Metadata Enrichment
Semantic resources Survey

Version:	1.4
Date:	14 December 2019
Prepared by:	Maral Dadvar (dadvar@ub.uni-frankfurt.de)
Contributors:	Kai Eckert (HDM), Rachel Heuberger (GUF), Pavel Kats (JHN)

Link to this document:	https://tinyurl.com/Europeana2-GUF-T2-1
References:	Project grant agreement - Activity 2 - Metadata Enrichment

2.1 Semantic resources survey
In this task GUF will create a detailed survey of controlled vocabularies, thesauri, ontologies and linked open data sources for Jewish cultural heritage materials. These will include generic resources applicable also to Jewish heritage materials, such as GeoNames or Library of Congress Headings, and domain-specific resources, such as the Encyclopedia of Russian Jewry. The survey will also contain recommendations as to which resources can be used to enhance metadata materials ingested during the course of the project and listed in the content survey prepared in the task 1.1.

This document is the deliverable of the task 2.1 Semantic Resources Survey. It presents a detailed survey of LOD resources which will be used for semantic enrichment of the datasets. In addition to the content of this document, there was a call to other partners and potential collaborators to provide the project with other potential resources which might be relevant to the Metadata Enrichment.

1. Introduction

Judaica Europeana 2.0 project is a revival of the efforts to aggregate Jewish heritage objects from some 30 institutions in Europe, the US and Israel into Europeana. This aggregator will contain several components (see Table 1) and the enrichment of the data sources will be one of them.

Table 1. Aggregator Components

ID	Name	Description
C1	Data storage layer	The core data layer of the system, storing incoming partners' data and enriched data
C2	Data storage API	External and internal APIs for using the data storage layer, allowing read/write access to all entities, batch access to datasets, and querying by pre-defined fields
C3	Harvesters	A series of custom harvesters, fetching data from the partners' systems and storing it in the storage layer
C4	Ingestion dashboard (incl analytics)	User dashboard for content partners to monitor their data on JHN
C5	Enrichment workflow	Generic API-based mechanism implementing a placeholder (stub) for scenarios to enrich the harvested data with data from external semantic resources
C6	Enrichment scenarios	Array of implementations to enrich specific metadata fields with information from external semantic resources
C7	Europeana connector	Mapping of the data to and from Europeana's EDM format
C8	Media storage	Online services for storing and managing media assets referenced by data
C9	User portal	Portal displaying ingested data to online users

Semantic enrichment is a process of creating additional objects and relations out of harvested data (for example, extracting named entities such as persons and places). The aggregator will provide mechanism for registering job trigger. Registered job will be triggered once dataset harvesting and conversion is completed. Dataset enrichment job will be able to retrieve the whole dataset from the database (via Data Storage API) and will invoke dataset enrichment process. The resulting enrichment objects and relations will be inserted into the database via Data Storage API.

This document describes the proposal activity 2, which aims to provide further information about the sources used for the enrichment process; i.e. providing richer context by linking the

extracted named entities, such as persons and geographical places to other sources containing relevant information. In the enrichment process, we start with enrichment of entities which already have an identifier (such as VIAF or GND ids), the next stage would be the entities which are explicitly described, for example prefLabel of an author, and at last we will identify and extract new entities within the metadata, for example entities mentioned in text in the 'description' attribute (Figure 1.), and then we will proceed with the rest of the enrichment process.

Figure 1. The words in red are the entities which potentially can be extracted from the text and be further enriched.

dc:description *Ilse Wunsch was born in 1911 in Berlin, Germany. She started studying piano at a very early age. After graduating from high school, Ilse Wunsch continued her studies at the Teachers Training College. In 1936 she left for Japan, for concertizing and teaching. She eventually moved to Chicago where she continued her studies at the Chicago Musical College, in due course, earning a Master of Music degree. After meeting Otto Mainzer, ...*

The task of enrichment consists of two steps: first, to identify entities unambiguously by means of stable URIs, second to find as much information (e.g., descriptions, links to related entities) as possible about the identified entities, usually by following links (such as owl:sameAs) to other data sources and this way by obtaining further URIs suitable for identification. Therefore, we must identify external resources and knowledge-bases which might contain the information that we are interested in.

