Linked Data Mapping Cultures: An Evaluation of Metadata Usage and Distribution in a Linked Data Environment

Konstantin Baierer
Humboldt-Universität zu Berlin, Germany
konstantin.baierer@ibi.hu-berlin.de

Evelyn Dröge
Humboldt-Universität zu Berlin, Germany
evelyn.droege@ibi.hu-berlin.de

Vivien Petras
Humboldt-Universität zu Berlin, Germany
vivien.petras@ibi.hu-berlin.de

Violeta Trkulja
Humboldt-Universität zu Berlin, Germany
violeta.trkulja@ibi.hu-berlin.de

Abstract

In this paper, we present an analysis of metadata mappings from different providers to a Linked Data format and model in the domain of digitized manuscripts. The DM2E model is based on Linked Open Data principles and was developed for the purpose of integrating metadata records to Europeana. The paper describes the differences between individual data providers and their respective metadata mapping cultures. Explanations on how the providers map the metadata from different institutions, different domains and different metadata formats are provided and supported by visualizations. The analysis of the mappings serves to evaluate the DM2E model and provides strategic insight for improving both mapping processes and the model itself.

Keywords: mapping evaluation; ontology evaluation; mapping varieties; DM2E model; Linked Data; Europeana

1. Introduction

Do mapping preferences of individual institutions influence the resulting data from a mapping process? In this paper, mapped datasets from eight different data providers (DP) processed by six different mapping institutions (MI) were analyzed. The primary aim of the analysis was an evaluation of the model to which the data is mapped. Based on the differences of mappings in the evaluation, different Linked Data mapping cultures emerged.

The evaluation of a dataset or data model provides insight into over- and underused parts of the model or misrepresented or misunderstood data mappings. Previous studies have looked at the distribution and usage of fields or model classes and properties and the mapping data in library catalogs (e.g. Seiffert, 2001; Smith-Yoshimura, Argus et al., 2010). These studies show that only a subset of the provided properties in data formats are used in practice. Palavitsinis, Manouselis & Sanchez-Alonso (2014) observed in their study of metadata quality in cultural collections that the "perceived usefulness for all elements of an application profile drops when the number of these elements rises" (p. 9). In Linked Data research, the focus has been on the analysis of certain vocabularies (e.g. Alexander, Cyganiak et al., 2009) and statistics on individual or aggregations of RDF datasets including data accessibility and coverage (Auer, Demter et al., 2012). Klimek, Helmich & Nacasky (2014) built a Linked Data Visualization Model (LDVM) which creates an analytical RDF abstraction and a visual mapping transformation.

This paper first introduces the DM2E model and its application context and then provides general statistics on the use of different model classes and properties by different providers and

https://doi.org/10.23106/dcmi.952136385

mapping institutions. Different data and model characteristics are discussed to provide an analysis of different mapping styles (cultures) and their consequences.

2. A Data Model for Cultural Heritage

Europeana[1] is the European digital library, which gives access to more than 30 million library, archive, museum and audio-visual objects from 36 countries. These objects are digitized and described by content providers in different metadata formats. National or domain aggregators deliver the object metadata to Europeana in the Europeana data model (EDM) (EDM Primer, 2013). Digitised Manuscripts to Europeana (DM2E)[2] is a domain aggregator contributing to the development of Europeana. Among other goals, DM2E collects, maps and delivers rich metadata about manuscripts to Europeana.

The metadata mapping and the ingestion of mapped data into Europeana are supported by a specialization of the EDM for manuscripts that was developed for DM2E. The EDM is very broad and generic in order to fit the different metadata standards like TEI or METS/MODS in which cultural heritage objects (also referred to as CHOs) are described by data providers. The model is RDF-based and can thus easily be extended by others as done in the DM2E project. The resulting specialization is called the DM2E model.

The DM2E model (Dröge, Iwanowa & Hennicke, 2014a) has been built as a specialization of the EDM in order to represent rich manuscript metadata in Europeana, which is also published as Linked Open Data (LOD) (Heath & Bizer, 2011). The development approach of the model was bottom-up: requirements from data providers as well as from technical partners were collected and new properties or classes were created or reused from external vocabularies. Properties and classes were added as subproperties / -classes to EDM resources when possible in order to enable backwards compatibility. In that way, the main structure of the EDM remains unchanged in the DM2E model. The core classes of both models are *edm:ProvidedCHO* for the cultural heritage object, *ore:Aggregation* for the provided metadata record and *edm:WebResource* for Web resources related to a CHO, e.g. an image of it. The class that is most extensively specialized in the DM2E model is *edm:ProvidedCHO*. More than 50 properties were added to this class to better describe the creator of a CHO, its contributors and concepts, places and time spans related to it. Similar to the EDM, the DM2E model mainly focuses on properties and not on classes to describe the provided data. Nevertheless, a small amount of classes were also added, e.g. to differentiate various types of CHOs like *dm2e:Page*, *bibo:Book* or *fabio:Article*. These classes are important to model hierarchical objects which are not yet fully supported in EDM.

