Converter: Enhancing Interoperability in Research Data Management

Sefika Efeoglu[1,3], Zongxiong Chen[2], Sonja Schimmler[1,2], and Bianca Wentzel[2]

[1] Technische Universität Berlin, Berlin, Germany
{sefika.efeoglu, sonja.schimmler}@tu-berlin.de
[2] Fraunhofer Institute FOKUS, Berlin, Germany
{zongxiong.chen, bianca.wentzel}@fokus.fraunhofer.de
[3] Freie Universität Berlin, Berlin, Germany

Abstract. Research Data Management (RDM) is essential in handling and organizing data in the research field. The Berlin Open Science Platform (BOP) serves as a case study that exemplifies the significance of standardization within the Berlin University Alliance (BUA), employing different vocabularies when publishing their data, resulting in data heterogeneity. The meta portals of the NFDI4Cat and the NFDI4DataScience project serve as additional case studies in the context of the NFDI initiative. To establish consistency among the harvested repositories in the respective systems, this study focuses on developing a novel component, namely the *converter*, that breaks barriers between data collection and various schemas. With the minor modification of the existing Piveau framework, the development of the converter, contributes to enhanced data accessibility, streamlined collaboration, and improved interoperability within the research community.

Keywords: Research Data Management · Interoperability · DCAT.

1 Introduction

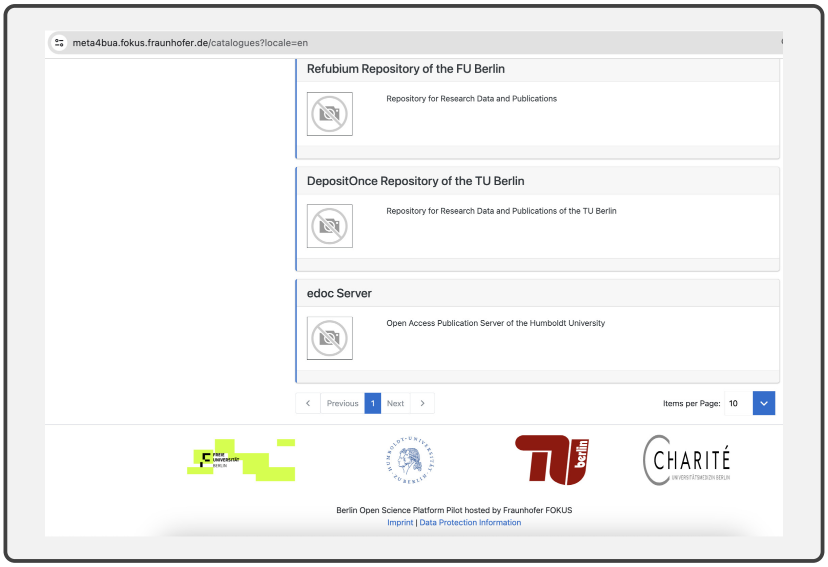
Research Data Management (RDM) plays a crucial role in the research field by facilitating the handling and organization of data. As the volume of data in research areas continues to expand, it becomes increasingly important to address the challenge of managing diverse metadata formats across different applications. One potential solution to this challenge is to standardize the general descriptive metadata into a common format, e.g., the Data Catalog Vocabulary Application Profile [4] (DCAT-AP) [1].

The Berlin University Alliance (BUA), a German excellency cluster, aims to foster collaboration and knowledge sharing among esteemed institutions: Freie Universität Berlin (FU Berlin), Humboldt-Universität zu Berlin (HU Berlin), Technische Universität Berlin (TU Berlin), and Charité - Universitätsmedizin

[4] DCAT: https://www.w3.org/TR/vocab-dcat-3/

Berlin [5]. The BUA's main goal is to establish a digital research space that follows the FAIR (Findable, Accessible, Interoperable, and Reusable) principles, catering to a diverse community of researchers, including professors, scholars, and students. To achieve this, we developed a user-friendly meetup portal called Berlin Open Science Platform (BOP) [6] (see Figure 1). This platform offers convenient access to data resources from the repositories of three universities, all on a single platform. It is based on Piveau [7], a well-established open source data ecosystem, which uses DCAT-AP. The portal brings together and organizes research data from the affiliated universities within the BUA. The research data stored in the BUA repositories encompasses diverse data types including images, doctoral theses, scientific papers, sounds, and audio.

