Semantic Web 0 (2014) 1
IOS Press

DM2E: A Linked Data Source of Digitised Manuscripts for the Digital Humanities

Editor(s): Christoph Schlieder, Universität Bamberg, Germany
Solicited review(s): Günther Görz, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany; Bernhard Haslhofer, Austrian Institute of Technology, Austria; Eetu Mäkelä, Aalto University, Finland & University of Helsinki, Finland & University of Oxford, UK

Konstantin Baierer [a], Evelyn Dröge [a], Kai Eckert [b,*], Doron Goldfarb [c], Julia Iwanowa [a], Christian Morbidoni [d], Dominique Ritze [e]

[a] *Berlin School of Library and Information Science, HU Berlin, Unter den Linden 6, 10099 Berlin, Germany*
E-mail: first.last@ibi.hu-berlin.de
[b] *Stuttgart Media University, Nobelstr. 10, 70569 Stuttgart, Germany*
E-mail: eckert@hdm-stuttgart.de
[c] *Austrian National Library, Josefsplatz 1, 1015 Vienna, Austria*
E-mail: fue@onb.ac.at
[d] *Dept. of Information Engineering, Università Politecnica delle Marche, Via Brecce Bianche, 60131 Ancona, Italy*
E-mail: christian.morbidoni@gmail.com
[e] *Research Group Data and Web Science, University of Mannheim, B6, 26, 68159 Mannheim, Germany*
E-mail: dominique@informatik.uni-mannheim.de

Abstract. The DM2E dataset is a five-star dataset providing metadata and links for direct access to digitized content from various cultural heritage institutions across Europe. The data model is a true specialization of the Europeana Data Model and reflects specific requirements from the domain of manuscripts and old prints, as well as from developers who want to create applications on top of the data. One such application is a scholarly research platform for the Digital Humanities that was created as part of the DM2E project and can be seen as a reference implementation. The Linked Data API was developed with versioning and provenance from the beginning, leading to new theoretical and practical insights.

Keywords: Linked Data, Dataset, Cultural Heritage, Digital Humanities, Digital Content, Europeana, EDM, DM2E

1. Introduction

The project "Digitised Manuscripts to Europeana" (DM2E)[1] was active from 02/2012 until 01/2015, funded under EU FP7. Its two primary goals were:

1. The transformation of various metadata and content formats describing and representing digital cultural heritage objects (CHOs) in the realm of digitized manuscripts from as many providers (cf. Section 2) as possible into the Europeana Data Model (EDM) to get it into Europeana, the European digital library.[2]
2. The stable provision of the data as Linked Data and the creation of tools and services to reuse the data in the Digital Humanities. The basis is the possibility to annotate the data, to link the data, and to share the results as new data.

*Corresponding author, e-mail: eckert@hdm-stuttgart.de
[1] `http://dm2e.eu/` (10.12.2015)
[2] `http://www.europeana.eu` (10.12.2015)

Table 1
DM2E data sources

Provider	Collection	ML[a]	CL[b]	Type[c]	Count	Vocab[d]	Format
Berlin Brandenburg Academy of Sciences	Deutsches Textarchiv	de	de	B	1547	1	TEI
University of Bergen	Wittgenstein Archive Bergen	en	de,en,var.	M	20		
Bulgarian Academy of Sciences	Codex Suprasliensis	en	cu	M	49	2	
Humboldt University Berlin	Polytechnisches Journal	de	de	J	42173	1	
ERC AdG EUROCORR	European Correspondence to Jacob Burckhardt	en	de	L	497	8	
University Library JCS Frankfurt am Main	Medieval Manuscript Collection	de	la,de,var.	M	634	1,2,5, 6,7	METS/ MODS
	Hebrew Manuscript Collection	de	he,var.	M	378		
	Modern Manuscripts	de	la,it,de,var.	M	279		
	Oriental Manuscripts	de	gez,la,cu,var.	M	29		
	Max Horkheimer Estate	de	var.	A	272		
Georg Eckert Institute for Textbook Research	GEI-Digital	de	de,la	B	3147	1,7	
Brandeis University Library via EAJC[e]	Spanish Civil War Posters	en	es	I	112		
Center for Jewish History via EAJC	YIVO Institute for Jewish Research Collection	en	he,yi,ru,var.	B/A	3987	4	MARC
	Leo Baeck Institute Collection	en	de,en,var.	M/B/J/A	7885		
National Library of Israel	Hebrew & various language Manuscripts	he	he,ar,var.	M	1296		
	Hebrew, Yiddish & various language Books	he,en	yi,he,var.	B	7722		
	Archival Material	he	de,he,en,var.	A	2775		
Berlin State Library	Personal Papers of Adelbert von Chamisso	de	de,fr,var.	M/L	4662	1	EAD
	Personal Papers of Gerhart Hauptmann	de	de,var.	M/L	14295		
	Publisher Archives of Gebauer & Schwetschke	de	de	M/L	43296		
	Western Manuscripts	de	de,la,var.	M	163		
Joint Distribution Committee (JDC) via EAJC	Records of the NYC Office of the JDC, 1914-18	en	en,var.	F	207		
Austrian National Library	Austrian Books Online	de	de,it,fr,var.	B/J	44425	1	MAB2
	Codices	de	la,de,tr,var.	M	175		
Max Planck Institute for the History of Science	Islamic Scientific Manuscripts Initiative	en	ar	M	763		Custom
	MPIWG Digital Rare Book Library	en	la,fr,de,var.	M/B	1264	1,3,4	
	The manuscripts of Thomas Harriot	en	en,var.	M	24	1	
Petőfi Literary Museum	A Tett Magazine	hu	hu	J	183	7	DC