3. Distribution of Classes and Properties

Ten datasets mapped to the RDF-based DM2E model describing manuscripts, books, letters and journal articles were analyzed. The total amount of RDF statements in the analyzed sample is 61,365,146. The data was delivered by eight data providers (DP) and mapped by six different mapping institutions (MI). The DPs, MIs and datasets were anonymized as the focus of the study does not lay in specifics of a single dataset but in the differences between the mapping behaviour of the six MIs. Our assumption is that not only the provided data but also the particular mapping approach influences the resulting data in the DM2E model. Table 1 shows the providers, datasets, the metadata format of the data before the ingestion and the responsible mapping institution. All data was mapped to the DM2E model version 1.1, latest revision (Dröge, Iwanowa et al., 2014b).

[1] Europeana website: http://europeana.eu/ (last accessed 22.04.2014).

[2] DM2E website: http://dm2e.eu/ (last accessed 22.04.2014).

[3] https://github.com/DM2E/dm2e-analysis/tree/master/sparql (last accessed 15.05.2014).

[4] https://github.com/DM2E/dm2e-analysis/blob/master/build_tables.py (last accessed 15.05.2014).

[5] https://developers.google.com/chart/ (last accessed 15.05.2014).

[2] DM2E website: http://dm2e.eu/ (last accessed 22.04.2014).

https://doi.org/10.23106/dcmi.952136385

The first aim of the analysis was to evaluate the DM2E model by identifying properties and classes that were not mapped. Unmapped resources could potentially be removed from the model to reduce its complexity. The analysis of the mappings could also be used to evaluate whether the model can cover different domains. Can a generic model like the EDM and its specializations be used to represent this data or do the Linked Data mapping cultures vary too much? Does a mapping reflect the institution that has mapped the data?

TABLE 1: Analyzed datasets.

Data Provider (DP)	Dataset	Metadata format	Mapping institution (MI)
DP I	Dataset 1	proprietary format	MI A
DP I	Dataset 2	proprietary format	MI A
DP II	Dataset 3	MAB2	MI B
DP II	Dataset 4	MAB2	MI B
DP III	Dataset 5	METS/MODS	MI C
DP IV	Dataset 6	METS/MODS	MI C
DP V	Dataset 7	TEI P5	MI D
DP VI	Dataset 8	EAD	MI D
DP VII	Dataset 9	TEI P5	MI E
DP VIII	Dataset 10	TEI P5	MI F

The evaluation reported in this paper is based on an automated analysis and visualizations. The RDF data in the triple store is organized in Named Graphs (Carroll et al., 2005), each Named Graph representing a specific ingestion of a specific dataset including full provenance. Using SPARQL, the latest ingestion of each dataset was determined. Then, a set of SPARQL queries was run on the data in these ingestions[3] to gather the raw counts for various quantifiable aspects of these datasets, including generic statistics such as number of statements, number of specific predicates, number of different ontologies, ranges of predicates, RDF types, as well as DM2E-specific statistics such as frequency of certain subclasses of *edm:PhysicalThing* or occurrences of predefined statement patterns. A Python script[4] then collated the raw tabular data, calculated means, sums and ratios within and across datasets and produced HTML with embedded SVG using the Google Chart data visualization API[5]. Unprocessed visualizations[6] and the source code[7] are available.

The providers or mapping institutions used a large variety of classes and properties of the DM2E model and produced rich mappings. Still, more than a half of all classes (24 out of 43) and about a third of all properties (47 out of 125) that the model offers were not used by any of the providers. The counts do not include classes and properties that are used for means beyond manuscript metadata, e.g. for external annotation tools or for tracking provenance within the DM2E interoperability infrastructure.

Figure 1 shows the distribution of all properties. The most frequently used properties are *dc:contributor*, *edm:rights*, *dc:format* und *dc:description*. Properties which must be used exactly once occur for each of the ca. 2.1 million CHOs: *dm2e:hasAnnotatableObject* (strongly recommended), *dc:language* (mandatory), *edm:dataProvider* (mandatory), *dc:type* (mandatory), *edm:aggregatedCHO* (connection between the CHO and the aggregation; this is mandatory and must occur once per object), *edm:type* (mandatory), *dm2e:displayLevel* (mandatory). The property *dc:title* is not mandatory and is used “only” 1,722,542 times in 2,134,934 CHOs. The strongly recommended properties were used almost as often as the mandatory ones. A major part

[3] https://github.com/DM2E/dm2e-analysis/tree/master/sparql (last accessed 15.05.2014).
[4] https://github.com/DM2E/dm2e-analysis/blob/master/build_tables.py (last accessed 15.05.2014).
[5] https://developers.google.com/chart/ (last accessed 15.05.2014).
[6] http://data.dm2e.eu/visualize/index.html (last accessed 24.07.2014).
[7] https://github.com/DM2E/dm2e-analysis (last accessed 15.05.2014).

https://doi.org/10.23106/dcmi.952136385

of the properties is used infrequently compared to the number of CHOs, a logical consequence because specific properties just fit particular datasets. About one third of the properties was not mapped. Both, DM2E-specific properties but also EDM properties, were not mapped. Properties from contextual classes, e.g. coordinates of places (*wgs84_pos:lat*, *wgs84_pos:long*), the date an institution started (*rdaGr2:dateOfEstablishment*) or ended (*rdaGr2:dateOfTermination*) are possibly simply missing in the data. SKOS properties like *skos:broader*, *skos:narrower* or *skos:notation* were not mapped. Uncommon properties like *dm2e:levelOfGenesis*, *dm2e:influencedBy* or *dm2e:misattributed* were not mapped even though they were explicitly requested by data providers. The distribution of properties mirrors previous findings from Seiffert (2001), who analyzed MAB fields of title data in libraries and showed that 58.46% of MAB fields for bibliographic data were unused. The same results could be found in an internal statistical analysis of EDM data at Europeana conducted in January 2014, which concluded that 40% of the fields remained unused.

