KAD Milestones 1 & 2: Application Design and Domain Conceptualization

John Can Lokman

VU University Amsterdam

Abstract. This report documents the initial design and prototype of an application that is being made for the Knowledge and Data course of Vrije Universiteit Amsterdam. This application combines various data sources into linked data using an ontology, serializes this ontology and its instances as a triple store, and allows users to query it with a user-friendly interface. In sections 1 and 2, a description of the application and its intended users were described and its initial design is summarized. In section 3 the ontology developed for the application is detailed.

Date: 18.10.2017

1 Description of the Application and Users

1.1 Goal

The current project aims to create a linked-data interface that allows users to query a database of scientific publications in order to get detailed metadata concerning publication patterns . Through the interface, it is particularly aimed to allow a data-driven investigation of interdisciplinary collaboration patterns. Some research questions —among other possible exploratory ones— are as following:

- Are there any collaboration patterns that are biased towards a certain disciplines? For instance, when medical researchers and computer scientists collaborate on research projects, do they tend to publish their research on journals that belong to one of their respective disciplines (i.e., a medical journal versus a computer science journal).
- Does interdisciplinarity of a research project affect its impact (e.g., as measured with number of outgoing citations).
- Are there any publication patterns that can explain researcher career trajectories. For instance, do variables such as interdisciplinarity of an author's lifetime research, number of overall collaborations with other researchers (i.e., network size), and other similar variables affect career-related variables of researchers, such as influence (e.g., as measured by number of citation) or tenure attainment.

The project is part of the 10-month research project 'Knowledge Flows in Interdisciplinary Research' [23], and as the investigation progresses over the next few months, the current linked-data interface is expected to shed light to these research questions, and also motivate new ones.

1.2 Users

The intended initial user base for the linked data interface are the researchers involved in the Knowledge Flows in Interdisciplinary Research project: Dr. Ali Khalili, Dr. Sascha Friesike, Prof. Peter van den Besselaar; and Academy Assistants Frederik König and John Can Lokman. As this research trajectory continues in the following years, more researchers and students who are involved in proceeding —or similar— projects in both Vrije Universiteit Amsterdam and other universities can be expected to join the user base.

2 Design

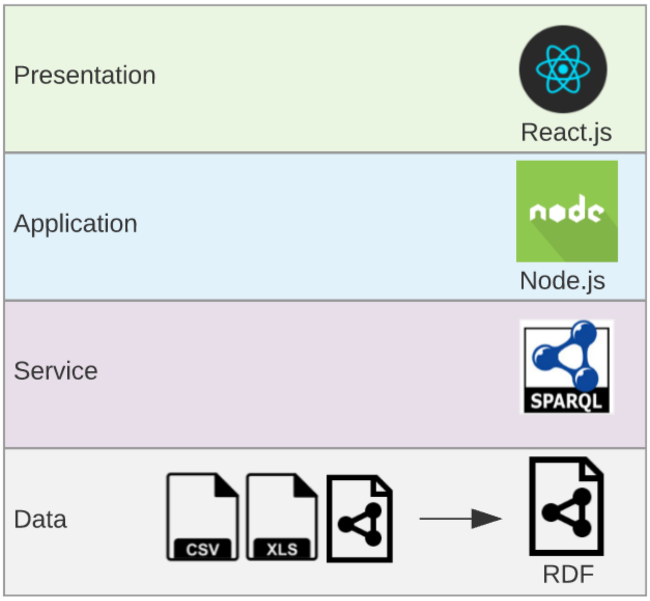

Fig. 1. The technologies we are planning to use for the project, distributed to four layers of application design.

2.1 Data Layer

In the background, the application will be based on an ontology of scientific collaboration, which will model the landscape of scientific publications (i.e., scientific journals) and collaborations between authors and scientific fields. The ontology will be populated with instances by incorporating data from multiple sources such as RISIS [19] and VU Research Portal of Vrije Universiteit Amsterdam—a service that is powered by Pure [8], Scopus [9], and Elsevier Fingerprint Engine [7]. More data sources that contain meta-scientific information could later be added as the project progresses. If such additional data sources do not come in RDF format, necessary transformations will be done using appropriate scripts or methods. (For a summary of application design, see Figure 1.)

2.2 Service Layer

The ontology and its instances will be serialized as a triple store and will be hosted online as linked open data.

2.3 Application Layer

Although this step is to be further specified, the application layer and SPARQL wrapper will likely be implemented using Node.js and Javascript, as part of Linked Data Reactor Framework [12].

2.4 Presentation Layer

Presentation layer will be implemented using mainly HTML and React.js as part of Linked Data Reactor Framework [12]. As the current application is primarily aimed for research purposes, it will be mostly implemented as a an accessible linked data browser designed with 'What you see is what you query' principle [13] (see Figure 3). A short walkthrough to the presentation layer is provided below.

Introduction and Step 1: Datasets Users will be greeted with an introduction page explaining the purpose of the application, and then will proceed to a page that shows them the datasets that will be used for their queries in the next page. This page will be visually similar to the one in Figure 2, and the users will likely be given the option to include or exclude databases with checkboxes.

Step 2: Linked Data Browser After selecting datasets to query, users will proceed to a linked data browser that will be similar to Figure 3. In this interface, users will be allowed to explore the data in a visual and accessible way.

Although the time constraints on the current course will likely will not allow incorporation of additional interfaces, a few possible ideas that may be realized after the course are summarized below.

