A specialisation of the Europeana Data Model for the representation of manuscripts: The DM2E model

Evelyn Dröge

Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany.

Email: evelyn.droege@ibi-hu-berlin.de.[1]

Julia Iwanowa

Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany.

Email: julia.iwanowa@ibi-hu-berlin.de.

Steffen Hennicke

Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany.

Email: steffen.hennicke@ibi-hu-berlin.de.

Abstract

The RDF-based Europeana Data Model (EDM) (EDM Primer, 2013) is used by Europeana, the European Digital Library, for representing heterogeneous data coming from museums, libraries, archives and galleries. The model combines various standards and existing ontologies and is very generic to suit many different cases. In order to represent rich metadata, the EDM can be specialised for specific domains as done by the Digitised Manuscripts to Europeana (DM2E) project for the domain of handwritten manuscripts with the DM2E model.

Before creating the DM2E model, decisions on a general modelling approach had to be made including the method of reusing external resources (Dröge, Iwanowa et al., 2013), decisions on the granularity of the specialisation and instruments of documentation. Model-related research questions are: What is the best way for creating a shared ontology for representing manuscripts in a digital library and how can diverse ontology requirements be combined without leading to a model which is too general? The first step in the model creation was to analyse the metadata about manuscripts coming from different data providers and in different formats like TEI, METS/MODS, MARC21 or provider-specific schemas. Furthermore, it was investigated if the data meets the mandatory requirements of the EDM. Additional properties, classes, resource definitions, restrictions and recommendations were added to the EDM which resulted in the DM2E model. The first operational version of the model was created in April 2013 and since then iteratively refined. New functions of the model include the representation of uncertain timespans and hierarchical objects.

The DM2E model will be discussed in its current representation. First mappings from provided data to the model will be analysed. Data mapped to the DM2E model is dereferenceable and will not only be delivered to Europeana but also be available via a LOD access point (Heath & Bizer, 2011).

Keywords: DM2E Model, Europeana Data Model, Linked Data, Ontology Development, Digital Libraries.

Introduction

Europeana[2] is the European digital library which provides a unified access to the cultural heritage of Europe. More than 30 million library, archive, museum and audio-visual objects from 36 countries are represented in Europeana[3]. These

[1] Corresponding author.

[2] Europeana website: http://www.europeana.eu/ [30.03.2014].

[3] Numbers as of November 2013. Europeana Professional website: http://www.pro.europeana.eu/web/guest/content [24.03.2014].

objects are delivered to Europeana by content providers via national aggregators like the German digital library (*Deutsche Digitale Bibliothek*, DDB)[4] or domain aggregators like the *Digitised Manuscripts to Europeana* project (DM2E)[5]. One of the major challenges for Europeana lays in finding a way to integrate the heterogeneity of objects provided and the metadata schemas describing these diverse objects. The current model used by Europeana to represent the provided data is the *Europeana Data Model* (EDM). This model was specialised by DM2E for the domain of manuscripts in order to enable rich mappings of the provided data. The specialisation, called the *DM2E model*, will be presented in the scope of this paper.

The paper is structured as follows: First, the data models currently used in Europeana, EDM and ESE, are presented. This section is followed by specialisations of the EDM in general and the detailed description of the DM2E model as a specialisation of the EDM for manuscripts in particular including the modelling approach, the reuse strategy and detailed insights on the build-up. The paper concludes with a first insight in the evaluation of the DM2E model and a brief look on future work.

Data representation in Europeana

The first principle solution in finding a way to integrate the diverse objects into Europeana was the creation of a common and simple schema, the *Europeana Semantic Elements* (ESE). The ESE represents the lowest common denominator in terms of semantics found in various metadata schemas which are used for the description of cultural heritage objects (ESE Specification, 2013). The schema provides a simple and flat representation for cultural heritage objects (often abbreviated as CHOs) based on the Dublin Core Elements Set[6]. As all data providers contributing to Europeana had to convert their metadata into this common schema, the previously existing interoperability problem was initially solved.

Although this approach worked well, there were also some serious drawbacks. Most importantly, the model was not easily extensible and did not provide sufficient semantics for describing many important details from the various metadata schemas, including the proper modelling of hierarchical or complex objects. Furthermore, since the ESE is based on XML, there is no easy way of linking objects to other objects or to other terminological sources.

The EDM has been developed as the successor of the ESE and as a response of its aforementioned shortcomings (Hennicke, Dröge et al., 2014). The EDM is similarly to the ESE a generic representation of the semantics in the cultural heritage domain (EDM Primer, 2013). However, it uses a different approach to data modelling and is much more expressive and flexible in terms of integration with other knowledge sources and semantic extensions.