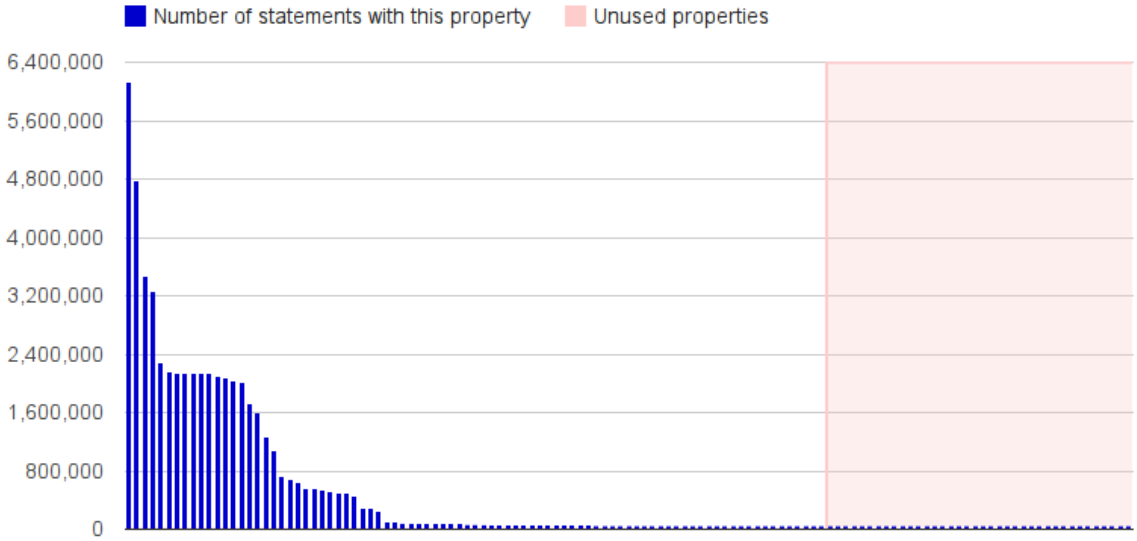

FIG. 1: Absolute frequency of all predicates. Properties on the right side of the vertical bar were never used in any dataset.

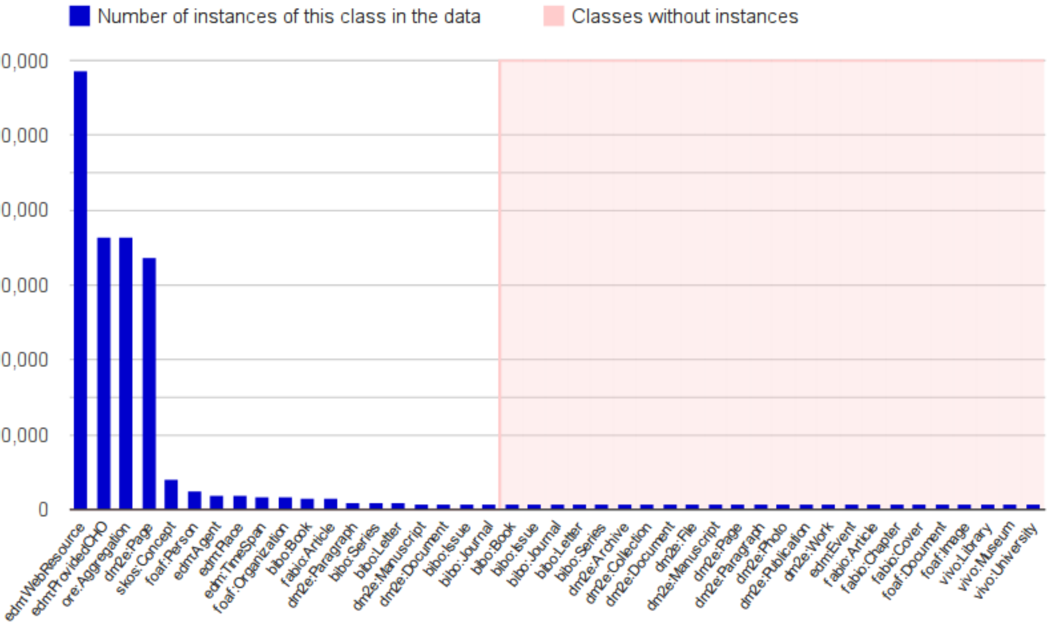

FIG. 2: Distribution of classes across datasets in DM2E.

https://doi.org/10.23106/dcmi.952136385

The most frequently used classes (as shown in fig. 2) are *edm:WebResource* (every CHO must point to at least one Web resource), followed by *ore:Aggregation* and *edm:ProvidedCHO*. They occur equally often, as there is always one aggregation per CHO, and are mandatory. Although contextual classes are not mandatory and less frequently mapped, they are very useful as they allow contextual data to become Linked Data representations with dereferenceable IRIs[8] as opposed to mere strings. The class *skos:Concept* (the fifth most mapped class) is used very unevenly: DP V-Dataset 7 uses it 138,440 times, DP I-Dataset 1, DP III-Dataset 5 and DP II-Dataset 4 do not use it at all. Subclasses of *foaf:Organization*, e.g. *vivo:Library*, *dm2e:Archive*, *edm:Event* were never used. Altogether, 24 of 43 classes are unused.

