OPEN ACCESS

KING'S College LONDON

King's Research Portal

DOI:
10.1108/RMJ-08-2019-0045

Document Version
Peer reviewed version

Link to publication record in King's Research Portal

Citation for published version (APA):
Arie-Erez, S., Blanke, T., Bryant, M., Rodriguez, K., Speck, R., & Vanden Daelen, V. (2020). Record Linking in the EHRI Portal. *Records Management Journal*, *30*(3), 363-378. https://doi.org/10.1108/RMJ-08-2019-0045

Citing this paper
Please note that where the full-text provided on King's Research Portal is the Author Accepted Manuscript or Post-Print version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version for pagination, volume/issue, and date of publication details. And where the final published version is provided on the Research Portal, if citing you are again advised to check the publisher's website for any subsequent corrections.

General rights
Copyright and moral rights for the publications made accessible in the Research Portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognize and abide by the legal requirements associated with these rights.

•Users may download and print one copy of any publication from the Research Portal for the purpose of private study or research.
•You may not further distribute the material or use it for any profit-making activity or commercial gain
•You may freely distribute the URL identifying the publication in the Research Portal

Take down policy
If you believe that this document breaches copyright please contact librarypure@kcl.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 08. Jan. 2025

Record Linking in the EHRI Portal

Sigal Arie Erez, Tobias Blanke, Mike Bryant, Kepa Rodriguez, Reto Speck, Veerle Vanden Daelen

Introduction

The creation of Linked Open Data (LOD) is a disruptive process within the knowledge industries, enabling new and innovative ways of working with large, dispersed datasets and yielding fresh insights into the structure of complex intellectual domains. Within the archival space, and specifically the area of Holocaust-related documentation, the European Holocaust Research Infrastructure (EHRI) project is invested in realising the vision of a "virtual observatory" that can, through leveraging interconnections between diverse trans-national sources, provide a more comprehensible picture of an archival landscape that is notoriously complex and difficult to navigate.

After describing why archival sources relating to the Holocaust tend to have an uncommonly complex history and reviewing related work on similar themes, this case study describes efforts within the EHRI project to link and integrate archival metadata that is available within EHRI's online portal. We cover our methods and progress in two specific areas: establishing and making navigable the connections between descriptions where material is related by archival provenance; and using co-referencing of subject and authority terms to improve users' ability to transparently browse descriptions sourced from many different institutions. In both cases we provide visualisations that seek to convey an overview of the degree of data integration yet achieved, and how much work still remains. Finally, we discuss how these efforts relate to the ongoing development of LOD-focused archival standards, and the accessibility of EHRI's data from a LOD perspective.

Holocaust Sources & Archival Fragmentation

The overall mission of the European Holocaust Research Infrastructure (EHRI) project is to build an infrastructure that advances trans-national and collaborative approaches to Holocaust research. Active since 2010 and with funding provided by the European Union, EHRI's activities are currently undertaken by a consortium of 24 partner institutions – Holocaust archives, libraries, museums, memorial sites and research institutions – located in 17 countries across Europe, Israel and the United States. A central component of EHRI's mission is to virtually integrate and interlink physically fragmented and dispersed archival collection descriptions relating to the Holocaust in the EHRI Online Portal (https://portal.ehri-project.eu).[1]

In order to understand the centrality of virtual integration and interlinkage for EHRI, a few words need to be said about current Holocaust archival landscape. The most prominent characteristic of this landscape is its dispersed and fragmented nature, brought about by historical conjectures. According to Grimsted, "[t]he Second World War – with the Nationalist-Socialist regime and accompanying Holocaust – wrought the greatest archival destruction and dislocation in history" (Grimsted, 2017). Archival

[1] For further background information on the EHRI project, see Speck et al (2014, pp. 157–177) and Blanke et al (2017).

Records Management Journal

collections shedding light on the Holocaust were particularly affected by such processes of destruction and dislocation for several reasons, including attempts by the perpetrators to destroy evidence about the crime, post-war refugees taking documentation to their new abodes, seizure of archival documents by occupying forces, post-war historical and juridical commissions assembling documentation thereby pulling it out of its original context, etc (Speck et al., 2014, pp. 157–158).

While EHRI's own identification work provides a global view of the dispersal of Holocaust archives, the activities of the Einsatzstab Reichsleiter Rosenberg (ERR) project illustrate the same problem in more depth. The ERR project attempted to survey all archival evidence pertaining to the activities of one Nazi agency – the Einsatzstab Reichsleiter Rosenberg (ERR) – which was engaged in widespread looting of cultural property in Nazi-occupied territories. As a consequence of processes of dispersal and fragmentation both during the war and its aftermath, the survey found documentation generated by the ERR itself and records by post-war agencies seeking to return the ERR loot to its legitimate owners in no less than 29 repositories located in 9 countries (Grimsted, 2012).

The dispersal and fragmentation of relevant archival material is the arguably biggest stumbling block to endeavours to study the Holocaust from truly trans-national perspectives. The material a historian needs to process for such a project is vast in size, complex in nature, and, due to its dispersed and fragmented nature, challenging, and often impossible, to locate and access. EHRI has attempted to alleviate this situation by following a two-pronged approach: first we have attempted to identify archives that hold Holocaust-related sources and to integrate information about such institution and archival descriptions of their material in the EHRI Portal. However, while surveying and virtual integration of information is clearly an important first step, it does not in itself address the challenge adequately. What is additionally required are lateral inter-linkages that virtually tie materials together that are related by either provenance or pertinence but physically dispersed. Only thus can research users of the EHRI Portal locate and contextually interpret all the sources that they need to tackle a particular (trans-national) research question.

