

Federated Querying of Scholarly Communication Infrastructures

Muhammad Haris^{1,*}, Sören Auer^{1,2} and Markus Stocker^{1,2}

¹L3S Research Center, Leibniz University Hannover, 30167, Hannover Germany

²TIB—Leibniz Information Centre for Science and Technology, Germany

Abstract

Exponentially increasing inter-related scholarly knowledge is being published on multiple scholarly communication infrastructures. Retrieving data from a single scholarly communication infrastructure is not sufficient to meet users complex requirements. Moreover, the manual linking of scholarly knowledge to produce inter-related outputs is a cumbersome task. Required are flexible and user-friendly mechanisms that retrieve inter-related data from distributed scholarly infrastructures. In the proposal presented here, we leverage a federated interface to access data from multiple scholarly communication infrastructures to answer complex user queries. Specifically, we use ORKG (Open Research Knowledge Graph), ORKG Ask, DataCite, OpenAIRE Graph and Semantic Scholar endpoints to access data from these infrastructures in a federated manner. We present the work for the information needs of diverse stakeholders to demonstrate the practicability of the federation, the straightforward implementation and the added value. The code of our service is publicly available on Gitlab^{1,2}.

Keywords

Federated Query, Machine Actionability, Open Research Knowledge Graph, (Meta)data-based Search

1. Introduction

Scholarly communication infrastructures provide a plethora of unstructured text documents (scholarly articles) [1], heterogeneous structured metadata, datasets in a variety of formats [2], among other artefacts. Such abundance of data creates the need to fetch information from multiple infrastructures. Retrieving information manually from multiple scholarly communication infrastructures is a tedious task [3, 4]. This problem creates the need for integrated retrieval of data from multiple scholarly infrastructures.

Towards this goal, we developed a GraphQL-based federated query service [5] that allows unified access to data of Open Research Knowledge Graph (ORKG) [6], DataCite¹ and GeoNames². ORKG supports the representation of scholarly knowledge in a machine actionable,

¹<https://gitlab.com/TIBHannover/orkg/orkg-graphql>


²<https://gitlab.com/TIBHannover/orkg/graphql-federation>

Sci-K'24: 4th International Workshop on Scientific Knowledge: Representation, Discovery, and Assessment. Companion conference of The 23rd International Semantic Web Conference (ISWC), Baltimore, MD

*Corresponding author.

✉ haris@l3s.de (M. Haris); auer@tib.eu (S. Auer); markus.stocker@tib.eu (M. Stocker)

ORCID 0000-0002-0877-7063 (M. Haris); 0000-0002-0698-2864 (S. Auer); 0000-0001-5492-3212 (M. Stocker)

 © 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

¹<https://datacite.org>

²<https://www.geonames.org/>

structured and semantic manner and enables retrieval of related scholarly content or derivatives thereof (e.g., comparisons). We propose that the (machine-readable) scholarly knowledge published in ORKG and other scholarly communication infrastructures can be leveraged to answer complex user queries (i.e., (meta)data-driven analysis). However, this federated system currently has some limitations: the scope of querying within the federated service is limited; it does not support a wide range of queries, which restricts the ability to retrieve diverse data types or conduct complex data searches across the integrated infrastructures. Furthermore, the current capabilities for filtering content within the federated query service is limited. While scholarly artefacts can be retrieved simultaneously from multiple scholarly infrastructures, the options for filtering these results based on specific criteria are not sufficiently comprehensive or user-friendly. Additionally, after retrieving the data, post-processing steps are required to refine the results according to the user's specific requirements.

To address these limitations, we extend the GraphQL-based federated system and integrate other scholarly communication infrastructures, namely, ORKG Ask³, OpenAIRE Graph [7, 8], and Semantic Scholar [9]. The extended federated system facilitates federated query execution and supports the integrated retrieval of scholarly information and knowledge, through the integration with ORKG Ask also generative content. Its primary objective is to enable the interconnection of scholarly knowledge and contextual information, and to allow filtering at the (meta)data level. We suggest that the machine-actionable scholarly knowledge published in ORKG could be utilized together with the scholarly content from other infrastructures to address complex user queries concerning bibliographic metadata, article content, or both.

