EasyChair Preprint

№ 2716

The EHRI Project - Virtual Collections Revisited

Mike Bryant, Linda Reijnhoudt, Reto Speck, Thibault Clérice and Tobias Blanke

February 19, 2020

The EHRI Project - Virtual Collections Revisited

Mike Bryant, Linda Reijnhoudt, Reto Speck, Thibault Clerice, and Tobias Blanke

Centre for e-Research, Department of Digital Humanities, King's College London
{michael.bryant,thibault.clerice,reto.speck,tobias.blanke}@kcl.ac.uk
Data Archiving and Networked Services
linda.reijnhoudt@dans.knaw.nl

Abstract. This paper introduces details of EHRI's approach to user-centric data integration across heterogeneous archival institutions using virtual collections. Virtual collections provide the means to re-unite archival material that has, through complex historical circumstances, been deposited in many physical locations. They also allow the creation of subject-specific groupings of material more closely comparable to archival research guides, and provide users with the ability to organise their own research in personalised ways.

1 Introduction

The overriding mission of the European Holocaust Research Infrastructure (EHRI) project[1] is to integrate into an online portal information on Holocaust-related archival documentation that is physically dispersed across repositories around the world [1]. This is a particularly challenging mission as archival sources on the Holocaust have, arguably more so than any other sources relating to contemporary history, undergone very extensive processes of destruction, fragmentation and dispersal: the Nazis endeavoured to destroy evidence about the crime; survivors migrated widely after the war and took important documentation with them; a wide variety of post-war historical commissions and projects have sought to reassemble surviving evidence, thereby frequently pulling material out of its original context, etc. All this has conspired to make historical research on the Holocaust a very complex undertaking. Indeed, relevant Holocaust source material can be found in more than 1,800 institutions across the world, and it is frequently not evident from available archival descriptions how the sources from one repository may relate to the ones of another [2].

One of the key challenges we faced in EHRI was therefore to establish a platform for forming virtual collections that allows the re-establishment of latent, lost or implied connections between archival material without further clouding the provenance and physical arrangement of such material . This paper offers

[1] http://www.ehri-project.eu

a concise outline of our approach. Section 2 provides an overview of the background behind this work, while section 3 describes the heterogeneous source data we encountered and includes an analysis of the main data integration challenges. Section 4 explains the specific rationale for EHRI's use of virtual collections to link together archival descriptions of physically dispersed material according to research themes, while section 5 describes our approach to the presentational issues we have faced. Section 6, finally, provides details about the technical implementation of virtual collections in the EHRI environment, including a sketch of some of the technical challenges we have encountered and a brief outlook of how we are planning to tackle these challenges in the future.

We believe that the platform we have established will enable researchers to virtually explore physically dispersed Holocaust collections, and to dynamically establish new connections, enabling the study of the Holocaust from a quantitatively increased and qualitatively more integrated empirical basis. At the same time, our approach to virtual collections will have general applicability to the challenge of how to develop interfaces to dispersed, fragmentary and complex historical collections that aim at offering researchers advanced search, browse and analysis capabilities across such collections.

2 Prior work

Virtual collections have been subject to extensive debates. They are frequently seen as one of the main benefits digitisation of resources can deliver to libraries and archives, with the digitisation of resources and tools meaning their organisation can now be conducted in a distributed, decentralised manner. Through virtual collections, users of archives and libraries can engage in what Terry Cook [3] has termed "community-based archiving", developing their own view onto holdings not bound to the organisation by the collection professional. It is thus no surprise that virtual collections have attracted a lot of interest from professional users and especially researchers [4]. Blanke et al [5] discuss the case of classicists working through digital libraries who can build up their own virtual collection bringing together resources from multiple data stores. Classicists with a common interest in certain research topics can share these virtual collections with each other.

