DELIVERABLE

Project Acronym: DM2E

Grant Agreement number: ICT-PSP-297274

Project Title: Digitised Manuscripts to Europeana

D2.2 – Intermediate Version of the Interoperability Infrastructure

Revision: 1.0

Authors:

Kai Eckert (UMA)
Evelyn Dröge (UBER)
Konstantin Baierer (EL)
Violeta Trkulja (UBER)
Vivien Petras (UBER)

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
0.1	01.07.2013	Kai Eckert	UMA	Initial Draft Version
0.2	12.07.2013	Kai Eckert	UMA	Included overview texts
0.3	12.07.2013	Evelyn Dröge	UBER	Revisions Section 2 and Section 3
0.4	13.07.2013	Kai Eckert	UMA	Provenance and Workflows
0.5	14.07.2013	Konstantin Baierer	EL	Text revisions
0.6	15.07.2013	Kai Eckert	UMA	Development and Conclusions
0.7	15.07.2013	Kai Eckert	UMA	Final revisions
0.8	16.07.2013	Violeta Trkulja	UBER	Some additions and final revision
0.9	19.07.2013	Kai Eckert	UMA	Feedback from WP2 list and incorporated feedback of Antoine Isaac
1.0	19.07.2013	Vivien Petras	UBER	Approval Final 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This deliverable contains original unpublished work except where clearly indicated otherwise.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

List of Tables

List of Figures

List of Abbreviations

API	Application Programming Interface
CHO	Cultural Heritage Object
EDM	Europeana Data Model
EDM+	Former Name of the EDM Specialisation by DM2E
EL	Ex Libris Germany
DM2E	Digitised Manuscripts to Europeana
DM2E Model	Specialisation of the EDM made by DM2E
DM2E v1.0	First operational version of the DM2E model
GUI	Graphical User Interface
JAAS	Java Authentication and Authorization Services
MPIWG	Max-Planck-Institut für Wissenschaftsgeschichte *(Max Planck Institute for the History of Science)*
NTUA	National Technical University of Athens
UBER	Humboldt-Universität zu Berlin *(Humboldt University Berlin)*
UMA	University of Mannheim
WP	Work Package
XSLT	Language for XML-to-XML transformations

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

1 Role and scope of this deliverable

This deliverable presents the current intermediate state of the interoperability infrastructure being developed in WP2. It is a continuation of deliverable D2.1 ("Initial Version of the Interoperability Infrastructure").

Since month 6, a prototype of the infrastructure has been developed and presented within the project. Based on the prototype, the actual interoperability infrastructure has been planned and developed, and will now be released in an intermediate version.

The intermediate version is used to transform and ingest metadata provided by WP1 and to publish the data as Linked Data to be consumed by WP3 (see D3.2 "Prototyping Platform Implemented").

Based on the experiences gained from the data ingestions and in cooperation with the development of the scholarly environment in WP3, the interoperability infrastructure will further be developed to its final version in the next 6 months.

2 WP2 Interoperability Infrastructure

2.1 Overview

Figure 1 shows an overview of the interoperability infrastructure, corresponding to Figure 2 in D2.1. It is updated based on the actual developments.

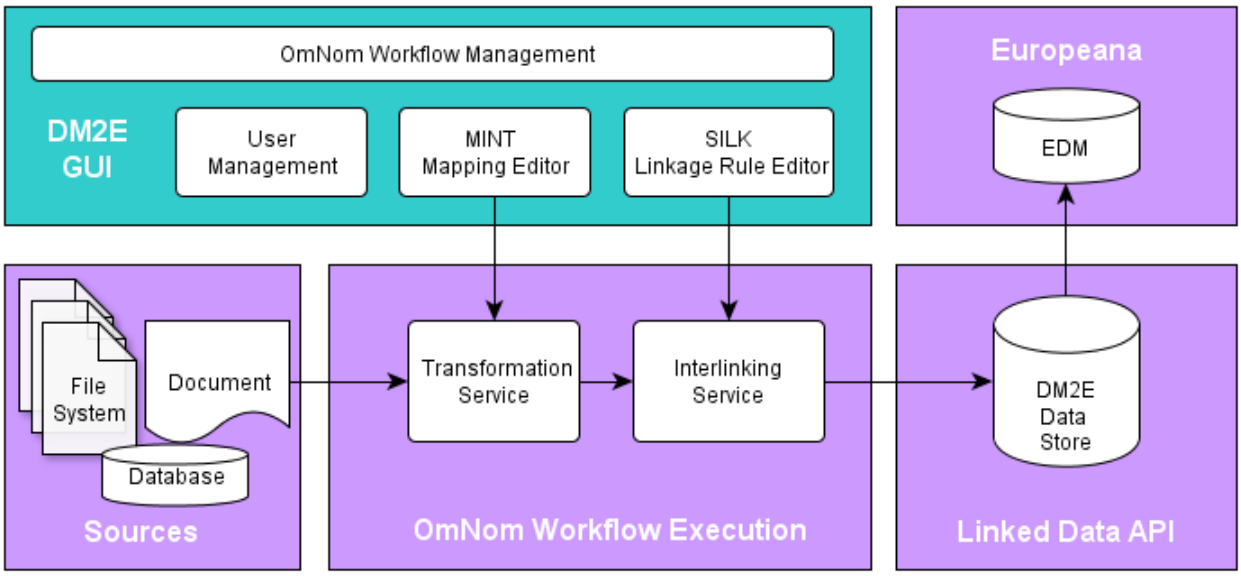

Figure 1: Overview of the interoperability infrastructure.

