The RESCS Ontology: linking Open Research Data from multiple sources to support interdisciplinary investigations

Kurt Baumann [a], Andrea Bertino [a,*], Laura Rettig [b], Sebastian Sigloch [a], Daniela Subotic [c], and Ivan Subotic [c]
[a] *SWITCH*
E-mails: kurt.baumann@switch.ch, andrea.bertino@switch.ch, sebastian.sigloch@switch.ch
[b] *eXascale Infolab, University of Fribourg*
E-mail: laura@exascale.info
[c] *DaSCH, University of Basel*
E-mails: daniela.subotic@dasch.swiss, ivan.subotic@dasch.swiss

Abstract. The availability of open repositories and the application of Semantic Web techniques are paving the way towards new usage scenarios for research data. This paper describes the ontology developed within the first phase of the Connectome project. The goal of the Connectome project is to make data from different providers interoperable and thus improve its use through both generic and discipline-specific services. On the basis of the RESCS (RESearch CommonS) ontology defined through an intensive exchange with various researchers, data providers and funders, we give a detailed description of the ontology. The paper concludes with a brief outlook on possible tools and applications which could take advantage of the Connectome knowledge graph in the future.

Keywords: Interdisciplinarity, Interoperability, Ontology, Open Research Data, Repositories

1. Introduction

Many kinds of scientific investigations require bringing together knowledge from different disciplines [1], but there are still cultural and institutional obstacles that are limiting the uptake of multi- and interdisciplinary scholarly research [2]. Even more challenging are technical barriers limiting the findability and reusability of resources and specifically, multi- and interdisciplinary investigations are faced by the lack of interoperability between repositories and the heterogeneity of data formats[1]. Open standards and semantic web technologies can improve this situation by transforming dataset descriptions (metadata) and repositories into coherent and interoperable infrastructures. This requires the construction of knowledge graphs, i.e., graphs of data "intended to accumulate and convey knowledge of the real world, whose nodes represent entities of interest and whose edges represent relations between these entities." [3] Structuring data and metadata in a knowledge graph enables both the discovery of regularities and the appreciation of relationships that would otherwise remain opaque. More specifically structuring data through an ontology into a knowledge graph opens up a unifying horizon of meaning for the interlinked entities and also new interdisciplinary investigations. We assume that these require new formal descriptions, i.e., new ontologies which can meet the needs of interdisciplinary research. In general, as far as scholarly research is concerned, the vision of a web of semantically linked data cannot consist in creating a unique and definitive ontology - which would be no

*Corresponding author. E-mail: kurt.baumann@switch.ch.

[1] Also in Switzerland, the repository landscape is characterized by a high degree of fragmentation[4]

more than the absolutization of a particular viewpoint - but rather in enabling a pluralism of research perspectives in the form of a variety of ontologies and knowledge graphs, as much interrelated as possible. This is even more true for interdisciplinary research approaches, and the creation of a knowledge graph for this specific kind of investigation is one of the central goals of the Connectome project. The Connectome project aims[2] to link and semantically enrich the metadata of relevant repositories and data providers through a new ontology (see Fig. 2). The harmonization of different open data and their integration in the knowledge graph will help to overcome the fragmentation and lack of interoperability of research repositories, and to pace the way to new ways to relate and investigate different regions of knowledge. The Connectome, even before being the result of a technological vision, intends to be the implementation of a specific governance which will be developed during the Connectome project and enable service providers to better address the needs of the different scientific communities.

In the pilot phase of the Connectome project, a partnership of seven Swiss organizations - DaSCH, EPFL Blue Brain, eXascale Infolab, FORS, SAGW, SATW, and SWITCH - focused on defining the needs of the different stakeholders in terms of how to increase the retrievability and reuse of linked data for research. In this paper, we first present the preliminary work carried out with scholars from varying disciplines to clarify common problems and needs, as well as the methodology used in the process. In the rest of the paper, we present the main features of the RESCS Ontology, focusing on the methodology followed for realizing it (Section 2); the aims, foundation and serialization of the RESCS Ontology (Section 3); the prototyping of the ontology in the developed Open Linked Data pipeline (Section 4) and, finally, a short overview of the potential developments and applications of this ontology and its knowledge graph (Section 5).