In addition to professional scholarly work, virtual collections also promise to support the integration of amateurs and armchair researchers, as they distribute access and means of data curation. In [6] a case study is presented where virtual collections help with involving amateurs in the digitisation work of museums. Other museum visitors seem to accept these kinds of virtual collections as useful. Neither Blanke et al [5] nor Terras [6], however, discuss the exact nature of what constitutes a virtual collection.

For archives, Bradley Westbrook [7] and William E. Landis [8] identify virtual collections as a tool for responding to the unmediated needs of users in ways that the descriptive aids developed by archivists themselves cannot. Traditional archival finding aids, concerned primarily with structure rather than substance,

were developed with the assumption that the archivists themselves would be able to direct a user with a subject-based query to the appropriate material in their provenance-based fonds. "In these online systems", however, Landis writes, "mediation is not something we can impose on end users the way we have been able to at our reference desks." [8]

Historian Alessandro Salvador goes even further and welcomes the potential of virtual collections for overcoming complexity in fragmented archival landscapes by integrating archival records, repository information, bibliographies, and other descriptive data into online research guides [9, slide. 11] . This view aligns closely with Candela and Straccia's notion of virtual collections as "user defined un-materialized views over very heterogeneous information space" [10], and can narrowly be interpreted as a focussed application of the virtual research environment (VRE) in facilitating the "linking, integration and subsequent analysis of data" [11]. Because virtual collections offer these opportunities to access data across repositories in a "heterogenous information space", it is not surprising that their development is often seen as one of the main benefits of Linked Data approaches. For instance, [12] uses virtual collections to make cultural heritage metadata and vocabularies interoperable.

This paper adds to this existing work by offering details of a concrete implementation of virtual collections in a data integration context, aimed at mitigating real-world problems faced by researchers in their use of digital tools.

3 The challenge of heterogeneous archival data

Existing descriptions of Holocaust-related material are heterogeneous and reflect the diversity of institutions that hold such material; spanning the whole spectrum from national archives and large dedicated Holocaust memory and research institutions to small communities archives and private repositories [13]. Despite the fact that relevant conceptual and technical standards exist for the description of archival materials,[2] these standards are frequently not adhered to in practice. A survey we undertook of EHRI partner archives found that fewer than half of repositories follow international descriptive standards, and throughout our work, we have encountered a great variety of descriptive paradigms [2].

In terms of integrating existing descriptions and establishing connections between related material, institutional diversity in the following areas has proved particularly challenging:

Depth of hierarchies in the descriptions

In many cases a particular archival fonds might be described broadly at the collection level (all items together), with specific descriptions for each item. In other cases there can be many more levels of description, as the fonds is broken into subfonds and then perhaps into series, subseries, and files. Even with fonds of broadly comparable size, the number of levels used in the description varies widely between repositories depending on their specific

[2] Principally those developed by the International Council on Archives (ICA).

organisational practices, and is not strongly guided by applicable standards.

Incompatible vocabularies
It is at present quite rare for archives that assign subject, place, and name (person, family, or corporate) classifiers to archival descriptions to do so from common vocabularies, such as the Library of Congress subject headings (LCSH)[3]. On the contrary, such "access points" (as the ICA refers to them) usually have a legacy basis within each institution.

Provenance vs. Pertinence
Organisation that *respect des fonds* by reflecting the provenance of the material, versus those that arrange collections on the pertinence principle, grouping together records according to subject content.

It very soon became apparent that a wholesale standardisation of the institutionally diverse descriptions prior to integration into the EHRI portal would be undesirable, and, indeed, infeasible. On the one hand, standardising to a common denominator would entail an unacceptable loss of information. On the other hand, institutionally idiosyncratic descriptions can, at times, reveal much about the complicated archival histories these collections have undergone, and as such constitute in themselves a valuable information resource.

Unlike other large-scale archival integration projects such as ApeX[4], we therefore decided to keep standardisation of structure to a minimum, and take a 'take it as it comes' approach to integrating data and building virtual collections. To enable this approach, we have dedicated much effort to establishing a platform that allows the expression of connections between related descriptions of archival items held in diverse repositories.

