DELIVERABLE

Project Acronym: DM2E

Grant Agreement number: ICT-PSP-297274

Project Title: Digitised Manuscripts to Europeana

D2.1 - Initial Version of the Interoperability Infrastructure

Revision: 1.0

Authors:

Evelyn Dröge (UBER)
Steffen Hennicke (UBER)
Julia Iwanowa (UBER)
Konstantin Baierer (EL)
Hannes Mühleisen (FUB)
Christian Bizer (FUB)
Nasos Drosopoulos (NTUA)
Dirk Wintergrün (MPIWG)
Klaus Thoden (MPIWG)

Project co-funded by the European Commission within the ICT Policy Support Programme		
Dissemination Level		
PP	Restricted to other programme participants (including the Commission Services)	X

Revision history and statement of originality

Revision	Date	Author	Organisation	Description
Draft 0.1	20.06.12	Hannes Mühleisen, Julia Iwanowa, Evelyn Dröge, Steffen Hennicke, Konstantin Baierer	Freie Universität Berlin, Humboldt-Universität zu Berlin, Ex Libris	Initial Draft
Draft 0.2	27.06.12	Steffen Hennicke	Humboldt-Universität zu Berlin	Incorporated feedback of Valentine Charles
Draft 0.3	29.06.12	Julia Iwanowa, Evelyn Dröge	Humboldt-Universität zu Berlin	Section 5.3 added
Draft 0.4	02.07.12	Evelyn Dröge, Steffen Hennicke, Konstantin Baierer	Humboldt-Universität zu Berlin, Ex Libris	Introduction added Refinements of chapters 4 – 6
Draft 0.5	04.07.12	Evelyn Dröge	Humboldt-Universität zu Berlin	Incorporated feedback of Christian Bizer
Draft 0.6	05.07.12	Hannes Mühleisen, Evelyn Dröge	Freie Universität Berlin, Humboldt-Universität zu Berlin	Section 3 added
Draft 0.7	15.07.12	Nasos Drosopoulos	National Technical University of Athens	Extension of section 5.1
Draft 0.8	15.07.12	Evelyn Dröge	Humboldt-Universität zu Berlin	Section 8 added
Draft 0.9	16.07.12	Dirk Wintergrün, Klaus Thoden	Max-Planck-Institut für Wissenschafts-geschichte	Section 7 added
Draft 1.0	17.07.12	Christian Bizer	Freie Universität Berlin	Refinements of sections 1, 3, 5, 7 & 8
Draft 1.1	20.07.12	Steffen Hennicke	Humboldt-Universität zu Berlin	Incorporated feedback of Valentine Charles
Draft 1.2	23.07.12	Evelyn Dröge, Steffen Hennicke, Konstantin Baierer	Humboldt-Universität zu Berlin, Ex Libris	Refinements of all sections
Draft 1.3	30.07.12	Julia Iwanowa	Humboldt-Universität zu Berlin	Incorporated feedback of Ewelina Rockenbauer
Draft 1.4	30.07.12	Steffen Hennicke, Julia Iwanowa	Humboldt-Universität zu Berlin	Incorporated feedback of Kilian Schmidtner
Draft 1.5	31.07.12	Evelyn Dröge	Humboldt-Universität zu Berlin	Incorporated feedback of Violeta Trkulja
Version 1.0	31.07.12	Evelyn Dröge, Steffen Hennicke	Humboldt-Universität zu Berlin	Version 1.0
Version 1.0	10.08.12	Stefan Gradmann	Humboldt-Universität zu Berlin	Approval version 1.0

Statement of originality:

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

Contents

1 Abstract 6

2 Introduction 7

3 Working approach 8

4 DM2E Extension of EDM 10

4.1 Basic Entity Structure in EDM 10

4.2 The EDM+ 12

4.2.1 ore:Aggregation 13

4.2.2 edm:ProvidedCHO 16

4.2.3 edm:WebResource 20

4.2.4 Contextual Resources 21

4.3 Conclusion and Next Steps 23

5 RDFisation, Mapping and Translation Workflow 25

5.1 MINT Platform 26

5.1.1 The Use of MINT within the Project 28

5.1.2 Target Schema Configuration 28

5.1.3 Evaluation and Next Steps 30

5.2 jMet2Ont 31

5.3 The D2R Platform 32

5.3.1 The Use of D2R within the Project 32

5.3.2 Requirements for the Mappings of Relational Databases with the D2R Server 33

5.3.3 D2R Mappings 33

5.4 ClioPatria, XMLRDF and Amalgame 35

5.5 Conclusion 36

6 Requirements for Input Data and Missing Information 37

6.1 Missing Elements in the Sample Data 37

6.1.1 Encoded Archival Description (EAD) 37

6.1.2 Machine-Readable Cataloging (MARC/XML) 39

6.1.3 Text Encoding Initiative (TEI/XML) 41

6.1.4 Individual Formats 42

6.2 Requirements and Suggestions 44

6.3 Conclusion and Next Steps 45

7 Contextualisation 46

7.1 Silk Link Discovery Framework 46

7.2 Conclusion and Next Steps 47

8 Conclusion and Next Steps 48

9 References 49

List of Tables

Table 1: ore:Aggregation properties in EDM+. 15
Table 2: edm:ProvidedCHO properties in EDM+. 19
Table 3: Missing mandatory elements SBB (EAD). 38
Table 4: Missing mandatory elements ONB, Codices (MARC). 39
Table 5: Missing mandatory elements ONB, Google Books (MARC). 39
Table 6: Missing mandatory elements NLI, books (MARC). 40
Table 7: Missing mandatory elements NLI, manuscripts (MARC). 40
Table 8: Missing mandatory elements BBAW (TEI P5). 41
Table 9: Missing mandatory elements UBER (TEI). 41
Table 10: Missing mandatory elements UIB (TEI). 42
Table 11: Missing mandatory elements, MPIWG, VLP database, table vl_essays. 43
Table 12: Missing mandatory elements, MPIWG, VLP database, table vl_people. 43

List of Figures

Figure 1: Basic tripartite structure for representing a CHO with EDM and minimal requirements by Europeana in terms of properties. 12
Figure 2: Combination of the workflow execution and management components that will be implemented by WP2. Sources on the left hand side will be formatted into RDF, mapped to EDM or EDM+ and enriched with contextualisation tools (centre) before being delivered to Europeana (right hand side). 25
Figure 3: MINT tool in action: This screen shows the creation of a mapping from a TEI flavour (in this case a document provided by BBAW) to the EUROPEANA target schema (EDM). In this particular example, an ore:Aggregation class is created for every local ID with an edm:aggregatedCHO property linking to the concatenation of a fixed base URL and the ID and some other properties set to literals found in the source text. 27
Figure 4: Preview of a transformation using a mapping created in MINT. The result is a set of RDF resources, serialised in RDF/XML, that represents the corresponding input record using the EDM vocabulary. 28
Figure 5: D2R accessed with the Firefox browser. This screenshot shows the resource “art10” with entities that are already mapped to the EDM and with some additional entities that are not yet mapped. 35

List of Abbreviations

BBAW	Berlin-Brandenburgische Akademie der Wissenschaften *(Berlin-Brandenburg Academy of Sciences and Humanities)*
CHO	Cultural Heritage Object
CJH	Centre for Jewish History
DFGA	Digitale Faksimile Gesamtausgabe *(Digital Facsimile Edition)*
DM2E	Digitised Manuscripts to Europeana
DNB	Deutsche Nationalbibliothek *(German National Library)*
DoW	Description of Work
EDM	Europeana Data Model
EDM+	Initial EDM Specialisation by DM2E
FUB	Freie Universität Berlin *(Free University Berlin)*
GND	Gemeinsame Normdatei *(Universal Authority File)*
GUI	Graphical User Interface
JDC	Joint Distribution Committee
LOD	Linked Open Data
MPIWG	Max-Planck-Institut für Wissenschaftsgeschichte *(Max Planck Institute for the History of Science)*
NLI	National Library of Israel
NTUA	National Technical University of Athens
ONB	Österreichische Nationalbibliothek *(Austrian National Library)*
PND	Personennamendatei *(Name Authority File)*
SBB	Staatsbibliothek zu Berlin *(Berlin State Library)*
UBER	Humboldt-Universität zu Berlin *(Humboldt University Berlin)*
UBFFM	Universitätsbibliothek Johann Christian Senckenberg Frankfurt am Main *(University Library Johann Christian Senckenberg Frankfurt am Main)*
UIB	Universitetet i Bergen *(University of Bergen)*
WP	Work Package

Please note that some translations of institution names are unofficial and do only serve a better understanding of the corresponding abbreviation.

1 Abstract

This deliverable on the “Initial Version of the Interoperability Infrastructure” reports on the results achieved by WP2 of DM2E during the first six months of the project and describes the next steps which will be carried out during the upcoming months.

The initial version of the infrastructure combines the D2R Platform for RDFisation of relational data, the Mint Tool for the RDFisation of XML data with the Silk Link Discovery Framework for the contextualisation of the RDFised data.

In order to verify that the chosen tools comply with the requirements that arise in the context of the DM2E project, WP2 has analysed the input data that was provided by the WP1 partners and has drafted a specialisation of the EDM (EDM+) which includes additional concepts that are necessary to represent the WP1 data as well as for using the RDFised data in the context of the annotation platform that is developed in WP3. Afterwards, the tools were tested by mapping the input data into the developed specialisation of the EDM. This test confirmed that the tools are capable of handling the required transformations, but also revealed some extensions to the tools that need to be implemented in order to comply with all aspects of the requirements.

Within this deliverable document, we report on the results of the analysis of the content providers' input data, the first draft of the developed specialisation of the EDM, the initial version of the interoperability infrastructure, the results of the mapping experiment as well as the extensions to the tools that will be implemented in the next months.

2 Introduction

The project “Digitised Manuscripts to Europeana” (DM2E) has three main objectives: (1) to provide rich data to Europeana in form of metadata about manuscripts modelled in accordance to the Europeana Data Model (EDM), (2) to develop an open-source tool chain for data mapping and conversion to EDM, and (3) to develop and implement new functionality for the Digital Humanities and to perform related research on the „Scholarly Primitives“, originally defined by John Unsworth.

The main aim of WP2 is to “provide the interoperability infrastructure for translating content from its current source formats into the Europeana Data Model (EDM).” The interoperability infrastructure will rely on existing tools which will be adapted and chained into a scalable translation workflow. This workflow will include three distinct steps: the RDFisation of existing content and metadata, the mapping into the Europeana Data Model, and the additional contextualisation and interlinking of this data with other resources (cf. DM2E DoW, 2012:14).

We report here on the activities and first outcomes achieved during the first six months of WP2. The deliverable is organised as follows:

Section 3 gives a description of our working approach along with the tasks we have worked on and the tools that we have used to accomplish the initial infrastructure version. This section gives an overview of the essential outcomes of our activities so far.

Section 4 provides a first draft of the DM2E specialisation of the EDM. The working title of this modification is EDM+. It is a response to requirements which came up during the initial sample data analysis of WP1 and WP2 and functional specification activities of WP3.

Section 5 deals with the toolchain evaluation and the definition of the first tool extension requirements. The first part of the toolchain evaluation focuses on the MINT tool for the RDFisation of content in XML format and the D2R server that maps database content into RDF. Mappings into the EDM can be provided by a MINT specialisation for DM2E that includes the initial data model from section 3. Before being able to map the source data of the projects’ data providers, we have to analyse the data first. The results of this analysis, including the listing of missing elements as well as requirements and optional demands on the source data, can be found in section 5.

Section 6 reports on the initial analysis of the data providers’ sample data, focusing on missing information with regard to the minimal and mandatory requirements of the EDM. It gives a set of requirements and recommendations to be considered by the data providers when mapping and converting their data for DM2E.

Section 7 takes a closer look at the contextualisation tool Silk.

Section 8 is a summary of the results of the activities of WP2 during the first six months and the next steps which will be carried in the upcoming months.

3 Working approach

The goal of WP2 for the first six months was to provide an initial version of the interoperability infrastructure. The focus lay on Task 2.1, the development of an infrastructure for RDF conversion and Task 2.2, the mapping of the metadata received from the data providers to the EDM. In addition, initial work has started in Task 2.3 which aims at enriching the transformed metadata by contextualising it to external data sources.

In order to verify that the chosen tools comply with the requirements that arise in the context of the DM2E project, we chose the following working approach:

1. First, we analysed the sample input data that was provided by the WP1 partners.
2. Based on this analysis, we developed a specialisation of the EDM data model which includes additional concepts that are necessary to represent the WP1 data as well as for using the RDFised data in the context of the annotation platform that is developed in WP3.
3. Afterwards, we tested the MINT tool and D2R platform by mapping the input data into the developed specialisation of the EDM. This test confirmed that the tools are capable of handling the required transformations, but also revealed some extensions to the tools that need to be implemented in order to comply with all aspects of the requirements.
4. In order to verify that the Silk Link Discovery Framework complies with the requirements that arise in the context of the DM2E project, we conducted an initial contextualisation experiment. This experiment already delivered promising results, but also showed that a more detailed evaluation will be necessary to ensure that all DM2E contextualisation requirements are met.

In the following, we provide more details on our working approach and the actual work that was carried out following it.

The data providers organised in WP1 provided WP2 with an extensive set of sample data describing digitised manuscripts. The data sets included metadata and object data about books, posters, photographs, audio files, movies, journals, autographs, and newspapers. The sample data were delivered in a large variety of formats which created a good opportunity to test our infrastructure early on.

In Task 2.1, we first began collecting tools in our WP2 Wiki which have the potential to facilitate the RDFisation and mapping task. For our initial tool analysis, we put the focus on the MINT mapping tool provided by NTUA and the D2R platform provided by FUB. Both of these tools allow the creation of so-called mappings that transform the various metadata formats used by the providers to RDF.

In order to evaluate those tools, we created mappings to EDM for the sample data offline and tried to apply them to the tools. Because it became apparent that the tools were not yet ready for covering all aspects of the mappings, we also created “mappings on paper”, i.e. we drew graph representations (mainly with the VUE tool[1] built by Tufts University) and manually produced EDM RDF data, serialised in Turtle and RDF/XML syntax.