2. Methodology

The Connectome project intends to link metadata of research objects from various data providers (both generic and domain-specific repositories) into a knowledge graph in order to enhance their discoverability and accessibility from many service providers. In this way, the Connectome ecosystem aims to pave the way to interdisciplinary investigations too.

There are many different approaches for building ontologies for new Knowledge Graph [5] (see also [6]) and these approaches can benefit from each other in a kind of "mutual fertilization" [7]. In order to realize and prototype a first version of the graph, a possible architecture of the Connectome ecosystem was elaborated using the following six-steps approach:

- Collect community feedback through semi-structured interviews.
- Analyze community feedback to derive Ontology Competency Questions.
- Review of best-practice ontologies.
- Define the ontology for the Connectome Knowledge Graph.
- Formalize and publish ontology for prototyping usage.
- Knowledge graph platform installation & prototype usage.

2.1. *Community Feedback & Ontology Competency Questions*

First, the Connectome partners carried out, together with Swiss researchers from various disciplines, a preliminary analysis of researchers' requirements regarding the potential benefits and problem-solving capabilities of open linked research data. To this aim, 25 semi-structured interviews were conducted - 21 interviews with Swiss researchers from 11 institutions, covering various research disciplines, 3 interviews with service providers and 1 with a representative of Swissuniversities, an umbrella organization of the universities in Switzerland.

The transcripts of these interviews were then analyzed by clustering emerging common needs and key issues. Based on this, the Connectome partners defined and evaluated the user-centric requirements for the Connectome Knowledge Graph in the form of 68 Ontology Competency Questions (OCQs). Such Ontology Competency Questions are "natural language questions outlining and constraining the scope of knowledge represented in an ontology" [8]. Here below are some samples of OCQ particularly relevant for the development of a knowledge graph in interdisciplinary perspective (see Appendix A for the complete overview of the OCQs):

- How can I discover interesting new material in other domains?

[2] https://www.switch.ch/about/open-science/

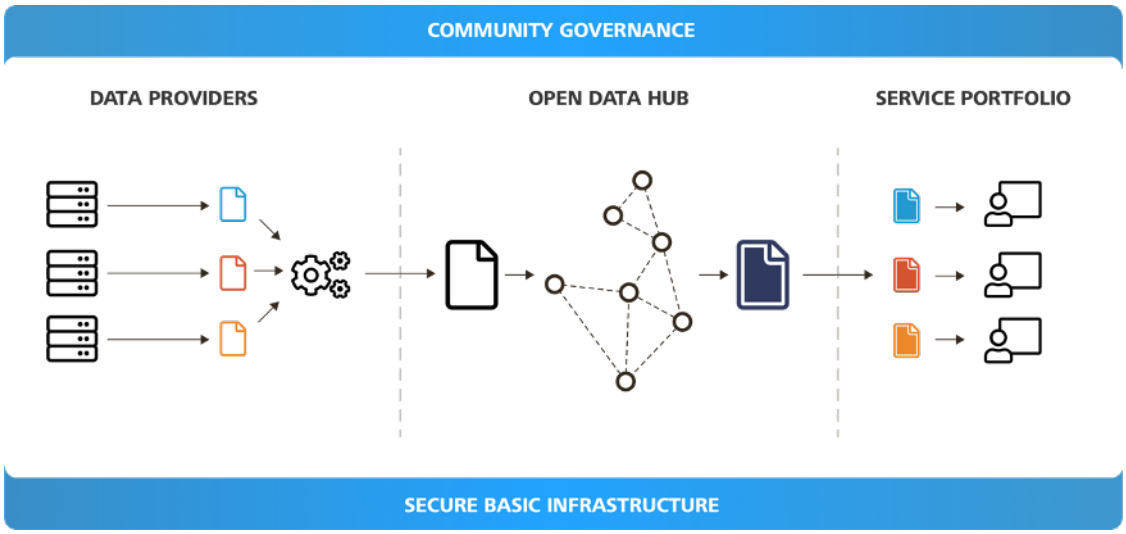

Fig. 1. Resources from different repositories are semantically enriched and linked together in the Connectome Knowledge Graph so that service providers can leverage open research data.

