



**European Holocaust Research Infrastructure
H2020-INFRAIA-2014-2015
GA no. 654164**

Deliverable 10.1

Collection Description Production Services

**René van Horik
DANS-KNAW**

**Boyan Simeonov
ONTOTEXT**

**Start: May 2015 [M1]
Due: October 2016 [M18]
Actual: April 2017 [M24]**



[EHRI is funded by the European Union](#)

Document Information

Project URL	www.ehri-project.eu
Document URL	n.a.
Deliverable	D10.1 Collection Description Production Services
Work Package	WP10
Lead Beneficiary	P1 DANS-KNAW
Relevant Milestones	MS1 + MS2
Dissemination level	Public
Contact Person	René van Horik / rene.van.horik@dans.knaw.nl / +31623297389
Abstract (for dissemination)	This deliverable describes the “EAD Creation Tool” Software. The aim of the software is to convert metadata in local format into the EAD metadata format. The conversion is based on a mapping between the local format and the EAD format.
Management Summary	n.a.

Table of Contents

1	Introduction	4
1.1	The EAD Creation Tool and the EHRI data infrastructure.....	5
1.2	Downloading the ECT software	6
2	Overview of the ECT Software	7
2.1	Requirements.....	7
2.1.1	System Requirements	7
2.1.2	Licensing	7
2.2	Run the ECT software as a desktop installation	7
2.2.1	On Windows OS.....	7
2.2.2	On Unix OS	7
2.2.3	On Mac OS.....	8
2.3	EAD system HOME directory	8
2.4	Input directory	8
2.5	Mapping directory.....	8
2.6	Output directory.....	8
2.7	XQuery directory	8
3	Using the ECT service.....	9
3.1	Add the data files you want to transform in the ~/input directory.....	9
3.2	Open http://localhost:8080 in a browser	9
3.3	Select your organization from the drop-down list	9
3.4	Select your files INPUT format	10
3.5	Select the transformation type.....	10
3.6	Preview/edit the mapping config file	10
3.7	Preview the input files	11
3.8	Start transformation.....	12
3.9	Explore the conversion report.....	12
3.10	Exploring the EAD validation inconsistencies	13
3.11	Validate the corrections	14

1 Introduction

Deliverable 10.1 is described in the DoA as follows “*The collection description production services will be an adaption of an existing open source tool. The production services will have a user-interface to map fields in the output of the archive’s database to those defined in the EHRI metadata guidelines.*”

The service developed is a software tool that converts metadata in a local format into metadata in the EAD format¹. This conversion is based on a mapping of metadata elements of the local format into metadata elements of the EAD standard. The target user community of the service are CHIs that have an information system to create and manage finding aids that are not formatted according to the EAD standard. The information system of the CHI must be able to export metadata. This export file is the input file of the ECT software.

The name of the software tool is “EAD Creation Tool” abbreviated as ECT. The result file created by the tool can be ingested in the EHRI portal. The software is developed by EHRI-partner Ontotext.

¹ EAD stands for Encoded Archival Description, and is a non-proprietary de facto standard for the encoding of finding aids for use in a networked (online) environment. Finding aids are inventories, indexes, or guides that are created by archival and manuscript repositories to provide information about specific collections. While the finding aids may vary somewhat in style, their common purpose is to provide detailed description of the content and intellectual organization of collections of archival materials. EAD allows the standardization of collection information in finding aids within and across repositories.

The EAD Metadata Schema and information related to the EAD standard can be found at: <https://www.loc.gov/ead/> [cited 24 March 2017].

