Task Force 3 starts off.
Mar 2023	Milestone 3: Task Force 3 wraps up and produces a survey on access protocols and standards to achieve information exchange. @ RDA P20 early results from the task forces will be presented.
Apr 2023	Task Force 4 and 5 start off. Deliverable “SKG Interoperability Framework v1.0” is released for public consultation.
May 2023	Deliverable “SKG Interoperability Framework v1.0” is finalised. Deliverable “SKG Interoperability Framework v2.0” is released for public consultation.
Jun 2023	Milestone 4: task Force 4 wraps up and delivers a high-level definition of the SKG profile. Deliverable “SKG Interoperability Framework v2.0” is finalised.
Jul 2023	Deliverable "SKG Profiles Handbook" is released for consultation.
Aug 2023	Deliverable "SKG Profiles Handbook" is finalised.

Date	Activity
Sep 2023	Task Force 5 wraps up on Core Information Model updates, and a new model release is planned. Dissemination of the achieved results.
Dec 2023	The WG closes its operation and plans maintenance of the Interoperability Framework as part of the OSG for FAIR data Interest Group and as an EOSC Interoperability Framework Working Group