[1] http://vue.tufts.edu [23.07.2012]

Based on these tests, we identified several requirements for the RDFisation and mapping tools to be used in the DM2E toolchain, e.g. the capability to create URLs from data in multiple places in the source data, the ability to define “default” values as well as support for dynamically creating contextual resources for describing people, places, and concepts. Both MINT and D2R were found suitable in principle to perform the RDFisation and mapping of the sample data provided by the project partners under the condition that we can adjust them according to our mapping requirements.

In Task 2.2, we have begun reviewing the EDM with special focus on the use case of digitised manuscripts. During the mapping exercises performed for Task 2.1 and two WP2 meetings, we extended and specialised the EDM towards supporting granular mappings of digitised manuscripts and for the purpose of rich client applications such as those being built by WP3. At the same time, we have adhered to best practices for publishing Linked Data on the Web. The working title of this modified version of the EDM is EDM+.

To capture the full wealth of semantics in the sample data provided, the EDM+ includes a subclass taxonomy extending the EDM and indicating the type of the described objects such as books, parts of books, articles as well as individual pages. In the same way, we have made the relationship between the described objects, contextual resources as well as their web representations more specific. Whenever possible, we reused existing vocabularies for the EDM modifications. We began collecting such vocabularies in our WP2 Wiki.

We also reviewed the sample data regarding the minimum requirements specified for metadata modelled with EDM. We have used the issues found to create a list of requirements and suggestions regarding the providers’ data, to extend a questionnaire for each data provider that was being prepared by WP1 as well as to further refine the EDM+ model.

Collecting requirements is, according to the DoW, entirely a task of WP1. Since we needed to have a first analysis of the metadata and requirements based thereupon for the next steps in WP2 earlier than WP1 could have provided them, we had to add this sub-task to our WP2 agenda. Thus, WP2 collected requirements from a top-down point of view whereas WP1, afterwards, collected requirements based on our work with a special focus on the data providers, which can be seen as a bottom-up approach. The results of these two approaches will be merged after month 6.

As initial work in Task 2.3, we have started to identify targets for contextualisation during our survey of the sample data from the data providers. So far, we have found that resources that represent organisations providing data or holding the described object may be linked to geographical information. Further questions are how conceptual and taxonomic terms can be linked to standardised vocabularies, how the location of a described object inside an institutions building can also be contextualised using internal locator identifiers, how authors and publishers can be linked to entities from standard vocabularies such as PND or VIAF and thereby enrich the information about the described object. We have used the Silk tool provided by FUB in order to test the contextualisation of the MPIWG data. The evaluation is ongoing, however, first results are presented.

In the following section, we are going to describe our EDM specialisation.

4 DM2E Extension of EDM

In the context of the DM2E project, we will collect metadata about digitised manuscripts, user generated annotations, and object data (e.g. TEI/XML files or page scans) from various cultural heritage institutions. Although these data foremost are about manuscripts, they are very disparate in terms of their representational model (e.g. METS/MODS, EAD, or MARC21). One of the main challenges is to transform these data into a unified model and at the same time to retain the richness and depth of the original metadata as much as possible in order to enable rich functionality on top of these data.

The Europeana Data Model (EDM) has been developed by the Europeana project for exactly this use case. The EDM is an open and cross-domain framework which is based on Semantic Web technology and Linked Data principles. It allows integrating heterogeneous cultural heritage data into a shared interoperability layer while still retaining the original semantics of the source data (cf. EDM Documentation, URL: http://pro.europeana.eu/edm-documentation). For this reason, the EDM has been chosen by DM2E to provide an interoperability layer and it is therefore a crucial building block of our interoperability infrastructure.

The first analysis of the provided sample data within the DM2E project has clearly shown that the current version of the EDM, on the one hand, is in principle able to accommodate all provided sample data but, on the other hand, has to be specialised in order to retain all provided information and semantics of the source data.

The DM2E project, therefore, creates a modified version of the EDM tailored towards the domain of digitised manuscripts as it is exemplified by the data provided to DM2E. The working title of this modified model is currently EDM+ and will probably be changed in the near future. The EDM+ essentially constitutes a subset of the full EDM model in terms of properties and a new set of specialisations of EDM properties and classes.

The first draft of the EDM+ is based on extensive mapping exercises of the provided sample data carried out by UBER and MPIWG as well as on results of a WP2 workshop held at the FUB, 7th May 2012, where decisions have been taken towards selecting necessary properties from the EDM and first extensions in terms of specialisations of properties and classes.

In the following sections, we first explain the basic entity structure as defined by the EDM for modelling a cultural heritage object and then describe the EDM+.

4.1 Basic Entity Structure in EDM

The EDM demands for each Cultural Heritage Object (CHO) to create a simple tripartite entity structure as its core representation. Note: The following description describes the original EDM requirements.

Each object (e.g. book, manuscript, journal) and each part of an object (e.g. chapters, pages, articles) as well as annotations (of the whole object or of parts which we can interpret as a CHO in its own right) are represented by three RDF entities, also called “resources”. These resources are instances of the following EDM classes:

- *edm:ProvidedCHO* - Is the class for the described CHO. The resource, which is an instance of the class *edm:ProvidedCHO*, represents the described CHO which can,

for example, be a book, a journal, a chapter within a book, or an article within a journal.

- The *ProvidedCHO* holds the descriptive metadata for the described CHO (like the title, or the creator). It is the anchor point to which other contextual resources from the Web can be connected to, for example places, people, concepts, or annotations.

- *ore:Aggregation* - Is the class for the "the complex constructs that are provided by contributors" (EDM Definitions 5.2.3, 2012:8), typically a metadata record. The resource, which is an instance of the class *ore:Aggregation*, represents the metadata record or a part of a metadata record which is about a CHO (e.g. a book) or a part of a CHO (e.g. a chapter of a book).
  - The *Aggregation* holds administrative, legal and provenance metadata about the described CHO and the metadata record itself. It is connected to the *ProvidedCHO* via the property *edm:aggregatedCHO* which expresses aboutness, i.e. "the metadata record (the *Aggregation*) is about or talks about this object (the *ProvidedCHO*)" (EDM Definitions 5.2.3, 2012:7).
- *edm:WebResource* - Is the class of a resource on the Web which in some form shows or is a view of the described CHO, like a page scan, a landing page, a thumbnail, a transcription, a table of content, an HTML page, an audio file, a text file or a viewer page. A resource on the Web, which is a view of the described CHO, is an instance of the class *edm:WebResource*.
  - A *WebResource* is attached to the Aggregation. For every *ProvidedCHO*, there must be at least one *WebResource* attached to the Aggregation. In the EDM this is via the property *edm:isShownAt* (URL to a view within an information context, like an HTML page or Digital Library viewer) or *edm:isShownBy* (URL to the "plain" view or image without any information context), in EDM+, the property *edm:isShownBy* is mandatory (details are given below).

In other words, we create a basic tripartite entity structure for every object we want to model in EDM: (1) One resource as an instance of the class *edm:ProvidedCHO*, (2) one resource as an instance of the class *ore:Aggregation,* which is connected to the *ProvidedCHO* via the property *ore:aggregatedCHO*, and (3) one resource as an instance of the class *edm:WebResource* which shows the described object in some way and which is attached to the ore:Aggregation via the property *edm:isShownAt* or *edm:isShownBy* (in EDM+, *edm:isShownBy* is mandatory).

Figure 1 shows this basic tripartite structure along with all mandatory EDM properties which will be explained later on. This constitutes the minimal EDM requirements towards representing a CHO. Note that the *Aggregation* links to the *ProvidedCHO* and the *WebResource*.

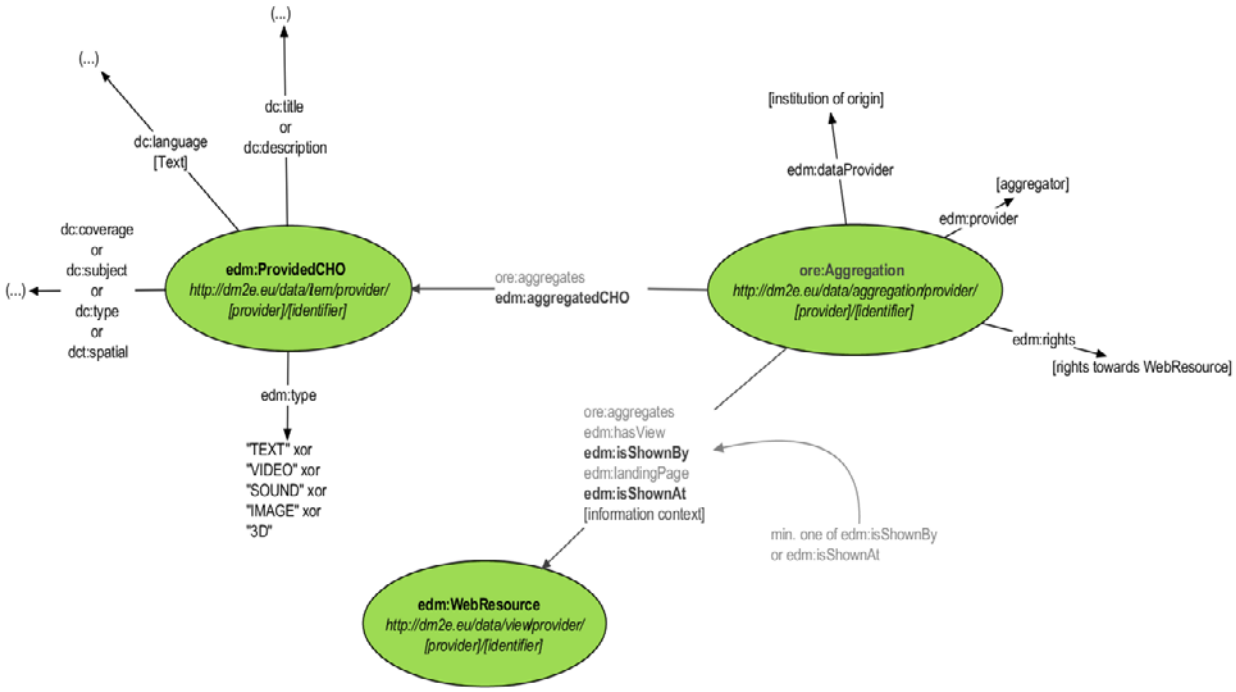

Figure 1: Basic tripartite structure for representing a CHO with EDM and minimal requirements by Europeana in terms of properties.

If a metadata record describes a complex object like a book with several chapters (a typical case are METS files about old printings which have been digitised), then this complex structure is also represented with the above described basic tripartite entity structure for each object which is deemed relevant. The relation between those objects (i.e. the parts of the complex object) is described by properties between the *edm:ProvidedCHO* (like *dct:isPartOf*) and the *ore:Aggregation* (like *ore:aggregates*).

In the following, we show how these entities are defined and which modifications we made referring to the EDM.

4.2 The EDM+

Within the DM2E project, our goal is to create mappings as rich as possible, representing the original semantics of the data providers metadata as closely as possible. Having taken a closer look at the EDM, we found out that we needed to make some adjustments in order to get a more homogeneous and rich metadata representation. The key points we have for achieving this are the following:

- Use URIs of web resources for properties when describing entities like authors, publishers, places, and topics instead of plain literals. The contextual classes *edm:Agent*, *edm:Place*, *edm:TimeSpan*, and *skos:Concept* are always optional in the EDM, but we strongly encourage and advertise to use them in DM2E as they are more expressive than simple literals.
- Specialise classes in order to capture specific entity types (e.g. fabio:*Book* (taken from http://purl.org/spar/fabio) instead of just *edm:PhysicalThing*).
- Specialise properties to capture specific relation types where needed.
- Extensions to the EDM (classes and properties) should be reused from existing namespaces, if possible. Only if no appropriate resource is found, we create a new

class or property in the DM2E namespace (this principle has not been followed in the early development stage in order to enable a dynamic and expansive approach – after consolidation of EDM+, however, a systematic check of the choices made will be a necessary step eventually replacing our own classes and properties with those available elsewehere as a consequence).

- For each CHO, one stable URL to a plain view or image of the object must be provided.

Additionally, we decided to use a subset of the full set of EDM properties in order to keep the EDM+ as clear and simple as possible for the data providers which will have to create the definite mappings of their data to EDM.

In the following sections, we will offer an insight into our first EDM modifications that are mainly made to the classes *ore:Aggregation* and *edm:ProvidedCHO*.

The extended EDM+ model, which is described in the following, includes more native EDM properties than those that will be implemented by Europeana in the first iteration. This also means that issues of data provision to Europeana have not yet been a primary concern. Not all native EDM properties are included and super-properties are not shown. This is a first draft and will be subject to future modifications (cf. section 4.3, "Next Steps").

In the tables below, the ranges indicate what kind of data value we prefer to receive (literal or resource). Properties in red have been created by DM2E (indicated by the namespace prefix "dm2e") whereas properties in blue are taken from an existing ontology. Existing ontologies were collected in the projects' Wiki, but not yet rated. Thus, they can be exchanged in later modifications of the model. The source ontology is indicated by the namespace prefix. The full namespace is given above the table. Note: All examples are fictional.

4.2.1 ore:Aggregation

The class *ore:Aggregation* is described by the EDM Definitions 5.2.3 as follows: "A set of related resources (Aggregated Resources), grouped together such that the set can be treated as a single resource. This is the entity described within the ORE interoperability framework by a Resource Map" (EDM Definitions 5.2.3, 2012:7).

The resource representing the metadata record provided to DM2E must be an instance of *ore:Aggregation*.

Specialisations

None.

Added Properties

DM2E extended the domain of a few EDM properties to *ore:Aggregation*. These properties are *dct:created*, *dct:modified*, and *dct:creator*.

The property *korbo:hasAnnotableVersionAt* is new and used for versions of annotable content for the prototype platform of WP3. For details on the allowed contents for this property, confer the DM2E WP3 Wiki or the DM2E deliverable D3.1.