Fig. 1. An overview of three repositories from the institutes of the Berlin University Alliance on the Berlin Open Science Platform.



meta4bua.fokus.fraunhofer.de/catalogues?locale=en

Refubium Repository of the FU Berlin

Repository for Research Data and Publications

DepositOnce Repository of the TU Berlin

Repository for Research Data and Publications of the TU Berlin

edoc Server

Open Access Publication Server of the Humboldt University

Items per Page: 10

Berlin Open Science Platform Pilot hosted by Fraunhofer FOKUS

Imprint | Data Protection Information

Previous
1
Next
FREIE UNIVERSITÄT BERLIN
HUMBOLDT-UNIVERSITÄT ZU BERLIN
TU berlin
CHARITÉ UNIVERSITÄTSMEDIZIN BERLIN

Despite utilizing the oai_dc format [8], the universities within the BUA leverage distinct vocabularies when publishing their data. Using different vocabularies in its schemas causes interoperability problems [2,3]. The current version of the Piveau harvester requires the adaption of its metadata importer when harvesting data from the different repositories within the BUA. It also lacks the capability to establish the corresponding mapping between DCAT and the data retrieved from the different endpoints. These problems raise the need to convert the schemas of the repositories into DCAT before sending the metadata to the Piveau harvester.

[5] This institution's research data is stored in the HU Berlin and FU Berlin repositories

[6] The metadata portal: `https://meta4bua.fokus.fraunhofer.de/`

[7] Piveau is available at `https://gitlab.com/piveau` and `https://github.com/piveau-data`

[8] OAI-DC Schema: `http://www.openarchives.org/OAI/2.0/oai_dc.xsd`

In this work, we developed a novel pipeline that integrates data from different sources and in different schemas. Specifically, we federated data within the BOP in the context of the BUA [9]. The approach can be utilized in other projects, including NFDI4Cat [10] and NFDI4DataScience within the NFDI initiative [11] to build a German National Research Data Infrastructure as well. By employing this pipeline, we can overcome the barriers posed by disparate data resources and harmonize the data into a cohesive and standardized framework.

In the following, we provide a detailed description of how the proposed converter is implemented and integrated into the Piveau framework. This integration is further extended to its application in the BOP project. The step-by-step process of implementing the converter and integrating it into Piveau is described, while its specific application within the BOP is highlighted. Finally, we summarize the contributions our implementation of the converter makes to the Piveau framework.

2 Methodology

We developed a *converter* [12], which finds the corresponding metadata between the schema of the harvested data and the DCAT vocabulary. The corresponding metadata between schemas is found by a schema matcher (see the GitHub repository). This *converter* facilitates the interoperability between DCAT and data resources using different schemas. After finding the corresponding metadata, it saves the harvesting data in the DCAT format replacing its original schema [13]. The *converter* acts as a bridge between repositories and the Piveau harvester [4,5] (see Figure 2), offering a set of different importers, transformers and exporters. Despite having its own transformers, the Piveau harvester needs maintenance for various schemas due to receiving metadata in a different format from the DCAT vocabulary. The Piveau harvester in Figure 2 exclusively receives data in the DCAT format after integrating the *converter*. As a result, there is no need for additional maintenance in the Piveau harvester, even when the incoming data from endpoints is in different schema formats.

With regards to BOP, an example of the significance of standardization can be observed in the BUA. This variation poses a challenge in achieving consistency across the harvested repositories within the BOP. The converter transforms the different data formats and schemas into a standardized format, ensuring that data from the general DSpace repositories [14] of HU Berlin, FU Berlin, and TU Berlin can be easily accessed and utilized, regardless of the specific vocabulary or schema employed by each university. For example, the repositories in the

[9] BUA: `https://www.berlin-university-alliance.de/`

[10] NFDI4Cat: `https://nfdi4cat.org/en/`

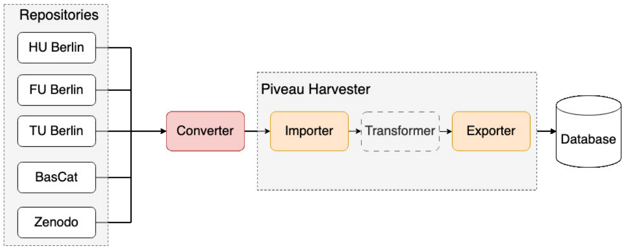
[11] NFDI: `https://www.nfdi.de`

[12] The converter is available at `https://github.com/sefeoglu/dcat-converter`

[13] The sample output of the converter is available at `https://github.com/sefeoglu/dcat-converter/blob/master/data/sample.rdf`

[14] The very commonly used repository solutions include DSpace, Zenodo and Dataverse (see `https://www.re3data.org/`.

Fig. 2. Pipeline: The converter communicates with different repositories and transforms different schemas and vocabularies into a standardized format, i.e., DCAT, and harvester employs importer to fetch metadata from converter and exports to persistent datastore.



