Ontology Enhanced FAIR Data Point Searches

Xiaofeng Liao[1,*], Coos Baakman[1], Kees Burger[2], Luiz Olavo Bonino da Silva Santos[3,4] and Peter A.C. 't Hoen[1,*]

[1]Radboudumc, Geert Grooteplein Zuid 26/28, 6500 HB Nijmegen, The Netherlands
[2]Health-RI, Jaarbeursplein 6, 3521 AL Utrecht, The Netherlands
[3]University of Twente, PO Box 217 7500 AE Enschede, The Netherlands
[4]Leiden University Medical Center, Postbus 9600, 2300 RC Leiden, The Netherlands

Abstract

The FAIR Data Point has an increasingly important role in efforts to meet FAIR principles. It provides machine readable access to the metadata of different types of digital objects. In this paper, we focus on metadata of datasets. Since its first reference implementation, more tailored implementations have been developed and deployed in the Health Care and Life Sciences domain. However, a problem coming with these increasing amount of FAIR Data Point instances and the datasets published is the Findability of relevant datasets from the large volume of resources. For efficient finding of relevant datasets we need to exploit the richness of their metadata and a good ranking algorithm.

In this paper we report the enhancements of the search and ranking capabilities of FAIR Data Point's reference implementation. Specifically, we improved its semantic search capability via creating association between class terms and the words frequently occur in the class description and labels. We also implemented a TF-IDF based ranking algorithm on the search results to present users the most relevant results.

With these two enhancements, the FAIR Data Point can respond to a user's search request with higher coverage and present the list with the more relevant results based on the Term Frequency - Inverse Document Frequency (TF-IDF) metric.

Keywords

FAIR Data Point, Ontology, Enhancement, Semantic Search, Ranking, TF-IDF

1. Introduction

FAIR Data Point (FDP) is a common approach to publish semantically-rich and machine-actionable metadata according to the FAIR principles[1]. A definition of its software architecture specifying its core components and services to register, index and allow users to search for metadata content of available was given in [2] and a reference implementation[1] was also presented. More tailored implementations have been developed including: The Netherlands eScience

SWAT4HCLS 2024: The 15th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences, February 26–29, 2024, Leiden, The Netherlands

*Corresponding author.
XiaoFeng.Liao@Radboudumc.nl (X. Liao); Coos.Baakman@radboudumc.nl (C. Baakman); kees.burger@health-ri.nl (K. Burger); l.o.boninodasilvasantos@utwente.nl (L. O. Bonino da Silva Santos); Peter-Bram.tHoen@radboudumc.nl (P. A.C. 't Hoen)
0000-0002-4706-1084 (X. Liao); 0000-0003-4317-1566 (C. Baakman); 0000-0002-5437-779X (K. Burger); 0000-0002-1164-1351 (L. O. Bonino da Silva Santos); 0000-0003-4450-3112 (P. A.C. 't Hoen)

[1]https://github.com/FAIRDataTeam/FAIRDataPoint

Center[2], LOVD[3] and The SURF Data Repository[4]. There are also softwares supports the FAIR Data Point protocol, including MOLGENIS software[3] and Castor EDC[5]. The work described in this publication is integrated in FAIR Data Cube [4].

More and more FAIR Data Points are being set up and running with various metadata of datasets from Health Care and Life Sciences domain are published to serve researchers. A detailed list of FAIR Data Point instances can be found in the FAIR Data Point HOME Server[6], where, at the time of this paper's writing, there were 41 active instances hosting metadatas of datasets and other types of digital objects.

An aspect that is currently under-investigated and important to increase the Findability of datasets published are improvements in the engines searching capabilities for relevant datasets. Dataset search is often complicated and inefficient when compared to a typical internet search, where algorithms use criteria of similarity between the potential keywords and the content and links included on websites. The current reference implementation of FAIR Data Point only allows for searching and ranking datasets in a primitive way. The main reason for this technological limitation is that links between datasets are still rare. This compromises the use of traditional web-based ranking algorithms.

In this work, we used the FAIR Data Point reference implementation as the basis for our enhancement work. We implemented a semantic search framework for datasets that can extend existing dataset search tools in two ways: 1. improving the semantic search capability over metadata via association with frequent words occur in class labels and description in ontology. 2. ranking the search results by applying the Term Frequency - Inverse Document Frequency (TF-IDF) [5] metric.