The *Resource Description Framework* (RDF)[7] is the representation language of the EDM. Information is no longer conceptualised in a tree-based way with attributes and literal values but in a graph structure with mostly explicit entities connected through meaningful relations. In this graph structure, information is broken down into statements in the form of triples which consist of a subject, the entity the statement is about, a predicate, the property connecting two entities, and the object, the value of the statement. An element in the triple may represent any imaginable entity which includes not only things on the Web, like websites, images or files, but also things outside the Web, like people, buildings and books, or even abstract concepts, like eras, ideas or terms. Subjects and predicates in triples must be resources; objects can be resources or literals. A resource is identified by a *Uniform Resource Identifier* (URI)[8] which is unique. This allows to connect and to integrate distributed information rather easily. The *Resource Description Framework Schema* (RDFS)[9] is used to define the actual ontology schema consisting of classes and properties.

The central classes of the EDM (see figure 1) are *edm:ProvidedCHO*, the class for the described cultural heritage object, *ore:Aggreation*, the class representing the metadata record provided for the described object and *edm:WebResource* which includes views of the described object like a thumbnail. Additional classes like *edm:Agent*, *edm:TimeSpan*, *edm:Place* or *skos:Concept* allow to represent contextual resources related to the described object. The properties provided by the EDM allow to describe how these things relate to each other, for example, by relating a book to a title with the property *dc:title* or to its creator with the property *dc:creator*.

[4] DDB website: https://www.deutsche-digitale-bibliothek.de/ [11.04.2014].

[5] DM2E website: http://www.dm2e.eu [11.04.2014].

[6] Dublin Core Elements Set: http://dublincore.org/documents/dces/ [30.03.2014].

[7] RDF 1.1 Primer. W3C working group note: http://www.w3.org/TR/rdf11-primer/ [30.03.2014].

[8] Uniform Resource Identifiers (URI): Generic Syntax: http://www.ietf.org/rfc/rfc2396.txt [30.03.2014].

[9] RDF Schema 1.1. W3C recommendation: http://www.w3.org/TR/rdf-schema/ [30.03.2014].

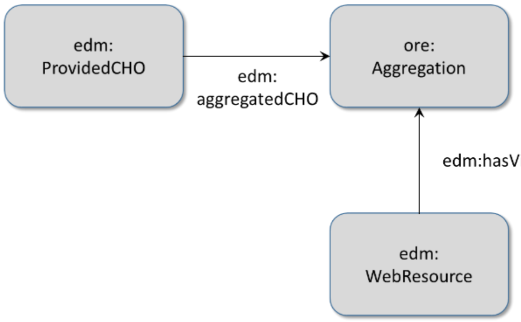

Figure 1: The EDM main classes: *edm:ProvidedCHO*, *ore:Aggregation* and *edm:WebResource*.

The EDM extensively reuses external vocabularies[10] like Dublin Core (elements and terms), OAI-ORE, SKOS and FOAF. Properties build the largest part of the model and are used to give detailed descriptions of the objects like its creator, contributors, a title or a description or of the metadata. A list and description of all elements in the EDM can be found in the latest Definition of the Europeana Data Model (2013).

Specialising the Europeana Data Model

An important feature of the EDM is the possibility to create specialisations (extensions and refinements) of the model. Specialisations are created by many projects in the context of Europeana, like the EDM refinement of *Europeana Libraries*[11] and the EDM extensions of *Europeana Creative*[12] or the specialisation of the DM2E-project, the DM2E model (Charles & Olensky, 2014). Not only Europeana projects but also institutions outside of the Europeana network take the EDM as a base for their data representations and specialise it. Two examples here are again the DDB and the *Digital Public Library of America* (DPLA)[13]. The DDB uses the EDM as base for facetted search in their portal and to deliver data to Europeana. The *DPLA metadata profile* (DPLA MAP) reuses diverse EDM resources but also resources from other vocabularies. The central build-up of the metadata profile resembles the one of the EDM: the core classes in the DPLA MAP are *ore:Aggregation*, *edm:WebResource* and *dpla:SourceResource*, which is a subclass of *edm:ProvidedCHO* (Digital Public Library of America Metadata Application Profile, 2013).

Specialisations, especially in form of refinements, allow the mapping of more specific semantics from other metadata schemas to EDM. The EDM represents the common generic layer of semantics through which all data is connected and the specialisations represent a semantically more expressive layer which allows the representation of more detailed information to the user, granular search and retrieval operations and more opportunities for external applications that may build on EDM data.

As the EDM unions various different vocabularies, an extension of the model has to consider how to handle the reuse of external vocabularies as well. Reusing can be done in different ways. Four possibilities were identified in Dröge, Iwanowa et al., 2013:

1. Direct adoption of external resources with their original URI in the current ontology.
2. Integration of external resources where URIs are adjusted to the namespace of the current ontology.
3. Indirect adoption of external resource with their original URI as a specialised subclass or subproperty of resources in the current ontology.
4. Direct adoption of external resources into the current namespace and pointing to the original resources via *owl:equivalentClass* or *owl:equivalentProperty*.