Our contributions are as follows:

1. We extend the GraphQL-based federated system to include the ORKG Ask, OpenAIRE Graph, and Semantic Scholar infrastructures, and to retrieve the fragmented scholarly content via a single endpoint to answer complex user queries.
2. We define a clear and straightforward mechanism for filtering results in a user-friendly manner. This enhancement includes filtering capabilities for data obtained from infrastructures beyond ORKG.

In order to demonstrate the practicability of federated scholarly infrastructures, the straightforward implementation and the added value, we propose the following scenarios reflecting diverse information needs of different stakeholder groups:

- Researchers: Filtering papers based on the specified criteria e.g., find all COVID-19 papers reported $R0$ estimate less than a threshold.
- Funders: Retrieve the most significant papers (both in terms of statistical impact and academic impact) published under a specific grant. This entails identifying research outputs that have advanced the field as well as demonstrated solid results for a particular problem.
- Bibliometricians: (1) Create a network of co-authors for a given researcher, focusing on those collaborations that have produced the most significant papers. This involves analyzing the content of researchers articles and identifying those that have reported

³<https://ask.orkg.org/>

statistically significant results. (2) Uncover the research focusing on novel problems that have received a high number of citations. This involves identifying such papers and assessing their citation counts.

In the context of above discussion, we thus focus on addressing the following research question:

How can we reuse interoperable scholarly knowledge to support complex (meta)data-driven analysis?

2. Related Work

Several services have been developed to process federated queries across different databases or scholarly communication infrastructures. BioThings Explorer [10] is an application designed to query a federated knowledge graph, which is formed by merging data from various biomedical web services. This tool uses detailed semantic annotations to understand the inputs and outputs of each source, enabling it to automatically link multiple web service calls to execute complex graph queries. Since there is no centralized knowledge graph to maintain, the proposed data explorer is a streamlined application that fetches data dynamically.

To access biological data, ROBOKOP (Reasoning Over Biomedical Objects linked in Knowledge-Oriented Pathways) [11] was proposed. ROBOKOP is a KG that supports the open biomedical question-answering application. Additionally, the ROBOKOP Knowledge Graph Builder (KGB) was presented, which constructs the KG and provides a rich framework to execute graph query on federated data sources. A biomedical query system named Translator Query Language (TranQL) [12] was proposed to search and analyze semantically connected knowledge graphs. Utilizing the Biolink data model, TranQL structures queries into a graph of Biolink elements. These queries are executed on federated knowledge sources, and the retrieved results are integrated into one knowledge graph. Similarly, BioCarian [13]— an user-friendly interface was proposed to enable querying heterogeneous biological databases in an exploratory manner. The interface is enriched with facets that enable better query construction, thus making it easier for users to filter data. BioCarian also provides a SPARQL endpoint where users can directly execute the federated queries to explore the disparate databases.

Tong et al. [14] proposed Hu-Fu system for effective and secure spatial query processing on federated data. Hu-Fu supports querying native SQL as well as various spatial databases, including PostGIS, Simba, GeoMesa, and SpatialHadoop. Comprehensive tests indicate that Hu-Fu outperforms state-of-the-art systems in execution speed and communication efficiency while ensuring robust security. Ontario [15] is a specialized federated query processing method designed for vast and diverse data systems. It offers efficient query processing across a collection of data sources within a data lake. Ontario utilizes RDF Molecule Templates which are abstract overviews of entity properties in a unified schema and their implementation in a data lake.

3. Federated Access to Scholarly Infrastructure

In this section, we present the approach to extend our federated query service. We integrate Open Research Knowledge Graph (ORKG), ORKG Ask, OpenAIRE Graph, DataCite and Semantic