Namespaces

```
edm:      <http://www.europeana.eu/schemas/edm/> .
dc:       <http://purl.org/dc/elements/1.1/> .
dct:      <http://purl.org/dc/terms/> .
korbo:    <http://purl.org/net7/korbo/vocab#> .
```

Property	Range	Description	Constraints
edm:aggregatedCHO	ProvidedCHO entity	Connects the ore:Aggregation with the edm:ProvidedCHO. It aggregates and it is about.	**mandatory** **not** repeatable
edm:provider	edm:Agent entity	Organisation (edm:Agent, see below) that provided this aggregation.	**mandatory** **not** repeatable
edm:dataProvider	edm:Agent entity	Organisation (edm:Agent, see below) that provided the source data for this aggregation (could be the same as edm:provider).	**mandatory** **not** repeatable
edm:rights	Remote resource	URL of a resource describing licensing rights from the Guidelines for the Rights in Objects submitted to Europeana (2012).	**mandatory** **not** repeatable
dc:rights	Remote resource	URL of any resource describing licensing rights of the Web resource.	optional repeatable
edm:isShownBy	WebResource (with stable URL)	Web resource leading to the **"plain" image** showing the ProvidedCHO.	**mandatory** (in DM2E) **not** repeatable
edm:isShownAt	WebResource entity	A Web resource in the original Digital Library showing the ProvidedCHO.	not mandatory but strongly recommended **not** repeatable
edm:hasView	WebResource entity	Additional Web resources showing, depicting or otherwise containing a view of the ProvidedCHO.	optional repeatable
edm:object	WebResource entity	Preview picture for the ProvidedCHO.	not mandatory but strongly recommended **not** repeatable
korbo:hasAnnotableVersionAt	WebResource entity	Content providers should provide **HTML representations** of their content including markup to identify named-content (Annotable Versions). An HTML representation of an object can include a single named-content (in this case it represents a single atomic piece of content, e.g. the transcription of a page) or multiple named contents (e.g. marking up each single paragraph or picture). Deciding the granularity of named-content is up to each content provider. For details on the allowed contents for this property confer the DM2E Wiki page:	not mandatory but strongly recommended **not** repeatable

Property	Range	Description	Constraints
		https://dm2e.hu-berlin.de/redmine/projects/wp3/wiki/Named_content_markup	
dct:created	xsd:dateTime literal	Creation date and time of this aggregation.	optional repeatable
dct:modified	xsd:dateTime literal	Modification date and time of this aggregation.	optional repeatable
dct:creator	Remote resource	URL of the creator of this aggregation (e.g. library staff member).	optional repeatable

Table 1: ore:Aggregation properties in EDM+.

Mandatory Properties

The mandatory properties for *ore:Aggregation* are *ore:aggregatedCHO*, *edm:rights*, *edm:provider*, *edm:dataProvider*, and *edm:isShownBy* (with a stable URL to the "plain" image showing the CHO).

URL Scheme

All *Aggregation* entities are identified by a URL with the following scheme:

```
http://dm2e.eu/data/aggregation/provider/[provider]/[identifier]
```

- [provider] is a short data provider identifier, e.g. sbb for Staatsbibliothek Berlin.
- [identifier] is a provider-unique identifier such as a signature or an internal database ID.

Example

```
@prefix    edm:  <http://www.europeana.eu/schemas/edm/> .
@prefix    ore:  <http://www.openarchives.org/ore/terms/> .
@prefix    dc:   <http://purl.org/dc/elements/1.1/> .
@prefix    dct:  <http://purl.org/dc/terms/> .

<http://dm2e.eu/data/aggregation/provider/sle/abc345> a
ore:Aggregation ;

edm:aggregatedCHO <http://dm2e.eu/data/item/provider/bla/12344> ;
edm:provider <http://dm2e.eu/data/provider/sle> ;
edm:dataProvider <http://dm2e.eu/data/provider/esl> ;
edm:rights <http://creativecommons.org/publicdomain/mark/1.0/> ;
dc:rights <http://loc.gov/data-license> ;
edm:isShownBy
<http://dm2e.eu/data/resource/provider/sle/abc345/scan.jpg> ;
edm:isShownAt
<http://dm2e.eu/data/resource/provider/sle/abc345/overview.html> ;
edm:hasView
<http://dm2e.eu/data/resource/provider/sle/abc34/text.txt> ;
edm:object
<http://dm2e.eu/data/resource/provider/sle/abc34/thumbnail.jpg> ;
dct:created ""^^xsd:time ;
```

```
dct:modified ""^^xsd:time ;
dct:creator <http://somelibrary.example/staff/brucewillis> .
```

4.2.2 edm:ProvidedCHO

The EDM Definitions 5.2.3 describes the *ProvidedCHO* as follows: “This class comprises the Cultural Heritage objects that Europeana collects descriptions about.” (Definition of the Europeana Data Model elements Version 5.2.3, 2012:14).

The resource representing the described cultural heritage object (CHO) must be an instance of *edm:ProvidedCHO*.

Specialisations

The resource which represents the described CHO can be made an instance of a second smaller class in order to be more precise about its type. The EDM provides only a limited set of generic classes for non-information resources. In the case of the CHOs aggregated by DM2E, we can only assign the class *edm:PhysicalThing*.

In DM2E, the type of the described CHO will be made more specific by creating additional subclasses for *edm:PhysicalThing*. The list of such possible specialisations is work in progress. Examples include *dm2e:Book*, *dm2e:Painting*, *dm2e:Manuscript*, *dm2e:Journal*, and *dm2e:Article*.

These specialisations, all of which are non-information resources, should be made subclasses of *edm:PhysicalThing* and not of *edm:ProvidedCHO*. The resource representing the described object must be an instance of *edm:ProvidedCHO*, but should additionally be made an instance of one of the aforementioned new subclasses of *edm:PhysicalThing*. In this way, the described object is typed as a non-information resource. The exact semantics and the detailed definition of those specialisations is not yet final at this point.

Added Properties

DM2E created three new properties with the domain *edm:ProvidedCHO*. These properties are *dm2e:titleTransliteration*, *dm2e:subtitleTransliteration*, and *dm2e:publishedAt*. Furthermore, the property *bibo:numPages* and *bibo:numVolumes* have been included from “The Bibliographic Ontology” (BIBO)[2]. The domain of the EDM property *edm:rights* has been added to *edm:ProvidedCHO*.

Namespaces

Note that the DM2E namespace does not exist yet.

```
edm:  <http://www.europeana.eu/schemas/edm/> .
dm2e: <http://dm2e.eu/schema/> .
ore:  <http://www.openarchives.org/ore/terms/> .
dc:   <http://purl.org/dc/elements/1.1/> .
dct:  <http://purl.org/dc/terms/> .
bibo: <http://purl.org/ontology/bibo/> .
```

[2] The Bibliographic Ontology, http://bibliontology.com [23.07.2012].

Property	Range	Description	Constraint
edm:type	Literal	Must be one of the following: TEXT, VIDEO, SOUND, IMAGE, 3D	**mandatory** **not** repeatable
dc:type	Specialisation of edm:PhysicalThing	The most specific type that applies to the CHO, repetition of rdf:type of Proxy	**mandatory** (in DM2E) repeatable
dc:title	Literal with language tag	Title	**mandatory** (either dc:title or dc:description) repeatable
dm2e:titleTransliteration	Literal with language tag	Subproperty of dc:title Title transliteration	optional repeatable
dc:description	Literal with language tag	A description of the CHO	**mandatory** (either dc:title or dc:description) repeatable
dcterms:alternative	Literal with language tag	Any form of the title used as a substitute or alternative to the formal title of the resource	optional repeatable
dm2e:subtitle	Literal with language tag	Subproperty of dc:title Any form of subtitle	optional repeatable
dm2e:subtitleTransliteration	Literal with language tag	Subproperty of dm2e:subtitle Subtitle transliteration	optional repeatable
dc:language	Literal with xsd:lang	Most prominent language of the CHO	**mandatory** (for CHOs of edm:type “TEXT”) repeatable
dct:issued	xsd:dateTime literal	Date of publication, also possible: dct:created, or dc:date for generic cases	optional **not** repeatable
dct:creator	edm:Agent entity	Creator of the CHO, possibly its author. Subproperties may be used to specialise relationship	optional repeatable
dc:publisher	edm:Agent entity	Publisher of the CHO	optional repeatable
dm2e:publishedAt	edm:Place entity	Subproperty of dct:spatial The place of publication	optional **not** repeatable
dc:identifier	Literal	Provider-local identifier of the CHO; *DM2E envisions several more specific subproperties which are meant to cover most domain or institution specific identifiers*	optional repeatable

Property	Range	Description	Constraint
dm2e:isbn	Literal	Subproperty of dc:identifier The ISBN number for the CHO	optional **not** repeatable
dm2e:callNumber	Literal	Subproperty of dc:identifier; The call number for some archival item	optional **not** repeatable
edm:currentLocation	edm:Place entity	Current location the of physical CHO, possibly a library building	optional **not** repeatable
edm:rights	Resource	URL of a resource describing licensing rights from the Guidelines for the Rights in Objects submitted to Europeana (2012); Use dc:rights if the URI of the license does not fit the demands of Europeana	optional repeatable
dc:subject	Resource	Subject of the CHO. Can be borrowed of another vocabulary	**mandatory** (in DM2E) repeatable
dct:extent	Literal	The size or duration of the resource. Currently, this property has two subproperties; *DM2E envisions several more specific subproperties which can be derived from the BIBO ontology*	optional **not** repeatable
bibo:numPages	Literal	Subproperty of dct:extent; Number of pages of the resource	optional **not** repeatable
bibo:numVolumes	Literal	Subproperty of dct:extent; Number of volumes of the resource	optional **not** repeatable
dct:tableOfContents	Resource or Literal	Any kind of table of contents for the CHO	optional repeatable
dct:provenance	Literal	Description of provenance of the CHO	optional repeatable
dc:format	Literal	The file format, physical medium, or dimensions of the resource	optional repeatable
edm:isDerivativeOf	Resource (ProvidedCHO entity)	Original version from which this object has been derived	optional repeatable
dct:hasVersion	Resource (ProvidedCHO entity)	Related resource that is an adaption of this resource	optional repeatable

Property	Range	Description	Constraint
dct:hasPart	Resource (ProvidedCHO entity)	Reference to a part of this CHO, e.g. a chapter of a book	optional repeatable
edm:isNextInSequence	Resource (ProvidedCHO entity)	Preceding same-level CHO, e.g. previous chapter.	optional repeatable
dct:references	Resource (ProvidedCHO entity)	Other CHO referenced in the content of this CHO.	optional repeatable

Table 2: edm:ProvidedCHO properties in EDM+.

Mandatory Properties

The mandatory properties for the *ProvidedCHO* are either *dc:title* or *dc:description*, one of *dc:coverage*, *dc:type*, *dc:subject*, or *dct:spatial* where *dc:type* is mandatory in DM2E, *dc:language,* if the CHO is a text object, and *edm:type*.

URL Scheme

All CHO entities are identified by an URL with the following scheme:

`http://dm2e.eu/data/item/provider/[provider]/[identifier]`

- [provider] is a short data provider identifier, e.g. sbb for Staatsbibliothek Berlin.
- [identifier] is a provider-unique identifier such as a signature or an internal database ID.

Example

Note that the DM2E namespace does not exist yet.

```
@prefix    edm:  <http://www.europeana.eu/schemas/edm/> .
@prefix    dm2e: <http://dm2e.eu/schema/> .
@prefix    ore:  <http://www.openarchives.org/ore/terms/> .
@prefix    dc:   <http://purl.org/dc/elements/1.1/> .
@prefix    dct:  <http://purl.org/dc/terms/> .
@prefix    bibo: <http://purl.org/ontology/bibo/> .

<http://dm2e.eu/data/item/provider/sle/DE-1a-904> a ore:Proxy,
dm2e:Book ;

edm:type "TEXT" ;
dc:type dm2e:Book ;
dc:title "On the Origin of Species"@en ;
dm2e:subtitle "by Means of Natural Selection"@en ;
dm2e:titleTransliteration "abc"@en ;
dm2e:subtitleTransliteration "xyz"@en ;
dct:alternative "On the Origin of Species by Means of Natural
Selection, or the Preservation of Favoured Races in the Struggle for
Life."@en;
dc:description "Darwin’s book introduced the scientific theory that
populations evolve over the course of generations through a process of
```

```
natural selection. It presented a body of evidence that the diversity
of life arose by common descent through a branching pattern of
evolution."@en ;
dc:language "en"^^xsd:lang ;
dct:issued "1859?11?12T00:00:00Z"^^xsd:dateTime ;
dct:creator <http://dm2e.eu/data/resource/provider/sle/darwin> ;
dc:publisher <http://dm2e.eu/data/resource/provider/sle/murray> ;
dm2e:publishedAt <http://dm2e.eu/data/place/provider/sle/42> ;
dc:identifier "DE-1a-904" ; # provider-local identifier
dm2e:isbn13 "978–0486450063" ; # more specific identifier
dm2e:callNumber "QE534.2.B64" ; # internal identifier
edm:currentLocation <http://dm2e.eu/data/place/provider/sle/12344> ;
edm:rights <http://creativecommons.org/publicdomain/mark/1.0/> ;
dc:subject <http://dewey.info/class/576.801> ;
dct:extent "502 Pages" ;
bibo:numPages "502" ;
dct:tableofcontents <http://example.com/resourcr/dasdas/toc.html> ;
dct:provenance "Added to Library in June 1965" ;
dc:format "Folio, Leather Bound" ;
edm:isDerivativeOf <http://dm2e.eu/data/item/provider/sle/abc9887> ;
dct:hasVersion <http://dm2e.eu/data/item/provider/sle/abc9886> ;
dct:hasPart <http://dm2e.eu/data/item/provider/sle/DE-1a-904-3> ;
edm:isNextInSequence
<http://dm2e.eu/data/item/provider/sle/DE-1a-903> ;
dct:references <http://dm2e.eu/data/item/provider/sle/DE-1a-923> .
```

4.2.3 edm:WebResource

According to the EDM Definitions 5.2.3, *WebResources* are “Information Resources that have at least one Web Representation and at least a URI.” (EDM Definitions 5.2.3, 2012:15).

The resource, which resembles any kind of view of the described CHO, is an instance of the class *edm:WebResource*. There must be at least one WebResource for each CHO.

In DM2E, a stable URL pointing the plain view or image of the CHO is **mandatory**. This resource is connected to the *Aggregation* via *edm:isShownBy*.

Specialisations

In DM2E, the type of a *WebResource* showing the described CHO will be made more specific by creating additional subclasses for *edm:WebResource*. The list of such possible specialisations is work in progress. An example may be *dm2e:PageScan*. The exact semantics and the detailed definition of those specialisations is not yet final at this point.

Added Properties

None. Currently, the full set of properties as stated in the EDM Definitions 5.2.3 is allowed.

Mandatory Properties

None.

