Archival Metadata Import Strategies in EHRI

Francesco Gelati

Belgian State Archives

Introduction

The EHRI portal aims to aggregate digitally available archival descriptions concerning the Holocaust. This portal is actually a meta-catalogue, or an information aggregator, whose biggest goal is to have updated information by means of building sustainable data pipelines between EHRI and its content providers. Just like in similar archival information aggregators (e.g. Archival Portal Europe or Monasterium), Encoded Archival Descriptions (from now, EADs) play a key role. EADs are the leading international standard for describing archival holdings in XML language and for digitally sharing archival information. EADs follow the ISAD(G) (General International Standard Archival Description) standard developed by the International Council of Archives (ICA).

Both proprietary and free software for describing archives are able to generate EADs from a given set of record. This is why EHRI gently asks partner institutions not to share .pdf inventories or finding aids, but to convert them, eventually by means of the the costless software ICA-AtoM, into EAD files. EAD files are indeed the only file format that can be ingested, both manually or automatically, in the portal. A manual one-time one-off ingest is always possible, but for big amounts of EAD files an automated ingest procedure was developed by means of the Open Archive Initiative.

EADs: from variety to uniformity

EADs 2002 (from now, simply EADs) are the most used, although not the most recent, EAD version. They have an extremely flexible structure, which represents both an advantage and a disadvantage. The uniqueness of an archival fonds can be described by a flexible schema with a higher degree of liability. On the other hand, the EAD files of two sets of records having the same hierarchical structure may contain the same piece of information in two different positions. I am not simply referring to a different linear order (piece of info A comes before B or vice-versa), but to a metadata embedded in different sections of the file creating a different, equally valid, information hierarchy. Another problem is represented by information

granularity: if we take the ISAD(G) field “physical description” as example, it can be noticed that some EHRI partner institutions put all details in the EAD level <physdesc>, whereas others go deeper and distinguish the physical facet <physfacet>, the <genreform> and the <dimensions>. In other words, both

```
<physdesc>3 typewritten files, 0,1 linear meters</physdesc>
```

and

```
<physdesc>3
        <physfacet>typewritten</physfacet>
        <genreform>files</genreform>
        <extent>0,1 linear meters</extent>
</physdesc>
```

are valid and equally represented.

Permanent identifiers (PIs from now) are a crucial issue. Sometimes partner institutions do not give a PI to each and every item of every level of description; sometimes this PI may be very long (more than 30 characters), thus not suitable. EHRI and the partner institution have to find an ad-hoc solution.
An .xml conversion matrix, named property file, is the solution for both PIs and EAD diversity. It defines on a given dataset (namely the dataload from an institution) which EAD fields will be shown where in the EHRI portal. It makes then possible to have a uniform result from the above-mentioned different describing practices, and to choose an EAD field (or tag) to be used as PI. A majority of EAD datasets has its own property file in order to get the best outcome and eventually to meet the institution’s wishes. For smaller number of institutions it is also possible to use the general property file.

Before the EADs: converting spreadsheets to EADs

What if an institution wants to export its metadata to the EHRI portal, but its collection management system does not generate EADs?
EHRI developed a free tool: the EHRI EAD Creation Tool. It allows to create an EAD file from an input file that may be both a .csv or an .xml one. In both cases a mapping file, saying which input field corresponds to which output field, is needed. The mapping file will also define

invariable information, such as the name of the institution, that will ends up in the first half of the EAD file, the <eadheader>.

Mapping information from an .xml file to an EAD-xml file might sound easier: but it is without considering that it may be necessary to modify the hierarchy in the information. A .csv input file has been so far more frequently adopted. If the institution can provide a set of spreadsheets, where each sheets is about one fonds (or collection), each row describes one item, and each column defines a ISAD(G)-compliant metadata field, the tool can be successfully run, the conversion can take place, and the EAD files are generated.

