Designing a Multilingual Knowledge Graph as a Service for Cultural Heritage – Some Challenges and Solutions

Valentine Charles
Europeana Foundation,
The Netherlands
valentine.charles@europeana.eu

Hugo Manguinhas
Europeana Foundation,
The Netherlands
hugo.manguinhas@europeana.eu

Antoine Isaac
Europeana Foundation,
The Netherlands
antoine.isaac@europeana.eu

Nuno Freire
INESC-ID, Portugal
nuno.freire@tecnico.ulisboa.pt

Sergiu Gordea
[3]AIT Austrian Institute of Technology, Austria
sergiu.gordea@ait.ac.at

Abstract

Europeana gives access to data from Galleries, Libraries, Archives & Museums across Europe. Semantic and multilingual diversity as well as the variable quality of our metadata make it difficult to create a digital library offering end-user services such as multilingual search. To palliate this, we are building an "Entity Collection", a knowledge graph that holds data about entities (places, people, concepts and organizations) bringing context to the cultural heritage objects.

The diversity and heterogeneity of our metadata has encouraged us to re-use and combine third-party data instead of relying only on those contributed by our own providers. This raises however several design issues. This paper lists the most important of these and describes our choices for tackling them using Linked Data and Semantic Web approaches.

Keywords: linked data; knowledge graph; Europeana.

1. Introduction

Europeana gathers over 50 million paintings, books, newspapers, audio recordings, etc., from more than 35 European countries and in more than 40 languages. With such a diversity, supporting users in their (multilingual) search and browsing activities is a challenge. The vision of Linked Open Data applied in the cultural sector (Gradmann, 2010) has led us into collecting more data about contextual entities such as people, places, concepts next to Cultural Heritage Objects' (CHOs) metadata. The Europeana Data Model (EDM) (Europeana, 2016) enables our data partners to describe contextual entities as Linked Data (LD) resources with their own URI identifiers instead of literals. In addition, to increase the semantic and multilingual coverage of its metadata, we perform automatic semantic enrichment of our dataset by linking literals found in the CHO metadata to linked open multilingual datasets such as GeoNames[1] and DBpedia[2] - see documentation and examples at (Europeana, 2018). The number of links between CHOs and contextual entities as well as of data containing multilingual labels has thus grown considerably. However, this richer data is still heterogeneous: different providers use resources with different, not necessarily entirely commensurate, semantic and multilingual characteristics, while others do not use any such resources at all.

To palliate this, we have begun to select and combine statements from various LD sources into an "Entity Collection" (EC), a knowledge graph (KG) centralising data about contextual entities.

[1] http://www.geonames.org/
[2] http://wiki.dbpedia.org/

https://doi.org/10.23106/dcmi.952138993

The EC is intended for use by several Europeana services, most immediately as a means to improve the users' experience in their search for CHOs (Hill et al., 2016a). It is designed to enhance:

- **Findability:** users can refine their search by filtering and browsing on people, places and subjects. Using the EC data for semantic enrichment reduces ambiguity in the CHO metadata, clarifying its meaning and improving its interlinking. Multilingual search benefits significantly from the multiple labels typically associated with each entity. For instance, an Entity auto-completion feature would use the EC to power search by keyword, returning a list of entities that have a label that matches what the use has typed, for any language available in the EC.
- **Contextualisation:** users can see additional contextual information related to specific CHOs. The EC can support annotation scenarios (semantic tagging) by suggesting entities to be used as tags instead as mere strings.
- **Exploration:** users can browse the relationships between CHO resources and entities. For instance, if an Entity created a CHO, a user could access the CHO via the page dedicated to that Entity, or access to more details about the Entity from the CHO item page.

The building of the EC has raised several challenges, motivating design decisions and solutions that we report in this paper. Section 2 presents related work on the activities involved in the creation, population, sharing and re-use of KGs. Building a KG such as the EC as an operational service requires well-designed processes for importing entities from external data sources and making the data available for exploitation, while maintaining data integrity and freshness as these sources evolve. The main activities and automatic processes involved are presented in Fig.1 and described in sections 3 and 4. We finish with a summary of our activities and future work.