Currently, all four ways can be found in existing vocabularies. The EDM reuses external resources by directly mixing and matching them (option 1) and provides additional definitions and mapping instructions for reused elements. During the initial modelling process in DM2E the third option was chosen. This seemed to be a cleaner way of reusing resources but led to an unnecessary complex model: not only demands this option the creation of many unwanted DM2E resources but also does the EDM part of the model have a different structure than the DM2E specialisations. In order to have a homogenous model, the approach was changed and the DM2E model now follows a similar reuse method as the EDM. A small difference is that a new property *dm2e:scopeNote* was introduced in the DM2E model to give detailed explanations for the usage of classes and resources in the scope of DM2E instead of reusing existing annotation properties for that purpose. This approach was chosen to avoid the multiple usage of popular properties like *skos:note* which may lead to conflicting descriptions of the same resource (real-world examples for conflicting descriptions and labels can be found in Dröge, Iwanowa et al., 2013).

Specialisations of the EDM are in cases of refinements also called application profiles (Charles & Olensky, 2014). An application profile mixes and matches existing resources from one or more namespaces for a specific local application (Heery & Patel, 2000). This includes the reuse practice. The goal of the recently started RDF Application Profile working

[10] Links to the specifications of the vocabularies can be found in table 2, section "Description of the DM2E model".

[11] Europeana libraries website: http://www.europeana-libraries.eu/web/ [11.04.2014].

[12] Europeana creative website: http://www.europeanacreative.eu/ [11.04.2014].

[13] DPLA website: http://dp.la/ [11.04.2014].

group[14] is to establish definitions and creation principles for RDF Application Profiles including best practices for publishing them as Linked Data. Specialisations of EDM like the DM2E model and the DPLA MAP are presented as use cases and will profit from the group's results.

Modelling approach in DM2E

The DM2E model is a specialisation of the EDM for the domain of manuscripts. The DM2E understanding of the term manuscript is very broad and therefore, the model covers the representation of medieval handwritten manuscripts but also typed books, like Ludwig Wittgenstein's *Brown Book*[15], or journals, like the 18th to 19th century *Polytechnische Journal*[16]. The model has been developed bottom-up based on the needs of the project's data providers. The first step of the specialisation process was to identify and analyse the requirements of the content providers. Simultaneously, the concordance between these requirements and the mandatory EDM elements was discovered. The EDM has only few mandatory elements but these are needed in order to provide a minimal representation of a cultural heritage object in Europeana. Mandatory elements are a Web representation of the object, metadata rights, the data provider and the aggregator, a type, subject, temporal or spatial characteristics of the provided object, a title or description of the object and the language in case of textual objects (EDM Mapping Guidelines, 2013). In order to check if the minimal requirements are fulfilled, the data providers in DM2E delivered sample data that was intellectually analysed. These datasets included metadata and object data about medieval manuscripts, manuscripts from philosophers, letters, journals and books including drawings. The sample data was represented in a large variety of metadata formats. Two surveys on the provided data were answered by the data providers and the metadata formats were collected and described in the project's Wiki. As it turned out, almost all content providers already worked with standardised metadata formats, like the interlibrary exchange formats MAB2[17] and MARC21[18], the archival standard format EAD[19], the full text encoding format TEI[20] and the METS/MODS[21] format for descriptive, administrative and structural metadata. Provider-specific formats based on individual database schemas did also occur and were processed with the D2R tool (Bizer & Cyganiak, 2006) to get RDF data. One of the main challenges the project is facing is to map these diverse datasets into a unified model without losing the richness and depth of the original metadata in order to enable rich functionalities on top of the data.

In addition to the surveys, the provided data was analysed based on intellectual mappings to the EDM. This has been done during mapping workshops attended by both, data providers and EDM experts. The aim of the preliminary mappings was to collect missing classes and properties that are needed in a later specialisation of the EDM and to check the completeness of the data regarding the EDM requirements. The results of the surveys and the mapping workshops have clearly shown that the current version of the EDM is, on the one hand, in principle able to accommodate all provided sample data but, on the other hand, has to be specialised in order to retain most of the provided information of the source data. One of the goals of the project was to enable mappings representing the original semantics of the provided metadata as closely as possible. This is important as the provided data is not only needed to display objects on Europeana but to create and provide rich Linked Data resources.

Figure 2 shows an excerpt of an exemplarily intellectual mapping from a metadata record provided by the Max Planck Institute for the History of Science which was created with the Visual Understanding Environment (VUE) by the Tufts University. Circles represent resources, boxes represent literals. Unmarked properties are part of the EDM but not mandatory (e.g. *dc:publisher*) whereas properties marked with plus are required by the EDM (e.g. *dc:title*). Properties marked with asterisk are needed in addition to EDM properties in order to provide clear and specialised mappings (e.g. *bibo:numPages*). To retain the backwards compatibility to EDM, requested extensions have been added whenever possible as subproperties or subclasses of existing EDM elements (see e.g. *dm2e:callNumber* as a proposed subproperty of *dc:identifier*).

[14] Wiki of the RDF Application Profiles working group: http://wiki.dublincore.org/index.php/RDF-Application-Profiles [11.04.2014].

[15] Wittgenstein's Brown Book is only at Wittgenstein source: http://www.wittgensteinsource.org/ [14.05.2014].

[16] Polytechnisches Journal website: http://www.polytechnischesjournal.de/ [14.05.2014].

[17] Specification of MAB2: http://www.ubka.uni-karlsruhe.de/hylib/mab/mab2.html [14.04.2014].

