The Holocaust Archival Material Knowledge Graph

Herminio García-González[1][0000−0001−5590−4857] and Mike Bryant[2][0000−0003−0765−7390]

[1] Kazerne Dossin, Goswin de Stassartstraat 153, 2800 Mechelen, Belgium
herminio.garciagonzalez@kazernedossin.eu
[2] King's College London, Strand, London WC2R 2LS
michael.bryant@kcl.ac.uk

Abstract. Holocaust research faces specific challenges, among other causes, due to the wide-dispersal of its sources. The EHRI portal, one of the EHRI project main outputs, offers a centralised platform for contextualising and integrating Holocaust-relevant archival material. In this work we explore a conversion of the EHRI portal data to Linked Open Data in order to explore the benefits that this non-centralised model could deliver to the field. We describe the transformation process from the EHRI portal data—following current International Council on Archives (ICA) specifications— to a Knowledge Graph aligned with the new ICA conceptual model, Records in Contexts (RIC). As part of this process we describe the challenges and limitations of this mapping alongside the future developments needed to cope with some of them. We envision this work as the first step in delivering Holocaust data to the Semantic Web, allowing partner institutions to evaluate its capabilities, and potentially implement them for their solutions, making the field more interconnected.

Keywords: Holocaust · Knowledge Graph · Shoah · Data transformation · Linked Open Data · Records in Contexts

1 Introduction

The creation of accessible, coherent and well integrated datasets has been demonstrated to be an important catalyst in enabling researchers to produce innovative and groundbreaking research [19]. In the Humanities, even before consideration is given to the interpretation of sources, their accessibility and complex provenances often present researchers with considerable logistical, organisational and accessibility challenges [22]. In research pertaining to the Holocaust and its historical legacy these challenges are particularly acute. For numerous reasons, including the intentional destruction of evidence [26] and the widespread dislocation of people and administrative bodies following the Second World War (WWII), Holocaust-related material and archival sources are highly fragmented and dispersed. In practice, this means that researchers seeking to access key

Holocaust sources must in many cases navigate a complex trans-national patchwork of archives with different mandates, cataloguing practices, and systems of arrangement.

Overcoming barriers to effective trans-national Holocaust research is one of the principal goals of the European Holocaust Research Infrastructure (EHRI)[3], an EU-funded research project, now in its third 4-year phase and soon to transition into a permanent organisation as a European Research Infrastructure Consortium (ERIC). For over a decade, EHRI has built tools to help researchers understand and navigate the complex landscape of Holocaust research [25], cataloguing sources across many hundreds of institutions and working with numerous archives, large and small, to integrate and contextualise their collections descriptions. A major part of these efforts is the EHRI Portal [5][4], an online database of Holocaust-related archival sources, which enables the integration and interlinking of archival descriptions and their associated metadata from around the world.

The development of the EHRI Portal, its technologies and APIs, along with various initiatives aimed at increasing the interconnectedness of its metadata have been described elsewhere [3]. In this paper we focus on our efforts to expose the rich metadata contained in the EHRI Portal, derived from institutions around the world as well as EHRI's own archival specialists, in a manner compatible with the Semantic Web and capable of better integrating with the emerging network of Linked Open Data (LOD) sources[5]. Semantic Web technologies offer a unique means where entities can be identified unambiguously, linked across databases, and where new data can be automatically inferred [4], attributes which have demonstrated to effectively support Digital Humanities activities [27]. The Knowledge Graph (KG) of Holocaust-related descriptions presented below, based on the EHRI portal data, serves as a first step to increasing the visibility of this kind of material and to facilitate other LOD publishers to link to EHRI's entities.

Producing and publishing LOD is a challenge common to many GLAM institutions [10,1], where datasets of research interest are frequently siloed in legacy databases and intermingled with more closely-held administrative data, not amenable to being made public. As described in [5], the EHRI Portal, while developed under an "open-first" approach, also includes many affordances for restricting the visibility and accessibility of material that is private to individual users, concealed from view for copyright reasons, or otherwise sensitive. We believe that the approaches described in this paper therefore have wide applicability to other practitioners who have an interest in expanding the openness of their data, particularly archival institutions. In addition, many archival institutions present a technological deficit making it very hard for them to adapt to new technologies and migrate old data [31]. This KG, therefore, could serve as an example for Holocaust-related institutions that want to experiment with

[3] `https://www.ehri-project.eu`
[4] `https://portal.ehri-project.eu/`
[5] `https://lod-cloud.net/`

Semantic Web technologies, and their possibilities, without being required to invest in a costly internal procedure. In the future, if more institutions decide to expose their data as LOD, connections could be made back and forth with this KG (as they use to do with the EHRI portal) allowing it to act as an authority hub for Holocaust material, facilitating the links between Holocaust-related institutions (see Section 5.3).

The rest of this paper is structured as follows: Section 2 describes related work; in Section 3 we outline EHRI's data and services and how the transformation was carried out; Section 4 introduces the KG and its main characteristics; in Section 5 we enumerate the challenges that arise from this work and how we intend to solve them in future. Finally, in Section 6 we draw the conclusions obtained from this work.

2 Related Work

Many works have addressed the modelling of historical data as KGs. One widely-cited example is Europeana [21], which offers metadata about different types of cultural heritage material. The level of detail offered by Europeana could, however, be insufficient for many researchers [29] and it does not seek to contextualise subject-specific material as EHRI does. Touching the Second World War as a whole, in [6] the authors investigated a linking algorithm to enrich WWII collections with events information modelled as LOD. Similarly, WarSampo [23] offers a Finnish KG for WWII integrating many different data sources[6] and offering them through a single web interface[7]. This KG models different perspectives such as events, persons, army units, places, etc. However, to the best of our knowledge, no KG has modelled the archival landscape of Holocaust-related material.

