Final Report
GESIS LOD Research Graph

27.12.2017

Benjamin Zapilko (benjamin.zapilko@gesis.org)
Amir Aryani (amir.aryani@ands.org.au)
Nadine Dulisch (nadine.dulisch@gesis.org)

Executive Summary

The aim of the GESIS LOD Research Graph project is creating a research graph that makes the connection between high-value collections (research datasets) and other scholarly works such as publications and grants discoverable. The focus is on a graph that connects European and Australian data-infrastructures. In this project, we aimed to explore the application of the outcomes of two RDA working groups, the Data Description Registry Interoperability (DDRI) WG and the RDA/WDS Publishing Data Services WG. We have adopted the Research Data Switchboard that is the outcome of the DDRI Working Group. This report summarizes the project results.

Objectives

The project GESIS LOD Research Graph aims to build a trusted research graph that makes the connection between high-value collections (research datasets) and other scholarly works such as publications and grants discoverable. The focus is on building such a graph across European and Australian data-infrastructures and on demonstrating the collaboration network in the data-driven research projects.

As part of this project, GESIS has implemented the Research Data Switchboard software[1] which was developed by the DDRI WG[2]. This software allows for aggregating, connecting and publishing the research information from international sources such as DataCite[3], SpringerNature SciGraph[4], and National Institutes of Health (NIH)[5]. Also, GESIS worked on adding a new component to their information system that enables connecting GESIS graph to the Scholix data collected by the RDA/WDS Publishing Data Services WG[6]. This is in collaboration with the National Computational Infrastructure (NCI)[7] and the Research Graph team[8], both in Australia. The GESIS Research Graph has been published online. Users are able to search and explore the graph with an graphical interface. Furthermore, GESIS is making this connected network of scholarly communications available to other data infrastructures using standardized web formats like Linked Open Data (LOD).

1 http://www.rd-switchboard.org/
2 https://www.rd-alliance.org/groups/data-description-registry-interoperability.html
3 https://www.datacite.org/
4 https://www.springernature.com/cn/researchers/scigraph
5 https://www.nih.gov/
6 https://www.rd-alliance.org/groups/rdawds-publishing-data-workflows-wg.html
7 http://nci.org.au/
8 http://researchgraph.org/

The outcomes of this project allow researchers of scientific domains to explore research information of their domain in order to identify connections between grants, research data and publications. Additionally, researchers/employees of libraries, archives and infrastructure organizations will benefit from the technical implementations of the Switchboard software and the information available by Research Graph.

Initial State

Driven by the rapid development of data storage technology and the increasing demand for open science, the number of research repositories is growing fast, and researchers have access to a new range of research information systems and data infrastructures. The problem is that these infrastructures are often operating in silos; that is, there is no easy way to make connections between their research results, especially unpublished datasets, and external research work. Even their publications will be isolated items if there are no cross-references in the literature.

Initiating a project that builds on the interoperability between global infrastructures requires working closely with international partners. One of the key value propositions of RDA is providing a platform that brings potential collaborators together. In this case, RDA enabled GESIS to join NCI and Research Graph Australia in the endeavour to build a trusted graph of scholarly works. This collaboration would never have happened without the RDA communications platform. Furthermore, RDA brings technologists, data specialists, policy makers, funders, and researchers together, and this fosters open dialogues about issues that largely affect the research community, such as access, discovery, and reuse of data. This platform has enabled GESIS to highlights the lack of connectivity between research data, publications, and grants. Also, RDA enables GESIS to get feedback from a large community and to call for further collaborations to extend the scope and outcomes of this project.

Project Outcomes

In this project, GESIS aims to address this issue of unconnected research repositories and infrastructures by linking German national research data records to grants, publications and researcher profiles across international repositories. Furthermore, this project enables a GESIS repository to be interoperable with the global network of Linked Open Data and controlled vocabularies.