BUA utilize the term “subject” in their schemas to define keywords about the data, in contrast to the corresponding term “keywords” in DCAT. Another example is that FU Berlin’s repository uses “abstract” to refer to the publication’s abstract, while the other repositories in the BUA use “description” in their schemas. We investigated schema alignment among metadata of dcat, oai_dc, dc_terms, and dc_elements, considering their metadata’s labels, comments, and definitions, along with prompting ChatGPT [15] by OpenAI [16], and computing cosine similarity with those models’ embeddings. We conducted experiments [17] about prompt templates in [6]. Leveraging this tool, BOP can efficiently harmonize and integrate data from all three universities. By addressing the challenge of data heterogeneity, the converter promotes a unified and cohesive research environment.

In the context of the NFDI initiative, we plan to harvest a variety of data repositories in the future. One of the domain-independent repositories we already harvest is the NFDI4Cat Zenodo [18] community, which is based on the repository software Invenio [19]. Another repository we harvest is a domain-specific Dataverse instance employed at the BasCat laboratory at TU Berlin. Both do not natively support the DCAT schema. In order to maintain a cohesive and standardized database, we can integrate new converter services that facilitate the transformation of arbitrary schemas into our targeted DCAT schema under the proposed framework. By utilizing this service into our data management workflow, we ensure that data from repositories like the Invenio and the Dataverse instance mentioned above can be harmoniously integrated into the NFDI4Cat Meta Portal using the DCAT format. This conversion process enables consistent data representation and enhances interoperability among the different NFDI projects. The flexibility of the converting script allows for the transformation of varying schemas, accommodating the unique characteristics and structures of different data sources within the research communities.

[15] ChatGPT refers to ChatGPT version 4.0

[16] https://chat.openai.com/

[17] The experiments and their results: https://github.com/sefeoglu/dcat-converter/tree/master/schema_matching_experiments

[18] Zenodo is available at https://zenodo.org/

[19] https://inveniosoftware.org

3 Conclusion

We developed a novel service, *converter*, which resolves the interoperability problem between the DCAT vocabulary and the harvested data before sending the retrieved data to the Piveau framework. Our main contributions are listed in the following.

- Without a converter, the Piveau importer is burdened with the task of managing multiple schemas, necessitating codebase adaption for each distinct schema. Moreover, the transformers currently available in the Piveau framework have limited capabilities in handling complex schema mappings. This limitation poses a challenge in effectively transforming and integrating data from different schemas.
- Our work offers a comprehensive solution for effectively managing different vocabularies within the same schema, e.g., oai_dc. Our proposal stands out due to its seamless deployment and easy integration as a pluggable service into the Piveau framework.
- Thanks to the converter introduced, we can eliminate the need for any extensive adaption of the Piveau harvester, streamlining the integration process and ensuring a smooth transition.

The proposed pluggable service, called *converter*, is initially used to demonstrate how the metadata of the institutes in the Berlin University Alliance (BUA) can be converted into DCAT format. However, it can be extended to convert the metadata of other universities into the same format within the metadata portal. In its future work, we are planning to include more institutes.

Acknowledgement. This research presented in this paper was supported by Berlin Uniersity Alliance (BUA) and DFG. Berlin Open Science Platform (Objective 3 and 5 of BUA) receives its funding from BUA, and NFDI4DataScience (460234259) and NFDI4Cat (441926934) receive their funding from DFG.

References

1. Dragan, A., Sofou, N.: DCAT application profile for data portals in europe (2019), https://joinup.ec.europa.eu/sites/default/files/distribution/access_url/2019-05/e3f7bcdf-eaad-4741-9bf6-dc61327f4eea/DCAT_AP_1.2.1.pdf
2. Efeoglu, S.: Graphmatcher: A graph representation learning approach for ontology matching. Ceur-Ws (2022), https://ceur-ws.org/Vol-3324/oaei22_paper7.pdf
3. Jiménez-Ruiz, E., Cuenca Grau, B.: Logmap: Logic-based and scalable ontology matching. In: Aroyo, L., Welty, C., Alani, H., Taylor, J., Bernstein, A., Kagal, L., Noy, N., Blomqvist, E. (eds.) The Semantic Web – ISWC 2011. pp. 273–288. SBH, Berlin, Heidelberg (2011)
4. Kirstein, F., Dutkowski, S., Dittwald, B., Hauswirth, M.: The european data portal: Scalable harvesting and management of linked open data. (2019)
5. Kirstein, F., Stefanidis, K., Dittwald, B., Dutkowski, S., Urbanek, S., Hauswirth, M.: Piveau: A large-scale open data management platform based on semantic web technologies. In: Harth, A., Kirrane, S., Ngonga Ngomo, A.C., Paulheim, H., Rula, A., Gentile, A.L., Haase, P., Cochez, M. (eds.) The Semantic Web. pp. 648–664. Springer International Publishing, Cham (2020)
6. Norouzi, S.S., Mahdavinejad, M.S., Hitzler, P.: Conversational ontology alignment with chatgpt. arXiv preprint arXiv:2308.09217 (2023)