URL Scheme

All *WebResource* entities created by DM2E are identified by an URL with the following scheme:

```
http://dm2e.eu/data/view/provider/[provider]/[identifier]
```

- [provider] is a short data provider identifier, e.g. sbb for Staatsbibliothek Berlin.
- [identifier] is a provider-unique identifier such as a signature or an internal database ID.

Example

Note that the DM2E namespace does not exist yet.

```
@prefix    edm:  <http://www.europeana.eu/schemas/edm/> .
@prefix    dm2e: <http://dm2e.eu/schema/> .
@prefix    dc:   <http://purl.org/dc/elements/1.1/> .
@prefix    dct:  <http://purl.org/dc/terms/> .

<http://dm2e.eu/data/view/provider/sle/abc345/scan.jpg> a
edm:WebResource, dm2e:PageScan ;

dc:description "Scan of the page" ;
dc:format "3D-PDF" ;
edm:rights <http://dm2e.eu/data/agent/provider/sle/sbb> ;
dct:hasPart <http://sle.example/scans/abc345/scan.jpg> .
```

4.2.4 Contextual Resources

In addition to the classes mentioned before, the EDM offers the possibility to use some contextual resources. These are *edm:Agent*, *skos:Concept*, *edm:Place*, and *edm:TimeSpan*. The EDM Definitions 5.2.3 define them as follows:

- edm:Agent: “This class comprises people, either individually or in groups, who have the potential to perform intentional actions for which they can be held responsible.” (EDM Definitions 5.2.3, 2012:9).
- skos:Concept: “A SKOS concept can be viewed as an idea or notion; a unit of thought. However, what constitutes a unit of thought is subjective, and this definition is meant to be suggestive, rather than restrictive. The notion of a SKOS concept is used to refer to specific ideas or meanings established within a knowledge organization system and describe their conceptual structure.” (EDM Definitions 5.2.3, 2012:9).
- edm:Place: “An 'extent in space, in particular on the surface of the earth, in the pure sense of physics: independent from temporal phenomena and matter' (CIDOC CRM)” (EDM Definitions 5.2.3, 2012:13).
- edm:TimeSpan: “The class of 'abstract temporal extents, in the sense of Galilean physics, having a beginning, an end and a duration' (CIDOC CRM)” (EDM Definitions 5.2.3, 2012:14).

These second-level entities describe places, people, time spans, and abstract concepts to which the more complex objects refer to.

Specialisations

In DM2E, the type of a contextualised resource will be made more specific by creating additional subclasses for each contextualised resource. The list of such possible specialisations is work in progress. Exemplarily subclasses for edm:Agent are *dm2e:Publisher*, *dm2e:Library*, and *dm2e:Author*. The exact semantics and the detailed definition of those specialisations is not yet final at this point.

Added Properties

None. Currently, the full set of properties as stated in the EDM Definitions 5.2.3 is allowed.

Mandatory Properties

None.

URL Scheme

All contextual entities created by DM2E are identified by an URL with the following scheme:

```
http://dm2e.eu/data/[agent|place|timespan|concept]/provider/[provider]
/[identifier]
```

- One of [agent|place|timespan|concept] according to the type of contextual resource.
- [provider] is a short data provider identifier, e.g. sbb for Staatsbibliothek Berlin.
- [identifier] is a provider-unique identifier such as a signature or an internal database ID.

Examples

Note that the DM2E namespace does not exist yet.

```
@prefix    edm:       <http://www.europeana.eu/schemas/edm/> .
@prefix    dm2e:      <http://dm2e.eu/schema/> .
@prefix    skos:      <http://www.w3.org/2004/02/skos/core#> .
@prefix    rdaGr2:    <http://rdvocab.info/ElementsGr2> .
@prefix    wgs84_pos: <http://www.w3.org/2003/01/geo/wgs84_pos#> .
@prefix    xsd:       <http://www.w3.org/2001/XMLSchema#> .

<http://dm2e.eu/data/agent/provider/sle/sbb> a edm:Agent, dm2e:Library ;

skos:prefLabel "Staatsbibliothek zu Berlin"@de ;
skos:altLabel "Staatsbibliothek Berlin"@de ;
wgs84_pos:lat "51.507222" ;
wgs84_pos:long "-0.1275" .

<http://dm2e.eu/data/agent/provider/sle/darwin> a edm:Agent,
dm2e:Author ;

skos:prefLabel "Charles Darwin"@en ;
skos:altLabel "Charles Robert Darwin"@en ;
edm:begin "1809-02-18"^^xsd:date ;
```

```
edm:end "1882-04-19"^^xsd:date ;
rdaGr2:placeOfBirth <http://dm2e.eu/data/place/provider/sle/42> ;
rdaGr2:placeOfDeath <http://dm2e.eu/data/place/provider/sle/42> .

<http://dm2e.eu/data/agent/provider/sle/murray> a edm:Agent,
dm2e:Publisher ;

skos:prefLabel "John Murray" ;
edm:begin "1768"^^xsd:year .

<http://dm2e.eu/data/place/provider/sle/42> a edm:Place ;

skos:prefLabel "London"@en ;
skos:altLabel "Londinium"@lat ;
wgs84_pos:lat "51.507222" ;
wgs84_pos:long "-0.1275" .
```

4.3 Conclusion and Next Steps

The first analysis of the heterogeneous sample data provided by the DM2E data providers and the mapping workshop carried out by WP2 clearly showed that, on the one hand, the EDM is in principle able to accommodate any structural and semantic aspect of the source data at a higher integration level but, on the other hand, needs to be specialised in order to practically retain the details of the structural and semantic intricacies.

Therefore, WP2 began drafting the EDM+ as a subset and semantic extension of the EDM. EDM properties, that are necessary to represent the source data at an interoperability level, have been selected and those properties and classes in need of specialisation have been identified. The EDM+ model has been drafted with clarity and ease of use in mind. EDM properties deemed unnecessary have been omitted of the description above.

The next steps foremost include the refinement of the EDM+ through further analysis of the source data mainly by the data providers themselves. This more extensive and detailed mapping to EDM by the data providers will especially allow us to create a more comprehensive list of necessary specialisations.

Furthermore, several important aspects need to be discussed and coordinated with WP1 and WP3 regarding the EDM+ ontological model in the coming months:

1. How can we integrate existing annotations (e.g. from TEI/XML documents) and model newly created annotations in Pundit? This work will be done in close cooperation with WP3 of DM2E.

2. How do we model object data, e.g. TEI/XML data? The EDM currently only focuses on metadata.

3. The EDM+ primarily focuses on rich data modelling for the purposes of DM2E (cf. WP3 activities). However, the issue of data provision to Europeana and possible necessary compromises need to be addressed as data delivery to Europeana is a central part of the DM2E agenda.

4. In this regard, the question of how to deal with Proxies and Named Graphs is important. Currently, DM2E plans to use Named Graphs for the WP3 platforms

Pundit and Korbo[3]. However, Europeana will probably require Proxies for the time being. We need to make sure that the transformation between Named Graphs and Proxies is seamlessly possible.

5. Additional requirements of the data providers towards EDM+ will arise from a questionnaire circulated by WP1 and, as already mentioned, from the detailed data mapping to EDM. WP3 will specify functional requirements which will have ramifications on the EDM+. This feedback needs to be discussed with WP1 and WP3 and then integrated into the EDM+ model.

6. Extension efforts of the EDM should be aligned to certain extent with other similar projects. The Europeana Libraries project, for example, modified the EDM for library specific data (Angjeli et al., 2011). Their view on how to model publication is of high relevance for EDM+.

[3] http://thepund.it and http://korbo.muruca.org [23.07.2012]. More information about the tools can be found in D3.1.

5 RDFisation, Mapping and Translation Workflow

The following part of the report provides a closer look at the tools we have evaluated and used so far for the RDF conversion and alignment with the EDM. These are the first building blocks to be used in the interoperability infrastructure that aims to provide a one-way generic production chain for migrating data from various sources to the EDM and its specialisations as well as for the contextualisation of the object representations. The general workflow is illustrated by Figure 2. In our evaluation, we have considered results from Europeana related activities as well as from generic Linked Open Data oriented development.

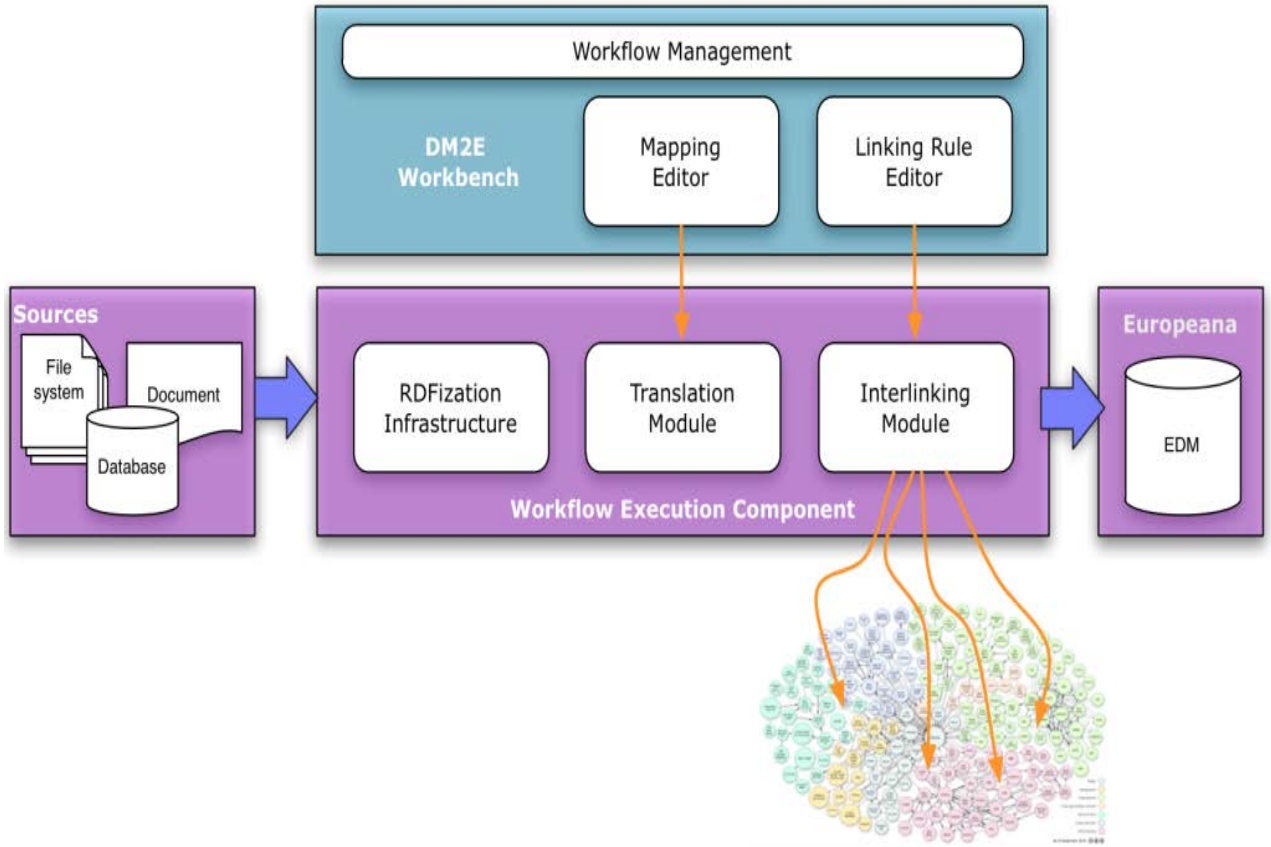

Figure 2: Combination of the workflow execution and management components that will be implemented by WP2. Sources on the left hand side will be formatted into RDF, mapped to EDM or EDM+ and enriched with contextualisation tools (centre) before being delivered to Europeana (right hand side).

The input to the workflow, seen on the left hand side of the figure, consists of metadata collections in various standards or proprietary formats which are received from data providers in txt-format, XML, or as relational database backends. Data providers in DM2E are ONB (Österreichische Nationalbibliothek), SBB (Staatsbibliothek zu Berlin), UIB (Universitetet i Bergen), NLI (National Library of Israel), MPIWG (Max-Planck-Institut für Wissenschaftsgeschichte), UBER (Humboldt-Universität zu Berlin), BBAW (Berlin-Brandenburgische Akademie der Wissenschaften), CJH (Centre for Jewish History), JDC (Joint Distribution Committee), UBFFM (Universitätsbibliothek Johann Christian Senckenberg Frankfurt am Main), and DFGA (Digitale Faksimile Gesamtausgabe). Their metadata are modelled using different formats, e.g. TEI, MAB2, MARC, METS/MODS, or stored in relational databases (see 5.3 for details). The workflow execution component, in the centre of the figure, includes the RDFisation infrastructure, the translation module that

is based on mapping editors and the interlinking module that is based on linking rule editors. Tools that enable the RDFisation of the input data are MINT (also offering a mapping editor) when dealing with XML files and D2R when the input data is stored in databases. All tools for RDFisation can be adapted to use the specialisations required by DM2E. For tools that convert RDF to RDF, the adaptations are trivial whereas XML-to-RDF tools have to be made aware of the extensions through adapted transformation tools, i.e. customised XML Schema Definitions and XSL stylesheets. The interlinking module is based on the Silk Link Discovery Framework. In addition, we might decide to include the Amalgame tool in the future. At this point of the project, we have only tested the Silk framework for contextualisation (see 7.1). The output of this component is enriched EDM content that will be integrated into Europeana and delivered to our own annotation platform, Korbo/Pundit.

Since Europeana is currently not able to handle the EDM+ extension of EDM directly, the ingestion process into Europeana will be two-fold: DM2E will store the rich and versatile data in the format described in section 4 on its own infrastructure and provide a simplified version that adheres to the scope of EDM currently supported by Europeana as data dumps or via a SPARQL endpoint.

5.1 MINT Platform

MINT (Drosopoulos et al., 2012; Kollia et al., 2012) is a framework for managing and transforming XML files. An easy-to-use interface and schema support allows the user to create re-usable XSL stylesheets without having to edit anything by hand.

