Extending Basic Dublin Core Elements for an Open Research Data Archive

Andias Wira-Alam, Dimitar Dimitrov, Wolfgang Zenk-Möltgen
GESIS – Leibniz Institute for the Social Sciences, Germany
{andias.wira-alam,dimitar.dimitrov,wolfgang.zenk-moeltgen}@gesis.org

Abstract

In our project "DATORIUM", we intend to provide a simple, open research data repository which focuses on social science research data. We encourage researchers to deposit their data and disseminate them among communities or academic partners. One of the key problems for long-term archiving is ensuring that the metadata elements are consistent and compatible with other standards. This paper discusses the use of basic Dublin Core elements with some simple extensions for structuring the data at study level. Moreover, we also depict the interplay between the emerging combination and the DDI metadata elements, particularly DDI-Lifecycle, and the possibility of using RDF to bring the data into the Linked Open Data Cloud.

Keywords: social science research data; DATORIUM; data documentation initiative (DDI); metadata; Dublin Core; Linked Open Data

1. Introduction

At our institute, GESIS—Leibniz Institute for the Social Sciences, we provide several services, including the Data Catalogue DBK[1] and the Registration Agency for Social Science Research Data da|ra[2], for making a large number of preserved studies[3] visible and available to users. The Social Science Research Data of the GESIS Data Archive includes empirical primary data from survey research, historical social research and texts for content analyses. The important characteristic of the Social Science Research Data is the life cycle process: study concept, processing, disseminating, analyzing, archiving, and repurposing. This process needs to be supported by the use of the metadata standards (as seen in Figure 1).

Currently, we are planning to establish another service called "DATORIUM" as an open repository system for researchers. We aim to support researchers publishing their data via that system by increasing visibility and availability. Research data urgently need interoperable metadata standards to be well understood, especially at study level. One of the most important and widely used metadata standards is Dublin Core, and it is also used not only for publications, but also for research data (Rice, 2008).

Meanwhile, the use of a comprehensive and rich metadata standard, such as the Data Documentation Initiative (DDI), attracts many researchers. These interoperability standards encourage research communities to share digital materials efficiently. DDI supports researchers in creation of high-quality metadata, facilitates reuse of metadata, and supports the life cycle of research data (Vardigan, 2008). For the purpose of this paper, we focus on the main parts of the DDI metadata elements at study level. The DDI elements have the potential to weave native Dublin Core elements into the DDI documents and this makes some aspects of our project easier. In addition, we deal with researchers in the field of social sciences; and, therefore, the use of the DDI metadata elements becomes necessary.

[1] http://www.gesis.org/en/services/research/data-catalogue/
[2] http://www.gesis.org/dara/en/home/about-dara/
[3] We use the term "study" in the sense of research data in the field of social sciences.

https://doi.org/10.23106/dcmi.952136035

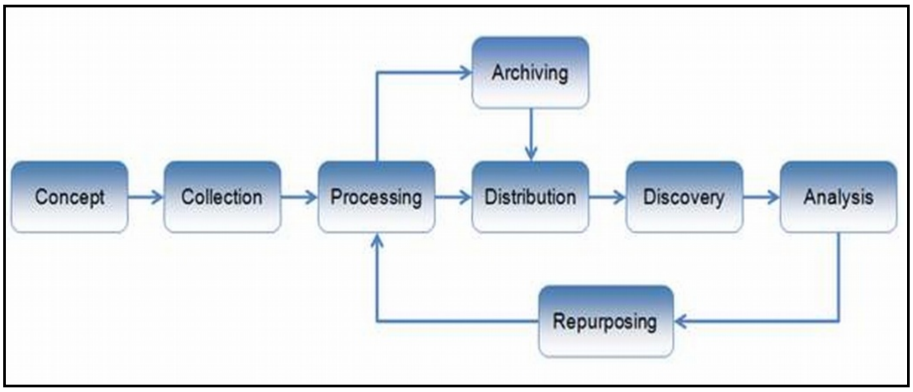


FIG. 1. A typical lifecycle process in the social science research data[4]

The design of the DATORIUM service appears as follows: Any independent researcher may register with the repository and submit studies. Nevertheless, we plan to conduct a review process in order to guarantee the quality of the submitted studies. This makes it possible to gather lots of different studies from all kind of users while simultaneously ensuring that the studies are relevant to social science researchers. In our existing services, we predominately cover larger studies whose research has been funded, e.g. ALLBUS, European Values Study, International Social Survey Programme or Eurobarometer, while also including mid-scale studies from academia. With the new DATORIUM service, we encourage researchers to publish their own smaller studies including data that has been collected for theses, as well as larger datasets that have been partially funded.