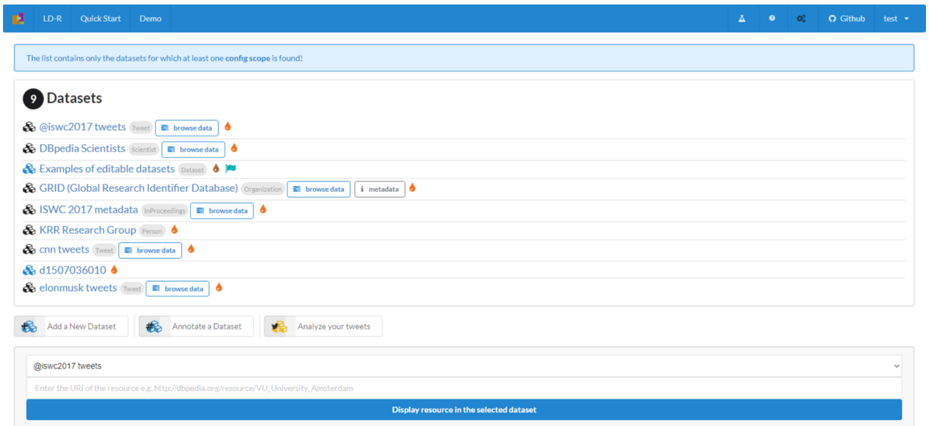

Fig. 2. An example page that lists available datasets. Screenshot taken from Linked Data Reactor [12]

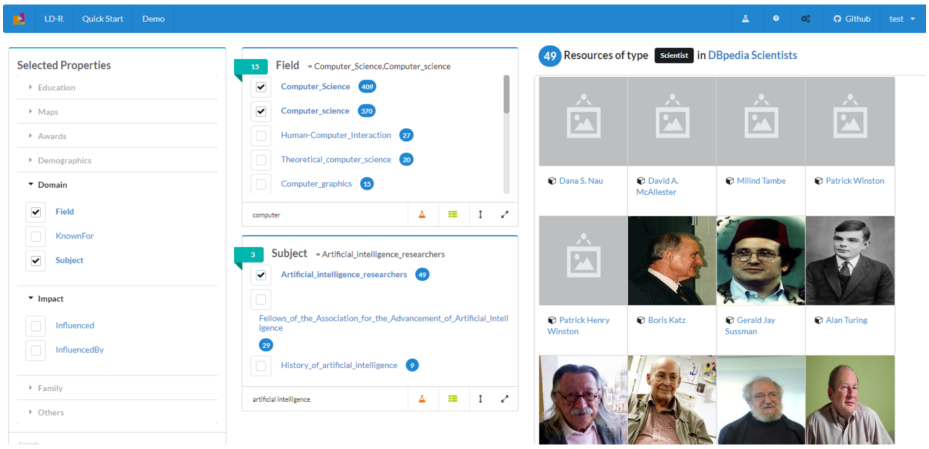

Fig. 3. A screen that closely resembles the envisioned linked data browser for the current project. Screenshot taken from Linked Data Reactor [12].

Possible Future Interface/Component - Search Engine In an alternative interface to that of Linked Data Reactor, a Google-like search engine built to search Meta-science could be used to greet users (see Figure 4). This search engine could allow searching for multiple research fields, researchers, and research projects. Alternatively, such a search box could also be added to the regular Linked Data Reactor browser in Figure 3.

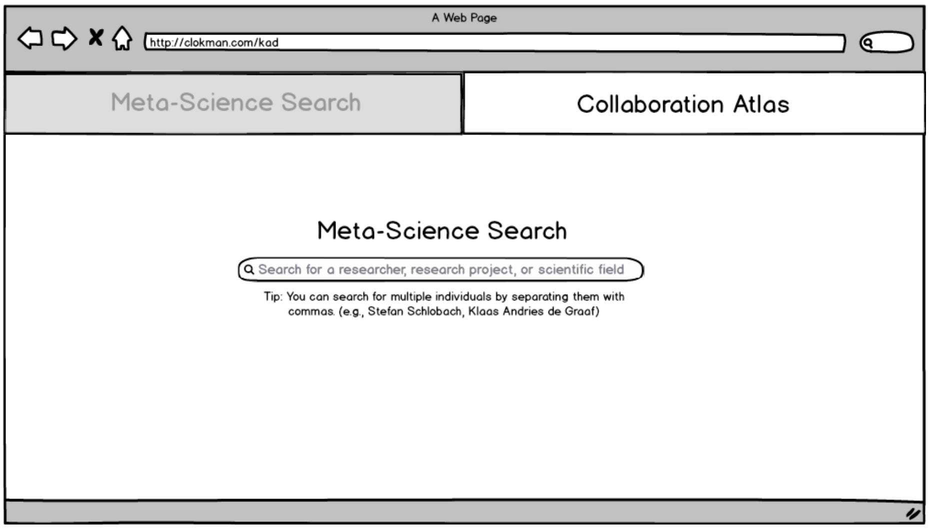

Fig. 4. A concept search engine to query a meta-scientific ontology and the instances it contains.

Possible Future Interface/Component - Linked-data-driven Visualizations The results of the search query could be returned as various interactive visualizations depending on the type of search queries entered (see Figures 5, 6, and 7). Such visualizations could be built using libraries such as D3 (for graphics) and React.js (for other interface elements and operations, such as removing keywords from search query on-the-go—i.e., without requiring users to go back to the initial search page).