[18] Specification of MARC 21:http://www.loc.gov/ standards /marcxml [14.04.2014].

[19] Definition of the EAD Schema: http://www.loc.gov/ead/eadschema.html [14.04.2014].

[20] Definition of the TEI guidelines: http://www.tei-c.org/Guidelines/P5/ [14.04.2014].

[21] Specification of METS/MODS: http://www.loc.gov/standards/mods/ [14.04.2014].

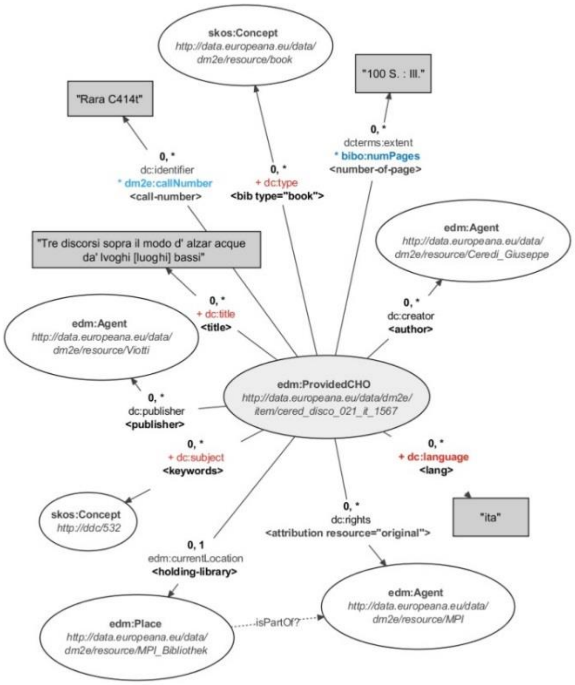

Figure 2: Excerpt of a conceptual mapping based on a record from the Max Planck Institute for the History of Science produced to analyse the provided data and to identify which specialisations of the EDM are needed.

Sample mappings as shown in figure 2 have been created for all provided datasets. Based on these mappings and the results of the metadata questionnaires, the first specialisation of the EDM has been drafted. To capture the full wealth of semantics in the provided sample data, subclasses were included that extend the EDM classes e.g. for indicating the type of the described objects such as book, journal or page. In the same way, the relationship between the provided CHO, contextual resources and Web representations have been defined in a more specific way. Linked Open Data repositories like LOV[22] and DataHub[23] were used to search for resources that could be reused in DM2E. However, not all resources that can possibly be used in the DM2E model were yet found as many vocabularies and many different search possibilities exist.

Reused and new resources are described via *dm2e:scopeNote* which holds a description for the use of external properties or classes in the context of DM2E. For example, it was decided to add new classes for book, cover and page as subclasses of *edm:PhysicalThing*. It was discovered that equivalents for book and cover could be reused from FaBiO, the FRBR-aligned Bibliographic Ontology, and Bibo, the Bibliographic Ontology, but not for page. In the next step, the classes *bibo:Book* and *fabio:Cover* were added as subclasses of *edm:PhysicalThing* and described in the DM2E context via the *dm2e:scopeNote* annotation property if an extra description is needed (see table 1). As the original definition is still valid, it is also shown in the DM2E model specification (Dröge, Iwanowa et al., 2014). Resources that are in the DM2E namespace, like the new subclass *dm2e:Page*, are also described via *dm2e:scopeNote* without having another original description

Table 1. Example of specialising classes in the DM2E model which are described with *dm2e:scopeNote* if needed and the original description if they are reused.

Class	DM2E scope note	Original definition
bibo: Book	cf. original scope note	A written or printed work of fiction or nonfiction, usually on sheets of paper fastened or bound together within covers.
fabio: Cover	ProvidedCHO of type cover. Can be part of another CHO, e.g. a book.	A protective covering used to bind together the pages of a document or the first, informative, page of a digital document.
dm2e: Page	ProvidedCHO of type page. A sheet of paper. Can be part of another CHO, e.g. dm2e: Manuscript.	-

After the first mappings of the data, the DM2E model was continuously refined based on the provider's feedback. The current version of the model, DM2E model 1.1, serves as base for the final content integration and includes most of the collected data provider's requirements as well as the requirements of the transformation, annotation and search components of the DM2E infrastructure.

The DM2E model schema

The DM2E model makes use of different namespaces for the schema, i.e. classes and properties, and the provided data, i.e. instances, to make a clear difference between the provided metadata and the way it is represented. The schema namespace is http://onto.dm2e.eu/schemas/dm2e/. Instances are stored in the data namespace http://data.dm2e.eu/data/. In former model versions, the schema namespace was versioned. The information about the model version and

[22] Linked Open Vocabularies (LOV) website: http://lov.okfn.org/dataset/lov/ [16.04.2014].

[23] The DataHub website: http://datahub.io/ [16.04.2014].

revision used for a mapping is provided during the data ingestion in the DM2E triple store and no longer part of the namespace URI since version 1.0.