[a] ML: Metadata Language [b] CL: Content Language
[c] M: Manuscripts / L: Letters / B: Books / I: Images / J: Journal Articles / A: Archival Items / F: Archival File
[d] 1: GND / 2: DBpedia / 3: DDC / 4: LCSH / 5: ZDB / 6: Geonames / 7: VIAF / 8: Freebase
[e] European Association for Jewish Culture

The Linked Data representation of the metadata as described in this paper can be accessed online[3] and is as part of the LOD cloud also registered on Datahub.[4] DM2E is a five-star Linked Data source adhering to the Linked Data Principles [3], i.e., it uses dereferenceable URIs, provides all metadata in RDF using proper content negotiation together with links to other Linked Data sources. The vocabulary achieves four stars of the Five Stars of Linked Data Vocabulary Use [12], cf. Section 3. The data is not only provided as an end in itself, but forms the basis for a scholarly research platform allowing scholars to access the underlying content to annotate it and link it to other sources (Section 4). In order to support the scholars in finding relevant content, the RDF data is enriched – contextualized – as part of the ingestion process (Section 5). A specialty is the provision of full provenance of the data and the support of versioning, as described in Section 6. All the RDF data provided by DM2E can be used without restrictions in accordance with the CC0 public domain dedication.[5]

The rights statements for the described digitized objects are individually assigned by the content providers who have to choose an appropriate statement from the options[6] offered by Europeana and attach this information to each individual item.

2. Sources

One major aspect of the DM2E project was publishing metadata about a number of international high profile collections both as Linked Data and through Europeana. Despite its name, DM2E is not restricted to manuscripts but contains also other historical resources like letters, books, images, journal articles, or archival items. Table 1 shows an overview on the content available as Linked Data, broken down by provider and collection name, metadata and content language, type of content, instance count, used reference authorities and metadata source format. The stated counts represent the respective number of instances for which the property

[3] http://data.dm2e.eu/data (10.12.2015)
[4] http://datahub.io/dataset/dm2e (10.12.2015)
[5] http://creativecommons.org/publicdomain/zero/1.0/ (10.12.2015)
[6] http://pro.europeana.eu/available-rights-statements (10.12.2015)

dm2e:displayLevel is set to true (see Section 3). As can be seen, this dataset is based on the integration of a variety of source metadata formats, reflecting the heterogeneity of the underlying materials, their international character and the flexibility of the DM2E model to effectively represent such diverse content.

Accordingly, there is no common workflow for providers to map their data to the DM2E model. Starting with the tools and the documentation provided by DM2E, providers therefore developed their own metadata transformations mainly based on XSLT, although some chose to directly implement export routines into their collection management systems. Initial consistency checks for mapped data revealed that despite the very detailed specification of the DM2E model some providers showed great creativity in individual interpretations of specific model features, most notably regarding the representation of hierarchies. In order to maintain a homegeneous data representation, specific mapping rules have been established for such cases and distributed amongst the providers in form of a recommendation document. In some cases, transformations created by one provider could be reused or adapted for other providers. This especially proved to be effective for the highly standardized library metadata formats such as MARC, METS/MODS and MAB2. The mapping recommendations and the resulting metadata crosswalks are documented on the DM2E wiki[7], a more detailed description of the individual transformation workflows is available as project deliverable [4].

3. Data Model

The DM2E model is an application profile of EDM, i.e., an application-specific specialization for the representation of manuscripts and similar historical content like old prints, posters, books and old journals [6]. EDM itself is very generic to represent resources provided by museums, libraries, archives and galleries all over Europe. It is based on top-level ontologies like OAI-ORE, Dublin Core and SKOS. Core classes are edm:ProvidedCHO for the described cultural heritage object (CHO), ore:Aggregation for the metadata record provided for the described CHO and edm:WebResource for views of the described CHO, such as images. CHOs can be further qualified by links to contextual resources being instances of edm:Agent, edm:TimeSpan, edm:Place, or skos:Concept.

The example of a manuscript by the philosopher Ludwig Wittgenstein shown in Figure 1 illustrates how DM2E data is built-up. Bold resources have been added in the DM2E model, others are part of the underlying EDM. ore:Aggregation shows, for example, who has created and mapped the metadata and where the CHO is shown on the Web, while edm:ProvidedCHO is about the physical object that is described. The DM2E model allows that CHOs have multiple hierarchical layers. The CHO here has two layers: the manuscript and paragraphs within the manuscript. The type of the CHO is given via dc:type in edm:ProvidedCHO. Agents are divided into organizations and persons and can be further described. dm2e:levelOfHierarchy "1" says that the manuscript is the highest hierarchical level of this object. Hierarchies are very collection-specific; usually the provider knows best which level is most relevant for scholars. Therefore, the property dm2e:displayLevel is used to give applications a hint, if this CHO for example should appear in a result list of a search interface or if it should show up in Europeana. Users searching for Wittgenstein usually do not want to see every paragraph of a Wittgenstein manuscript in the search results.