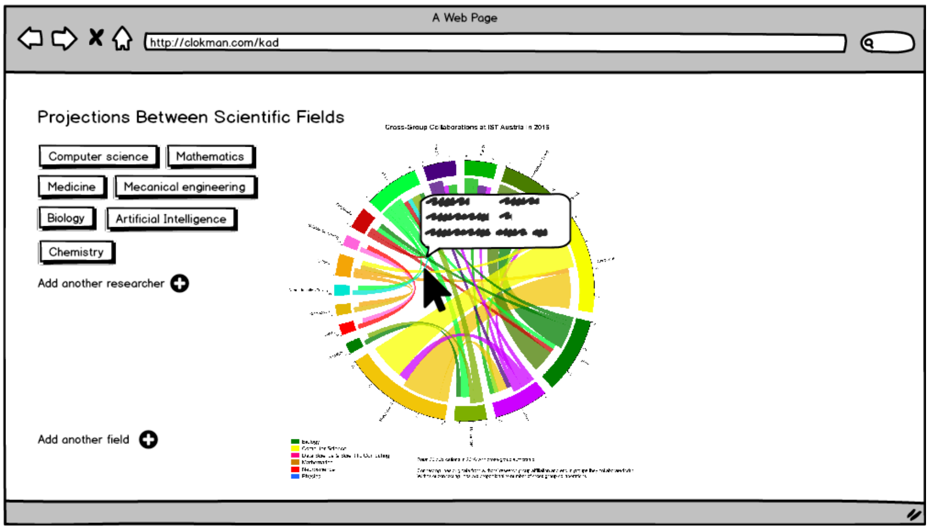

Fig. 5. A concept interactive visualization that could be generated as a response to a search query that relates to scientific fields or topics.

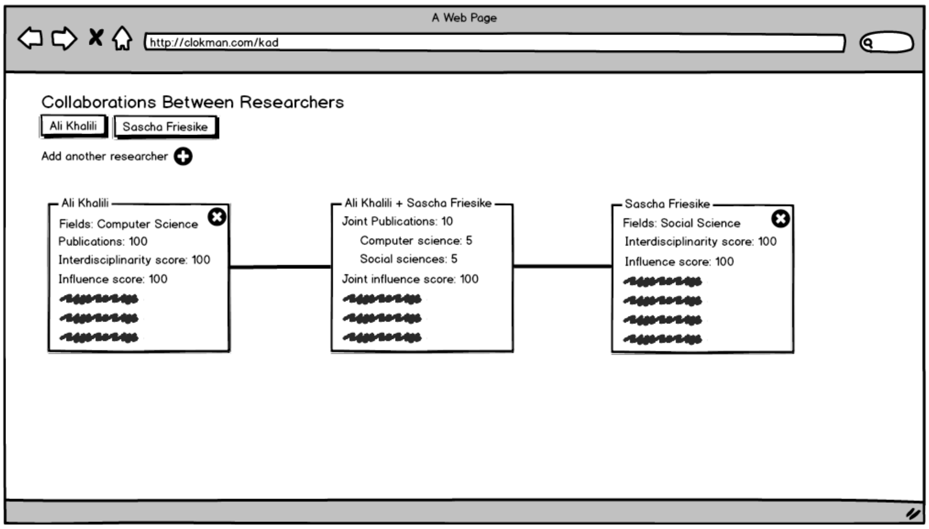

Fig. 6. An interactive visualization that shows collaboration between two researchers as well as other related information for each researcher individually.

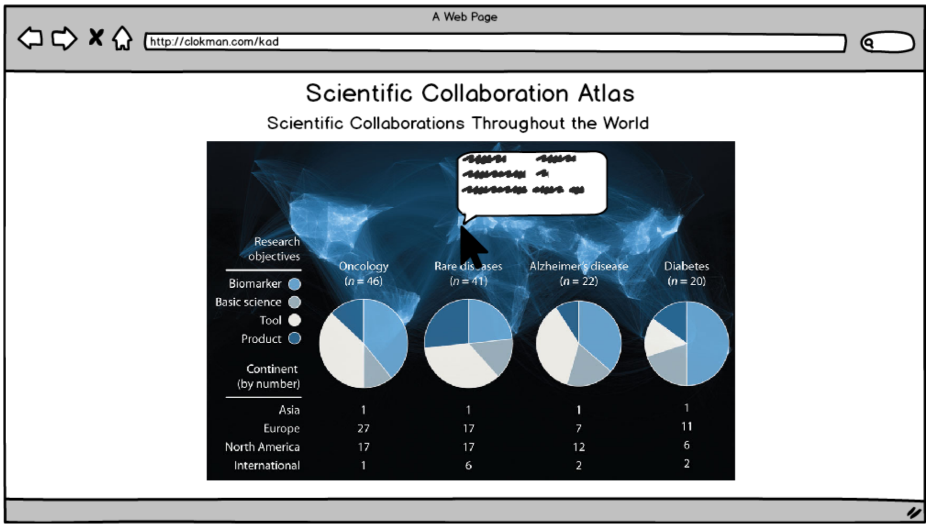

Fig. 7. A concept interactive visualization that maps collaborations based on author locations and research topic. Image taken from [14], and background visualization from [3].