The central element is the DM2E GUI, indicated in the earlier depiction as DM2E Workbench. The DM2E GUI subsumes and integrates four different GUIs: OmNom, MINT, Silk, and the User Management. The user accesses the system via the OmNom Workflow Management System, logs into the system and can then create, edit, and use various workflows to ingest data to the DM2E data store. A typical workflow consists of a transformation based on a mapping created in MINT and the generation of links to external sources based on linkage rules created in the SILK Workbench. All GUIs are interlinked and provide single-sign-on, i.e., once a user logged in to one GUI, she or he is automatically logged in to other GUIs as well; the user experiences an integrated system. The single-sign-on is realized via a central authentication server providing an additional GUI for the user management (registration of new users, editing of accounts, retrieval of forgotten passwords etc.).

The workflows are executed by the OmNom backend. The modules are actually web services, hence the indication as services. The RDFisation takes place in the transformation, as the MINT mappings are used to create RDF/XML directly from XML input, which is currently the only input provided by WP1. If non-XML data should be ingested, additional services can be provided to transform the data to XML or RDF first. RDF data, as delivered for example by the D2R server, can be mapped by MINT as well.

An important difference to the original workflow is the provision of the data via the DM2E data store. The data is not directly ingested to Europeana for a simple reason: in DM2E, cultural heritage objects (CHOs) can be further granularized. Every single page of a manuscript, a chapter, an article in a journal, every transcription of a page or parts of a page like paragraphs or even individual sentences can be represented as single CHOs. The level of granularity and the identification of the single CHOs are determined by the providers during the mapping process. This granularity usually varies from provider to provider and

from dataset to dataset but is generally much more fine-grained than the granularity of the CHOs that are to be ingested into Europeana, which is usually the level of whole manuscripts. Notable exceptions from this rule occur, when individual pages or images are deemed noteworthy enough to be presented in Europeana as a CHO of its own, with an individual page in the Europeana portal. This means that access to the DM2E data is provided via a dedicated Linked Data API backed by the DM2E data store. A subset of this data will subsequently be transformed from the DM2E model (formerly known as EDM+) to the Europeana Data Model (EDM) and then ingested into Europeana. This transformation is straight forward as the DM2E model is designed as a full specialisation of EDM.

2.2 Workflow Creation and Execution with OmNom

Figure 2 shows the data flow through the interoperability infrastructure and gives a more detailed and technical view compared to Figure 1. The provided metadata is transformed and ingested into a central triple store, where each ingestion results in a different dataset, internally organised in different named graphs (a named graph here refers to a subset of triples in a triple store that is identified by a URI). When the ingested data is contextualised, i.e., when links to external sources are generated, these links are stored in separate linksets. As a consequence, different approaches for the link generation can be applied in parallel and the results can be used independently from the ingested data.

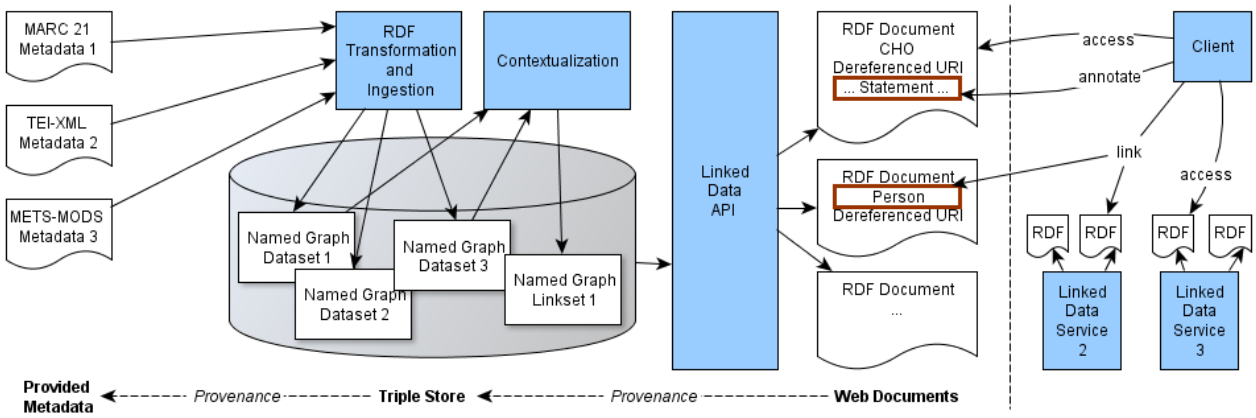

Figure 2: Data Flow through the interoperability infrastructure.

The access to the data is possible via the Linked Data API, a RESTful API following the Linked Data principles. There are two features distinguishing the DM2E infrastructure:

1. The DM2E data model incorporates a metalevel that connects the data to data resources and these resources to datasets. The datasets are versioned and stable URIs can therefore be provided for all data that is provided by DM2E. This model is described in more detail in Section 3.

2. The OmNom workflow system is an orchestration of different webservices that can be deployed everywhere on the web and easily be extended with new webservice implementations. The webservices are described using RDF and all intermediate transformation results are identified by URIs and described by RDF metadata. Together, this enables the representation of the full provenance chain of the data back to the original data. This system is further explained in Section 4.