An additional factor adding further complexity to the Holocaust archival landscape is the existence of significant copy archives. In the aftermath of World War II, institutions dedicated to collecting Holocaust documentation were established in several countries. As part of their collection missions, institutions such as Yad Vashem in Jerusalem, the United States Holocaust Memorial Museum in Washington DC or the Mémorial de la Shoah in Paris have endeavoured to obtain full or partial copies of Holocaust materials from archives across the globe, and integrated them into their own holdings. This has resulted in the situation that many, and often central, documents are today available to researchers in copied form in several repositories, without, however, any clear indications about the location and context of the originals, rendering their interpretation hazardous. The copy-holding institutions do not necessarily organize the copies in the same way as the originals, and reorganisations within the original-holding archival institution or the original collection are typically not reflected in the descriptions of the copy-holders.

Records Management Journal

As a consequence, Holocaust-relevant documentation can today be found in a many, notably diverse collection-holding institutions, spread across a very large geographic area, and is often located in very surprising locations. The scope of the fragmentation and dispersal challenge is illustrated by EHRI's own identification and integration work to date. Since we started work in 2010, we have identified and described more than 2,100 institutions that hold Holocaust material located in 59 countries, and we have so far integrated more than 350,000 descriptions of Holocaust-related archival units held by 756 institutions.

While EHRI's identification and data integration work is ongoing and far from complete, we have recently started to focus on establishing methodologies for uncovering and expressing such interlinkages within our integrated data store, and to develop and test new methods for visualising the results. The following sections will detail our work in this regard. Here it suffices to note a few overriding challenges we had to address:

1. Move beyond *respect des fonds* as traditionally understood: one of the foundational principles of archival science is "respect des fonds", which mandates that archival records must be managed and organised according to provenance, that is to say the entity by which they were produced.[2] However, the widespread fragmentation of Holocaust sources makes it impossible to apply a unitary, provenance-based view on the Portal's integrated data content. Consider, for instance, the situation of the documentation relating to the Einsatzstab Reichleiter Rosenberg, mentioned above. The ERR fonds today is split across no less than 29 repositories, and it is likely that each repository has organised its ERR holdings in a different fashion. When all this information is integrated into an aggregation such as the EHRI Portal, multiple and overlapping relationships and views on the integrated data content need to be supported: some of these may be provenance based, while others may not.

2. Incomplete data: while the EHRI Portal contains an unparalleled amount of information about dispersed and fragmented Holocaust sources, it is far from complete. While our identification and description of relevant institutions is close to comprehensive in some countries, for instance Austria, Belgium, the Netherlands and Poland, it remains patchy in many others. With regard to descriptions of archival materials, much remains to be done. To date, we have integrated archival descriptions for approximately 1/3 of all identified institutions. Such limitations in terms of coverage of course also limit our ability to express and visualise inter-relationships between fragmented Holocaust collections (Vanden Daelen et al., 2019).

3. Heterogeneity: existing descriptions of Holocaust-related archival materials are marked by very significant heterogeneity. Heterogeneity manifests itself in different ways, most notably in terms of languages – the EHRI Portal incorporates descriptions expressed in 23 different tongues – but also in terms of structure and content. In fact, adoption of international standards such as Encoded Archival Description (EAD) remains relatively low in our domain, and even in cases where descriptive standards

[2] While there is a rich theoretical literature exploring the shortcomings of *respect des fonds* as an organisational principle, especially for born-digital records — see, e.g. Bearman and Lytle (1985) — it remains the standard approach for paper-based records, at least in EHRI's domain.

Records Management Journal

are followed, their interpretation and implementation varies widely between different institutions (Speck and Links, 2016). Such heterogeneity adds another layer of complexity to any attempt to interlink and visualise integrated Holocaust collections.

Related Work

Many related initiatives have also sought to build integrated portals of archival finding aids from many institutions. In Europe, ArchivesHub, AIM25, and Archives Portal Europe (APE) and, in the United States, ArchiveGrid, all follow this general idea, albeit with a primarily geographic rather than subject-based focus. A survey of these sites shows that the primary axis on which archival descriptions from different institutions are connected is access points (e.g. subject terms, creators, and referenced places), yet only AIM25, with its relatively small and geographically constrained set of institutions (all within the London area) was able to index their descriptions and provide the means to browse material from many sources using a common controlled vocabulary (Cosgrave, 2003). Elsewhere, we find that a user's experience of connectedness between diverse material comes only from the use of free-text search and faceted browsing, which we discuss below in relation to EHRI's case.

One aspect many of these integrated portals — and EHRI's — have in common is the use of EAD — and its underlying ISAD(G) conceptual model — as the main data transport encoding (Bredenberg and Jagodzinski, 2014; Bron et al., 2013; Hill, 2002). The strengths and limitations of EAD (and it's related schema, the *Encoded Archival Context for Corporate Bodies, Persons, and Families* — EAC-CPF) in the Linked Open Data (LOD) context have been well discussed elsewhere, including Elizabeth Shaw (2001), Jennifer Bunn (2013), and Richard Gartner (2015), each of whom note the inherent tensions between the document-centric XML schema and the atomistic database-like structure of RDF. We see below, in discussing references between archival entities, one way in which this limitation manifests itself in practice. While the release of the ICA's Expert Group on Archival Description (EGAD) initial draft of the Records in Context Conceptual Model (RiC-CM) and accompanying ontology (RiC-O) — intended to provide a standard for Semantic Web-friendly archival description (EGAD, 2016) — arrived too late to take into consideration for most of the work described below, we have noted in the discussion below how it fits with the copy-original linking cases covered here.

Background - Different Linking Approaches

From the user's perspective, the EHRI Portal has a hierarchical structure, with the topmost level being individual countries, for which the project has prepared extensive documentation in textual form to serve as a high-level guide to the historical context and general archival situation as it applies to Holocaust-related material. From the country level, users of the portal can browse collection holding institutions (subsequently, CHIs) and from there, descriptions of archival collections, also structured hierarchically.