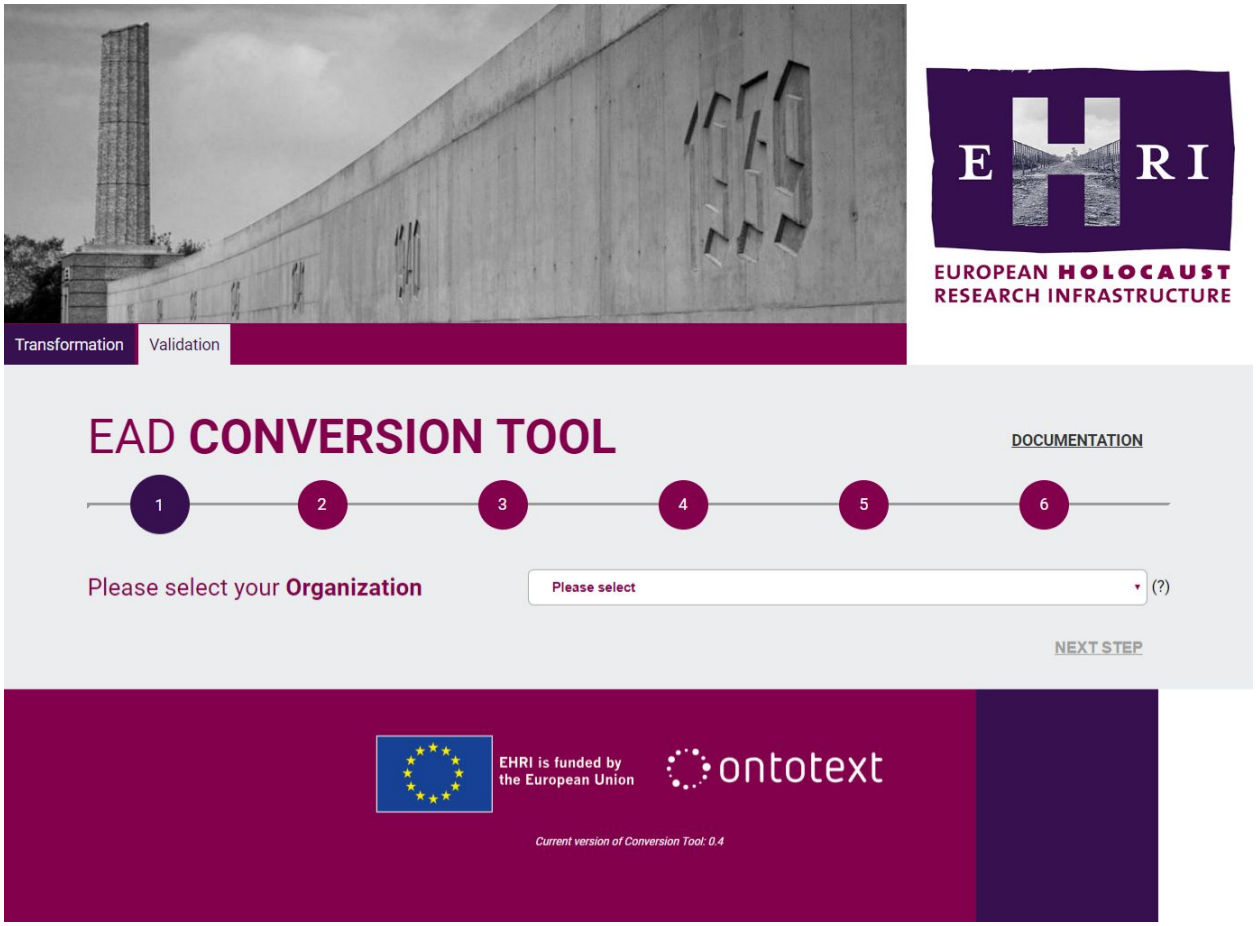

Fig. 1: EAD Conversion Tool, screenshot

OAI protocols: publishing the EADs on the web

«Once we have valid EAD files, I will send you the outcome and you will ingest it in the EHRI portal» may propose the partner institution collection manager. «No please» would reply EHRI, struggling for establishing sustainable connections and for avoiding one-time one-off imports.

A sustainable connection is by all means the ultimate goal of each import strategy: it is the only way both to have an automated import procedure, and to be sure that data may be easily and quickly updated. Some institutions publish their EADs on their servers on the web by means of the Open Archive Initiative - Protocol for Metadata Harvesting (from now, OAI-PMH): this endpoint webpage may be harvested by EHRI harvesting tool.

If the institution does not have an OAI-PMH endpoint, a second tool, named Metadata Publishing Tool, was developed by EHRI to permit the institution to publish EADs on its servers according to the Open Archives Initiative ResourceSync Framework Specification (from now, OAI-RS). The OAI-RS is more recent than the OAI-PMH, has more features, but they are both based on the same principles, and give a compatible outcome. This is why the OAI-RS was adopted instead of the OAI-PMH by the Metadata Publishing Tool.

EHRI can harvest data anyhow published according to the OAI-RS: an institution may of course decide to autonomously create its OAI-RS endpoint.