1.1 The EAD Creation Tool and the EHRI data infrastructure

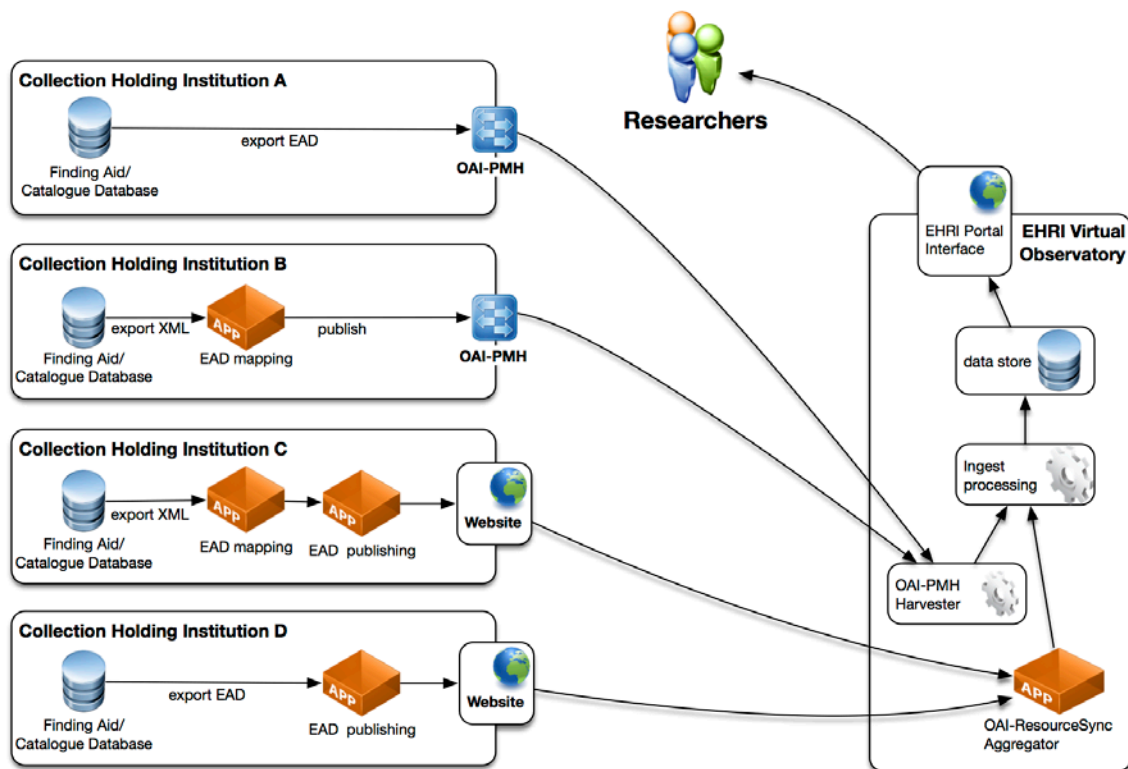


Figure 1. EHRI data infrastructure

The EHRI project has developed software services to assist the data integration process. The software services are represented by the orange boxes in Figure 1.

To what extent the service is usable for a CHI depends on the way the local data infrastructure is organised. E.g whether metadata on archival holdings are available in a digital form, its format and how the information infrastructure is able to communicate with the outside world.

Two standards form the core of the data integration workflow: (1) the EAD metadata format for archival finding aids (see footnote 1) and (2) the OAI-PMH protocol for metadata harvesting².

Within EHRI we make a distinction between several types of information infrastructures.

- CHI type A = CHI that can export metadata in the EAD format and supports the OAI-PMH metadata harvesting protocol, so the EHRI harvester can automatically gather the metadata from the CHI.
- CHI type B = This CHI supports the OAI-PMH harvesting protocol. The metadata itself, however, is not available in the EAD format. A local format is used that can be exported in XML. A tool is available to convert the local metadata format into EAD. This is the EAD Conversion Tool
- CHI type C = This CHI does not have metadata available in a local format. So the metadata has to be converted to the EAD standard. For this a tool is available. The CHI does not have an OAI-PMH data provider installed. EHRI has developed a metadata publisher (the Metadata Publishing Tool (MPT-tool), covered in Deliverable D10.2) that implements the ResourceSync Framework. This framework describes a

² See: <<https://www.openarchives.org/pmh/>> [cited 14 July 2017]

synchronization framework for the web that allows third-party systems to remain synchronized with a server's evolving resources.

- CHI type D: is capable of exporting metadata in EAD format, but does not have a OAI-PMH service. So it also needs the MPT-tool (that supports the ResourceSync framework).

The EHRI portal contains a service that integrates the metadata provided by the CHIs (type C and D). This service is part of the EHRI portal.

The ECT software carries out the “EAD mapping” actions as part of the use cases “Collection Holding Institution A” and “Collection Holding Institution B” in figure 1. The EAD mapping activity is a component of a data integration chain.

1.2 Downloading the ECT software

The software can be downloaded from: ftp://ftp.ontotext.com/pub/EHRI/conversion_tool/

Use the “Guest” FTP account.