The class *dm2e:Page* is used most often as the aggregation level of an object (see table 2). While DM2E prepared for different types and aggregation levels, the data appears to be aggregated almost exclusively on the page level. However, in the mappings, several levels are used. Most datasets make use of two different levels of hierarchy within a CHO. This can not only be explained with the provided metadata. For example, chapters are never mapped but exist in the provided books. Which and how many levels of a hierarchical object are mapped seems to be mostly based on the mandatory elements in the model and on the decisions of the MI.

TABLE 2: Different CHO types (subclasses of *edm:PhysicalThing* or *skos:Concept*).

Dataset	bibo: Series	bibo: Book	dm2e: Manuscript	dm2e: Paragraph	bibo: Journal	bibo: Issue	fabio: Article	bibo: Letter	dm2e: Page
Dataset 1	-	-	24	-	-	-	-	-	10,427
Dataset 2		1,251	10						530,314
Dataset 3	4,552	39,873	-	-	-	-	-	-	-
Dataset 4	-	-	175	-	-	-	-	-	46,006
Dataset 5	-	-	1,012	-	-	-	-	-	307,202
Dataset 6	-	2,916	-	-	-	-	-	-	472,994
Dataset 7	-	1,295	-	-	-	-	-	-	416,172
Dataset 8	-	-	-	-	-	-	-	3,630	34,596
Dataset 9	-	-	-	-	1	346	42,173	-	159,277
Dataset 10	-	-	20	9,635	-	-	-	-	-
Total	4,552	45,335	1,241	9,635	1	346	42,173	3,630	1,976,988

Only few mappers use *edm:Agent* (DP IV-Dataset 6: 2,919; DP II-Dataset 3: 11,796; DP VIII-Dataset 10: 35). In the same datasets where *edm:Agent* is used, *foaf:Organization* and *foaf:Person* are mapped as well. *foaf:Organization* and *foaf:Person* are mapped by everyone. In some datasets, they are rarely mapped (DP I-Dataset 1: 2 organizations, 3 persons and 0 agents; DP II-Dataset 4: 0 agents, 33 organizations, 275 persons), in other datasets they are very often mapped (DP II-Dataset 3: 11,796 agents, 21,592 persons, 175 organizations). Here, it seems that these mappings of agents do not depend on the mapper but on the provided data.

4. Linked Data References vs. Literal Statements

Broadly speaking, an RDF statement can have either a literal (a possibly typed string) or a reference to a resource (an IRI, a blank node or an RDF container type). Since the DM2E model strongly recommends using literals and IRI exclusively, the relationship between statements referring to literals or resources and the total number of statements in a dataset reveals differences in the datasets as can be seen in figure 3. When the datasets are grouped by the percentage of

[8] Internationalized resource identifier. An extension of URI allowing unencoded Unicode characters in most places of a URI (RFC 3987).

https://doi.org/10.23106/dcmi.952136385

literal statements, clusters of similar percentages appear according to the respective MI - independent of the metadata content.

For example, the percentage of literal statements in DP V-Dataset 7 (28.273%) and DP VI-Dataset 8 (28.270%) is almost equal, yet the content is vastly different (collection of digitized prints of various genres and ages vs. personal correspondence of an 19th century scholar), the metadata originally created by different data providers (research project vs. library) and in different formats (TEI vs. EAD). The only commonality between the datasets is that the same organization (MI D) created the mappings to DM2E. Therefore, we put forth the correlation between the ratio of literal statements and the mapping institution is much stronger than between ratio of literal statements and similarity of the original data.

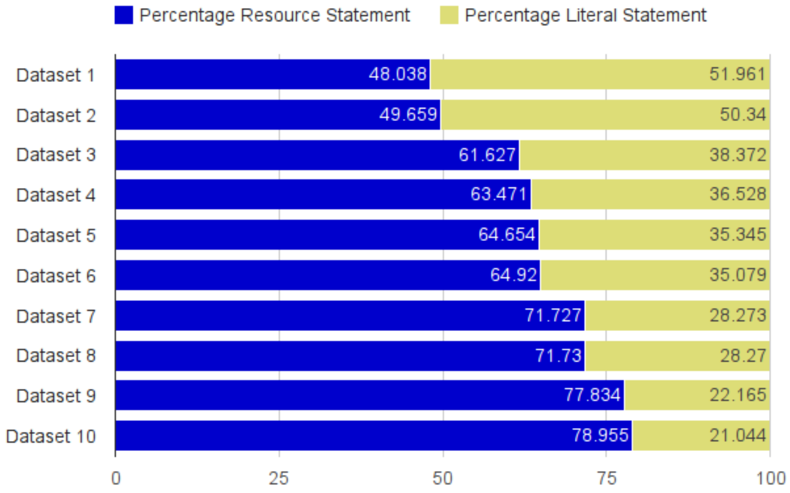

FIG. 3: Ratio of statements with literal statements to resource statements per dataset.