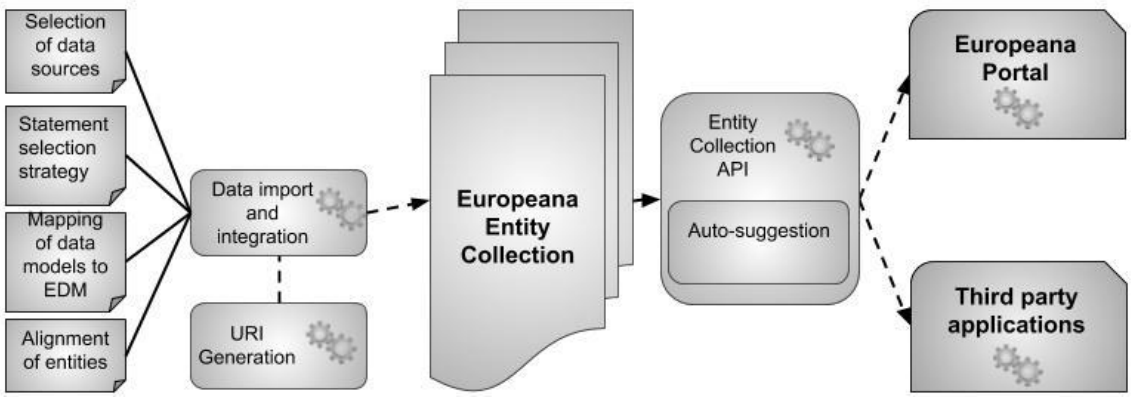

FIG. 1: Overview of Entity Collection processes in Europeana

2. Related Work

KGs have been created to solve data heterogeneity and quality issues, to structure and organise back-end datastores, and to provide advanced end-user services. They are typically intended to unify and enhance existing data, providing a centralised service capable of addressing issues of (query) disambiguation, responsiveness, relevance ranking, data enrichment, etc.

The best-known KG implementation is perhaps Google's Knowledge Graph, which exploits information extracted from a number of web sources (Dong et al., 2014). In the LD community, DBpedia has long played a key role for providing a large, open body of knowledge that others can re-use and link to (Auer et al., 2007). Wikidata[3] is another example of a (crowdsourced) open database, which is also used as a data source of Google's own KG.

[3] http://wikidata.org

https://doi.org/10.23106/dcmi.952138993

Gabrilovich & Usunier (2016) presents the many research aspects involved in the creation of KGs: relation extraction, conversion and mapping, ontology matching, etc. Not all of these, however, are relevant for Europeana. For example, Dong et al. (2014) and Szekely et al. (2015) focus on the problems of knowledge extraction and merging from large set of automatically extracted data, including unstructured and structured sources. We do not aim to operate at such scale, instead focusing on building a KG on top of already extracted and structured knowledge.

DBpedia and Wikidata integrate different sources too. But their information-orientation is different. DBpedia extracts data from semi-structured sources in one information space (Wikipedia). Wikidata sources data from the crowd. In both cases there is no range of pre-existing external 'official' sources. In particular, the modelling of the data can be decided based on what is available (and needed) in the 'information ecosystem', which is directly at hand. There can be conflicts in the data though, i.e., statements reflecting views of different Wikidata contributors (or their sources). To address this, Wikidata handles provenance at a very granular level (individual statements). Multilingualism - a key issue for us - is also a focus in both initiatives: DBpedia separates language editions, but seeks to interconnect them as much as possible, while Wikidata starts with language-neutral resources and adds language-specific information about them.

BabelNet4 is another KG that heavily focuses on multilingualism. It links some 16 million entities across 284 languages. In terms of data integration, it sits 'above' Wikidata, including it as a dataset alongside many other data sources, including GeoNames and Wordnets for various languages (Navigli & Ponzetto, 2012). Like some other KGs, it is also not open enough: its license prevents the sort of partial re-publication Europeana performs to provide its (open) services.

Other relevant work includes efforts on tackling specific problems of KG creation. A lot of work in the domain focuses on ETL aspects, such as mapping and conversion of one dataset into a KG (Pellissier et al., 2016): but unlike many of these efforts, our EC is not about publishing legacy data as LD. Rather, we are re-users of already published and curated data. In addition, we do not need to represent all the information from the data sources that we re-use for our KG: we can and should focus on the most useful parts for us and our re-users5. We expect that designing our EC needs to combine automatic and manual processes where the organizational setting is clear and that it will in the first instance benefit from wider discussions on management of data flows such as versioning, archiving and on the documentation of changes, along the lines of the OAIS reference model ("Open Archival Information System ", n.d.).

More directly relevant to our case, considerable work has been devoted to "reconciliation" (aka. "matching" or "alignment") of entities across datasets. This is a vital concern for Europeana, as the sources we seek to use can have overlapping scopes. Automatic matching (Euzenat & Shvaiko, 2013) as well as manual and semi-automated approaches (Ossenbruggen et al., 2011) can be relevant here. The problem can be also mitigated by selecting sources (or parts thereof) with very limited (or no) overlap.

