Applying a Linked Data Compliant Model: The Usage of the Europeana Data Model by the Deutsche Digitale Bibliothek

Stefanie Rühle
Göttingen State and University Library, Germany
sruehle@sub.uni-goettingen.de

Francesca Schulze
German National Library, Germany
f.schulze@dnb.de

Michael Büchner
German National Library, Germany
m.buechner@dnb.de

Abstract

In 2013/14 the Deutsche Digitale Bibliothek (DDB) changed its data model from the CIDOC conceptual reference model to the Europeana Data Model (EDM). This decision was taken against the background of two major mandates the DDB has to fulfill: as a portal and as a platform the DDB is providing access to digital objects from German cultural heritage and research institutions. The DDB also aims to become the German aggregator for Europeana. Using EDM as the internal DDB data model was considered the most reasonable solution to meet these challenges. The DDB uses the data model for all portal functions that require semantic links between metadata (search facets, hierarchies, links between authority files and digital objects). The application of EDM for the DDB portal created some difficulties since not all necessary classes and properties had been entirely implemented in Europeana-EDM at that time. Therefore, DDB defined a metadata model which is based on the Europeana Data Model Definition but contains additional extensions. The DDB publishes metadata under the CC0 Public Domain Dedication license in EDM-RDF/XML via an OAI-PMH interface to serve Europeana and also via an Application Programming Interface (API) for external users to develop new applications on the basis of metadata harmonized by the DDB.

Keywords: Deutsche Digitale Bibliothek; German Digital Library; Europeana Data Model; CIDOC Conceptual Reference Model; metadata model; metadata mapping; metadata interoperability; linked data

1. Introduction

The Deutsche Digitale Bibliothek (DDB) provides a portal and a platform providing access to digital objects from German cultural heritage and research institutions. It brings together specialists from archives, museums, libraries as well as research, monument protection and media institutions in a Competence Network, funded by federal, state and local authorities. The full version of the portal was launched in March 2014. Besides being the main access point to digitized cultural and academic objects from Germany the DDB aims to become the German aggregator for Europeana, the central access point to Europe´s digitized cultural heritage. Europeana is operated by the Europeana Foundation and provides public services like the Europeana portal[1]. It accumulates and distributes metadata on digital collections from data providers across Europe, for example the DDB. Europeana encouraged the DDB to change the basis for its internal metadata model from CIDOC-CRM to the Europeana Data Model (EDM). EDM is a linked data compliant model developed by Europeana. It uses properties and classes of different namespaces, i. e. terms of the Dublin Core Metadata Element Set, the DCMI Terms and the OAI-ORE (EDM Definition, 2013). The DDB metadata model also uses properties and classes defined by Europeana taking into account the event-based modelling of object lifecycles

[1] URL to Europeana portal: http://www.europeana.eu/portal/

in CIDOC-CRM, however, these descriptions are less complex than in CRM (CRM Definition, 2014). In 2013/14 the DDB replaced CRM with EDM. As a result, mappings to the internal DDB format became less complex which reduces costs of metadata transformations. Using EDM also enables the reusability of Europeana tools. This report presents different applications on the basis of EDM in the DDB and describes the extensions of the model for DDB purposes. With this example, we want to illustrate that EDM is suitable as a domain model for the representation of digital cultural heritage. This model can also be used beyond the purpose of delivering metadata to Europeana. Other projects which adapted or extended EDM for their purpose are for instance The European Library[2], Digitised Manuscripts to Europeana[3] or Europeana Fashion[4].

2. Use of EDM in the DDB

The requirements of the DDB concerning the data model are a result of the EDM triples' functions in the DDB. EDM in the DDB (in the following called DDB-EDM) is used

- for an advanced and facet-based search in the DDB portal,
- to represent the hierarchical organization of the digitized objects,
- to interlink objects and authorities, and
- to publish the data via OAI-PMH and an Application Programming Interface (API).

2.1. Facets

The facet-based search enables users to filter their search results by means of defined categories.