Adopted recommendations from RDA

Data Description Registry Interoperability (DDRI) Working Group:

- Implementing the Research Data Switchboard source code and the research data interoperability recommendation by the DDRI Working Group

RDA/WDS Publishing Data Services Working group:

- Creating a connection between the Data Literature Linking (DLI) Services and GESIS infrastructure.
- Linking GESIS records to Scholix[9] data
- Collaborate with Research Graph team to contribute the GESIS data-literature links to the Scholix hubs.

Task 1: Implementation of Research Data Switchboard at GESIS

The Research Data Switchboard has been developed in the DDRI WG and has been implemented at GESIS. This component allows a new capability for GESIS to create Research Graph clusters that show connections between research datasets,

[9] http://www.scholix.org/

publications, grants and researcher profiles. The initial cluster which was set up includes GESIS data with information on publications, research data and grants for the social science domain. An inclusion and connection to data of other research domains in order to identify the cross discipline collaboration opportunities has been done in Task 3.

Task 2: Connecting RD-Switchboard to Research Graph in Australia

In order to create a global network of Research Graph clusters, the GESIS cluster has been augmented with the Research Graph data of other clusters. In the first step, the GESIS cluster has been linked to the Research Graph registry. Based on that, the GESIS cluster has been augmented with data from other data repositories such as ORCID[10], Research Data Australia, and NCI. This component is the outcome of the direct collaboration with NCI and the Research Graph team in Australia. Figure 1 illustrates the connection between the Research Graph clusters of ANDS and GESIS with ORCID.

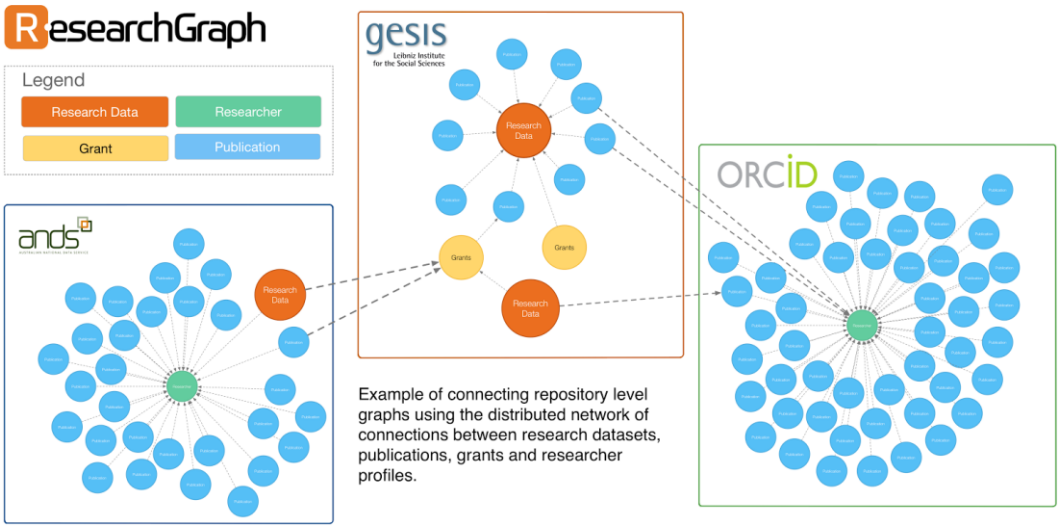

Figure 1. Example of connections between Research Graph clusters.