- What information from other not academic sources (e.g., Government, GLAM) relate to my information?
- How can I trust information?
- What is the body of knowledge, where my research is embedded in?
- What are my adjacent research disciplines?
- What kind of interdisciplinary research matches with my research domain?

The elaborated OCQs helped the Connectome partners to determine the data model of the Connectome Knowledge Graph from an end-users' perspective so that the relevant entities (nodes in a graph) and their respective relationships (edges in a graph) can contribute to realize the use cases of the stakeholders belonging to the Connectome ecosystem, (see Fig. 2 in Section 4). Additionally, the Connectome partners verified, whether or not a given OCQ, could be answered algorithmically within the elaborated Connectome Knowledge Graph and clarified which metadata should be provided to satisfy the identified user needs.

2.2. *Review of Best Practices*

Already existing ontologies can be used to connect data from different disciplines through standardized metadata and an open web-architecture. Therefore, we did not design a completely new ontology, but rather tried to extend well-established standards with specific features that were lacking according to our analysis. [6] The choice of reusing and extending a specific ontology depended on the following set of criteria:

- Structure, allowing to infer (new) knowledge fragments.
- Shared vocabulary, enabling interoperability to other ontologies.
- Established community, and governance allowing co-design new features.
- Dissemination and long-term availability.
- Coverage in relation to our Ontology Competency Questions (quality vs. quantity of metadata).

Using these criteria, the Connectome partners reviewed the data models and best practices of the following initiatives: Researchgraph.org[3], Scholix[4], DCAT[5], Dublin Core[6], OAI[7], OpenAIRE[8], Freya PID Graph[9], PROV-O[10] and Schema.org[11]. Considering the identified Ontology Competency Questions (see Section 2.1), as well as the review criteria above, the schema.org and PROV-O ontologies were chosen as a basis for our ontology.

[3] https://researchgraph.org/
[4] https://www.w3.org/TR/rdf-schema/
[5] https://www.w3.org/TR/vocab-dcat/
[6] https://dublincore.org/
[7] https://www.openarchives.org/
[8] https://www.openaire.eu/
[9] https://www.project-freya.eu/en
[10] https://www.w3.org/TR/prov-o/
[11] https://www.schema.org

2.2.1. Schema.org

Schema.org is a standardization effort founded by major search engines companies to create, maintain and foster the use of schemas on the World Wide Web. Its data model is derived from the *Resource Description Framework (short: RDF)*[12], and the standard supports various serialization formats (e.g., RDFa or JSON-LD). Many existing entities of Schema.org were reused together with new relevant entities and properties (the used entities are documented in the following Section). The prioritized entities and properties were then the starting point for a profiling and interpretation on Schema.org and the corresponding *SHACL*[13] formalization. The Shapes Constraint Language (SHACL) is a language for validating RDF graphs against a set of conditions. The Connectome partners used the features of Blue Brain Nexus, an open-source knowledge graph ecosystem designed by the EPFL Blue Brain Project[14], to validate incoming data in the RESCS Ontology using SHACL.

2.2.2. PROV-O

The PROV-O Ontology provides a set of classes, properties, and restrictions that can be used to represent and interchange provenance information generated in different systems and under different contexts. PROV-O is incorporated in the Blue Brain Nexus tool and is used to achieve provenance descriptions for entities. We extended upon these two ontologies since they did not capture all our data needs sufficiently; namely, the relationships between entities in a research data context are more general in Schema.org than we need for the Research Data Connectome. In order for the Connectome to be useful and usable, we found the declaration of relationships between for example the "Dataset" and "ScholarlyArticle" classes, two core pieces of the Connectome, to be lacking. In future work, incorporating more domain-specific data will require further additions to the ontology, such as metadata about research methodologies used to create a dataset. Defining an ontology that is extensible with the required properties allows us to do so.