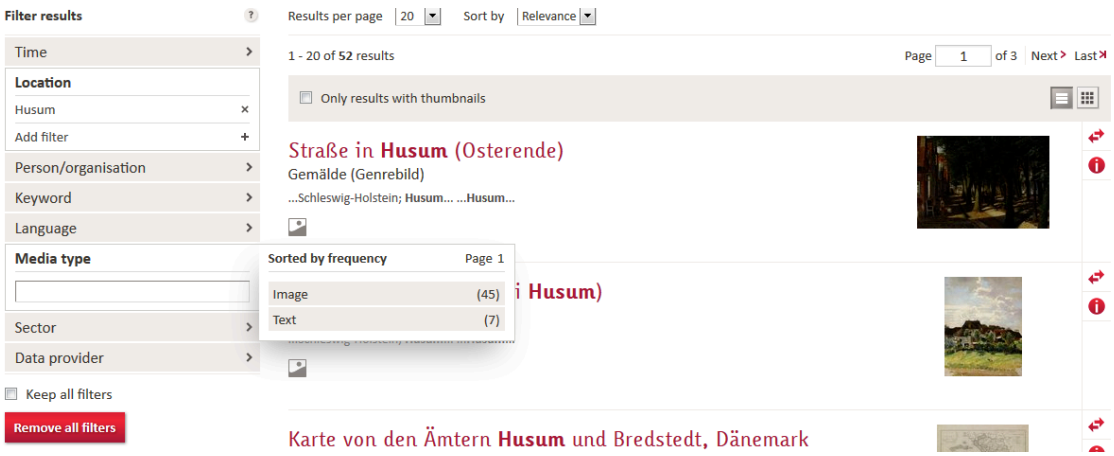

FIG. 1. Facets in the DDB Portal

The categories are based on the classes edm:TimeSpan for time, edm:Place for location, edm:Agent and dcterms:ProvenanceStatement for person/organization and data provider, skos:Concept for keyword, media type and sector, and dcterms:LinguisticSystem for language. In a next step, some of these categories will be refined using triples and controlled terms specifying the relation between an object and a place, time, person or organization.[5] This will allow users to distinguish between the "aboutness" of an object and information concerning its lifecycle and help them to differentiate whether it is the time and place of creation or modification, whether a person was involved in the finding or the destruction of an object etc.

[2] For a project description, see http://dm2e.eu.
[3] For a project description, see http://www.theeuropeanlibrary.org/tel4/.
[4] For a project description, see http://www.europeanafashion.eu/portal/home.html.
[5] For the specification of these relations the DDB uses URIs of the event vocabulary developed by the LIDO Community, see http://terminology.lido-schema.org/eventType.

2.2. Hierarchy

To describe the hierarchical relations between objects – e. g. the hierarchy of resources from libraries or archives – we use two classes to express the different nodes of a hierarchy in the user interface: the edm:ProvidedCHO for objects with a proper name or description (e. g. monographs, journals, articles, illustration) and the edm:PhyscialThing for nodes that are not described with a proper name or description but are needed to express the hierarchical structure (e. g. an issue). We use a domain specific property called `ddb:hierarchyPosition` for the description of the order of resources inside the hierarchy. Besides this property, DDB-EDM includes `edm:isNextInSequence` for compliance with the EDM used in Europeana.[6]

```
− <edm:ProvidedCHO ns3:about="http://www.deutsche-digitale-bibliothek.de/item/FUSPFK23HIF5KRO5OSUMZK6OAW77TB5G">
    <edm:currentLocation ns3:resource="http://d-nb.info/gnd/4023118-5"/>
    <edm:hasType ns3:resource="NLEJVNIMM7RRZIESAF4TEOQB6YOJZU27"/>
    <edm:type>IMAGE</edm:type>
    <dc:type>Illustration</dc:type>
  − <dc:title>
      2. Eiserne Dingstöcke im Altonaer Museum V. l. n. r.: Süderdithmarschen; Gegend v. Bordesholm (?); Gegend v. Ostenfeld, Kr. Husum; Gegei
    </dc:title>
    <ddb:hierarchyType>htype_015</ddb:hierarchyType>
    <ddb:aggregationEntity>false</ddb:aggregationEntity>
    <ddb:hierarchyPosition>DMDLOG_0018</ddb:hierarchyPosition>
    <edm:isNextInSequence ns3:resource="http://www.deutsche-digitale-bibliothek.de/item/I7HEYGWOL5DCZJIWPGGBZ243IQ77TYB7"/>
    <dcterms:isPartOf ns3:resource="http://www.deutsche-digitale-bibliothek.de/item/D2PNNJXEOTEDCQKIIRGJFTVSHDWYNZLH"/>
  </edm:ProvidedCHO>
```