The description of studies is a key factor not only in increasing the availability and usability of a study on the Web, but also for long-term preservation of the data. Therefore, it is urgent to use a widely-used metadata standard that supports high interoperability across domains. The use of this standard may potentially attract researchers from other fields. By comparing the typical metadata elements of a study and publication, we discover that there is an intersection between Dublin Core elements and DBK elements that we currently use, e.g. title, creator, or abstract. However, to meet our requirements, we have to extend the intersection with additional elements.

To support our project, we use DSpace[5] as a repository framework. The main reason for this lies in the flexibility and usability of the metadata schemas within DSpace. Furthermore, DSpace supports Dublin Core elements by default and has a flat metadata schema that helps us as developers to maintain the data. DSpace is also open source and has a wide and active community. According to DSpace's website, there are 1289 institutions that have registered to use DSpace for their repository application[6].

2. Metadata Schema

Inspired by Hausstein (2011), we use a subset of our existing metadata from the "Data Catalogue" (DBK) whose schema is nearly flat—i.e., there is no hierarchical structure. The same concept has been implemented in DSpace, where DSpace supports flat metadata schemas which do not have many nested elements. We list the metadata elements that are needed for the DATORIUM project here:

- Title
- Creator
- Abstract
- Distributor
- Format
- ModeOfCollection
- SamplingProcedure
- SpatialCoverage
- StudyUnitID

[4] http://www.ddialliance.org/
[5] http://www.dspace.org/
[6] As of March, 30 2012.

https://doi.org/10.23106/dcmi.952136035

- Description
- DOI
- AccessTypeDescription
- AccessTypeName
- CaseQuantity
- DataCollector
- Note
- Publication
- Publisher
- ReferenceDate (StartDate)
- ReferenceDate (EndDate)
- Subject
- Universe
- VariableQuantity
- Version
- VersionDate

Since we want to leverage the Dublin Core elements, we must first discover which elements from DBK can be mapped into Dublin Core. We also consider other elements to be mapped. Moreover, for long-term preservation, we need to fulfill a requirement to facilitate an export into other standards, e.g. DDI 2.1 or DDI 3, as depicted in Figure 2. Other metadata standards for long-term preservation, including METS, MARC or PREMIS, and SDMX for the purposes of data distribution, could be taken into account (Jensen, 2011). However, since DDI is used at the GESIS Data Archive for long-term preservation and workflow support, we focus on DDI as our main reference point. Therefore, we also need to map the DBK elements to DDI elements which has in fact been done (Zenk-Möltgen, 2012). In addition, Figure 3 depicts the details about the mapping between DATORIUM elements and DDI 3 elements.

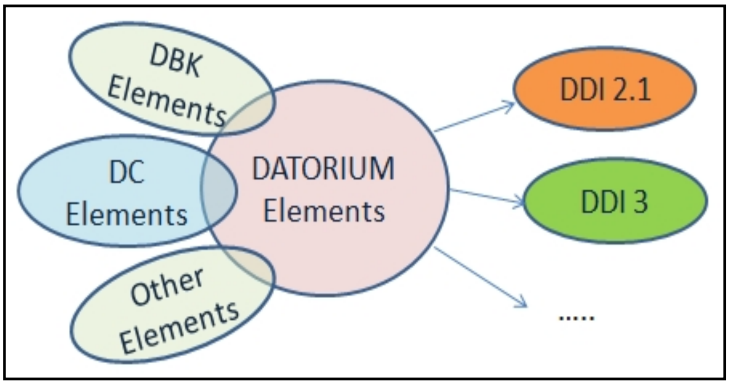
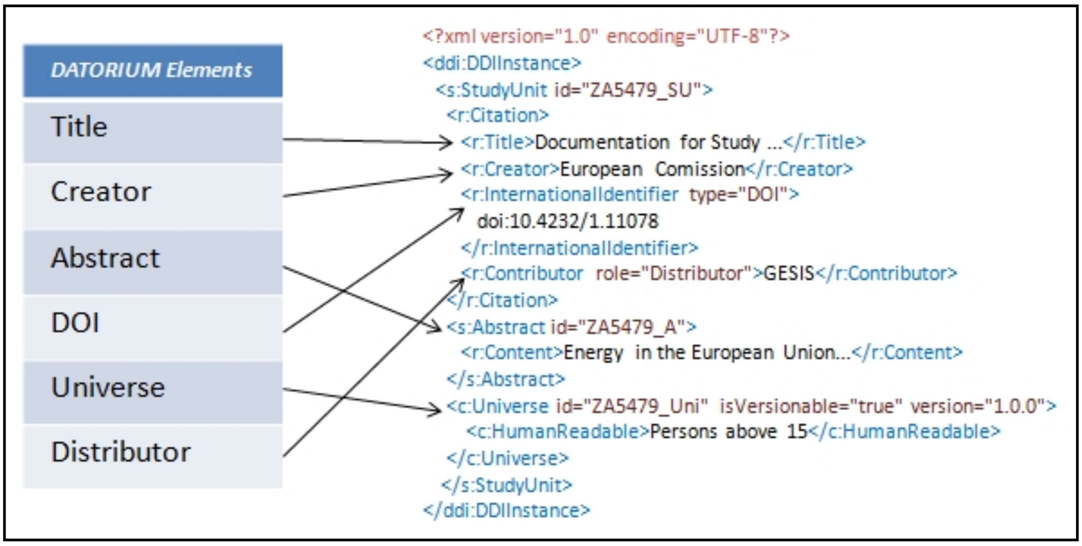