2.3 Mapping Creation with MINT

MINT is used in DM2E as a mapping editor (Figure 3). The user can access the mappings that have been created in MINT from OmNom and use them in workflows to transform the data to the DM2E data model. With one click in OmNom, MINT is opened to edit an existing mapping.

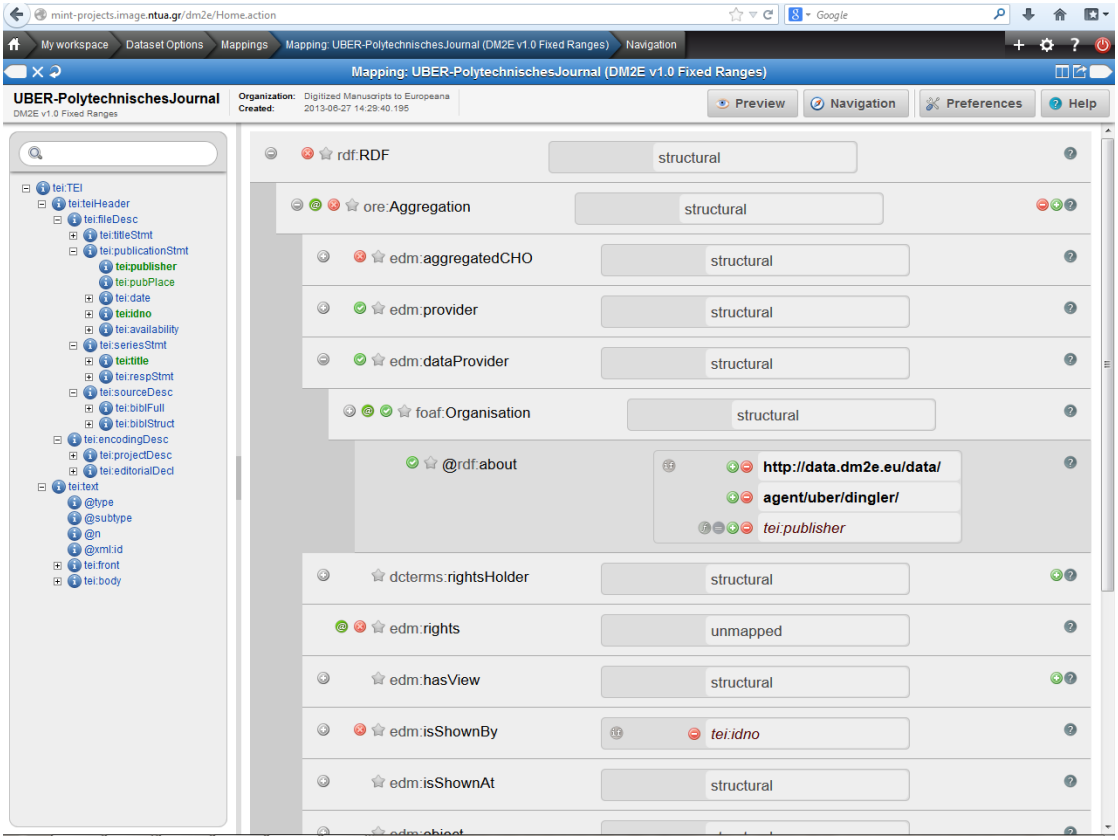

Figure 3: The DM2E MINT instance.

MINT was the first part of the DM2E infrastructure that has been released in production mode. In May, the DM2E MINT installation was updated to the operational version of the DM2E data model and providers can use it to create mappings for their data. A process has been established at NTUA that ensures that existing mappings are now transferred and updated if further changes to the DM2E data model should take place.

The released DM2E MINT installation has been showcased in London during the All-WP meeting, with dedicated workshops on the second day to provide hands-on examples for the data providers. The providers have since been creating the mappings in MINT, as well as with custom XSLT scripts (see Section 2.5), so that the data ingestion can start immediately, when OmNom is released.

2.4 Link Generation with Silk

Silk, like OmNom, consists of two parts:

1. the backend that generates links based on input data and a linkage rule file, and
2. the frontend (Silk Workbench, Figure 4).

The Silk backend is already available as an OmNom-compatible webservice and can be included in ingestion workflows. The workbench has been installed and will be integrated, like the MINT GUI, as special editor for linkage rule files.

A current limitation of Silk is that it only works on structured data with data elements describing identifiable resources. Input data for DM2E often contains textual descriptions where entities are described. These entities first have to be extracted from the text in order to create links to resources that correspond with these entities.

