DDB-KG: The German Bibliographic Heritage in a Knowledge Graph

Mary Ann Tan[1,2], Tabea Tietz[1,2], Oleksandra Bruns[1,2], Jonas Oppenlaender[1,2], Danilo Dessí[1,2] and Harald Sack[1,2]

[1]*FIZ Karlsruhe – Leibniz Institute for Information Infrastructure, Germany*
[2]*Karlsruhe Institute of Technology, Institute AIFB, Germany*

Abstract

Under the German government's initiative "NEUSTART Kultur", the German Digital Library or *Deutsche Digitale Bibliothek* (DDB) is undergoing improvements to enhance user-experience. As an initial step, emphasis is placed on creating a knowledge graph from the bibliographic record collection of the DDB. This paper discusses the challenges facing the DDB in terms of retrieval and the solutions in addressing them. In particular, limitations of the current data model or ontology to represent bibliographic metadata is analyzed through concrete examples. This study presents the complete ontological mapping from DDB-Europeana Data Model (DDB-EDM) to FaBiO, and a prototype of the DDB-KG made available as a SPARQL endpoint. The suitabiliy of the target ontology is demonstrated with SPARQL queries formulated from competency questions.

Keywords

Cultural Heritage, Digital Library, Ontology, Knowledge Graph

1. Introduction

The German Digital Library or *Deutsche Digitale Bibliothek*[1] was officially launched in 2014, not only to serve as Germany's contribution to the Europeana[2] project, but also to make the country's rich cultural heritage (CH) available to a much broader audience online. The DDB's collection currently includes close to 38M cultural heritage objects (CHOs) or artifacts in the form of metadata, submitted by around 500 providers hailing from the GLAM sectors (Galleries, Libraries, Archives, Museums), cultural sites and research institutions. The DDB has fulfilled its first goal as a national aggregator. However, as a cultural heritage portal, it failed to entice the general public due to imprecise search results [1]. Nevertheless, the DDB has been included in the German government's initiative "NEUSTART Kultur"[3] to improve the country's cultural infrastructure.

HistoInformatics 2021 – 6th International Workshop on Computational History, September 30, 2021, online
{ann.tan | firstname.lastname}@fiz-karlsruhe.de (M. A. Tan)
http://www.fiz-karlsruhe.de/de/forschung/information-service-engineering (M. A. Tan)
0000-0003-3634-3550 (M. A. Tan); 0000-0002-1648-1684 (T. Tietz); 0000-0002-8501-6700 (O. Bruns); 0000-0002-2342-1540 (J. Oppenlaender); 0000-0003-3843-3285 (D. Dessí); 0000-0001-7069-9804 (H. Sack)

[1]Deutsche Digitale Bibliothek, https://www.deutsche-digitale-bibliothek.de
[2]Europeana, https://www.europeana.eu
[3]NEUSTART Kultur, https://bit.ly/3CKD53a

CH portals are faced with several challenges: they have to provide representations and storage, and enable the discovery and retrieval of a huge number of complex artifacts. They are faced with scalability and veracity issues that often hound big data systems. Fortunately, some of these challenges can be addressed by the adoption of Semantic Web (SW) technologies and Linked Open Data (LOD). Particularly in enhancing user-experience, construction of a Knowledge Graph (KG) from historical collections has been shown to be quite effective [2]. A KG for the DDB can enhance the search and retrieval functions of the current portal by facilitating semantic search and exploration [3]. To address the above mentioned challenges and undertake the construction of a KG, this paper initially sets its focus on the metadata coming from the library sector and outlines the following contributions:

- A complete mapping[4] of bibliographic metadata from DDB-EDM to FaBiO [4] to address the limitations of the current model.
- A DDB-KG[5] prototype for the library sector.
- Proof of concept through sample competency questions (CQ) and their corresponding SPARQL queries.

This paper is organized as follows. Section 2 reviews other CH portals, their implementations and exploration techniques. Section 3 presents the current state of and hurdles in the DDB, including an in-depth analysis of the DDB's data model and dataset. Section 4 explores the proposed changes to the current implementation, the initial steps in constructing the DDB-KG and a selection of SPARQL queries to illustrate the addressed challenges. Finally, Section 5 presents conclusions from this study and discusses succeeding work.

