HAL
open science

Archival Metadata Import Strategies in EHRI

Francesco Gelati

To cite this version:

Francesco Gelati. Archival Metadata Import Strategies in EHRI. Archives et Bibliothèques de Belgique - Archief- en Bibliotheekwezen in België, In press, Trust and Understanding: the value of metadata in a digitally joined-up world, ed. by R. Depoortere, T. Gheldof, D. Styven and J. Van Der Eycken (106), pp.15-22. hal-02124632

HAL Id: hal-02124632

https://hal.science/hal-02124632v1

Submitted on 9 May 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

ARCHIVAL METADATA IMPORT STRATEGIES IN EHRI

Francesco GELATI

Introduction[1]

The *European Holocaust Research Infrastructure* (EHRI) portal website 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 up-to-date information by means of building sustainable data pipelines between EHRI and its content providers, which are mostly collection holding institutes. Just like in similar archival information aggregators (e.g. *Archives Portal Europe* or *Monasterium*), *Encoded Archival Descriptions* (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 *General International Standard Archival Description* or ISAD(G) developed by the *International Council on Archives* (ICA).

Both proprietary and open-source software for describing archives are able to generate EADs from a given set of records. In order to operate a standardized ingest strategy, EHRI requests partner institutions not to share .pdf inventories or finding aids, but to convert them, eventually by means of 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 EHRI 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. This EAD type has an extremely flexible structure, which presents us with both an advantage and a disadvantage. On the one hand, the uniqueness of an archival fonds can be described via 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

---

[1] All images are covered by intellectual property rights belonging to the EHRI consortium and to their respective authors. 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. The author especially wishes to thank the reviewers and the editorial board of the publication.

at two different locations. I am not simply referring to a different linear order (piece of information 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. A second problem is information granularity: if we take the ISAD(G) field *physical description* as an example, we notice how some EHRI content providers present all details in the EAD tag <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.

Persistent and unique identifiers (PIs from now) are a crucial element for creating valid EAD and for optimizing the automatic ingest. Some institutions do not give a PI to every item or every description level of each collection; sometimes the PI may be very long (more than 30 digits), or filled-in with special characters, and thus error-prone. PIs played for instance an important role when EHRI first engaged with Kazerne Dossin - *Memorial, Museum and Documentation Center on Holocaust and Human Rights* in Mechelen, Belgium. Kazerne Dossin would become the pilot partner for the automatic ingest process, but Kazerne Dossin also faced the broader challenge of adopting standardized archival workflows and developing its digital accessibility based on the needs identified by EHRI. In 2013 Kazerne Dossin still worked with the item-level catalogue developed in 1994 for its predecessor, the *Jewish Museum of Deportation and Resistance*. This catalogue was much closer to a registrar's output than an archival inventory. Although every item in the catalogue had its own unique PI, collection descriptions (and therefore PIs for collections) were never created. All researchers depended on the item level descriptions to find what the needed. No contextual information was provided. Thanks to EHRI's support, Kazerne Dossin undertook the important step of creating collection level descriptions and assigning PIs to all of them, which led to «the double advantage of (1) allowing to process and provide information on archival collections much faster (since writing descriptions on a collection level are considerably less time-intensive than on item level) and (2) allowing more broadly

sharable descriptions as collection-level faces fewer challenges on privacy regulation than item level descriptions»[2].

Once collection level descriptions (compliant with ISAD(G)) were introduced at Kazerne Dossin, EHRI assisted in extracting valid EAD from the software used by Kazerne Dossin and make the .xml files EHRI portal compliant. An .xml conversion matrix, *named property file*, is the solution for both PIs and EAD diversity. It defines on a given dataset (namely the data load from an institution) which EAD fields will be shown where in the EHRI portal. It allows then 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 a smaller number of institutions a general property file was used.

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 valid EADs?

