Half-day satellite event of RDA and EOSC Future at the RDA 20th Plenary Meeting
Building blocks of Global Research Commons: Europe and beyond
Session: EOSC and its alignment from a bottom-up perspective

Enhancing Generic Data Descriptors with Discipline Specific Metadata

Prof. Vaidas Morkevičius

March 20, 2023

Lithuanian Data Archive for Social Sciences and Humanities (LiDA)/Kaunas University of Technology
Context

- RDA recommendation
  Guidelines for publishing structured metadata on the Web

- FAIR principles refer frequently to metadata as a key enabler in discoverability, but also having a major role in accessibility, interoperability and reusability

- Publishing structured metadata makes data more discoverable by web search tools

- It also enables rich display of a search result, making it easier for data seekers to judge the relevance of the presented results in terms of the data behind them
**RDA Supporting Outcome Data Discovery Paradigms: User Requirements and Recommendations for Data Repositories**

- As data repositories make more data openly available it becomes challenging for researchers to find what they need either from a repository or through web search engines.

- **Recommendation 3**: Make it easier for researchers to judge relevance, accessibility and reusability of a data collection from a search summary.

- **User requirement**: Fully annotated data (To judge validity, need to know where and when the data was measured, and the basic experimental and instrumental parameters. These are more important than e.g. who created the data.)
Context

- FAIR data principles
  - F2 principle requires that data be described with rich metadata as digital resources and objects that are “not well-described cannot be accurately discovered” (Jacobsen et al., 2020)
  - In addition, R1.3 requires that (meta)data meet domain-relevant community standards
Problem

- **Generic descriptors** such as Dublin Core, DataCite or OpenAIRE provide **basic and rather general information** about datasets
  - **Strength**: can be used for describing **various types and formats** of the data in different disciplines
  - **Weakness**: they provide **little detailed information** about the **specifics** of the data objects that are pertinent to the specific domains and are used in search queries by the users
Problem

- **Generic descriptors** such as Dublin Core, DataCite or OpenAIRE provide **basic and rather general information** about datasets.
  
- For example, finding survey data about attitudes of Lithuanian population towards immigrants in 2007 would be hardly possible without the detailed metadata provided alongside the survey data (for example, in the DDI Codebook).
Problem

- Generic data catalogs and repositories, such as Harvard Dataverse Repository, Dryad Digital Repository, or Figshare Repository, commonly use generic metadata schemes for curating their data.

- Even if descriptions in more specific metadata standards are available in some of them, they are just transformations from and/or additions to more generic formats.
Problem

- EOSC Marketplace Resources Catalogue as envisioned in its architecture would be a Web portal that facilitates searching, discovering and ordering of services from various providers across domains in European countries (EOSC Executive Board, 2021: 7)
- This requires harvesting not only the generic metadata from the research data providers, but also richer domain-specific metadata.
Project

In order to investigate the possibilities of including rich (discipline specific) data descriptions into generic metadata schemes (and in turn repositories), so that generic data descriptors become more compatible with the FAIR Guiding Principles, members of Lithuanian National RDA Node implemented a project Framework for Increased Discoverability of Social Science Data Objects in the EOSC Portal Service Catalogue.
The project aimed to recommend a framework for harvesting and delivering for discovery rich metadata of social science data objects for the generic repositories. The proposed framework allows enriching generic descriptors of data with relevant additional information depending on the types of the social science data object. This conditional model allows for flexibility and comprehensiveness at the same time, as its main operating framework is based on the most widely used generic metadata descriptors, at the same time, integrating additional elements from domain specific metadata descriptors.
Solution

• Studied three generic descriptors: Dublin Core, DataCite and OpenAIRE

• Among the more specific outputs and recommendations targeted at the social science domain we also developed a generic framework for including domain specific metadata into generic data repositories
Solution

• Two elements (terms) that could be employed for enriching generic data descriptors with discipline specific metadata
  • Type (Resource Type)
    • Suitable for differentiating the types of datasets
    • Would allow to produce harvesting algorithms that are able to collect discipline and dataset specific metadata and display them correctly in generic repositories
  • Among the three standards the OpenAIRE Guidelines contain most advanced specification of resource types (1)
Solution

- **Two elements (terms) that could be employed for enriching generic data descriptors with discipline specific metadata**
  - **Type (Resource Type)**
    - Among the three standards the OpenAIRE Guidelines contain most advanced specification of resource types (2)
    - It both requires to identify the general type of the resource in the attribute `resourceTypeGeneral`, and also demands to use COAR Resource Type Vocabulary (with `uri` attribute linking to a vocabulary term) for describing the data set more precisely
  - This standard may also be employed in other generic data descriptors
Solution