[12] https://www.w3.org/TR/rdf-schema/

[13] https://www.w3.org/TR/shacl/

[14] https://actu.epfl.ch/news/blue-brain-nexus-an-open-source-tool-for-data-driv/

3. The RESCS Ontology

3.1. Aims of the Ontology

The RESCS Ontology aims to relate metadata from different disciplinary fields to each other and promote the discovery and exploration of resources in interdisciplinary research. In addition, this ontology aims to facilitate the discovery and exploration of data from non-academic institutions in order to provide important support for those empirical and multidisciplinary research efforts whose goal it is to identify data from highly heterogeneous domains. As far as the use of data from different repositories is concerned, just consider the enormous need in the humanities for data from museum collections, galleries and other cultural heritage databases and archives as well as the countless sources of national and international governmental and non-governmental data that can be used for sociological, economic and legal research. In general, increasing the discoverability and usability of open data and open access publications provides added values for open resources - it will support the adoption of open access principles and FAIR[15] data management by scholars and scientific organizations.

3.2. Foundation of the Ontology

Reusing existing ontologies such as Schema.org and PROV-O offers the benefit of a stable and well known set of entities and properties. Nevertheless, considering the Ontology Competency Questions, the Connectome partners defined a new set of types and properties, alongside their cardinalities, that extend Schema.org and PROV-O into the new RESCS Ontology with the key types (see Table 1).

All types are listed in the form of a types tree, accompanied by a list of all properties. Figure 2 visually illustrates the relationships between the key types in Table 1 of the RESCS Ontology.

None of the previously considered ontologies capture domain-specific research methodologies, such as information on dataset creation (sampling methodologies, social, geographic contexts). Some domain-specific dataset sources (e.g., DaSCH[16]) that are integrated into the Research Data Connectome provide this information and our aim with this ontology is to not lose such details. The extensible RESCS Ontology

[15] https://www.go-fair.org/fair-principles/

[16] https://dasch.swiss/

Table 1
RESCS Ontology Key Types

Types	Description
schema:CreativeWork	The in Schema.org defined type was extended with a set of PROV-O properties and represents "the most generic kind of creative work, including books, movies, photographs, software programs, etc."
schema:ScholarlyArticle	The in Schema.org defined type was extended with a set of new properties and represents publications and/or articles as a result from a project.
schema:Dataset	The in Schema.org defined type was extended with a set of new properties and represents a standalone or publication-related dataset.
schema:Grant	The in Schema.org defined type was extended with a set of new properties and represents the typically financial aspects of ResearchProjects.
rescs:ResearchProject	This new type was defined with a set of new properties and represents the organization of a ResearchProject that can result in schema:CreativeWork outputs.
rescs:Organization	This new type was defined with a set of new properties and represents the organization to which a ResearchProject, Grant, ScholarlyArticle, Dataset or Person is affiliated with. Many properties reuse grid.ac.
rescs:Person	This new type was defined with a set of new properties and represents a Person such as an author of a ScholarlyArticle or Creator of a Dataset.

gives us the precision that future data source integration and applications on top require.

3.3. Serialization of the Ontology

Once the ontology was elaborated, the Connectome partners formalized it into a machine-readable Turtle syntax and file format[17] for expressing metadata in RDF. There is a wide range of tools to collaboratively document, extend and publish ontologies. The Connectome partners explored selected open source tools for documentation (Protégé[18] and OntoDocs[19]) and collaboration (Zazuko[20], GitHub[21]) purposes. The Connectome partners decided to formalize the RESCS Ontology in OntoDocs and visualize it in Zazuko SPEX[22]. OntoDocs is a Python command line application allowing the description of ontologies encoded and formalized in RDF. Zazuko SPEX introspects data residing in SPARQL endpoints. It is using the self-describing nature of RDF based data to give a visual representation of its schema. Correspondingly, the RESCS Ontology is documented on https://www.rescs.org (see Fig. 3) and visualized on Zazuko (see Fig. 2).

[17] https://www.w3.org/TR/turtle/
[18] https://protegewiki.stanford.edu/wiki/Main_Page
[19] https://pypi.org/project/ontodocs/
[20] https://zazuko.com/
[21] https://en.wikipedia.org/wiki/GitHub
[22] https://github.com/zazuko/SPEX

4. Prototyping through Linked Data Pipeline in Blue Brain Nexus

Once formalized and documented, the corresponding ontology files (.ttl Turtle file and SHACL shapes) were imported into Blue Brain Nexus[23], which the Connectome partners installed using a new architecture (see Appendix B) on a prototyping Kubernetes Cluster[24] running on a SWITCHengines OpenStack infrastructure[25]. Blue Brain Nexus consists of the following three distinct components:

- *Nexus Forge*, based on a generic python framework that enables data scientists, data and knowledge engineers to build and search on a knowledge graph.
- *Nexus Delta*, it's a scalable and secure service to store and leverage data organized in a knowledge graph. There is an API available that allows performing data management operations.
- *Nexus Fusion*, it's an extensible web application, which hosts different apps enabling various use cases, it will support workflows for collaborative data and knowledge discovery. Fusion runs on top of Delta.