2. Related Work

The Semantic Web, as envisioned by Tim Berners-Lee, is an enhancement of the World Wide Web (WWW), in which web content is made intelligible to software agents for the purpose of data sharing, discovery, integration and reuse [5]. To make machines understand real-world knowledge, ontologies aim to formally define concepts. In an ontology, domain-specific constraints are encoded together with the data, for example, when data is structured by a graph. The application of knowledge graphs in large scale enterprises has shown that they are ideal for representation, storage, and exploration of data from diverse domains [6].

Hyvönen [7] enumerated the advantages of publishing CH content using semantic web standards: (1) global view of heterogeneous and distributed content, (2) automatic content aggregation, (3) semantic search, (4) exploratory search. These qualities are equally beneficial to both end-users and content-providers.

To this end, the authors advocated the use of ontologies and LOD. CultureSampo is a cross-domain CH portal which makes use of an ontology architecture model called FinnOnto [8], where domain-specific ontologies were made interoperable through the existence of an upper ontology. By utilizing a knowledge graph, their approach allowed users to explore their collection via a thematic graphical interface.

[4]DDB-EDM to FaBiO Mapping, https://bit.ly/3qBxxCo
[5]DDB-KG SPARQL Endpoint, http://ddbkg.fiz-karlsruhe.de

Similarly, ArCO [9] used a network of ontologies to represent Italy's cultural heritage through catalogue records. By identifying the activities involved in cataloging and preservation of CHOs,[6] they were able to formulate modular ontologies with the aid of design patterns.

refer [2] is a semantic annotation and exploration tool for textual content in wordpress platforms.[7] The tool allows to (semi-)automatically annotate text with DBpedia entities. The annotated text can be visualized and explored by means of a relation browser. This navigation interface enables a serendipitous exploration of the entire content of the platform, which is especially useful for platforms focused on cultural heritage data [10].

These CH portals leverage the expressive power of ontologies in representing knowledge and the effectiveness of knowledge graphs in accumulating and conveying knowledge. This study share the same domain and goals as the aforementioned ones. By virtue of these similarities, adopting similar approaches to address corresponding challenges are expected to yield the same benefits.

3. The German Digital Library

Currently, the DDB contains approximately 38M metadata of CHOs from seven different sectors: GLAM, multimedia libraries, research institutions, cultural sites and other public institutions. This chart[8] illustrates the proportion of each sector's contribution. Sectors with only a minimal fraction of CHOs as compared to the others are not visible in this illustration. Digitized objects from the library sector make up a fifth of the entire DDB collection and amount to around 8M objects. Thus, this study will focus on the library sector. The subsequent sections discuss the challenges that influence the current limitations of the DDB. In addition, the DDB-EDM is analyzed with respect to these challenges.

3.1. Challenges

Current issues facing the DDB can be illustrated with a sample search. This sample is representative of the usual search conducted by regular users. When searching for "Schillers Räuber",[9] a flood of vaguely-related results inundate the search page. The top search results point towards physical copies of a 19th century commentary on "*Die Räuber*", followed by a collection of images from a live performance of the play by the Baden State Theater. Editions of the text only show up on the third page of the search results. Upon closer inspection, the retrieved objects are heterogeneous i.e., objects originating from libraries, museums, archives are mixed together. The results also show a high granularity, results for different acts of the tragedy are displayed separately, and not grouped together, as one would expect when searching a digital library. At the very least, these results would need further processing; in their current state, they just reveal the challenges facing CH portals: heterogeneity, high granularity, representational

[6] In ArCO, the term "property" refers to the cultural heritage object.

[7] Example: http://scihi.org/

[8] Pie Chart of the DDB's Collection,https://git.io/J0J0B

[9] "*Die Räuber*" or "The Robbers" is Friedrich Schiller's first play. Link to sample search for "Schillers Räuber", https://bit.ly/3m7aoHY

complexity, and volume. In addition, the current implementation does not distinguish between copies, editions or issues of the same publication.

Objects from libraries have complex bibliographic relationships [11] e.g., a book of illustrations based on a text carrying the same title, an adaptation of a novel as a screenplay, editions of the same publication. Unlike in museums where all objects are considered unique, books having the same creative content may have been translated into several languages, or translated by several scholars into the same language at different times, re-issued due to printing mistakes, bundled in a series, and so on. These relationships need to be made clearly distinct before refined search results become even possible.