3 Domain Modeling

3.1 Domain

Interdisciplinarity in research is generally seen as desirable, and it is likely to be an important factor that can bring about new perspectives and solutions to our increasingly sophisticated and multi-faceted research pursuits today. However, the impact of interdisciplinarity —or to put simply, the effect of diversity of research in an article, journal, or institute— on the scientific quality and merit is a matter of debate, and there does not seem to be conclusive findings. Some authors suggest that 'distance' between disciplines may play a crticial role in the effectiveness of interdisciplinarity [11,27], and some claim a 'U-shaped' relationship [25], while the discussion also includes various other theories and findings [26,2]. The ongoing debate and possible impact of results on policy making invites more studies in this direction.

As was described in previous section, the goal of the current study is to investigate the domain of scientific publications form an interdisciplinarity perspective and examine any patterns in scientific research. For instance, are there fields that seem to yield higher-impact results when they collaborate, or if being an interdisciplinary researcher lead to more publications? In order to explore the effects of interdisciplinarity and discover scientific collaboration patterns, we aim to model the domain of of scientific publications using an ontology, and populate it using a bibliometric databases such as Elsevier's Pure [8], RISIS [19], and Web

Of Science [5]. In interdiciplinarity research, like other meta-scientific research topics, this bibliometric approach is an often used and considered an effective method [20,16,28,17,4].

3.2 Methodology

Ontology Creation and Revisions In order to create a model of the domain, a first prototype of an ontology was developed during the past month as part of the course, and this model was revised and improved through project meetings. As the ontology progressed —and I gained more experience—methodological changes ocurred and the ontology was significantly changed between revisions. Most notably, the range and class restrictions that were often applied with 'rdfs:range' and 'rdfs:domain' properties were entirely removed due to the reasoning errors and inflexibility they lead to, and also as per expert recommendations (now, more delicate range and domain restrictions are applied through equivalency and subclass relationships where needed). The latest version of the ontology features more sophisticated class definitions (see Fig. 8) through equivalency statements and this results in a more stable ontology and more reliable inferences. Besides the technical advancements, the structure of the ontology was updated based on project meetings. Therefore, the current version consists of a more comprehensive, accurate, and stable model of the domain.

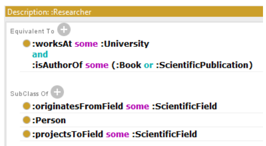

Fig. 8. A class definition that resembles natural language, made by using equivalency and subclass assertions.

Populating the Ontology with Instances In order to populate the ontology with individuals, a bibliography database was initially obtained from VU's *Pure* service [8]. Due to database being in .bibtex format, and due to lack of sufficiently high quality .bibtex to RDF conversion packages in Python, a parser and .ttl converter was programmed using Python (and 'pybtex' package). Although the data preparation part of the ontology creation process had cost most of the available time in this way, the time investment was seen as necessary due to *Pure* dataset being an important element for the project. These scripts can be reached at '`https://github.com/clokman/KAD/tree/master/Milestone_2`'. The reader is encouraged to view the Python scripts as they would likely be able to demonstrate a good degree of understanding of the concepts of the course and may be relevant for evaluation (related files will also be attached together with the submission of this report).

It should also be noted that due to practical concerns (i.e., lack of time and computing power)), a truncated version of the actual Pure database was used for importing instances, and the number of instances imported to the ontology was limited to roughly 375 instances (124 KB in size). The original *Pure* bibliography contains more than 1.5 million lines (about 100 MB), and depending on feasibility of computing power and time, it may be used in the final assignment.

Adding External Classes Although a few example scientific domains (e.g., computer science) were used as placeholders during previous assignments, a comprehensive and accurate domain map of scientific fields were necessary for the purposes of this assignment and for the project in general. Therefore, as a convenient and trustworthy way of categorizing fields of science, ‘ Web Of Science Category Terms’ [22] was added to the ontology as classes through parsing and conversion to .ttl. (This Python script is named ‘d_web-of-science-categories.py’ in the GitHub repository and is also among submission files.)

Future Enhanements for the Ontology Although instances and classes were successfully imported to the ontology, the external files worked with were not in RDF or similar format, and therefore, entities from them they were incorporated to the ontology’s name space rather keeping their own namespaces. Therfore, future versions of the ontology could be more open, and either use a different namespace for imported entities, and directly use external ontologies (in the current version of ontology, there are only one or a couple of such external links, such as ‘foaf:knows’).

3.3 Conceptualization and Realization

The current version of the ontology includes 28 classes for describing the domain of scientific publications and academic research output in general:

- *Publications*: 9 classes including journal articles, conference proceedings, and books (fig. 10).
- *Persons*: 6 classes for authors, researchers, editors, and collaborators (fig. 11).
- *Institutions*: 8 classes for organizations such as universities and publishers 12
- *Scientific fields*: A superclass with over a hundred subclasses for scientific fields imported from Web of Science, this will likely be the main way of defining scientific areas of researchers, journals, and institutions in the future iterations of the app (see fig. 13).

Through using properties (fig. 14), a conceptual network between these classes has been established. 25 object properties describe:

– Citations and publications
– Work and collaboration associations

A visual summary and explanations of the structure of the ontology can be seen below.