While the relationship between resource and literal statements gives some insight into how MIs structure the data, it does not answer questions pertaining to the quality and usefulness of literal statements. To tackle this problem, the literal statements containing properties with literals allowed as their range were clustered into three groups (see fig. 4). The "preferred" literal statements (properties that are either mandatory, recommended or increase the descriptive content)[9] are a sign of data quality since they enhance the descriptiveness of the data, improve the search and browse experience and granularize textual information. The "neutral" literal statements are those neither preferred nor unwanted, i.e. properties where it is not important for contextual information if they refer to resources or literals. Lastly, the "deprecated" literal statements are statements with those properties that allow both literals and resources in their range, yet the data providers chose to use literals.[10] Even though the label implies it, it is not necessarily a wrong choice to use literals when they are allowed as an alternative to an IRI. However, inconsistent usage is detrimental to the homogeneity of the data, requiring data consumers to use more complex queries to capture both types of statements and are often a sign for poor structure within the data.

As can be seen in figure 4, there is some evidence that the relationship of the number of preferred and deprecated literal statements is correlative with the mapping institution. For

[9] Preferred properties in literal statements: *skos:prefLabel*, *rdfs:label*, *skos:altLabel*, *dc:description*, *dm2e:displayLevel*, *edm:type*, *dc:title*, *dm2e:subtitle*, *dc:language*, *dc:format*, *dc:identifier*.

[10] "Deprecated" properties in literal statements: *dc:rights*, *dcterms:created*, *dcterms:modified*, *dcterms:issued*, *dcterms:temporal*, *rdaGr2:dateOfBirth*, *rdaGr2:dateOfDeath*, *rdaGr2:dateOfEstablishment*, *rdaGr2:dateOfTermination*. The model recommends for time-related properties the use of *edm:TimeSpan* resources but also allows *xsd:dateTime*/*xsd:gYear* or *rdf:Literal*.

https://doi.org/10.23106/dcmi.952136385

example, the data produced with mappings by MI A (Dataset 1 and 2) and MI C (Dataset 5 and 6) is very coherent in this regard. However, for the datasets produced by MI B (Dataset 3 and 4) we see a slight variance, for the datasets produced by MI D (Dataset 7 and 8) even a significant variance in the ratios. Taking the more specific grouping into account, the preferred-deprecated ratio is much more influenced by the original metadata than the overall literal-resource ratio. Considering the data produced by MI D, it is remarkable that the one dataset (Dataset 7) contains the largest proportion of deprecated literal statements within the set of datasets, whereas the other dataset (Dataset 8) contains no deprecated statements at all.

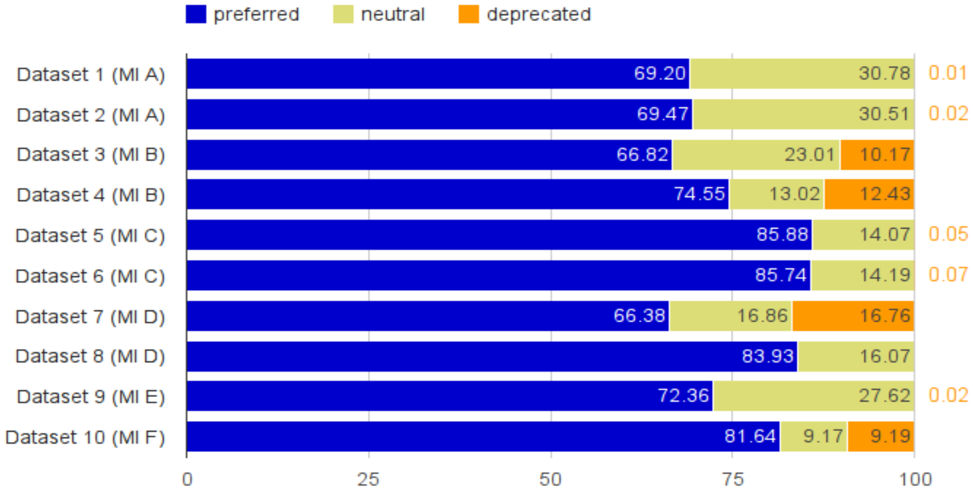

FIG. 4: Distribution of "preferred", "neutral" and "deprecated" literal statements within the datasets.