The software and documentation is available in a .zip file. The name of the zip-file is “conversion_tool-x.x.zip” (where x.x. is a version number)

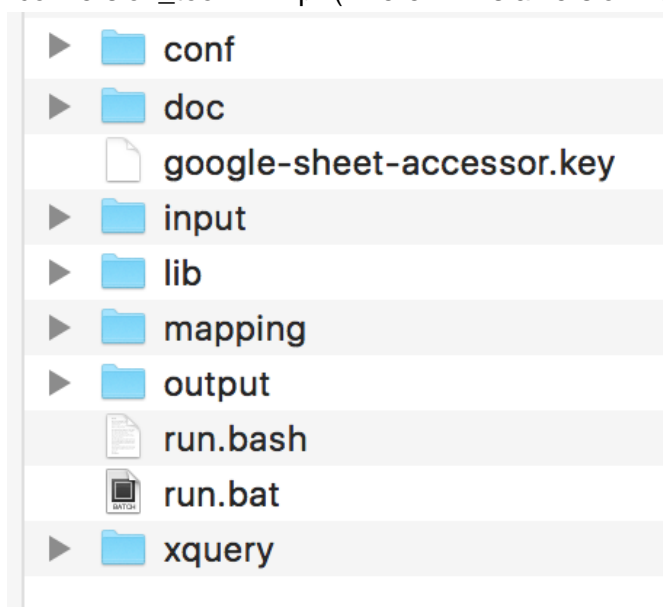


Figure 2. Directories created after downloading the .Zip file.

Figure 2 shows the directory structure after opening the .Zip file

2 Overview of the ECT Software

The EAD Creation Tool (ECT) is a web-based data transformation and validation tool, created in the scope of the European Holocaust Research Infrastructure (EHRI) project. It can be used for transforming XML, XML-EAD1, and CSV data in a well-formed EAD 2002 format by mapping, correcting and validating it in accordance to the guidelines of the EAD standard and the harvesting/ingest workflow.

ECT enables you to:

- Choose the mapping configuration file of your organization or use your own;
- Edit the mapping configuration to suit your needs;
- Use a custom transformation type;
- Convert your data to the EAD 2002 format;
- Preview all validation inconsistencies;
- Generate a well-formed EAD 2002 data file.

Supported formats:

- Input files (/input directory) – XML, XML EAD 1, CSV;
- Output files (/output directory) – EAD 2002;
- Mapping files (/mapping directory) – XLS, XLSX, Google Sheet.

2.1 Requirements

2.1.1 System Requirements

- Microsoft Windows 7 SP1, Windows 8, and Windows 10
- Linux
- Mac
- [Java 8](#) or later

2.1.2 Licensing

ECT is available as open-source software.

2.2 Run the ECT software as a desktop installation

The ECT software setup and running is easy and straightforward.

2.2.1 On Windows OS

1. Download and unzip the installation file.
2. Click the run.bat file.
3. The ECT tool GUI automatically opens at <http://localhost:8080>.

2.2.2 On Unix OS

1. Download and unzip the installation file.
2. Click the run shell script file.
3. The ECT tool GUI automatically opens at <http://localhost:8080>.

2.2.3 On Mac OS

1. Download and unzip the installation file.
2. Click the run shell script file.
3. The ECT tool GUI automatically opens at <http://localhost:8080>.

2.3 EAD system HOME directory

When started the ECT software automatically creates four sub-directories in its HOME directory for storing data and configurations.

2.4 Input directory

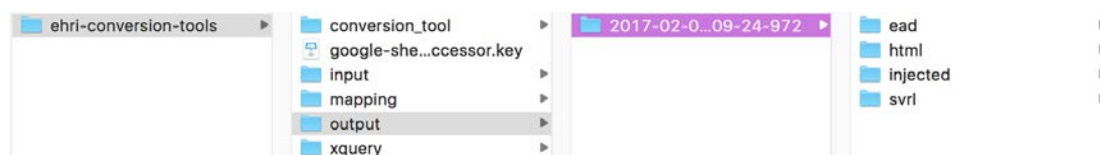
The */input* data directory is where you add the data files you want to transform.

2.5 Mapping directory

The */mapping* directory is where you can add your own mapping configuration files or, in some cases, the edited default mapping config, after correcting the validation inconsistencies from the conversion.