The above-mentioned challenges can be addressed by modeling heterogeneous library objects with a domain-specific ontology suitable for the application profile. The choice of ontology must consider the following requirements:

1. The ontology must be able to adequately represent and clearly distinguish between the different types of CHOs while maintaining interoperability.
2. It must be possible to organize the objects in a hierarchical manner while simultaneously abstracting their granularity for better retrieval, e.g., only return the cover page of a book rather than its multiple chapters.
3. Bibliographic relationships must be taken into account.
4. Bibliographic information must be encoded into non-generic properties.
5. Linked open data must be prioritized over literals and decentralized controlled vocabulary.

3.2. DDB-EDM

As one of Europeana's national aggregators, cultural and scientific objects in the DDB are modeled using an extension of the EDM called DDB-EDM. The core design principles of the EDM favor flexible and simple CHO representations. The aggregators of Europeana are empowered to adapt their choice of metadata element sets [12]. To foster interoperability in EDM, all CHOs are instances of the class *edm:ProvidedCHO*, regardless of the objects' classification according to UNESCO:[10] tangible vs intangible, or movable vs immovable. A movable object may have different locations through the course of its lifetime, but this is not the case with cultural sites. Characteristics of tangible objects may vary depending on their specific types: a monograph and a rare handwritten manuscript are both tangible objects, however, the former may have been re-issued while the latter may be unique. For these reasons, the DDB-EDM does not fulfill the 1st requirement outlined in Section 3.1.

Object properties, such as *edm:hasType* or *dc:type*, are used to indicate classification. However, the concepts encoded in these properties are often inconsistent. Assigned values belong to several conceptual equivalences: document type, document structure, production process, purpose, manifestation, and subject headings. Hence, these properties are still insufficient for conveying the semantics corresponding to an object's classification.

The hierarchical arrangement of CHOs from the library sector is modelled by the data properties *ddb:aggregationEntity*, *ddb:hierarchyType* and *ddb:hierarchyPosition*. Incidentally, the values assigned to *ddb:hierarchyType* often coincide with the values assigned to object properties

[10]What is meant by "cultural heritage"? (UNESCO), https://bit.ly/2VPOFZR

meant for type classification. As for granularity, the whole-part object property *dcterms:isPartOf* is being utilized. Because all objects belong to *edm:ProvidedCHO* regardless of their position in the hierarchy, parts of an object are treated equally during retrieval time.

To indicate relationships between objects, the DDB-EDM makes use of the following object and data properties: *dc:relation*, *dcterms:isReferencedBy*, *dcterms:isReplacedBy*, and *dcterms:isRequiredBy*. However, as it can be seen from Tillet's illustration of bibliographic relationships [11], the existing attributes are not sufficient. Being able to represent distinct bibliographic relationships facilitates the discovery of links among objects during exploration. Using *dcterms:isReferencedBy* to indicate the relationship between the original play and (1) a commentary, (2) an illustration, (3) an adaptation of the play for a theater season, or (4) a collection of plays, is insufficient to convey the underlying semantics that exist in these distinct bibliographical relationships. Moreover, details that are specific to bibliographic metadata, such as edition, volume number, and formats, are assigned to generic properties *dc:description* and *dc|dcterms:extent*. The same is true for bibliographic identifiers such as ISBN, call number, OCLC[11] identifiers, etc. Different identifiers are stored in *dc:identifier*. With such generic properties, crucial information that provides the necessary context for understanding specific bibliographic objects cannot be properly encoded.

Finally, for the purpose of seamless integration to LOD, the DDB extended EDM by using Dublin Core Metadata Terms (dcterms) in conjunction with Dublin Core Element Set (dc). However, because the EDM does not restrict its aggregators in using the same range of values for most of the data and object properties, data providers of the DDB are assigning a mixture of literals, controlled vocabulary and Uniform Resource Identifiers (URIs) from external sources.

4. DDB-KG

As mentioned in section 2, ontologies are necessary as well as crucial to the encoding domain-specific knowledge in graph-structured data. Existing ontologies can be leveraged to adequately model different types of CHOs. Taking previous studies as a guide, a middle-out approach to ontology mapping is selected to find the most suitable ontology for each of the sectors. Due to the limitations of the DDB-EDM, there is a need to find alternatives suitable for publishing and retrieval of bibliographic metadata. The subsequent subsections present the mapping from DDB-EDM to FaBiO, construction of the DDB-KG, and a selection of competency questions with SPARQL queries to illustrate the advantages of adapting FaBiO.

