In-place Synchronisation of Hierarchical Archival Descriptions

Mike Bryant
Department of
Digital Humanities,
King's College London
United Kingdom
michael.bryant@kcl.ac.uk

Linda Reijnhoudt
Data Archiving &
Networked Services
The Hague
The Netherlands
linda.reijnhoudt@dans.knaw.nl

Boyan Simeonov
Ontotext
Sofia
Bulgaria
boyan.simeonov@ontotext.com

***Abstract*—This short paper describes work undertaken by the European Holocaust Research Infrastructure (EHRI) project to achieve reliable and repeatable harvesting of hierarchical archival metadata that is robust to structural changes and reorganisation of the source material.**

***Index Terms*—Archives, Structured data, Harvesting.**

I. Introduction

One of the goals of the European Holocaust Research Infrastructure (EHRI) project is to virtually integrate metadata about Holocaust-related archival collections in a way that makes it easier for historians and researchers to understand the provenance and consequence of fragmented, partial, and physically dispersed material. The project's main focus in these efforts is the EHRI portal[1], an online database containing archival metadata from over 500 institutions around the world. Whilst a substantial subset of collection descriptions in the EHRI portal have been authored by EHRI staff, the overall majority are derived from structured data provided by partner institutions, typically in the form of Encoded Archival Description (EAD) documents.

One area of focus within the project is creating links between material across different institutions, denoting, for example, the existence of shared provenance that could be of relevance to researchers. Copy collections, where duplicates of the same physical material exists in multiple locations, often with original finding aids, are common in Holocaust research due to the policies of institutions such as Yad Vashem, Memorial de la Shoah, and the United States Holocaust Memorial Museum (USHMM). Highlighting the existence of copies and/or alternative collection descriptions that might provide easier access for the user is one area where cross-institutional data integration can aid the research process.

[1]https://portal.ehri-project.eu

Attempting to integrate collection descriptions from many diverse institutions is challenging in several respects. Approaches to the creation of finding aids for archival material vary widely, notwithstanding the existence of conceptual standards providing guidance on the semantic level[2], and the technical resources necessary to provide this data to third-parties such as EHRI in structured form are usually very limited. This short paper focuses on one specific area: integrating hierarchical collection descriptions in a manner that is repeatable and robust. The next section explains why this presents a technical challenge in view of the contextual enrichment the project is attempting to provide.

II. Annotating hierarchical descriptions

Notwithstanding the various discussions about the suitability and usability of the traditional multi-level archival description format – fonds down to item level with no repeated information – in the digital environment (e.g. [1]), most of EHRI's partner archives catalogue their material in this manner and the project has opted to maintain this structure in its integrated database. Whilst this has entailed the tackling of considerable extra complexity verses an approach that either omitted descriptive levels beneath that of the fonds, or flattened hierarchies into a single record, it was deemed necessary due to the wealth of information often available in lower levels of the descriptive hierarchy depending on the style of finding aids created by the source institution.

When EHRI receives structured collection descriptions from partner institutions it is most commonly in a superset of the Encoded Archival Description (EAD) 2002 (see [2]) suitable for ingest into EHRI's database.

[2]Most notably, the General International Standard for Archival Description (ISAD(G)) and related standards from the International Council on Archives (ICA).

Once ingested, EHRI's database administration tools allow project staff to create connections between individual units of description at all levels, denoting such providential relationships as touched upon in the previous section. These relationships between units of description are thus additional metadata that is "layered" upon EHRI's proxy of the institution's original hierarchical finding aid.

There are two other cases where the EHRI portal layers additional information upon harvested collection metadata. The first is user-generated content in the form of public or private notes that users of the EHRI portal can apply to individual descriptive units (highlighting, for example, data inconsistencies or aspects of the material relevant to their research.) The second involves the forming of material into "virtual collections" – digital reconstructions of data with a shared theme or provenance – described in [3].

The difficulty arises when trying to keep EHRI's harvested proxy of its partner's archival metadata up-to-date, whilst preserving the additional "annotations" layered upon the latter in the form of connections, notes, and virtual finding aids. If the collection descriptions produced by EHRI's partner archives were purely static and never changed this would not be an issue, but in the real world, material can be reorganised or re-described in ways that significantly affect the structure of collection descriptions, sometimes involving hundreds or thousands of individual records. At the time of writing such reorganisation has occurred at several of EHRI's partner institutions, leaving the EHRI portal's data out-of-sync with its source.

A naive approach to this problem would be to simply expunge EHRI's entire proxy of an institution's metadata and replace it with the new version. Whilst appealingly simple, such an approach has significant downsides, chief among them being that metadata layered upon the original proxy becomes "orphaned" (detached from its target) or, in the worst case, refers to the wrong target. Administrative metadata relating to the digital proxy – telling the user when a description has been created and/or revised, for example – is also lost in this case. An additional issue with deletion/re-ingest of an entire fonds is that without knowing which individual descriptive units may have been moved, re-parented, or deleted, we have no way to maintain URLs for such items, potentially producing a lot of broken links and compromising the citability of data on the EHRI portal.

The next section briefly reviews some related problems and technical approaches.