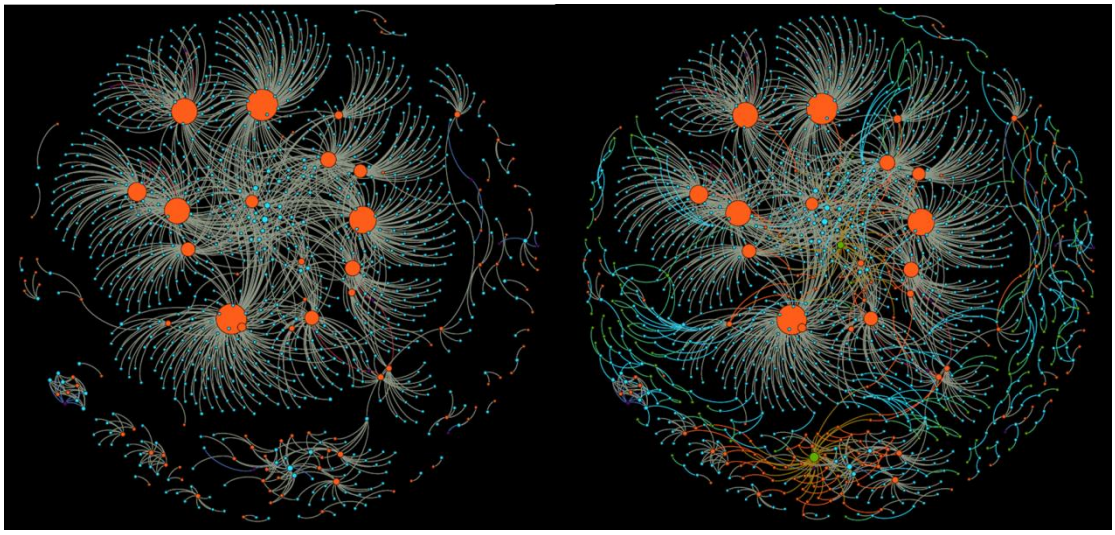

Figure 2. GESIS graph before and after the augmentation with ORCID data.

Figure 2 shows the visualization of the GESIS graph cluster before and after its augmentation with ORCID data. The benefit of the augmentation process can be seen that originally not connected resources can be connected via the data from ORCID in

[10] https://orcid.org/

most cases. Persistent identifiers like those from ORCID as well as DOIs in the data repositories play a major role in connecting the different data repositories with each other.

Task 3: Linking GESIS datasets and research information from data sources inside the EU and provision of new connections in RD-Switchboard implementation at GESIS

While it was initially planned to include research information, in particular data about grants, from organizations like Deutsche Forschungsgemeinschaft (DFG) and EU, We had to change our plan give two major obstacles: (a) we could not access the grant data from these organizations during our limited project duration (in case of DFG) and (b) we did not hold required metadata information, e.g. like persistent identifiers, in order to build basic connections to other research information repositories (in case of EU). Because of this, we considered other data sources as part of our GESIS graph cluster: SpringerNature SciGraph and National Institutes of Health (NIH). We have selected these sources because they provide a wide range of interdisciplinary research information and their information is highly connected.

Also, GESIS has been building a new component for finding connections between the GESIS graph and the Scholix records. To achieve this goal, GESIS has been working on a mapping between the Scholix data model - developed by of the RDA/WDS Publishing Data Services Working group - and the Research Graph data model.

Task 4: Add additional graphs (organizations, controlled vocabularies) to RD-Switchboard

Information on organizations and controlled vocabularies including e.g. research topics are an important value addition to research information. In this task, we have developed a novel approach for connecting the graph elements to controlled vocabularies using dynamic linking. This allows for an exploration of the graph based on e.g. research topics. In addition, we have explored the connections between the existing Research Graph clusters and organization identifiers including but not limited to the GRID database[11] by Digital Science.

The main challenge of connecting the current graph to organization information is the excessive complexity of the resulting graph. Controlled vocabularies with organizations tend to attract a large number of relations in the graph and hence create super nodes. Such super nodes make the graph difficult to process and visualize, furthermore, they create a challenge for clustering the graph and find related items using multiple degrees of separation. We developed a concept to address these issues by adding axillary nodes in the case a Research Graph object is connected to more than one organization or research topics. Additionally, we apply Dynamic Linking which enables to introduce a large number of connections from Research Graph objects (grants, publications, datasets and grants) to other nodes at runtime. These relations can be added or removed using Cypher commands and in bulk. The main advantage is that these connections can be only presented in the graph when needed as part of an analysis such as exporting GraphML files for Gephi.

By the time of writing this report, we are writing an article about this work which will be submitted to the Joint Conference on Digital Libraries 2018[12] (submission deadline 15th January 2018).

[11] https://www.grid.ac/

[12] https://2018.jcdl.org/

Publication and availability of results

The generated GESIS Research Graph cluster is available online at http://sora.gesis.org/graph. It is published on a website where users can interactively search and explore the graph and its connections. Figure 3 shows a screenshot of the website.