Even though no KG has yet taken a holistic view of Holocaust-related archival material, a number of relevant initiatives have appeared in recent years focused on a particular region or country[8]. Others, with a more trans-national perspective that address similar topics (e.g., Jewish material), do inevitably overlap with EHRI's scope, such as the Yerusha platform, which offers a centralised access for Jewish archival heritage[9]. Unfortunately, to date, there is a dearth of linkages between these platforms. This complicates both users' access to the information in navigating many overlapping sources, and the task of the holding institutions in keeping their metadata up-to-date in multiple places. This plethora of siloed alternatives gives traction to an alternative semantic landscape where data could be more interoperable and authority hubs (former aggregators) could act as linking facilitators (see Section 5.3.)

In Cultural Heritage a number of conceptual models, vocabularies and ontologies (some of them related to a conceptual model) have emerged aiming to cover

[6] https://seco.cs.aalto.fi/projects/sotasampo/en/#datasets
[7] https://www.sotasampo.fi/en/
[8] https://www.oorlogsbronnen.nl/
[9] https://yerusha.eu/

different aspects of the field, e.g., CIDOC-CRM [12], PROV-O [24], FRBR[10], NIE-INE[11], ROAR[12] or ARKIVO [28], among others. As relates specifically to archives, a number of attempts have been made to address the mapping from the Encoded Archival Description (EAD) vocabulary to these aforementioned ontologies. For example, converting from EAD to CIDOC-CRM has been addressed, among others, by [7,34,33,15] with different levels of EAD semantic coverage. However, CIDOC-CRM was originally intended for museum objects interoperability with some links to archives or libraries, which in some cases hinders the establishment of metadata equivalents. Moreover, due to these differences in scopes, domain experts will be always more comfortable with a domain-specific model that could then integrate with a larger scope, and that effectively covers and unifies the International Council of Archives (ICA) standards [18]. More recently, a transformation tool from EAD to Records in Contexts Ontology (RiC-O) has been released [14], using XSLT stylesheets as the base for the mapping. As explained later, EHRI expands ICA standards to fit some specific needs making us opt for a domain specific conversion which can be later shared as an EAD to RiC-O mapping for the whole community based on the shared commonalities.

2.1 Prior attempts in the EHRI project

Inside the EHRI project there have been a number of existing cases where semantic and/or RDF technologies were employed, in addition to those mentioned below relating to EHRI's data model. As we have written about previously [5], EHRI uses a graph database (Neo4j) as its underlying data store, and while it functions as a "property graph" rather than a native triplestore, it has some common characteristics. We have on two occasions experimented with automatic mapping from the internal Neo4j schema to a LOD format, one using an interface to the SAIL (Storage and Inference Layer) API[13], and the other using the NeoSemantics (n10s) Neo4j plugin[14]. While both approaches showed promise in some respects, we did not put them into production due to either compatibility issues stemming from tightly-coupled dependencies, or limitations in query performance and scalability resulting from the on-the-fly translation approach.

A more recent undertaking aimed to enrich data already in the portal relating to controlled vocabularies for camps and ghettos, linking them with Wikidata and georeferencing them against GeoNames [2]. Although the goal of this work was not to fully convert EHRI portal data to RDF it established some of the foundations that we build on here. Inside the wider EHRI consortium we also want to highlight the Holocaust Victims Names database[15] implemented by

[10] https://repository.ifla.org/bitstream/123456789/811/2/ifla-functional-requirements-for-bibliographic-records-frbr.pdf
[11] https://github.com/nie-ine/Ontologies
[12] https://leonvanwissen.nl/vocab/roar/docs/
[13] https://rdf4j.org/documentation/reference/sail/
[14] https://neo4j.com/labs/neosemantics/
[15] http://dati.cdec.it/

CDEC [8] in which they developed a Shoah ontology[16] reusing and extending existing ontologies like FOAF[17] and BIO[18] (extended in bio-ext[19]) and model the information about these victims using it. This experience motivated us to offer EHRI portal data as LOD so initiatives such as this from partner institutions could be linked and jointly queryable by users.

3 EHRI's data and transformation

3.1 EHRI's data model

Data in the EHRI portal is based around three main entities: countries, archival institutions, and archival descriptions. Countries constitute an entry point and provide information on the situation of Holocaust research in a specific country. Collection-holding institutions (CHIs)—typically archives or bodies with similar mandates—are grouped within their host country and include relevant contact details along with additional context and information pertaining to their holdings—as described in the International Standard for Describing Institutions with Archival Holdings (ISDIAH)[20]. Archival descriptions are contained within their holding institution and store the information aligned with the General International Standard Archival Description (ISAD(G).)[21] One notable characteristic of archival descriptions is that they can be nested to arbitrary depth to form a hierarchy, modelling the physical arrangement of the described materials (e.g., fonds, series, subseries, collections, etc.)

In addition to these three main entities, the EHRI portal also employs entities for enriching and indexing archival metadata. Authority sets are collections of people, families, or corporate bodies, as described in the International Standard Archival Authority Record for Corporate Bodies, Persons and Families (ISAAR (CPF))[22], whilst a number of controlled vocabularies contain collections of content-specific keywords (i.e., subject headings, and historical places). These descriptive entities are linked from the access points and creators sections of the archival description, serving as a connecting point between collections and allowing a thematic search.