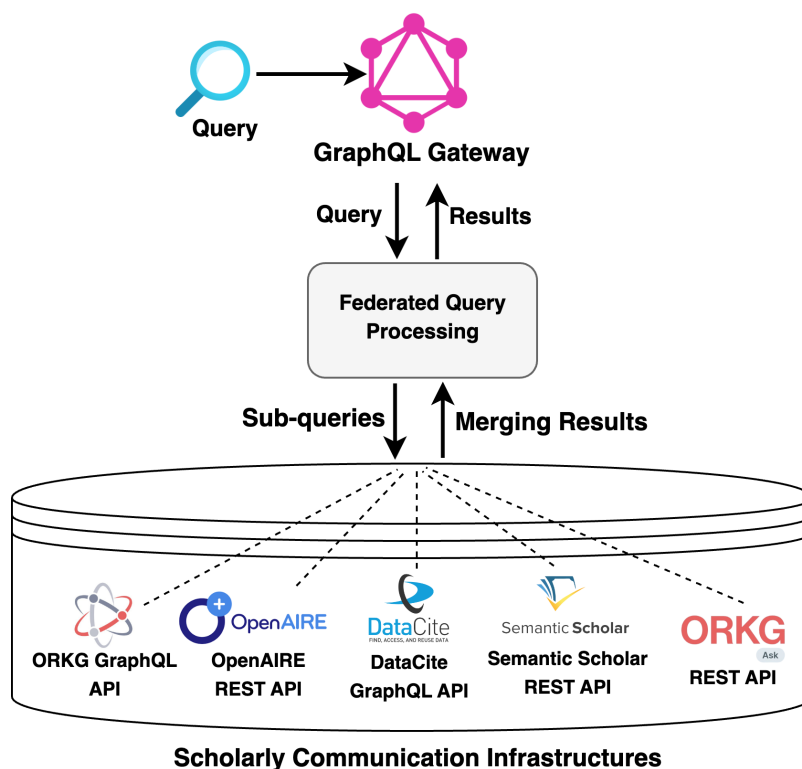


Figure 1: Overview of the virtually integrated APIs of multiple scholarly infrastructures (ORKG, ORKG Ask, OpenAIRE Graph, DataCite, and Semantic Scholar). The figure illustrates the process of executing a federated query: The query is divided into sub-queries, which are executed on the respective infrastructures. The results are then retrieved, aggregated, and presented to the client.

Scholar infrastructures in a federated GraphQL service to retrieve and integrate the fragmented scholarly content to enable complex data-driven analysis.

DataCite offers a GraphQL endpoint for the PID Graph, connecting persistently identified resources from DataCite, ORCID, ROR, and other sources, and providing standardized metadata for these resources. Similarly, we implement a GraphQL endpoint to enable access to ORKG content. Additionally, we also integrate the ORKG Ask, OpenAIRE Graph, and Semantic Scholar REST APIs to fetch diverse information regarding different scholarly artefacts. ORKG Ask is an advanced search system that helps to find and extract valuable information from a vast corpus of research articles. It revolutionizes the way researchers navigate scholarly literature by leveraging cutting-edge Large Language Model (LLM) technologies to deliver precise and relevant answers according to research queries. By federating also ORKG Ask, the system obtains the ability to leverage generative content, which provides a possibility to address missing data. When certain pieces of information are not available in the ORKG papers, the system can use generative models to predict the missing details, thereby enhancing the comprehensiveness and utility of the search results. OpenAIRE Graph provides metadata about various scholarly artefacts (articles, datasets, software) and other objects, including projects, organizations, and researchers. Semantic Scholar is an academic search engine developed by the Allen Institute

for AI. It uses artificial intelligence and machine learning to help researchers find relevant information efficiently.

Our federated query service facilitates cross-walking between metadata about artefacts (such as articles and datasets) with their context (such as people and organizations) and the content of the articles, thereby supporting complex (meta)data-driven analysis. Figure 1 illustrates the enhanced federated architecture. The proposed federated graph plays a crucial role in enabling complex (meta)data-driven analysis. In addition to metadata analysis, infrastructures can utilize article's content from the ORKG to perform new kind of analysis on data. For example, a researcher may discover all research outputs (papers) published under a particular grant that have impact (high number of citations) as well as have significant results (in a statistical sense of $p < .001$).

4. (Meta)data-driven Analysis

4.1. Researcher

A researcher discovered an ORKG COVID-19 comparison for the virus' basic reproduction (R_0) estimates. This comparison can be accessed through the federated GraphQL endpoint using its DOI: <https://doi.org/10.48366/r44930>. The researcher is interested in filtering studies that report R_0 estimate less than a specified threshold and utilize the method `generalized growth model`. While the comparison includes details about methods used to conduct experiments, not all papers include them. We thus leverage ORKG Ask to complete this missing data in a generative manner. The federated endpoint is designed to facilitate such complex queries by executing the relevant components at designated endpoints, thereby enabling the essential metadata-driven analysis for the research. Listing 1 shows the query executed on ORKG and ORKG Ask in a federated manner. ORKG provides the details of a comparison whereas semantic details (methods, and results) of compared papers are retrievable from ORKG Ask. We can also include additional filters to refine the search results, for example search only those studies that have reported results for the European region. For such a scenario, the GeoNames endpoint within the federated service will be queried to retrieve all countries included in the EU region. The results can be filtered by adding the `location` property to the `where` clause: `property: "location", value: "EU"`.