FIG. 2. Schema combination and export



```
<?xml version="1.0" encoding="UTF-8"?>
<ddi:DDIInstance>
 <s:StudyUnit id="ZA5479_SU">
  <r:Citation>
   <r:Title>Documentation for Study ...</r:Title>
   <r:Creator>European Comission</r:Creator>
   <r:InternationalIdentifier type="DOI">
    doi:10.4232/1.11078
   </r:InternationalIdentifier>
   <r:Contributor role="Distributor">GESIS</r:Contributor>
  </r:Citation>
  <s:Abstract id="ZA5479_A">
   <r:Content>Energy in the European Union...</r:Content>
  </s:Abstract>
  <c:Universe id="ZA5479_Uni" isVersionable="true" version="1.0.0">
   <c:HumanReadable>Persons above 15</c:HumanReadable>
  </c:Universe>
 </s:StudyUnit>
</ddi:DDIInstance>
```

FIG. 3. Schema mapping between DATORIUM elements and DDI 3 elements (simplified)

https://doi.org/10.23106/dcmi.952136035

3. Schema Mapping

In order to increase the availability of the research data, we need to have high-quality structuring and interoperability of metadata. Consequently, the data dissemination has to follow the common standard format used on the Web, namely RDF. Since Dublin Core elements are useful and established for describing resources on the Web and also available in RDF, we use them as a foundation for our metadata elements. As seen in Table 1, we show the mapping details of our metadata elements for DATORIUM with Dublin Core and other elements. Overall, several metadata elements can be mapped with Dublin Core elements. The remaining elements that cannot be mapped have to be further investigated in order to find suitable vocabularies from existing ontologies, e.g. OWL[7], BIBO[8], SWRC[9], DDI RDF[10].

TABLE 1: Mapping between DBK elements and DC elements

DATORIUM Elements	DBK Elements	DC Elements	Other Elements
Title	dbk.Title	dc.title	swrc.title
Creator	dbk.PrincipalInvestigator	dc.creator	?
DOI	dbk.DOI	-	bibo.doi
Abstract	dbk.Abstract	dc.abstract	?
AccessTypeDescription	dbk.AccessAvailabilityDescription	-	?
AccessTypeName	dbk.Availability	-	?
CaseQuantity	dbk.NumberOfUnits	-	?
DataCollector	dbk.DataCollector	-	swrc.organization
Description	-	dc.description	?
Distributor	-	-	bibo.distributor
Format	dbk.DataType	dc.format	?
ModeOfCollection	dbk.ModeOfDataCollection	dc.AccuralMehod	?
Note	dbk.Note	-	swrc.note
Publication	dbk.Publication	-	swrc.publication
Publisher	*(fixed)*	dc.publisher	swrc.publisher
ReferenceDate (StartDate)	dbk.ReferenceDate.StartDate	dc.date *(as time interval)*	swrc.startDate
ReferenceDate (EndDate)	dbk.ReferenceDate.EndDate		swrc.endDate
SamplingProcedure	-	-	?
SpatialCoverage	dbk.GeographicCoverage	dc.spatial	?
StudyUnitID	dbk.StudyNo	dc.identifier	?
Subject	dbk.TopicClassification	dc.subject	?
Universe	*(in DBK, SelectionMethod and GeographicCoverage contain information about Universe)*	dc.coverage	?
VariableQuantity	dbk.NumberOfVariables	-	?
Version	dbk.Version	-	owl.hasVersion
VersionDate	dbk.VersionDate	-	?