While this hierarchical structure itself does, in a literal sense, consist of connections between archival descriptions from different source institutions, it is also an attempt to reflect the physical organisation of the original material and thus not included in this paper as a linking activity (elsewhere we describe

Records Management Journal

EHRI's technical approach to maintaining archival hierarchies, see: (Bryant et al., 2018).) For the same reasons, we also do not cover the portal's textual search functionality, although the ability to search material from multiple sources is perhaps the primary way that archival collections catalogued by EHRI are (implicitly) connected.

Instead, we focus below on the process of manifesting latent connections relating to intellectual and physical provenance and the integration of archival taxonomies that together provide ways for users to explore material from different sources and better understand its historical and archival context.

Virtual Collections

Virtual collections within the EHRI Portal provide the means to group archival descriptions together in a manner distinct from the hierarchy that reflects physical organisation within their respective holding institutions. Moreover, this synthetic organisational structure is itself hierarchical and can thus imitate an artificial (or "virtual") *fonds*, to use the archival terminology. A full description of EHRI's virtual collection work is described in Bryant et al (2015).

Virtual collections as originally envisioned had two main purposes within EHRI. The first was to allow the creation of virtual finding aids that contained descriptions drawn from multiple institutions, allowing fragmented archival sources to be presented in a more coherent and user-friendly manner. The second was to facilitate the creation of research guides, bringing topically-related material, along with additional descriptive aids, together in a manner resembling — and functionally compatible with — a standard hierarchical collection description. This latter role was also envisioned as a component of the EHRI Portal's virtual research environment (VRE) and an activity that individual users could partake in themselves to create private — or publicly shared — virtual fonds, not dissimilar to a bookmark management system with multiple levels of nesting.

The primary way in which virtual collections contribute to EHRI's linking activities from the perspective of a user of the portal is by placing archival descriptions within the same context for both browsing and searching activities. When a user browses the EHRI Portal by following the artificial hierarchy of the virtual collection they remain within the same "virtual space" unless specifically opting to view a particular description in its original context. Descriptions within a particular VC could be seen as linked in multiple distinct contexts.

The EHRI Portal currently contains a number of virtual collections that were central to integration activities in the project's first phase (2010-2014): two are structured as research guides, bringing together topical material relating to the Terezin Ghetto and the archive of the Jewish Community in Vienna. Three additional VCs are dispersed fonds, providing more intuitive ways to browse material that, for historical reasons, does not exist in the same physical location.

A number of factors, however, have deterred us from investing scarce technical resources to expand the use of virtual collections in the project's second phase, including — beyond an internal trial — incorporating the functionality in the portal's VRE. One factor is the lack of dedicated expertise from

Records Management Journal

subject matter experts that we could devote to the creation of VCs in light of the project's many other objectives. Another was the relatively low engagement with other VRE functionalities by users of the EHRI Portal, and the reduced emphasis on the VRE in general, making the public availability of this more complex and elaborate end-user system less of a priority.

As a result, while virtual collections are still in use in the EHRI Portal, future plans involve consolidation and streamlining rather than expanding their use as a linking mechanism. Instead, we have focused on two alternative mechanisms of linking dispersed material: materialisation of original-copy relationships and integration of access points. The next section introduces the general mechanism of linking archival descriptions in the EHRI Portal on which these two specialisations are built.

Links within the EHRI Portal

Links in the EHRI Portal are implemented using the Open Annotation data model (Sanderson et al., 2013), with each of the connected entities being a "target" of the link annotation. The body of the link is either textual (as for copy links, and other associative assertions) or can refer to a particular part of an archival description that makes a reference to an external entity, which are typically access points that are derived from subjects, corporate body, person or family (CPF) authorities, or place references within the <controlaccess> section of EAD-encoded material. In cases where links are directional — as with copies — we extend the Open Annotation model to include a "source" assertion between the outgoing entity and the link annotation.

Links between content items (institutions, authorities, or archival units) can be created for a number of reasons, with relationship types derived from the International Standard for Archival Authority Records (ISAAR) (Vitali, 2004), section 5.3.2:

- Hierarchical: the entities have a superior/subordinate relationship
- Temporal: one entity succeeded or is succeeded by the other
- Family: the entities belong to the same family
- Associative: the entities are related in some other way not covered above

Although copy links originally used the "Associative" type, we have extended ISAARs relationship type categories with a dedicated "Copy" type, since these relationships are so prominent in EHRI's field, and distinguishing them from other associative relationships enables dedicated functionality and analytics for this type. (In contrast, access point links still use the "Associative" type, since they are already distinguished by the presence of a non-textual body.)

The next section looks at the first of our linking case studies: materialising links between descriptions of copy collections — material physically copied from one archive to another — and descriptions of the source (or sources) from which they derive.

Copy-Original Links

As discussed above, a feature of the Holocaust-related archival landscape is that there is substantial duplication of important source material, in large part due to the existence of institutions like Yad Vashem, the United States Holocaust Memorial Museum (USHMM), and Mémorial de la Shoah, which

Records Management Journal

have mandates to preserve and improve access to specifically Holocaust documentation. Such institutions hold a substantial quantity of material copied — in electronic or physical form — from other institutions, often regional archives, and subsequently (re-)organised and catalogued using in-house procedures by, for instance, grouping multiple fonds sourced from particular archives together as a collection.

Materialising these copy-original connections via discrete, structured links has several distinct advantages for EHRI and users of the portal. It allows browsing between related items, and doing so in a bi-directional manner, even when the information from which a link was derived is contained within only one of the connected item descriptions. In cases where a copy collection derives from *part* of another fonds — for example, a specific series or sub-series — this can be made explicit with the targets of the links, and such links can themselves hold dedicated metadata, incorporating information such as when and why a particular copy was made. Perhaps most significantly, structured links allow making the provenance of material more explicit, which is of particular concern in an environment like the EHRI Portal which takes metadata out of its native context, easily obscuring the fact that two archival descriptions might actually refer to the same underlying material, an important consideration for researchers for whom physical access is required.