We envision our KG as being built by in-house specialists in cultural-sector data, and we count on our active network of data partners to flag relevant data sources to integrate, e.g., because their scope would match well their datasets. Instead, related work in search can be more relevant for our attempts to provide discovery services, especially searches for entities, ranked by their relevance, as e.g. Google provides for their KG (Google, 2018). (see Section 4.3 for our choices on ranking)

[4] http://babelnet.org/

[5] For example, the DBpedia to EDM mapping only captures the information Europeana needs: https://github.com/europeana/tools/blob/master/europeana-enrichment-framework/enrichment/enrichment-framework-knowledgebase/src/main/resources/dbpedia2agent.xsl

https://doi.org/10.23106/dcmi.952138993

General best practices for publishing data are also relevant. The W3C recently published Data on the Web Best Practices (Farias Lóscio et al., 2017) with recommendations such as "reuse vocabularies, preferably standardized ones", which especially argues for not re-inventing the wheel in terms of the classes and properties used to express structured data. Europeana does not refrain from minting its own classes and properties when needed. But the position of our EC as a service built on top of existing data and which needs to remain interoperable with the data others publish in our community, raises a strong requirement for re-using existing ontologies. This is a difference with e.g. DBpedia and Wikidata, which create specific ontologies and align them afterwards with existing vocabularies when possible. Szekely et al. (2015) have adopted an existing ontology, Schema.org6, which is also used by Google. Another recommendation is to "make data available through an API". We aim to make available, at a minimum, an entity discovery service, alongside raw access to data via LD content negotiation for entities, batch dump access and an expert (and difficult to maintain) SPARQL endpoint. Like Google, DBpedia provides a simple text-based entity look-up service. Wikidata provides the full MediaWiki API, geared towards the retrieval of Wiki pages; access to data is chiefly handled through the LD content negotiation, dumps and a full SPARQL query service.

The sector of Galleries, Libraries, Archives & Museums (GLAM) has recognized early the potential of Linked Open Data and several efforts have been carried out, which can be compared to ours. Organizations have released contextual entities from their legacy vocabularies, gazetteers and authority lists. Concepts, person names and place names from the Getty Museum Art and Architecture Thesaurus (AAT), Union List of Artist Names (ULAN) and Thesaurus of Geographic Names (TGN) are available via content negotiation and a SPARQL endpoint (Getty, 2018). The German National Library has published its reference set of resources (GND) as LD (DnB, 2018a) similarly to the French, American and Spanish National Libraries.

While these efforts chiefly aim at publishing data from relatively isolated (institutional) information spaces, they try to create links to other datasets, starting with their peers. Some projects are dedicated to 'network' reference datasets. OCLC's Virtual International Authority File7 (VIAF) merges person and organization data from authority lists from more than 50 national libraries and agencies. It serves a unified description of each authority next to links and the original data from each library, see for example: http://viaf.org/viaf/9847974.rdf. The German National Library runs the Entity Facts service serving GND data combined with other datasets, including VIAF (DnB, 2018b). The SNAC project8 has performed a merging of data for persons found in archive collections. It connects its data to others, such as Getty's ULAN. Cross-datasets links can already be present in the original data or require semi-automatic reconciliation. Often a mixture of both happens, i.e., legacy identifiers from external datasets are found in the records of a source dataset and these implicit links need to be made explicit as URI references (e.g. https://www.europeana.eu/portal/en/record/90402/SK_A_4691.html which has identifiers from the Rijksmuseum and Europeana).. This renders the alignment processes often very specific to the data at hand - say, library and archive records could use quite different matching scripts.

The thematic project Europeana Food and Drinks has performed an interesting experiment, selecting relevant concepts from general datasets like DBpedia and linking them to institutional datasets to form a common "classification" for the project (Alexiev, 2015). They compared the multilingual interest of the various options available. This is similar to what we intend for our EC. We need to address a wider scope across subjects and types of collections, however, as well as publish our data in channels that can serve more purposes.

Note that despite their specificities we can benefit from these GLAM-related efforts from a data representation perspective, as most of them adhere to the principle of re-using existing ontologies. Some are also great examples regarding the distribution of the data. For example, the

[6] http://schema.org
[7] http://viaf.org/
[8] http://socialarchive.iath.virginia.edu/

https://doi.org/10.23106/dcmi.952138993

STW thesaurus for economics has a web service[9] that is exemplar of the way SKOS-like concept vocabularies can be served via a web API. DigitalNZ, a GLAM aggregator like Europeana, provides a Concepts API for its data re-users (DigitalNZ, 2015). Finally, OCLC's Worldcat Identities project (O'Reilly, 2007) is a good example of how entities can be used to provide novel ways to find and explore objects.

3. Building and Making Available a Knowledge Graph for Europeana

Europeana data experts and officers take the strategic decisions needed to import, integrate and manage data in our KG, including criteria to select data sources, and maintain our data model to represent and map the entities to the data. They perform the configuration and regularly execute the import and update of entities, which are then made available through a dedicated API.