III. Related problems and technical approaches

EHRI's need to preserve layered metadata on proxies of change-prone hierarchical descriptions is conceptually similar to the more general problem of annotation in the web environment, where the hierarchy is the tree structure of HTML (or XHTML) documents that have a URL (assumed to be persistent) but no guarantee of immutability. Instead of the targets in this case being individual items (pages with their own URL), they are instead parts of the page, identified by a specific HTML node, such as a paragraph tag.

A related, more focused area of research — and one with particular relevance to EHRI due to its use of EAD — is on annotating and detecting changes in XML documents, where (due to well-formedness guarantees) each node is identifiable by a specific XPath. While a full review of the literature would take up too much space in this short paper, Horo et al. [4] provide a good overview of web annotation issues and the robustness of XPath-based annotation, with one of the conclusions being that XPath alone (and by extension the native structure of XML itself) is not by itself sufficient to maintain robust content-level annotation.

This limitation has important implications for EHRI, since it follows that, given two EAD documents describing the same fonds in a hierarchical manner, one with structural modifications relative to the other, we cannot use XML semantics alone to determine what has changed and how, in order to re-target our annotations accordingly. If a set of descriptive units representing files within a series have been relocated to a sub-series — that is, from one `<c>` node to a more deeply-nested `<c>` node — we cannot tell them apart from newly-created items.

In addition to XML semantics, however, we can also lean on EAD semantics to help us concretely identify descriptive units in a way that persists across structural changes. Just as the ISAD(G) field 3.1.1 prescribes reference codes that "identify uniquely the unit of description", EAD has `<unitid>` fields for this purpose. Problematically, however, these reference codes (including those given as examples in the ISAD(G)) are themselves often derived from an item's place in the descriptive hierarchy, and correspondingly subject to change when that that hierarchy changes. If the identifiers are not persistent and change when reorganisations occur, we cannot use them to track material across changes.

It follows, then, that the bare minimum requirement for EHRI to understand and mirror structural changes in its archival description proxies is that unit identifiers within the source descriptions are both unique (within the entire scope of the sync operation) and persistent

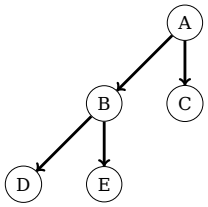

Fig. 1. A tree of descriptive units.

Institution native ID (NID)	EHRI hierarchical ID (HID)
A [fonds]	gb-001234-A
B [series]	gb-001234-A-B
D [item]	gb-001234-A-B-D
E [item]	gb-001234-A-B-E
C [series]	gb-001234-A-C

TABLE I
Example HID generation for a UK-based institution with EHRI numerical identifier "001234"

across structural changes. Fortunately, we have found this to be the common case in data received from our partner institutions, since the unit identifiers typically derive from database primary keys used by internal metadata management software, or can be derived thus for the purposes of data transfer. The next section discusses how EHRI derives its own hierarchical identifiers from institution-specific — "native" — identifiers.

IV. Hierarchical identifier generation

EHRI's primary means of identifying units of description at all levels is a hyphen-delimited hierarchical identifier (HID) derived from:

1) the ISO 3116 alpha-2 country code of its holding institution
2) an EHRI-derived numerical institution identifier
3) the native unit identifier(s) (NID) used by the original material

In deriving the HID from institution-endowed identifiers further transformations are done to remove shared prefixes (common in identifiers that "extend" that of the parent item) and non-URL-safe characters. The HID is thus a one-way transformation, as shown in table I, i.e. we cannot convert from the HID back to native IDs.

The most visible manifestation of the HIDs is their use in URLs in the EHRI portal, and as the primary handle for retrieving items from one of the EHRI APIs. Not only does it allow us to identify items uniquely within their containing scope (the institution, for a top-level descriptive unit, or the parent unit for one in a lower level), the HID is generally readable and contains information about an item such as its position within the hierarchy.

The corollary to being derived directly from third-party information is that EHRI HIDs are also subject to

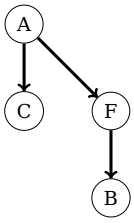

Fig. 2. Tree of descriptive units to be ingested.

change: if an institution reorganises their finding aids to place a given (uniquely-identified) unit elsewhere the descriptive hierarchy, EHRI's HID-generation algorithm will produce a different result.[3]

V. Synchronisation procedure

EHRI's automated procedure for synchronising archival descriptions is transactional (either succeeding completely or leaving the data untouched) and can operate at one of two levels:

1) all descriptive units within a given fonds
2) all archival descriptions held by a given institution

The first scenario is intended for ad-hoc updating of a single collection from data contained within one EAD file. The second is designed for periodic, automated synchronisation of all material belonging to a given archive (including top-level collections that may potentially have been removed) and works using multiple EAD files as input data.

1) Pre-ingest: The synchronisation procedure begins prior to data ingest by defining a bijective function H from existing NIDs to existing HIDs $H : N \to E$ within the subject *scope* (EHRI's proxy of the fonds or the archive), so for every $\nu \in N$ there is exactly one $\varepsilon \in E$ and vice versa. See Table I for the value pairs (ν, ε) for our example tree, as shown in Fig 1.