Many useful data sources exist that can be used for data enrichment in the domain of Jewish heritage and studies, e.g., domain specific online encyclopedias such as the YIVO Encyclopedia of Jews in Eastern Europe. In contrast to sources such as Wikipedia, they describe all entities in depth from the domain perspective, i.e., with respect to Jewish history, which makes them more useful for our task . There are also general purpose knowledge-bases with structured data access via Linked Open Data representations, such as DBPedia or Yago, which also contain relevant information about Jewish culture and history next to other topics. Additionally, there are highly relevant data sources such as the Integrated Authority File (GND) of the German National Library providing mainly unique identifier to the described entities, but also brief additional contextual information, usually of a very high quality.
In brief, for finding and employing relevant resources for contextualization we will be facing one of the following situations; either unstructured data sources like online encyclopedias which need to be made available as structured data with stable URIs, or relevant subsets of general purpose knowledge bases such as DBpedia have to be identified and extracted to fill the gaps between the specialized resources and to provide further context.

2. Datasets

2.1 Datasets to be used for contextualization

Based on the entities which have to be enriched, we need a variety of data sources for providing further information and linking. To this end, as described by Task 2.1, we have identified several data sources, including a variety of encyclopedias, book sections, as well as sub-graphs extracted from general-purpose knowledge bases, which can be used for enrichment purposes.

In Table 1, we have presented a list of these resources which are specific to the domain of Jewish cultural heritage. This table also provides a brief description of these resources and their contents.

These sources reflect our work on JudaicaLink, a domain-specific knowledge base for Jewish culture. These sources are available as ready to use datasets on JudaicaLink, and depending on the original source, relevant information, such as birth/death date and location, occupation and identifiers are extracted from the sources, and structured as Linked Open Data[1]. Therefore, in this project we will use JudaicaLink as the hub where we collect all the datasets which will be utilized for data enrichment.

Table 1. Resources for enrichment

Source	Description
The Library of Haskala	Contains 525 books in German and Hebrew identified by leading scholars as Haskala literature. The database is hosted and run by the Judaica Division of the Frankfurt University Library. The author's information have been extracted and the created dataset is further enriched with the corresponding GND information of each author.
Persons of National Library of Israel	This dataset contains the authors extracted from the authority file of National Library of Israel. When available, there is extra information, such as the date of birth and death and list of publications of each author also added.
Persons of University Library of Frankfurt Judaica collection	This dataset contains authors extracted from Judaica collection of University Library of Frankfurt. Moreover, when available, the authors are enriched with extra information from other resources such as common authority file (Gemeinsame Normdatei, GND) of the German National Library.

[1] http://www.judaicalink.org/datasets

Stolpersteine in Mainz	The stumbling blocks are a project of the artist Gunter Demnig, which began in 1992. In May 2018, there were around 69,000 bricks in Germany and 23 other European countries. The stumbling blocks are the largest decentralized memorial in the world. In Mainz (including Mainz Kastel) since 2007, more than 200 stumbling blocks have been laid. This dataset is based on the Wikipedia page list of stumbling blocks in Mainz and contains the address of the Stolpersteins, the date of installation, the founder, comments and coordinates.
Persons from DBPedia	Contains persons extracted from DBpedia. This dataset is a subgraph of DBPedia which includes persons, such as Rabbis who are related to jewish culture and studies. When available these persons were linked to the common authority file of the German National Library.
Eine Jüdische Familie aus Aschaffenburg	The information regarding Hirsch family is gathered by Jüdische Leben in Unterfranken - Biographische Datenbank e.V. (Jewish Life in Lower Franconia - Biographical Database e.V.). The Association of Jewish Life in Lower Franconia - Biographical Database e.V. systematically researches the city and monastery archives in Aschaffenburg for sources on Jewish history. Origin: Biographical database - Jewish Lower Franconia
Persons from GND	Contains persons extracted from the Common Authority File (Gemeinsame Normdatei, GND) of the German National Library.This dataset is a subgraph of GND which includes persons, such as Rabbis who are related to jewish culture and studies. When available these persons were linked to other resources.
Biographisches Handbuch der Rabbiner	The Biographisches Handbuch der Rabbiner is an online encyclopedia provided by the Salomon L. Steinheim-Institute for German-Jewish history at the University of Duisburg-Essen, edited by Michael Brocke and Julius Carlebach. The goal of this encyclopedia is to be a complete directory of all rabbis who lived and worked in or originated from German-speaking areas since the age of enlightenment. The encyclopedia consists of two parts: Part 1: Die Rabbiner der Emanzipationszeit in den deutschen, böhmischen und großpolnischen Ländern (1781 – 1871), edited by Carsten Wilke. Part 2: Die Rabbiner im Deutschen Reich (1871 – 1945), edited by Katrin Nele Jansen.
Geographical Coordinates	This dataset contains all the geo-coordinates of all the cities/countries stated in the JudaicaLink datasets as birth