4.1. DDB-EDM to FaBiO

To tackle the above-mentioned usability challenges, DDB-EDM is mapped to FaBiO, an ontology for describing bibliographic records and modelling the bibliographic universe. FaBiO is selected amongst all the other data models because (1) it is aligned with the Functional Requirements for Bibliographic Records (FRBR), (2) it is intended for the publication of bibliographic metadata, and (3) it has an extensive and hierachical class structure. FaBiO's alignment with FRBR enables the representation, storage and retrieval of objects on several abstraction levels. At the highest level,

[11]Online Computer Library Center, http://www.oclc.org

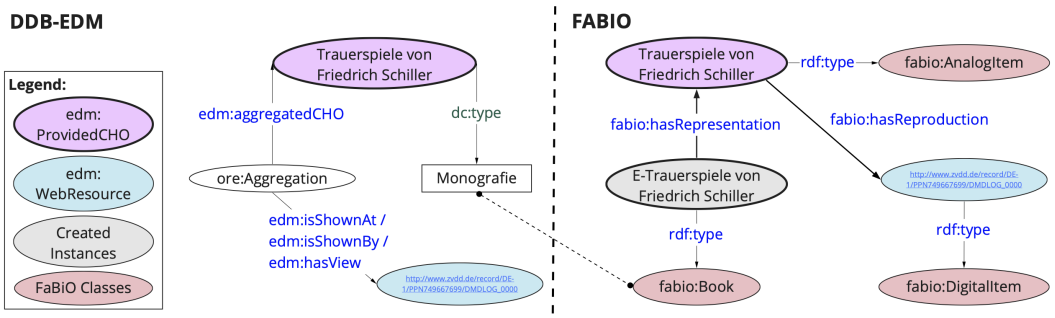

Figure 1: Alignment from DDB-EDM to FaBiO

the **Work** (*fabio:Work*) captures the "the essence of a distinct intellectual or artistic creation" [4]. Works are linked via *fabio:hasRealization* to one or more Expressions (*fabio:Expression*). An **Expression** is the form of a Work when its "content is 'realized' in physical or electronic form" [4]. A **Manifestation** (*fabio:Manifestation*) is a an embodiment (*fabio:hasEmbodiment*) of an Expression and describes the particular format in which the Expression is stored (identified, for instance, by an ISBN). Last, an **Item** (*fabio:Item*) is a particular physical (or electronic) copy as found, for instance, on the shelves of a library. Manifestations are linked to their items through *fabio:hasExemplar*.

FaBiO defines a set of hierarchical sub-classes under each of the endeavors which allows for retrieval of heterogeneous objects. The four-level abstraction also can be used to simplify the retrieval of CH objects. For instance, Items can be omitted if the user is only interested in learning about the abstract Work. In addition, such omissions are especially important in the context of libraries with rare manuscripts where the representations only exist on the Work and Item levels. FaBiO allows these omissions by defining additional bibliographic relationships between non-adjacent abstractions levels: (1) *fabio:hasManifestation* between Work and Manifestation, (2) *fabio:has Portrayal* between Work and Item, (3) *fabio:hasRepresentation* between Expression and Item.

Each instance of *edm:ProvidedCHO* is assigned to *fabio:AnalogItem*, while its digital representations are considered digital reproductions. Figure 1 shows an example using "*Trauerspiele von Friedrich Schiller*" (*TvFS*).[12] Subsequently, individuals are created for each of the abstraction levels according to the original metadata attributes of the CHO (See gray ellipse in Figure 1). A *Work* instance is created when the *dc:title* is considered a Work (*Werk*) in the German Integrated Authority File aka *Gemeinsame Normdatei* (GND).[13] An *Expression* instance is always created by default, while a *Manifestation* instance is created when a publishing event is specified in the metadata. When the CHO is not determined to be a *Work* and is part of another CHO, indicated by *dcterms:isPartOf*, no other instances are created other than *fabio:AnalogItem* and *fabio:DigitalItem*. An example of this would be a chapter or section of a book that is considered as a separate instance of *edm:ProvidedCHO* in the DDB. However, when the title of this object is determined to be a *Work* in the GND, then the primary object is considered as a *fabio:WorkCollection*.