The DM2E model adds mostly subclasses and -properties for the domain of manuscripts to existing elements of the EDM. Main additions have been made for person roles, e.g. dm2e:composer, pro:author and more properties that specialize dc:creator, classes that are used to indicate the CHO's type like bibo:Letter, dm2e:Manuscript or fabio:Article or properties to describe specifics of manuscripts or other document types of the model, e.g. dm2e:incipit for the opening words of a manuscript or dm2e:receivedOn for a date on which a letter was received. An hierarchical object is described on every level using edm:ProvidedCHO and ore:Aggregation as the metadata vary between pages or even paragraphs. Additionally, this allows to refer to every CHO as an independent object if desired. The model was created and refined in an iterative, agile process taking into account several mapping workshops and constant feedback by the data providers and application developers. This feedback led to model changes where properties or classes have been added or dropped, or property ranges

[7] http://wiki.dm2e.eu (10.12.2015)

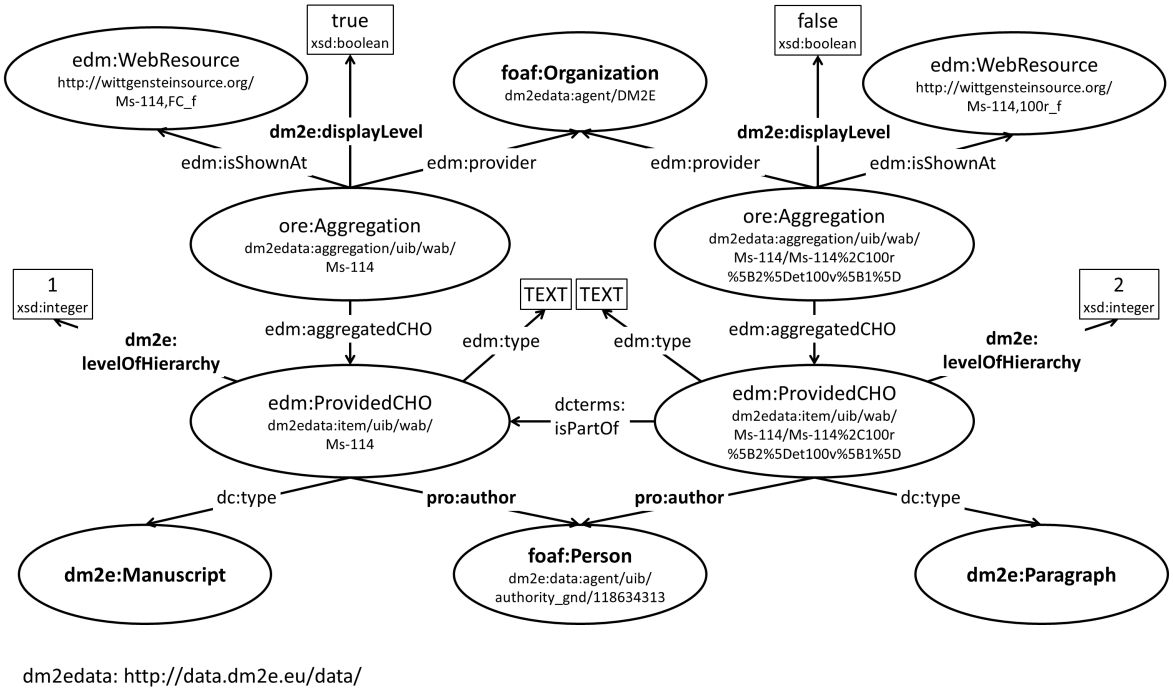

Fig. 1. Example of the DM2E model in use: representation snippet of a Wittgenstein manuscript.

have been adapted. After the intellectual collection of requirements and resulting initial versions of the DM2E model, an evaluation of the mapped data [2] has been conducted to get information about the actual usage of the model. As a result, many properties and classes of the model have been removed despite originally been asked for by providers as they have not been used. In the appendix (Table 3), we provide an overview on the usage of the metadata fields for the class `edm:ProvidedCHO` across the datasets.

Whenever possible, established vocabularies have been reused, precisely: BIBFRAME, BIBO, CIDOC-CRM, FABIO, PRO, rdaGr2, VIVO and VoID. DM2E-specific usage guidelines for each reused element are provided via `dm2e:scopeNote`. On the five stars scale for LOD vocabulary use proposed by [12], the model gets four stars, as it is dereferencable and machine-readable, linked to other vocabularies, has metadata about it but is not (yet) linked to by other vocabularies. As of January 20, 2015, the DM2E model contains 65 additional properties and 23 additional classes. The DM2E dataset currently includes descriptions for 2,670,996 cultural heritage objects, 2,478,765 of which representing single annotatable pages, while 182,259 have displayLevel set. Regarding contextual resources, 33,080 objects of type `skos:Concept` are available, 37,772 are typed as `edm:TimeSpan`, 21,304 as `edm:Agent`, 3,751 as `foaf:Organization`, 104,779 as `foaf:Person` and 27,266 as `edm:Place`. Compared to the EDM data that is available via the Europeana LOD pilot [11], the specialized DM2E data forms a smaller, complementary dataset with RDF statements at a very detailed level. The namespace URI for the DM2E model schema is `http://onto.dm2e.eu/schemas/dm2e/` and `data.dm2e.eu/data/` for instances. The full documentation of the model, including detailed changelogs between model versions, can be accessed via the DM2E wiki.