The DM2E model specialises the EDM mainly via subclasses and subproperties of existing EDM classes and properties, e.g. by adding the subproperty *pro:author* to the existing property *dc:creator*. However, the model also offers a few additional options that are not specialising EDM resources. This is the case when the DM2E model covers functions that are not offered by, e.g. the property *dm2e:hasAnnotatableContent* which points to an annotatable object that is needed for the semantic annotation tool Pundit[24] (Grassi, Morbidoni et al., 2013). As opposed to the EDM, the DM2E model makes use of named graphs instead of proxies for data provenance.

Overview

The main motivations in the DM2E project are not only to deliver data to Europeana but to create *Linked Open Data* (LOD) and to build new LOD-based tools. Linked Data is described by Berners-Lee in the Linked Data Design Issues as data that is made available on the Web, that can be accessed by human users and tools, is linked to other data and dereferencable via stable identifiers (Berners-Lee, 2006; Heath & Bizer, 2011). Ideally, LOD is represented in RDF.

The third Linked Data principle, linking to other data, is fulfilled by reusing resources. Resources in DM2E originate from diverse vocabularies, like Dublin Core, Bibo, FaBiO or the OAI-ORE specification. External vocabularies, from which resources were reused, are listed in Table 2. A large amount was already used in EDM.

Table 2. External vocabularies that are reused in the DM2E model in alphabetical order of the vocabulary prefixes.

Prefix	**Namespace**
bibo	http://purl.org/ontology/bibo/
crm	http://www.cidoc-crm.org/cidoc-crm/
dc	http://purl.org/dc/elements/1.1/
dcterms	http://purl.org/dc/terms/
edm	http://www.europeana.eu/schemas/edm/
fabio	http://purl.org/spar/fabio/
foaf	http://xmlns.com/foaf/0.1/
ore	http://www.openarchives.org/ore/terms/
pro	http://purl.org/spar/pro/
rdaGr2	http://rdvocab.info/ElementsGr2/
skos	http://www.w3.org/2004/02/skos/core#
vivo	http://vivoweb.org/ontology/core#
void	http://rdfs.org/ns/void#
wgs84_pos	http://www.w3.org/2003/01/geo/wgs84_pos#

Altogether, 103 new resources, 79 properties and 24 classes, were introduced in the DM2E model (see Table 3). The numbers of resources that are in the DM2E namespace indicate that there may still be resources left that could already be described by another vocabulary and reused. During the modelling process, it was decided, to integrate all properties and classes of the data providers that were needed to represent their objects, even if they are on different levels of granularity. If a later evaluation of the model identifies many unused resources some of them will be excluded from the model to reduce its complexity.

Table 3. Number of new resources in the DM2E model. The numbers on the left side of the slashes are resources in the DM2E namespace whereas resources on the right side were reused.

Class	**New Properties DM2E/Other**	**New Classes DM2E/Other**
ore: Aggregation	2 /5	-
edm:ProvidedCHO	39 /19	-
edm:PhysicalThing		4 /5
edm:Agent	0 /2	0 /2
foaf:Person	2 /0	-
foaf: Organization	-	1 /3
edm:NonInformationResource	-	2 /1
skos:Concept	0 /2	3 /3
edm:Place	0 /1	-
edm:TimeSpan	0 /2	-

Most of the new properties, 58 out of 79, are added to the class *edm:ProvidedCHO*. Other properties were added to the classes *ore:Aggregation*, *edm:Agent* and *foaf:Person*[25],

[24] Pundit website: https://thepund.it/ [20.04.2014].

[25] *foaf:Person* and *foaf:Organization* are not part of the EDM but of the DM2E model. However, the EDM offers properties that should only be used with either persons or organisations. In the DM2E model, these properties have the domain of the new subclasses *foaf:Person* and *foaf:Organization*.

skos:Concept, *edm:Place* and *edm:Timespan*. The most broadly specialised classes are *edm:PhysicalThing* and *skos:Concept* for the further description of CHOs.

Like the EDM, the DM2E model includes some mandatory elements that are the minimal requirement for a valid mapping. Mandatory elements are needed on the one hand to fulfil the requirements of the EDM and thus to produce a valid EDM mapping and on the other hand to meet the requirements that tools offering further functionalities based on DM2E data have, like search and browse or text annotation functions. Additional mandatory properties in the DM2E model are *dm2e:displayLevel*, *dc:type* that points to a subclass of *edm:PhysicalThing* or *skos:Concept*, *dc:format* for annotatable resources and *skos:prefLabel* for *edm:Agent*, *skos:Concept*, *edm:Place*, *edm:TimeSpan* and *edm:Event*. The provided metadata is very diverse, so mandatory elements were not often used. Resources that increase the quality of a mapping a lot were marked as "highly recommended" instead.