> *"MINT services compose a web based platform that was designed and developed to facilitate aggregation initiatives for cultural heritage content and metadata in Europe. It is employed from the first steps of such workflows, corresponding to the ingestion, mapping and aggregation of metadata records, and proceeds to implement a variety of remediation approaches for the resulting repository. The platform offers a user and organization management system that allows the deployment and operation of different aggregation schemes (thematic or cross-domain, international, national, or regional) and corresponding access rights. Registered organizations can upload (http, ftp, oai-pmh) their metadata records in xml or csv serialization in order to manage, aggregate and publish their collections."*[4]

The general workflow in MINT starts with the ingestion of structured or semi-structured data and continues with the establishment of crosswalks to a reference schema in order to take advantage of a well-defined, machine understandable model. The underlying data serialisation is in XML, while the user's mapping actions are registered as XSL transformations. The common model functions as an anchor, to which various data sources can be attached and become, at least partly, interoperable.

The key functionalities include:

- Organisation and user level access rights and role assignment
- Collection and record management (XML serialisation)
- Direct import and validation according to registered schemas (XSD)
- OAI-PMH based harvesting and publishing
- Visual mapping editor for the XSLT language

[4] http://mint.image.ece.ntua.gr/redmine/projects/mint/wiki/Introduction [23.07.2012]

- Transformation and previewing (XML and HTML)
- Repository deployment and remediation interfaces

Users define their metadata crosswalks with the help of a visual mapping editor for the XSL language (Figure 3). The mapping is performed with simple drag-and-drop or input operations which are then translated to the corresponding code. The mapping editor visualises both, the input and target XSD, in an intuitive interface that provides access and navigation of the tree structure and data (statistics and value lists) of the input schema, and the structure, documentation and restrictions of the target one.

It supports string manipulation functions for input elements in order to perform 1-n and m-1 (with the option between concatenation and element repetition) mappings between the two models. Additionally, structural element mappings are allowed, as well as constant or controlled value (target schema enumerations) assignment, conditional mappings (with a complex condition editor), and value mappings between input and target value lists. Mappings can be applied to ingested records, and be edited, downloaded and shared as templates between users of the server.

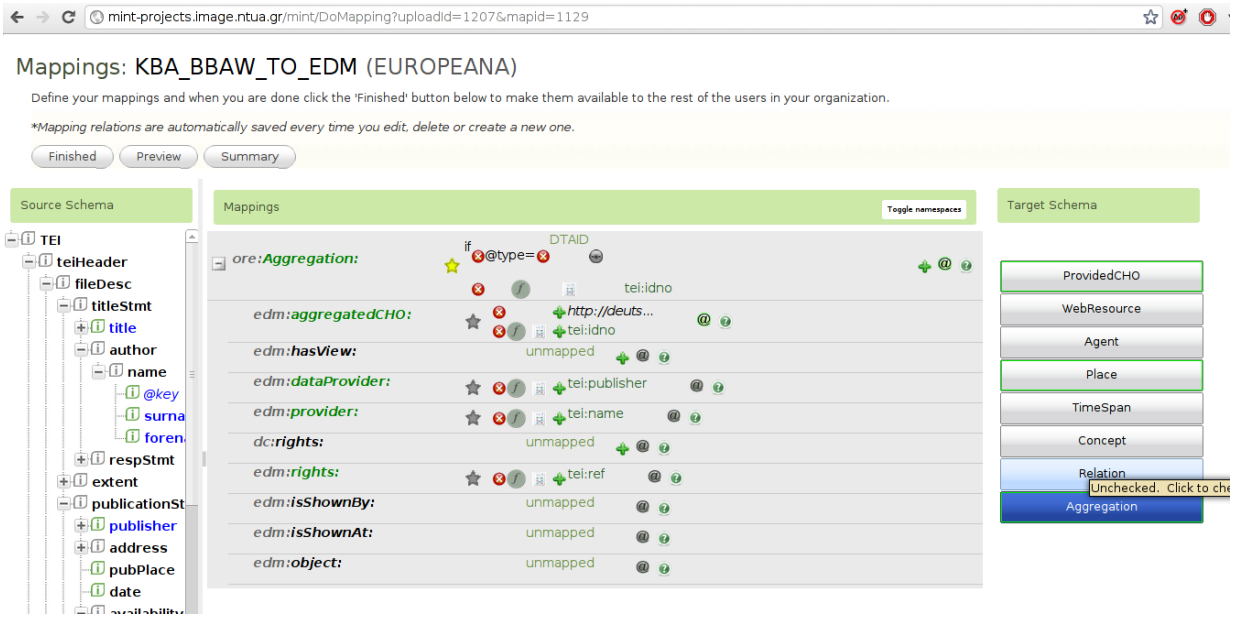

Figure 3: MINT tool in action: This screen shows the creation of a mapping from a TEI flavour (in this case a document provided by BBAW) to the EUROPEANA target schema (EDM). In this particular example, an ore:Aggregation class is created for every local ID with an edm:aggregatedCHO property linking to the concatenation of a fixed base URL and the ID and some other properties set to literals found in the source text.

Preview interfaces, like the one in Figure 4, present to users the steps of the aggregation including the current input XML record, the XSLT of their mappings, the transformed record in the target schema, subsequent transformations from the target schema to other models and available HTML renderings of each XML record. Users can transform their selected collections using complete and validated mappings in order to publish them in available target schemas for the required aggregation and remediation steps.

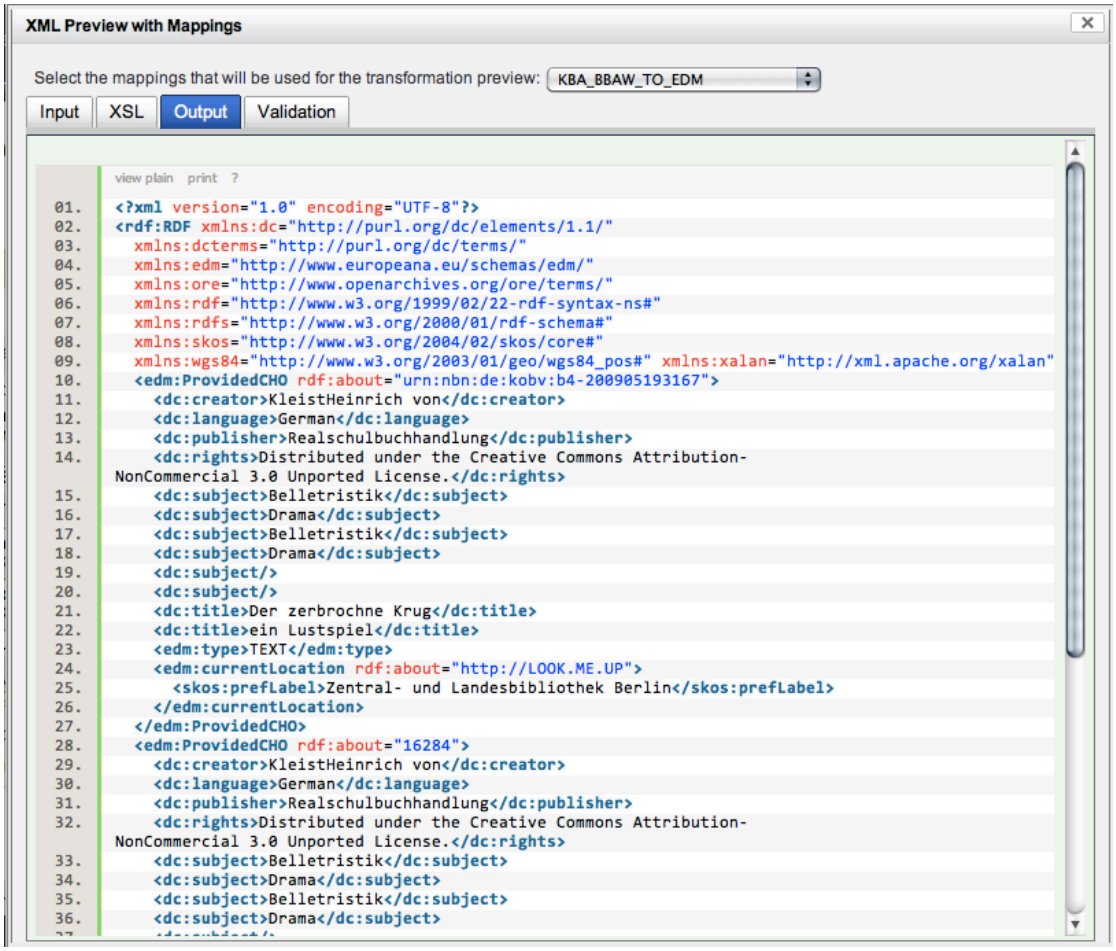

Figure 4: Preview of a transformation using a mapping created in MINT. The result is a set of RDF resources, serialised in RDF/XML, that represents the corresponding input record using the EDM vocabulary.

5.1.1 The Use of MINT within the Project

During this first deployment, MINT was used to create concept and example mappings to transform sample provider data into RDF. The following provider formats were used to create respective mappings to the current EDM XML schema:

- MARC/XML, xmlns=http://www.loc.gov/MARC21/slim (ONB, NLI)
- MAB (ONB)
- MAB2 (SBB)
- Max-Planck-Institut für Wissenschaftsgeschichte model (MPIWG)
- TEI/XML, xmlns=http://www.tei-c.org/ns/1.0 (UIB, UBER)
- TEI P5 (BBAW)

The result guided the evaluation of DM2E providers' input data (see section 6 for a complete analysis) and the requirement analysis regarding the specialisation of EDM. DM2Es MINT specialisation is accessible via http://mint-projects.image.ntua.gr/dm2e.

5.1.2 Target Schema Configuration

MINT transforms one form of XML into another form of XML using XSLT. The stylesheet rules are created using a GUI by dragging and dropping input XML elements to target XML

elements. Those target XML elements are defined by an XSD schema and a MINT-specific configuration in JSON to define the GUI. Users with specific rights in MINT can add those XSD output schemas, so providers can create mappings and apply them to the input XML. Once the XSD schema has been created, it can be adapted on the server and re-loaded by clicking a button.

Step 1: XSD

With MINT being heavily based on XSD, the first step in creating a target schema usable in MINT is to create an EDM+ XSD schema. There are quite a lot of Europeana related ontologies already available as XSD schemas in the released version of the MINT tool, as well as schemas for well-known ontologies. The following is a non-exhaustive list of target schemas supported in MINT.

- carare
- dc
- dcmitype
- dcterms
- dcwrap
- EBU
- eclap
- EDM
- EDMDC
- EDMDCTERMS
- EDMSchema
- ESE
- EUScreen
- lido
- MARC
- mods
- OWL
- RDAGR
- RDF
- simpledc
- SKOS
- WGS
- xml

Since XSD supports inheritance, customised schemas can be created by combining several XSD schema files and since DM2Es EDM+ (see Section 4) is based on widely-used, standard ontologies such as BIBO, DCTERMS, and SKOS, which are in part already covered in MINT, creating the EDM+ target schema in MINT is mostly about creating schemas for those ontologies not yet covered in MINT (improving the reusability of the toolchain in the process), combining those schemas to a unified new schema and copying/pasting/renaming some class definitions for the various intended specialisations of *edm:ProvidedCHO*, *Agent*, *Topic*, *Location*, *TimeSpan*, and additional classes (e.g. *dm2e:Book*, *dm2e:Author*).

Step 2: JSON

To set up an XSD schema as a target schema in MINT, a thin configuration layer has to be created in the form of an accompanying JSON document with the following elements:

- version - Configuration file version
- xsd - The XSD filename
- namespaces - An array containing target namespaces and prefixes
- wrap
- item
- paths
  - item - The item level XPath
  - label - The item label XPath
  - id - The item record ID XPath
- groups
- navigation
  - type
    - label
    - group
  - name - The name of a group
  - label - The label used for this button
  - hide - An array of elements to hide from button
  - include - An array of elements to include in button.
- preview
- publication
  - type - The type of transformation during publication (values: xsl)
  - value - The parameter for publication transformation (i.e. the xsl file for transformation)
- customisation

This file defines how the GUI items are arranged, what type of mapping is offered to the user and which labels are used for naming the target classes and properties. Furthermore, XSD elements can be grouped in a way that is better from a usability perspective rather than adhering rigidly to the original hierarchy.

We are currently working on setting up target schemas required for the EDM+. This is usually a task that is not meant for the average user, but for an aggregation manager, as this is also one of the less documented parts of the MINT source. Nevertheless, through the project's efforts regarding specialisation of EDM, there may be cases where faster test deployment of a target schema will be required and in this regard, NTUA will look into revising the code and offering a more intuitive administration.

5.1.3 Evaluation and Next Steps

MINT is a fitting tool for XML-to-RDF transformation for several reasons: First of all, it is based on XSLT, a standardised language for the transformation of XML documents that can handle almost all cases of mapping many-to-many elements with support for pattern matching. Secondly, its very intuitive web-based user interface with drag-and-drop

mapping and instant previews requires little knowledge of the technical details from the providers and allows a rapid development of mappings. Thirdly, the main developers of the tool are DM2E partners which are working on adapting the tool to DM2E's and Europeana's requirements. Last but not least: MINT has been serving numerous cultural heritage aggregation projects, such as ATHENA[5], ECLAP[6], JUDAICA[7], and DCA[8], and has been proven to work for the XML-to-EDM mapping scenario in other Europeana-related projects, such as EUScreen[9], CARARE[10], and Linked Heritage[11]. Finally, MINT is used for the prototyping of EDM, while also being an integral part of Europeana's United Ingestion Manager (UIM), serving as a mapping, processing, reconciliation, and remediation engine for XML records and RDF resources.

Even though the RDFisation workflow can be split into RDFisation and contextualisation, there is the need for custom callbacks to programmatic code for advanced features like sequence extraction, intra-linking elements, resource creation from literals and more. Because MINT is written in Java, the task should be as simple as writing the custom functions in Java and providing those functions to the XSLT engine in a special namespace. However, MINT uses the limited free version of the Saxon toolkit, which does not support custom functions. There are three different ways to handle this: Either by adding preprocessing and/or postprocessing steps to programmatically alter the XML (and duplicating and de-generalising the features of XSLT), by upgrading Saxon to the commercial edition (and running into all kinds of legal issues about re-distribution and open sourcing MINT) or by changing the XSLT framework to another free software like the Apache-licensed Xalan (and removing other vital functionality that only the XSLT 2.0 capable Saxon supports). This issue will continue to be addressed as we gather more information through subsequent experiments with real data and transformations. The availability of local unique identifiers in the input is considered essential and will guide the URI generation policy and implementation.