4. Application

Two applications consuming the DM2E Linked Data have been implemented in the DM2E project to provide the scholarly research platform. The first one is a faceted browser that allows scholars to make sense of the DM2E collections and navigate them along several dimensions, iteratively restricting the search results by language, author, publishing institutions, and

other metadata fields. Facets are derived from a SOLR[8] index populated by running queries against the DM2E SPARQL endpoint. To populate such an index we used the approach and implementation described in [13]. This is based on SPARQL configurations to derive document id, facets and corresponding values from pre-defined graph patterns. It is easy for example to get all datatype properties of a resource and turn them into corresponding facets. In this way facets names (fields in Solr) are not known a priori, so we use dynamic fields: in the following example fields ending with `ss` are multivalued string fields derived from a RDF property. In the prototype implementation we use the `dm2e:displayLevel` property to exclude hierarchical levels that should not show up in search results. As we currently do not provide a public SPARQL endpoint, the RESTful Solr search API is also an important building block to search the DM2E data programmatically. It straightforwardly provides a fast full text search over the main metadata fields, e.g. by using `q=Bartolomeo` as URL query string.[9] Furthermore it can be used to perform non relational queries, such as "all the cultural objects that are written in Italian and issued at a precise year, e.g. 1831", by using the following query string `fq=language_ss:it&fq=issued_ss:18*&wt=json`.[10]

The API response in a JSON array, from which the dereferenciable URL of Linked Data resources are easily retrievable by looking at the "id" slot. Solr does not provide a direct way to get all the facets used in the index, which would facilitate the developer in writing queries. However such a list can be retrieved with a work-around with the following query: `q=*:*&wt=csv&rows=0&facet`.[11]

Based on the Solr API, a end-user faceted browser was developed by customising Ajax-Solr. Each resource shown in the faceted browser can be opened in its provider's own digital library (following the EDM property `edm:isShownAt`) and its Linked Data representation can be reached by clicking on "see RDF data" that points to its dereferenciable URL. The DM2E faceted search is available online[12] and its usage demonstrated in a short online screencast.[13]

Furthermore, for datasets containing links to annotatable digital objects, users are provided with "Annotate with Pundit" links that direct them to the DM2E semantic annotation environment. The latter constitutes the second web application built on top of the data and is based on Pundit and Feed, two software components developed as part of the DM2E project.

Pundit[14] is an annotation tool that allows users to enrich web pages with semantically structured data. In DM2E substantial improvements was done over the previous versions [9] [10], leading to a completely new user interface and additional annotation functionalities [14]. Annotations in Pundit encode machine readable semantic connections among images, texts and LOD entities in form of RDF triples (using the Open Annotation data model),[15] consumable via SPARQL or a dedicated REST API.[16]

On the other hand, the Feed REST API provides access to Pundit "as a service" by using the URL of a Web page to be annotated as a call parameter. An extension to the Feed API has been developed in DM2E, allowing this call parameter to also be a dereferenceable URL of an RDF description of a digitized object. Feed parses this RDF description to create a customized annotation environment. While the faceted search application only works on resources at display level (e.g. entire books), the annotation environment allows users to go deeper into the hierarchy. The `dcterms:isPartOf` and `edm:isNextInSequence` properties are used to provide basic navigation functionalities such as reaching a specific page of a manuscript or going to the next/previous pages. The actual digital contents that users can annotate are retrieved by following `dm2e:hasAnnotatableVersionAt` links or, alternatively, the `edm:object` links. There can be multiple annotatable objects associated with a resource as, for example, the facsimile image and its HTML transcription. In this case, both contents are shown and made annotatable.[17] In case of the presence of links to

[8] `http://lucene.apache.org/solr/` (10.12.2015)

[9] `http://141.20.126.236:8080/solr-dm2e/collection1/select?q=Bartolomeo`

[10] `http://141.20.126.236:8080/solr-dm2e/collection1/select?q=*:*&fq=language_ss:it&fq=issued_ss:18*&wt=json`

[11] `http://141.20.126.236:8080/solr-dm2e/select?q=*:*&wt=csv&rows=0&facet`

[12] `http://purl.org/net/dm2e/search` (10.12.2015)

[13] `https://youtu.be/_rQ_7NhewhQ` (10.12.2015)

[14] `https://thepund.it/` (10.12.2015)

[15] `http://www.openannotation.org/spec/core/` (10.12.2015)

[16] Pundit server API documentation: `http://net7.github.io/pundit2/rest-api.html` (10.12.2015)

[17] Annotatable page example: `http://bit.ly/1FIu6dA` (10.12.2015)

popular LOD datasets such as DBpedia,[18] Feed is able to gather additional metadata (e.g. full names, descriptions, categories) in order to provide additional context to scholars. For this purpose, we established automated contextualization processes, as described in the next section.

By annotating digital objects with Pundit, users in fact create additional RDF knowledge. This could be, for example, links connecting a geographical map of Metz, depicted in a manuscript page, to the city of Metz in DBpedia, or links from a sentence of a manuscript transcription to a DBpedia entity that is mentioned in such a sentence. Such RDF data, in turn, can be indexed to enrich the faceted search interface, thus improving search and discovery. Demonstrative examples of such end-user information enrichments can be seen in an online screencast.[19]

As of September 30, 2014, about 6,600 annotations for about 900 digital objects from the DM2E dataset have been created by scholars using our research platform.

5. Contextualization

Linking our datasets to external sources like GND,[20] DBpedia, Geonames,[21] or the Library of Congress Subject Headings[22] enables to easily get information about a resource, either directly by following the link to the external source or by detecting connections between resources based on the same links. While the links to GND often are already present in the original metadata, links to all other sources are generated automatically. To create the links, we use the link discovery framework Silk.[23] Silk generates links based on a linkage rule that is provided by the user. Such a linkage rule specifies the conditions which have to hold to create a link, e.g. the names of two resources need to have a Jaccard measure value above 0.8. All links that are currently in our system are generated with the same configuration which compares the labels using the Jaro Winkler distance and requires a confidence value of 0.9 aiming at a high precision while tolerating spelling variations. This might seem like a very simple method, but in most cases we have no further information in the metadata besides a simple string. Our evaluation suggests, however, that even this simple method leads to good results, as many of these strings are not ambiguous. Table 2 shows the number of generated links to each external sources. We link agents, places, and subjects.