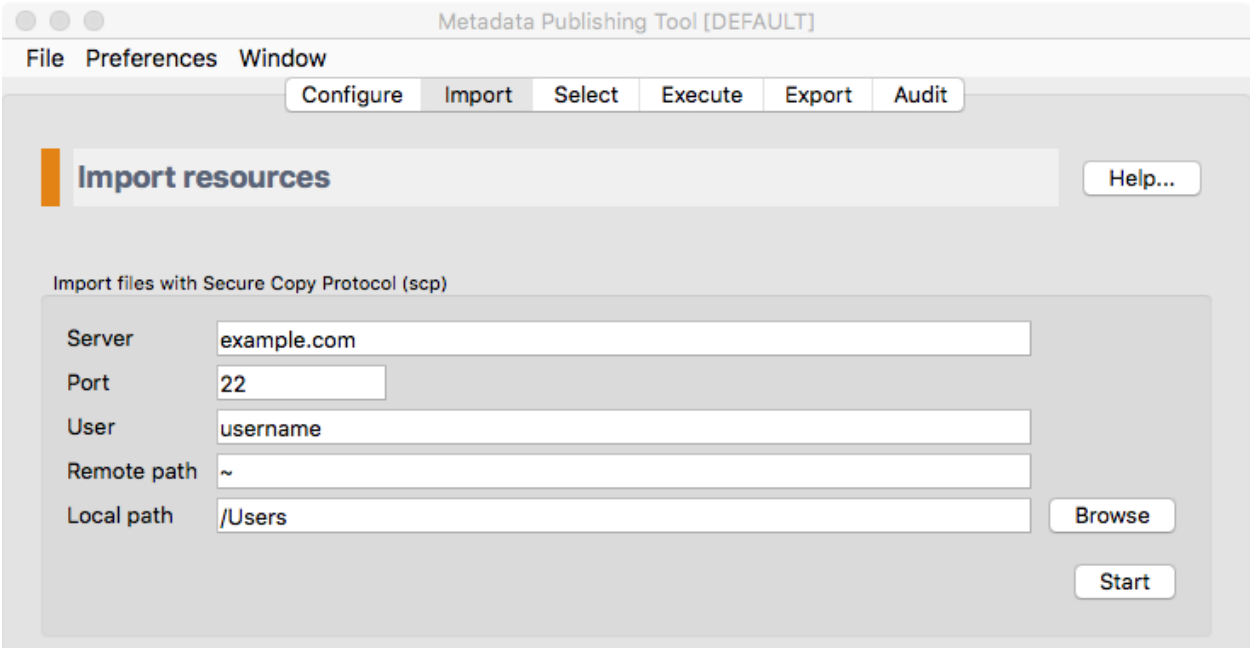

Fig. 2: Metadata Publishing Tool, screenshot

Four import strategies

In order to have a more systematic overview of the archival metadata import strategies, the reader can find below EHRI import workflows, based on the different metadata conversion and

publication variables. Four types of institutions are represented:

- type A: this institution can export metadata in the EAD format and supports the OAI-PMH, so the EHRI harvester can automatically gather the metadata from the CHI.
- type B: this institution supports the OAI-PMH harvesting protocol. The metadata is, however, not available in the EAD format. Metadata needs to be converted into EAD, EHRI suggests to use the EAD Conversion Tool.
- type C: this institution does not have EADs and does not publish its metadata in an OAI protocol (OAI-PMH nor OAI-RS). EHRI suggests to use both EHRI tools.
- type D: this institution can export metadata in the EAD format, but does not have an OAI-PMH nor an OAI-RS endpoint. EHRI suggests to create an OAI-RS endpoint by means of the Metadata Publishing Tool.