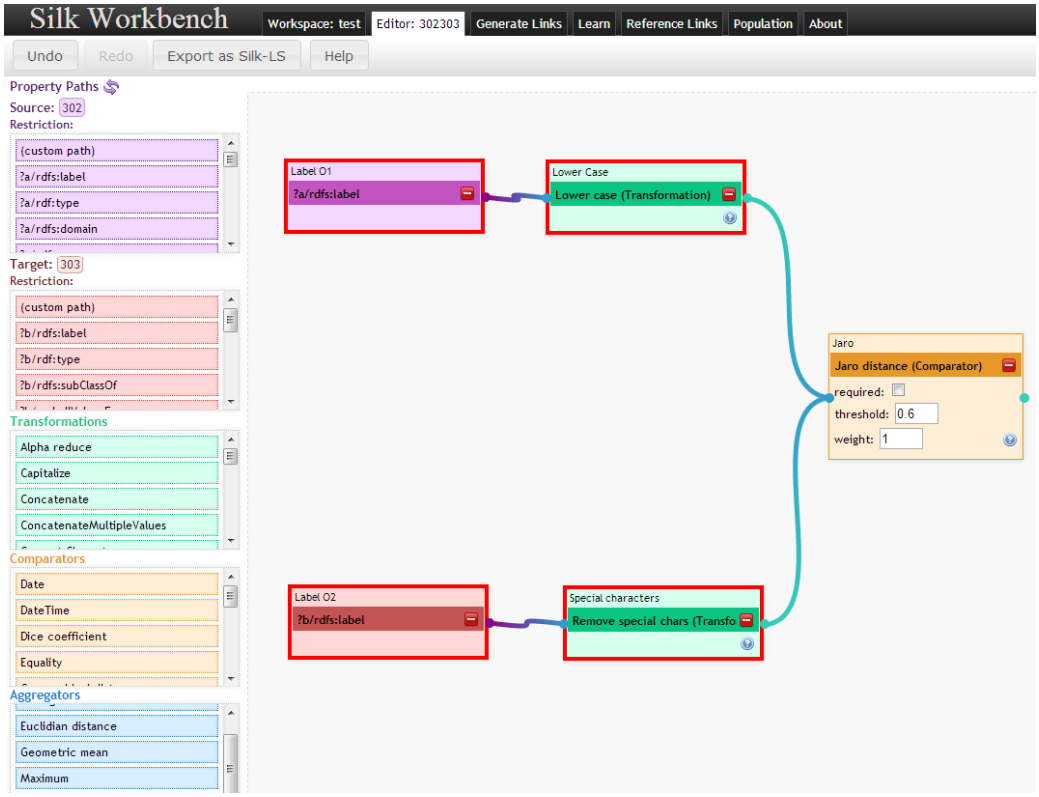

Figure 4: Silk Workbench.

Silk is currently extended at UMA to support entity extraction from texts as a pre-processing step. Another approach would be the use of text mining tools like GATE or RapidMiner in the contextualisation workflows. A webservice that executes RapidMiner processes has been implemented for testing purposes.

2.5 Support for custom XSLT

As XSLT is the predominant transformation technology in the cultural heritage domain and many providers are experienced with the creation of XSLT scripts, a generic XSLT transformation service has been implemented that can not only work with MINT mappings,

but also with simple XSLT scripts and complex multi-file XSLT scripts provided as ZIP containers.

An interesting way for the creation of mappings demanded by the providers is also the use of MINT for most parts of the mappings, followed by manual adjustments to the resulting XSLT for more complex cases where MINT provides not enough flexibility.

2.6 Linked Data API for WP3

The data in DM2E is accessed via a RESTful Linked Data API, following the Linked Data principles. Table 1 lists the most important URIs. All URIs, in particular the URIs of RDF data and datasets, represent resources described in the DM2E data model (Section 3).

URI (http://data.dm2e.eu/data/...**)**	**Type**	**Content**
dataset/[provider]/[datasetID]/**[version]**	void:Dataset	Description of the ingested dataset
resource/[provider]/[datasetId]/[identifier]	edm:ProvidedCHO	303 to latest version of the resource map
aggregation/[provider]/[datasetID]/[identifier]	ore:Aggregation	303 to latest version of the resource map
resourcemap/[provider]/[datasetID]/ [identifier]/**[version]**	ore:ResourceMap	RDF data, linked to a void:Dataset of the same version
linkset/[provider]/[linksetID]/**[version]**	void:Linkset	Description of a linkset
linkset/[provider]/[linksetID]/**[version]**/ [provider]/[datasetID]/**resource**/[identifier]	dm2e:DataResource[1]	RDF data, links for a specific resource

Table 1: Linked Data API.

The provided RDF data remains stable as it is always related to a specific, versioned dataset or linkset. This is important to allow statements about the content of the metadata as well. The described resources are never versioned and URIs representing them are 303 redirected to the latest version of their description. The description contains links to earlier versions, if they exist.

The content itself (scanned pages of the manuscripts, the textual transcriptions, images, etc.) is hosted by the providers and linked from the metadata. For the hosting of the content, a technical specification has been developed in DM2E (WP1, WP2, WP3; “DM2E Annotatable Content”) to ensure that the content remains stable and is prepared for the consumption by applications like the scholarly environment developed in WP3.

2.7 User Management

As different GUIs and systems are integrated in DM2E, a central user management and authentication is necessary. For the intermediate version, the focus is on the user authentication in the DM2E GUIs. Authentication and authorisation is performed for three reasons:

[1] A dm2e:DataResource is a non-abstract information resource that provides RDF data. Therefore, it is a specialization of a foaf:Document. In DM2E, every dm2e:DataResource is connected to a void:Dataset by means of void:inDataset.

1. Security: to prevent outsiders from changing and executing workflows and data ingestions.
2. Provenance tracking: to have a user name associated with all created resources and executed workflows.
3. Personalisation: to improve the usability, e.g., by highlighting to the user those MINT mappings or workflows that have been created by this user.

The authentication server provides GUI elements like a login screen that can be incorporated in the DM2E GUIs. Currently, the user simply is redirected to the login if he or she is not authenticated (Figure 5). The GUIs all implement single-sign-on using the Java Authentication and Authorization Services (JAAS) with the JOSSO[2] framework. Other APIs can be added as needed, for instance to support non-Java applications.