Finally, this structure is augmented by *annotations* and *links*, both modelled as first-class entities that can connect and add additional information to those discussed above. In the current EHRI portal, vocabularies, annotations, and links are the only parts of the data model derived from and partially aligned

[16] `http://dati.cdec.it/lod/shoah/shoah.rdf`
[17] `http://xmlns.com/foaf/0.1/`
[18] `https://vocab.org/bio/`
[19] `http://dati.cdec.it/lod/bio-ext/`
[20] `https://www.ica.org/en/isdiah-international-standard-describing-institutions-archival-holdings`
[21] `https://www.ica.org/en/isadg-general-international-standard-archival-description-second-edition`
[22] `https://www.ica.org/en/isaar-cpf-international-standard-archival-authority-record-corporate-bodies-persons-and-families-2nd`

with RDF, namely the Simple Knowledge Organisation System (SKOS)[23] in the case of vocabularies, and the Web Annotation Model framework for annotations and links. EHRI's use of the relevant standards for linking and indexing metadata records is discussed further in [3].

3.2 Schema alignment

As noted above, EHRI's data is primarily aligned with the conceptual standards from the ICA. As a result, import and export of metadata pertaining to archival descriptions from the EHRI portal was designed around EAD [30], the most well established format derived from ISAD(G)[24]. However, while EAD is widely adopted in the archival field, it inherits the limitations that non-semantic XML technologies present, as discussed in [17], along with other issues stemming from its flexibility as an encoding medium [32].

Seeking to address said limitations, the ICA has been working on a new conceptual model of the archival domain, using a graph as data model. Dubbed Records in Contexts-Conceptual Model (RiC-CM) [20], it is currently on its second draft version, v0.2, released in 2021, and offers a companion ontology for modelling the data in RDF, called RiC-O[25]. As this specification is intended to supersede EAD in the future, we have used it as our base ontology for the transformation of EHRI's data into semantic form.

Using RiC-O 0.2 as a foundation has distinct benefits. It allows us to implement a version of EHRI's data using Records in Contexts (RiC) on top of the existing implementation, letting us test the new data model before the stable version is released. It presents a future common alignment point for other institutions that are currently using ISAD(G) (and/or ISAAR) for data publication and will likewise, in future, seek to make a similar transition, potentially building on EHRI's mapping rules for their own use cases. And it constitutes a zero-cost demonstration for EHRI partner institutions of how RiC works and its potential benefits, without them having to make a substantial investment themselves in mapping or adapting their in-house data sources.

Since not all of our required semantics are covered by the current RiC-O draft, however, it has been necessary for us to extend the ontology in some respects. Following best practice in ontology modelling we have tried to reuse other ontologies or vocabularies as much as possible, using schema.org[26] to complete some fields missing from RIC-O. Schema.org offers a set of classes dedicated to archives since its version 3.5. These classes and their fields complement and align very well to those in RiC-O. For those fields still missing, but necessary from our data perspective, we have included them as properties of a future EHRI ontology (e.g., `https://lod.ehri-project-test.eu/history`).

[23] `https://www.w3.org/2004/02/skos/`
[24] In addition to its counterpart schema for authority information, the Encoded Archival Context (EAC)
[25] `https://www.ica.org/en/records-in-contexts-conceptual-model`
[26] `https://schema.org/`

3.3 Data transformation

In order to construct the KG we opted for a batch transformation of the data using the ShExML language [16] and engine[27]. The transformation was executed for each entity in succession, following the paginated structure of the EHRI APIs in order to permit resumption of the transformation if required. This division was deemed necessary given the amount of data present in the EHRI portal, exceeding 400,000 archival descriptions[28]. The transformation is additionally divided into two main processes: harvesting and transformation.

For the harvesting process we have made use of the existing EHRI API endpoints[29], as a more open and reproducible alternative to requiring privileged access to the internal database. Therefore, we have used the JSON API for harvesting the countries, archival institutions and archival descriptions and the GraphQL API [9] for the links between archival descriptions, and vocabularies and authority sets. For the vocabularies the information is already available as RDF[30] so only the links are downloaded and the already existing RDF triples are incorporated later in the process of building the complete KG. The harvesting script is written in Python and can be consulted on Github[31].

As mentioned above the transformation process is divided per entity and per page, so the transformation occurs in the following steps:

1. Countries and archival institutions: one-step conversion.
2. Archival descriptions: mapping rules are executed against items in a batched manner.
3. Vocabularies (links): mapping rules are executed against each page of the GraphQL API.
4. Authority sets: mapping rules are executed against each page of the GraphQL API.

The execution of these mapping rules produce several Turtle files that are then merged together, using the RDF compositional property, along with the pre-existing SKOS-format vocabularies. All the materials and resources used for this process can be openly consulted on Github as Open Source[32].

4 Dataset

The KG consists of 6,571,095 triples that in Turtle format comprise 767MB of data[33]. We have published this KG using Apache Jena Fuseki as the triple store[34]

[27] `https://github.com/herminiogg/ShExML`
[28] Consulted on 04/04/2023
[29] `https://portal.ehri-project.eu/api`
[30] See for an example: `https://portal.ehri-project.eu/vocabularies/ehri_terms/export?format=TTL`
[31] `https://github.com/herminiogg/EHRI2LOD/blob/main/src/downloader.py`
[32] `https://github.com/herminiogg/EHRI2LOD`
[33] Statistics consulted on 04/04/2023
[34] `https://jena.apache.org/documentation/fuseki2/`

and the LodView viewer[35] in order to allow exploration of the data[36]. The KG also provides a SPARQL endpoint for more complex queries[37].