5. Variance of Statements and Redundancy of Data in Triples

To measure the redundancy of data in triples, we introduce the measure of Predicate-Object-Equality-Ratio (POER-*n*), which is defined as the percentage of triples that share the same predicate and object with at least *n* other statements. In other words, POER-*n* measures how many statements state the same facts about different subjects. The smallest possible POER-*n* of the datasets in DM2E, POER-1, ranges from 0.08% (Dataset 5) to 2.48% (Dataset 3). While impressive as a signifier of structural redundancy, using POER-*n* to assess data-intrinsic redundancy proves to be much more difficult. First of all, there is a lot of duplication required by the triple structure of RDF, i.e. *rdf:type* statements have a limited range of possible values defined by the DM2E model. Certain literal properties have even smaller ranges. Other areas of redundancy can be explained by the original metadata, such as manuscripts being published in the same year or by the same author. Some redundancies, however, can point to problems. For example, redundancies in *dc:subject* statements will, when passing a certain frequency threshold, not be discriminatory for any kind of search (e.g. assigning the keyword "philosophy" to any CHO). Redundant *dc:title* statements can show mapping errors or missing content. For example, if many *dc:title* statements contain the text "Untitled Page" or just a page number, the content may have been mapped incorrectly.

Hence, the usefulness of POER-*n* is very dependent on the value of *n*. Whereas the bulk of the statements contained in POER-1 or even POER-100 can be discarded as arbitrary similarities, a high POER-1000 or POER-10000 cannot be easily explained with random chance. If the same fact is stated about 10,000 different subjects within a dataset, this is a strong indicator that either the original metadata is very homogenic (e.g. by the same author or released in the same year) or that the data is not properly internally aligned (e.g. hundreds of different auto-generated *skos:Concept*s with the same *skos:prefLabel*). Instead of setting *n* to an arbitrary number, a lot can be gained by using the number of instances of certain classes as the threshold, for example, in

https://doi.org/10.23106/dcmi.952136385

the case of DM2E, the number of *ore:Aggregation*/*edm:ProvidedCHO* instances. The exact mechanics of how to fine-tune POER-*n*, finding proper threshold values and visualizing both, the POER-*n* and the statements it represents, is still subject to further research.

Figure 5 presents the average number of statements per instance of a class within a dataset (ANOS). We see that the data mapped by MI C is very homogenous with regards to the ANOS, for both *ore:Aggregation*/*edm:ProvidedCHO* and contextual classes. Obviously, the workflow for the RDFization of the original data used by MI C is organized in such a way (e.g. by reusing the same XSLT scripts) that the resulting RDF follows a relatively rigid structure.

For the *edm:ProvidedCHO* instances, we see a significant higher ANOS for data mapped by MI D. Since the data is generated from very different input formats, the deciding factor here is apparently MI D's thorough mapping process, producing more statements by normalizing unstructured fields, adding alternative titles, different languages etc.

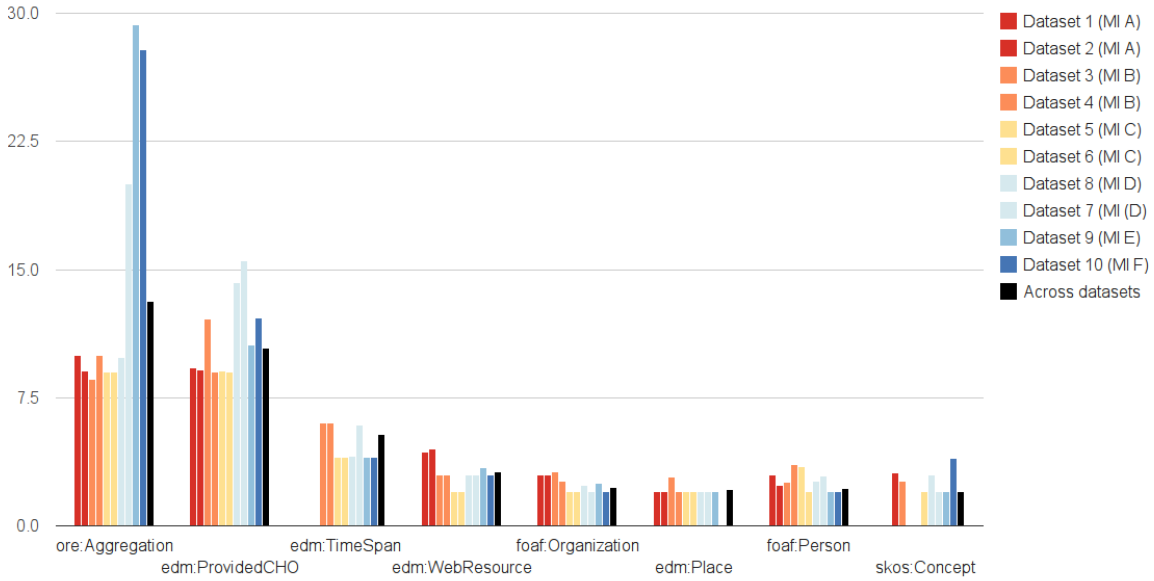

FIG. 5: Average number of statements per class per dataset.

The three outliers with significantly more-than-average ANOS for *ore:Aggregation*s are all generated from TEI data. Apparently, TEI's exhaustive mechanics for adding metadata to the header of a TEI document heavily and positively influences the richness of the metadata on aggregation level. While still slightly above average, the ANOS for *edm:ProvidedCHO* from TEI data is much lower than for *ore:Aggregation*, leading to the conclusion that TEI is a top-heavy format, inciting TEI producers to create exhaustive meta-metadata describing the provenance of the TEI document rather than the manuscript itself.