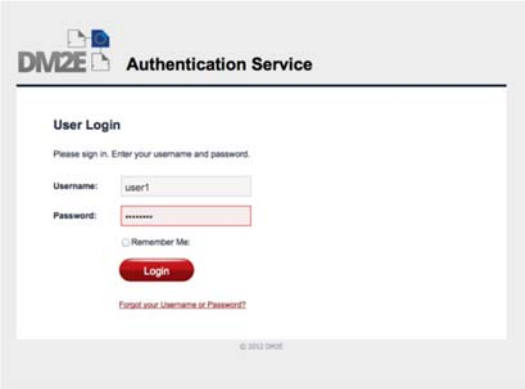

Figure 5: DM2E authentication service.

It is planned to extend the authentication to the interaction of webservices as well, but this is of lower priority. For now, critical services like the actual ingestion service and the workflow execution are only secured on system level to prevent an unauthorised execution from the outside which would lead to unsupervised ingestions.

[2] JOSSO is an Open Source Internet SSO solution for rapid and standards-based (SAML) Internet-scale Single Sign-On implementations, allowing secure Internet access to the Web-based applications or services of customers, suppliers, and business partners http://www.josso.org.

3 The DM2E Model

3.1 Published Operational Version 1.0

The DM2E model (formerly known as EDM+) is the central building block of the infrastructure that determines how the data is represented for consumption by WP3. It has been published on April 15th 2013 in its currently used first operational version 1.0.[3] The DM2E data model fulfils three goals:

1. It provides enough details and flexibility to support the use cases as determined by WP1 and WP3.
2. It is designed as a specialisation of the Europeana Data Model and therefore fully compatible. All data can be ingested with little to no transformation effort into Europeana.[4]
3. It provides a metalevel to represent further information about the ingested data, following the Linked Data principles. In particular, it does not rely on the concept of proxies, as introduced in OAI-ORE and as used in Europeana.

The last point is crucial: we here have a direct connection between the data and the infrastructure, as it is to be expected in a Linked Data context.

3.2 Provenance

Table 2 shows the four levels of abstraction that are available in the DM2E model. The two upper rows in red indicate the levels that are managed by the DM2E infrastructure and build the metalevel that is used to represent the provenance information of the ingested data. The lower two rows in green are two different levels that contain the actually provided data. The class ore:Aggregations can be seen as a metalevel for providers as it is about the provided metadata of the CHO.

The provenance of the RDF data is provided using the Vocabulary of Interlinked Datasets (VoID).[5] Each ingestion and each link generation results in a new and stable void:Dataset that is linked from all data resources (foaf:Document, dm2e:DataResource, ore:ResourceMap) that are published via the Linked Data API. Therefore, the URI of the dataset determines permanently the identity of all statements contained in this dataset.

This allows the stable annotation of these statements. An investigation of the possibilities and limitations of Linked Data for provenance-aware applications has been conducted for the development of the DM2E model and the results have been accepted for publication at the International Conference for Dublin Core and Metadata Applications 2013 (DC-2013), held from September 3rd to September 5th in Lisbon.

[3] See http://onto.dm2e.eu/schemas/dm2e/1.0/ and http://pro.europeana.eu/documents/1044284/0/DM2E+Model+V+1.0+Specification for a full documentation of DM2E v1.0.

[4] The transformation is straightforward, but the CHOs to be ingested have to be identified first, as not all DM2E CHOs are suitable for Europeana. In most cases, however, this can be determined simply on class base, i.e., all dm2e:Manuscripts, dm2e:Books etc. are ingested. Plans for case-by-case propagation of noteworthy CHOs to Europeana using DM2E-specific properties exist but are not yet production-ready. Properties that do exist in DM2E but not in EDM can be mapped to the more generic EDM properties.

[5] http://www.w3.org/TR/void/ (12.07.2013).

Class	Uplink	Metadata
void:Dataset (Named Graph)	---	About the RDF data, DM2E perspective
ore:ResourceMap dm2e:DataResource foaf:Document	void:inDataset	
ore:Aggregation	ore:isDescribedBy	About the provided metadata, providers perspective, record level
edm:ProvidedCHO	ore:isAggregatedBy	About the content

Table 2: Levels of abstraction in the DM2E Model.

The provenance for each dataset is provided using simple Dublin Core, as well as by means of the PROV ontology[6] that is used to connect the datasets with the workflow system, as further described in Section 4.

3.3 Versioning

In DM2E, every void:Dataset refers to a stable, unique version of the data resulting from one specific ingestion. As the stability of the published data is ensured by means of versioning, additional links have to be provided that relate the different versions of a dataset. Table 3 lists the properties that are currently used in DM2E for this purpose. An inclusion of or relation to other vocabularies like the currently developed ADMS vocabulary[7] is considered.

Property	Description
dm2e:previousVersion	Links to the previous version of this dataset.
dm2e:firstVersion	Links to the oldest available version of this dataset.
dm2e:version	Serial number of this version, starting with 1.
dm2e:versionName	Provides a human-readable name for this version.
dm2e:nextVersion	Links to the next version of this dataset.
dm2e:latestVersion	Links to the latest available version of this dataset.[8]
dm2e:availableVersions	Number of available versions of this dataset.

Table 3: Versioning of the DM2E model.

These links are automatically created during the ingestion. The Linked Data API always resolves URIs of described entities like CHOs or Persons to the latest version of a dataset that contains statements about it. Therefore, all applications that do not care about the stability of the content and the prior versions can just ignore the links to the former versions. Details of the versioning process are likely to change until the final version, as the representation of the versions depends on the requirements raised by WP3.