The next steps towards the use of MINT for the transformation of provider data to DM2E's specialisation of EDM involve finalising the EDM+ data model and publishing a provisionally authoritative version WP-wide. Based on this informal EDM+ ontology, we can formalise EDM+ to XSD. Finally, we will collect further requirements for specific interface settings (i.e. hiding, wrapping etc.) in a Wiki page, in order to know how to adapt the configuration file.

5.2 jMet2Ont

In addition to MINT, we have also evaluated the jMet2Ont tool[12]. jMet2Ont (Walkowska & Sielski, 2012) is a Java-based tool for transforming XML to RDF, developed by Justyna Walkowska and Krzysztof Sielski at the Poznan Supercomputing and Networking Center. It is available under the GNU General Public License version 3. The tool is different from XSLT frontends such as MINT in that the transformation target is actually RDF, not RDF/XML. That implies that this tool works with the semantics of RDF graphs and is not just using a simple syntax and hierarchy as it is the case with XML in general and with

[5] http://www.athenaeurope.org [23.07.2012]
[6] http://www.eclap.eu [23.07.2012]
[7] http://www.judaica-europeana.eu [23.07.2012]
[8] http://www.dca-project.eu [23.07.2012]
[9] http://lod.euscreen.eu [23.07.2012]
[10] http://www.carare.eu [23.07.2012]
[11] http://www.linkedheritage.org [23.07.2012]
[12] http://fbc.pionier.net.pl/pro/jmet2ont/index.html [23.07.2012]

XSD/XSLT in particular. Furthermore, the transformation is based on graph paths ("Ontology paths"), not XPath. jMet2Ont handles multiple levels of distinguishable entities in the target schema which is useful since XML records and their hierarchy do not map well to RDF resources. The path approach allows generating "empty"/"parent" resources on demand. External information in the form of XML files can be merged into the RDF document at mapping time.

One major issue with jMet2Ont is the syntax of the transformation file: It is obviously very different from XSLT and not very well documented (though documentation is excellent considering that the tool is in an early development state). The heavy reliance on Regular Expressions with capturing groups does not make it easier to use. There is no GUI at the moment, the tool can be used by calling the Java class with a custom properties file and the source file or directly be accessed from other Java code that has to be developed first. It is unrealistic to teach the configuration file syntax to the data providers.

Relying on jMet2Ont at this point would require active development of the tool and a frontend from our part, which is at the moment not sensible. Other tools might also not be easy to use, but those tools can directly be upgraded and adapted to our purposes as their developers are also contributing to DM2E. We should keep an eye on the development of jMet2Ont though, as this might work for cases where MINT will not do.

5.3 The D2R Platform

The D2R platform enables mappings of data from relational databases to RDF (Bizer & Cyganiak, 2006). The platform was built by Christian Bizer and Richard Cygnaniak at FUB. It is coded in Java and can be accessed via command line. By using D2R Server, which is part of the D2R platform, one can additionally navigate the content of databases with an HTML or RDF browser or by making SPARQL queries. The transformed RDF data can thus be browsed and searched (Bizer & Cyganiak, 2006).

5.3.1 The Use of D2R within the Project

One of the projects aims is to directly address databases of the providers and extract metadata without losing any information and with data that is always up-to-date. This can be done with the use of D2R. We tested D2R with a data dump of the MPIWG VLP (The Virtual Laboratory)[13] database.

The Virtual Laboratory is a real live platform where scientists publish and discuss their research on experimentation in life science, art, and technology. It is based on collections and presentations of texts and images concerning various aspects of the experimentalisation of life. The digitised objects, that build the collections, are representations of instruments, experiments, sites, and people between 1830 and 1930. Thus, the platform operates as an archive as well as an laboratory for current research that can be presented and shared by researchers and students.

The database, that forms the backbone of the platform, consists of 12 tables. Not every table contains data that is suitable for being published as Linked Data. The ones that are used with D2R are thus the tables vl_categories, vl_concepts_kwindex, vl_essays, vl_experiments, vl_images, vl_literature, vl_movies, vl_objects, vl_people, vl_sites, vl_technology, and vl_transcript. The structure of the database includes eight different

[13] http://vlp.mpiwg-berlin.mpg.de/index_html [23.07.2012]

object types which can be modelled as entity of the class edm:ProvidedCHO. Each digital object is stored in one table of the underlying database.

5.3.2 Requirements for the Mappings of Relational Databases with the D2R Server

By creating a default mapping for the MPIWG sample database with the D2R Server, we found out that most of the tables do not have a primary or foreign key defined. Without unique identifiers, we can neither map content to the EDM nor use D2R properly. Therefore, some parts of the database have to be adjusted before we can start the mappings. In the case of primary keys, this was easy to achieve. For example, the table vl_people has a unique value identifier for each tuple in the reference attribute that could be defined as primary key for this table. Similar unique values were found in the other tables as well.

Other preparatory work that had to be done was:

- Check ID consistency

  There are *NULL* values in the “reference” attribute of the vl_concepts_kwindex table that we wanted to use as a primary key. The "reference" attributes in the other tables can be used as primary keys, rows with *NULL* values in vl_concepts_kwindex had to be deleted first. Otherwise, the IDs were consistent.

- Create primary keys

  The tables already have identifiers (attribute "reference") which have to be changed into primary keys or marked as identifiers in D2R. That was done indirectly with D2R.

- Create foreign keys

  There are identical attributes in different tables, but they are not declared as foreign keys. Sometimes, they have a slightly different syntax (e.g. lit12345 and 12345). This has to be checked and then we have to add them as foreign keys. Those indirect foreign key attributes are mostly named "source".

- Check and delete columns without values

  There are some attributes that do not store anything and which can be deleted. Sometimes, they are NULL, sometimes they contain empty strings.

After completing this first steps, we were able to work with the tool.

5.3.3 D2R Mappings

First of all, a Postgresql database was created for test issues. The data dump was successfully imported into the new database. The D2R-Server for the MPIWG data was then installed on the projects’ server and is running at http://dm2edev.hu-berlin.de/d2r-mpiwg2.

In the following, we will show exemplary on the table vl_essays how the mappings were conducted.

After having indicated the adjusted values of the “reference” column as primary keys, mappings to *ore:Aggregation* were proceeded. All entities of *ore:Aggregation* will be identified by an URL with the following scheme:

```
http://dm2e.eu/data/aggregation/provider/[provider]/[identifier]
```

- [provider] is a short data provider identifier, e.g. mpiwg for the Max-Planck-Institut für Wissenschaftsgeschichte. We have two possibilities to deal with that: either we make manual mappings with D2R or we create an ID directly in the database.
- [identifier] is a provider-unique identifier such as a signature or internal database ID. In our case, we used the values of the reference column.

The scheme is the same as the one that we proposed in section 4 of this report where we explained our slightly modified EDM+. At this state of the mapping procedures, we have not yet used this scheme. Instead, the test mappings were created with the following temporary URI of our developing server:

```
http://dm2edev.hu-berlin.de/d2r-[provider]/resource/[table]/[identifier]
```

Note that this scheme was created in a first text scenario and will be changed if we will use the tool for further mappings. It still remains an open question if we should include the name of tables into the definite RDF namescheme or if we can leave them out without losing important connections between resources.

Mappings to *edm:ProvidedCHO* will be created analogue to mappings to *ore:Aggregation*. Thus, all *edm:ProvidedCHO* entities will be identified by an URL with the following naming scheme:

```
http://dm2e.eu/data/item/provider/[provider]/[identifier]
```

Classes apart from *ore:Aggregation* and *edm:ProvidedCHO* will be created as described in section 4.2.4. As mentioned before, we have to figure out if the name of the table should be added to the nameschemes that we decided to use for our RDF representation.

Figure 5 provides a look on a resource with first mappings to the EDM and remaining provider specific entities. The exemplary resource “art10” is not yet mapped to all elements that are required by the model (see for example the mandatory elements of the EDM in Figure 1), but those can be added with default values in D2R.

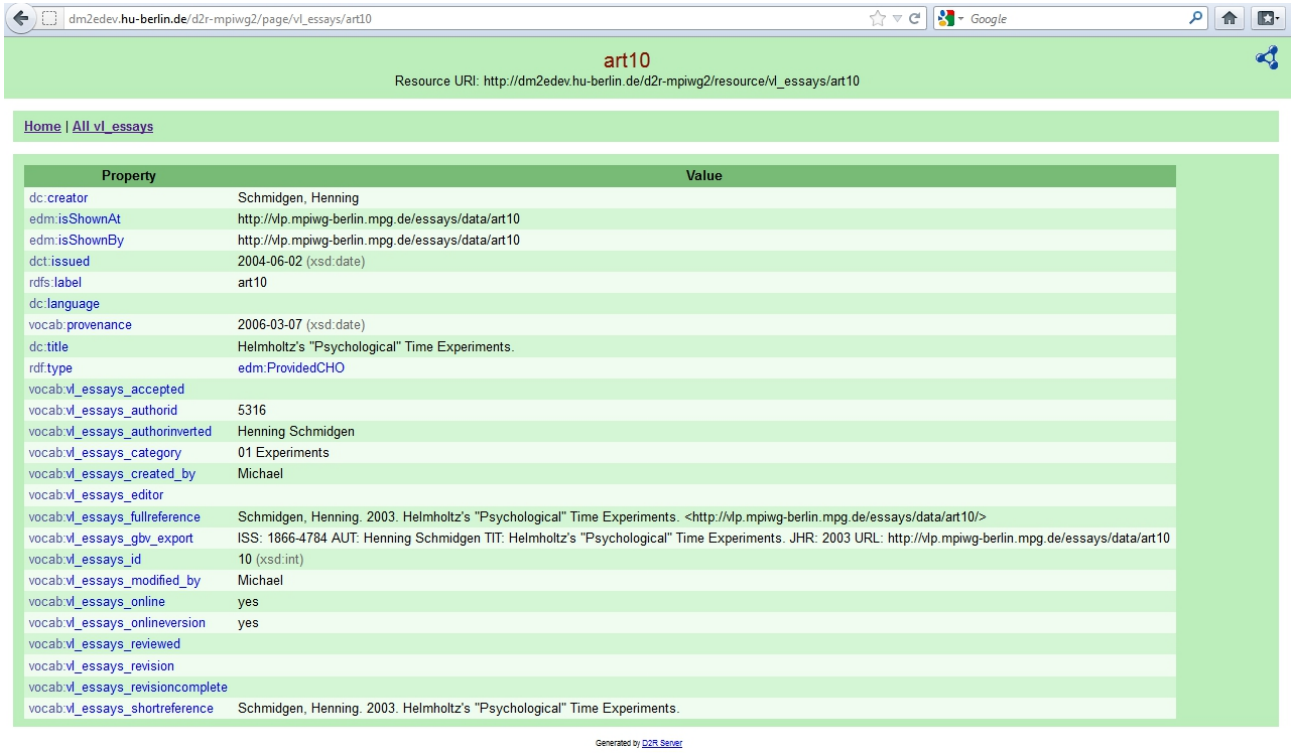

Figure 5: D2R accessed with the Firefox browser. This screenshot shows the resource "art10" with entities that are already mapped to the EDM and with some additional entities that are not yet mapped.

The first D2R mappings indicate that the later mappings of databases to the EDM or EDM+ can be very challenging. We were able to produce RDF data but, as you can see in the output in Figure 5, many triples are not yet mapped. The database that we have tested is very detailed which may lead to a lot of adjustments of our EDM+ model. By producing RDF out of databases, we have some more steps to take. Not only do we need tools like D2R for transforming the datasets, but we also have to check the data and handle things like missing keys, information stored in large text fields or empty fields and columns.

The first steps that we have taken with D2R are not that large, but they show how the transformation from databases to RDF could look like and that mappings to the EDM are possible.

5.4 ClioPatria, XMLRDF and Amalgame

Apart from the mentioned tools, there are still some tools left that were not yet tested in the first months, but which could nevertheless be useful for our purposes and be part of the later toolchain. Tools that we might analyse later in addition to the tools that have been tested so far are Amalgame and XMLRDF.

Amalgame and XMLRDF (Wielemaker et al., 2011) are components of a workflow for bringing XML metadata and vocabulary information to Europeana. They are built on top of the ClioPatria framework (Schreiber et al., 2006), a software suite for handling RDF data and creating HTTP web services which is written in SWI/Prolog. They were further developed and used in the Europeana Connect Project[14].

[14] http://www.europeanaconnect.eu [23.07.2012]

XMLRDF[15] is, as the name implies, a tool for converting XML to RDF, developed by Jan Wielemaker from 2008 to 2011 and refined by Victor de Boer, Steffen Hennicke, and Antoine Isaac.

Amalgame[16] serves the purpose of enriching RDF data with external information contained within local XML files, relational databases or structured information on the Web such as Linked Data. It is developed by Jacco van Ossenbruggen, Michiel Hildebrand, Antoine Isaac, and Victor de Boer.

5.5 Conclusion

The initial version of the infrastructure currently combines the D2R Platform for RDFisation of relational data, the Mint platform for the RDFisation of XML data with the Silk Link Discovery Framework for the contextualisation of the RDFised data.

In this section, we reported about the experiments that were conducted with the tools that are part of the DM2E interoperability infrastructure or might be added to the infrastructure in the future.

Our experiments showed that the MINT platform as well as the D2R platform are in principle capable of translating the input data from the WP1 content providers into the EDM+ model, but still need to be extended in order to cope with all aspects of the transformation. These extensions will be implemented in the next months.

[15] http://semanticweb.cs.vu.nl/xmlrdf [23.07.2012]
[16] http://semanticweb.cs.vu.nl/amalgame [23.07.2012]

6 Requirements for Input Data and Missing Information

The first round of analysis of the provided sample data has shown that the metadata we received lacks certain information which are necessary to meet the minimal requirements of the EDM. Those minimal requirements are shown in Figure 1 and further explained in the EDM documentation (Definition of the Europeana Data Model elements Version 5.2.3, 2012). During these first mapping exercises we also looked at the possibilities to either link to existing contextual resources or to create new ones (also cf. section 4).

In the following, we will give a review of the providers' sample data and point out information which is missing or is incomplete but necessary to meet the EDM's minimal requirements. This review is preliminary in that it will be revised by the data providers which will create the definite mappings of their data to the EDM in the coming months.