4.2. Funders

A funder seeks to collect all articles that have employed Named Entity Recognition across various domains, particularly those reporting an accuracy exceeding, for example, 65%. This process will allow the funder to understand the domains where NER has been applied, the methodologies used, and their corresponding accuracies. Additionally, the funder aims to determine if a retrieved artefact is peer-reviewed or not. Such a complex scenario can be addressed by querying OpenAIRE Graph, DataCite, and ORKG in a federated manner. The federated query outlined in Listing 2 is first executed on OpenAIRE Graph to gather all papers acknowledging a specific grant (859136⁴), then on DataCite to retrieve peer review information,

⁴<https://cordis.europa.eu/project/id/819536>

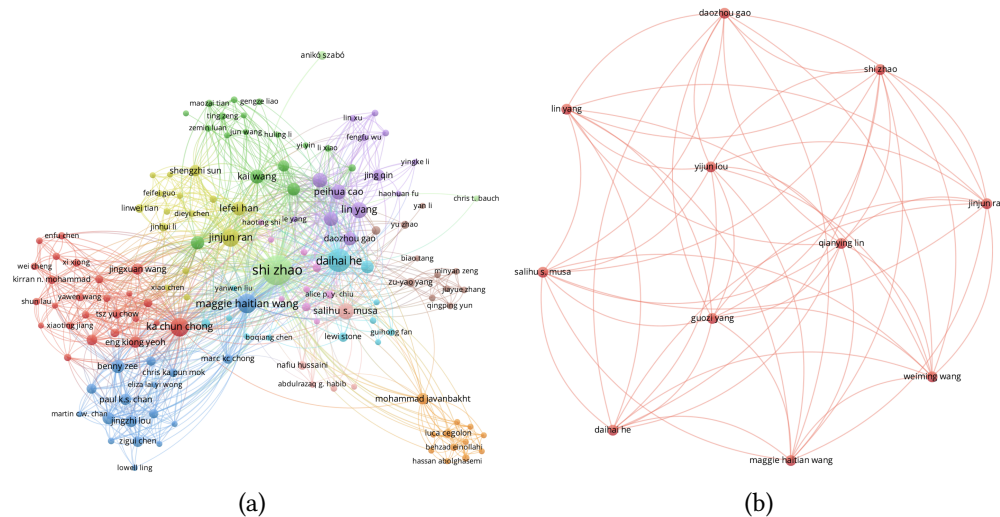


Figure 2: Co-author network of a researcher Shi Zhao; A) Showing all authors with whom he has published articles, retrieved by querying the DataCite PID Graph using the ORCID; B) Highlighting all COVID-19 research articles reporting a basic reproduction (R_0) estimate less than 4.

and finally on ORKG to obtain the corresponding machine-actionable contribution descriptions. These papers are then filtered based on the research problem addressed and the reported accuracy. Additionally, the funder can generate a comparison of all research contributions and analyze the significance of the funded work, considering both its impact and statistical significance.

4.3. Bibliometricians

4.3.1. Co-author Network Analysis

A bibliometrician wants to find all co-authors of a particular researcher who is working on the estimation of COVID-19 reproduction and published highly significant papers i.e., reported average 95% Confidence Interval (CI) less than some threshold. Thus, she performs a network analysis of the researcher's collaborations and also determines highly significant studies in the epidemiology research. This task relies on the DOI of articles that reported 95% CI values, retrievable from ORKG thanks to machine actionable content representation. The set of articles of a researcher can be retrieved using ORCID ID by leveraging the DataCite services. Finally, the most important articles meeting the criteria (COVID-19 articles reported CI 95% < 4) can be fetched from ORKG. Listing 3 shows the query executed on DataCite and ORKG in a federated manner. Figure 2 shows the co-authors network of a researcher with whom he has published COVID-19 articles.

4.3.2. Citation Analysis

A bibliometrician is exploring studies that address fairness in machine learning models, with a specific focus on identifying articles that have reported significant outcomes in fairness metrics

and are also highly cited. This task involves accessing articles that have reported fairness metrics, available through ORKG. With a relevant set of articles identified, the number of citations for each can be retrieved from Semantic Scholar. Listing 4 outlines the query executed on ORKG and Semantic Scholar in a federated manner. Lines 3-9 in the query are executed on the ORKG API to fetch the papers' DOIs, and line 12 is executed on Semantic Scholar to retrieve the citations of each paper.