Table 2
Number of links per external source

dbpedia	freebase	geonames	judaica	lcsh	geodata	nytimes	GND
12287	1868	1571	1474	141	5770	570	22698

Altogether, about 24,000 links have been automatically generated. With a manual analysis of 150 random links from agents to DBpedia and 150 random links from places to Linked Geodata, we evaluate the quality. For agents, 125 correct links have been detected which results in a precision of 0.83. Since DBpedia covers several labels, we can for example correctly link "Jakobä" to the DBpedia agent "Jacqueline Countess of Hainaut." The incorrect links result either from ambiguous names, e.g. "Heinrich Fischer" who refers to a Swiss rower in DBpedia and not to an author, or from incomplete information, e.g. if only the first name or surname is given.

For places, 128 correct links can be detected, resulting in a precision of 0.85 which is similar to the precision for agents. Since Linked Geodata includes labels in various languages, even places with a German label such as "München" can be linked to "Munich". The reasons for incorrect links, however, are the same to the ones for agents, e.g. the German city "Heidelberg" is mapped to the city "Heidelberg" located in South Africa due to identical labels.

Across all datasets, about 18% of all agents and 60% of all places are linked on average. With a different linkage rule it is possible to detect more links but with the risk to reduce the precision. Further, the amount of detected links as well as their quality highly depend on the popularity and currency of the resources. Since more than one linkset can be available in our system and the user can track their provenance, more liberal linkage rules can also be applied and the user can be informed about its quality.

[18] http://wiki.dbpedia.org/ (10.12.2015)
[19] https://youtu.be/tUrdLm43CMA (10.12.2015)
[20] http://www.dnb.de/DE/Standardisierung/GND/gnd_node.html (10.12.2015)
[21] http://www.geonames.org (10.12.2015)
[22] http://id.loc.gov/authorities/subjects.html
[23] http://silkframework.org/ (10.12.2015)

6. Implementation

At the core of DM2E's infrastructure is a Jena TDB[24] triplestore, accessible by a Jena Fuseki SPARQL endpoint. After evaluating a few different RDF storage solutions, this combination offered the perfect balance between maintainability, scalability and versatility. In fact, all DM2E's internal applications and the infrastructure are interfacing with the data exclusively through the SPARQL endpoint, making, in theory, the actual triplestore implementation interchangeable. The RDF data is partitioned into Named Graphs that correspond to individual ingestions (see also Section 6.3 – Dataset Provenance); exporting and importing N-Quad dumps of the full data store as well as specific subsets is straight-forward. Dumps for the DM2E metadata collection[25] and for the contextualized external entities[26] are available to the public.

6.1. Data Ingestion

There are two user interfaces that allow data providers or mapping institutions to deliver data to DM2E: A Linked Data-based workflow engine with an HTML5 web interface allows casual users to test their transformations and the ingestion process (Omnom)[27] while a set of command line tools is targeted at power users doing large-scale ingestions and conversions (dm2e-data.sh).[28]

Omnom is centered on the idea that RDF's flexible graph-based structure combined with the semantic expressivity of ontologies[29] not only allows the definition and execution of intelligent workflows, automating tedious, long-running and error-prone tasks, but solves the problem of tracking data provenance [8]. Combined with the simple Web User Interface,[30] Omnom can be very helpful for the technically-non-too-savvy to understand the processes of data mapping, data transformation and data ingestion and iteratively improve their own workflow, though Omnom's approach to use and persist RDF for all data does lead to suboptimal performance when doing full-scale transformations/ingestions.

The command line suite of tools is developed with a server environment in mind and consists of a set of Java tools for DM2E validation, provenance-tracking data ingestion, DM2E-EDM-conversion and EDM validation, as well as shell scripts encapsulating XSLT transformers and RDF serializers and for orchestrating the various operations.

The authoritative source of the DM2E model is the textual/tabular *DM2E Model Specification*, which contains not only the definitions of all properties and classes to be used, but illustrates their usage with examples. The specs are synchronously formalized as an dereferenceable OWL ontology. However, the DM2E model puts restrictions on the usage of properties and classes that cannot be expressed under OWL's Open World Assumption. These restrictions are targeted towards structural validation of subgraphs of DM2E data rather than inference of new facts. While DM2E is involved in the development of community standards for RDF validation [5], we implemented a custom solution using Java, available on GitHub.[31] While the validation tool is "hard-wired" to DM2E's model, it is rather meticulous and has proven useful not only for discovering outright model violations (e.g. wrong cardinality of properties or missing conditional statements) but stylistic problems such as unwise characters in URIs and labels or variations in the UTF-8 normalization.

6.2. Delivery to Europeana

Being a domain aggregator for Europeana, DM2E has a strong focus on interoperability with the EDM, both on the model and data level. The DM2E model is a specialization of the EDM, i.e., after RDFS inference on the data and removing any statements with properties not contained in EDM, every DM2E-compliant subgraph is an EDM-compliant subgraph. DM2E uses this technique to convert the DM2E data into pure EDM to make the ingestion as easy as possible for the Europeana side, using a synthesis of the two models

[24] `http://jena.apache.org` (10.12.2015)

[25] DM2E collection metadata dump: `http://data.dm2e.eu/dm2e-fuseki-direct.2016-01-07.final.nquads.gz` (13.01.2016)

[26] DM2E contextualized links dump: `http://data.dm2e.eu/dm2e-fuseki-direct.2016-01-07.linksets.nquads.gz` (13.01.2016)

[27] `http://omnom.dm2e.eu`

[28] `https://github.com/DM2E/dm2e-ontologies/blob/master/src/main/bash/dm2e-data.sh` (10.12.2015)