As mentioned above, we have mainly used RiC-O as the modelling ontology, with some fields missing from RiC-O aligned to schema.org. In these cases, we have double-classed the instances that combine predicates from both specifications, allowing for better discoverability and data completeness. These double typed classes are country (`rico:Place` and `schema:Country`) and archival institution (`rico:CorporateBody` and `schema:ArchiveOrganization`). In the future EHRI ontology this will be made explicit with a dedicated class that inherits from both super classes. At the same time, and following the same principle, we have added the three possible name predicates, i.e., `rdfs:label`, `schema:name` and `rico:name`, allowing for a more standardised access from existing agents.

Inverse relations are always provided where possible as the RiC-O specification suggests, letting users navigate the graph in bidirectional fashion and making the graph more predictable. Examples of this are `rico:hasOrHadHolder`[38] and `rico:isOrWasHolderOf`[39] or `rico:hasInstantiation`[40] and `rico:isInstantationOf`[41].

In order to better interconnect with existing or future KGs and to allow users to expand the data we have comprehensively provided the following links. For countries we have connected each country to its DBpedia instance (e.g., `ehri-country:gb owl:sameAs dbr:United_Kingdom`). In the case of archival institutions we have linked them to the main institution webpage that will potentially show the same information or more information in semantic technologies. In many cases, they do not yet provide semantic information but it is still good to provide the link for the future. In addition, camps and ghettos controlled vocabulary entities (that were already in RDF) provided a link to Wikidata [13] (using `rdfs:seeAlso`) pointing to the equivalent entity [11]. A class diagram can be consulted in Figure 1.

4.1 Post-transformation enrichment

Apart from the triples and links generated from the batch process, there are other kinds of links that can be included per case, and that are out of the scope of the batch transformation as they could require manual verification and update. For now, we perform two post-tranformation enrichments: language links with their counterparts in DBpedia and links of EHRI authorities (persons and corporate bodies) to their counterparts in the CDEC dataset.

In the case of DBpedia, languages are easily linked based on the label similarity against `dbo:Language` instances. For this purpose a federated `CONSTRUCT`

[35] https://github.com/LodLive/LodView
[36] https://lod.ehri-project-test.eu/
[37] https://lod.ehri-project-test.eu/query/
[38] https://www.ica.org/standards/RiC/RiC-O_v0-2.html#hasOrHadHolder
[39] https://www.ica.org/standards/RiC/RiC-O_v0-2.html#isOrWasHolderOf
[40] https://www.ica.org/standards/RiC/RiC-O_v0-2.html#hasInstantiation
[41] https://www.ica.org/standards/RiC/RiC-O_v0-2.html#isInstantiationOf