3.1. Selection of Data Sources

Selecting data sources (or parts thereof) to integrate in the EC requires an intellectual effort prior to the actual harvesting and import of the data. It implies analysis of external data by a data expert and application of selection criteria. Europeana's strategy relies on leveraging existing linked open datasets and vocabularies and the following criteria to evaluate and select data sources (Isaac et al., 2015):

- **Availability and Access:** The datasets should be available on the Web and compliant with the LD recipes. They should be re-usable under an open license.
- **Granularity and Coverage:** The datasets should have the same coverage or should obviously complement each other. Reconciling resources that are semantically too far from each other could introduce ambiguities or semantic flaws for entities. For Europeana the data sources should answer to Who?', 'What?', 'When?', 'Where?' questions that are the most relevant to the cultural heritage domain as they help contextualise CHOs. Language coverage is also a key requirement: we aim to support over 29 languages in which Europeana receives metadata as reported in (Hill et al., 2016b). Ideally a dataset should provide labels in all the languages supported by Europeana or contribute with the labels necessary to reach such coverage. Generic data sources in terms of coverage or granularity are also likely to introduce semantic flaws during manual or automatic enrichment processes (see below on 'size').
- **Quality:** This includes intrinsic aspects of the dataset that can be manually or automatically assessed, such as the structure and representation of values and languages.
- **Connectivity:** The richness of the EC will be improved if the selected datasets have incoming and outgoing links to other datasets.
- **Size:** Depending on the size of the selected dataset, the number of entities is a criterion of selection. A high number of resources and statements is preferable, if the alignment process can deal with the greater ambiguity (i.e., higher number of entities associated with a given name) that larger sizes tend to generate. For example, GeoNames has 7.5M place names. The name "Guadalajara" limited to Mexico returns over 15 places, a lot of them are small *pueblas* with population under 15.

The need for a consistent and value-adding EC dictates a careful strategy for balancing domain-specific sources with more generic ones while addressing issues of semantic grain mismatch. We tend to choose general "pivot" datasets to cover as many entities as possible. For instance, Europeana might favour Wikidata over domain specific vocabularies such as Getty's AAT. Yet, in some cases we may want to give precedence to complementary datasets for more specific entities. Complementarity is not only relevant for entity-level data but also for CHO-level metadata: for instance, a dataset that includes metadata for CHOs could be used to create abstract "work"-level entities for our own CHOs, as it is often the case in library metadata. Note that the question of selecting pivot data sources vs. complementary (or domain) ones is

[9] http://zbw.eu/beta/econ-ws/about

https://doi.org/10.23106/dcmi.952138993

independent from the actual alignment of entities in the EC (whether merging entity resources or representing matches between them as links, which preserves the original data).

The next step is to choose entities to be imported in the EC. The manual selection of individual entities from a data source is time-consuming and unfeasible for large sources. A query scenario is therefore envisioned, where a user can define the selection by designing queries to a data source (if a query service is available) that implement the appropriate selection criteria. For instance, in order to only import in the EC DBpedia data related to artists, a filter query would be created based on the statement pattern *anEntity rdf:type dbp:Artist* .

3.2. Data Modelling, Mapping and Statement Selection

Building a KG requires data to be represented in a consistent way. Each linked entity in the EC is an instance of a contextual class as defined in the EDM for representing people (*edm:Agent*), places (*edm:Place*), concepts (*skos:Concept*), time periods (*edm:Timespan*) or organizations (*foaf:Organization*). Mappings are created between the data model of a selected data source and EDM[10]. Custom mappings to EDM are needed to select the relevant information and the properties for given entities. This process is made easier (if not trivial) when the data sources are based on SKOS (Simple Knowledge Organisation System) (Miles & Bechhofer, 2009) which EDM re-uses for describing concepts and also preferred and alternatives labels for people, places, time periods and organizations. Note that besides the top-level classes above, most of the EDM elements[11] come from ontologies used in (cultural heritage) linked datasets, such as Dublin Core, RDA, and FOAF. EDM also seeks to adhere to the W3C best practice "choose the right formalization level": we refrain from adding too many formal axioms that would make mappings harder and perhaps disqualify good data sources without a serious reason besides elegance of modelling.

We also use mappings to select statements to be imported in the EC, e.g. by filtering out properties, (sub-)types of entities or specific resources (URIs), if they are irrelevant for Europeana. Note that Europeana does not need every statement from the selected datasets, e.g., labels for languages that it does not support (in GeoNames) or entities not relevant for Cultural Heritage such as modern pop stars (in DBpedia)[12].

3.4. Data integration, Reconciliation, Alignment and Curation

After being imported in the EC, the new entities need to be integrated with the existing EC entities. This step consists in the following workflow – some components of which have been already implemented as part of the semantic enrichment mentioned earlier:

Integration and reconciliation of entities. Imported entities are integrated with existing EC entities (i.e., the statements about these two entities are merged) or new corresponding entities are created (i.e., a new Europeana URI is minted). This is supported by the execution of automated background data-processing jobs, with scheduling, notification and reporting functionalities. Entity data will be previewed before integration into the EC for quality control purposes. The integration strategy may be influenced by the selected data sources. For instance, using Wikidata as a pivot data source for all the Europeana entities would make it easier to reconcile entities within the EC, as it is very rich in alignments to datasets in our sector (e.g., VIAF). Wikidata would then be used as a source from which Europeana could access other vocabulary alignments.