Looking at the distribution of ANOS for *edm:WebResource* instances, clusters of very similar ANOS defined by the respective MI emerge. The explanation for this is that most information assigned to *edm:WebResource* instances is boilerplate (format and rights information mostly) with only the IRI of the *edm:WebResource* instance itself changing.

In general, the distribution of ANOS across datasets is more homogenous for contextual classes (*foaf:Person*, *foaf:Organization*, *edm:Place*, *edm:TimeSpan*, *skos:Concept*) than for manuscript-related classes (*ore:Aggregation*, *edm:ProvidedCHO*). The main reason for this is that ANOS for the former is significantly smaller than for the latter, i.e. relatively few statements are asserted about instances of contextual classes (the highest ANOS for contextual classes is 3.96 for *skos:Concept* in Dataset 10). On the other hand, this is also a sign that there is still potential for possible improvement on account that, e.g. digitization projects focusing on the written legacies of individuals tend to have extensive dossiers about the context (like places, persons and concepts). Apparently, the full richness of this data is not yet fully ported over to the RDFized data.

6. Being Linked Open Data - Usage of different Ontologies

The Linked Data principles recommend using existing namespaces and ontologies. The DM2E model included a number of other ontologies and encouraged data providers to map their data using properties from them. Figure 6 shows the ontologies and their number of properties referenced by the DM2E model as well the number of properties used by data providers.

Every ontology is used, however, not all properties are used: of DM2E, slightly more than 50% of the offered properties are used, around 66% of EDM. Most of the properties of the DC and BIBO ontologies are used (75%). Vocabularies like DC and DCTerms have fewer resources in the model than DM2E but they are more often used. Other ontologies like rdaGr2 provide very specific properties for very specific contextual classes which are also often not mapped (e.g. the already mentioned *rdaGr2:dateOfEstablishment*). Even though the two CIDOC-CRM properties in the model, *crm:P79F.beginning_is_qualified_by* and *crm:P80F.end_is_qualified_by*, are also very specific, they serve an important case: they are used to indicate how accurate a timespan is.

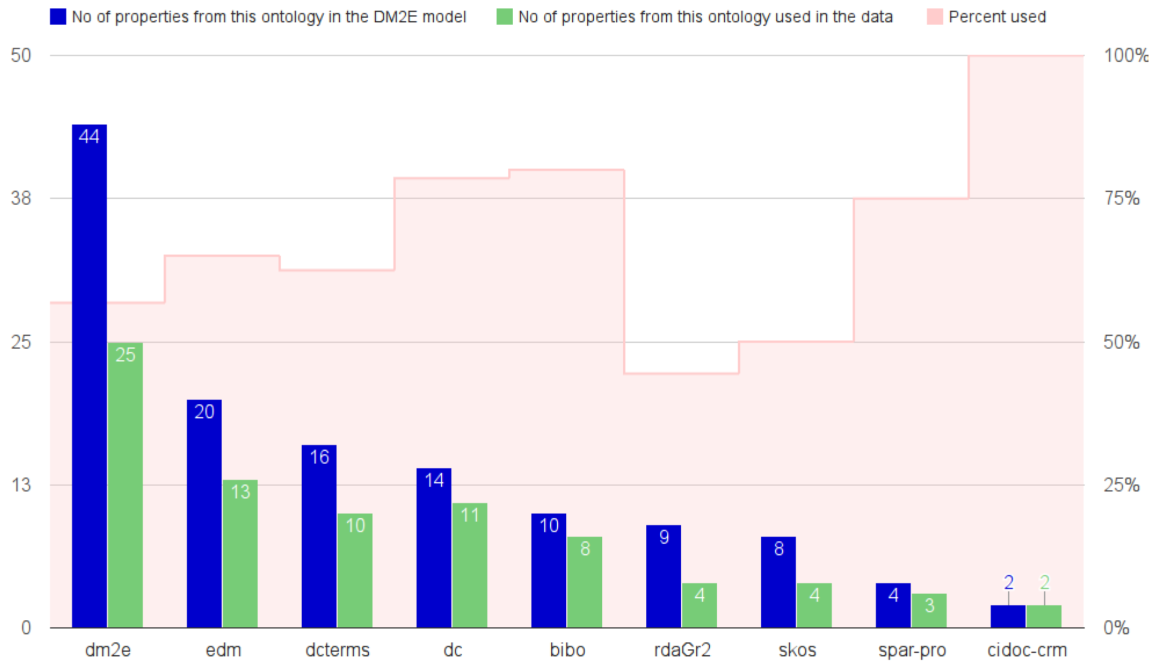

FIG. 6: Number of properties defined in the DM2E model vs. number of properties actually used in the data, by referenced ontology.

The fact that only half the properties defined in the DM2E model are actually used (see also fig. 1) deserves closer scrutiny, however. Because the ontology is being developed by DM2E for DM2E, this cannot be explained with the specificity of the domain of the ontology, but with the dynamics of the process of ontology development: In the early stages, the intricate knowledge of data providers about the details of their data led them to require increasingly semantically narrow properties from the DM2E ontology engineers (e.g. *dm2e:honoree* or *dm2e:wasStudiedBy*). However, when the MI (which do not necessarily coincide with the DP, see table 1) started implementing the mappings, many of those requirements were dropped due to the specific properties being hard to map or not being readily discernible from the original metadata. Over the

https://doi.org/10.23106/dcmi.952136385

course of many cycles of mapping, data ingestion and refinement of the data model, new properties have been added but unused properties were never dropped.