Creating Copy-Original Links

Copy-original links in the EHRI Portal have to date been created only via manually, or via semi-automated batch processes with manual validation. There are two means of fully automated link creation that we have considered:

1. Creating links encoded unambiguously in externally-sourced structured data (e.g. EAD)
2. Creating links automatically via computational inference

In the first instance we have been limited by a lack of structured data. EHRI's primary means of structured data input is the EAD 2002 XML format (Pitti, 1999). This XML schema — the characteristics and limitations of which have been extensively explored, including by Shaw (2001), Gartner (2015), and EHRI's own report on metadata standards (Riondet et al., 2017) — in theory provides enough semantic scaffolding to encode in a machine-readable way references to the original archival sources for copied material. Specifically, the ISAD(G) fields 3.5.1 "Location of Originals" (LO) and 3.5.2 "Location of Copies" (LC) can, when translated to EAD XML (fields <originalsloc> and <altformavail> respectively), contain both descriptive text and/or pointers to external entities.

Our first problem was simply institutions taking different semantic interpretations of the ISAD(G) guideline and using different fields to record collection provenance. At Yad Vashem, for example, the field "Scope and Content" was used, USHMM used "Archival History", and Cegesoma the field "Biographical History". We realized that authority records, and most specifically the creator of the archival unit is particularly relevant to determine the relationships between different archival units. However, in practice, this field is often missing in archival descriptions.

Records Management Journal

Secondly, while many of EHRI's partner archives did record collection provenance to some extent, there was considerable variation in style and levels of detail, and no institutions encoded the information in a structured manner. Even were they to have provided structured references, the dearth of appropriate controlled vocabularies or machine-readable handles (in the semantic web sense, URIs: a universal way of referring to an archival institution or record) other than standard web URLs would have made automatic entity resolution difficult.

Without the means of taking unambiguous copy-original references directly from partner-provided data we were left with the option of extracting links purely via machine resolution of entities from plain text. Accuracy concerns, and the relative ease of manual validation given the modest size of the overall data set, mitigated in favour of a semi-automated approach, however. The next sections explore our formalisation of copy-original links and the process for link creation.

Formalisation of Copy-Original Links

While the ideal case, in linking terms, is to connect an archival description of some copied material to another description representing its source, this is often not possible due to a lack of specificity in the source data, or due to one or the other being ambiguous, uncatalogued or otherwise not referenceable. In this case, we have to refer to the holding institution itself as the link target. An even more general case is where we know that an archive holds copies of material from another institution but we do not have detailed information about exactly what it is. While imperfect, the latter two situations are common and recording them can still provide valuable information to users of the EHRI Portal, as well as to the archival institutions concerned. We therefore need to account for four different types of linkage, from most to least specific:

- Copy collection to original collection
- Copy collection to original repository
- Copy repository to original collection
- Copy repository to original repository

Since the two items connected by a copy-original link are not equivalent — one being derived from the other — these links are directional. Since material may have many copies but a copy generally only has one original, we take the direction of these links to "point" from the derivation to its source.

A corollary of this directional assumption is that we treat the Location of Originals and Location of Copies as the logical inverse of each other and only create links in the style of the former field, regardless of which field the source information derives.

Records Management Journal

Batch Link Creation

Copy-original link creation is a batch process in which tabular data describing a set of links is ingested into the EHRI Portal. The preparation of the tabular input data involves resolving plain text content extracted from Location of Originals (LO) or Location of Copies (LC) fields associated with an archival unit into a counterpart entity representing the corresponding referenced item.

The first set of copy-original links created in bulk — that is, excluding those added as part of manual cataloguing processes — was a set of over 2,500 references from USHMM archival descriptions to 591 distinct sources, of which 344 referenced original holding institutions (the remainder being mostly private donors and thus non-linkable.) Once in tabular form, these textual fields were resolved manually by EHRI staff by inserting the identifier of the EHRI institution record to which they referred. Once all resolvable entities were added, the dataset was ingested into the EHRI Portal resulting in the creation of 1,767 links from USHMM archival descriptions to their original holding institutions. Aside from the importance of USHMM as an aggregator of Holocaust-related archival material and its subsequent importance as a source of copy-original links in the EHRI Portal, this set of links was significant because, with a set of textual references manually resolved to their target entities, it provided useful text data for subsequent attempts to add greater automation to the process of resolving copy-original references from free text, described in the next section.

Suggesting Candidate Links via Automatic Matching

By adding a degree of automated mining of LC and LO text for potential connections between archival entities (institutions or collection descriptions), we hope to both lower the workload for EHRI staff performing manual cataloguing duties (of which creating copy-original links is a part), and be able to more consistently discover links within data provided third-parties.

The current system for suggesting "candidate" links builds on our existing Apache Solr-based search infrastructure, with some tuning of parameters to better handle the type of text content found in LO or LC fields, as opposed to a directly-formulated search query. For detecting *either* institution or archival description references a two-stage process is used, first performing a search of institution records using specifically the authorised form of name, alternate names (including parallel translations), and address fields and then, if a match is found, performing a secondary search for identifiers of archival units held by that institution.

This works well for certain common cases, such as fields with contents such as "[Institution name] [accession number]" — for example "Yad Vashem M.52", for which a correct link target would be that fond's EHRI identifier, "il-002798-m_52". Nonetheless, variations are plentiful, with most ambiguity deriving from:

- Text incorporating names of both copy and original, e.g " X was copied from Y"
- Multiple references within a single text
- Highly generic collection identifiers: e.g. "a" or identifiers with inexact punctuation
- Ambiguous institution names, e.g. the many distinct branches of the Bundesarchiv

Records Management Journal

- Cases where the location of copies text includes the name of the copy collection, which in turn referenced the archive from which is was sourced

In the future, in the absence of unique identifiers that would make fully-automated linking more feasible, increased standardisation (or recommendations) as to how LO or LC fields are to be completed by archivists could make reliable computer-assisted inference more practical.