[29] `http://onto.dm2e.eu/omnom` (10.12.2015), `http://onto.dm2e.eu/omnom-types` (10.12.2015)

[30] `https://github.com/DM2E/dm2e-gui` (10.12.2015)

[31] `https://github.com/DM2E/dm2e-ontologies` (10.12.2015)

expressed in OWL.[32] As the last step before delivery to Europeana, the produced EDM representations are validated using a combination of XML Schema and Schematron.[33]

Due to its ubiquitious deployment in the GLAM sector and its proven track record for scalability, DM2E and Europeana agreed on OAI-PMH as the preferred mode of delivery of data for ingestion into Europeana. Using a multi-step process of extracting per-ore:Aggregation-subgraphs from the triplestore, validation against the DM2E model, conversion to EDM, data massaging and validation against the EDM model,[34] an EDM dump of all data in DM2E is created monthly. With OAI-PMH set names corresponding to datasets, these EDM RDF/XML files are then served using the Repox OAI-PMH repository.

6.3. Linked Data API

The Linked Data API is implemented using a significantly advanced version of Pubby.[35] The source code for this DM2E-specific version is available via GitHub.[36] An integration of the additional features – which are of general interest – into the main branch of Pubby is planned. The basis for all of them is unleashing the power of SPARQL by allowing arbitrary URI patterns to be mapped to customized SPARQL queries. In the following, we describe how the requirements regarding data access have been accomplished for the DM2E data within Pubby.

Multiple resource handling. DM2E implements the OAI-ORE resource map, i.e., whenever the URI of a resource or an aggregation is requested, the client gets redirected to the URI of a resource map. The resource map contains both information about a resource and information about the aggregation – which roughly represents a metadata record in EDM. This implementation also follows practical considerations from the point of view of application developers, as, more often than not, the data about a resource and the data about the aggregation are used together. So this leads to a substantial reduction of necessary requests to the API.

Versioning. All DM2E data is versioned, i.e., the data provided under the URI of a resource map never changes. When updated data is ingested, the API redirects to the new resource map, but the new resource map gets a new URI and contains links to earlier versions of the data in the form of `prov:wasRevisionOf`. This allows the stable identification of triples within the data, a prerequisite for the data to become a trusted subject of scholarly work.

Dataset provenance. The full provenance of the DM2E data is provided by linking resource maps and other data pages to superordinate datasets using the VoID vocabulary [1]. The datasets are versioned and all data in a dataset shares the same provenance, following the idea of a common provenance context to support provenance-aware Linked Data applications [7]. The version of a resource map and the provided provenance information then simply corresponds to the version and provenance of the dataset. Versioned datasets are implemented as Named Graphs.

Statement-level provenance. To support contextualized resources with statements from various enrichment processes, a special approach has been implemented using statement annotations [7]. Subject URIs are created for all statements and these URIs are linked to the datasets the statements originate from. The statement URIs are identified and described as statements using RDF reification. All reification triples are created on the fly, only where necessary, and can safely be ignored by applications not interested in the provenance of the statements. The HTML representation of the contextualized resources makes use of this information and provides an "Oh Yeah?" button for all – possibly wrong – links to external resources, leading to the provenance information of the statement.[37] To the best of our knowledge this is the first implementation of this button as envisioned by Tim Berners-Lee [3].

7. Discussion

Several aspects make the DM2E dataset an interesting and unique source of information. First of all –

[32] `https://github.com/DM2E/dm2e-ontologies/blob/master/src/main/resources/edm/edm.owl` (10.12.2015)

[33] `https://github.com/DM2E/edm-validation`

[34] `https://github.com/DM2E/dm2e-ontologies/blob/master/src/main/bash/dm2e-data.sh` (10.12.2015)

[35] `http://wifo5-03.informatik.uni-mannheim.de/pubby/` (10.12.2015)

[36] `https://github.com/dm2e/pubby` (10.12.2015)

[37] For example the city Nancy: `http://data.dm2e.eu/data/html/place/onb/abo/Nancy` (10.12.2015)

following the goals of the DM2E project – it contains data from many, carefully selected collections of not only manuscripts, but also old prints, posters, books and old journals with historic value. The data model was developed specifically for this domain where no suitable comprehensive data models existed yet. The DM2E model is also an example of an application profile, an application-specific specialization of the EDM. As such, the data blends well with the huge amount of EDM data available through Europeana. In contrast to many other Linked Datasets, the model and the API have both been tailored to the original data as well as to consuming applications. From a technical point of view, the use of multiple resource representations, versioning and the provision of a full provenance chain have to be mentioned, particularly the proper separation of original, curated metadata from data enrichments generated by automated processes with varying quality. The main short-coming is arguably the lack of a publicly available SPARQL endpoint, mainly due to performance considerations. Fast response times for the scholarly research platform have higher priority. The SOLR-based search and browse interface, however, is provided as convenient entry point to the data and provides a RESTful search API sufficient for most use cases. The data itself also has some shortcomings due to the heterogeneity of the original data. The quality of the metadata ranges from rich descriptions with unambiguous identifiers from authority files for agents, places and subjects to sparse descriptions with few information hidden in free-text fields. It is insofar a dilemma that the contextualization works best for the better data and particularly the poor data is hard to improve. A remedy might be the feedback of data from the annotations provided by the scholars. We plan to investigate this as part of our future work, when more annotations will hopefully be available.

References

[1] Keith Alexander, Richard Cyganiak, Michael Hausenblas, and Jun Zhao. Describing linked datasets – on the design and usage of voiD, the "vocabulary of interlinked datasets". In Christian Bizer, Tom Heath, Tim Berners-Lee, and Kingsley Idehen, editors, *Proceedings of the WWW2009 Workshop on Linked Data on the Web, LDOW 2009, Madrid, Spain, April 20, 2009.*, volume 538 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2009. Available at `http://ceur-ws.org/Vol-538/ldow2009_paper20.pdf`.