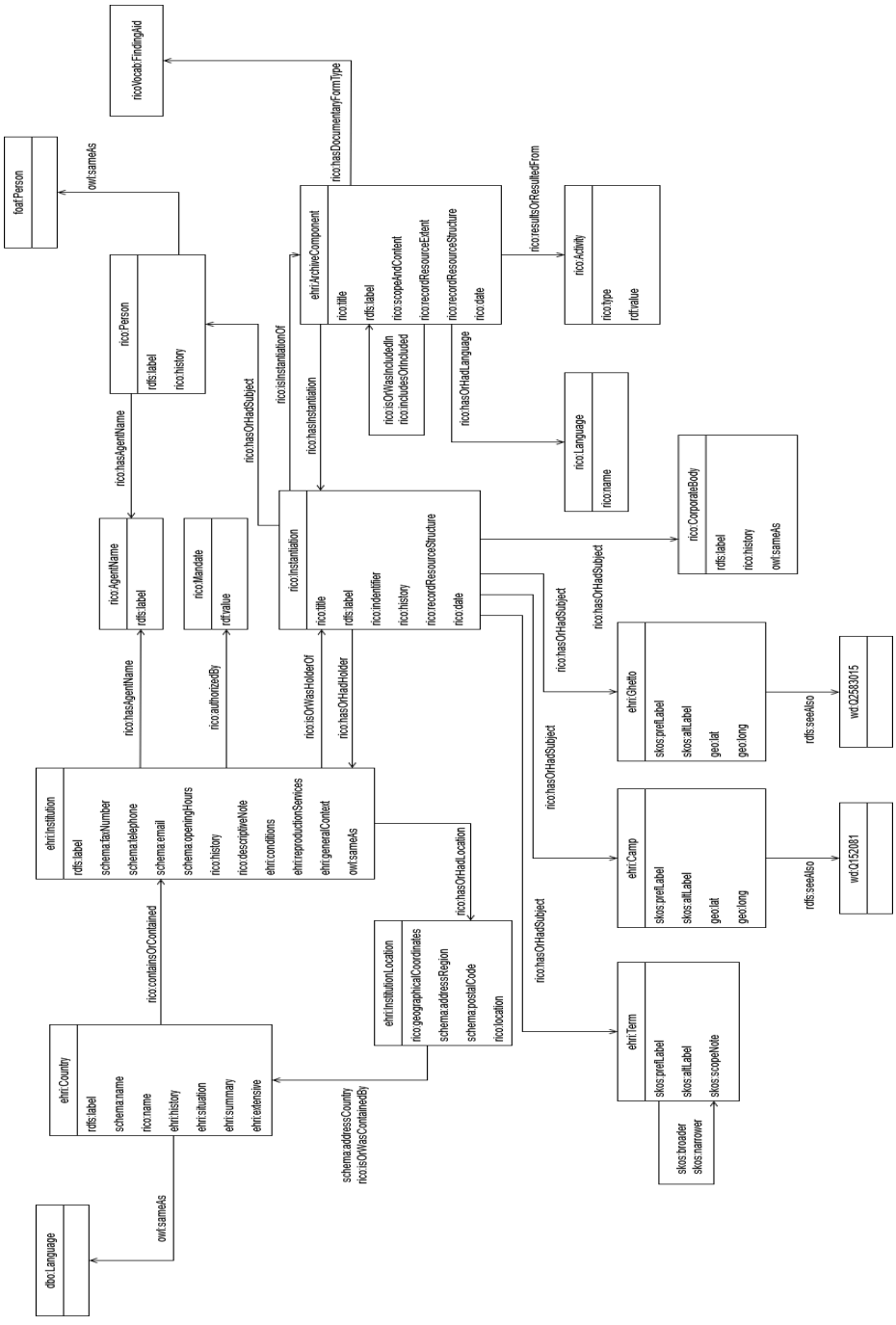

Fig. 1. Class diagram representing the data model followed in the conversion to RDF using RiC-O and schema.org as the base ontologies.

SPARQL query is run on the resulting KG[42] and the results are supervised by content experts. For CDEC person database links we run another federated query that, similar to that used with DBPedia, establishes the links between EHRI and CDEC authority files[43]. These triples are verified by CDEC staff and then are retained for future use, such that only previously unseen relations are required to be validated. Both generated links datasets are uploaded to the main triple store and added to the KG. These post-transformation enrichments allow to execute SPARQL Federated queries over multiple KGs letting users answer more complex questions like the example given below in Listing 1. More examples can be found in the EHRI KG landing page[44].

5 Challenges and Future Work

5.1 Mapping copies and originals

Even though the majority of the data in the EHRI portal is mapped using the techniques described in this paper there are still some aspects where the available ontologies do not provide us with satisfactory solutions. In other cases, solutions will require further consensus from the community.

One significant challenge pertaining to Holocaust-related material is the amount of copying of collection material that has been carried out by different archives around the world, who have proceeded to describe the same underlying material using their own specific in-house style. From the very beginning, the EHRI portal has had, as one of its main goals, the recontextualisation of the Holocaust sources. In this sense in the EHRI-2 phase a copy linking system was introduced, allowing to interlink copies to their originals [3]. Researchers can now have a clearer view of these different versions of the original material[45].

The EHRI portal supports four types of links: 1) copy archival unit to original archival unit (the archival unit was copied from this specific original archival unit); 2) copy archival institution to original archival institution (the institution holds copies from another institution without specifying which); 3) copy archival unit to original archival institution (the archival unit was copied from the mentioned archival institution, without knowing from which exact collection it was copied); and 4) copy archival institution to original archival unit (the archival institution holds copies of this original archival unit without knowing which copied archival unit holds the copies.) All links can be interpreted bidirectionally, for example, X archival unit was copied from Y original archival unit or Y original archival unit was copied into X archival unit.

[42] `https://github.com/herminiogg/EHRI2LOD/blob/main/src/auxFiles/linksLanguagesDBpedia/linksLanguagesDBpediaFederatedQuery.rq`

[43] `https://github.com/herminiogg/EHRI2LOD/blob/main/src/auxFiles/linksToCDEC/queryToMatchToCDEC.rq`

[44] `https://lod.ehri-project-test.eu/`

[45] See for an example: `https://portal.ehri-project.eu/units/us-005578-irn524242`

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX fo: <http://www.w3.org/1999/XSL/Format#>
PREFIX db: <http://dbpedia.org/>
PREFIX dbp: <http://dbpedia.org/property/>
PREFIX rico: <https://www.ica.org/standards/RiC/ontology#>
PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX schema: <http://schema.org/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX ehri:  <http://lod.ehri-project-test.eu/>
PREFIX ehri_institutions: <http://lod.ehri-project-test.eu/institutions/>
PREFIX ehri_units: <http://lod.ehri-project-test.eu/units/>
PREFIX ehri_countries: <http://lod.ehri-project-test.eu/countries/>
PREFIX bio-ext: <http://dati.cdec.it/lod/bio-ext/>
PREFIX shoah: <http://dati.cdec.it/lod/shoah/>


