Open Science Graphs Must Interoperate!

Amir Aryani, Martin Fenner, Paolo Manghi, Andrea Mannocci, Markus Stocker
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
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 draw 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 draw 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
<table>
<thead>
<tr>
<th>Research Entities</th>
<th>Applications</th>
<th>Data Sources</th>
<th>Added value</th>
<th>Data Export</th>
<th>FAIRness</th>
<th>Openness</th>
</tr>
</thead>
<tbody>
<tr>
<td>PID graph</td>
<td>Discovery, Research Impact, Open Science Monitor, Brokering, Reporting to funders, Statistics</td>
<td>PID providers</td>
<td>Standard GraphQL query interface with client libraries available in many languages. Strong support for Jupyter notebooks</td>
<td>API: GraphQL</td>
<td>Findable: Graphs are searchable by PIDs and metadata with GraphQL API</td>
<td>Redistributed free of charge</td>
</tr>
<tr>
<td>OpenAIRE Research Graph</td>
<td>Discovery, Research Impact assessment, Open Science Monitoring, Brokering, Reporting to funders, Statistics</td>
<td>Any data source trusted by scientists: repositories, archives, registries, databases, publishers</td>
<td>Enrichment of metadata and relationships by full-text mining, User-feedback, Inference by context propagation, deduplication, Provenance tracking</td>
<td>Format: OpenAIRE XML format APIs: LOD, OAI-PMH, Dumps, REST Search APIs</td>
<td>Findable: searchable on Zenodo, accessible by DOI</td>
<td>Redistributed free of charge under CC-BY licence</td>
</tr>
<tr>
<td>ORKG</td>
<td>At the granularity of items of scholarly literature content, discovery, comparison, recommendation, visualization, reuse</td>
<td>Literature, research data repositories, terminologies, LIMS/ELIN</td>
<td>Multimodal infrastructure for the acquisition, curation and publishing of machine-actionable scholarly knowledge.</td>
<td>Format: JSON and RDF serializations APIs: REST API, SPARQL</td>
<td>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. No4j 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. No4j dumps or RDF exports can be made available for reuse.</td>
<td>Data and Software are released under CC-BY-SA and MIT licences, respectively.</td>
</tr>
<tr>
<td>Research Graph</td>
<td>Supporting repositories and research infrastructures</td>
<td>PID providers, data repositories, publishers, funders, discovery services and aggregators</td>
<td>Identity resolution, metadata enhancement, topic modeling, clustering, text mining and GIS mapping</td>
<td>Format: XML and JSON APIs: Cloud Hosted Services, REST API and GraphQL</td>
<td>Findable: Metadata available via researchgraph.org</td>
<td>Controlled Access</td>
</tr>
<tr>
<td>Schoexplorer</td>
<td>Discovery, Statistics, PID Resolution</td>
<td>DataCite, CrossRef Event Data, EMBL-EBI (EMPC), OpenAIRE, Scholix-compliant data sources</td>
<td>Deduplication by PID, provenance tracking</td>
<td>Format: Scholix [?] APIs: Dumps on Zenodo, REST Search APIs</td>
<td>Findable: searchable on Zenodo, accessible by DOI</td>
<td>Redistributed free of charge under CC-BY licence</td>
</tr>
</tbody>
</table>
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](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](https://www.rd-alliance.org/groups/global-open-research-commons-ig)
Thank you!

Any question?

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