For the purpose of extending the Dublin Core elements, we have chosen the ontology "Semantic Web for Research Communities" SWRC (Sure, 2005), BIBO, and OWL. There are seven elements of the DATORIUM that could be mapped according to the SWRC vocabulary; four of them have already been mapped according to the Dublin Core elements. However, even if elements are already mapped to Dublin Core, it still makes sense in making additional mappings, because all of them will establish connections to other resources that are only available with the

[7] http://www.w3.org/TR/owl-ref/
[8] Bibliographic Ontology: http://www.bibliontology.com/
[9] http://www.ontoware.org/swrc/
[10] This is an ongoing work under the DDI community to create ontology based on DDI elements.

https://doi.org/10.23106/dcmi.952136035

SWRC vocabularies. In the second and third column, (-) means that there is no corresponding element to the DBK and Dublin Core elements respectively. In the fourth column, (?) means that we have to further investigate the possibility of mapping with other elements from existing ontologies. At this point, there are also tools to discover the correspondences and make the connections between resources in the Linked Open Data cloud[11].

4. Conclusion and Outlook

Dublin Core is suitable for the efficient sharing of research data; therefore we use DC elements for our purpose. We investigated the possibility of mapping between DATORIUM elements and DC elements, with its possible extension into RDF. Therefore researchers are able to share their data among communities efficiently. This mapping also brings the benefit of increasing the availability of the research data in general. Furthermore, the use of DSpace also has advantages since the metadata schemas are easy to maintain and flexible when extended. We also demonstrated the interplay between our DATORIUM metadata elements and the DDI 3 metadata elements.

As a matter of fact, Dublin Core elements are one the most used vocabularies in the Linked Open Data Cloud (Ell, 2011). In the future, we plan to map the remaining elements with existing ontologies. We are currently working with the DDI community in order to produce DDI vocabularies in RDF, thus we can also leverage these vocabularies in the future (e.g., Bosch 2011). Our contribution might also be considered by the Dublin Core community to further develop the metadata vocabularies for describing Research Data. Publishing research data in the Linked Open Data Cloud increases the availability and visibility of the research data.

Acknowledgements

This project is funded by GESIS – Leibniz Institute for the Social Sciences. We thank Monika Linne who worked on project planning and requirements analysis for the project and the DSpace developers for their effort to continually develop a great tool as open source software. Finally, we also thank the reviewers for their useful inputs and advice for improving this work.

References

Bosch, T., Wira-Alam, A., Mathiak, B. (2011). Designing an ontology for the Data Documentation Initiative. Poster Presentation at Extended Semantic Web Conference, ESWC 2011

Ell, B., Vrandečić, D., Simperl, E (2011): Labels in the Web of Data. In Proceeding of the 10th International Semantic Web Conference, 2011.

Hausstein, B., Zenk-Möltgen, W., Wilde, A., Schleinstein, N. (2011). da|ra Metadatenschema: Version 1.0. GESIS Working Paper. Retrieved March 20, 2012 from http://nbn-resolving.de/urn:nbn:de:0168-ssoar-282661

Jensen, U., Katsanidou, A., Zenk-Möltgen, W. (2011): Metadaten und Standards. In: Büttner, Stephan; Hobohm, Hans-Christoph; Müller, Lars; FH Potsdam, FB5 Informationswissenschaften (Hrsg.): Handbuch Forschungsdatenmanagement, Bad Honnef: Bock u. Herchen. S. 83-100, 2011.

Martinez, Luis. (2008). The Data Documentation Initiative (DDI) and Institutional Repositories. Retrieved March 21, 2012 from http://www.disc-uk.org/docs/DDI_and_IRs.pdf

Rice, Robin. (2008). Applying DC to Institutional Data Repositories. Proceedings of the International Conference on Dublin Core and Metadata Applications, 2008, 212

Vardigan, Mary, Pascal Heus, Wendy Thomas. (2008). Data Documentation Initiative: Toward a Standard for the Social Sciences. The International Journal of Digital Curation 3, 1 (2008)

[11] One of the authors of this paper currently also works on this topic. The work was presented at IASSIST Conference 2012 (see http://www.iassist2012.org/indexfolder/program/index.php?show=session:O).

https://doi.org/10.23106/dcmi.952136035

Wackerow, Joachim. (2008). The Data Documentation Initiative (DDI). Proceedings of the International Conference on Dublin Core and Metadata Applications, 2008, 206

Zenk-Möltgen, W., Habbel, N. (2012): Der GESIS Datenbestandskatalog und sein Metadatenschema. Version 1.8. GESIS Technical Reports 2012/1. Retrieved June 21, 2012 from http://nbn-resolving.de/urn:nbn:de:0168-ssoar-292372

https://doi.org/10.23106/dcmi.952136035