Upper level of the model

Both, the EDM as well as the DM2E model, are used by different providers that may describe the same resources (e.g. the same CHO or the same creator). In order to allow several statements about the same resource, which can even be contrary, the EDM has introduced the class *ore:Proxy*. *ore:Proxy* is used to make statements on the provided content. The DM2E model also aims at providing this possibility, but has chosen another way to do that. By introducing *Named Graphs* (Carroll, Bizer et al., 2005), in which a fourth position is added to a statement, an RDF triple can be further described. By making a quadruple out of a triple, one can gather triples and make additional statements about them. Named Graphs allow us thus to make statements about statements or descriptions. RDF graphs created from the input data of a provided collection are identified by an URI and belong to the class *void:Dataset*. They are not mapped by the provider but automatically added in the data ingestion process.

Not only the provided data but the whole DM2E infrastructure is based on Linked Data principles.

> "Linked Data is the paradigm that drives the whole DM2E infrastructure. The DM2E model reflects this by explicitly defining classes for datasets and published data resources. This way, the meta-level of resource descriptions becomes a first-class member of the data model and can be used for annotations and provenance tracking."

(Dröge, Iwanowa et al., 2014: 12)

When a provider ingests data into the project's triple store, additional RDF is produced in this process. Interactions of providers inside the ingestion platform, like uploading files or creating workflows, are also represented in RDF but described with additional vocabularies and not with the DM2E model (Eckert, Ritze et al., 2014)

Specialised classes and properties

Properties and classes, which are represented in more detail in the sample data, as well as properties and classes from which was assumed that they should have a more detailed representation in the manuscript domain, have been specialised. Most of the introduced specialisations are used in a similar way in well-established standards like those provided to the project. An example: the DM2E model introduces the property *dm2e:incipit* which is used for representing the opening words of a manuscript. This property is similar to the MAB field 661 and a mapping can easily be made between both representations.

Some properties and classes are specialised in more detail than others. Properties with many new subproperties are e.g. *dc:creator*, for different types of creators of the CHO (see figure 3), and the very general property *edm:hasMet* which points to agents, events, places, timespans or concepts.

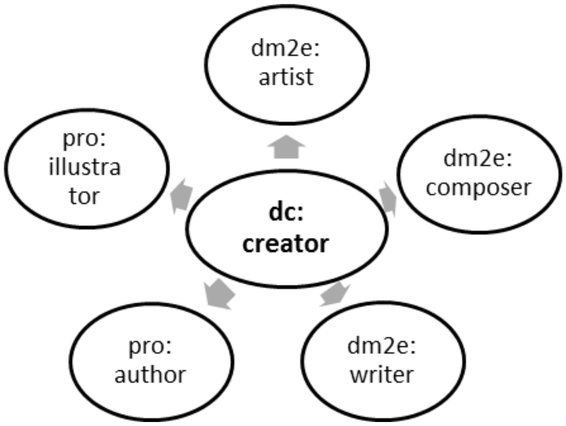

Figure 3: Specialised properties for *dc:creator*.

Classes in the DM2E model are used to further distinguish the type of physical objects and of agents. Every CHO in DM2E must indicate the object's type via *dc:type*. The property points in the scope of DM2E to a subclass of *edm:PhysicalThing* or *skos:Concept*. Physical things are objects or parts of an object like a book, a manuscript or a page. Concepts are conceptual units of an object, like a chapter or a paragraph. These resources are needed to compare or distinguish CHOs. The example in figure 4 shows extended classes that are defined as subclasses of *edm:PhysicalThing*. The dark grey ovals illustrate the new classes which were added to the model. The figure shows where they are semantically meaningful integrated based on the DM2E-specific needs. The same way the *edm:Agent* class was further specialised via subclasses in order to enable the distinction between persons or organisations as agents.

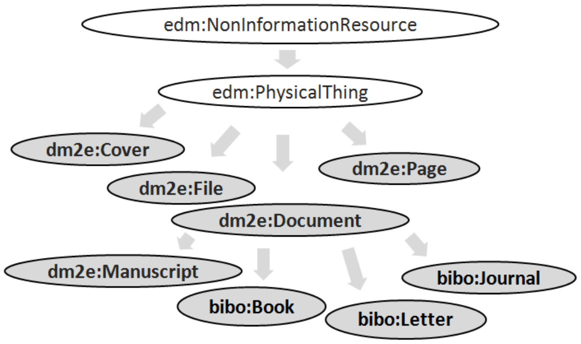

Figure 4: Subclasses of *edm:PhysicalThing* in the DM2E model.

In summary, many new resources were needed to describe agents (especially persons) more detailed, e.g. agents who created or contributed to the CHO or who are mentioned or are related to the CHO or its aggregation in a different way than those supported by the EDM. As it was the aim to have a specialisation for the EDM that has a similar build-up, the DM2E model mainly makes use of properties instead of classes to further describe resources.