The API of Nexus Delta was used to store, manage, validate and index by the Connectome Linked Data Pipeline uploaded metadata. This pipeline, elab-

[23] https://bluebrainnexus.io
[24] https://kubernetes.io
[25] https://www.switch.ch/engines/

Fig. 2. RESCS Ontology visualization based on Turtle file using Zazuko Spex.

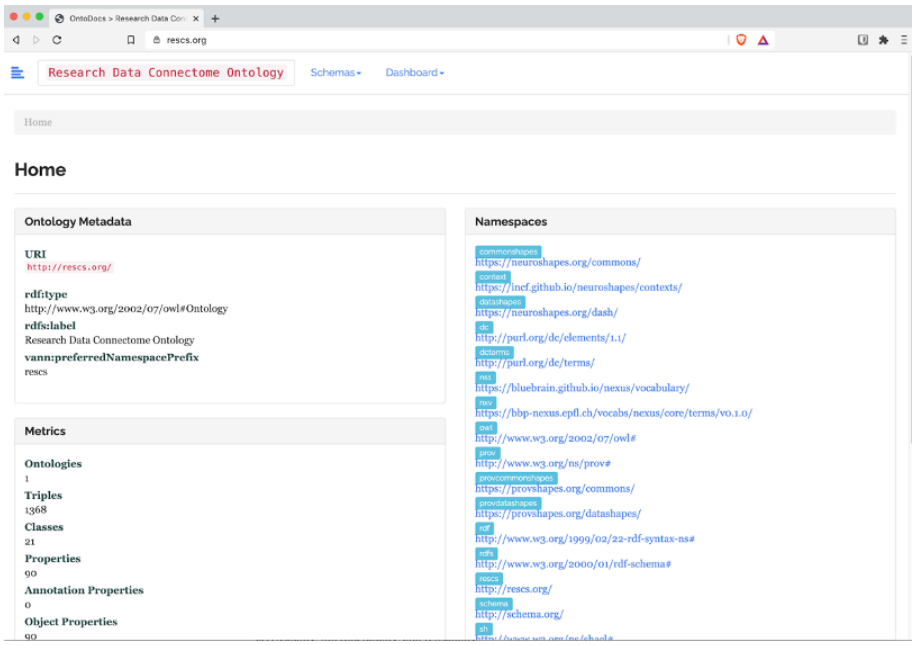

Fig. 3. Documentation of RESCS Ontology on https://www.rescs.org using OntoDocs.

orated by the eXascale Infolab as part of a SWITCH Innovation Lab, aimed to harvest, extract and harmonize metadata from various data providers into the metadata-structure of the RESCS Ontology by using the Nexus Forge features. This activity allowed us to test capabilities of the RESCS Ontology and our estab-

Fig. 4. Screenshot showing an OpenAIRE dataset in the linked date pipeline.

lished Connectome Knowledge Graph running on Blue Brain Nexus. The respective metadata for the *Connectome Linked Data Pipeline* was obtained by integrating metadata from Aramis, DaSCH, Elsevier, Europeana, Eurostat, FORS, OpenAIRE (see Fig. 4) and opendata.swiss.