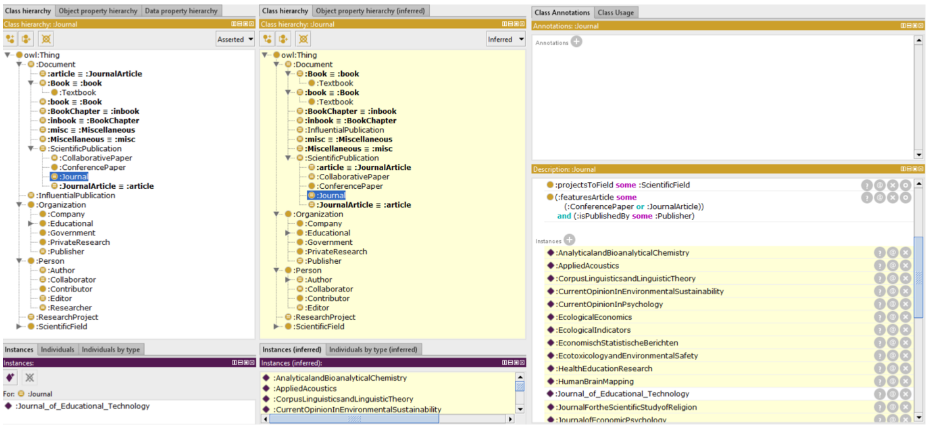

Fig. 9. A screenshot from Protege for a more 'formal' view of the ontology compared to the visalizations in other figures.

3.4 Inferencing

Inferencing The ontology used plenty of class restrictions and has been able to make meaningful inferences on the imported VU-Pure data. For instance, fig. 15 shows imported articles serving as a clue for inferring which journal they belong to. In the earlier versions of the ontology, there were, in fact, more inferences being made due to somewhat more liberal class definitions being in use (fig. 16, also see fig. 14). As the external data tuned better and better to the ontology with the development of the Python scripts used to prepare them, the bibliographical information imported also became more detailed over time, and general inferences (e.g., "all things that has an author is a document") were replaced by more precise assertions that came with the imported file (e.g., "this_instance has type article"). In future, more experimentation could be made to increase the number of inferences, although well-prepared and pre-aligned files may, once again reduce the need for inferencing for crucial information in future work as well. And unfortunately, the prototype is not yet advanced enough to use the *scientific field* or *organization* (i.e., institution) classes (e.g., it cannot yet say an

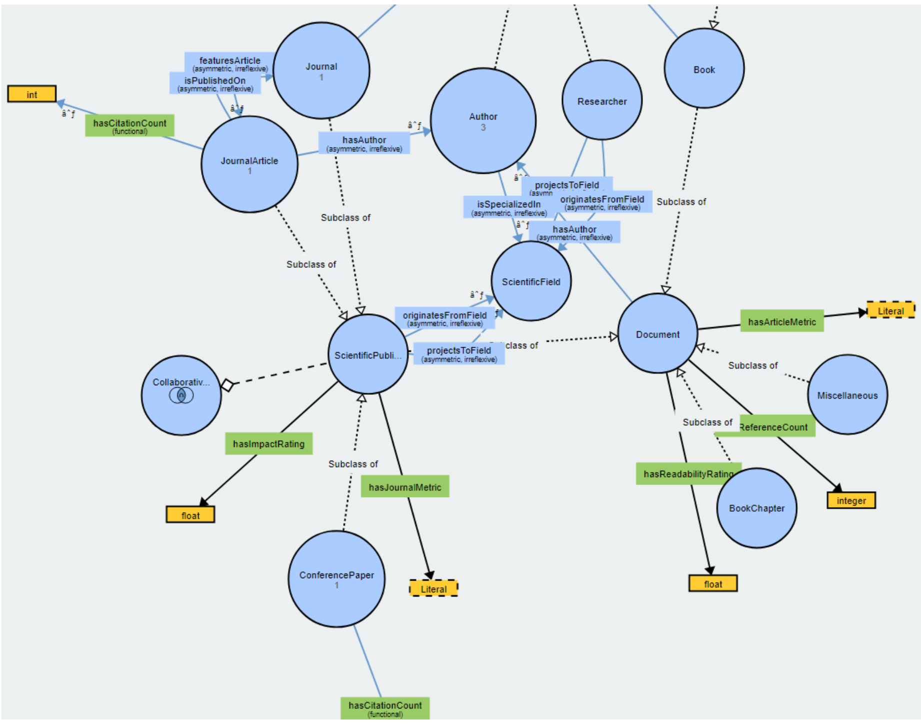

Fig. 10. A model of documents and scientific publications in the ontology.

article belongs to a scientific field). This (crucial) feature will be either implemented next iterations, or if the time constraints prevent it, after completion of the course as part of the research project.