[2] Konstantin Baierer, Evelyn Dröge, Vivien Petras, and Violeta Trkulja. Linked data mapping cultures: An evaluation of metadata usage and distribution in a linked data environment. In William E. Moen and Amy Rushing, editors, *Proceedings of the 2014 International Conference on Dublin Core and Metadata Applications, DC 2014, Austin, Texas, USA, October 8-11, 2014*, pages 1–11. Dublin Core Metadata Initiative, 2014. Available at `http://dcpapers.dublincore.org/pubs/article/view/3699`.

[3] Tim Berners-Lee. Linked Data. `http://www.w3.org/DesignIssues/LinkedData.html`, 2006.

[4] Kristin Dill, Evelyn Dröge, Øyvind Liland Gjesdal, Doron Goldfarb, Esther Guggenheim, Julia Iwanowa, Marko Knepper, Gerhard Müller, Alois Pichler, Kilian Schmidtner, Klaus Thoden, and Jorge Urzúa. D1.2 – final integration report. `http://dm2e.eu/files/D1.2_2.0_Final_Integration_Report_140214_final.pdf`, 2014.

[5] Evelyn Dröge, Thomas Bosch, Valentine Charles, Robina Clayphan, Mark Matienzo, Stefanie Rühle, Adrian Pohl, Mika Alonen, Lars Svensson, and Karen Coyle. Report on the current state: use cases and validation requirements [editor's draft]. Deliverable 1, DCMI RDF Application Profiles Task Force, 2014. Available at `http://wiki.dublincore.org/index.php/Deliverable_1`.

[6] Evelyn Dröge, Julia Iwanowa, and Steffen Hennicke. A specialisation of the Europeana Data Model for the representation of manuscripts: The DM2E model. In *Libraries in the Digital Age (LIDA) Proceedings*, volume 13, 2014. Available at `http://ozk.unizd.hr/proceedings/index.php/lida/article/view/117`.

[7] Kai Eckert. Provenance and annotations for Linked Data. In Muriel Foulonneau and Kai Eckert, editors, *Proceedings of the 2013 International Conference on Dublin Core and Metadata Applications, DC 2013, Lisbon, Portugal, September 2-6, 2013*, pages 9–18. Dublin Core Metadata Initiative, 2013. Available at `http://dcpapers.dublincore.org/pubs/article/view/3669`.

[8] Kai Eckert, Dominique Ritze, Konstantin Baierer, and Christian Bizer. RESTful open workflows for data provenance and reuse. In Chin-Wan Chung, Andrei Z. Broder, Kyuseok Shim, and Torsten Suel, editors, *23rd International World Wide Web Conference, WWW '14, Seoul, Republic of Korea, April 7-11, 2014, Companion Volume*, pages 259–260. ACM, 2014. doi:10.1145/2567948.2577347.

[9] Marco Grassi, Christian Morbidoni, Michele Nucci, Simone Fonda, and Giovanni Ledda. Pundit: Semantically structured annotations for web contents and digital libraries. In Annett Mitschick, Fernando Loizides, Livia Predoiu, Andreas Nürnberger, and Seamus Ross, editors, *Proceedings of the 2nd International Workshop on Semantic Digital Archives, Paphos, Cyprus, September 27, 2012*, volume 912 of *CEUR Workshop Proceedings*, pages 49–60. CEUR-WS.org, 2012. Available at `http://ceur-ws.org/Vol-912/paper4.pdf`.

[10] Marco Grassi, Christian Morbidoni, Michele Nucci, Simone Fonda, and Francesco Piazza. Pundit: Augmenting web contents with semantics. *Literary and Linguistic Computing*, 28(4):640–659, 2013. doi:10.1093/llc/fqt060.

[11] Antoine Isaac and Bernhard Haslhofer. Europeana Linked Open Data - data.europeana.eu. *Semantic Web*, 4(3):291–297, 2013. doi:10.3233/SW-120092.

[12] Krzysztof Janowicz, Pascal Hitzler, Benjamin Adams, Dave Kolas, and Charles Vardeman. Five stars of Linked Data vocabulary use. *Semantic Web*, 5(3):173–176, 2014. doi:10.3233/SW-140135.

[13] Christian Morbidoni. Linked data and facets to explore text corpora in the humanities: a case study. In Matthew Horridge, Marco Rospocher, and Jacco van Ossenbruggen, editors, *Proceedings of the ISWC 2014 Posters & Demonstrations Track a track within the 13th International Semantic Web Conference, ISWC 2014, Riva del Garda, Italy, October 21, 2014.*, volume 1272 of *CEUR Workshop Proceedings*, pages 413–416. CEUR-WS.org, 2014. Available at `http://ceur-ws.org/Vol-1272/paper_125.pdf`.

[14] Christian Morbidoni and Alessio Piccioli. Curating a document collection via crowdsourcing with Pundit 2.0. In Fabien Gandon, Christophe Guéret, Serena Villata, John G. Breslin, Catherine Faron-Zucker, and Antoine Zimmermann, editors, *The Semantic Web: ESWC 2015 Satellite Events - ESWC 2015 Satellite Events Portorož, Slovenia, May 31 - June 4, 2015, Revised Selected Papers*, volume 9341 of *Lecture Notes in Computer Science*, pages 102–106. Springer, 2015. doi:10.1007/978-3-319-25639-9_20.