5. Conclusion & Outlook

The RESCS Ontology designed for the Connectome shows how aggregating open data from different disciplines and repositories into a knowledge graph can create added value for the different stakeholders, allowing a tailored expansion and adaptation to the needs of different scientific communities. The Connectome Knowledge Graph, realized on the basis of the RESCS Ontology by linking and mapping metadata from generic and domain specific repositories, will start supporting use cases coming from SSH research projects from SSH making use of contributions from empirical sciences or data collections not generated in an academic environment. We are thinking, for example, of research projects on the social, economic, or cultural impacts of climate change, or studies of the economic and social impacts of certain religious practices. Other research areas, such as for example the use of botanical data in the study of the prehistoric and ancient world, but also the multiple intersections between economics and historical research could be of particular interest to us. To name a few more research fields, the relationship between philosophy and neuroscience or ethics in the biomedical field could also be an interesting area. In addition, we believe that the RESCS will make a significant contribution to the pluralism of research methodologies by continuously merging and interconnecting different ontologies. One of the aims of the project is to base this process on a close exchange with an open community of users and a scientific committee representing the whole spectrum of academic research. To ensure this, a project governance will be developed to allow members of the scientific community to propose changes and developments to the ontology. Ad hoc developed special interest groups involving different kinds of external stakeholders will also allow continuous exchange with different service providers interested in increasing the exploitation of the Connectome Knowledge Graph generated by the RESCS Ontology. These services will be both generic - i.e., mainly search and visualization functions for metadata - as well as thematic, supporting, for example, the creation of private, discipline-specific knowledge bases and their exploitation through specific data analysis tools. In addition, the integration of research data with "data for research" from extra-academic sources like repositories of galleries, libraries or museums, as well as archives of governmental institutions, administration, private industries, NGOs, etc. represents a powerful catalyst for interdisciplinary and applied research, as well as a significant contribution to citizen science activities.

Acknowledgements

We would like to thank Samuel Kerrien (EPFL Blue Brain), Brian Kleiner (FORS), Manuel Kugler (SATW), Bogdan Roman (EPFL Blue Brain), Christiane Sibille (SAGW), Elfie Swerts (FORS), Mohameth François Sy (EPFL Blue Brain) and Bojana Tasic (FORS) for their contribution to the development of the RESCS Ontology.

Appendix A. List of the Ontology Competency Questions

Table 2

List of the Ontology Competency Questions

Nr.	Ontology Competency Question (CQ) (for research and development)	Nr.	Ontology Competency Question (CQ) (for research and development)
1	How can I access old publications?	36	Who has used a given dataset?
2	How can I access previously locked / withheld information (data and publication)?	37	Who is interested in a dataset?
3	How can I find relevant information?	38	What other datasets were built on the basis of this dataset?
4	How can I discover interesting new material in other domains? (e.g., mathematics)	39	What is the social circle of a dataset?
5	What information from other sources (such as government) relate to my information?	40	Is this dataset trustworthy?
6	What are the activities of my research domain?	41	How is this dataset extendable?
7	How can I find willingly "hidden" yet officially public information?	42	What standards is this dataset using?
8	Where can I find my relevant articles?	43	Has a given dataset been re-used?
9	What methods are my peers using for such a research question?	44	Is a given dataset easy integratable?
10	How can I trust information?	45	In what research was a dataset used?
11	How do I know, that information is not "blacked-out" / censored?	46	Who, Where, When, What and How of the dataset?
12	What is the body of knowledge, where my research is embedded in?	47	How is copyright being handled for this dataset?
13	What are my adjacent research disciplines?	48	What kind of material/data can be found in this domain?
14	What kind of interdisciplinary research matches with my research domain?	49	How may I use the data?
15	Which related topic of research activities exist in my research domains?	50	How can I cite the data?
16	Who are my research peers?	51	Who assembled the data?
17	What is the work of my research peers?	52	What is the status of a project?
18	What are my peers working on?	53	When were the data published?
19	How can I contact my peers?	54	Is there any related publication?
20	What is happening in my research domain?	55	In which language(s) is the content written?
21	What is the most popular research domain?	56	When were the data collected?
22	How many papers are published in my research domain?	57	Which period and/or which geographical area does it cover?
23	How can I get an overview of available data?	58	Which funding does the data have/had?
24	Can I combine various information into a joint whole?	59	Is there any additional information about the project?
25	How can I get an overview of relevant unpublished data?	60	Is there any additional information about the researcher?
26	Who would value my data? / How can I share my data?	61	Is there any additional information about the organization?
27	How can I share my data?	62	How was the data collected?
28	What are relevant case-studies for comparison purposes?	63	What methodologies were used?
29	Is a given dataset objective?	64	Where do my research peers work?
30	Is a given dataset reliable?	65	Can you send me an alert about new relevant datasets?
31	What modes of access are there?	66	Which institutions are focusing on a particular research subject?
32	Can I access company data?	67	Who is successful in obtaining grants in my research area?
33	What articles do you recommend me to read?	68	What is the most popular research domain in my discipline?
34	How can I be informed of new relevant datasets?	69	What are the access conditions of a dataset?
35	Who is working with a given dataset?		