[12]Friedrich Schiller's Dramas, https://bit.ly/365OBHz
[13]GND is maintained by the German National Library, https://bit.ly/2SHmATp

In order to address the 1st modeling requirement mentioned in section 3.1, instances of the CHOs representation in other abstraction levels are assigned to FaBiO Endeavor sub-classes based on the object type terms encoded in *dc:type* and *edm:hasType*. Since the object type terms refer to several concepts, sub-class assignment guided by these considerations: (1) An object having terms that refer to document types is assigned to FaBiO sub-classes on the Work and Expression levels; (2) object type terms that refer to production process and manifestation determine the sub-classes under *fabio:Manifestation*; (3) a term without corresponding FaBiO sub-class, use the default superclass per abstraction level; and (4) it is possible for a term to have a union of multiple sub-classes.

4.2. Object and Data Properties

In addition to *dc* and *dcterms*, the DDB-KG also makes use of the following namespaces to encode the CHO attributes and provenance information: Friend-Of-A-Friend (foaf) to model agents, BIBliographic Ontology (*bibo*) and BIBFrame (*bf*) to encode bibliographic information mentioned in Section 3.1, and PROVenance Ontology (prov-o). Mapping of the object and data properties from DDB-EDM to the DDB-KG triples is presented in this Google Sheet[4] entitled "WEMI Fields". The complete mapping is also saved in a JSON file and hosted on GitHub.[14]

Attributes of the CHOs original metadata encoded in *edm:ProvidedCHO* are copied to the instances created on the different abstraction levels according to the specifications of the Europeana [13]. For instance, *dcterms:publisher* and *dcterms:issued* are copied from on Item instances up to the Manifestation instances, while *dcterms:title* appears in all abstraction levels.

4.3. Querying the KG

The modeling proposed in the preceding section results in the first version of the DDB-KG that covers randomly selected bibliographic metadata. The KG includes 2.06M RDF triples based on 22K library objects that were extracted from the DDB.

Utilizing FaBiO enables a more refined search, e.g. when looking for *"Die Räuber"* the user is able to specify what output he/she expects via indicating the level of abstraction. Thus, if the user is only interested in obtaining a list of Friedrich Schiller's works, the results only include instances of *fabio:Work* associated with the author (see Figure 2). However, if the user is interested in e.g., translations of *"Die Räuber"* or a screenplay based on the work, the library objects from *fabio:Expression* are required (see Figure 3).

The DDB-KG supports mapping of works and agents to external resources, e.g. Integrated Authority File (GND) [14] via *owl:sameAs*. Such links extend and enrich the knowledge about Germany's cultural and scientific heritage. For example, the KG is able to answer the following questions: *What are the works of Thomas Mann?, How many languages was Johann Wolfgang von Goethe's Faust translated into?, How many books were published by publishers located in Leipzig?*, etc.[15]

[14]Mapping guide in JSON, https://git.io/JRpP4

[15]See more example SPARQL queries on Github, https://ise-fizkarlsruhe.github.io/ddbkg/docs/examples/

```
#For PREFIX definitions, refer to additional sample SPARQL queries
linked in the footnote.
SELECT DISTINCT ?title ?work ?gnd ?author
WHERE {
    ?work rdf:type fabio:Work ;
        dcterms:creator ?author ;
        dcterms:title ?title ;
        owl:sameAs ?gnd .
    ?author foaf:name ?author_name
        FILTER (regex(str(?author_name), "Schiller, Friedrich", "i")
        || regex(str(?author_name), "Friedrich Schiller", "i")) .
}
```

Figure 2: The SPARQL query to get a list of Friedrich Schiller's Works.

```
SELECT DISTINCT ?title ?ddbitem ?type
WHERE {
  ?work rdf:type fabio:Work ;
        fabio:hasRealization ?expression .
  ?expression fabio:hasRepresentation ?ddbitem .
  ?ddbitem dcterms:type ?type ;
        dcterms:title ?title
        FILTER regex(str(?title), "Die Räuber", "i") .
}
```

Figure 3: The SPARQL query to search for all expressions of Friedrich Schiller's "The Robbers".