• Two elements (terms) that could be employed for enriching generic data descriptors with discipline specific metadata
  • Description
    • Suitable for including detailed information about the various important aspects related to the context, conditions and process of data collection
    • DataCite Metadata Schema already allows more detailed specification of the description types with sub-property descriptionType (1)
Solution

- Two elements (terms) that could be employed for enriching generic data descriptors with discipline specific metadata
  - Description
    - DataCite Metadata Schema already allows more detailed specification of the description types with sub-property descriptionType (2)
    - This sub-property is essential for including different descriptors from discipline specific metadata schemes and needs to be specified, preferably, as a controlled vocabulary
Solution

- **Two elements (terms)** that could be employed for enriching generic data descriptors with discipline specific metadata
  - **Description**
    - [DataCite](http://www.datacite.org) Metadata Schema already allows more detailed specification of the description types with sub-property `descriptionType` (3)
    - Element Description in generic data descriptors could be specified as having attribute `descriptionType` with recommendations for **controlled vocabulary** use, when relevant (with `uri` attribute linking to a vocabulary term)
Studied three generic descriptors: Dublin Core, DataCite and OpenAIRE.

Among the more specific outputs and recommendations targeted at the social science domain we also developed a generic framework for including domain-specific metadata into generic data descriptors.

Solution:

**Cntr for Data Analysing and Archiving**

**Metadata export (publication via OAI-PMH)**

- **LIDA oai: openaire special descriptor (standard):**
  - OpenAIRE Guidelines for institutional and thematic Repository Managers v4.1
  - Generic elements (according to the standard):
    - <datacite:title>Barometer of the Public Opinion Research Center, September - October 1999</datacite:title>
  - Resource Type element (according to standard):
    - <oaire:resourceType>dataset</oaire:resourceType>
  - Description element (modified standard, special crosswalk, for all the detailed descriptions relevant to dataset type, identified according to Resource Type):
    - <description xml:lang="en-US">Abstract: The purpose of the study: assess the opinion of the Lithuanian population on the issues relevant to the society during the rapid political, social, and economic transformation in the country.</description>

**Metadata Aggregator: Ex Libris Primo library discovery service**

**LVB Metadata Aggregator: Ex Libris Primo library discovery service**

**LVB Search Service: Ex Libris Primo library discovery service**

Structured display of search results

**Metadata export (publication via OAI-PMH)**

- **Repository of Data Archive**
  - SSD objects: native or any other metadata standard
  - Metadata export:
    - DOI Codebook v2.5
    - DataCite Metadata Schema v4.0
    - OpenAIRE Guidelines for Data Archives v2.0
    - DC Metadata Element Set v1.1
    - DCMi Metadata Terms
    - Schema.org Etc.

- **DataCite Metadata Aggregator**
  - FDIC/Repository of Data Archive
  - SSD objects: native or any other metadata standard
  - Metadata export:
    - DOI Codebook v2.5
    - DataCite Metadata Schema v4.0
    - OpenAIRE Guidelines for Data Archives v2.0
    - DC Metadata Element Set v1.1
    - DCMi Metadata Terms
    - Schema.org Etc.

**Metadata Aggregator: EOSC Portal Service Catalogue**

**Search Service: EOSC Portal Service Catalogue**

Structured display of search results

- **OpenAIRE/DataCite/DC generic descriptor:** Import via OAI-PMH
  - Indexing and preparation for search services

- **OpenAIRE generic descriptor:** Import via OAI-PMH
  - Indexing and preparation for search services
Solution

Generalized

Domain specific repository

Repository of Data Archive

SSD objects: native or any other metadata standard

Rich metadata in generic descriptors

Metadata export (publication via OAI-PMH)
- Modified generic descriptors: OpenAIRE Guidelines for institutional and thematic Repository Managers
- DataCite Metadata Schema
- Dublin Core terms

- Generic elements (according to the standard):
  
  <datacite:title>Barometer of the Public Opinion Research Center, September - October 1989</datacite:title>
  
- Resource Type element (partially implemented in OpenAIRE language attribute may be added), DataCite and Dublin Core could be updated following OpenAIRE example:
  
  <datacite:resourceType resourceTypeGeneral="dataset" xmlns="en-US" url="http://opendata.o3research.org/REDD-12345" resourceTypeURI="dataset">survey data</datacite:resourceType>

- Description element (partially implemented in DataCite or and language attribute may be added), OpenAIRE and Dublin Core could be updated following DataCite example:
  
  <description descriptionType="Abstract" xmlns="en-US">The purpose of the study: assess the opinion of the Lithuanian population on the issues relevant to the society during the rapid political, social, and economic transformation in the country.
  