EHRI developed a free tool: the *EAD Conversion Tool*, or ECT. This tool creates an EAD file from an input file that may be both a .csv or an .xml file. In both cases a mapping file, indicating 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 end up in the first half of the EAD file, the <eadheader>. Again, Kazerne Dossin was an important case study for EHRI. Comprehensive and consistent digitally-available archival descriptions allowed Kazerne Dossin to transform by means of the EAD Conversion Tool the raw xml-export from its collection management system into valid EAD files. Mapping information from an .xml file to an EAD-xml file might sound easier. However, we must take into account that it may be necessary to modify the hierarchy in the information. A .csv input file has so far been adopted more frequently. If the institution can provide a set of spreadsheets with one fonds (or collection) per sheet, each row describes one item, and each column defines an ISAD(G)-compliant metadata field, the EAD conversion tool can run successfully, the conversion can take place, and the EAD files are generated.

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, is what the partner institution collection manager might say. *No thank you*

[2] D. Styven, M. Caragea, V. Vanden Daelen, "The Learning Curve in Sharing Data with the EHRI Project: The Example of a Memorial Site, Kazerne Dossin, Mechelen", in *EHRI Document Blog*, 2018. https://blog.ehri-project.eu/2018/06/19/kazerne-dossin-mechelen.

would be EHRI's reply, struggling to establish sustainable connections and to avoid one-time one-off imports.

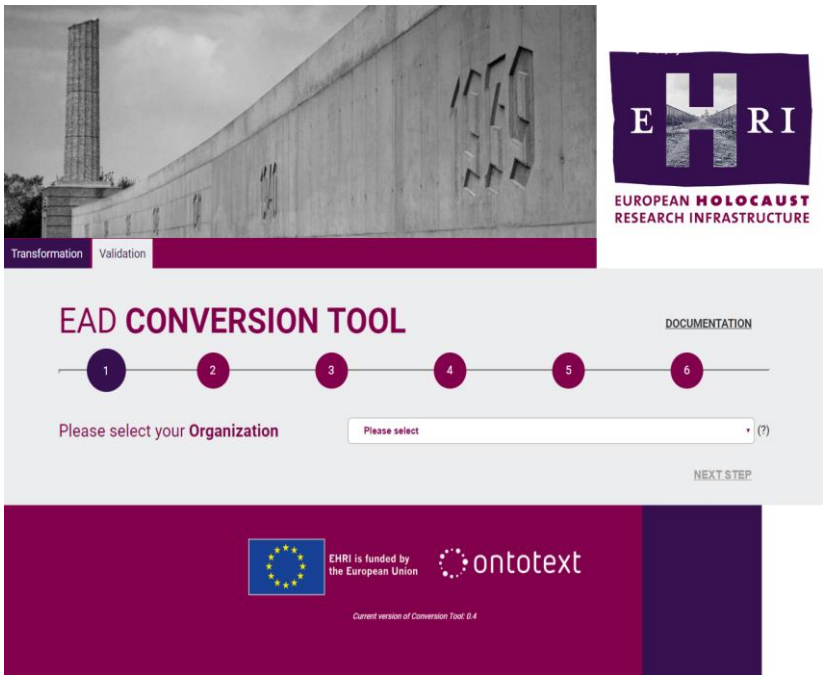

Fig. 1: EAD Conversion Tool, screenshot

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

If the collection holding institute does not have an OAI-PMH endpoint, a second tool, the *Metadata Publishing Tool*, also developed by EHRI, permits the institute to publish EADs on its server according to the *Open Archives Initiative ResourceSync Framework Specification* (from now, OAI-RS). In our example of Kazerne Dossin, this institute took the opportunity to share online its archival descriptions by means of the OAI-RS protocol, building a standard-compliant web-page freely accessible to the whole academia: Kazerne Dossin's partner institutes and other projects can now harvest and import the latest version of every collection description in .xml form because of the EAD Conversion Tool. The files are automatically updated at the endpoint every time Kazerne Dossin's IT team runs the *EAD Conversion Tool*. The OAI-RS offers more features than the OAI-PMH,

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 during the creation of the *Metadata Publishing Tool*.