Alignment of entities. The detection of duplicates within the EC is currently based on the co-referencing information found in the data (*owl:sameAs* or *skos:exactMatch* links). We do not

[10] The mappings we use for the EC source datasets (DBpedia, GeoNames, etc) can be found at https://github.com/europeana/tools/tree/master/europeana-enrichment-framework/enrichment/enrichment-framework-knowledgebase/src/main/resources

[11] See a full listing at https://github.com/europeana/corelib/wiki/EDMObjectTemplatesEuropeana

[12] See for example the list of filtered agents: https://docs.google.com/spreadsheets/d/1Wu8gPsgdtwnDN-GSuettT8WwqmvTeHaeAlqBF8-_joE

https://doi.org/10.23106/dcmi.952138993

exclude the possibility of creating alignments using (semi-)automatic or manual tools such as Mix'n'match[13] and CultuurLink, following up on recent experiments (Manguinhas et al., 2016). We have found that despite selecting large datasets we are still missing a lot of coreferencing information to other datasets (chiefly domain vocabularies, but also reference datasets such as VIAF).

M*anual curation of entities and/or data*. As an additional step to maintain integrity, curators from Europeana staff will be able to edit the data for a Europeana entity by adding, changing or removing statements (including alignments), without preventing future updates from the imported data sources. Existing entities may also be deprecated.

These workflows will also benefit from additional normalisation and cleaning rules to apply to the data collected for each entity, as hinted from some "matching rules" presented in the documentation of Europeana's automatic semantic enrichment (Europeana, 2018). For instance, labels and values are not always accurate, and are sometimes even missing.

3.5 Data Integration Strategies

The management of the data within the EC has raised key data integration problems, which we are still discussing at the time of writing.

The main issue concerns when descriptions coming from different sources require merging, i.e. whenever two or more resource descriptions exist for the same entity. A choice is needed regarding which statements will be prioritised to become part of the description for the resulting Europeana entity. We have identified several options:

- **Unification**. The simplest strategy is to unify all statements coming from the different datasets into a single description. However, this strategy may lead to inconsistencies, e.g. cases where more than one statement exists for the same property when only one is allowed (e.g. the birthplace of a Person is stated in source A to be a country while source B is more granular and states the city) and contradictory statements (e.g. two distinct birth dates for the same Person).
- **First come / first serve**. This strategy considers an order (for the source datasets) while selecting the statements for the entity description. While copying a statement in the EC, the cardinality constraints defined for a given property are enforced by skipping the statement once the maximum is reached. The order in which the source datasets are merged may be defined to reflect the distinction between the pivot and complementary datasets, so that a pivot takes precedence by being the first to be considered for merging.
- **Most representative**. This strategy chooses among conflicting statements based on the number of source datasets that contain them. This assumes that if a statement is found in more datasets, it is more likely to be "true". However, there can be situations where incorrect statements may be spread, as many datasets integrate data from other sources, replicating the issue. Also, the strategy does not define how a statement can be chosen in case of a tie.
- **Differentiated most representative**. This more complex strategy tries to balance pros and cons from the previous strategies by distinguishing the datasets into two explicit groups (pivot and complementary). For competing statements *within* a group, this strategy may apply the "most representative" or the "first come / first serve" strategies. Then, statements from the pivot group are copied, and statements from the second group are added - while preserving cardinality constraints.

Any chosen integration strategy will be supported by provenance and attribution information capturing the source of a given entity or statement (e.g. tracking the source URIs in an *owl:sameAs* or *skos:exactMatch* for an entity or RDF Graphs for statements).

[13] http://tools.wmflabs.org/mix-n-match/

https://doi.org/10.23106/dcmi.952138993

3.6. Data in the Entity Collection

The current data available in the EC inherits from the data sources previously harvested to underpin Europeana semantic enrichment. As of May 2018, the EC contains data for:

- 215.802 Places: a subset of **Geonames**, corresponding to places part of European countries and of a specific feature class[14]. (i.e. "A", "P.PPL", "S.CSTL", "S.ANS", "S.MNMT"...)
- 165.005 Agents: a subset of **DBpedia** corresponding to most of the instances of *dbp:Artist* with some exceptions, and integrated from 49 DBpedia language editions. All locale DBpedias that match the list of languages supported by Europeana have been harvested from which a selection is made to enrich concepts and persons.
- 1.572 Concepts: a subset of **DBpedia** comprising a handful of WWI battles, the "World War I" category and other categories[15] being used for Europeana Collections. And also two vocabularies: one for music genres, forms and compositions obtained from Wikidata and the photography vocabulary maintained by the Photo Consortium.
- 599 Organizations: data about Europeana's data partners collected through our Customer Relationship Management (CRM) system. Co-references to Wikidata were added when available and represented as *owl:sameAs* relations.

We will add more entities, first from the data sources we already ingest, and then extending to other data sources, especially Wikidata (see Section 3.1 for our motivations), as well as time spans, which are not yet represented in the EC.

4. Accessing the Entity Collection Data

The EC is made available via an API ("Europeana Entity API", n.d.), which powers the search query auto-completion and the entity pages in Europeana.