Sample data encoded in the following metadata formats have been analysed for this report:

- Encoded Archival Description (EAD): SBB
- Machine-Readable Cataloging (MARC/XML): ONB, NLI
- Text Encoding Initiative (TEI/XML): BBAW, UBER, UIB
- Provider-specific models: MPIWG

They cover most of the models used by the data providers in the project. The following metadata formats have not been analysed for this report:

- Maschinelles Austauschformat für Bibliotheken (MAB2): SBB, ONB
- Open Archives Initiative for Metadata Harvesting (OAI-PMH): UBFFM

At the time this report was written there was no proper MAB2/XML export available. An automatic translation of the MAB2 txt-file was not easily possible. As UBFFM is a new data provider who joined DM2E at a later stage, their data will be mapped later.

6.1 Missing Elements in the Sample Data

This section gives for each sample data set a brief characterisation of its contents, a summary of the missing information values with regard to mandatory EDM properties and an assessment of the possibilities to either link to existing contextual resources or to create new resources from existing data values. We also point out general issues which arose during the mappings like, for example, a lack of documentation.

6.1.1 Encoded Archival Description (EAD)

Data provider: SBB

The EAD data provided by SBB was a test export which is still work in progress. Furthermore, MAB2 probably will be the main export format for the SBB data. The mapping of MAB2 to EDM is described in D1.1. The tested example was rich and had most necessary elements included, but other examples may not be that rich, as the provider SBB already indicated. Therefore, the missing elements mentioned below are an educated guess.

Missing elements:

ore:Aggregation	edm:rights edm:provider edm:dataProvider	for the sub-levels: edm:isShownBy or edm:isShownAt
edm:ProvidedCHO	dc:coverage, dc:type, dc:subject or dct:spatial	

Table 3: Missing mandatory elements SBB (EAD).

The test mapping of the sample data to EDM showed that EDM is able to represent the complex and hierarchical structure of a finding aid. This result confirms earlier and similar findings of mapping of EAD data to the EDM (Hennicke et al., 2011).

The EAD sample data contains several fields which are candidates for creating resources. As already mentioned, the completeness and quality of the records will vary.

Several XML elements and attributes contain literals without any ID pointing to a concept. However, because their values appear to originate from a controlled vocabulary, those XML elements are candidates for creating concept resources. Therefore, it will be easy to create new contextual resources from those values. Examples are <genreform>, <c level="item">, and <physfacet type="levelOfGenesis">.

Especially the level information is important. This information is a candidate to create a small-scale vocabulary from it, i.e. a taxonomy of resources describing the hierarchical relation between the levels of the hierarchical archival description found in the EAD help files. Each level in the EAD hierarchy has its own unique ID which can be used to create a URI for the resource in EDM.

Some XML elements have attributes with an additional type of information in the form of a code (e.g. <genreform type="S">Standortkonvolut</genreform>). We need explanations for those codes. For some XML elements, the relation type is given in an attribute (e.g. the attribute role in the element <persname normal="Ranke, Leopold von" role="creator" authfilenumber="118598279" source="GND"/>).

For creating *Places*, candidates with literal values like <geogname role="placeOfOrigin">Frankfurt/Oder</geogname> can be used. The relation type is given in an attribute.

For creating *Agents*, we have IDs from the Gemeinsame Normdatei (GND)[17]. However, the source of the ID is not always given (for example, for <corpname> this information is missing). Otherwise, it is easy to expand the GND ID to a working URI pointing to the appropriate GND resource.

The SBB provides the metadata records in MAB2 format as well. As the MINT mapping tool has been developed to handle XML encoded data files, MAB2 files currently cannot be transformed with this tool. Therefore, we have worked in the initial data analysis with the EAD/XML formatted data. Because of the deeper granularity of the MAB2 format and semantically richer data, it is necessary to provide and develop a transformation workflow for MAB2 encoded data as well. Therefore, the SBB is currently working on the development of such a transformation tool, which can be used for transformation purposes from MAB2/TXT into MAB2/XML by the other project partners.

[17] The GND is a universal authority file that combines bibliographic data of several other authority files like the PND, SWD or GKD. It is used and managed by German-speaking libraries and held at the German National Library. More information can be found here: http://www.dnb.de/EN/Standardisierung/Normdaten/GND/gnd_node.html [23.07.2012].

6.1.2 Machine-Readable Cataloging (MARC/XML)

Data provider: ONB

Example: 170 Codices

The sample data about the 170 codices from the ONB are described and delivered to DM2E in MARC/XML format. Each record has an unique identifier which can be used by DM2E. We can create some contextual resources like *edm:Agent* and *edm:Place* from the source data as well.

Missing elements:

ore:Aggregation	edm:rights edm:dataProvider edm:isShownBy or edm:isShownAt
edm:ProvidedCHO	edm:type

Table 4: Missing mandatory elements ONB, Codices (MARC).

For further contextual resources, e.g. edm:TimeSpan, it is necessary to normalise the input data because of some inconsistencies. For example, the metadata field including information about time spans is delivered in two different ways: “1. Hälfte 5. Jhdt., 400-450” and “1580-1599, Ende 16. Jhdt.”. It would be very useful if we had the possibility to extract the numerical time span in order to make this information searchable. In this case, we need a consistent string which we can handle with regular expressions. The MINT tool as well as the D2R server support regular expression functionality and were able to extract the necessary information.

Furthermore, we have a problem to indicate the right URL that can be used as a web representation of the object in the ONB data. It appears, that the same data fields store more than one URL. For example, the data field tag with the number 856u contains the following links: http://www.onb.ac.at/sammlungen/hschrift/bibliographie.htm, http://www.bildarchiv.at/ProfiSzettel.aspx?a=b&wort=Cod%2015&Wien and http://www.manuscripta-mediaevalia.de/hs/katalogseiten/HSK0751a_b0002_jpg.htm. Only the last one leads to a proper landing page for the digitised manuscript page. This can only be sorted out manually.

Additionally, regarding the EDM, it is necessary to have information about the rights of the data and the document type, that means providing either "TEXT", "VIDEO", "IMAGE", "3D", or "SOUND". This information can be added manually in the preparation of the mappings with MINT or D2R.

Example: 50,000 Google Books

Missing elements:

ore:Aggregation	edm:rights edm:dataProvider edm:isShownBy or edm:isShownAt
edm:ProvidedCHO	edm:type

Table 5: Missing mandatory elements ONB, Google Books (MARC).

The second data collection provided by the ONB includes 50,000 online books. The metadata records of the Austrian online books have the same ambiguity here as in the

sample data examples described above. Semantically different information is stored in the same data fields. For example, the data field “tag 650a” includes terms like "Gewaltenteilung", "Politische Philosophie", "Bellarmino, Roberto", "Kritik". We are not able to explain relations between these terms. In order to retain this information, we need more specific input from the data provider.

In this data set, we are missing the mandatory input data about the URIs of the digitised documents showing the described object. When the digitisation process is finished, the missing information will be included into the metadata records as well.

The ONB provides all metadata in MAB2 format as well. For the initial data analysis we have decided to work with the MARC/XML formatted data because of the input requirements of the MINT Tool. Some of the MAB2 fields cannot be mapped to MARC/XML without losing semantically important information of the CHO. That way, one of our future objectives in WP1 and WP2 will be the discovery, improvement and implementation of additional importing and mapping functions within the interoperability infrastructure that can handle the MAB2 format as well.

Data provider: NLI

Example: Books

Missing elements:

ore:Aggregation	edm:rights
edm:ProvidedCHO	

Table 6: Missing mandatory elements NLI, books (MARC).

Example: Manuscripts

Missing elements:

ore:Aggregation	edm:rights edm:isShownBy or edm:isShownAt
edm:ProvidedCHO	

Table 7: Missing mandatory elements NLI, manuscripts (MARC).

The NLI provides books and manuscripts to DM2E. By analysing their data, we could only find missing elements in *ore:Aggregation*. We need to know which rights the metadata have. This information can be committed as a default value. More problematic are missing web resources that can be mapped to *edm:isShownBy* or *edm:isShownAt* for the manuscripts. It would be great if the URLs could be added to the metadata records.

6.1.3 Text Encoding Initiative (TEI / XML)

Data provider: BBAW

The BBAW provides digitisations of German books from the 16th to the 20th century. Almost all necessary elements are represented in the metadata record. The only missing elements are:

ore:Aggregation	edm:isShownBy or edm:isShownAt
edm:ProvidedCHO	

Table 8: Missing mandatory elements BBAW (TEI P5).

There is no URL of a web resource for the CHO in the record. Without this URL, the mappings to the EDM are not complete because Europeana cannot link to the object. Linking to a concatenation of "http://deutschestextarchiv.de/" and the tei:idno[@type='DTAID'] does work, but we do not know for sure if these URLs will remain stable.

Smaller problems may be that *Agents* are not modelled in a consistent way. The author of a book (*dct:creator*) has a forename and a surname as well as an identifier from a controlled vocabulary, the PND. Providers or data providers can also be represented as agents. In the case of BBAW, these are institutions, so they do not have a fore- and surname but a physical address and an email address. The physical address is represented by a string value including a street, house number, postcode, and town. It would be better to split this information into separate entities before performing the mappings.

Data provider: UBER

The UBER metadata describes articles of the German "Polytechnisches Journal". Each record consists of two CHOs: an article and the concrete journal in which the article is published. As every *edm:ProvidedCHO* is exactly linked to one *ore:Aggregation*, each metadata record includes always two CHOs and two aggregations. The connections between those CHOs, as well as the connections between articles of the same volume that are represented by different metadata records, have to be discussed.

Missing elements:

ore:Aggregation	edm:isShownBy or edm:isShownAt (regarding the journal)
edm:ProvidedCHO	dc:coverage, dc:type, dc:subject or dct:spatial (article and journal) dc:language (article and journal) edm:type (article and journal)

Table 9: Missing mandatory elements UBER (TEI).

The missing mandatory elements are URLs of web resources for *edm:isShownBy* or *edm:isShownAt* in the class *ore:Aggregation* for the journal and *dc:coverage*, *dc:type*, *dc:subject*, or *dct:spatial*, *dc:language,* and *edm:type* for the journal, as well as for the concrete article. The missing Web resource for the journal exists, but is not part of the metadata record. It can be mapped by hand.

Other issues include diverse metadata creators with slightly different roles like supporters or cooperating organisations. It is not always trivial how to map them. Titles do not seem to be normalised. For example, there were text fields with "Polytechnisches Journal" and

"Dingler-Online | Das digitalisierte Polytechnische Journal". It is not clear which spelling should be favoured.

For the class *edm:Place*, there are no additional information like coordinates. Thus, it is not always clear which place is exactly meant, but getting coordinates is nothing that we can really expect. People, stored in *edm:Agent*, are not made explicit with identifiers or connections to other controlled vocabularies.

Data provider: UIB

The mandatory elements in a TEI record can usually be found in its header. Unfortunately, the TEI headers of the UIB records are not that expressive in opposite to the very rich text annotations that are great for the use in the WP3 platforms. This leads to elements that are not always missing but difficult to extract:

ore:Aggregation	
edm:ProvidedCHO	dc:title or dc:description dc:coverage, dc:type, dc:subject, or dct:spatial edm:type

Table 10: Missing mandatory elements UIB (TEI).

The header mainly consists of administrative metadata, like, for example, personal and organisational responsibilities for metadata creation, funding, and digitisations. Descriptive metadata about the CHO is almost completely missing here: There is an author and Web resources for each page of the Wittgenstein archive but we do not have titles or descriptions for the CHOs. Elements that we could map to *dc:coverage*, *dc:type*, *dc:subject*, or *dct:spatial* are also missing.

It is also not clear what the boundaries of a single CHO should be as each metadata record describes several pages of the Wittgenstein archive. If each page or each annotation snippet should be mapped solely, we need metadata for all of them. It seems that we have a lot of default values for these records that are just not represented in the records itself. UIB must provide a list of default values.

6.1.4 Individual Formats

Data provider: MPIWG

The sample database, which we are trying to connect and map via D2R, is called “The Virtual Laboratory - Essays and Resources on the Experimentalization of Life” and is a real life platform where historians publish and discuss their research in the fields of life sciences, arts, and technology. The structure of the database includes eight different tables or object types which can be modelled as entities of the class *edm:ProvidedCHO*. Each digital object is stored in exactly one table of the back-end database. The tables are named vl_essays, vl_experiments, vl_technology, vl_objects, vl_sites, vl_people, vl_concepts_kwindex, and vl_literature.

In the following, we provide a look on two of them.

VLP Database, Table vl_essays

Having a closer look at this table we have discovered some missing information for the mapping to the EDM like data that can be mapped to *edm:rights*, *edm:provider*, and *edm:dataProvider*. After having processed mappings with D2R, each essay object got an unique URI and can be described as *edm:ProvidedCHO* or *ore:Aggregation*.

ore:Aggregation	edm:rights edm:provider edm:dataProvider
edm:ProvidedCHO	dc:language edm:type

Table 11: Missing mandatory elements, MPIWG, VLP database, table vl_essays.

VLP Database, Table vl_people

The table vl_people includes some important person information which we can use in our further contextualisation work. For example, we have data about birth places and death places as well as birth dates and death dates of persons.

We propose to handle this data as follow: By including constructs from other vocabularies like elements of the RDA Group 2 (rdaGr2)[18], we have the ability to retain the semantics of the data and to keep the specific content. For example, we can make use of the *rdaGr2:placeOfBirth* and *rdaGr2:placeOfDeath* properties to form a triple of the type *edm:Agent - rdaGr2:placeOfDeath - edm:Place*.

Unfortunately, people that are described in this table do not have any relations to other tables which also include those person names. For example, "Schmidgen, Henning" is an author of the essay with the URI http://vlp.mpiwg-berlin.mpg.de/essays/data/art10, but he cannot be found in vl_people. At the same time, there are many relations between "Schmidgen, Henning" and persons from vl_people which can be seen by taking a look at a publication from Schmidgen in vl_essays[19].

The basic problem with the table structure of MPIWG is that the information pieces in separate tables are not connected with each other. Furthermore, for instance, person names which are used within an essay are part of one large text entry in the essay table and not specifically marked as person names. This makes it necessary to perform an additional extraction step for such information while converting those tables to RDF and EDM. We still need to find a way to extract this information and to interlink it properly.

Missing elements:

ore:Aggregation	edm:rights edm:provider edm:dataProvider
edm:ProvidedCHO	dc:language edm:type

Table 12: Missing mandatory elements, MPIWG, VLP database, table vl_people.