5. Conclusion and Future Work

In a digital library, efficient retrieval and seamless exploration of its holdings are crucial to increase uptake in the community. By adopting SW technologies in the DDB, the current challenges can be overcome to improve user-experience. In particular, a suitable ontology is required to encode the complexity and semantics of Germany's cultural heritage from the library sector. Using FaBiO to represent bibliographic metadata, the following benefits are expected: (1) distinction between different types of object, (2) representation of bibliographic relationships through FRBR entities and relationships, (3) encoding of bibliographic details (edition, volume, call number) into non-generic object and data properties and (4) linkage to the continuously expanding LOD world. The advantages of adapting a suitable ontology and graph-structured data especially in avoiding information overload are exemplified by comparing the results of the initial sample search in the DDB and the SPARQL queries.

At present, the DDB portal employs a keyword-based search and retrieval function. The DDB-KG is intended to supplement the current search features of the portal by employing query expansion through SPARQL queries. Future work will assess retrieval efficiency when querying specific endeavors or levels of abstraction. Moreover, the process of restructuring the DDB requires further analysis of metadata from the other sectors, and adoption or creation of ontolo-

gies to adequately model their intricacies. Simultaneously, efforts will be targeted at defining a top-level ontology to describe cross-sector relationships for the purpose of interoperability. The re-structuring of the DDB into a KG will provide the much needed support for the organization, access, and exploration of CHOs, thereby fostering a better understanding of German cultural history.

References

[1] D. Fiene, Ein Millionenprojekt ist gescheitert, [online], 2014. https://bit.ly/3sj53yq.

[2] T. Tietz, J. Jäger, J. Waitelonis, H. Sack, Semantic annotation and information visualization for blogposts with refer., in: VOILA@ ISWC, 2016, pp. 28–40.

[3] M. A. Tan, T. Tietz, O. Bruns, J. Oppenlaender, D. Dessì, H. Sack, DDB-EDM to FaBiO: The Case of the German Digital Library, The Semantic Web – ISWC 2021 (2021).

[4] S. Peroni, D. Shotton, FaBiO and CiTO: Ontologies for describing bibliographic resources and citations, Journal of Web Semantics 17 (2012) 33–43.

[5] T. Berners-Lee, J. Hendler, O. Lassila, The semantic web: A new form of web content that is meaningful to computers will unleash a revolution of new possibilities, ScientificAmerican.com (2001).

[6] N. Noy, Y. Gao, A. Jain, A. Narayanan, A. Patterson, J. Taylor, Industry-scale knowledge graphs: Lessons and challenges, Commun. ACM 62 (2019) 36–43. URL: https://doi.org/10.1145/3331166. doi:10.1145/3331166.

[7] E. Hyvönen, Semantic Portals for Cultural Heritage, in: S. Staab, R. Studer (Eds.), Handbook on Ontologies, International Handbooks on Information Systems, Springer, 2009, pp. 757–778. doi:10.1007/978-3-540-92673-3.

[8] E. Hyvönen, K. Viljanen, J. Tuominen, K. Seppälä, Building a national semantic web ontology and ontology service infrastructure –the finnonto approach, in: S. Bechhofer, M. Hauswirth, J. Hoffmann, M. Koubarakis (Eds.), The Semantic Web: Research and Applications, Springer Berlin Heidelberg, Berlin, Heidelberg, 2008, pp. 95–109.

[9] V. A. Carriero, A. Gangemi, M. L. Mancinelli, L. Marinucci, A. G. Nuzzolese, V. Presutti, C. Veninata, ArCo: The Italian Cultural Heritage Knowledge Graph, The Semantic Web – ISWC 2019 (2019) 36––52. URL: http://dx.doi.org/10.1007/978-3-030-30796-7_3. doi:10.1007/978-3-030-30796-7_3.

[10] F. Windhager, P. Federico, G. Schreder, K. Glinka, M. Dörk, S. Miksch, E. Mayr, Visualization of cultural heritage collection data: State of the art and future challenges, IEEE transactions on visualization and computer graphics 25 (2018) 2311–2330.

[11] B. Tillet, What is FRBR?: A Conceptual Model for the Bibliographic Universe, 2004.

[12] S. Peroni, F. Tomasi, F. Vitali, Reflecting on the Europeana Data Model, in: IRCDL 2012, 2012, pp. 228–240.

[13] A. Angjeli, M. Bayerische, et al., D5.1 Report on the alignment of library metadata with the European Data Model (EDM) Version 2.0., Technical Report, Europeana, 2012.

[14] R. Behrens-Neumann, B. Pfeifer, Die gemeinsame normdatei—ein kooperationsprojekt, Dialog mit Bibliotheken 1 (2011) 37–40.