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

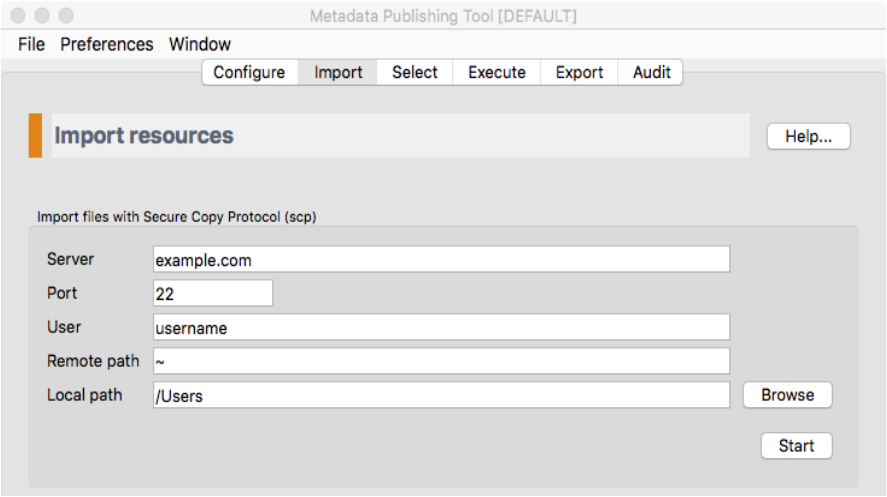

Fig. 2: Metadata Publishing Tool, screenshot

Four import strategies

In order to achieve a more systematic overview of the archival metadata import strategies, the reader can find below a visualisation of EHRI's import workflows, based on the different metadata conversion and publication variables.

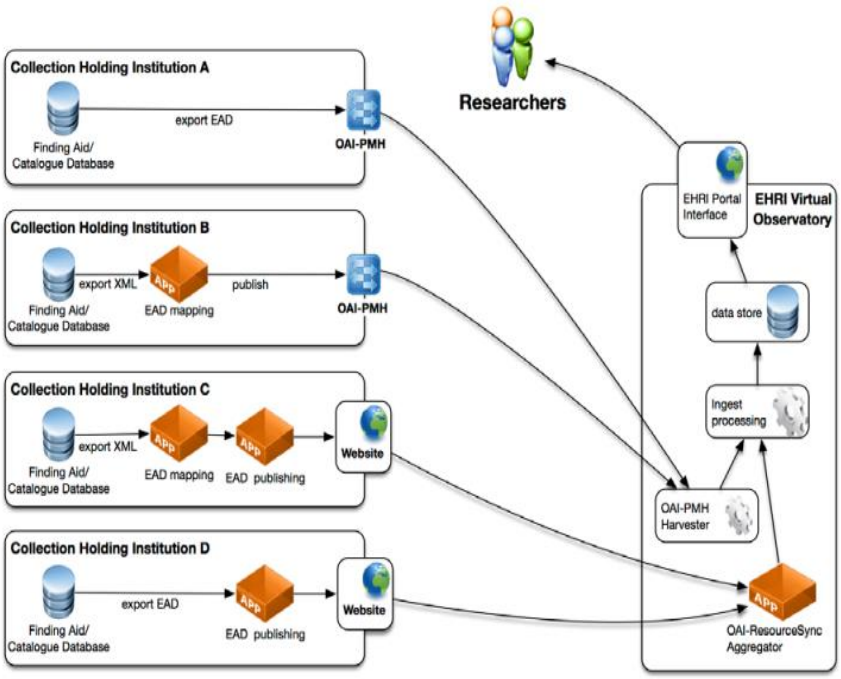

Fig. 3: EHRI data import workflows,
from Henk van den Berg and Boyan Simeonov (2017).

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 collection holding institute.
- type B: this institution supports the OAI-PMH harvesting protocol. The metadata is, however, not available in an 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.

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 to regularly run one or both tools, thus refreshing the output, or the provider may prefer to run the tools' user interfaces on a desired time frequency.

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/
- Dorien Styven, Marius Caragea, Veerle Vanden Daelen, "The Learning Curve in Sharing Data with the EHRI Project: The Example of a Memorial Site, Kazerne Dossin, Mechelen" in *EHRI Document Blog*, 2018. https://blog.ehri-project.eu/2018/06/19/kazerne-dossin-mechelen/
- 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
- https://blog.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