FIG. 2. Description of an edm:ProvidedCHO in DDB-EDM

2.3. Interlinking with Authorities

We use EDM to interlink DDB objects with resources from external data sources. As a first step, we connected DDB objects with person authority files from the Integrated Authority File (Gemeinsame Normdatei, GND). To establish the relations we exploit only GND URIs which are delivered in the original metadata. For persons, who play a role in the lifecycle of an object (e. g. author), we extended EDM with the CIDOC-CRM-Property `P11_had_participant`. For the inverse relation, i. e. from a GND person to DDB objects, we use the EDM property `edm:wasPresentAt`. Furthermore, we use the Dublin Core property `dcterms:subject` for persons, who are described or depicted by the object. To exploit information behind respective GND URIs and to offer person pages in the DDB portal, we apply the web service Entity Facts[7] offered by the German National Library. It allows other applications to integrate and interlink information from GND entities with their data sources. Entity Facts is implementing data enrichment, therefore different data sources (e. g. external links from BEACON files or images of persons from Wikipedia) are merged into a simple and easy-to-use JSON-LD fact sheet. The first version delivers information on entities of the GND entity type Person via an API. Subsequent versions will supply information on places and corporations as well. The GND is widely used in the library community and less represented in other sectors. Therefore, the DDB is developing an assessment tool[8] that will support users to compare, match and map their domain-specific vocabularies to the GND in a semi-automatic way.

2.4. Publication as Linked Data

We provide metadata of the cultural heritage institutions in the DDB-EDM RDF/XML format by applying linked data principles. We use URIs to uniquely identify different resources and their relations in RDF. Therefore, we transfer URIs from the original metadata records during the mapping to EDM whenever possible. Apart from the GND, we take URIs from vocabularies

[6] For information about hierarchies in Europeana see Task Force on hierarchical objects, 2013.
[7] For an example see the query for "Johann Wolfgang von Goethe" at http://hub.culturegraph.org/entityfacts/v1/118540238.
[8] The assessement tool is developed by digiCULT, a project partner of the DDB, see http://www.digicult-verbund.de/.

which are available as Linked Open Data, like Iconclass[9], Dewey Decimal Classification[10] or the Library of Congress vocabularies[11]. We also create URIs, for instance by adding a namespace to a code or identifier provided in the original metadata record (e.g. ISO 639-2 code "eng" to "http://id.loc.gov/vocabulary/iso639-2/eng"). Moreover, we include URIs from the ddb-vocnet namespace into EDM properties to receive controlled terms for the search in the DDB portal. This affects mostly properties, which express the type of a resource (e.g. type of a digital representation of an object). For the identification of some resources, however, it was necessary to additionally establish DDB-internal URIs. These URIs have a DDB-namespace and are created on the basis of common rules for respective DDB resources (e. g. resource class name/ISIL[12]/local identifier). In order for external users to recognize non-resolvable DDB-internal URIs they are encoded by a hash (e.g. EO5NPTOTBJL4V3RXVRLXE7YME7HY6DCW as can be seen in figure 2).