Generic repository

Metadata Aggregator: EOSC Portal Service Catalogue

OpenAIRE/DataCite/DC generic descriptor: Import via OAI-PMH

Indexing and preparation for search services

Users searching data

Search Service: EOSC Portal Service Catalogue
Structured display of search results

Lithuanian Data Archive for Social Sciences and Humanities

https://lida.dataverse.it
Modified generic descriptors:
OpenAIRE Guidelines for institutional and thematic Repository Managers
DataCite Metadata Schema
Dublin Core terms

Generic elements (according to the standard):
<dc:title>Barometer of the Public Opinion Research Center, September - October 1989</dc:title>
https://vocabularies.coar-repositories.org/resource_types

Resource Type element (partially implemented in OpenAIRE (language attribute may be added), DataCite and Dublin Core could be updated following OpenAIRE example):
<oaire:resourceType resourceType General="dataset" xml:lang="en-US"
uri="http://purl.org/coar/resource_type/NHD0-W6SY">survey data</oaire:resourceType>

Description element (partially implemented in DataCite (uri and language attribute may be added), OpenAIRE and Dublin Core could be updated following DataCite example):
<descriptions><description descriptionType="Abstract" xml:lang="en-US">The purpose of the study: assess the opinion of the Lithuanian population on the issues relevant to the society during the rapid political, social, and economic transformation in the country.</description></descriptions>
<descriptions><description descriptionType="GeographicCoverage" xml:lang="en-US">Lithuania</description></descriptions>
Use case

• Use case (prototype) was developed that involves servicing of metadata of social science data objects from the domain specific Dataverse repository if the Lithuanian Data Archive for Humanities and Social Sciences (LiDA) for the generic portal of the Lithuanian Academic Electronic Library (LVB), which performs metadata harvesting, indexing and publishing (report)

• This use case could be extended to other generic data repositories that aggregate and present metadata (EOSC etc.)
Use case

- The LVB portal, which works on Ex Libris Primo library discovery service, employs OpenAIRE metadata standard for describing its records.
- Therefore, transformation of fields from Dataverse native JSON descriptor into OpenAIRE Guidelines compliant metadata scheme was developed.
- Since we could only use unmodified OpenAIRE metadata several “work-arounds” were implemented in order to produce results similar to the ones intended in general framework.
Use case

• How it works?
  • 1. Go to the LVB English interface
  • 2. Select the resource Lithuanian Data Archive for Social Sciences and Humanities (LIDA) in the list of search resources
  • 3. Enter lida in the search field
  • 4. From the list of results of short records follow a link to a full record (for example, the fifth record is Emigracija iš Lietuvos pagal šalis, 1919-1940 m., or use permalink)
  • 5. Explore the metadata fields of the detailed record (especially, in the Description field)
## Use case

### Details

<table>
<thead>
<tr>
<th>Kind of Data</th>
<th>Aggregated data : <a href="http://purl.org/coar/resource_type/ACF7-8YT9">http://purl.org/coar/resource_type/ACF7-8YT9</a></th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>Emigracija iš Lietuvos pagal šalis, 1919-1940 m.</td>
</tr>
<tr>
<td>Title [eng]</td>
<td>Emigration from Lithuania by Country, 1919-1940</td>
</tr>
<tr>
<td>Access Rights</td>
<td>License / Data Use Agreement : <a href="http://test-lida.dataverse.lt/dataset.xhtml?persistentId=hdl:21.12137/C0WSSS&amp;version=2.0&amp;selectTab=termsTab">http://test-lida.dataverse.lt/dataset.xhtml?persistentId=hdl:21.12137/C0WSSS&amp;version=2.0&amp;selectTab=termsTab</a></td>
</tr>
<tr>
<td>Dataset Version</td>
<td>2.0</td>
</tr>
<tr>
<td>Creators</td>
<td>Vaskela, Gediminas</td>
</tr>
<tr>
<td>Affiliations</td>
<td>Lithuanian Institute of History, Lithuania</td>
</tr>
<tr>
<td>Data Collection</td>
<td>Vaskela, Gediminas</td>
</tr>
<tr>
<td>Data Curation</td>
<td>Vaskela, Gediminas</td>
</tr>
<tr>
<td>Data Curators</td>
<td>Štreimikis, Antanas</td>
</tr>
<tr>
<td>Distributor</td>
<td>Lithuanian Data Archive for Social Sciences and Humanities</td>
</tr>
<tr>
<td>Contact Person</td>
<td>Žvaliauskas, Giedrius</td>
</tr>
<tr>
<td>Identifiers</td>
<td>hdl:21.12137/C0WSSS</td>
</tr>
<tr>
<td>Other IDs</td>
<td>LiDA_HistatData_0030</td>
</tr>
<tr>
<td>Publication</td>
<td>Kaunas : Lithuanian Data Archive for HSS (LiDA), 2011.</td>
</tr>
<tr>
<td>Creation / Issued Date</td>
<td>2011</td>
</tr>
</tbody>
</table>
# Use case