3.1 Case study: Integration of material by Hans G. Adler

Hans Günther Adler was a Czech Jew born in Prague who, during the course of the war, was imprisoned in Theresienstadt, Auschwitz, and Buchenwald. Following liberation he worked at the Jewish Museum in Prague before emigrating to the United Kingdom in 1947. A prolific writer throughout his life, Adler's works and letters, both original manuscripts and copies, are distributed in many different institutions throughout Europe and beyond, including the EHRI partners Jewish Museum Prague, the Institute for War and Holocaust Studies (NIOD) in Amsterdam, the International Tracing Service (ITS) in Bad Arolsen, King's College London, and Yad Vashem in Jerusalem. Due to this wide distribution, it is difficult for contemporary researchers to gain a coherent overview of the output of this important figure in Holocaust scholarship.

Using virtual collections, we facilitate the creation of such integrated views on material of specific research interest; allowing descriptions of materials from

[3] http://id.loc.gov/authorities/subjects.html
[4] http://www.apex-project.eu

many archives to be aligned with each other. In addition to material from the same *fonds* but physically separated, there is also a case for including in such virtual collections that, from a purely archival sense, should be separately organised, such as letters sent by one individual to another. For example, the letters sent by Adler to Dora Philippson and kept at Beit Theresienstadt in Israel could belong within a virtual collection based on Adler's work.

4 Virtual collections

Virtual collections serve therefore three main purposed within EHRI:

1. To assemble virtual fonds from multiple dispersed archives, reuniting material that belongs together under the provenance principal (figure 1).
2. To support "research guides" in EHRI's portal, overlaying the material itself with a higher-level thematic overview written from the perspective of the historian or the researcher (figure 2).
3. To allow users to create and organise personal lists of items, termed bookmark sets. Bookmark sets are, by default, private to individual users but can be made public and shared at the user's discretion.

Whereas archival finding aids often focus primarily on the structure of the material within a particular repository, research guides typically take a more subject-oriented approach, explicitly tailored to the user of the archives and often prepared by historians in the form of a book or pamphlet comprised of long-form narrative text.

There is, however, invariably some overlap between the descriptive finding aid and the archival research guide. This overlap can often be seen expressed in the areas that finding aids venture into narrative and research guides into structure, as they both must necessarily do. Virtual collections in EHRI embrace this overlap, allowing descriptive "glue" to complement structure in places where this would aid the understanding of the material itself.

Just as material within physical archives is organised hierarchically (from collection to item level), the EHRI research guides can provide their own nested sections that may be wholly or in-part comprised of references to physical documentary units. This arbitrary nesting allows items from separate physical collections to be combined in ways that maintain the coherence of their structure, regardless of the level at which the original archivists have chosen to place the descriptive detail. These intermediate virtual sub-levels are directly analogous to the sub-fonds, series, and sub-series commonly used by archivists, and the descriptive information that can be associated with individual components of EHRI virtual collections takes the same ISAD(G)-based format as standard archival descriptions. In this respect, virtual collections within EHRI more literally comprise virtual *finding aids*, composed of digital surrogates and not, as Westbrook [7] has proposed, of "discrete digital objects and digital objects borrowed from their established collection contexts."