5. Discussion

Federated search is a widely used approach to retrieve data from distributed sources. To address our research question, we presented a federated architecture that crosswalks the metadata and data from DataCite, ORKG, ORKG Ask, OpenAIRE Graph, and Semantic Scholar infrastructures. The proposed federated architecture leverages the machine-actionable scholarly knowledge published in the ORKG, the generative content capabilities by ORKG Ask, and classical (bibliographic) metadata to answer complex user queries. Our federated endpoint enables different stakeholders to pose queries to meet their complex information needs. We have shown the practicability of federated search by presenting different federated queries: (i) retrieving statistically significant research articles (ii) retrieving highly impactful research articles published under a specific grant (iii) supporting bibliometricians in creating network of co-authors who have published articles with significant results.

5.1. Future directions

We aim to develop ORKG Commons (Figure 3), a web-based platform that will enable interactive exploration of data served by the federated GraphQL service presented here. This platform will allow users to seamlessly navigate and analyze interconnected artefacts retrieved from various scholarly communication infrastructures in a federated setup. A key feature of ORKG Commons will be to allow users to write queries in natural language, which will be automatically converted into GraphQL queries to facilitate broader accessibility and ease of use. Additionally, ORKG Commons will feature dynamic facets for content filtering. These facets will be generated based on the metadata and the content of scholarly artefacts –providing users with enriched options to refine their search results. This interface will provide users with appropriate options to refine their search results and facilitate navigation through the vast and interconnected scholarly knowledge available through various scholarly communication infrastructures.

6. Conclusion

Federated search is a well-established method for retrieving data from heterogeneous and distributed sources. This paper introduces a federated architecture for a system designed to support cross-walks between scholarly metadata and data. The federated system described has the potential to help users in conducting advanced scientific inquiries. By enabling the formulation of complex information needs, this system supports the exploration and analysis of scholarly

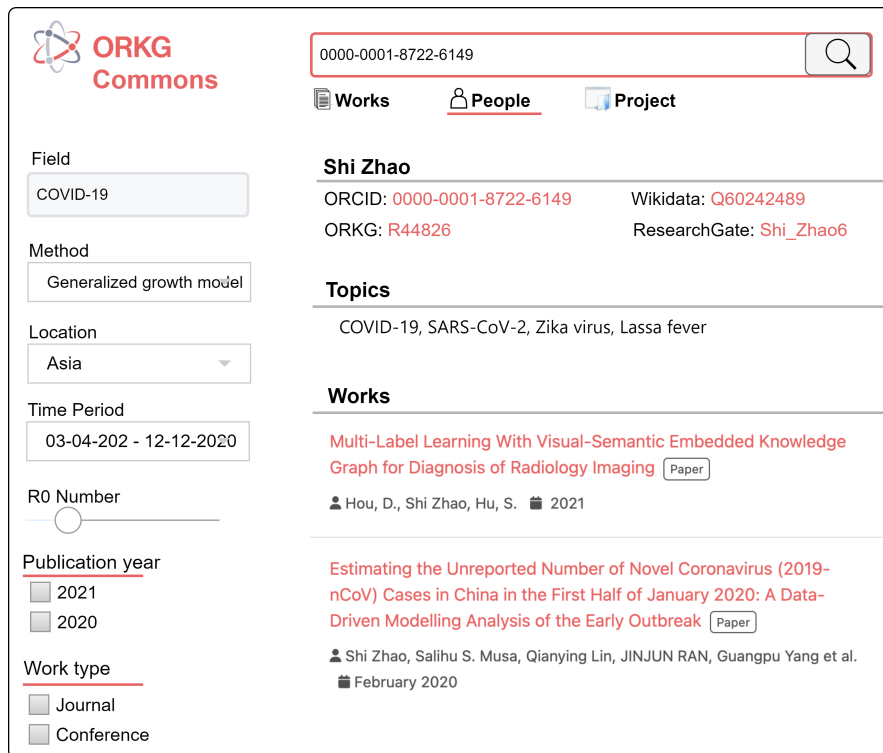


Figure 3: Mock-up of the planned ORKG Commons displaying user profile information by ORCID ID, including topics and works retrieved in a federated manner from diverse scholarly communication infrastructures and supporting dynamic faceting based on the structured content source from ORKG, enabling filtering content based on both metadata and data expressing scholarly knowledge.

knowledge published in scholarly articles together with contextual research information. This, in turn, can lead to more insightful, data-driven discoveries across various scientific domains.