	location or death location. All these coordinates as well as their GND identifiers are extracted from GND entries of the corresponding cities/countries when available.
Encyclopedia of Russian Jewry	Rujen.ru provides an Internet version of the Encyclopedia of Russian Jewry, which is published in Moscow since 1994, giving a comprehensive, objective picture of the life and activity of the Jews of Russia, the Soviet Union and the CIS. The encyclopedia is structurally divided into three parts: 1. biographical information, 2. local history of the Jewish community in pre-revolutionary Russia, the Soviet Union and the CIS, and 3. thematic information on concepts related to Jewish civilization, the contribution of the Jews of Russia in various fields of activity, various Jewish social, scientific, cultural organizations, etc. The originally published volumes contain more than 10,000 biographies and more than 10,000 place names. The electronic version contains corrections and additions in the form of new articles.
Das Jüdische Hamburg	Das Jüdische Hamburg contains articles in German by notable scholars about persons, locations and events of the history of Jewish communities in Hamburg. Das Jüdische Hamburg is a free online resource based on the book "Das Jüdische Hamburg - Ein historisches Nachschlagewerk", ISBN: 978-3-8353-0004-0, Wallstein, Göttingern (2006). For this datasets we have implemented a Person recognition and have been able to identify 196 person descriptions among the articles. When available, the persons are described by occupation, date of birth and death as well as place of birth and death.
Yivo Encyclopedia	The YIVO Encyclopedia of Jews in Eastern Europe, courtesy of the YIVO Institute of Jewish Research, NY. The only resource of its kind, this encyclopedia provides the most complete picture of the history and culture of Jews in Eastern Europe from the beginnings of their settlement in the region to the present.

There are other sources (see Table 2 for examples) which can be of great value for our enrichment purposes. Some of these datasets are not accessible due to copyright issues and are still in negotiation purposes, some we are waiting to receive the data from the provider and some are still under preparation.

Table 2. Potential resources for enrichment

Source	Description
Encyclopedia Judaica	Includes more than 21,000 entries on Jewish life, culture, history, and religion in 22 volumes.
Jewish Encyclopedia	This website contains the complete contents of the 12-volume Jewish Encyclopedia, which was originally published between 1901-1906. The Jewish Encyclopedia, contains over 15,000 articles and illustrations. This online version contains the unedited contents of the original encyclopedia. Since the original work was completed almost 100 years ago, it does not cover a significant portion of modern Jewish History (e.g., the creation of Israel, the Holocaust, etc.). However, it does contain an incredible amount of information that is remarkably relevant today.
KIMA - Historical Hebrew Gazetteer	Each entry in this database consists of preferred forms of a toponym (both in Hebrew-script and in its English normalized form), variant Hebrew-script names and their transcriptions, together with their extant historical attestations, a calculated historical span of use, and geographical coordinates where available. The Kima entities were matched to existing open knowledge resources of contemporary and historical places: Geonames and Wikidata entities. Kima currently holds 27,239 Places, with 94,650 alternate variants of their names and 236,744 attestations of these variants.

Throughout the project based on the requirements and the dataset entities which need to be enriched, JudaicaLink knowledge base might be further enhanced with additional resources. We also will conduct a survey among project partners to identify further relevant data sources, which do not necessarily have to be publicly available, but could also be in-house created data.

2.2 Datasets to be contextualized - Preliminary Analysis and Exploration

The first step for metadata enrichment is to exactly know the data, the available attributes, records to be enriched and attributes which can be used to extract further entities for enrichment.