4.1 Entity Collection Look-up API

Two API methods are available to look up for entities in the EC. The first one uses content negotiation to deliver data in HTML or JSON-LD formats, according to the client preferences indicated through the HTTP request header. A known entity can be accessed using its URI; the content negotiation service will automatically redirect the request either to the Entity API endpoint, or to the Entity Page in Europeana16.

The second method enables to look up an EC entity using an alternative URI that is recorded in the source dataset. This lookup uses the owl:sameAs and skos:exactMatch co-reference statements available within the entity data and returns a redirection in line with common HTTP best practices. This method is a key requirement for semantic integration of Europeana KG with the existing linked data repositories.

The default format chosen for representing the entities and facilitate the re-use of the data in the EC is JSON-LD (Sporny at al. 2014), the JSON representation for LD. This format was chosen as it is commonly used in Web-based programming environments, to build interoperable Web services. It can also be used when data is integrated in other pieces of JSON data, such as the ones returned by the autosuggestion API (see Section 4.3). To make the JSON-LD serialisation more compact, we have defined a JSON-LD context, which defines abbreviations for the namespaces used in EDM and specific data types (e.g., http://rdvocab.info/ElementsGr2/gender can be simply referred to as "gender"). The data thus becomes better understandable by (third party) web developers without affecting the underlying semantics. Some EDM properties can be used with several values in different languages, such as

[14] http://www.geonames.org/statistics/total.html

[15] See: https://docs.google.com/spreadsheets/d/1qjyyneg6aMoPC2v5hwC8YinmHKNyJtvTJp1HJdnnPc8

[16] For instance, http://entity.europeana.eu/entity/agent/base/146741?wskey=apidemo. NB: at the time of writing one still needs a key to de-reference these URIs. This will be changed later.

https://doi.org/10.23106/dcmi.952138993

skos:prefLabel, skos:altLabel, foaf:name. To facilitate standardisation, this context is available as a separate resource17 which can be referenced in the JSON-LD serialization of the contextual entity.

RDF/XML will be also supported as it is commonly used, especially for CHO metadata ingestion at Europeana.

4.2 Generation of URIs in the EC

An important aspect of data integration in the EC is the generation of URIs for every entity. Our design is based on a LD scenario where URIs must be (i) **Dereferenceable**, both humans and user-agents must be able to meaningfully resolve the URI (ii) **Unambiguous**, a URI should not refer to two distinct resources (iii) **Immutable**, it should not change in time. As Europeana holds data which is not available elsewhere as a whole, it needs to create URIs in its own namespace (data.europeana.eu), so that a data consumer can access and retrieve the data. Identifiers need to be both easy to assign and future-proof. URIs follow the pattern: *http://data.europeana.eu/{entity_class}/{scheme}/{localID}*

- {entity_class} corresponds to the types of EDM contextual entities (Agent, Place, Concept and Organizations).
- {scheme} represents a sub-division under each entity class. A special division with the name "base" will contain all entities that are integrated from external data sources.
- {local_id} is the local identifier for the entity.

For the local identifier we chose to generate a sequential identifier for entities that are collected from external sources since it is the type that requires less effort to assign and maintain (Archer et al., 2012). The choice of minting human readable URIs was discussed and rejected within our community (Europeana, 2015) as it increases complexity for both maintenance and data consumption. Such URIs could be envisioned as alternative URIs. A more practical alternative to human readable URIs is to have URLs that, after content negotiation, contain a human readable part. This would have no impact on data consumption and would require considerably less effort to implement and maintain.

4.3 Discovery of Entities

The API provides another two methods for discovery and retrieval of entities in the EC:

- *entity auto-completion:* implementing quick search by entity names. This type of discovery, integrated in Europeana to support end-users to formulate more precise search queries, is based on entity labels only (i.e. *skos:prefLabel, skos:altLabel, edm:acronym*).
- *entity search:* supporting retrieval of entities by using free querying on all properties or on (a combination of) individual properties. The latter enables advanced search scenarios, e.g. finding cities in a given country (using *edm:isPartOf*), or fashion designers born in the XIXth century (e.g by using *rdagr2:professionOrOccupation* and *rdagr2:dateOfBirth*)

Recommending entities for search auto-completion is a challenge, given the requirement for achieving a high precision for suggestions in the top 10 list. Moreover, the multilinguality of the EC and the search queries (users often search in Europeana using their native language) add to the difficulty. The ranking of individual entities uses a formula that integrates and normalizes two measures: relevance and popularity. The relevance of an entity is computed as the number of Europeana records that contain one of the entity labels, while its popularity is computed using the Wikidata PageRank, as calculated across 133 of its languages versions (Diefenbach & Thalhammer, 2018). Preliminary testing has indicated that this approach yields good results, though the need for cross-linguistic matching due to the modest average multilingual coverage

[17] http://www.europeana.eu/schemas/context/entity.jsonld

https://doi.org/10.23106/dcmi.952138993

currently limits the precision of the suggestions. Future work will investigate the employment of a Learning-To-Rank approach to improve the ranking of individual entities based on the information captured within the Europeana access logs.