Acknowledgments

This work was co-funded by the European Research Council for the project ScienceGRAPH (Grant agreement ID: 819536) and the German Research Foundation (DFG) project NFDI4DS (PN: 460234259).

References

- [1] T. Kuhn, C. Chichester, M. Krauthammer, N. Queralt-Rosinach, R. Verborgh, G. Gianakopoulos, A.-C. N. Ngomo, R. Vigiante, M. Dumontier, Decentralized provenance-aware publishing with nanopublications, *PeerJ Computer Science* 2 (2016) e78. doi:10.7717/peerj-cs.78.
- [2] J. Hendler, Data integration for heterogenous datasets, *Big data* 2 (2014) 205–215. doi:10.1089/big.2014.0068.

- [3] Y. Zhou, S. De, K. Moessner, Implementation of federated query processing on linked data, in: 2013 IEEE 24th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), 2013, pp. 3553–3557. doi:10.1109/PIMRC.2013.6666765.
- [4] A. Schwarte, P. Haase, K. Hose, R. Schenkel, M. Schmidt, Fedx: Optimization techniques for federated query processing on linked data, in: International Semantic Web Conference, 2011. doi:10.1007/978-3-642-25073-6_38.
- [5] M. Haris, K. E. Farfar, M. Stocker, S. Auer, Federating scholarly infrastructures with graphql, in: H.-R. Ke, C. S. Lee, K. Sugiyama (Eds.), Towards Open and Trustworthy Digital Societies, Springer International Publishing, Cham, 2021, pp. 308–324. doi:10.1007/978-3-030-91669-5_24.
- [6] M. Stocker, A. Oelen, M. Y. Jaradeh, M. Haris, O. A. Oghli, G. Heidari, H. Hussein, A.-L. Lorenz, S. Kabenamualu, K. E. Farfar, et al., Fair scientific information with the open research knowledge graph, FAIR Connect 1 (2023) 19–21. doi:10.3233/FC-221513.
- [7] P. Manghi, Bolikowski, N. Manola, J. Schirrwagen, T. Smith, Openaireplus: the european scholarly communication data infrastructure, D-Lib Magazine 18 (2012). doi:10.1045/september2012-manghi.
- [8] P. Manghi, N. Houssos, M. Mikulicic, B. Jörg, The data model of the openaire scientific communication e-infrastructure, in: J. M. Doderio, M. Palomo-Duarte, P. Karampiperis (Eds.), Metadata and Semantics Research, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, pp. 168–180. doi:10.1007/978-3-642-35233-1_18.
- [9] S. Fricke, Semantic scholar, Journal of the Medical Library Association: JMLA 106 (2018) 145. doi:10.5195/jmla.2018.280.
- [10] J. Callaghan, C. H. Xu, J. Xin, M. A. Cano, A. Riutta, E. Zhou, R. Juneja, Y. Yao, M. Narayan, K. Hanspers, A. Agrawal, A. R. Pico, C. Wu, A. I. Su, BioThings Explorer: a query engine for a federated knowledge graph of biomedical APIs, Bioinformatics 39 (2023). doi:10.1093/bioinformatics/btad570.
- [11] C. Bizon, S. Cox, J. Balhoff, Y. Kebede, P. Wang, K. Morton, K. Fecho, A. Tropsha, Robokop kg and kgb: integrated knowledge graphs from federated sources, Journal of chemical information and modeling 59 (2019) 4968–4973. doi:10.1021/acs.jcim.9b00683.
- [12] S. Cox, S. C. Ahalt, J. Balhoff, C. Bizon, K. Fecho, Y. Kebede, K. Morton, A. Tropsha, P. Wang, H. Xu, et al., Visualization environment for federated knowledge graphs: development of an interactive biomedical query language and web application interface, JMIR Medical Informatics 8 (2020) e17964. doi:10.2196/17964.
- [13] N. Zaki, C. Tennakoon, Biocarian: Search engine for exploratory searches in heterogeneous biological databases, BMC Bioinformatics 18 (2017) 435. doi:10.1186/s12859-017-1840-4.
- [14] Y. Tong, X. Pan, Y. Zeng, Y. Shi, C. Xue, Z. Zhou, X. Zhang, L. Chen, Y. Xu, K. Xu, et al., Hu-fu: Efficient and secure spatial queries over data federation, Proceedings of the VLDB Endowment 15 (2022) 1159. doi:10.14778/3514061.3514064.
- [15] K. M. Endris, P. D. Rohde, M.-E. Vidal, S. Auer, Ontario: Federated query processing against a semantic data lake, in: Database and Expert Systems Applications: 30th International Conference, DEXA 2019, Linz, Austria, August 26–29, 2019, Proceedings, Part I 30, Springer, 2019, pp. 379–395. doi:10.1007/978-3-030-27615-7_29.