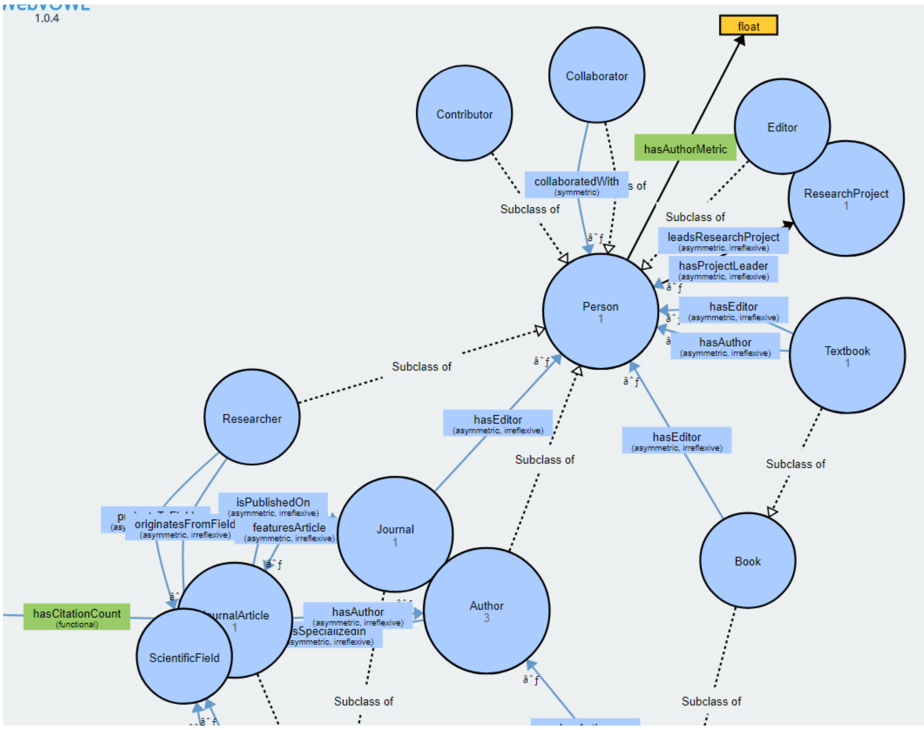

Fig. 11. Part of the ontology that is showing person-related classes and properties.

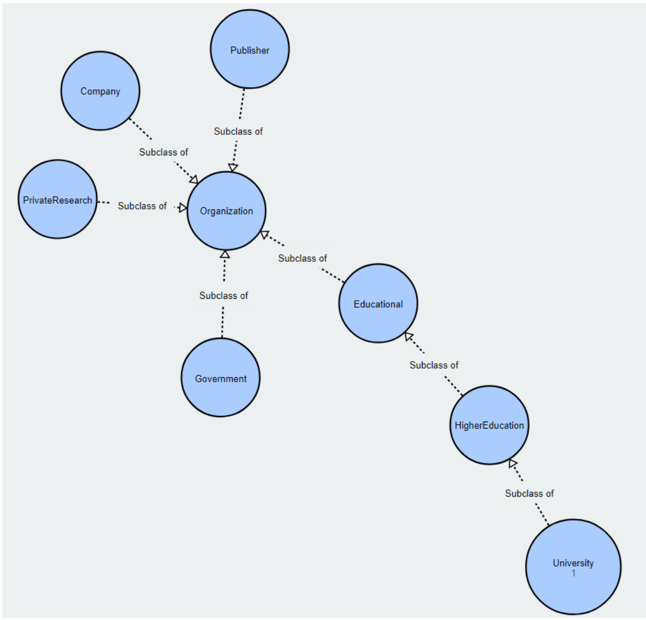

Fig. 12. Organization (i.e., institution) part of the ontology.

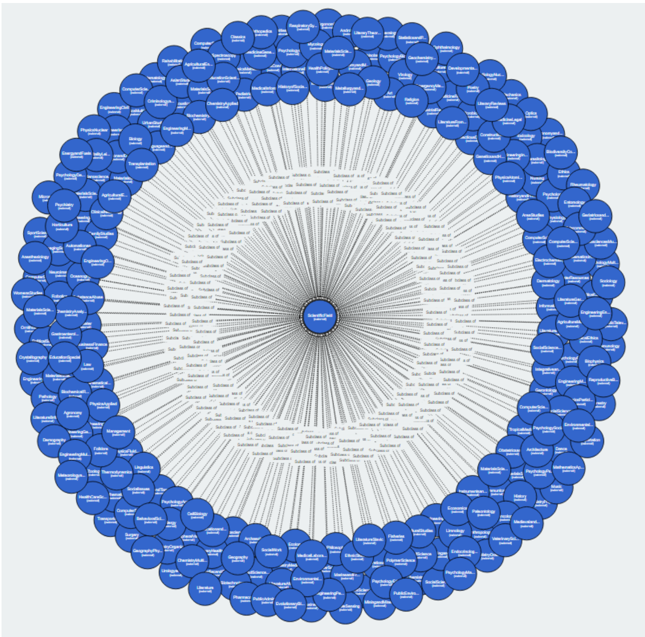

Fig. 13. The superclass 'ScientificField' in the middle, and scientific fields as subclasses on the sides. Subclasses parsed, transformed to Turtle format, and imported to the ontology from Web of Science Category Terms [22]

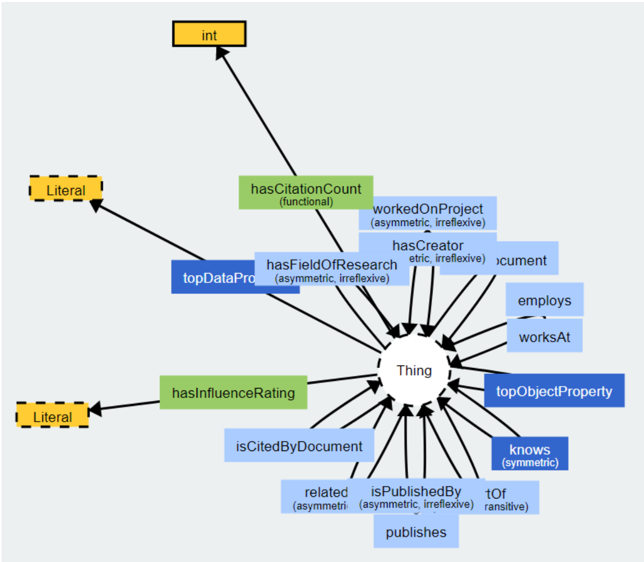

Fig. 14. Some properties in the ontology. At times, broad range and domain of these properties allowed these to be more involved in inferencing.