2. Design and Implementation

To imkprove the search capability of the FAIR Data Point's reference implementation, we designed a prototype including:

- A Semantic Query Enhancer (SQE) component, which enhance queries by associating the user's search keywords with terms occur in class labels and descriptions in the pre-loaded ontology.
- A ranking algorithm based on TF-IDF metric to rank the results retrieved in the previous step.

This is done in the following steps as depicted in Figure 1

- index metadata from the ontologies
- retrieve and store associations for the search query words
- find documents for each associated words that has sufficient relevance
- score and rank the documents, before returning them

[2]https://github.com/fair-data/fairdatapoint
[3]https://github.com/LOVDnl/fdp.lovd.nl
[4]https://repository.surfsara.nl/
[5]https://www.castoredc.com/
[6]https://home.fairdatapoint.org/

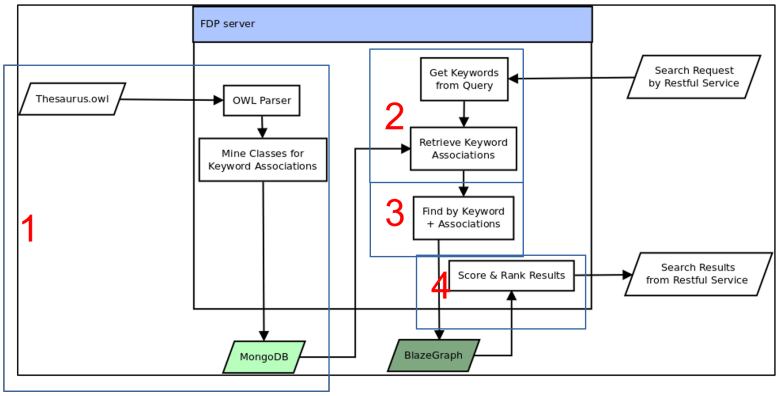


Figure 1: The 4 steps to enhance search query words by association terms from an ontology.

Ontology For ontology, we chose *Thesaurus.owl* and generate association by linking the word to the class description. An example is given in the Figure 2 where "disease" is associated with "pain", because they occur in the same class description. The NCI Thesaurus [6] serves as a widely employed reference terminology designed to enhance translational research in cancer, encompassing both basic and clinical science. Comprising nearly 110,000 terms distributed among approximately 36,000 concepts, the Thesaurus is organized into 20 subdomains. These subdomains encompass diverse areas such as diseases, drugs, anatomy, genes, gene products, techniques, and biological processes. By doing association, we get a result of 2449259 word associations.

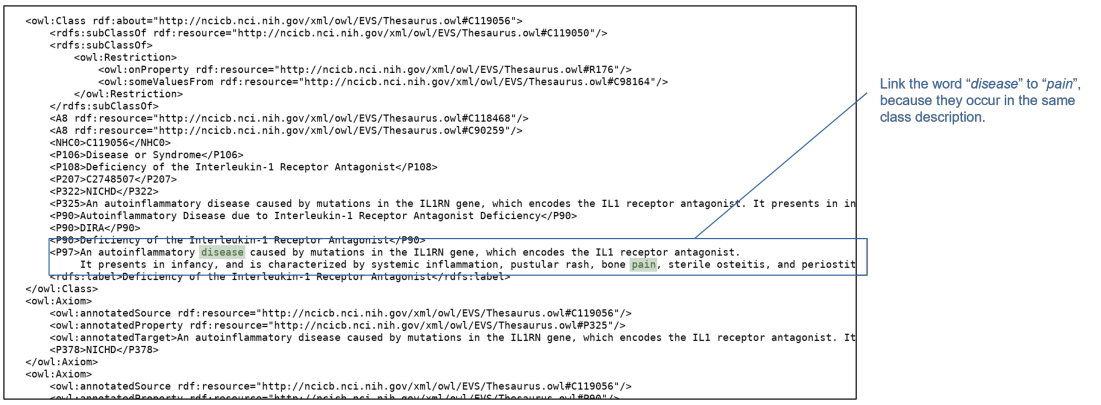
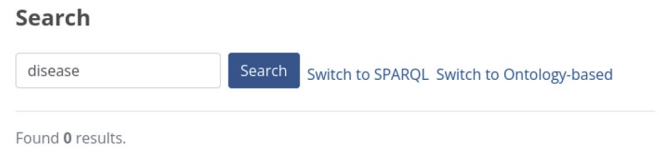


Figure 2: Association is generated via link keywords with terms occurred at class description