A. Federated GraphQL Queries

Listing 1: Retrieving COVID-19 comparison from ORKG and other semantic details from ORKG Ask as well as filtering the results according to the specified condition.

```
1 {
2   comparison( doi: "10.48366/r44930"
3     where: [{property: "METHOD", value: "generalized growth model" }]) {
4     doi label
5     contributions {
6       id label
7       paper { doi label }
8       ORKG Askdetails { method }
9       data { propertyLabel label}
10  } } }
11
12 Result (shortened):
13 { "data": {
14   "comparison": {
15     "doi": "10.48366/r44930",
16     "label": "COVID-19 Reproductive Number Estimates",
17     "contributions": [{
18       "paper": {
19         "doi": "10.1101/2020.03.08.20030643",
20         "label": "Transmission potential of COVID-19 in Iran"
21       },
22       "ORKG Askdetails": {
23         "method": "The study statistically estimated the cCFR and the basic reproduction number
24           using the exponential growth rate of the incidence."
25       },
26       "data": [...]
27     }
28   }
29 }
```

Listing 2: GraphQL query executed on OpenAIRE Graph, DataCite and ORKG to obtain the details of papers published under ScienceGraph project (Grant ID: 819536).

```
1 { # OpenAIRE Graph query
2   project(
3     id: "819536", where: [
4       { property: "P32", value: "named entity recognition" }
5       # search for numeric value
6       { property: "P45075", _GTE: 0.65 }
7     ]) {
8     #DataCite Query
9     peerReview { type }
10    papers { # ORKG query
11      label doi
12      contributions {
13        id label
14        data { propertyId label }
15      }
16    }
17  }
18 Result (shortened):
19 { "data": { "project": {
20   "papers": [{
```

```

21         { "type": "PeerReview" }
22         "label": "Agriculture Named Entity Recognition-Towards FAIR,
23         Reusable Scholarly Contributions in Agriculture",
24         "contributions": [...]
25     }]
26 ]}  }}

```

Listing 3: Results obtained by querying the PID Graph and ORKG federation about COVID-19

```

1 Query:
2 { # DataCite query
3   person(id: "https://orcid.org/0000-0001-8722-6149") { id name
4     # ORKG query
5     papers( where: [
6       {
7         property: "P32" #research problem
8         value: "Determination of the COVID-19 basic reproduction number"
9       }
10      { property: "HAS_VALUE", _LT: 4 }
11    ]) { label id} # showing paper title and doi
12 } }
13
14 Result (shortened):
15 { "data": { "person": { "id": "https://orcid.org/0000-0001-8722-6149", "name": "Shi Zhao",
16   "papers": [{
17     "id": "R44910",
18     "label": "Estimating the Unreported Number of Novel Coronavirus
19     (2019-nCoV) Cases in China in the First Half of January 2020:
20     A Data-Driven Modelling Analysis of the Early Outbreak"
21   }]
22 } } }

```

Listing 4: Results obtained by querying the Semantic Scholar and ORKG federation about the study Fairness in machine learning and has reported Balanced accuracy > (75%).

```

1 Query: {
2   # ORKG query
3   papers(
4     where: [
5       { property: "P32", value: "fairness in machine learning" }
6       { property: "P140012", _GTE: 75 }
7       { MinCitationCount: 35 }
8     ]) {
9     doi label
10
11     # retrieving citations from Semantic Scholar
12     citationCount
13   }
14 }
15
16 Result (shortened): {
17   "data": {
18     "papers": [{
19       "doi": "10.1145/3357384.3357974",
20       "label": "AdaFair: Cumulative Fairness Adaptive Boosting",
21       "citationCount": 64
22     }]
23 } }

```