Visualising Copy-Original Links

Figure 1 provides a visualisation of our copy-original linking activities as a force-directed graph, with institutions as graph nodes and links between them as edges. Since the majority of our links derive from structured data provided by USHMM, that institution is unsurprisingly very central to this network. Smaller clusters, however, can be seen around several other institutions. Nodes are sized according to weighted in-degree, that is, the number of times they are referenced as the original holding institution for material at other archives. At present, the Arolsen Archives (formerly the International Tracing Service (ITS)) and the Jewish Joint Distribution Committee (JDC) are the most significant sources of copy collections. As EHRI's data on copy collections gets more comprehensive we would expect this visualisation to become less centralised around USHMM, with distinct clusters around other large aggregators of Holocaust-related documentation such as Yad Vashem and Mémorial de la Shoah. Currently, copy-original links connect 17.5% of the 2,167 CHIs listed in the EHRI Portal.

Records Management Journal

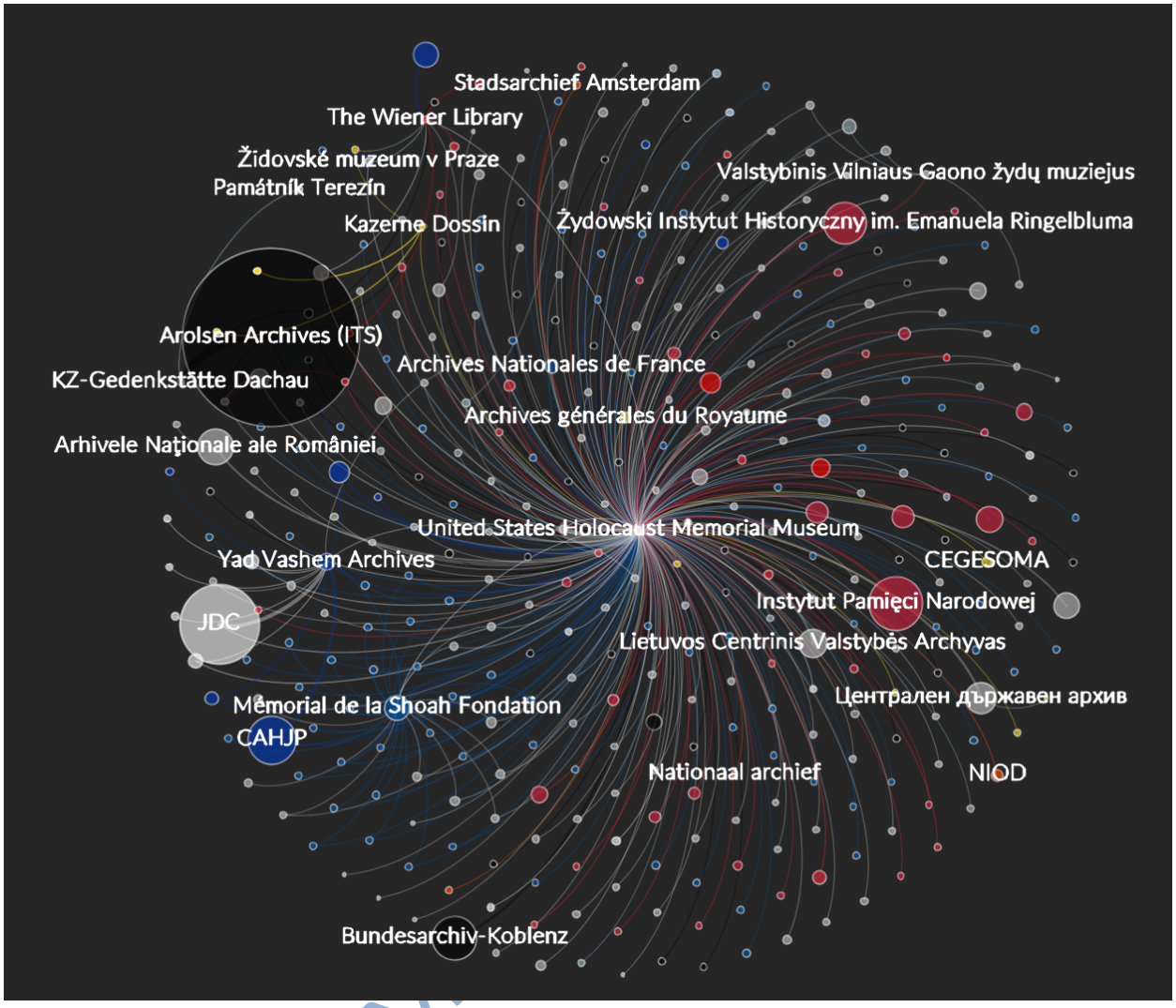

Figure 1: Visualisation of archival institutions connected by copy-original connections in the EHRI Portal. A large number derive from USHMM data, hence its central location.

The next section describes EHRI's second method of linking archival data: integrating multilingual access points across data from different institutions.

Records M

Subject & Authority Co-Referencing

The use of subject, place, and authority terms to index archival descriptions is one of the primary ways to make them discoverable by researchers. And indeed, such terms - access points - are commonly used by archival institutions with which EHRI deals: of those institutions whose metadata was available to EHRI in structured form, over 90% of the descriptions were indexed using some set of subject, authority, or place terms.[3] In data aggregated by EHRI there was, however, little standardisation of these terms: many institutions with significant holdings used their own internal language-specific controlled

[3] In total, 90.4% of provided fonds included access points, in 59.2% of unit descriptions at all levels.

vocabularies, and of those that employed vocabularies such as Library of Congress Subject Headings (LCSH) or Faceted Application of Subject Terminology (FAST), no two used the same one in broadly compatible ways. Moreover, as noted by Shaw (2001), EAD does not have sufficient structure to represent commonly-used features of structured index terms such as subject sub-fields, and as a consequence such subtleties were often lost in transit, making what was received much harder to decode and resolve.