Properties for additional functionalities

Some properties that are introduced in the DM2E model are not added as subproperties to existing EDM properties. These properties were primarily needed for additional technical functionalities that the EDM does not cover and are less content-related. Examples are properties for search functionalities for hierarchical objects, the property *dm2e:scopeNote* for additional comments on a resource and the property *dm2e:hasAnnotatableContent* that is needed for annotations with Pundit. Not only data providers but also developers responsible for search functionalities or for the annotation tools suggested additions to the DM2E model. Additional search functionalities in the project were needed as Europeana cannot display the granularity of the mapped objects on small levels like pages. Furthermore, it is not yet possible to display extensions of the EDM in the portal. Therefore, it was decided to provide additional search functionalities next to Europeana for not only providing Linked Data via a SPARQL endpoint which is mainly for developers but to make the data easy accessible and browsable for the casual user. For having an entry point to the data and to not overstrain users with several thousand pages of the same author in a search result list, *dm2e:displayLevel* was introduced. The property enables a selective view for hierarchical objects in the search and browse interface. Only CHOs marked with *true* are displayed as a browsing entry point into the whole collection. This leads to higher performance of the search engine and to better usability for the end user. The property *dm2e:hasAnnotatableContent* requires a specific type of CHO representation that can be annotated with Pundit. Annotatable content can be a specific type of image, like PNG or JPEG, or text. Whenever the property is used, the type of content should be indicated by using one of the permitted mime-types. In order to represent uncertainty in time spans, the properties *crm:P79F.beginning_is_qualified_by* and *crm:P80F.end_is_qualified_by* from CIDOC-CRM were reused. The values “uncertainty_data” or “uncertainty_granularity” can be added with these properties to indicate whether a time span was estimated or the exact limitations of a time span are unknown.

Model documentation

In order to make it easier for others to reuse the DM2E model, it was important to properly document the model. During the iterative development phase and specialisation process, the documentation of all intermediate versions was updated continuously. The documentation of the DM2E model is currently available in three different formats. The textual description of the model helps providers for their mappings. It can be found on Europeana Pro[26], in the project’s Wiki[27] or on the DM2E website. Individual classes and properties defined in the DM2E namespace are made accessible through the vocabulary publishing platform Neologism[28] via the schema namespace of the model and the individual class and property URIs. The full model including reused resources can be seen and downloaded as an OWL file via an account on GitHub[29]. Specific recommendations for the representation of DM2E metadata to support content providers in creating concrete RDF representations of metadata mapped to the DM2E model have been also published. The recommendations include specific guidelines for encoding certain aspects of the data such as time information, URI design or representation of subject terms and hierarchies.

Evaluation of the DM2E model

An evaluation of the DM2E model based on a mapping analysis has recently started in order to reduce the models complexity and to make it less detailed where the current level of granularity is not needed. A first step in the evaluation was to figure how often classes and properties are used in the provider mappings. Although the evaluation is still ongoing, it could already be seen that there are classes and properties in the DM2E model that are never used. These are not only new properties or classes introduced by DM2E but also resources defined in EDM.

Nine datasets from seven different data providers including 61 million RDF statements were examined. Only

[26] Europeana Pro website: http://pro.europeana.eu/ [04.05.2014].
[27] DM2E Wiki: http://wiki.dm2e.eu/Main_Page [16.04.2014].
[28] Neologism website: http://neologism.deri.ie/ [04.05.2014].
[29] OWL-files of the DM2E model on GitHub: https://github.com/DM2E/dm2e-ontologies/tree/master/src/main/resources/dm2e-model [16.04.2014].

about half of the classes that the DM2E model offers were used during the mappings in at least one dataset. Classes that were often used are the core classes *edm:WebResource*, *edm:ProvidedCHO* and *ore:Aggregation*, *dm2e:Page* and *skos:Concept*. Not used are subclasses of *foaf:Organization*, *edm:Event* and some specific CHO types like *dm2e:Document*, *dm2e:File* or *fabio:Chapter*. The properties and classes that were used to describe individuals as well as the way they are represented vary between datasets. There are a lot of differences in the mapped datasets which have to be further analysed. The analysis of the properties showed that about a third of them are not mapped in any dataset. A consequence for the DM2E model is that the unused resources will be removed from the model if it can be assumed that they will also not be used for other mappings in the manuscript domain. This will hopefully reduce the complexity of the model without prohibiting the providers from creating rich mappings. Further analyses based on the mappings are ongoing work.

Outlook and conclusion

The DM2E model has been built as a specialisation of the EDM in order to represent rich manuscript metadata on Europeana and to be published as Linked Open Data. The build-up approach was bottom up. Whenever feasible, external resources were reused. Provider feedback, mappings to the model as well as a first evaluation of the model based on the mappings have shown that the DM2E model covers the provided manuscript metadata sufficiently. Nevertheless, the model can still be improved: the evaluation has shown that many resources, classes as well as properties, were not used in the mappings. Additionally, data represented by the archival format EAD was not yet analysed and included into the model. Thus, the focus of further developments of the model will mainly lie on extending it regarding the requirements for the EAD format, on further linking to other vocabularies, on reducing the model's complexity by removing unused resources and on improving the model to meet potentially additionally upcoming results of the evaluation.

ACKNOWLEDGMENTS

This work was supported by a grant from the European Commission "ICT Policy Support Programmes" provided for the DM2E project (GA no. 297274). The authors would like to thank all colleagues of DM2E as well as the European Commission for their support.

REFERENCES

Berners-Lee, T. (2006). Linked Data - Design Issues. W3C Website. Retrieved from: W3C website. http://www.w3.org/DesignIssues/LinkedData (14.05.2014).