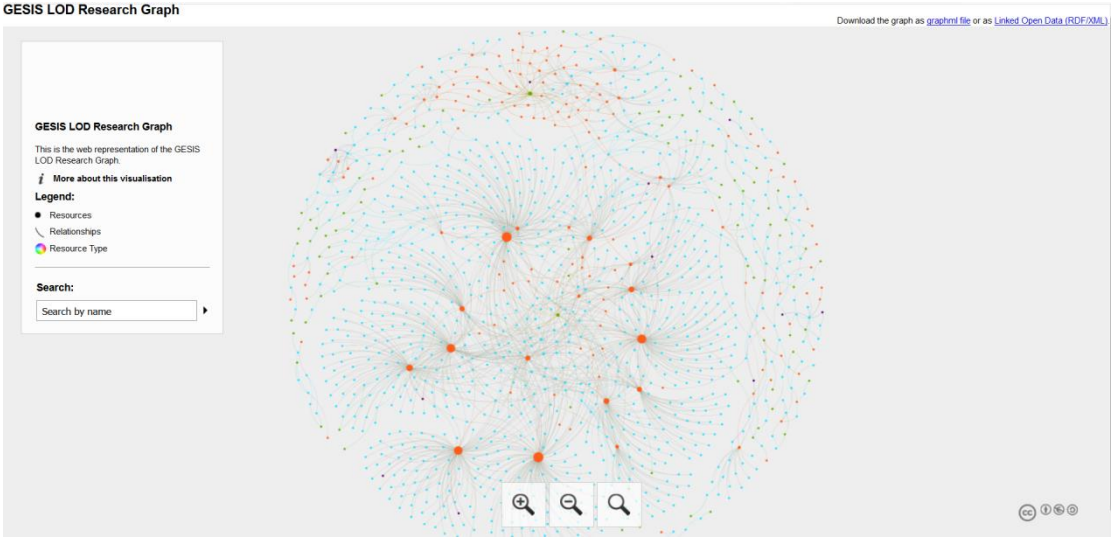

Figure 3. Screenshot of the GESIS LOD Research Graph website.

The website is generated by the Sigma.js export for Gephi[13] and creates an interactive network graph visualization. Additionally, we provide the GESIS Research Graph cluster in standardized web formats like Linked Open Data on that website in order that it can be reused by other data infrastructures. Currently, the data from NIH and SciGraph is preprocessed for a publication on the website.

The developed source code of this project has been published in the Research Graph GitHub repository at https://github.com/researchgraph in order to provide the code for third parties in general and for all collaborators in the context of Research Graph in particular. The source code consists primarily of methods for data mappings, data conversions and data ingest processes.

Dissemination Activities / Publications

The following dissemination activities and publications are planned:

#	Event	Date	Activity
1	RDA Webinar	February 2018	Webinar about the project results
2	RDA Plenary Meeting	March 21- 23 2018	Presentation of the outcomes of this collaboration project
3	JCDL 2018	June 3 – 6 2018	Article about adding additional graphs (see Task 4)

[13] https://gephi.org/

Summary & Conclusions

In this project, a research graph has been created that holds the connection between datasets, publications and grant information aggregated by National Computational Infrastructure (NCI) in Australia and GESIS from European and Australian research institutions. The GESIS Research Graph cluster includes currently data from GESIS, NIH and SpringerNature SciGraph. The graph has been integrated in the ResearchGraph.org website. Additionally, the graph is available online at http://sora.gesis.org/graph where the graph can be explored, searched and is provided as GraphML files and as Linked Open Data.

The generated graph allows researchers of scientific domains to explore research information of their domain in order to identify connections between grants, research data and publications.

This collaboration project marks a starting point for GESIS for further collaborations with Research Graph. GESIS will continue to work and augment the GESIS LOD Research Graph cluster and will participate in further collaboration activities. Therefore, we will apply for additional funding, e.g. at Deutsche Forschungsgemeinschaft (DFG).