2) Ingest: During data ingest we check that NIDs for each descriptive unit are unique within the sync scope. If they are not the process is aborted. During ingest new HIDs are generated as described in section IV, forming the input for our post-ingest NID-to-HID bijective function $H' : N' \to E'$. In our example, we are ingesting the archival tree shown in Fig 2.

3) Post-ingest: At this point, we have a mix of old and new archival descriptions side-by-side within our database, some representing duplicate nodes. All descriptive units exist in one of four states:

[3]We thus have the paradoxical situation that whilst EHRI relies upon partner institutions to use unique identifiers for each unit within each of their fonds', we ourselves generate identifiers that relate strictly to the hierarchy and are thus non-persistent, a situation that could create issues for third-parties seeking to harvest EHRI's data in the same repeatable manner that we seek to do from our source archives.

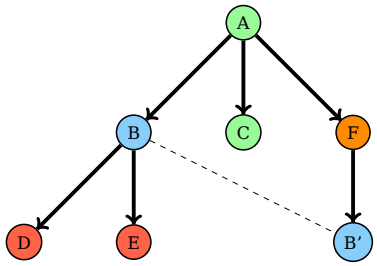

Fig. 3. Tree representing the descriptive units after ingest of the new tree with structurally modified data and prior to cleanup.

Static The unit has remained in the same place in the descriptive hierarchy, though its contents may have changed.[4] This applies to $\{A, C\}$ in Fig 3

$$\{\nu \in N \cap N' | H(\nu) = H'(\nu)\} \tag{1}$$

Moved The unit has been moved from one place in the descriptive hierarchy to another, with contents that may have also changed. In the tree these units appear twice. This applies to $\{B, B'\}$, with the duplication marked as a dashed line.

$$\{\nu \in N \cap N' | H(\nu) \neq H'(\nu)\} \tag{2}$$

Created The unit has been newly-created in the revision of the descriptive hierarchy. This applies to $\{F\}$.

$$N' \setminus N \tag{3}$$

Deleted The unit has been removed from the descriptive hierarchy. This applies to $\{D, E\}$.

$$N \setminus N' \tag{4}$$

Fig 3 represents this intermediate state, with nodes A and C being static (albeit with properties that might have been updated), node B being re-parented beneath new node F and its child nodes D and E being deleted.

Having determined which proxy items within the hierarchy represent items that have been relocated from one place to another we can then proceed to transfer the EHRI-specific metadata referenced in section II, re-targeting connections, user annotations, and membership of virtual collections from the original node to the new one. Finally, the original node is marked for deletion along with any items that have been completely removed from the hierarchy.

The synchronisation procedure generates a report which tells us how many items were created anew, updated, or left unchanged, plus a set of old HIDs mapped to new HIDs representing relocated items. Using this information we can perform some additional cleanup tasks at levels closer to the user interface of the EHRI portal, including the generation HTTP permanent redirects from URLs containing old HIDs to the new ones, and updating the portal's search engine for the given sync scope.

[4]A form of metadata versioning is incorporated into these in-place property updates, although this does not represent a true snapshot in the version-control sense.

VI. Conclusion

This short paper has reported on the approach the EHRI project has adopted for synchronising proxies of hierarchical archival descriptions in a manner that makes it possible to preserve additional metadata added to enrich and provide cross-institutional context to the source information. Section II has described the problem and explained why, in a purely digital environment, the project is endeavouring to tackle the complexity that descriptive hierarchies involve. Section III has briefly reviewed related (and somewhat broader) problems, and explained why it is only possible to do (unsupervised) restructuring of data hierarchies if the source data meets certain criteria, namely contains identifiers for each descriptive unit that are unique within the scope of the sync operation and persistent between restructuring operations. Section IV describes the technique EHRI uses to generate hierarchically-aware identifiers for individual descriptive units within the portal database. Finally, section V describes the transactional synchronisation procedure and the manner in which we determine which descriptive units have been freshly created, moved from one place to another, or deleted, allowing us to maintain intra-fonds contextual metadata added by the project and its users. At the time of writing the technique described above has been employed successfully by EHRI to synchronise a number of large fonds and institution holdings and represents a step towards robust automated archival interoperability.

References

[1] Sarah Higgins, Christopher Hilton, and Lyn Dafis. Archives context and discovery: Rethinking arrangement and description for the digital age. In *2nd annual conference of the International Council on Archives*, pages 11–15, 2014.

[2] Laurent Romary and Charles Riondet. Ead odd: a solution for project-specific ead schemes. *Archival Science*, 18(2):165–184, Jun 2018.

[3] Mike Bryant, Linda Reijnhoudt, Reto Speck, Thibault Clerice, and Tobias Blanke. The ehri project - virtual collections revisited. In Luca Maria Aiello and Daniel McFarland, editors, *Social Informatics*, pages 294–303, Cham, 2015. Springer International Publishing.

[4] Masahiro Hori, Mari Abe, and Kouichi Ono. Robustness of external annotation for web-page clipping: Empirical evaluation with evolving real-life web documents. In *International Conference on Knowledge Capture*, pages 65–72, 2003.