As a result, while EHRI did ingest access points from partner institutions, most were initially "unlinked" and thus primarily useful for free-text search purposes rather than traditional keyword-based browsing. Had EHRI, in contrast, opted to treat every *subject* access point encountered in partner data as an index term (not including CPF (corporate body/person/family) authorities or places), this would have resulted in over 80,000 distinct terms for subjects alone — not an optimal browsing experience.

The project did, however, index descriptions created either manually or with direct partner involvement to a number of *Holocaust-specific* controlled vocabularies, namely:

- EHRI's thesaurus, consisting of 913 hierarchically-organised Holocaust-specific terms, translated into 10 languages, derived in substantial part from Yad Vashem's subject index terms.
- a set of over 1,300 ghettos, including geographical location information, derived from Yad Vashem's Encyclopedia of Ghettos and USHMM's Encyclopedia of Camps and Ghettos.
- a set of over 2,000 camps, derived from USHMM's Encyclopedia of Camps and Ghettos along with data from the Arolsen Archives, Wikidata, Wikipedia, and Bundesarchiv.
- a set of over 3,000 Holocaust-related person and corporate body authorities maintained by the project, catalogued in alignment with ISAAR(CPF) and EAC-CPF.

Given the existence of these datasets — which received significant attention and improvement over the course of the project's second phase (Cooey, 2019; Nispen and Jongma, 2019) — efforts were latterly made to integrate them, on a partner-by-partner basis, with access points from third-party descriptions.

Method of Co-Referencing Access Points

Access points in the EHRI Portal consist of simple text strings within one of several categories (person, corporate body, place, subject, and genre), a set of which can be part of textual descriptions associated with an archival unit. Materialising an access point involved creating a link between the archival unit to which it belongs and a particular item within a controlled vocabulary. Unlike copy-original links, however, the body of the link is not just textual but refers directly to the originating access point.

The matching process involves three steps, beginning with a textual similarity comparison between the original access point text and the multilingual labels belonging to controlled vocabulary items. For items where no matches were detected, a secondary pass retries the similarity comparison using permutations of word order (frequently, first and last names.) Finally, where ambiguities remain — commonly for person authorities where dates of birth or death are missing or imprecise — the institution from which the original description originated will be contacted for confirmation.

Records Management Journal

Once a match has been made between a particular institution's usage of a term and an EHRI-specific subject term, place or authority, a link is created and the mapping between the access point text and vocabulary item retained so that structured data subsequently received from the institution can be pre-processed to encode the reference directly. This latter step ensures that when new material is received from that institution, or existing material updated, the connections are not lost.

Access Point Statistics and Visualisation

Table 1 shows the number of controlled vocabulary items connected to distinct access point strings, along with the number of archival units and archival institutions this connects. The percentages in the latter two columns show the proportion of items connected relative to items linked through other means (either manually or via structured data already encoded with references to EHRI vocabularies.) Since co-referencing is done on a per-institution basis, relatively few institutions have been connected via this method. The high percentage of archival units connected, however, shows that by focussing on institutions with large holdings, co-referencing is an efficient method of integrating access points in bulk.

Vocabulary	# Terms	# Distinct APs	# Archival Units	# Institutions
Subject Headings	583	2212	25075 (81.2%)	15 (3.5%)
Camps	466	807	7340 (97.8%)	9 (20.9%)
Ghettos	312	502	3212 (93.9%)	5 (15.2%)
Persons	1583	2224	4492 (89.3%)	10 (13.5%)
Corporate Bodies	557	937	14882 (73.3%)	13 (3.7%)

Table 1: Access points linked via co-referencing with percentage of total archival units and institutions connected to controlled vocabularies.