Appendix B. Graphical View of the Connectome Architecture

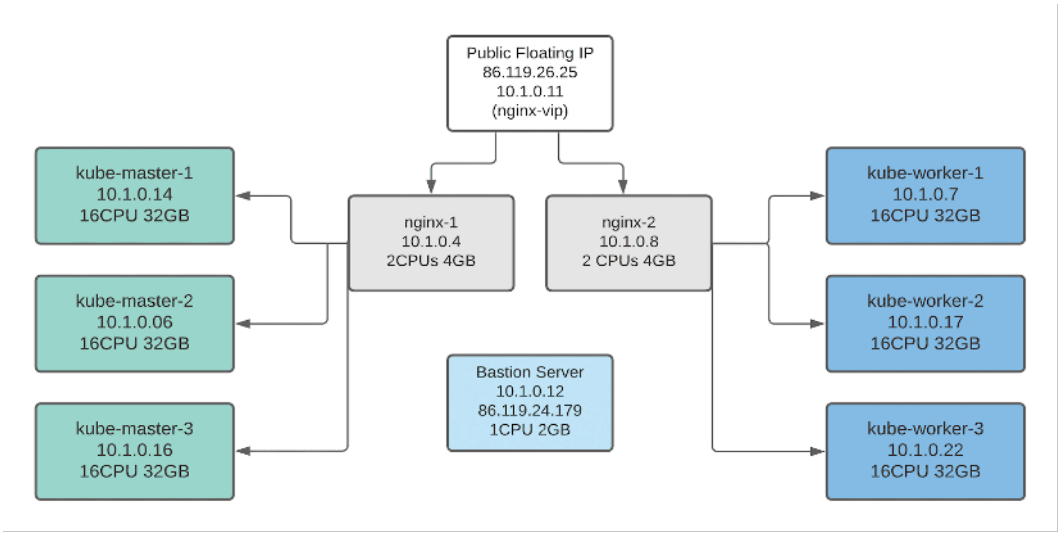

Fig. 5. Graphical View of the Connectome Architecture.

References

[1] D. Pedersen, Integrating social sciences and humanities in interdisciplinary research. Palgrave Communication 2, 16036 (2016), https://doi.org/10.1057/palcomms.2016.36

[2] U. Wiesmann et al. Enhancing Transdisciplinary Research: A Synthesis in Fifteen Propositions. In: Hadorn G.H. et al. (eds) Handbook of Transdisciplinary Research. Springer, Dordrecht, 2008, https://doi.org/10.1007/978-1-4020-6699-3_29

[3] A. Hogan et al. Knowledge Graphs, in: ArXiv, 2020, arXiv:2003.02320

[4] M. Kindling et al. The landscape of research data repositories in 2015: A re3data analysis, in: D-Lib Magazine, March/April 2017, Volume 23, Number 3/4, https://doi.org/10.1045/march2017-kindling

[5] D. Fensel et al., How to Build a Knowledge Graph, in: Knowledge Graphs. Springer, Cham. 2020, https://doi.org/10.1007/978-3-030-37439-6_3

[6] P. Cudré-Mauroux, Design Considerations on SWITCH's Connectome Vision, https://www.switch.ch/export/sites/default/about/innovation/.galleries/files/SWITCHInnovationLab_eXascale-InfoLab_Results.pdf accessed 12/2020

[7] S. Auer, J. Lehmann, Creating knowledge out of interlinked data: making the web a data washing machine, in *WIMS* '11: Proceedings of the International Conference on Web Intelligence, Mining and Semantics, 2011, 1–8

[8] Analysis of Ontology Competency Questions and their formalizations in SPARQL-OWL, in: Journal of Web Semantics, Volume 59, 2020, http://dx.doi.org/10.2139/ssrn.3490821