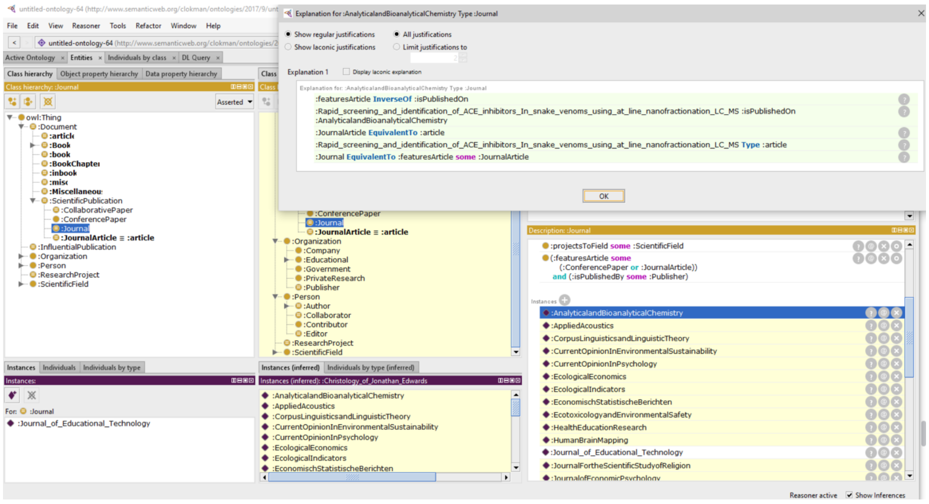

Fig. 15. The inferences on the external data. The journal instances are assigned through the interpretation rules (the reasoning can be seen on the 'Explanation' window).

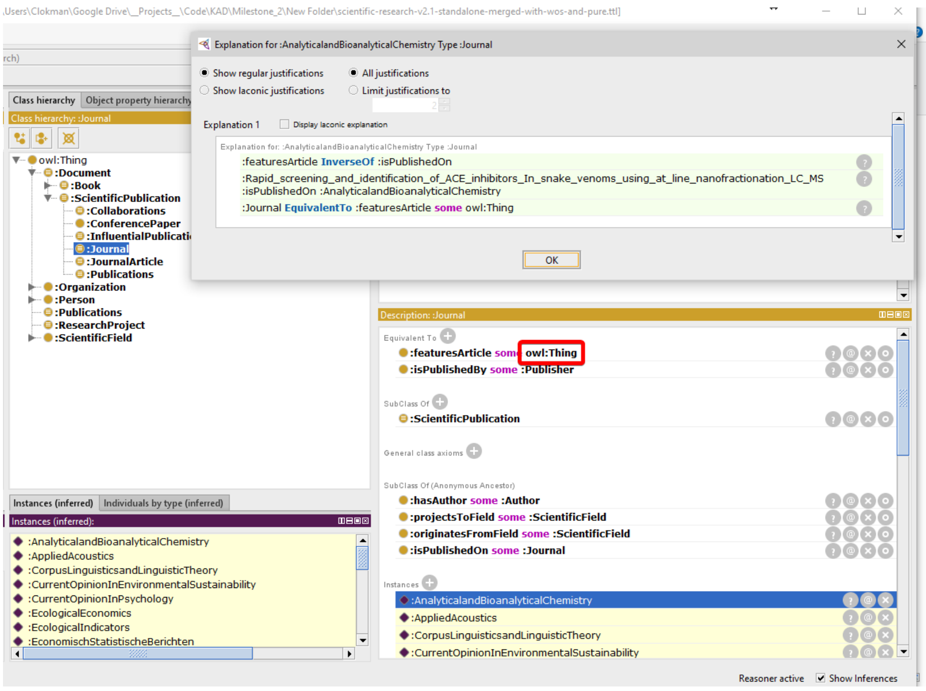

Fig. 16. A previously set, more liberal class definition that allowed a higher number and variety of inferences on the external data (the reasoning can be seen on the 'Explanation' window) , but is highly possible that it also led to less accurate categorizations.