[18] http://metadataregistry.org/schema/show/id/15.html [31.07.2012]
[19] http://vlp.mpiwg-berlin.mpg.de/essays/data/art10 [23.07.2012]

The mapping of the database is still work in progress as already mentioned in section 5.3.3. That means that there may still be some additional requirements that we have not indicated yet.

6.2 Requirements and Suggestions

Based on the first round of analysis of the provided sample data we are able to give requirements and suggestions for data providers on how they should prepare and map their data to the EDM. Of course, those requirements are preliminary as they will be further revised by subsequent mappings of additional data sets carried out mainly by the data providers themselves.

The following points are requirements:

- Provide information for each mandatory EDM element (see above). In the case of default values this can be done without adding the information explicitly to the source data but by communicating the default value through mapping instructions.
- In case of complex objects: Be sure to provide the mandatory information for each sub-part. Clearly indicate through mapping instructions if values are inherited through the hierarchy or sequence. Otherwise, information for mandatory EDM elements must be made explicit for each object within the complex object.
- Provide a URI or an ID from which an URI can be generated for each object which is meant to be represented in EDM. This is especially important in the case of complex objects. Each sub-part needs to have its own unique URI or ID.
- Provide a URL to the “bare image” depicting the described object. Also confer the EDM+ section.
- Indicate which type applies to the object according to the DM2E object types (e.g. dm2e:Book or dm2e:Journal).

In addition to the requirements, we make the following strong recommendations:

- Provide a URI of a contextual resource (e.g. from GND or VIAF) instead of a string value or an ID alongside a string value whenever possible. In the case of an ID indicate its source. If the ID is pointing to a concept from an internal controlled vocabulary, then, ideally, we would collect the internal controlled vocabularies and convert them to SKOS.
- In case of literals: Especially *Agents*, *Places*, and *Concepts* should be normalised and unambiguous.
- In the case of complex objects with hierarchies and sequences: The hierarchy and sequence between parts of a complex object must be either apparent from the XML structure and thus clearly stated in the mapping instructions (e.g. are the siblings in an EAD finding aid in a meaningful sequence?) or must be explicitly indicated in some element or attribute. The type of the relation between the parts should be made explicit as well.
- In the case of complex objects with hierarchies and sequences: Be sure to either explicitly insert every desired (non-mandatory) information for each sub-part into the source data or to clearly indicate through the mapping instructions which values are inherited through the hierarchy.
- Data provider need to explain codes and shortcuts in their data (e.g. by making the documentation available or through clear descriptions in the mapping instructions).
- Avoid mixing different types of information within one element when only one type is allowed!

- Provide information about the language of string values (e.g. for description fields, place names, concepts etc.), ideally using standard language codes.
- Default values for mandatory elements (e.g. information about the data provider, the *edm:type*, or *edm:rights*) do not need to be in every metadata record but must be clearly communicated through mapping instructions. In the case of complex objects make sure that the default values adhere to all parts (e.g. *edm:rights* information about web resources or the *edm:type*).

6.3 Conclusion and Next Steps

While doing the test mappings, we encountered some problems. In some cases, a metadata record of a provider describes more than one object. For example, a record describes a journal and an article it contains at the same time. In such a case, the EDM advises to distinguish between those two objects and to create two different *edm:ProvidedCHO* resources. However, as both object representations need to meet the minimal set of mandatory elements, it is not always easy to identify the necessary information in the original metadata record.

There are still some questions we had while dealing with the provider data that remain unanswered: How do we deal with implicit information, i.e. inheritance of information within an hierarchical complex object? Can we assume that the language tag of the parent object is also valid for the child? At this point, we need more provider feedback to proceed with the mappings. As the data providers will soon pick up the mappings for their data, this feedback will be available.

The requirements formulated here need to be picked up by the data providers and to be further refined during the definite mappings. Additionally, the requirements as they are outlined here need to be aligned with the requirements towards the source data and the EDM+ as they are formulated by WP1 and WP3.

7 Contextualisation

The contextualisation of our data is done in Task 2.3. As this is the last task that we have started, we have tested only one tool, Silk[20], at this. The first experiments on contextualisation were done using person names. The next steps will be contextualisation by location (e.g. place names, institutions) and, in the case of the MPIWG data, it is intended to generate overarching research relevant keywords and include controlled vocabularies like the Getty-AAT.

7.1 Silk Link Discovery Framework

The first tool we have used was the Silk Link Discovery Framework. The tool is under development at the Freie Universität Berlin. Silk is a tool for discovering relationships between data items within different Linked Data sources. Data publishers can use Silk to set RDF links from their data sources to other data sources on the Web (Volz et al., 2009).

The aim is to find similarities between metadata in RDF which is stored in a triple store with a SPARQL endpoint. Using Silk, new relations between these objects can be discovered and stored again in the triple store. A flexible configuration language, the “Silk Link Specification Language”, allows the configuration of which objects should be compared and which properties are to be used (Isele & Jentzsch, 2012).

The objects can be provided by different data sources. The test case is to identify authors and editors of the digital documents provided by the MPIWG and link them to the PND provided by the Deutsche Nationalbibliothek (DNB). Currently, the metadata provided by the MPIWG only contains information about the authors as text formatted as “LASTNAME, FIRSTNAME”. In order to contextualise the authors, the existing metadata of the MPIWG (index.meta) was transferred in to a preliminary RDF version (an OWL version is currently under preparation, as well as a mapping to EDM+). For the test case the following relations are relevant:

- MPIWG_MD:has_bibl_metaData
  relates a MPIWG digital object to the bibliographical data MPIWG:BibData
- MPIWG_MD:author
  relates the bibliographical data to an author (in the form LASTNAME, FIRSTNAME)

The MPIWG keeps a copy of the RDF version of the GND in its own triple store for internal use. Each person of the GND is related to an internal person ID used at the MPIWG (ni:Person).

The data contains names of authors and persons described as

- FOAF:firstName first name
- FOAF:lastName last name
- FOAF:name fullname of the person

The sample set contains approximately 25,000 datasets. The MPIWG person database contains about 2,000,000 datasets. As mentioned above, most of them are imported from the GND dataset.

[20] http://www4.wiwiss.fu-berlin.de/bizer/silk [23.07.2012]. Authors: Robert Isele, Anja Jentzsch, Christian Bizer, Julius Volz.

Silk was used to identify MPIWG_MD:author and names described in the person repository. Silk allows different settings for the measure used to identify similar strings and the combination of different properties of input and source. In the current initial phase of working with Silk, the emphasis was more on the question if the tool can deal with the large amount of data to be analysed on local available hardware than on optimising the mapping strategies and on the possible workflows for using the tool.

Therefore, no systematic tests with the different settings have been done up to now. The best results so far could be obtained by using a Jaccard measure and threshold of 0.4. Here, roughly a third of all authors where correctly identified, another third were marked as to “be checked”, and a third was not recognised. What was checked was at first the exact identity of name and first name, then the name and the initial of the first name, and lastly only the last name.

The relative low amount of exact results is partly due to ambiguous names, but mostly because of incomplete names in the existing metadata like “Einstein, A.” instead of “Einstein, Albert” and because of entries which contain more than one author. A change in the threshold parameter increases the result of possible candidates, but manual checking becomes time consuming. It would be helpful to have a workflow tool similar to Amalgame, where it is possible to refine the search results step by step.

The latter problem will be resolved by a new version of the conversion from indexMeta-XML to indexMeta.rdf. The former is a more general problem, where an improved user interface for Silk would be helpful.

7.2 Conclusion and Next Steps

Silk allows the comparison of large collections. Under the given constraints of real data, an interactive improvement of the results would be helpful. This could be a first run with exact matching and a second one using heuristics on parts of the not matched entries. These steps are easily possible with Amalgame using a workflow editor.

In the future, more systematic research on Silk is necessary, which parameters yield the highest success rate under which conditions. Use cases have to be defined and example settings for these should be developed and made available to the content providers.

The Silk Workbench provides a graphical user interface for the configuration of Silk. For the iterative improvement of the Silk Link Specifications a graphical editor would be useful. As the next step, the Silk Workbench should be evaluated concerning the suitability for the DM2E use case.

Furthermore, a tool for the correction and validation of the results has to be developed. Additionally, candidates for using Silk should be chosen and provided by the content providers. Another tool, that will be tested next to Silk for the intermediate interoperability version, is Amalgame (cf. section 5.4).

8 Conclusion and Next Steps

In the last six months, WP2 created the initial version of the DM2E interoperability infrastructure. The initial version of the infrastructure combines the D2R Platform for RDFisation of relational data, the Mint Tool for the RDFisation of XML data with the Silk Link Discovery Framework for the contextualisation of the RDFised data.

In order to verify that the chosen tools comply with the requirements that arise in the context of the DM2E project, we have analysed the input data that was provided by the WP1 partners and have developed a specialisation of the EDM data model, called EDM+, which includes additional concepts that are necessary to represent the WP1 data as well as for using the RDFised data in the context of the annotation platform that is developed in WP3.

Afterwards, the MINT and D2R Platform were tested by mapping input data into the developed specialisation of the EDM. This test confirmed that the tools are capable of handling the required transformations, but also revealed some extensions to the tools that need to be implemented in order to comply with all aspects of the requirements. These extensions will be implemented by the WP2 participants over the next months.

Regarding the data mapping and modelling with EDM+, our next step is to unify the requirements of WP2, which are derived from mapping sample data to the EDM, with the requirements which come from WP1 and WP3, i.e. from the data providers and the Pundit prototype platform and its envisioned functionality. All these requirements need to be integrated into the EDM+ and implemented by the data providers in terms of data provision and their mappings and by WP3 in terms of Pundit development.

In summary, we can say that we have achieved our goal to create an initial version of the DM2E interoperability structure.

As the next step and as well as according to the DM2E work plan, we will develop the DM2E interoperability infrastructure further by extending the chosen tools with missing functionality and by implementing an easy-to-use graphical interface which will facilitate the DM2E data providers to use the infrastructure on their own.

9 References

(2011). *Europeana Data Model Primer, v26/10/2011*. Retrieved from http://pro.europeana.eu/documents/866205/13001/EDM_Primer_111026.pdf

(2012). *Definition of the Europeana Data Model elements, Version 5.2.3, 24/02/2012*. Retrieved from http://pro.europeana.eu/documents/900548/bb6b51df-ad11-4a78-8d8a-44cc41810f22

(2012). *DM2E - Digitised Manuscripts to Europeana. Description of Work.*

(2012). *Guidelines for the Rights in Objects submitted to Europeana, 09/02/2012*. Retrieved from http://pro.europeana.eu/documents/900548/1037382/Europeana_rights_201202.pdf

Angjeli, A., Baumgartner, M., Chambers, S., Charles, V., Clayphan, R., Deliot, C., … (2011). *Europeana Libraries, D5.1 Report on the alignment of library metadata with the European Data Model (EDM), Version 1.0.* Retrieved from http://www.europeana-libraries.eu/documents/868553/1eade085-34ac-487f-82af-d5cd2545e619

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

Bizer, C., Heath, T., Berners-Lee, T., & Idehen, K. (Eds.) 2009. *Linked Data on the Web (LDOW2009).*

Cruz, I., Decker, S., Allemang, D., Preist, C., Schwabe, D., Mika, P., … (Eds.) 2006. *The Semantic Web - ISWC 2006: 5th International semantic web conference, ISWC 2006, Athens, GA, USA, November 5-9, 2006 : proceedings*. *Lecture Notes in Computer Science*. Berlin, Heidelberg, New York: Springer. Retrieved from http://www.worldcat.org/oclc/496562772

Drosopoulos, N., Tzouvaras, V., Simou, N., Christaki, A., Stabenau, A., Pardalis, K., … (2012, April). *A Metadata Interoperability Platform*, San Diego, USA.

Griesbaum, J., Mandl, T., & Womser-Hacker, C. (Eds.) 2011. *Information und Wissen: global, sozial und frei?: Proceedings des 12. Internationalen Symposiums für Informationswissenschaft (ISI 2011) ; Hildesheim, 9. - 11. März 2011*. Boizenburg: Hülsbusch. Retrieved from http://www.worldcat.org/oclc/719426134

Hennicke, S., Olensky, M., Boer, V., Isaac, A., & Wielemaker, J. (2011). A data model for cross-domain data representation. The "Europeana Data Model" in the case of archival and museum data. In J. Griesbaum, T. Mandl, & C. Womser-Hacker (Eds.), *Information und Wissen: global, sozial und frei? Proceedings des 12. Internationalen Symposiums für Informationswissenschaft (ISI 2011) ; Hildesheim, 9. - 11. März 2011* (pp. 136–147). Boizenburg: Hülsbusch.

Isele, R. & Jentzsch, A. *Link Specification Language*. Retrieved from http://www.assembla.com/wiki/show/silk/Link_Specification_Language

Kollia, I., Tzouvaras, V., Drosopoulos, N., & Stamou, G. (2012). A Systemic Approach for Effective Semantic Access to Cultural Content. *Semantic Web Journal*, *3*(1), 65–83.

Schreiber, G., Amin, A., van Assem, M., Boer, V. de, Hardman, L., Hildebrand, M., … (2006). MultimediaN E-Culture Demonstrator. In I. Cruz, S. Decker, D. Allemang, C.

Preist, D. Schwabe, P. Mika, … (Eds.): *Lecture Notes in Computer Science, The Semantic Web - ISWC 2006. 5th International semantic web conference, ISWC 2006, Athens, GA, USA, November 5-9, 2006 : proceedings* (pp. 951–958). Berlin, Heidelberg, New York: Springer.

Volz, J., Bizer, C., Gaedke, M., & Kobilarov, G. (2009). Silk – A Link Discovery Framework for the Web of Data. In C. Bizer, T. Heath, T. Berners-Lee, & K. Idehen (Eds.), *Linked Data on the Web (LDOW2009).* Retrieved from http://ceur-ws.org/Vol-538/

Walkowska, J. & Sielski, K. *jMet2Ont User Documentation*. Retrieved from http://fbc.pionier.net.pl/pro/jmet2ont/documentation.html

Wielemaker, J., Boer, V. de, Isaac, A., van Ossenbruggen, J., Hildebrand, M., Schreiber, G., & Hennicke, S. (2011). *EuropeanaConnect, D1.3.1 Semantic workflow tool available, Version 1.0*. Retrieved from http://pro.europeana.eu/documents/12117/1000137/Semantic+workflow+tool+available+at+SourceForge