Records Management Journal

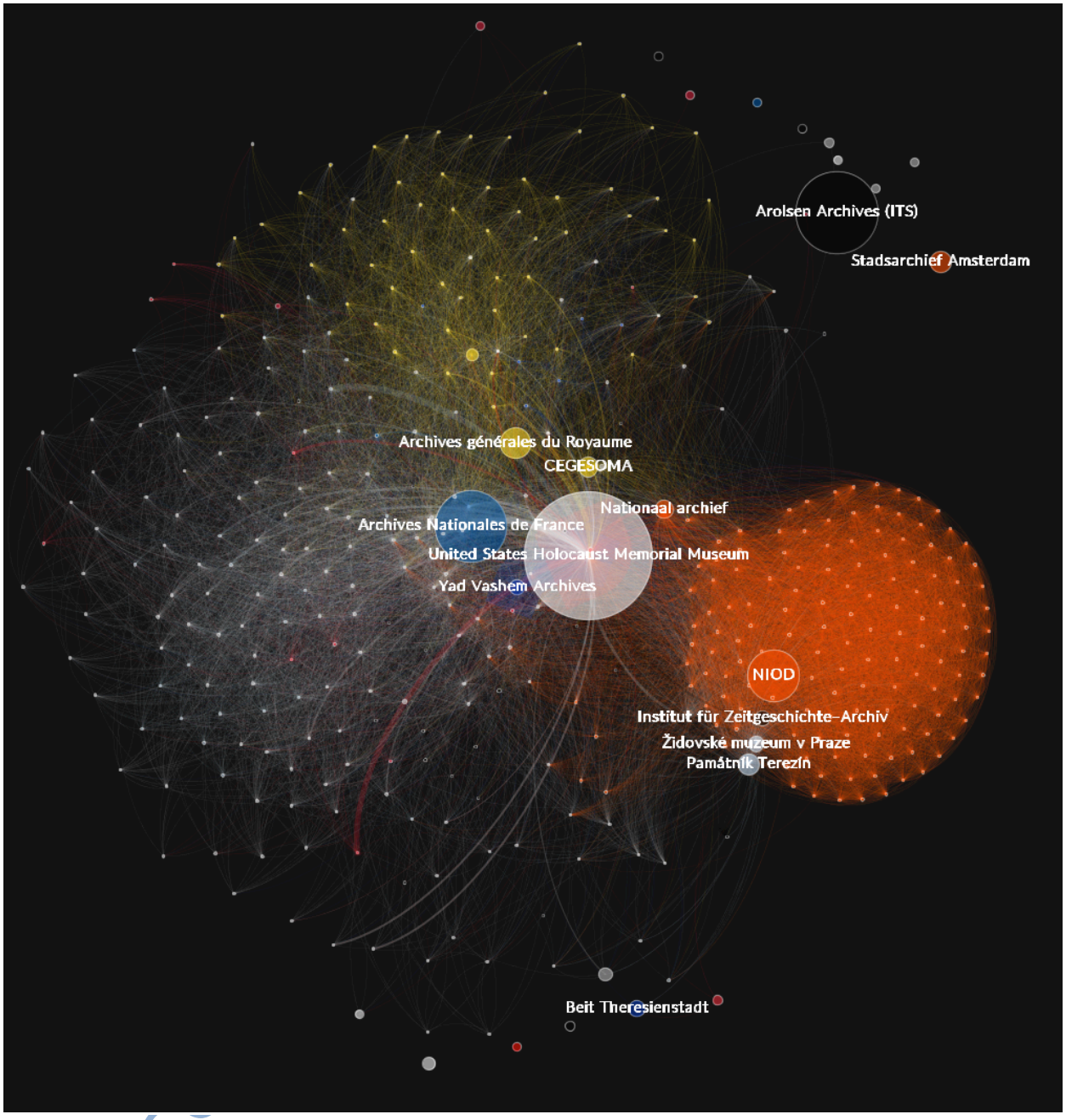

Figure 2: Visualisation of subject access point co-occurrence among archival institutions catalogued by EHRI.

Figure 2 provides a visualisation of our subject access point integration activities, encompassing all linking methods. Each node in the image represents an archival institution that has *either* more than 100 archival units, or more than 8 co-referenced subject terms with archival units from other institutions (these are cutoff points that remove noise from the visualisation.) Nodes are sized according to the number of archival units listed in the EHRI Portal. The figure shows several characteristics — and limitations — of EHRI's integration activities:

1) There is a high degree of correlation between *institutions within* two particular countries: Belgium (in yellow) and — to an even greater degree — the Netherlands (orange). We might expect this if such access points were purely language-specific, but in fact they are all references to terms within EHRI's multilingual thesaurus, albeit in two largely disjoint sets. The actual reason for this is that a significant amount of data concerning Holocaust-related material in both Belgian and Dutch archives derived not from the archives themselves (with their typically diverse cataloguing practices) but from aggregated sources: Sources pour l'histoire des populations juives et du judaïsme en Belgique/Bronnen voor de geschiedenis van de Joden en het Jodendom in België, 19de-21ste eeuw; and Network Oorlogsbronnen (https://www.oorlogsbronnen.nl/) respectively (for more details, see the EHRI reports for each country.) If these aggregate sources had more overlap in the index terms they used we would see less spatial divergence between the heavily yellow and heavily orange areas of the visualisation.
2) Some institutions with a significant number of archival units, e.g. Arolsen Archives (formerly known as the International Tracing Service (ITS)) are not integrated at all in this respect due to the lack of subject access points within their archival descriptions.
3) While there are relatively few U.S.-based archival institutions within the EHRI Portal, USHMM is somewhat centrally located due to its large number of subject-indexed archival units.

There is still considerable work to be done in order to fully integrate access points within the EHRI Portal. One area of considerable unexploited potential lies in expanding the number of geospatially indexed descriptions available, in addition to those currently indexed to EHRI's ghettos and camps vocabularies.

Discussion

Our experience with copy-original references has some relevance to the ongoing development of new archival standards. For example, the Records-in-Context Conceptual Model draft available at the time our work took place (RiC-CM 0.1) did not permit us to easily describe the copy/original relationships we typically encountered *in the wild*, since its *is-copy-of* (RiC-R6) and *has-copy* (RiC-R1) relationships could only be applied to **Record** entities (the RiC equivalent of an item-level archival unit) and not **Record Set** aggregations. The substantially revised RiC-CM 0.2, released while this article was in review, has largely addressed this particular issue by moving the domain of *is-copy-of* relationships (along with many others) to a base class common to Records and Record Sets entities. While this extra semantic complexity undoubtedly entails additional intellectual overhead, in this specific case it has made alignment between EHRI's data models and RiC-CM, and thus eventual implementation of RiC-O in a linked data context, somewhat easier.

While we anticipate the maturing of RiC-CM and RiC-O, the provision of means to query EHRI's collection metadata in structured ways — enabling bidirectional feedback between EHRI, its partner institutions, and other stakeholders — has been an ongoing focus. At present, copy/original information added by EHRI is incorporated into EAD-format descriptions that can be exported from the portal (albeit it in a somewhat semantically unfriendly form), and are also queryable via its GraphQL API (Bryant,

Records Management Journal

2017) or downloadable as a dedicated tabular dataset. Likewise, information about co-referenced access points is accessible in exported EAD and EAC-CPF, and via the EHRI portal's APIs, meaning that if an EHRI partner institution wishes to know whether EHRI's data enrichment efforts have involved their own collections they are able to determine this using structured data, with the automation possibilities this provides.