References

1. Ang, H.M., Kwan, Y.H.: Bibliometric analysis of journals in the field of geriatrics and gerontology. Geriatrics & Gerontology International 17(2), 357–360 (feb 2017), https://doi.org/10.1111%2Fggi.12880
2. Barry, A., Born, G., Weszkalnys, G.: Logics of interdisciplinarity. Economy and Society 37(1), 20–49 (feb 2008), https://doi.org/10.1080%2F03085140701760841
3. Beauchesne, O.H.: Map of scientific collaboration between researchers. http://olihb.com/2011/01/23/map-of-scientific-collaboration-between-researchers/ (2011), http://olihb.com/2011/01/23/map-of-scientific-collaboration-between-researchers/, accessed on Fri, October 13, 2017
4. Cardona, M., Marx, W.: Vitaly L. Ginzburg: A Bibliometric Study. In: On Superconductivity and Superfluidity, pp. 217–232. Springer Berlin Heidelberg, https://doi.org/10.1007%2F978-3-540-68008-6_7
5. Clarivate Analytics: Web of Science. http://wokinfo.com (2017), http://wokinfo.com, accessed on Wed, October 18, 2017
6. van Eck, N.J., Waltman, L.: Software survey: VOSviewer a computer program for bibliometric mapping. Scientometrics 84(2), 523–538 (dec 2009), https://doi.org/10.1007%2Fs11192-009-0146-3
7. Elsevier: Elsevier Fingerprint Engine. https://www.elsevier.com/solutions/elsevier-fingerprint-engine (2017), https://www.elsevier.com/solutions/elsevier-fingerprint-engine, accessed on Fri, October 13, 2017
8. Elsevier: Pure. https://www.elsevier.com/solutions/pure (2017), https://www.elsevier.com/solutions/pure, accessed on Fri, October 13, 2017
9. Elsevier: Scopus. https://www.scopus.com/home.uri (2017), https://www.scopus.com/home.uri, accessed on Fri, October 13, 2017
10. Fitzgerald, D., Callard, F.: Social Science and Neuroscience beyond Interdisciplinarity: Experimental Entanglements. Theory Culture & Society 32(1), 3–32 (jun 2014), https://doi.org/10.1177%2F0263276414537319
11. Jensen, P., Lutkouskaya, K.: The many dimensions of laboratories' interdisciplinarity. Scientometrics 98(1), 619–631 (sep 2013), https://doi.org/10.1007%2Fs11192-013-1129-y
12. Khalili, A., Loizou, A., van Harmelen, F., Andries de Graaf, K., Albert Merono-Penuela, Pek van Andel, P.v.: Linked Data Reactor. http://ld-r.org (2017), http://ld-r.org, accessed on Sat, October 14, 2017
13. Khalili, A., Merono-Penuela, A.: WYSIWYQ – What You See Is What You Query (2017), http://research.ld-r.org/papers/wysiwyq.pdf, accessed on Sat, October 14, 2017
14. Lim, M.D.: Consortium Sandbox: Building and Sharing Resources. Science Translational Medicine 6(242), 242cm6–242cm6 (jun 2014), https://doi.org/10.1126%2Fscitranslmed.3009024
15. Morooka, K., Ramos, M.M., Nathaniel, F.N.: A bibliometric approach to interdisciplinarity in Japanese rice research and technology development. Scientometrics 98(1), 73–98 (sep 2013), https://doi.org/10.1007%2Fs11192-013-1119-0
16. Mugabushaka, A.M., Kyriakou, A., Papazoglou, T.: Bibliometric indicators of interdisciplinarity: the potential of the Leinster–Cobbold diversity indices to study disciplinary diversity. Scientometrics 107(2), 593–607 (feb 2016), https://doi.org/10.1007%2Fs11192-016-1865-x

17. Perianes-Rodriguez, A., Waltman, L., van Eck, N.J.: Constructing bibliometric networks: A comparison between full and fractional counting. Journal of Informetrics 10(4), 1178–1195 (nov 2016), https://doi.org/10.1016%2Fj.joi.2016.10.006
18. Prathap, G.: Quantity quality, and consistency as bibliometric indicators. Journal of the Association for Information Science and Technology 65(1), 214–214 (sep 2013), https://doi.org/10.1002%2Fasi.23008
19. RISIS Consortium: RISIS: Research Infrastructure for Science and Innovation Studies. http://risis.eu (2017), http://risis.eu, accessed on Fri, October 13, 2017
20. Roessner, D., Porter, A.L., Nersessian, N.J., Carley, S.: Validating indicators of interdisciplinarity: linking bibliometric measures to studies of engineering research labs. Scientometrics 94(2), 439–468 (oct 2012), https://doi.org/10.1007%2Fs11192-012-0872-9
21. Stardog Union: Stardog: the Enterprise Knowledge Graph. http://www.stardog.com (2017), http://www.stardog.com, accessed on Fri, October 13, 2017
22. Thomson Reuters: Web of Science Category Terms. http://images.webofknowledge.com/ (2017), http://images.webofknowledge.com, accessed on Wed, October 18, 2017
23. VU Network Institute: Academy Assistants and Projects. http://www.networkinstitute.org/academy-assistants/academy-projects-17/ (2017), http://www.networkinstitute.org/academy-assistants/academy-projects-17/, accessed on Fri, October 13, 2017
24. Wang, J., Thijs, B., Gllnzel, W.: Interdisciplinarity and Impact: Distinct Effects of Variety Balance and Disparity. SSRN Electronic Journal (2014), https://doi.org/10.2139%2Fssrn.2548957
25. Wang, J., Thijs, B., Glänzel, W.: Interdisciplinarity and Impact: Distinct Effects of Variety Balance, and Disparity. PLOS ONE 10(5), e0127298 (may 2015), https://doi.org/10.1371%2Fjournal.pone.0127298
26. Yegros-Yegros, A., Rafols, I., D'Este, P.: Does Interdisciplinary Research Lead to Higher Citation Impact? The Different Effect of Proximal and Distal Interdisciplinarity. PLOS ONE 10(8), e0135095 (aug 2015), https://doi.org/10.1371%2Fjournal.pone.0135095
27. Zhang, L., Rousseau, R., Glänzel, W.: Diversity of references as an indicator of the interdisciplinarity of journals: Taking similarity between subject fields into account. Journal of the Association for Information Science and Technology 67(5), 1257–1265 (feb 2015), https://doi.org/10.1002%2Fasi.23487
28. Zulueta, M.A., Bordons, M.: A global approach to the study of teams in multidisciplinary research areas through bibliometric indicators. Research Evaluation 8(2), 111–118 (aug 1999), https://doi.org/10.3152%2F147154499781777612