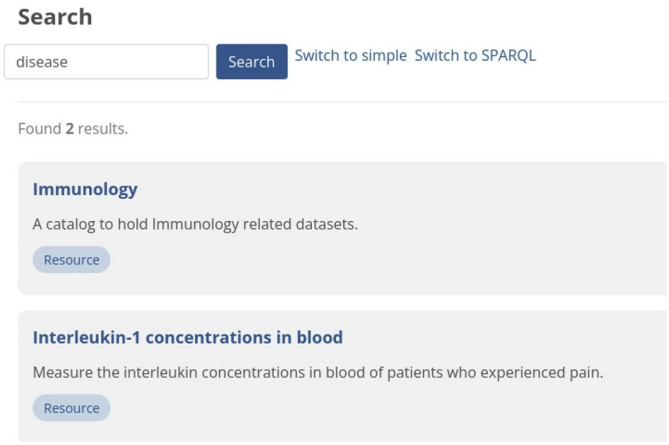
Rank We used a basic TF-IDF [5] algorithm to rank the dataset. The reference implementation applies no ranking algorithm on the result list but only the primitive result from a SPARQL query against the triplestore behind the FAIR Data Point server.

3. Result

On the portal of our ontology enhanced implementation[7], a user can use the ontology enhanced search capability by clicking ”Switch to Ontology-based” link, as shown in Figure 3a. A simple comparison of the search capability between the reference implementation and our ontology enhanced implementation is given in Figure 3. Specifically, in Figure 3a, a search of the keyword ”disease” in the reference implementation gives 0 results. However in Figure 3b, the keyword ”disease” found 2 results. The reason of this difference attributes to the association between ”disease” and ”immunology”/”interleukin-1”, which co-occurred in the class description in the *Thesaurus.owl* ontology.



(a) Results of searching ”Disease” in the reference implementation.



(b) Results of searching ”Disease” in the Ontology Enhanced implementation

Figure 3: The comparison of search results between the reference implementation and the ontology enhanced implementation.

4. Discussion

Due to lack of user logs on the reference implementation of FAIR Data Point server, it is hard to apply a machine learning based ranking algorithm. These user logs supposed to contain the search keywords a user entered and the target results the user clicked. We plan to log the user

[7]http://145.38.186.66/

behaviors at our enhanced FAIR Data Point portal to capture the keywords a user entered and the datasets the user clicked. With these logs, it is possible to train and apply a learning to rank algorithm. With more datasets being submitted and published to our running FAIR Data Point instance, a more detailed evaluation on the search performance in terms of precision and recall would be available.

In the existing setup, the singular ontology *Thesaurus.owl* is employed, given its status as a comprehensive ontology in the field of cancer research. However, it is worth noting that the incorporation of multiple ontologies is feasible, provided that technical challenges such as memory consumption are effectively addressed. In our upcoming implementation, a configuration option will be introduced to enable users to select ontologies based on their specific requirements.

Acknowledgments

This work was supported by SURF-DCC via the pilot:"Enhancing FAIR Data Point's Search Capability as a FAIR Service v2."

References

[1] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al., The fair guiding principles for scientific data management and stewardship, Scientific data 3 (2016) 1–9.

[2] L. O. B. da Silva Santos, K. Burger, R. Kaliyaperumal, M. D. Wilkinson, FAIR Data Point: A FAIR-Oriented Approach for Metadata Publication, Data Intelligence 5 (2023) 163–183. URL: https://doi.org/10.1162/dint_a_00160. doi:10.1162/dint_a_00160.

[3] K. J. van der Velde, F. Imhann, B. Charbon, C. Pang, D. van Enckevort, M. Slofstra, R. Barbieri, R. Alberts, D. Hendriksen, F. Kelpin, et al., Molgenis research: advanced bioinformatics data software for non-bioinformaticians, Bioinformatics 35 (2019) 1076–1078.

[4] X. Liao, A. Niehues, C. de Visser, J. Huang, T. H. Ederveen, C. Doornbos, P. Kulkarni, K. J. van der Velde, M. A. Swertz, M. Brandt, A. J. van Gool, P. A. '. Hoen, Fair data cube, a fair data infrastructure for integrated multi-omics data analysis, medRxiv (2023). URL: https://doi.org/10.1101/2023.04.23.23289000. doi:10.1101/2023.04.23.23289000.

[5] A. Rajaraman, J. D. Ullman, Data Mining, Cambridge University Press, 2011, p. 1–17. doi:10.1017/CBO9781139058452.002.

[6] G. Fragoso, S. de Coronado, M. Haber, F. Hartel, L. Wright, Overview and utilization of the nci thesaurus, Comparative and functional genomics 5 (2004) 648–654.