SELECT ?record WHERE {
    ehri_institutions:it-002845 rico:isOrWasHolderOf ?instantiation .
    ?instantiation rico:isInstantiationOf ?record .
    ?record rico:hasOrHadSubject ?personEHRI .
    ?personEHRI owl:sameAs ?person .
    SERVICE <http://lod.xdams.org/sparql> {
        ?person a foaf:Person ;
                shoah:persecution ?persecution .
        ?persecution shoah:toNaziCamp ?camp .
        ?camp rdfs:label "Auschwitz" .
    }
}
```

Listing 1: SPARQL Federated query to get the records referring to people deported to Auschwitz.

Looking into the current RiC-O draft, the properties `rico:hasCopy`, `rico:isCopyOf`, `rico:hasOriginal` and `rico:isOriginalOf` seem to cover the same semantics explained above. However, if we look at the domain and range of these properties we see that they are bound to `rico:RecordResource` meaning that the relation can only be established between two entities of this type or its descendants. Ultimately, this translates to being able to only map the first type of link out of the four supported links in the EHRI portal. A future envisioned solution will be to introduce these custom properties used in the EHRI portal as new properties of the planned EHRI ontology.

In addition, the RiC-CM puts the emphasis on the distinction between a Record Resource and an Instantiation, the latter being the representation of the record in a digital or physical form. In this sense, we can see copies as different instantiations of the same record where, for example, the original may be a deed and the copy a microfilm. But in essence both refer to the same original material. Looking at the already mapped data, however, this presents an issue, as archival units (Record Resources in RiC-CM) are assumed to be held by only one institution in the EHRI portal, with identifiers derived from this hierarchy. In order to maintain this information, therefore, we are compelled to continue creating only one instantiation per Record Resource and make the links between them.

One alternative would be to use the `owl:sameAs` property to indicate that in fact the resource is the same. Unfortunately, this creates some additional verbosity in our mapped data, potentially affecting new users navigating the data as it hinders the clarity of the graph. While it does not restrain the use of the ontology for our mapped data, it is true that clarifying the semantics for these cases when using RiC-O will benefit data producers and consumers as exposed with this case. Thus, we will follow the development of the proposed RiC-O closely to adapt our batch conversion if this point becomes clearer in future versions.

5.2 Handling updates

As mentioned in Section 3.3, we opted for a batch approach for the conversion of the data. This means that at some point data could be added, updated or deleted in the EHRI portal making parts of the KG obsolete or incomplete. In order to cope with this issue many strategies could be taken. One possible approach would be to execute the batch process as a nightly task and exchange the old KG for the newly generated one. However, given the size of the dataset this process would be time consuming and fairly inefficient. We have therefore designed a workflow that, while based on the batch approach, incorporates only updates that took place since the previous harvesting operation without impacting the overall performance. Our envisioned solution is to process change events from the EHRI portal as a stream that are incorporated into an append-only historic log where all changes since the creation of the KG can be tracked. This would facilitate not only processing the changes as they arrive, but also to reconstruct update events in cases where it is needed to replicate them for further

migrations or installations, or recover from down time. From each of these events it is possible to download the new contents from the harvesting source (JSON API or GraphQL API) and, depending on the type of event (creation, deletion, update), run the necessary `INSERT` and/or `DELETE` SPARQL queries against the SPARQL endpoint. This workflow can be seen in Figure 2. We will undertake the implementation of the proposed data update architecture as future work in order to more efficiently keep the KG up-to-date in a timely manner.

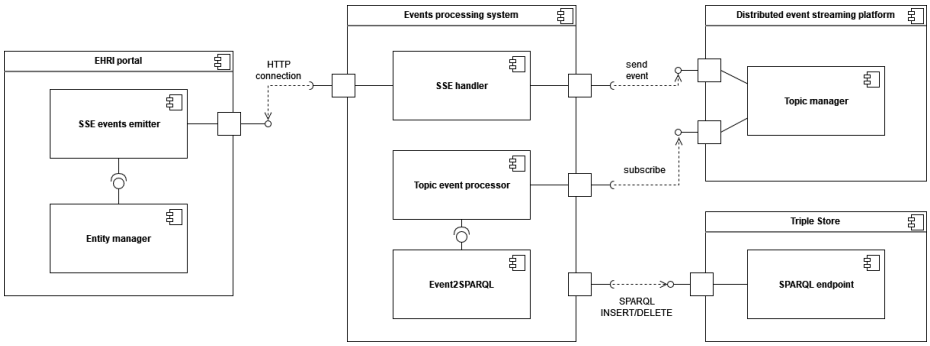

Fig. 2. Architecture of the envisioned stream-based incremental update system to align the KB with the EHRI portal's live dataset.

5.3 EHRI KG as an authority hub

The EHRI portal acts as an aggregator for information about Holocaust documentation, allowing users to seamlessly access metadata about collections, and institutions to contextualise their own records in a larger, trans-national landscape. This enhanced contextualisation happens in the portal, where EHRI portal users can benefit from it, but the metadata providers themselves are not able to easily reflect it in their own data. In this sense, this centralised approach presents challenges when it comes to improving agents' access and data reusability (in this case, by the own contributing institutions.)

Federated approaches, however, pose other challenges like how to actively manage the links between different nodes, how to manage widely-used persistent and unique identifiers, and how to foster node discoverability. These challenges, however, can be mitigated by an aggregator which can promote visibility, link an institution's data to other data in the network, and is able to manage coherent and consistent identifiers across the network. Therefore, we should take those advantages and put them to work in a federated manner. In this regard, we see this KG as a first step on establishing an authority hub (as opposed to an aggregator) where the source of truth lies on institutions' own data. This allows for a more lightweight KG where only general metadata information about collections is served, with the rest being available on-demand (via semantic web