In this project datasets will be gathered from 25 different providers; museums, libraries, national archives and etc. The following analysis is done on 10 of the available datasets at the

time of writing this report (September 2019)[2]. These data sources were still on their original format as received from the providers. The enrichment process will be applied on the datasets when mapped into the EDM data model. Table 3 shows the dataset names along with their abbreviations used in the rest of this text.

Table 3. Name abbreviations

2048612_20170209_150009	221
AIU	AIU
Akadem_Europeana	AE
Archivio di Stato di Venezia	ASV
Hungarian Jewish Archive	HJA
Jewish Museum London	JML
Jewish Museum Prague	JMP
Medem	Med
Musei Sefardi Toledo	MST
National Library Israel	NLI

We have run several analysis on the datasets, to understand the structure and content of each dataset. Table 4 illustrates all the attributes extracted from each dataset. As shown in the table, datasets have different attributes and therefore different information. In this table, 'X' means that the certain attribute exists in the mentioned datasets. For example, datasets 221, AE, MST, and NLI all have the attribute 'creator' while the attribute 'geoname' only exists in JMP dataset. At the same time, we should keep in mind that sometimes attributes with different names, present the same information.
Each attribute can be useful for the enrichment process and therefore it is important to make sure that we don't lose any practical information and semantics during the mapping process.
Sometimes the value of attributes can be enriched directly. As in the following example;

```
<edm:ProvidedCHO rdf:about="#REB01: 000253485">
<dc:contributor xml:lang="fre">Gayus, Eliya</dc:contributor>
```

the 'contributor' attribute in the ProvidedCHO taken from one of the project datasets (AIUJE1_MARC21), can be enriched with further information about the stated person (Gayus, Eliya). The details of the enrichment process will be explained in the next sections.
Some of the attributes might be used to extract further entities and information which can be used for enrichment of other entities. For example we can use the 'date' attribute, which

[2] The analysis are based on the datasets which were made available by the providers at the time of the writing (Sep2019). We will update this section including table and analysis on February 2020.

represents the publication date for disambugation when there are several options for a title or a contributor. In Table 4, the blue cells are the attributes which were identified as the ones which can be enriched and the green cells are the attributes which can be informative for the enrichment process, as explained earlier. This selection can be modified as required.

Table 4. Attributes - The blue cells are the attributes which were identified as the ones which can be enriched and the green cells are the attributes which can be informative for the enrichment process. 'X' means the attribute was present in the corresponding dataset. The highlighted cells are the attributes which will be used for the contextualization.

	221	AE	AIU	ASV	HJA	JML	JMP	Med	MST	NLI
Agent	X									
aggregatedCHO	X									
Aggregation	X									
alternative	X	X						X	X	
altLabel	X									
archdesc							X			
author							X			
begin	X									
biographicalInformation	X									
broader	X									
c01							X			
c02							X			
Concept	X									
contributor	X	X							X	X
controlaccess							X			
created		X	X	X	X	X				
creation							X			

creator	X	X	X		X	X			X	
dao							X			
dataProvider	X	X	X	X	X	X		X	X	X
date							X	X	X	X
dateOfBirth	X									
dateOfDeath	X									
description	X	X	X		X	X		X	X	
did							X			
dsc							X			
ead							X			
eadheader							X			
eadid							X			
end	X									
exportedRecords	X	X	X	X	X	X		X	X	X
extent	X	X			X	X			X	
filedesc							X			
format	X	X	X					X	X	
geogname							X			
hasMet	X									
head							X			
identifier	X	X		X	X	X			X	X
isFormatOf									X	
isPartOf			X					X	X	

isReferencedBy									X	
isShownAt	X	X	X	X	X	X		X	X	
isShownBy				X	X	X		X	X	X
issued	X	X						X		
langmaterial							X			
language	1	X			X	X	X	X	X	X
langusage							X			
metadata	X	X	X	X	X	X		X	X	X
note	X									
object		X	X		X	X		X	X	X
p							X			
persname							X			
physdesc							X			
Place	X									
prefLabel	X									
profiledesc							X			
provenance			X							
ProvidedCHO	X									
provider	X	X	X	X	X	X		X	X	X
publicationstmt							X			
publisher	X	X						X	X	
RDF	X									
record	X	X	X	X	X	X		X	X	X

rights	X	X	X	X	X	X		X	X	X
scopecontent							X			
source	X			X	X	X			X	
spatial	X	X	X		X	X		X	X	X
subject	X	X	X	X	X	X	X	X	X	X
tableOfContents	X	X							X	
temporal									X	
TimeSpan	X									
title	X	X	X	X	X	X		X	X	X
titleproper							X			
titlestmt							X			
type	X	X	X	X	X	X		X	X	X
unitdate							X			
unitid							X			
unittitle							X			
unstored					X					
WebResource	X									

