Representing Data Use Restrictions in Descriptive Metadata

Carley Meredith[1]

[1] *Bank of Canada, 234 Wellington St. W, Ottawa, ON K1A 0G9 Canada*

Keywords
metadata schema, DataCite, data use restrictions, descriptive metadata, data catalogue

1. Introduction

The Bank of Canada Library licenses data to support the research of its economists. These data assets are described in a data catalogue that is available Bank-wide. Our descriptive metadata records aim not only to facilitate search and discovery, but also end user compliance with the contractual terms and conditions on data use. With the increasing prevalence of open LLMs and chatbots like ChatGPT, more vendors are including explicit restrictions in their data license agreements on the use of their data in relation to AI. The current project seeks to update our metadata schema for data resources to better describe AI-related limitations and permitted uses.

2. Background

The Bank's data catalogue's default system configuration is based on the DataCite Metadata Schema, mapped to MARC. DataCite's schema is designed primarily to describe "research data and other research outputs" [1]. The Bank therefore developed an internal metadata schema that is based on DataCite but expanded with additional properties—influenced by the Data Documentation Initiative (DDI)—to better describe the technical characteristics and licensing rules of the Bank's data assets. In particular, the Bank's schema expands the Rights section from DataCite to capture more nuances within our vendors' data license agreements.

The Rights property in DataCite, mapped to 540 in MARC, encompasses "any rights information for [the] resource" [2]. The property allows for free text input and can accommodate linking to external license URIs, such as Creative Commons licenses. In the case of data licensed to the Bank, we use this field to link to a redacted copy of the data license agreement provided to us by the vendor, stored in our EDMS. However, we also want to draw out specific aspects of the license agreement to make the most important usage information directly available within the metadata record, encouraging user compliance.

meca@bank-banque-canada.ca (C. Meredith)
0009-0002-4391-3056 (C. Meredith)

https://doi.org/10.23106/dcmi.952456331

To accomplish this, we added additional properties within the Rights wrapper, including Access, Location, Redistribution, and Republication. These fields use controlled vocabularies that are intended to cover all potential scenarios that arise in the agreements. Access and Location describe who at the Bank is permitted to access a given data asset and where it is stored. Redistribution defines whether the licensed user may share the data to other employees within the Bank, and Republication defines whether they may publish the data, such as on the Bank's public website or in a research paper.

A sample record may include the following metadata:

Rights: Data Use Agreement (with link)
Access: Limited Users – Departmental Access for X Department Only
Location: Departmental R:/
Redistribution: Only Registered Users May Redistribute Aggregate Data
Republication: Source Data Fully – Aggregates Only

3. Updating the Schema to Describe AI Use Limitations

The current project aims to further enhance the Bank's metadata schema by developing a property specifically to describe AI use limitations. The need to formalize how we represent these restrictions was flagged by contract management librarians at the Bank, after noticing a trend in vendors adding explicit clauses to their data license agreements restricting the input of their content into AI tools. My project report will outline the following considerations.

First, we needed to determine if the information could be captured in one of our existing metadata properties. For example, we considered adding the restrictions to the 540 Rights field along with the link to the redacted license agreement, or simply using a 500 Note field. We decided that introducing a new property would be preferable for end users, since it would allow the information to be highlighted and flagged by a more relevant field label.

Next, we considered input conventions including the option of using a controlled vocabulary versus free text. While free text offers more flexibility to describe unique cases, we believed that a controlled vocabulary would provide greater simplicity and predictability to the end user. To inform our decision, we compiled all known and potential license scenarios that would need to be covered by the property. We decided to use a controlled vocabulary with three primary options. Feeding licensed content into AI tools is a) not permitted b) permitted, but only if the data remains on a Bank of Canada server or c) users must seek explicit permission in advance.

4. Conclusion

We identified the need for a metadata property that would enable us to describe contractual limitations on inputting licensed data into LLMs. The project report will include the draft schema entry for our new AI restrictions property, implementation updates, expanded use cases, and a more detailed description of the existing data usage properties in our metadata schema for data resources.

https://doi.org/10.23106/dcmi.952456331

Acknowledgements

The views expressed in this paper are solely those of the author and may differ from official Bank of Canada views. Thanks for providing historical context on the Bank's metadata schema goes to Jennifer Trower, Michele Sura, Tyler Brooks, and Elizabeth Smith.

References

[1] DataCite Metadata Working Group, DataCite Metadata Schema for the Publication and Citation of Research Data and Other Research Outputs, Version 4.5 (2024) 1-3. doi:10.14454/g8e5-6293.

[2] DataCite Metadata Working Group, DataCite Metadata Schema for the Publication and Citation of Research Data and Other Research Outputs, Version 4.5 (2024) 39. doi:10.14454/g8e5-6293.

https://doi.org/10.23106/dcmi.952456331