The OpenScience Slovenia metadata dataset

Mladen Borovič*, Marko Ferme, Janez Brezovnik, Sandi Majninger, Albin Bregant, Goran Hrovat, Milan Ojsteršek

University of Maribor, Slovenia

ABSTRACT

The OpenScience Slovenia metadata dataset contains metadata entries for Slovenian public domain academic documents which include undergraduate and postgraduate theses, research and professional articles, along with other academic document types. The data within the dataset was collected as a part of the establishment of the Slovenian Open-Access Infrastructure which defined a unified document collection process and cataloging for universities in Slovenia within the infrastructure repositories. The data was collected from several already established but separate library systems in Slovenia and merged into a single metadata scheme using metadata deduplication and merging techniques. It consists of text and numerical fields, representing attributes that describe documents. These attributes include document titles, keywords, abstracts, typologies, authors, issue years and other identifiers such as URL and UDC. The potential of this dataset lies especially in text mining and text classification tasks and can also be used in development or benchmarking of content-based recommender systems on real-world data.

© 2019 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

* Corresponding author.
E-mail address: mladen.borovic@um.si (M. Borovič).
1. Data

The OpenScience Slovenia metadata dataset comprises of metadata and author data of diploma, master and doctoral theses and research publications of Slovenian universities and research institutions. The data is available in CSV and JSON formats. The dataset consists of four tables: metadata, authors, organizations and typology. The metadata table consists of the following attributes:
| Attribute          | Description/remark                                      |
|--------------------|--------------------------------------------------------|
| ID                 | Metadata identifier; table linking attribute           |
| UDC                | Universal Decimal Classification                      |
| Typology           | Typology according to the COBISS.SI system; table linking attribute |
| Language           | Primary language of the document                       |
| Title              | Title in the primary language                          |
| Subtitle           | Subtitle in the primary language (can be missing)      |
| Abstract           | Abstract in the primary language (can be missing)       |
| Keywords           | Keywords in the primary language (semicolon “;” separated) |
| AlternativeLanguage| Alternative language of the document (can be missing)  |
| AlternativeTitle   | Title in the alternative language (can be missing)     |
| AlternativeAbstract| Abstract in the alternative language (can be missing)  |
| AlternativeKeywords| Keywords in the alternative language (semicolon “;” separated, can be missing) |
| IssueYear          | Year of issue/publication                              |
| URL                | URL link to the publicly available file (links can be dead) |
| OrganizationID_COBISS| Organization identifier according to the COBISS.SI system (https://plus.cobiss.si/opac7/help/cobib/fc/codelist) |
| OrganizationID_eVS | Organization identifier according to the eVS (https://www.gov.si/assets/ministrstva/MIZS/Dokumenti/Visoko-solstvo/eVS-evidenca-VSZ-in-SP/eVS_VSZ_31052019.xlsx); table linking attribute |
| COBISSID           | COBISS.SI system identifier; can be used to query the COBISS.SI system or to link metadata entries in this dataset with the KAS dataset |
| NRID               | National repository identifier; can be used to query the OpenScience Open Data web API |

The authors table consists of the following attributes:

| Attribute   | Description/remark              |
|-------------|---------------------------------|
| MetadataID  | Metadata identifier; table linking attribute |
| AuthorID    | Author identifier               |
| Name        | Name of the author              |
| Surname     | Surname of the author           |

The organizations table consists of the following attributes:

| Attribute   | Description/remark              |
|-------------|---------------------------------|
| eVSID       | eVS identifier; table linking attribute |
| Abbreviation| Organization abbreviation       |
| Name        | Organization name in Slovenian  |
| EnglishName | Organization name in English    |

The typology table consists of the following attributes:

| Attribute   | Description/remark              |
|-------------|---------------------------------|
| Typology    | Typology designation according to the COBISS.SI system (https://plus.cobiss.si/opac7/help/cobib/fc/codelist); table linking attribute |
| Language    | Language of the typology designation description (can be either “slv” for Slovenian or “eng” for English) |
| Description | Typology designation description |
2. Experimental design and methods

As project partners in establishing the Slovenian Open-Access Infrastructure [1], all universities in Slovenia and other included college, higher education and research institutions agreed to provide us with access to their existing digital library systems in order to obtain documents and their metadata. The data collection process consisted of three phases: the existing repository analysis, document and metadata acquisition, and metadata deduplication and enrichment.

2.1. Existing repository analysis

The first phase was an analysis of institutions that should be included in the Slovenian Open-Access Infrastructure and their digital library systems capabilities. We analysed what types of documents are stored in these digital library systems and existing options for data and metadata transfer. The analysis included different digital library systems such as COBISS.SI [2], ePrints and dLib.si with their data and metadata transfer services such as SRU (Search/Retrieve via URL), SRW (Search/Retrieve Web Service) and OAI-PMH (Open Archives Initiative - Protocol for Metadata Harvesting). The formats used were COMARC/B XML and Dublin Core.

2.2. Document and metadata acquisition

The second phase was the acquisition of documents and their metadata from analysed digital library systems. First, the document metadata was acquired using services in existing digital library systems. For metadata acquired from the COBISS.SI system, SRU and SRW services were used. Metadata acquired from ePrints and dLib.si based systems was acquired using OAI-PMH services provided by these systems. The metadata from the COBISS.SI system was retrieved in COMARC/B XML format and metadata acquired via OAI-PMH services was retrieved in Dublin Core format. The documents were acquired using identifiers found in the metadata with the majority of documents being publicly available online in PDF format. URL, URN and DOI identifiers were used in this part to resolve the online source of these documents which were then downloaded to the Slovenian Open-Access Infrastructure servers. The documents which were not publicly available online were either made publicly available online in one or more of the analysed digital library systems or were transferred to our servers using custom-built software after being provided to us on FTP servers and storage devices.

2.3. Metadata deduplication and enrichment

The document metadata was stored in different schemas, depending on the source of the metadata. In the third phase, our custom-built software was developed and used to resolve any duplicate metadata entries describing the same document. Additionally, the same custom-built software was used to detect missing metadata across different metadata schemas and enrich the metadata entries with missing fields using metadata from other metadata schemas. The matching criteria during the merging process were based on titles, abstracts, publication years, authors and other identifiers (URL, UDC). In case of duplicate metadata entries, fields obtained from COBISS.SI services were deemed better quality and were considered final, while metadata from other sources were used to enrich missing metadata fields.

Acknowledgments

The provided data stemmed from the establishment of the Slovenian open access infrastructure. The Slovenian open access infrastructure is partly financed by the European Union, European Regional Development Fund, and the Ministry of Education, Science and Sports of the Republic of Slovenia within the framework of the Operational Programme for Strengthening Regional Development Potentials for Period 2007–2013.
Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] M. Ojsteršek, J. Brezovnik, M. Kotar, M. Ferme, G. Hrovat, A. Bregant, M. Borovič, Establishing of a Slovenian open access infrastructure: a technical point of view, Program Electron. Libr. Inf. Syst. 48 (4) (2014) 384–412, https://doi.org/10.1108/PROG-02-2014-0005.

[2] M. Seljak, T. Seljak, The development of the COBISS system and services in Slovenia, Program, Electron. Libr. Inf. Syst. 36 (2) (2002) 89–98, https://doi.org/10.1108/00330330210429316.