DDB-EDM RDF/XML records contain the results of our normalization and enrichment processes. An example is the use of DDB license URIs for both the metadata record and the digital object. The DDB licenses, which are compliant with the Europeana Licensing Model, give external users information whether and how they can reuse the metadata and digital objects. The DDB publishes its metadata records under the CC0 Public Domain Dedication license via its API[13]. This allows the development of further applications by using DDB metadata. Even though the DDB-API supplies the metadata in different XML formats (source format, DDB-EDM, DDB-View), DDB-EDM is considered as the most harmonized, interlinked and enriched representation of the metadata describing the objects. An application on the basis of the DDB-API is "Archivportal-D[14]" – a portal which provides a view on the DDB content and metadata from an archival perspective. DDB also delivers EDM metadata sets under the license CC0 via an OAI-PMH interface to Europeana. The interface is open to the public as well.

3. Mapping Workflow

The workflow to integrate metadata sets from institutions into the DDB consists of three main steps: 1) clarification of formal and content-specific aspects, 2) data clearing, and 3) ingest. An institution willing to participate has to fill out a content questionnaire including information about the holding/collection and the metadata format (MARC, METS/MODS, ESE, EAD, LIDO et al.). The data clearing begins with the analysis of test data and the adjustment of mapping rules. The original metadata is transformed with XSLT scripts to all DDB target formats, comprising EDM. All metadata representations of an object record are structured in the container format Cortex defined by the DDB. After the ingestion into the DDB test system, data experts review the quality of the transformation result in the test portal and in an XML preview. To support quality control, the DDB is implementing a validation tool. Data clearing is an iterative process with several circles of reviews and adjustments. After approval by the data provider the complete data contribution is ingested into the DDB backend and published via the DDB frontend (portal) and other public interfaces.

The switch from CRM to EDM had a strong impact on our mapping workflow and back-end operations. Since the data sets from all providers that were published via the DDB at that time had to be represented in the new DDB-EDM data format we had a big one-time effort to adjust all respective steps in our workflow. These were: a) the definition of new rules to map the elements and their contents from seven source formats to EDM, b) the indication of provider specific

[9] A classification system for art and iconography. For further information see http://www.iconclass.nl/home.
[10] See http://dewey.info/.
[11] See http://id.loc.gov/.
[12] ISIL is an acronym for International Standard Identifier for Libraries and Related Organisations. The registration for German institutions is managed by the German ISIL and Library Codes Agency at the Staatsbibliothek zu Berlin.
[13] The API of the DDB is documented in the wiki space "API der Deutschen Digitalen Bibliothek", available under the URL: https://api.deutsche-digitale-bibliothek.de/doku/display/ADD/API+der+Deutschen+Digitalen+Bibliothek.
[14] The development of Archivportal-D is funded by Deutsche Forschungsgemeinschaft (DFG). The portal will be launched publicly in September 2014. For a project description in German language see http://www.landesarchiv-bw.de/web/54267.

information in the mappings, c) the adaptation of the transformation tools including the programming of new XSLT scripts, d) the adjustment of the SOLR schema, e) the configuration of the search facets and hierarchies for the frontend, f) the transformation, ingestion and indexing of the complete DDB holdings which comprised around six million records in 2013.

Even though we installed a process that ensured that CRM and EDM records could be ingested in parallel, a few concessions had to be made. For instance, we prioritized the change of the published data sets to EDM. This resulted in a slower increase of content in the DDB since little resources were left for new ingests.

However, the introduction of DDB-EDM decreased the workload for the conceptual and technical mappings considerably. The establishment of mappings to CRM required expert knowledge. Our domain experts, however, were more familiar with EDM because they were already involved in mapping activities for contributing metadata to Europeana via other projects. Furthermore, with EDM the mappings became less complex and less error-prone, because in CRM a statement can be expressed in many ways which often resulted into a series of triples. For example, to state that an object is about a person the mapper had to opt for one of the following paths in CRM:

- E89 Propositional Object (or Subclass) P67F refers to E39 Actor (or Subclass)
- E89 Propositional Object (or Subclass) P129F is about E39 Actor (or Subclass)
- E24 Physical Man-Made Thing (or Subclass) P62F depicts E39 Actor (or Subclass)

We map this statement to DDB-EDM as follows:

- edm:ProvidedCHO dcterms:subject edm:Agent

This example shows that we lost precision in DDB-EDM regarding semantic relations, because the CRM properties "refers to", "is about" and "depicts" were merged into the single EDM property "dcterms:subject". But this generic property is sufficient to distinguish the "aboutness" from the lifecycle of an object which is the crucial requirement for our search facets. This decision was also reasonable regarding the time saved for mappings, the processing of records and thus the ingestion of data contributions.