[6] http://www.w3.org/TR/prov-o (12.07.2013).
[7] https://dvcs.w3.org/hg/gld/raw-file/default/adms/index.html
[8] Another possibility is the link to an abstract placeholder URI subsuming the actual versions which could be used to provide links to the latest version.

4 OmNom Workflows and Provenance Chain

The distinguishing feature of the OmNom workflow system is the representation of all components involved as resources on the web with dereferenceable URIs. In particular, these are

1. the user,
2. the workflows,
3. the workflow executions,
4. every webservice,
5. every webservice execution,
6. all configuration parameters,
7. all inputs,
8. all outputs, and
9. all intermediate results.

Figure 6 illustrates this. The resources are interlinked using the PROV ontology, thus forming a full provenance chain. As described in Section 3.2, the starting point for provenance is usually a dataset representing the result of one specific data ingestion (DM2E-RDF). The dataset is linked to a job (Publishing-Wittgenstein-Job) that created the dataset. The job represents the execution of a webservice (Publishing-Service) as part of a specific workflow execution (Wittgenstein-Ingestion-Job) which is linked to the actual workflow (Wittgenstein-Ingestion-Workflow).

The Publishing-Wittgenstein-Job used the resource RDF-XML as input, which was generated by the XSLT-Wittgenstein-Job, which used the original input file (TEI-XML) and a specific MINT-Mapping. The MINT-Mapping was created by The Human in Charge, who also executed the Wittgenstein-Ingestion-Workflow.

The provenance chain is meant to be followed backwards from the result back to the origin. In OmNom this representation is also used to describe and execute workflows from the input file to the final output, the ingested dataset, which is indicated by the top-to-bottom organisation in Figure 6. Therefore, a workflow ontology has been developed that describes the above mentioned entities, as well as their relations.

All entities use URIs and support content negotiation, i.e., RDF descriptions for all entities with links to the generating workflow can be obtained by dereferencing the URIs of those entities. The webservices use the RDF description to describe themselves, i.e., which parameters they require and what they produce. OmNom consumes this information for the proper orchestration of the workflows and improving the usability of the GUI. Additionally, these webservice descriptions become part of the provenance chain, as illustrated in Figure 6.

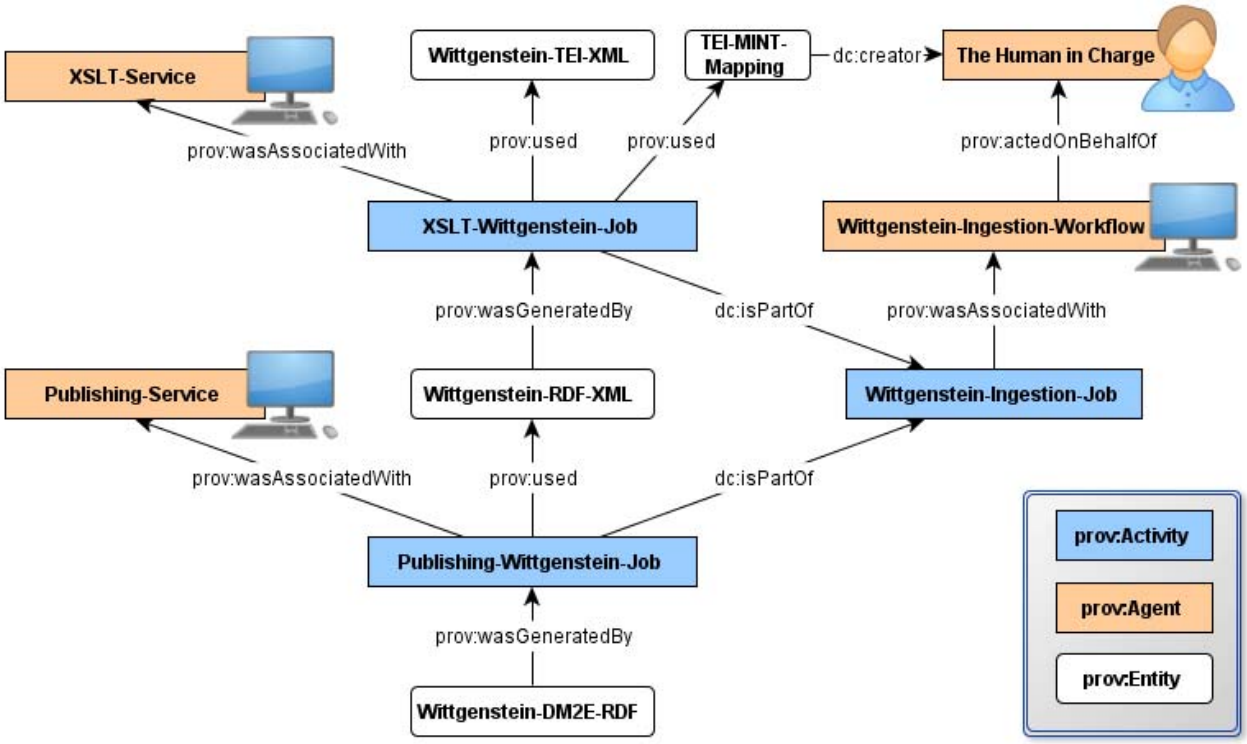

Figure 6: Provenance chain of a dataset in DM2E.