The entity search has a generic implementation, allowing API users to formulate complex queries following the Solr query syntax ("SOLR Query Syntax," n.d.). Built-in statistics on EC are made available via facet profiles. For example, the faceted field on the property type provides, in real-time, the number of Agents, Concepts, Places and Organizations available in the EC (see also Section 3.6). The presentation of the search results uses pagination as specified by the Linked Data Protocol (Speicher et al., 2015). Applications that integrate search can thus easily fetch all results by issuing a chain of calls for the *next* (page) URL, which is available in every response.

5. Conclusion and Future Work

This paper has presented different requirements, highlights challenges and proposes solutions to adopt when building a knowledge graph for cultural heritage.

Solutions to some of the problems and questions raised in the paper have been found sufficient to allow the creation of a first version of the EC. However, some decisions still need to be taken to ensure the coherence of the EC over time

Data coverage and Extensibility. Europeana needs to expand its EC to cover as many CHOs as possible and support 'client' Europeana services. Future work includes the sourcing of suitable datasets to represent times periods as well as named events.

Data integration strategy. Both automatic and manual curation approaches need to be considered. Future work includes the improvement of the quality of current data by removing statements with no or faulty language tags, filtering unwanted statements or entities, refining the data mappings to include new statements, etc.

Enrichment. The EC will be used to enrich the Europeana metadata still represented as literals (the process mentioned above still uses a separate database).

Discoverability. The mapping work from Schema.org to EDM (Wallis et al., 2017) will allow the entities to be indexed by search engines and therefore more discoverable for the users. For instance the inclusion of owl:sameAs links from the Google Knowledge Graph in Schema.org markup would maximise the chance of the Europeana content to be displayed in KG cards.

References

V. Alexiev. (2015). Europeana Food and Drinks Deliverable D2.2 Classification Scheme. http://foodanddrinkeurope.eu/wp-content/uploads/2015/12/D2.2-Classification-scheme.pdf

P. Archer, S. Goedertier, N. Loutas. (2012). D7.1.3 - Study on persistent URIs, with identification of best practices and recommendations on the topic for the MSs and the EC. Project Interoperability Solutions for European Public Administrations. https://joinup.ec.europa.eu/sites/default/files/document/2013-02/D7.1.3%20-%20Study%20on%20persistent%20URIs.pdf

S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z. Ives. (2007). DBpedia: A Nucleus for a Web of Open Data. In: The Semantic Web. Lecture Notes in Computer Science, vol 4825. Springer, Berlin, Heidelberg.

D. Diefenbach and A. Thalhammer. (2018). PageRank and Generic Entity Summarization for RDF Knowledge Bases. In: Lecture Notes in Computer Science, vol 10843. Springer, Berlin, Heidelberg.

DigitalNZ. Introducing the DigitalNZ Concepts API (2015). Retrieved August 13, 2018 from https://digitalnz.org/blog/posts/introducing-the-digitalnz-concepts-api

DnB - Deutsche National Bibliothek. (2018a). Linked Data Service of the German National Library. Retrieved August 13, 2018 from http://www.dnb.de/EN/Service/DigitaleDienste/LinkedData/linkeddata_node.html

DnB - Deutsche National Bibliothek. (2018b). Entity Facts. Retrieved from http://www.dnb.de/EN/Wir/Projekte/Abgeschlossen/entityFacts.html

X. Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, T. Strohmann, S. Sun,and W. Zhang. (2014). Knowledge vault: a web-scale approach to probabilistic knowledge fusion. In Proc. 20th ACM SIGKDD

https://doi.org/10.23106/dcmi.952138993

International Conference on Knowledge Discovery and Data Mining (KDD). 24-27 August, 2014, New York, USA. 601-610.

Europeana Foundation. (2015). Staying persistent: EuropeanaTech community help to pave the way for new Europeana URIs. Retrieved August 13, 2018 from http://pro.europeana.eu/blogpost/staying-persistent-europeanatech-community-help-to-pave-the-way

Europeana Foundation. (2016). Definition of the Europeana Data Model elements v5.2.7. Retrieved August 13, 2018 from http://pro.europeana.eu/edm-documentation

Europeana Foundation. (2018). Automatic Semantic Enrichment at Europeana. Retrieved August 13, 2018 from https://pro.europeana.eu/page/europeana-semantic-enrichment#automatic-semantic-enrichment

(n.d.). Europeana Entity API (alpha). In: Europeana Pro. Retrieved August 13, 2018 from https://pro.europeana.eu/resources/apis/entity

J. Euzenat and P. Shvaiko. (2013). Ontology Matching. 2nd Edition, Springer, 2013.

B. Farias Lóscio, C. Burle, N. Calegari (eds). (2014). Data on the Web Best Practices. W3C Recommendation. 31 January 2017. https://www.w3.org/TR/dwbp/