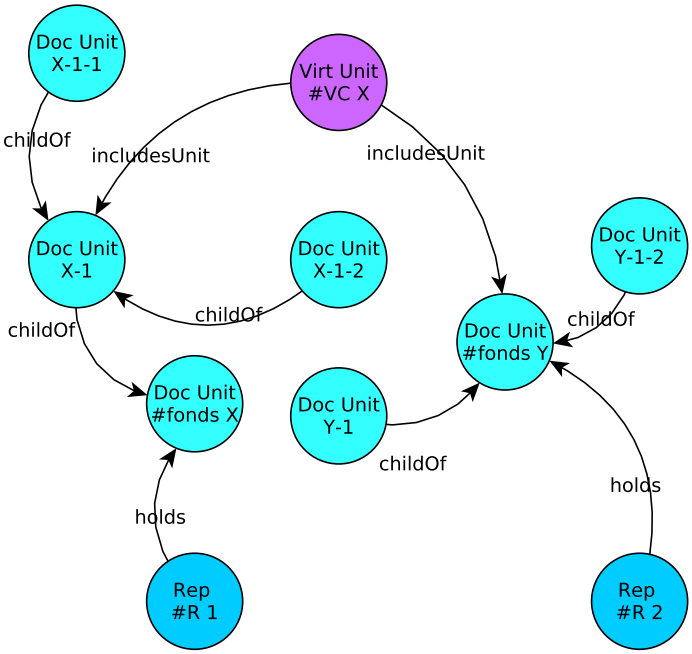

Fig. 1. Dispersed archives: repository R1 holds a fonds X, with a part missing, which is held by repository R2 there called fonds Y.

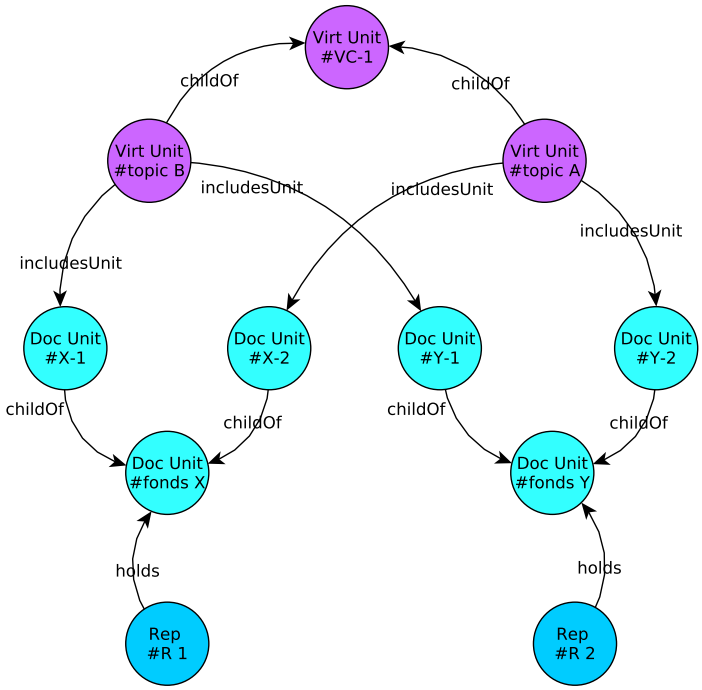

Fig. 2. Thematic collection: a new structure, based on topics, with units from different archives, and possibly different levels.

Figures 1 and 2 illustrate two of the ways that virtual collections can be assembled, as, respectively, the gathering together of material from the same source into a dispersed virtual archive, and as a thematic collections organised by topics. Figure 1 shows the need to freely combine items at different levels of description (*VC X* incorporated both *fonds Y* in its entirety along with a child item from *fonds X*) in order to account for institutional differences in descriptive style. Figure 2, on the other hand, uses a structure solely based upon pertinence, with material from multiple different fonds included within subject-based intermediate levels.

5 Representational Challenges

As outlined in section 3, a key challenge for EHRI is to maintain clarity about the provenance and identity of information displayed in its online portal. Simultaneously, we wanted to avoid the "structural opacity" that Wendy Scheir [14] has identified as a major stumbling block for users of electronic finding aids. With material that can be either aggregated from many different archives, created directly by EHRI, or contributed by individual users, there exists potential for misattribution and confusion. Virtual collections, by presenting material in new and different contexts to that represented by the archive in which they physically reside, heighten this danger.

There were two situations we identified that revealed the problems with identity and context in offering virtual collections alongside standard physically-derived digital finding aids, both involving navigation through item hierarchies via different contextual entry points.