The webservices have a very simple RESTful interface: they are executed by putting a URI of a configuration resource to them. For once, the mechanism of invoking a webservice is intuitive and straightforward to implement. More importantly, however, the actual configuration (i.e., the assignment of values to the input parameters) is made explicit as a resource of its own, which also becomes part of the provenance chain. A webservice always and immediately answers the start request with a job URI representing the actual webservice execution. The description of the job contains information about the current status of the job, links to the configuration and webservice being used, and, when finished, links to the resulting resources. Figure 7 shows the connections between entities that form part of a workflow in OmNom. The overall picture is very complex, but as all entities are represented by dereferenceable URIs, this system allows entering the workflow and provenance chain at any point, providing a rich interlinking to gather more information using the follow-your-nose principle.

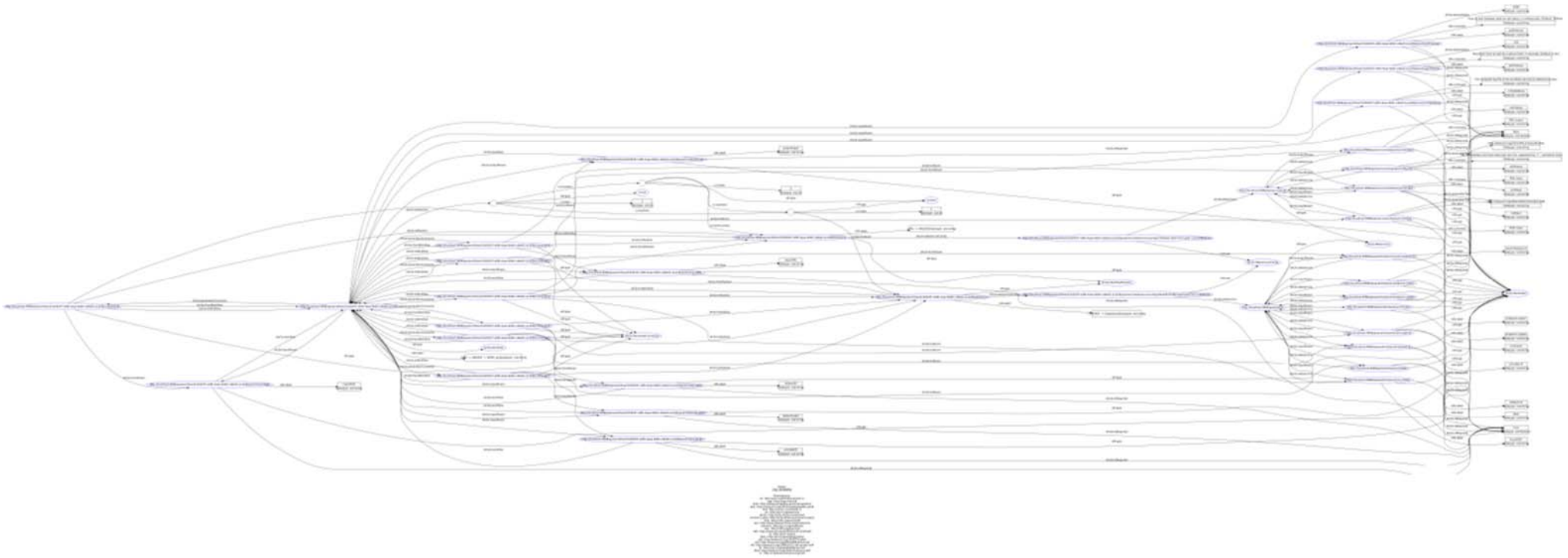

Figure 7: Connections between entities consumed and produced by OmNom.

5 Development and Deployment

5.1 Development

The source code of MINT, the Linked Data API, the OmNom workflow system, all transformation services, and the Silk service is available via GitHub (https://github.com/DM2E). The user management server is currently developed at MPIWG and will be available on GitHub as well.

5.1.1 OmNom

As OmNom is the component in the DM2E interoperability infrastructure that is newly developed, some more information on the development process should be given.

After a quick survey among the developers in the DM2E project regarding their preferences and experiences, it was decided that the system will be developed in Java. The Linked Data publication is based on Pubby,[9] a Linked Data frontend for SPARQL endpoints. The webservices are implemented using JAX-RS 2.0. Unit tests and integration tests are written and performed for all parts of the workflow system. OmNom is built using Maven, which keeps track of all dependencies and allows a very easy setup for new developers.

5.1.2 Grafeo

OmNom does not only create RDF data, it also communicates via RDF data and describes all resources in RDF. Therefore, a dedicated API has been developed called Grafeo. Grafeo is based on Apache Jena[10] and provides high-level functionalities tailored to OmNom, like serialisation and deserialisation of Java objects to RDF and JSON, encapsulation of SPARQL queries in high-level APIs, the publishing of RDF resources, or the consumption of RDF data from external sources. Several additional APIs have been considered for reuse, especially for the serialization and deserialization, but they were either too complex, not flexible enough, or lacked stability.

5.2 Deployment

The intermediate interoperability infrastructure will be deployed on several servers. The vocabulary server and the MINT instance are already deployed and in production. The OmNom Workflow Server will be deployed next, so that the ingestion of the data can start as soon as the mappings from the providers are available.

5.2.1 Servers

Figure 6 shows all servers that are used for the deployment of the intermediate interoperability infrastructure in DM2E.