4. The DDB-EDM Model

The decision to minimize the transformation costs by using EDM in the DDB raised some difficulties. Coming from the event based CIDOC-CRM, the DDB needed properties and classes to describe the events in the lifecycle of the digitized resource. Such properties and classes were available in EDM, but at that time Europeana had not yet implemented them entirely, especially not the necessary event class and its associated properties. Therefore we developed a DDB-EDM model that was an extension of the implemented Europeana EDM described in the Europeana Mapping Guidelines (EDM Mapping Guidelines, 2013).

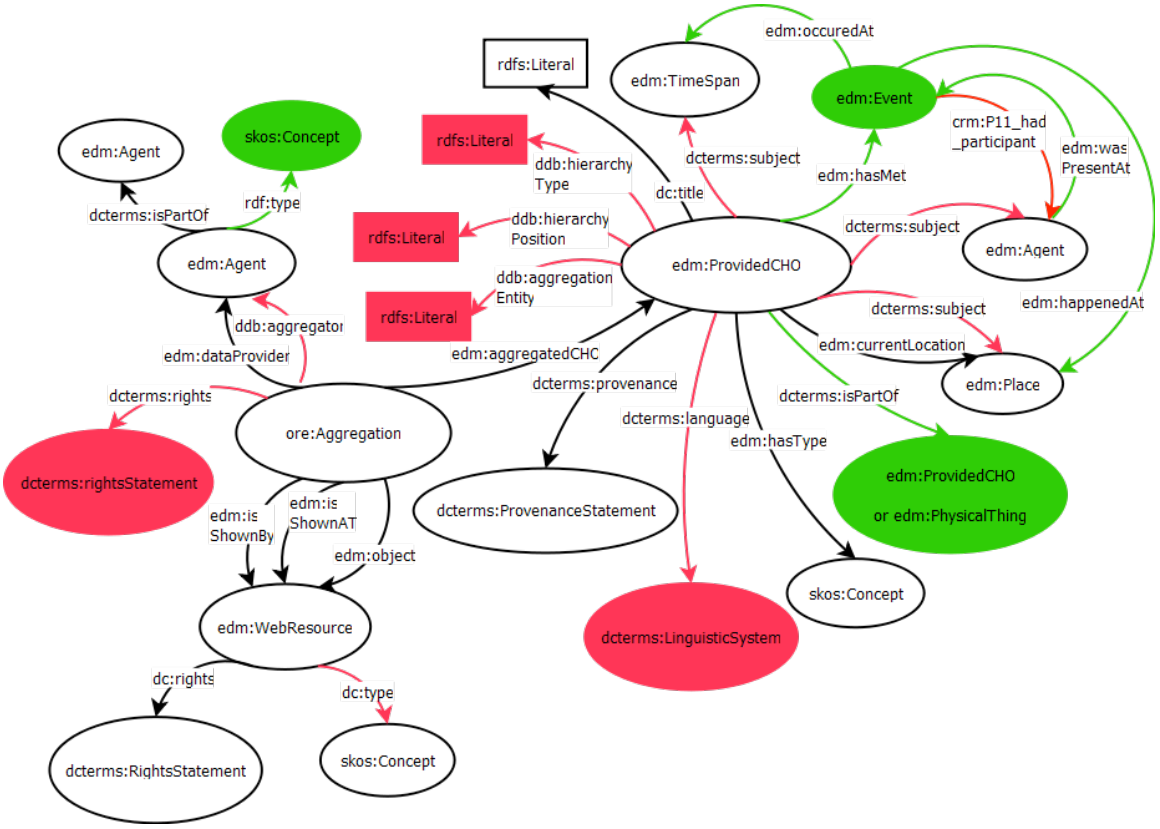

FIG. 3. DDB-EDM model

Figure 3 gives an overview over the properties and classes used in the DDB. Properties and classes used in the DDB, but not implemented in Europeana in 2013, are colored green. This concerns all statements about `edm:Event` and `edm:PhysicalThing`. Properties and classes used in the DDB, but not compliant with the Europeana Model, are red. We use these terms for domain specific requirements. These are:

- `dcterms:rights` with `ore:Aggregation` as domain and `dcterms:RightsStatement` as range, used for rights statements about the metadata. Depending on the value of this property the metadata will be provided by the DDB to Europeana or not[15],
- `ddb:aggregator` with `ore:Aggregation` as domain and `edm:Agent` as range, used for the aggregator providing data to the DDB[16],
- `dcterms:rights` with `edm:WebResource` as domain and `dcterms:RightsStatement` as range, used for DDB specific rights statements,
- `dcterms:language` with `edm:ProvidedCHO` as domain and `dcterms:LinguisticSystem` as range, used to describe the language of the resource with non-literal values[17],
- `dcterms:subject` with `edm:ProvidedCHO` as domain and a non-literal value as range which may be an instance of one of the EDM conceptual classes `edm:Agent`, `edm:Place`, `edm:TimeSpan` etc.[18],
- `ddb:hierarchyType` with `edm:ProvidedCHO` as domain and a literal value as range, used to describe the object type of an `edm:ProvidedCHO` or `edm:PhysicalThing` as part of a

[15] Metadata are only exposed to Europeana or others when the value is "CC0".
[16] Europeana uses edm:provider for Europeana aggregators which in our case is the DDB. Because the property is not repeatable the DDB needs a domain specific property for the description of DDB aggregators.
[17] Europeana uses dc:language and allows the use of literal and non-literal values whereas the use of URIs in the DDB is mandatory.
[18] Europeana uses dc:subject and allows the use of literal and non-literal values whereas the use of URIs in the DDB is mandatory.

hierarchy (e.g. journal, volume, article, illustration). Values used here are based on a vocabulary that will be published as Linked Open Data in the future, which will result in a revision of the DDB-EDM model,

- `ddb:hierarchyPosition` with `edm:ProvidedCHO` as domain and a literal value as range, used to describe the order of an `edm:ProvidedCHO` or `edm:PhysicalThing` in a hierarchy,
- `ddb:aggregationEntity` with `edm:ProvidedCHO` as domain and a literal value as range, used to distinguish between hierarchical levels with proper descriptions and levels without such descriptions (e.g. an issue that is only identified by the number),
- `rdf:type` with `edm:Agent` as domain and `skos:Concept` as range, used to describe the relation between a corporate body and the type of sector it belongs to, and
- `crm:P11_had_participant` with `edm:Event` as domain and `edm:Agent` as range, used to describe that there is a relation between an event and an agent (e. g. the creation event and the creator).

5. Conclusion and Outline

The implementation of EDM has turned out to be the most effective way to serve the requirements of the DDB portal for functions based on linked data principles and external applications like Europeana. Prospectively, DDB-EDM will also contain the results of further enrichment and normalization processes the DDB is currently establishing for authority data and controlled vocabularies which will subsequently improve the portal as well.

References

CRM Definition (2014). Definition of the CIDOC Conceptual Reference Model, Version 5.1.2. Retrieved April 24, 2014 from http://cidoc-crm.org/docs/cidoc_crm_version_5.1.2.pdf

EDM Definition (2013). Definition of the Europeana Data Model, version 5.2.4. Retrieved April 24, 2014 http://pro.europeana.eu/edm-documentation

EDM Mapping Guidelines (2013). Europeana Data Model – Mapping Guidelines, Version 2.0. Retrieved April 29, 2014 from http://pro.europeana.eu/edm-documentation

Task Force on hierarchical objects (2013). Recommendations for the representation of hierarchical objects in Europeana. Retrieved April 29, 2014 from http://pro.europeana.eu/web/network/europeana-tech/-/wiki/Main/Taskforce+on+hierarchical+objects