Getty Research Institute. (2018). Getty Vocabularies as Linked Open Data. Retrieved August 13, 2018 from http://www.getty.edu/research/tools/vocabularies/lod/

Google. (2018). Google Knowledge Graph Search API. Retrieved August 13, 2018 from https://developers.google.com/knowledge-graph/

E. Gabrilovich and N. Usunier. (2016). Constructing and Mining Web-scale Knowledge Graphs. Tutorial. In Proc. 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. Pisa, Italy, 17-21 July, 2016

S. Gradmann. (2010). Knowledge = Information in Context: on the Importance of Semantic Contextualisation in Europeana. Europeana Whitepaper. Retrieved August 13, 2018 from http://pro.europeana.eu/publication/knowledgeinformation-in-context

T. Hill, D. Haskiya, A. Isaac, H. Manguinhas, and V. Charles. (2016a). Europeana Search Strategy. Europeana Whitepaper. Retrieved August 13, 2018 from: http://pro.europeana.eu/publication/europeana-search-strategy

T. Hill, A. Isaac, V. Charles, N. Freire and H. Manguinhas. (2016b). Europeana Search Strategy. Europeana Whitepaper. Retrieved August 13, 2018 from: http://pro.europeana.eu/publication/europeana-search-strategy

A. Isaac, H. Manguinhas, V. Charles, J. Stiller (eds). (2015). Selecting target datasets for semantic enrichment. Companion document to the report of the EuropeanaTech Task Force on Enrichemnt and Evaluation. Retrieved August 13, 2018 from https://pro.europeana.eu/project/evaluation-and-enrichments

H. Manguinhas, V. Charles, A. Isaac, T. Miles, A. Lima, A. Néroulidis, V. Ginouvès, D. Atsidis, M. Hildebrand, M. Brinkerink, S. Gordea. (2016). Linking subject labels in Cultural Heritage Metadata to MIMO vocabulary using CultuurLink. In: Proc. 15th European Networked Knowledge Organization Systems Workshop (NKOS 2016) co-located with the 20th International Conference on Theory and Practice of Digital Libraries 2016 (TPDL 2016). Hannover, Germany, 9 September (2016).

A. Miles, S. Bechhofer, (eds.). (2009). SKOS Simple Knowledge Organization System – Reference. W3C Recommendation. https://www.w3.org/TR/skos-reference/

R. Navigli and S. Ponzetto. (2012). BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. Artificial Intelligence, 193, Elsevier, 2012, 217-250.

(n.d.). Open Archival Information System. In Wikipedia. Retrieved August 13, 2018 from https://en.wikipedia.org/wiki/Open_Archival_Information_System

T. O'Reilly. (2007). WorldCat Identities. In: O'Reilly Radar. [Blog post]. Retrieved August 13, 2018 from http://radar.oreilly.com/2007/02/worldcat-identities.html

J. Ossenbruggen, M. Hildebrand, V. de Boer. (2011). Interactive vocabulary alignment. In: Proc. 15th International Conference on Theory and Practice of Digital Libraries, Berlin, Germany, 26-28 September, 2011, 296-307.

T. Pellissier Tanon, D. Vrandečić, S. Schaffert, T. Steiner, and L. Pintscher. (2016). From Freebase to Wikidata - The Great Migration. In Proceedings of the 25th International Conference on World Wide Web (WWW 2016), 11-15 April, 2016, Montreal, Canada, 1419-1428.

(n.d.). Solr Query Syntax. In Solr Wiki. Retrieved August 13, 2018 from https://wiki.apache.org/solr/SolrQuerySyntax

S. Speicher, J. Arwe, A. Malhotra (eds.). (2015). Linked Data Platform 1.0. W3C Recommendation. https://www.w3.org/TR/ldp/

M. Sporny, D. Longley, G. Kellogg, M. Lanthaler, N. Lindström. (2014). JSON-LD 1.0, A JSON-based Serialization for Linked Data. W3C Recommendation. https://www.w3.org/TR/json-ld/

https://doi.org/10.23106/dcmi.952138993

P. Szekely, C. Knoblock, J. Slepicka, C. Yin, A. Philpot, A. Singh, D. Kapoor, P. Natarajan, D. Marcu, K. Knight, D. Stallard, S. Karunamoorthy, R. Bojanapalli, S. Minton, B. Amanatullah, T. Hughes, M. Tamayo, D. Flynt, R. Artiss, S. Chang, T. Chen, G. Hiebel, and L. Ferreira. (2015). Using a knowledge graph to combat human trafficking. In: The Semantic Web - ISWC 2015. Lecture Notes in Computer Science, vol 9367. Springer, Cham.

R. Wallis, A. Isaac, V. Charles, and H. Manguinhas. (2017). Recommendations for the application of Schema.org to aggregated Cultural Heritage metadata to increase relevance and visibility to search engines: the case of Europeana. Code4Lib Journal, 37 (April 2017). Available at http://journal.code4lib.org/articles/12330.

https://doi.org/10.23106/dcmi.952138993