Conclusions

One of our main goals within the EHRI project is providing a virtual view of an archival landscape that adequately captures the complexities and messiness that have emerged from the history of the Holocaust and its aftermath. Doing so puts us in conflict with the urge to simplify, uniformalise, and abstract over differences in organisational approach, and in some cases, can be difficult to reconcile with mainstream archival practice. The existence of copy archives and the amount of copied (and alternately-described) material is one reason why we have focused on documenting such connections between archives and their holdings, in addition to more conventional approaches such as integrating descriptive index terms. In the upcoming third phase of the EHRI project, starting in 2020, we aim to both continue these tasks, refining the workflows described above, and expand and improve the access to collection metadata in LOD-compatible ways. In doing so, and with the creation of tools that can better take advantage of this enhanced metadata in querying and visualising the Holocaust as an information domain, we will come closer to realising the potential of this trans-national dataset.

Acknowledgements

EHRI is a consortium with many partners across numerous countries, and many individuals were involved in the work described herein in addition to the co-authors. In particular we would like to thank Anna Ullrich and Giles Bennett at the Institut für Zeitgeschichte (IfZ) for their efforts in resolving and validating USHMM copy-collection original holding institutions, Linda Reijnhoudt at DANS for her work on formalising copy-original links, the EHRI-teams at Yad Vashem, CEGESOMA, Kazerne Dossin, and the EHRI Data Identification and Integration Work Package in general.

Bibliography

Bearman, D.A., Lytle, R.H., 1985. The power of the principle of provenance. Archivaria 21, 14–27.

Blanke, T., Bryant, M., Frankl, M., Kristel, C., Speck, R., Daelen, V.V., Horik, R.V., 2017. The European Holocaust Research Infrastructure Portal. J Comput Cult Herit 10, 1:1–1:18. https://doi.org/10.1145/3004457

Bredenberg, K., Jagodzinski, S., 2014. Archives Portal Europe–A Challenge of Harmonization and Outreach, in: 9th European Conference on Archives. Girona.

Bron, M., Proffitt, M., Washburn, B., 2013. Thresholds for Discovery: EAD Tag Analysis in ArchiveGrid, and Implications for Discovery Systems. Code4Lib J.

Bryant, M., 2017. GraphQL for archival metadata: An overview of the EHRI GraphQL API, in: 2017 IEEE International Conference on Big Data (Big Data). IEEE, pp. 2225–2230.

Bryant, M., Reijnhoudt, L., Simeonov, B., 2018. In-place Synchronisation of Hierarchical Archival Descriptions, in: 2018 IEEE International Conference on Big Data (Big Data). Presented at the 2018 IEEE International Conference on Big Data (Big Data), pp. 2685–2688.

Records Management Journal

https://doi.org/10.1109/BigData.2018.8622517
Bryant, M., Reijnhoudt, L., Speck, R., Clerice, T., Blanke, T., 2015. The EHRI Project - Virtual Collections Revisited, in: Aiello, L.M., McFarland, D. (Eds.), Social Informatics, Lecture Notes in Computer Science. Springer International Publishing, pp. 294–303.
Bunn, J., 2013. Developing descriptive standards: a renewed call to action. Arch. Rec. 34, 235–247. https://doi.org/10.1080/23257962.2013.830066
Cooey, N., 2019. Leveraging Wikidata to Enhance Authority Records in the EHRI Portal. J. Libr. Metadata 19, 83–98. https://doi.org/10.1080/19386389.2019.1589700
Cosgrave, R., 2003. The AIM25 project. J. Soc. Arch. 24, 159–174. https://doi.org/10.1080/0037981032000127025
EGAD, 2016. RiC-CM-0.1.pdf | International Council on Archives [WWW Document]. URL https://www.ica.org/en/egad-ric-conceptual-model-ric-cm-01pdf (accessed 11.8.19).
Gartner, R., 2015. An XML schema for enhancing the semantic interoperability of archival description. Arch. Sci. 15, 295–313. https://doi.org/10.1007/s10502-014-9225-1
Grimsted, P.K., 2017. Pan-European Displaced Archives in the Russian Federation: Still Prisoners of War on the 70th Anniversary of VE Day, in: Displaced Archives. Routledge, pp. 140–167.
Hill, A., 2002. Bringing Archives Online through the Archives Hub. J. Soc. Arch. 23, 239–248. https://doi.org/10.1080/0037981022000006408
Nispen, A. van, Jongma, L., 2019. Holocaust and World War Two Linked Open Data Developments in the Netherlands. Um. Digit. 3. https://doi.org/10.6092/issn.2532-8816/9048
Pitti, D.V., 1999. Encoded archival description: An introduction and overview.
Riondet, C., Romary, L., Nispen, A. van, Rodriguez, K.J., Bryant, M., 2017. Report on Standards (report).
Sanderson, R., Ciccarese, P., Van de Sompel, H., Bradshaw, S., Brickley, D., a Castro, L.J.G., Clark, T., Cole, T., Desenne, P., Gerber, A., others, 2013. Open annotation data model. W3C Community Draft.
Shaw, E.J., 2001. Rethinking EAD: Balancing flexibility and interoperability. New Rev. Inf. Netw. 7, 117–131. https://doi.org/10.1080/13614570109516972
Speck, R., Blanke, T., Kristel, C., Frankl, M., Rodriguez, K., Daelen, V.V., 2014. The past and the future of holocaust research: From disparate sources to an integrated european holocaust research infrastructure. ArXiv Prepr. ArXiv14052407.
Speck, R., Links, P., 2016. Who holds the key to Holocaust-related sources? Authorship as subjectivity in finding aids. Holocaust Stud. 22, 21–43.
Vanden Daelen, V., Bennett, G., Ullrich, A., Pohl, D., Babeş, A., Haultain-Gall, M., 2019. Final report on data identification and integration (Deliverable No. D9.5).
Vitali, S., 2004. Authority Control of Creators and the Second Edition of ISAAR(CPF), International Standard Archival Authority Record for Corporate Bodies, Persons, and Families. Cat. Classif. Q. 38, 185–199. https://doi.org/10.1300/J104v38n03_15

Records Management Journal