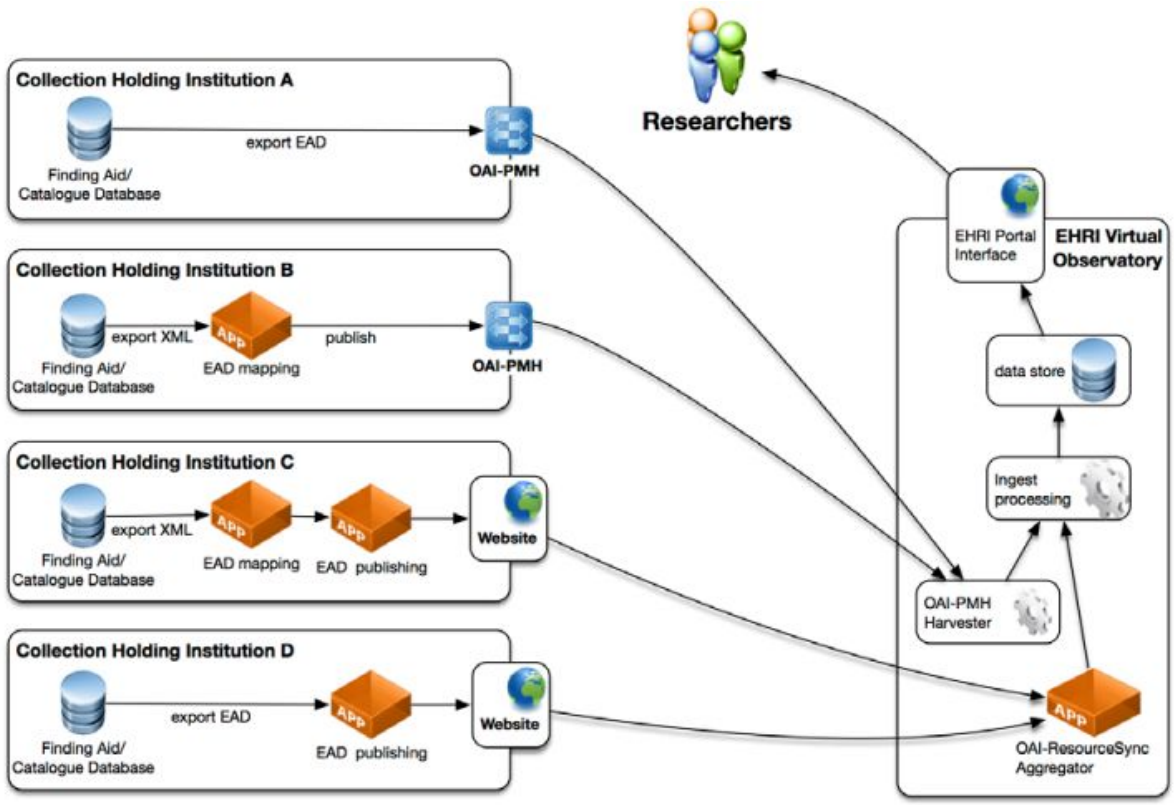

Fig. 3: EHRI data import workflows, from Henk van den Berg and Boyan Simeonov (2017).

Conclusions

The described workflows create a sustainable connection requiring no intervention from the EHRI-side and little or no manual intervention from the content provider side. The latter may in fact set a script regularly running one or both tools, thus refreshing the output, or may prefer to run user interfaces on a desired time frequence. The length of the paper did not allow me to present more specific features or case studies, nor to deeper describe the rich variability of solutions concerning descriptive metadata in the EADs. I hope that this overall view could still stress the benefits of a sustainable connection in comparison to (often quicker) one-off one-time ingests.

Credits

The EAD Conversion Tool was developed by Ontotext for EHRI; the Metadata Publishing Tool was developed by DANS (Data Archiving and Networked Services) for EHRI.

Bibliography

Laura Brazzo and Reto Speck, "Introduction" in *Quest. Issues in Contemporary Jewish History. Journal of Fondazione CDEC*, 13, 2018 (thematic issue *Holocaust Research and Archives in the Digital Age*, edited by Laura Brazzo and Reto Speck). http://www.quest-cdecjournal.it/

Veerle Vanden Daelen, Jennifer Edmond, Petra Links, Mike Priddy, Linda Reijnhoudt, et al., “Sustainable Digital Publishing of Archival Catalogues of Twentieth-Century History Archives” in *Open History: Sustainable digital publishing of archival catalogues of twentieth-century history archives*, Brussels, 2016. https://hal.inria.fr/hal-01281442

Henk van den Berg and René van Horik. European Holocaust Research Infrastructure Deliverable 10.2: Collection Description Publishing Services. S.L, 2017. https://ehri-project.eu/sites/default/files/downloads/ehri_downloads/D10%201%20Collection%20Description%20Production%20Services.pdf

Henk van den Berg and Boyan Simeonov. European Holocaust Research Infrastructure Deliverable 10.1: Collection Description Production Services. S.L., 2017. https://ehri-project.eu/sites/default/files/downloads/Deliverables/D10%202%20Collection%20Description%20Publishing%20Services.pdf

Sitography

Last consulted on 12 September 2018

http://portal.ehri-project.eu/

http://monasterium.net/
http://www.archivesportaleurope.net/
http://github.com/EHRI/manuals/tree/master/ECT
http://rspub-gui.readthedocs.io/
http://www.openarchives.org/pmh/
https://github.com/EHRI/ehri-rest/tree/master/ehri-io/src/main/resources
http://www.openarchives.org/rs/
http://eadiva.com/2/
https://ehri-project.eu/ehri-for-institutions#Automated