Bizer C., & Cyganiak R. (2006). D2R server – Publishing Relational Databases on the Semantic Web. Poster at the 5th International Semantic Web Conference, Athens, GA, USA.

Carroll, J., Bizer, C., Hayes, P., & Stickler, P. (2005). Named Graphs. In: Journal of Web Semantics, 3 (2005), 247-267.

Charles, V., & Olensky, M. (2014). Report on Task force on EDM mappings, refinements and extensions. Retrieved from: Europeana Professional website. http://pro.europeana.eu/documents/468623/bca65b72-fb8f-4b4f-802d-1072690ae33a (11.04.2014).

Definition of the Europeana Data Model, v5.2.4 (2013). Retrieved from: Europeana Professional website. http://pro.europeana.eu:9580/documents/900548/0d0f6ec3-1905-4c4f-96c8-1d817c03123c (16.04.2014).

Digital Public Library of America Metadata Application Profile, Version 3 (2013). Retrieved from: DPLA website. http://dp.la/info/wp-content/uploads/2013/04/DPLAMetadataApplicationProfileV3.pdf (11.04.2014).

Dröge, E., Iwanowa, J., Hennicke, S., & Eckert, K. (2014). DM2E Model V1.1 Specification. Europeana Professional website. http://pro.europeana.eu/documents/1044284/0/DM2E+Model+V+1.1+Specification (25.03.2014).

Dröge, E., Iwanowa, J., Trkulja, V., Hennicke, S., & Gradmann, S. (2013). Wege zur Integration von Ontologien am Beispiel einer Spezifizierung des Europeana Data Model. In H.-C. Hobohm (Ed.): Informationswissenschaft zwischen virtueller Infrastruktur und materiellen Lebenswelten. Proceedings des 13. Internationalen Symposiums für Informationswissenschaft, pp. 273-284. Glückstadt: VWH.

Eckert, K., Ritze, D., Baierer, K., & Bizer, C. (2014). RESTful open workflows for data provenance and reuse. In proceedings of the companion publication of the 23rd international conference on World wide web companion, pp. 259-260.

Europeana Data Model Primer, v14/07/2013 (2013). Retrieved from: Europeana Professional website. http://pro.europeana.eu/documents/900548/770bdb58-c60e-4beb-a687-874639312ba5 (04.05.2014).

Europeana Data Model Mapping Guidelines, v2.0 (2013). Retrieved from: Europeana Professional website. http://pro.europeana.eu:9580/documents/900548/60777b88-35ed-4bae-8248-19c3696b81fb (11.04.2014).

Europeana Semantic Elements Specification and Guidelines v17/07/2013 (2013). Retrieved from: Europeana Professional website. http://pro.europeana.eu/documents/900548/2eee7beb-b9d8-4532-a089-8e8d6df38ce7 (04.05.2014).

Grassi, M., Morbidoni, C., Nucci, M., Fonda, S., & Piazza, F. (2013). Pundit: augmenting web contents with semantics. In: Literary and Linguistics Computing, 28(4), 640 – 659.

Heath, T., & Bizer, C. (2011). Linked Data: Evolving the Web into a Global Data Space. Synthesis Lectures on the Semantic Web: Theory and Technology (Vol. 1). Morgan & Claypool.

Heery, R., & Patel, M. (2000). Application Profiles: mixing and matching metadata schemas. In: Ariadne 25(2000). Retrieved from: Ariadne website. http://ariadne.ac.uk/issue25/app-profiles (10.04.2014).

Hennicke, S., Dröge, E., Trkulja, V., & Iwanowa, J. (2014). From ESE to EDM and Beyond: How Europeana Provides Access to its Cultural Heritage Objects. In M. Ockenfeld (Ed.): Informationsqualität und Wissensgenerierung. Proceedings der 3. DGI-Konferenz, 66. Jahrestagung der DGI, pp. 129-140. Frankfurt am Main: DGI.

Curriculum Vitae

Evelyn Dröge works as a research assistant at the Berlin School of Library and Information Science (IBI) at Humboldt-Universität zu Berlin. She has studied information science and language technology at the Heinrich-Heine-University Düsseldorf and is currently doing her PhD in library and information science which focuses on the evaluation of ontology matching tools. She works in the Digitised Manuscripts to Europeana (DM2E) project and is responsible for the DM2E model.

Julia Iwanowa is a research assistant at the Berlin School of Library and Information Science (IBI) at Humboldt-Universität zu Berlin. She is currently working for Digitised Manuscripts to Europeana (DM2E) where she is responsible for the DM2E model and for mappings from TEI to DM2E. Julia has studied Applied Computer Science in the Humanities, German and Slavonic at the University of Cologne.

Steffen Hennicke is a research assistant at the Berlin School of Library and Information Science (IBI) at Humboldt-Universität zu Berlin. He studied history, political science, and media science at the University of Potsdam, Sussex University (UK), and the Free University of Berlin and received his Magister Artium (MA) in 2007. Steffen has been involved in EuropeanaConnect and is currently working for Digitised Manuscripts to Europeana (DM2E).