Table 3

DM2E metadata term usage in `edm:ProvidedCHO` per dataset

	sbb/kpe_DE-Ha179_37172	onb/abo	uber/dingler	sbb/kpe_DE-1a_8535	cjh/lbiarchive	nli/books	sbb/kpe_DE-1a_995	gei/gei-digital	cjh/yivolibrary	nli/archives	bbaw/d ta	cjh/yivoarchive	mpiwg/rara	nli/manuscripts	cjh/lbilibrary	ub-ffm/msma	jdc/nyar1418	ecorr/burckhardtsource	ub-ffm/horkheimer	ub-ffm/mslebr	mpiwg/ismi	ub-ffm/msmo	onb/codices	sbb/manumed	pim/kassak	cjh/lbiperiodicals	bas/codsupra	brandeis/scwp	ub-ffm/msor	mpiwg/harriot	uib/wab
dc:subject		73326	42173		55729	9574		1559	3181	2725	1799	10531	1098	1445	3374		2949		2583					491		608					20
dc:language	43296	44608	42173	14311	7390	7996	5008	3147	2566	2781	1547	1540	1264	1624	789	689	207	497	272	415	763	333	193	163	183	73	49	112	48	24	20
dc:type	43296	44425	42173	14295	7079	7722	4662	3147	2477	2775	1547	1510	1264	1296	741	634	207	497	272	378	763	279	175	163	183	65	49	112	29	24	20
edm:type	43296	44425	42173	14295	7079	7722	4662	3147	2477	2775	1547	1510	1264	1296	741	634	207	497	272	378	763	279	175	163	183	65	49	112	29	24	20
rdf:type	43296	44425	42173	14295	7079	7722	4662	3147	2477	2775	1547	1510	1264	1296	741	634	207	497	272	378	763	279	175	163	183	65	49	112	29	24	20
dc:title	43296	44426	42173	14295	7091	7722	4677	3147	2488	2775	1552	1596	1264	1296	747	634	207		272	378	763	279	175	163	200	65	96	112	29	24	20
dc:identifier	43296	28398		14295	7079	7722	4662	3147	2477	2775	1547	1510	1268	1296	741	634	207		272	378	763	279	175	163		65	49	112	29	37	20
dm2e:levelOfHierarchy	43296		42173	14295			4662				1547		1264				207				763			163	183		49			24	20
dcterms:issued	40208	29101		13111	4273	5987	3961	2648	1529		1548	1053	1261		458	586				311		265	333	130	17	64			26		
dc:contributor	82424	3980		714	3988	2458	464		1706	1		745	1	1014	634	32				14		7	1	28		120			2		
dcterms:extent	43289			14294	6815	7688	4652		2429		1130	1498	1229	1237	737		206							161		65					
dm2e:writtenAt	61372			14199			3229							241				450						102							
dm2e:holdingInstitution	43296			14295			4662				1265					634			272	378		279		163					29		
dm2e:subtitle		1104	49179		777	6973			2111		1434	120		115	647								6	148	1	24					
dcterms:isPartOf		17758	42173					1541												26		24			182		77		3		
dm2e:shelfmarkLocation		39873									1265					634			272	378		279	175	163					29		
dc:publisher		26176		1	105	7721	4	3153	2457		1361	234	1193	4	460											50					
bibo:pages			42173																						165		48				
dc:creator		24712			4538	7239	2		2029			349		590	676											28					
pro:author			24599	126			772	4740			1388		1354			283		507		196		131	199	94	154		23	112	4	24	20
dm2e:publishedAt		27813						3192					1261			374				106		141						20	16		
dc:description				1196	12996	2	1769	4191	49	7746		631	438	1	89	1199	414	497	325	363		584				60	52		70	16	20
dm2e:mentioned		154	17003	1687			4872																10								
bibo:recipient	2482	3		15183			3901											497					2								
dm2e:composer				17128			3777																								
dm2e:genre	3258	5807		3642			1381				2182																				
bibo:number		16027																													
dcterms:tableOfContents		1603	8455		1821	323		2546	128	4	519	19		14	44	102				64		33	55			2			4		
edm:isNextInSequence		11882																		21		16			165		47		2		
dm2e:pageDimension		129			33	6936			2062			1092		838	75								174			3					
dcterms:alternative		1287			1203	2314			725	2467	1547	100	672	7	54	29				608		3	1			47	45				
dm2e:callNumber					7072								1262		721											64				16		
bibo:numPages		3195									1547		1229										175				1				
dcterms:spatial	3831			131			772										777														
dcterms:created										2767				643				492	272												
dc:format								2630								586				312		265							26		
dc:rights					1152				2			1238			558											1					
dm2e:printedAt		494			1	4			8		1609	2			1								1			2					
dcterms:provenance						1								1296				492					133								
edm:currentLocation																634			272	378		279						112	29		
dm2e:support														785									175				1				
dm2e:illustration		642																					85								
pro:printer		671			1	19			9			2			4											2					
bibo:volume		481																							17						
dm2e:sentOn																		492													
bibo:editor								162			98		212							3					1				16		
dm2e:copyist																160				124		123							13		
dm2e:writer				264			71																18								
dm2e:receivedIn																		256													
dm2e:receivedOn																		253													
dcterms:temporal																	207														
dm2e:previousOwner																							165								
dm2e:watermark							159																								
dm2e:cover																							149								
dm2e:explicit														49									19				48				
pro:translator							2						99										2								
dm2e:artist				8			69																								
dm2e:incipit																							20				48				
bibo:numVolumes		23																													
edm:hasMet		19																													
dm2e:principal							4																7								
dm2e:honoree				6																											
pro:illustrator		1					1																4								
dm2e:painter																							3								