In the standard mode of the EHRI portal the context of archival units (descriptions of material) is explicit in a hierarchy that typically reflects the way the material is physically stored, in a given collection or fonds, and by extension within a particular repository. Since a unit of material can only exist in one place at a time (copies notwithstanding) its archival context can be made fairly unambiguous.

Within the EHRI environment, however, a documentary unit can exist within *many different virtual collections* simultaneously, collections which may have been created for entirely different purposes and thus represent very different contexts. Respecting these different contexts implies respecting the manner in which a given item was arrived at by a particular user.

Similarly, we wanted to avoid a situation where a user, starting at the top level of a virtual collection, navigated through successive levels until they reached a non-virtual item and were thereupon removed from the context from which they began. In other words, when viewing a non-virtual documentary unit that was *discovered* through the context of a virtual collection, the discovery context should take precedent over that of the physical context.

The corollary of this situation is that, from an interface perspective (and indeed an interpretive one), the *identity* of a physical documentary unit viewed

within the context of a virtual collection is different from its canonical identity. This distinction is, in practice, manifested in two ways:

- The URL for an item viewed in a virtual context always encapsulates its discovery path, allowing perma-linking to permit sharing of items with explicit context.
- User interface components such as "breadcrumbs" display the discovery path, rather than an item's physical context.

As an example, compare the physical path to an item (of which there can be only one) to one of potentially many virtual discovery paths:

Physical path:

- Czech Republic → Terezin Memorial → Photographic & Film Material → Dr. Weiglovi

Example virtual paths:

- Terezin Collection → Research Guide → Dr. Weiglovi
- Notable Czechs → Dr. Weiglovi

In implementing virtual collections in a web interface our key concern was to ensure that an item viewed in a virtual context was a web "resource" like any other, and did not depend on maintaining browser-side state to determine the path a user had taken to arrive at a given page, where many such paths potentially exist. For this reason, the path to a virtual context is encoded into the URL for the page and should always be consistent and shareable if the virtual collection is publicly visible.

6 Technical implementation

While a detailed description of EHRI's technical implementation is beyond the scope of this paper, we include some details that may be of interest. Since a large portion of the data EHRI is integrating is in some form hierarchical (e.g. archival collections and subject vocabularies) a graph database was chosen as our primary data store [15] [5]. Archival descriptions are modelled as nodes, which are connected via edges to other descriptions, the repositories that hold them, and many other layers of the data model, such as researcher annotations and archival thesaurus terms. The advantages of this approach, in purely practical terms, include simple and fast traversal of potentially unbounded node paths (e.g. from leaf item to root) and the ability to easily incorporate new and revised assumptions in the data model due to the lack of an explicit database schema.

Hierarchical virtual collections are one area where we feel the graph model particularly shows its strengths relative to traditional relational databases. In particular, it provides a very low-overhead environment in which to manage and

[5] Specifically Neo4j (http://www.neo4j.org), which adheres to the pragmatic property-graph model, consisting of nodes, edges, and (typically scalar) property values which can be assigned to either.

reorganise tree structures without sacrificing performance when either navigating the hierarchy (for example, when traversing from an item-level unit to its top-level collection) or inserting new items (the typical trade-off in relational databases when employing optimisations such as the nested set model or adjacency lists.)

Our current implementation of hierarchical virtual collections *does* involve potentially expensive data retrieval queries. Due to the much more dynamic structure of VCs in comparison to the (largely static) underlying data, one particular implementation challenge is providing full-text search within the hierarchy of specific virtual collections. Our current approach depends on determining a set of the top-most virtual and non-virtual items within a given VC and applying a search constraint to the union of these items and their "descendants" (child, grandchildren, and so on.) This approach scales poorly, however, due to the unbounded number of items that can exist at each level of the VC hierarchy, and in future we may move to a system employing individual indexes specific to virtual collections.