<table>
<thead>
<tr>
<th>Languages</th>
<th>Lithuanian</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>English</td>
</tr>
<tr>
<td>Subjects / Keywords</td>
<td>gyventojai</td>
</tr>
<tr>
<td></td>
<td>emigracija</td>
</tr>
<tr>
<td></td>
<td>Lietuva</td>
</tr>
<tr>
<td></td>
<td>Istorija</td>
</tr>
<tr>
<td></td>
<td>Migracija</td>
</tr>
<tr>
<td>Subjects / Keywords [eng]</td>
<td>population</td>
</tr>
<tr>
<td></td>
<td>emigration</td>
</tr>
<tr>
<td></td>
<td>Lithuania</td>
</tr>
<tr>
<td></td>
<td>History</td>
</tr>
<tr>
<td></td>
<td>Migration</td>
</tr>
<tr>
<td>CESSDA ELSST Keywords population</td>
<td><a href="https://elsst.cessda.eu/id/c80d7b6c-7b89-4a0e-ab43-941d89888b02">https://elsst.cessda.eu/id/c80d7b6c-7b89-4a0e-ab43-941d89888b02</a></td>
</tr>
<tr>
<td></td>
<td>emigration</td>
</tr>
<tr>
<td>CESSDA Topics</td>
<td>History</td>
</tr>
<tr>
<td></td>
<td>Migration</td>
</tr>
<tr>
<td>Funder Name</td>
<td>European Social Fund, according to the activity “Improvement of Human Resources Quality in Scientific Research and Innovations” of Measure No. 2.5</td>
</tr>
<tr>
<td>Award Number</td>
<td>VP1-3.1-ŚMM-02-V-02-001</td>
</tr>
</tbody>
</table>
## Use case

**Abstract [eng]**

This dataset contains data on emigration from Lithuania by country in 1919-1940.

**General notes:**
The language of the data sources is Lithuanian (lit) and Russian (rus). Dataset (data and metadata) “Emigration from Lithuania Emigration Countries, 1919-1940” from 2011 to 2021-10-01 was available on the portal of Lithuanian Data Archive for Humanities and Social Sciences (LiDA) www лидата.eu; object PID: www.leadat.eu/data/histat/LiDA_HISTDEMX_0030; produced by Šteimikis, Antanas in 2011.

**Time Period Covered:**
- **Start:** 1919
- **End:** 1939

**Date of Collection:**
- **Start:** 2011
- **End:** 2011

**Series Name:** Demography

**Series Information:** dataverse collection “Demography” contains data about population size, density, population size by place of residence (city or village), gender, confession, ethnic or national, social or estate cast; demographic historical data (birth rates, mortality, marriages, divorces, sickness rate, etc.); data about population migration; census data, etc.

**Geographic Coverage:** Lithuania (1919-1940)

**Unit of Analysis:**
- Individual: emigration (DDI Alliance Controlled Vocabulary for AnalysisUnit)
- Political-administrative area (DDI alianso AnalysisUnit kontroliuojamas žodynas)

**Time Method:**
- Time series (DDI Alliance Controlled Vocabulary for TimeMethod)

**Data collector:**
- Central Statistical Bureau under the Ministry of Finance, Lithuania

**Collection Mode:**
- Transcription; Compilation/Synthesis; Other: Calculations (DDI Alliance Controlled Vocabulary for ModeOfCollection)

**Social Science Notes:**
The data for 1940 are for January-May.
Use case

• How it works?
  • 6. Compare metadata fields with those available in the original data object (https://hdl.handle.net/21.12137/C0WSSS)
Half-day satellite event of RDA and EOSC Future at the RDA 20th Plenary Meeting
Building blocks of Global Research Commons: Europe and beyond
Session: EOSC and its alignment from a bottom-up perspective

Thank you!

Contact: vaidas.morkevicius@ktu.lt

This work received funding from the EOSC Future project as part of the RDA Open Calls Programme aimed at on strengthening the work of RDA Communities in EOSC