Our analysis also shows, the datasets have different formats and content type. For example, there is only one dataset originally available in EDM format which has Provided CHO object and as tables 5 illustrates the types of dataset content, some of them are for example text (such as books or newspaper articles) or images (such as photographs of art or historic photos).

Table 5. Type across datasets

Dataset	Values
221_type	TEXT
AE_type	VIDEO
AIU_type	IMAGE
ASV_type	TEXT
HJA_type	Archival document
HJA_type	TEXT
HJA_type	photograph
HJA_type	IMAGE
HJA_type	card
JML_type	audio recording
JML_type	SOUND
JML_type	TEXT
Med_type	TEXT
MST_type	Libros
MST_type	TEXT
MST_type	Manuscritos
NLI_type	TEXT

Sample Datasets - Complementary Analysis

We also did some further analysis on the sample datasets which were already mapped to the EDM data model (Table 6). If the entity which is intended to be enriched, such as an Agent, has further information provided, for example an identifier, then identifying it in and linking it to other resources can be easier and more accurate. However, if there are not any extra information available, then for the enrichment process, we need to entity disambiguation and try to extract some distinctive information from other attributes such as 'description'.

Table 6. ProvidedCHO and Agent properties

edm: ProvidedCHO	dc: subject	**edm: Agent**	rdaGr2: dateOfBirth
	edm: hasView		rdaGr2: placeOfBirth
	dc: temporal		rdaGr2: dateOfDeath
	dc: object		rdaGr2: placeOfDeath
	dcterms: spatial		dc: identifier
	dc: description		owl: sameAs
	dc: date		skos: prefLabel
	owl: sameAs		skos: altLabel
	dc: creator		rdaGr2: gender
	dc: title		
	edm: rights		
	dc: language		
	edm: dataProvider		
	edm: aggregatedCHO		
	dc: publisher		
	edm: provider		
	dc: format		
	edm: isShownAt		

In the following we have selected two attributes from the sample datasets, ‘contributor’ and ‘spatial’ and have illustrated as examples that what extra information would be added to them, if enriched with the JudaicaLink data sources.

dc:contributor ‘Caceres, Abraham’	***dcterms:spatial ‘Paris’***
rdf:type Person skos:altLabel Abraham Caceres (es) Abraham Caceres (en) Caceres, Abraham (de) Caseres, Avraham Casseres, Abraham de Casseres, Avraham Cáceres, Avraham jl:Occupation jld:occupation/composer skos:prefLabel Caceres, Abraham owl:sameAs <http://dbpedia.org/resource/Abraham_Caceres> <http://es.dbpedia.org/resource/Abraham_Caceres> <http://rdf.freebase.com/ns/m.0fqlq84> <http://viaf.org/viaf/38500608> <http://wikidata.dbpedia.org/resource/Q4668794> <http://www.wikidata.org/entity/Q4668794> <http://yago-knowledge.org/resource/Abraham_Caceres> <http://d-nb.info/gnd/1070357049> <http://d-nb.info/gnd/1070357049/about> <http://hub.culturegraph.org/entityfacts/1070357049> <http://viaf.org/viaf/315583232> <http://data.judaicalink.org/data/gnd/1070357049> <http://data.judaicalink.org/data/rdf/dbpedia/Abraham_Caceres> gndo:gndIdentifier 1070357049 dc:Subject <http://dbpedia.org/resource/Category:Jewish_composers>	rdf:type Concept geo:as WKT Point (+002.348800 +048.853409) gndo:gndIdentifier 4044660-8 skos:prefLabel Paris Same As <http://d-nb.info/gnd/1019630-4> <http://sws.geonames.org/2988507> <http://viaf.org/viaf/265224327>