7 Conclusion

We have outlined above the use of virtual collections in the EHRI project as a means to harmonise heterogeneous data from different archives, present coherent, thematically-based research guides, and to allow users to organise data in ways that best fit their research.

In section 2 we introduce the concept of virtual collections through some of the prior work and discussion addressing the topic in the context of archival integration and virtual research environments. Section 3 outlines the key characteristics of the data EHRI has encountered, including the fact that whilst hierarchical organisation is pervasive, data from individual archives varies greatly in both the depth of these hierarchies, the location of the descriptive detail within them, and the classification vocabularies used. Section 3.1 presents a case study of data integration in a fragmented archival environment

Section 4 describes the manner in which virtual collections can be used to lend coherence to thematic aggregations of data which span multiple physical archives, allowing differences in descriptive style to be harmonised via the use of synthetic groupings of material. Section 5 describes some of the representational challenges associated with this approach with regard to the provenance of items and the context in which they are discovered and viewed. We describe our attempts to mitigate these issues by including discovery context in our conception of a virtual collection "resource", providing an unambiguous handle to a given item within one of potentially many virtual contexts.

Finally, section 6 gives a brief overview of the technical architecture behind EHRI's platform and our plans for future work in this area.

Bibliography

[1] Reto Speck, Tobias Blanke, Cony Kristel, Michal Frankl, Kepa Rodriguez, and Veerle Vanden Daelen. The past and the future of holocaust research: From disparate sources to an integrated european holocaust research infrastructure. *arXiv preprint arXiv:1405.2407*, 2014.

[2] Tobias Blanke and Conny Kristel. Integrating holocaust research. *International Journal of Humanities and Arts Computing*, 7(1-2):41–57, 2013.

[3] Terry Cook. Evidence, memory, identity, and community: four shifting archival paradigms. *Archival Science*, 13(2-3):95–120, 2013.

[4] Tobias Blanke and Mark Hedges. Scholarly primitives: Building institutional infrastructure for humanities e-science. *Future Generation Computer Systems*, 29(2):654–661, 2013.

[5] Tobias Blanke, Leonardo Candela, Mark Hedges, Mike Priddy, and Fabio Simeoni. Deploying general-purpose virtual research environments for humanities research. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 368(1925):3813–3828, 2010.

[6] Melissa Terras. Digital curiosities: resource creation via amateur digitization. *Literary and linguistic computing*, page fqq019, 2010.

[7] Bradley D. Westbrook. Prospecting virtual collections. *Journal of Archival Organization*, 1(1):73–80, 2002.

[8] William E. Landis. Nuts and bolts. *Journal of Archival Organization*, 1(1):81–92, 2002.

[9] Alessandro Salvador. They're reading our minds: humanities research and digital thinking with cendari, 2013. Accessed: 2014-08-20.

[10] Leonardo Candela and Umberto Straccia. The personalized, collaborative digital library environment cyclades and its collections management. In *Distributed multimedia information retrieval*, pages 156–172. Springer, 2004.

[11] Michael Fraser. Virtual research environments: Overview and activity. *Ariadne*, (44), 2005.

[12] Guus Schreiber, Alia Amin, Lora Aroyo, Mark van Assem, Victor de Boer, Lynda Hardman, Michiel Hildebrand, Borys Omelayenko, Jacco van Osenbruggen, Anna Tordai, et al. Semantic annotation and search of cultural-heritage collections: The multimedian e-culture demonstrator. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6(4):243–249, 2008.

[13] Reto Speck and Petra Links. The missing voice. *International Journal of Humanities and Arts Computing*, 7(1-2):128–146, 2013.

[14] Wendy Scheir. First entry: Report on a qualitative exploratory study of novice user experience with online finding aids. *Journal of Archival Organization*, 3(4):49–85, 2006.

[15] Tobias Blanke, Michael Bryant, and Mark Hedges. Back to our data—experiments with nosql technologies in the humanities. In *Big Data, 2013 IEEE International Conference on*, pages 17–20. IEEE, 2013.