[9] http://wifo5-03.informatik.uni-mannheim.de/pubby
[10] http://jena.apache.org/

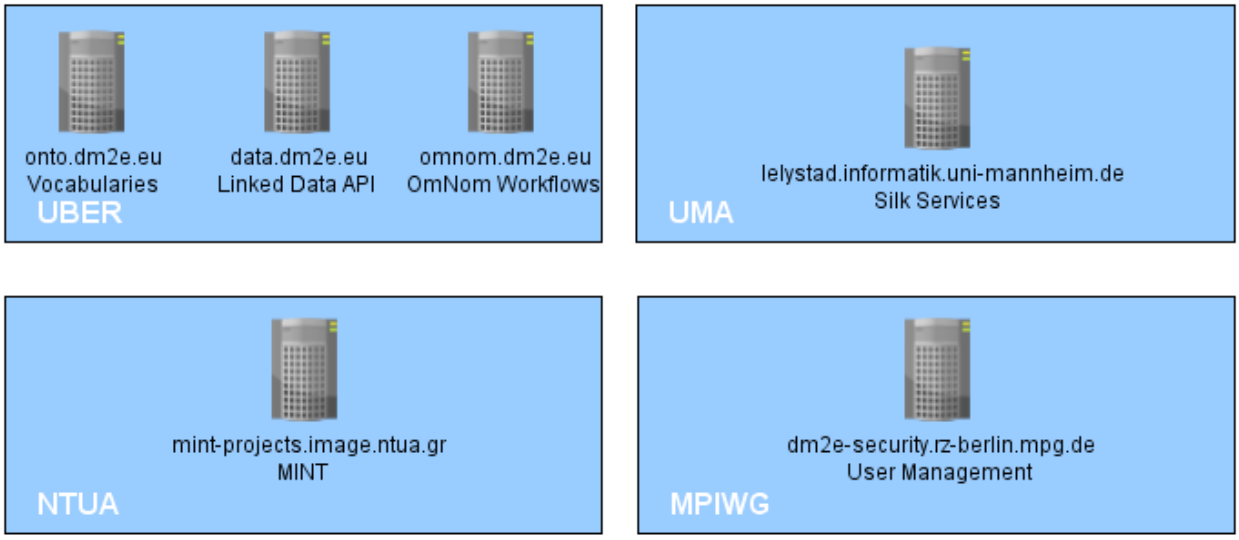

Figure 8: Deployment of the intermediate interoperability infrastructure.

The main part is hosted on servers at UBER:

- The vocabulary server (onto.dm2e.eu) dereferences all URIs of the DM2E data model (Section 3) and URIs of the OmNom workflow system (Section 4).
- The data server (data.dm2e.eu) provides the actual data via the Linked Data API.
- The OmNom server (omnom.dm2e.eu) hosts the OmNom GUI and the OmNom workflow execution system.

At UMA, the Silk service is currently deployed for testing purposes.

MINT is hosted at NTUA.

MPIWG hosts the central user directory.

5.2.2 Namespaces

The deployment on different servers is connected to the namespaces that are publicly used in DM2E:

- http://onto.dm2e.eu/... is used by the DM2E data model, but also for the OmNom workflow ontology and other vocabularies that are created in DM2E.
- http://data.dm2e.eu/... Is used to create URIs for all ingested datasets, CHOs, aggregations and linksets. These URIs are dereferences by the Linked Data API.
- http://omnom.dm2e.eu/... These URIs refer to workflow components like webservices, but also files, parameters, configurations etc.

5.2.3 Packaging and Documentation

For now, there is no out-of-the-box packaging for the installation of the whole DM2E system. The installation and usage of MINT is documented in the MINT Wiki.[11] OmNom is easy to

[11] http://mint.image.ece.ntua.gr/redmine/projects/mint/wiki

deploy via Maven, the source code is documented using Javadoc. A properly packaged version will follow, together with a full documentation. Silk provides an own documentation.[12] The Silk contextualisation service in OmNom is provided as Maven project as well, it installs all necessary dependencies automatically.

For the final version, it is planned to provide all components as independent packages for an easy installation, as well as a virtual machine image including all components as a ready-to-use solution for own tests.

[12] http://wifo5-03.informatik.uni-mannheim.de/bizer/silk

6 Conclusion and Next Steps

With the intermediate interoperability infrastructure, an important step in the DM2E project has been reached. The whole infrastructure has been designed and developed based on extensive surveys and user studies, as well as the experiences gained from the OmNom prototype.

The operational version of the DM2E model is published and the mapping of the data to this model using MINT already started. A first evaluation of the MINT installation currently takes place in WP1.

The OmNom workflow system for the ingestion is now fully functional and will be deployed during the next days in order to start the actual data ingestion.

The development team of WP2 is in contact with the developers of the Unified Ingestion Manager (UIM) and other Europeana projects, namely Europeana Inside, Creative, and Cloud. The goal is to maximise synergy effects and minimise unnecessary parallel developments. The open architecture of OmNom, the support for a full provenance chain and the integration with Linked Data are seen as the most innovative aspects of OmNom that potentially can be integrated in the other projects.

For the final version, the focus is on the improvement of the OmNom GUI and the integration and use of Silk and other tools for data contextualisation. Another step will be the migration of the triple store to OWLIM, provided by OntoText. Further adaptions to the Linked Data API based on experiences from WP3 are also expected. Apart from these developments, the documentation and improvement of the code will have a high priority, in order to provide a stable, easy-to-install, and easy-to-use final version of the DM2E interoperability infrastructure.