Open Science Graphs Must Interoperate!

Amir Aryani, Martin Fenner, Paolo Manghi, **Andrea Mannocci**, Markus Stocker

Consiglio Nazionale delle Ricerche

ISTITUTO DI SCIENZA E TECNOLOGIE DELL'INFORMAZIONE "A. FAEDO"

Outline

- Introduction: Open Science and Open Science Graphs
- Drive for a change
  - A classification framework for OSGs
  - An OSGs interoperability framework
- Conclusions

Introduction

The **Open Science movement** is urging researchers to leave traces of their daily quest

- Deposition of multiple entities **beyond the traditional literature**
  - Research data
  - Software
  - Ideas & concepts
  - Tools & protocols
  - Methods & results
  - etc.
- Such information ends up in a **plethora of different services** scattered across the Web

Introduction

Rich set of relations among entities and towards authoritative registries

- Data registries (e.g. re3data)
- Authors and authorships (e.g. ORCID)
- Affiliation (e.g. GRID.ac, ROR)
- Funding (e.g. CORDA)

Great interest and unprecedented opportunities in contributing/consuming such information for

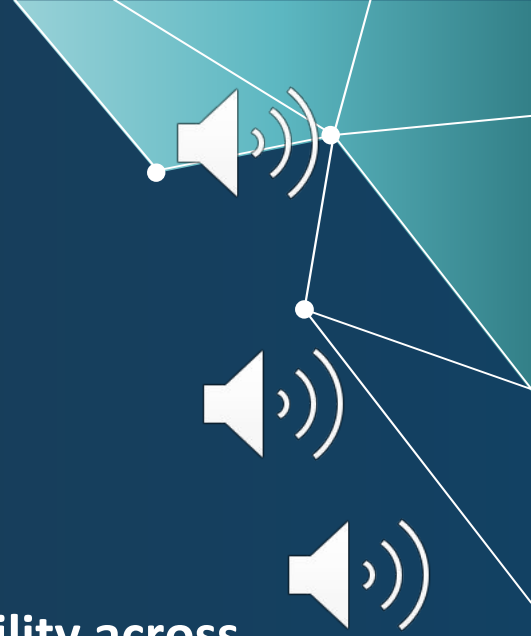
- understanding and monitoring Open Science
- assessing the impact of research
- performing Quantitative Science Studies & Science of Science

Open Science Graphs

Many initiatives spawned so to serve specific user needs and applications

Scopus®

Open Science Graphs



Nowadays OSGs suffer from

- High fragmentation
- Isolation, information silo
- Duplication of effort, low synergy

To target such problems, we advocate for **interoperability across OSGs,** whose main challenges can be identified as:

- Need to define a classification for OSGs that supports assessing their value, compare their features, and identify differences.
- Need to define an agreed-upon framework enabling a seamless exchange of information across OSGs.

A classification framework for OSGs

We started from analysing **5 representatives OSGs**:

- **FREYA PID graph** (https://www.project-freya.eu)
- **OpenAIRE research graph** (https://zenodo.org/communities/openaire-research-graph)
- **Open Research Knowledge Graph** (https://www.orkg.org)
- **Research Graph** (https://researchgraph.org)
- **Scholexplorer** (http://scholexplorer.openaire.eu)

A classification framework for OSGs

We drawn a **classification framework** across seven dimensions

1. **Research entities modelled**
2. Applications served
3. Data sources integrated
4. Added value
5. Data export and provisioning
6. FAIRness
7. Openness

A classification framework for OSGs

We drawn a **classification framework** across seven dimensions

1. Research entities modelled
2. **Applications served**
3. Data sources integrated
4. Added value
5. Data export and provisioning
6. FAIRness
7. Openness

A classification framework for OSGs

We drawn a **classification framework** across seven dimensions

1. Research entities modelled
2. Applications served
3. **Data sources integrated**
4. Added value
5. Data export and provisioning
6. FAIRness
7. Openness

A classification framework for OSGs

We drawn a **classification framework** across seven dimensions

1. Research entities modelled
2. Applications served
3. Data sources integrated
4. **Added value**
5. Data export and provisioning
6. FAIRness
7. Openness

A classification framework for OSGs

We drawn a **classification framework** across seven dimensions

1. Research entities modelled
2. Applications served
3. Data sources integrated
4. Added value
5. **Data export and provisioning**
6. FAIRness
7. Openness

A classification framework for OSGs

We drawn a **classification framework** across seven dimensions

1. Research entities modelled
2. Applications served
3. Data sources integrated
4. Added value
5. Data export and provisioning
6. **FAIRness**
7. Openness

A classification framework for OSGs

We drawn a **classification framework** across seven dimensions

1. Research entities modelled
2. Applications served
3. Data sources integrated
4. Added value
5. Data export and provisioning
6. FAIRness
7. **Openness**

	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

A framework for OSGs interoperability

Major drivers for OSGs interoperability:

- OSGs **inherent diversity and plurality**
  - No one-size-fits-all
  - OSGs serve different applications and have different scopes
  - Overlapping and/or complementary information
- OSGs are often initiatives with **no clear sustainability plan**
  - No loss of information capital
  - Redundancy, federation and sharing is key
- OSGs should be the **backbone of modern Open Science** and scholarly communication
  - Disruptive impact on traditional scholarly communication led by publishers

A framework for OSGs interoperability

Capitalise on synergies and non-negligible efforts for content acquisition, integration, enrichment performed locally at OSGs.

This can be achieved by agreeing upon a **lingua franca** so to support information flow on two level of abstraction

- **information model** (maximise information exchange and flexibility)
- **technological** (standards, exchange formats, primitives, APIs, etc.)

Scholix (http://www.scholix.org) is a successful example in this sense, despite targeting a much simpler scenario (i.e. literature-data linking).

EOSC as an optimal channel for OSGs interoperability

We envisage the **European Open Science Cloud (EOSC,** https://www.eosc-portal.eu**)** 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.

- system of system architecture
- local autonomy and diversity are considered added value and thus are fostered

Conclusions

- Outlined the State of the Art in **Open Science Graphs**
- Proposed a first **classification framework for OSGs** based on seven different dimensions
- Advocated for the establishment of an agreed upon **Interoperability Framework** for OSGs

- We invite to take part of the discussion and follow the **RDA IG on Open Science Graphs for FAIR Data**, https://www.rd-alliance.org/groups/global-open-research-commons-ig

Consiglio Nazionale delle Ricerche

ISTITUTO DI SCIENZA E TECNOLOGIE DELL'INFORMAZIONE "A. FAEDO"

Thank you!

Any question?

Andrea Mannocci
ISTI-CNR
andrea.mannocci@isti.cnr.it