technologies) based on users' requirements. The authority hub, then, would have the responsibility to maintain the links among different providers' items, allowing institutions to search across the network throughout the hub or even re-utilising the data for their own systems.

Moreover, as more institutions start to work following these principles, fewer and fewer data integration procedures will be required, avoiding the costly and not always fruitful endeavour to keep aggregated metadata sufficiently up-to-date.

6 Conclusions

Given that the RiC conceptual model has not yet reached its first non-draft release, the work described here is also evolving. We have described in Section 3 the general shape of EHRI's data and how we have approached schema alignment, and where it has been necessary to extend or work around limitations with the ontology. Likewise we have described how the transformation is implemented, using EHRI's existing APIs and the ShExML mapping language. The resulting dataset, described in Section 4, is further enriched with connections to more general KBs, such as DBpedia, or others within the same domain, such as CDEC's person database. In Section 5 we described a number of planned advancements to the EHRI KB, including the incorporation of more information about the provenance of Holocaust sources.

The vision described above in Section 5.3—of a distributed LOD environment where each custodian of Holocaust-related material can publish its own metadata, integrating with a common set of vocabularies and authorities that are curated by domain-specific entities like EHRI, or more general ones like Wikidata—is appealing for many reasons. Researchers can benefit enormously from efforts to bring coherence and a deeper level of contextualisation to domains like Holocaust research which are, as discussed in the introduction to this paper, fraught with historical and organisational complexity. Centralised approaches to data integration, whilst necessary with today's level of LOD adoption in the archival domain, are complex to administer and invariably compromised in how up-to-date and comprehensive they can manage to be.

By expanding EHRI's LOD capabilities, building on efforts by the creators of RiC and other such systems, we can hope to foster a greater degree of knowledge interoperability in the domain of Holocaust research. If more data providers can justify the necessary technical investments to eventually publish their own linked datasets, perhaps using the techniques described here as a blueprint with which to do so, this will correspondingly benefit EHRI's goals in contextualising Holocaust sources and bringing greater clarity to the domain.

Supplemental Material Availability: The presented Knowledge Graph and the accompanying documentation are available for consultation on: `https://lod.ehri-project-test.eu/`. The source code for the conversion can be openly consulted on: `https://github.com/herminiogg/EHRI2LOD`.

Acknowledgements

This work has been carried out in the context of the EHRI-3 project funded by the European Commission under the call H2020-INFRAIA-2018-2020, with grant agreement ID 871111 and DOI 10.3030/871111.

References

1. Alexiev, V.: Museum linked open data: Ontologies, datasets, projects. Digital Presentation and Preservation of Cultural and Scientific Heritage (VIII), 19–50 (2018)
2. Alexiev, V., Nikolova, I., Hateva, N.: Semantic Archive Integration for Holocaust Research. The EHRI Research Infrastructure. Umanistica Digitale (4) (2019). https://doi.org/10.6092/issn.2532-8816/9049
3. Arie Erez, S., Blanke, T., Bryant, M., Speck, R., Rodriguez, K., Vanden Daelen, V.: Record linking in the EHRI portal. Records Management Journal **30**(3), 363–378 (2020). https://doi.org/10.1108/RMJ-08-2019-0045
4. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. Scientific american **284**(5), 34–43 (2001)
5. Blanke, T., Bryant, M., Frankl, M., Kristel, C., Speck, R., Daelen, V.V., Horik, R.V.: The European Holocaust Research Infrastructure Portal. Journal on Computing and Cultural Heritage (JOCCH) **10**(1), 1–18 (2017). https://doi.org/10.1145/3004457
6. Both, J., de Hooge, D., IJff, R., Inel, O., de Boer, V., Aroyo, L.: Linking Dutch World War II Cultural Heritage Collections with Events Extracted by Machines and Crowds. In: Joint Proceedings of SEMANTiCS 2017 Workshops co-located with the 13th International Conference on Semantic Systems (SEMANTiCS 2017), Amsterdam, Netherlands, September 11 and 14, 2017. CEUR-WS (2017), `http://ceur-ws.org/Vol-2063/events-paper3.pdf`
7. Bountouri, L., Gergatsoulis, M.: The semantic mapping of archival metadata to the CIDOC CRM ontology. Journal of Archival Organization **9**(3-4), 174–207 (2011). https://doi.org/10.1080/15332748.2011.650124
8. Brazzo, L., Mazzini, S.: From the Holocaust Victims Names to the Description of the Persecution of the European Jews in Nazi Years: the Linked Data Approch and a New Domain Ontology. Book of abstract of DH (2015)
9. Bryant, M.: GraphQL for archival metadata: An overview of the EHRI GraphQL API. In: 2017 IEEE International Conference on Big Data (Big Data). pp. 2225–2230. IEEE (2017). https://doi.org/10.1109/BigData.2017.8258173
10. Candela, G., Sáez, M.D., Escobar Esteban, M., Marco-Such, M.: Reusing digital collections from GLAM institutions. Journal of Information Science **48**(2), 251–267 (2022). https://doi.org/10.1177/0165551520950246
11. Cooey, N.: Leveraging Wikidata to enhance authority records in the EHRI portal. Journal of Library Metadata **19**(1-2), 83–98 (2019). https://doi.org/10.1080/19386389.2019.1589700