7. Conclusion: Linked Data Mapping Cultures

The analyses have shown that the particular mapping institution plays an important role in the way that data actually is represented after a mapping process. Datasets mapped by the same MIs have similar characteristics in the various analyzed aspects, e.g. which resources are used for the mappings and which are not. The representation of the data before the mapping has a less significant influence on the structure of the mapped data as has the domain or CHO types. The source format is reflected in the number of provided statements, e.g. whenever TEI is used (where the full text of an object is also annotated and can be used for mappings), many more statements are produced.

As already identified in previous model evaluations, mapping institutions do not make use of the full range of possible ontology elements that could be mapped. Models, including the DM2E model, could be reduced (especially when only a small percentage of specific vocabularies is used as shown in the last figure). Contextual resources are not mapped as thoroughly as the core classes for the representation of the object (*edm:ProvidedCHO*) and its metadata record (*ore:Aggregation*).

From a user's perspective, the Linked Data representation should be derived from the source data by a function of the source data and not strongly be influenced by the specifics of the mapping process. While technical means such as the quantitative analyses presented here help make the skew more evident, it can eventually only be rectified by a more agile development process that involves all stakeholders balancing semantic expressivity with data interoperability, peer-review of mappings or ongoing evaluation of mappings and mapped data, improved and extended mapping guidelines with a strong focus on reusability and sustainability of data and data model. From a Linked Data mapping cultural perspective, our conclusion is that ontologies should not just be extended to fit new requirements but also pruned from over-specific bloat regularly and that this can only be achieved when ontologists, data providers, mapping institutions, developers and data consumers incessantly communicate, compromising between semantic accuracy and technical feasibility.

Acknowledgements

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

Alexander, Keith, Richard Cyganiak, Michael Hausenblas, and Jun Zhao. (2009). Describing Linked Datasets. On the Design and Usage of VoID, the “Vocabulary of Interlinked Datasets”. In Bizer et al. (Eds.), Proceedings of the Linked Data on the Web Workshop (LDOW2009), Madrid, Spain, April 20, 2009, CEUR Workshop Proceedings. Retrieved, May 14, 2014, from http://ceur-ws.org/Vol-538/.

Auer, Sören, Jan Demter, Michael Martin, and Jens Lehmann. (2012). LODStats – An Extensible Framework for High-Performance Dataset Analytics. In ten Teije et al. (Eds.), Knowledge Engineering and Knowledge Management. 18th International Conference, EKAW 2012, Galway City, Ireland, October 8-12, 2012, Proceedings (pp. 356-362). Berlin, Heidelberg: Springer. doi: 10.1007/978-3-642-33876-2.

Carroll, J. Carroll, Christian Bizer, Pat Hayes, and Patrick Stickler. (2005). Named Graphs. In Journal of Web Semantics, 3, 247-267.

Dröge, Evelyn, Julia Iwanowa, and Steffen Hennicke. (2014a). 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. Retrieved, July, 24, 2014, from http://ozk.unizd.hr/proceedings/index.php/lida/article/view/117.

Dröge, Evelyn, Julia Iwanowa, Steffen Hennicke and Kai Eckert. (2014b, March). DM2E Model V1.1 Retrieved, May 12, 2014, from http://pro.europeana.eu/documents/1044284/0/DM2E+Model+V+1.1+Specification.

https://doi.org/10.23106/dcmi.952136385

Europeana Data Model Primer, v14/07/2013. (2013, July). Retrieved from: Europeana Professional website. Retrieved, April 28, 2014, from http://pro.europeana.eu/ documents/900548/770bdb58-c60e-4beb-a687-874639312ba5.

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

Klimek, Jakub, Jirí Helmich, and Martin Necasky. (2014). An analysis supported by numerous visualizations Application of the Linked Data Visualization Model on Real World Data from the Czech LOD Cloud. Linked Data on the Web (LDOW 2014) Workshop. Retrieved, May 14, 2014, from http://events.linkeddata.org/ldow2014/papers/ldow2014_paper_13.pdf.

Palavitsinis, Nikos, Nikos Manouselis, and Salvador Sanchez-Alonso. (2014). Metadata quality in digital repositories: Empirical results from the cross-domain transfer of a quality assurance process. Journal of the Association for Information Science and Technology. doi: 10.1002/asi.23045.

Seiffert, Florian. (2001). Eine Analyse der Verbunddaten des HBZ. ABI-technik 21(2): 125-146.

Smith-Yoshimura, Karen, Catherine Argus, Timothy J. Dickey, Chew Chiat Naun, Lisa Rowlison de Ortiz, Hugh Taylor. (2010, March). Implications of MARC Tag Usage on Library Metadata Practices, OCLC Online Computer Library Center, Inc. Retrieved, May 14, 2014, from http://www.oclc.org/research/publications/library/2010/2010-06.pdf.

https://doi.org/10.23106/dcmi.952136385