12. Doerr, M.: The CIDOC conceptual reference module: an ontological approach to semantic interoperability of metadata. AI magazine **24**(3), 75–75 (2003). https://doi.org/10.1609/aimag.v24i3.1720
13. Erxleben, F., Günther, M., Krötzsch, M., Mendez, J., Vrandečić, D.: Introducing Wikidata to the Linked Data Web. In: The Semantic Web–ISWC 2014: 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014.

Proceedings, Part I 13. pp. 50–65. Springer (2014). https://doi.org/10.1007/978-3-319-11964-9_4
14. Francart, T., Clavaud, F., Charbonnier, P.: RiC-O Converter: a Software to Convert EAC-CPF and EAD 2002 XML Files to RDF Datasets Conforming to Records in Contexts Ontology. In: Proceedings of Linked Archives International Workshop 2021 co-located with 25th International Conference on Theory and Practice of Digital Libraries (TPDL 2021). pp. 30–36 (2021), https://ceur-ws.org/Vol-3019/LinkedArchives_2021_paper_14.pdf
15. Gaitanou, P., Bountouri, L., Gergatsoulis, M.: Automatic generation of crosswalks through CIDOC CRM. In: Metadata and Semantics Research: 6th Research Conference, MTSR 2012, Cádiz, Spain, November 28-30, 2012. Proceedings 6. pp. 264–275. Springer (2012). https://doi.org/10.1007/978-3-642-35233-1_26
16. García-González, H., Boneva, I., Staworko, S., Labra-Gayo, J.E., Lovelle, J.M.C.: ShExML: improving the usability of heterogeneous data mapping languages for first-time users. PeerJ Computer Science **6**, e318 (2020). https://doi.org/10.7717/peerj-cs.318
17. Gartner, R.: An XML schema for enhancing the semantic interoperability of archival description. Archival Science **15**(3), 295–313 (2015). https://doi.org/10.1007/s10502-014-9225-1
18. Gueguen, G., da Fonseca, V., Pitti, D., Grimoüard, C.: Toward an international conceptual model for archival description: a preliminary report from the International Council on Archives' Experts Group on archival description. The American Archivist **76**(2), 567–584 (2013). https://doi.org/10.17723/aarc.76.2.p071x02401282qx2
19. Hyvönen, E.: Using the Semantic Web in digital humanities: Shift from data publishing to data-analysis and serendipitous knowledge discovery. Semantic Web **11**(1), 187–193 (2020). https://doi.org/10.3233/SW-190386
20. International Council on Archives (ICA): Records in Contexts-Conceptual model (RiC-CM) 0.2 (2021), https://www.ica.org/sites/default/files/ric-cm-02_july2021_0.pdf, accessed 03/04/2023
21. Isaac, A., Haslhofer, B.: Europeana Linked Open Data – data.europeana.eu. Semantic Web **4**(3), 291–297 (2013). https://doi.org/10.3233/SW-120092
22. Khan, N.A., Shafi, S., Ahangar, H.: Digitization of cultural heritage: Global initiatives, opportunities and challenges. Journal of Cases on Information Technology (JCIT) **20**(4), 1–16 (2018). https://doi.org/10.4018/JCIT.2018100101
23. Koho, M., Ikkala, E., Leskinen, P., Tamper, M., Tuominen, J., Hyvönen, E.: WarSampo knowledge graph: Finland in the Second World War as Linked Open Data. Semantic Web **12**(2), 265–278 (2021). https://doi.org/10.3233/SW-200392
24. Lebo, T., Sahoo, S., McGuinness, D., Belhajjame, K., Cheney, J., Corsar, D., Garijo, D., Soiland-Reyes, S., Zednik, S., Zhao, J.: Prov-o: The prov ontology (2013), https://www.w3.org/TR/prov-o/
25. de Leeuw, D., Bryant, M., Frankl, M., Nikolova, I., Alexiev, V.: Digital Methods in Holocaust Studies: The European Holocaust Research Infrastructure. In: 2018 IEEE 14th International Conference on e-Science (e-Science). pp. 58–66. IEEE (2018). https://doi.org/10.1109/eScience.2018.00021
26. Malka, T.D.: Missing persons and world war ii: Between personal and national loss. War in History **29**(3), 641–663 (2022). https://doi.org/10.1177/09683445211038600
27. Meroño-Peñuela, A., Ashkpour, A., Van Erp, M., Mandemakers, K., Breure, L., Scharnhorst, A., Schlobach, S., Van Harmelen, F.: Semantic technolo-

gies for historical research: A survey. Semantic Web **6**(6), 539–564 (2015). https://doi.org/10.3233/SW-140158
28. Pandolfo, L., Pulina, L., Zielinski, M.: Towards an Ontology for Describing Archival Resources. In: Proceedings of the Second Workshop on Humanities in the Semantic Web (WHiSe II) co-located with 16th International Semantic Web Conference (ISWC 2017). pp. 111–116 (2017), https://ceur-ws.org/Vol-2014/paper-12.pdf
29. Peroni, S., Tomasi, F., Vitali, F.: Reflecting on the europeana data model. In: Digital Libraries and Archives: 8th Italian Research Conference, IRCDL 2012, Bari, Italy, February 9-10, 2012, Revised Selected Papers 8. pp. 228–240. Springer (2013). https://doi.org/10.1007/978-3-642-35834-0_23
30. Pitti, D.V.: Encoded archival description: An introduction and overview **5**, 61–69 (1999). https://doi.org/10.1080/13614579909516936
31. Ruest, N., Lin, J., Milligan, I., Fritz, S.: The archives unleashed project: technology, process, and community to improve scholarly access to web archives. In: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020. pp. 157–166 (2020). https://doi.org/10.1145/3383583.3398513
32. Shaw, E.J.: Rethinking EAD: balancing flexibility and interoperability. New Review of Information Networking **7**(1), 117–131 (2001). https://doi.org/10.1080/13614570109516972
33. Stasinopoulou, T., Bountouri, L., Kakali, C., Lourdi, I., Papatheodorou, C., Doerr, M., Gergatsoulis, M.: Ontology-Based Metadata Integration in the Cultural Heritage Domain. In: Asian Digital Libraries. Looking Back 10 Years and Forging New Frontiers: 10th International Conference on Asian Digital Libraries, ICADL 2007, Hanoi, Vietnam, December 10-13, 2007. Proceedings 10. pp. 165–175. Springer (2007). https://doi.org/10.1007/978-3-540-77094-7_25
34. Theodoridou, M., Doerr, M.: Mapping of the encoded archival description DTD element set to the CIDOC CRM (2001), https://cidoc-crm.org/sites/default/files/ead.pdf