2.6 Output directory

The */output* data directory is where the ECT software stores all transformed data files. They are organized into subdirectories, which names reflect their creation time. Each subdirectory contains four other folders – */ead*, */html*, */injected*, and */svrl*. The ones of your interest are the first two folders, as they contain the newly generated EAD 2002 files, as well as the results from the EAD validation, in HTML format.



2.7 XQuery directory

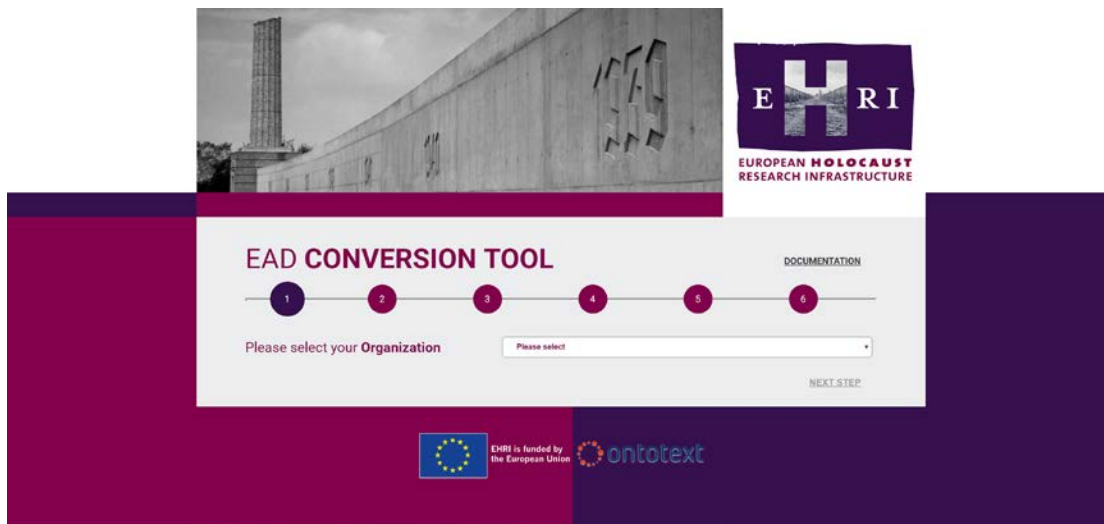
The */xquery* directory is where you can add a custom *.xqy* file to transform data files into a format different from the default EAD 2002 standard.

3 Using the ECT service

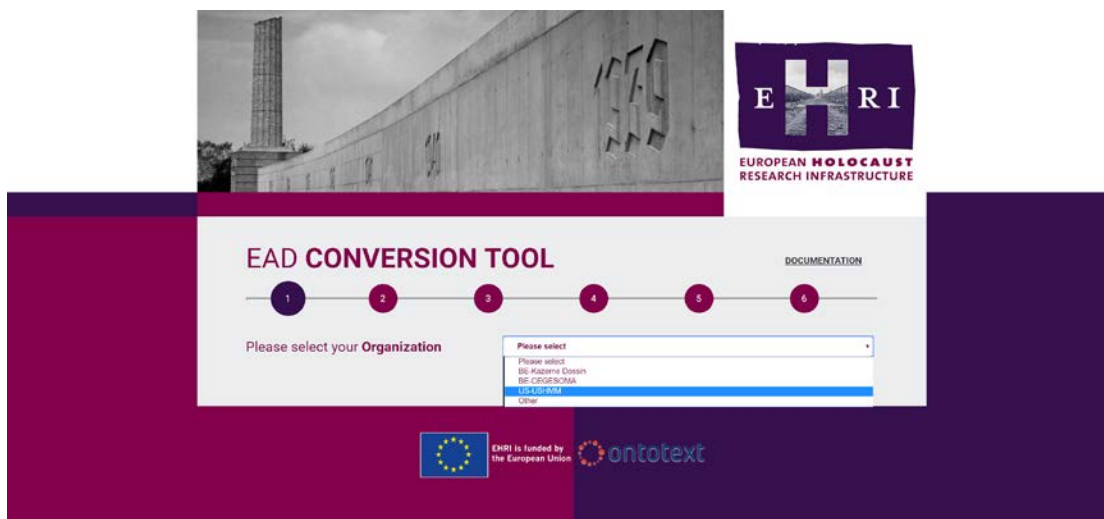
To transform your data into a well-formed EAD 2002 file, follow the steps:

3.1 Add the data files you want to transform in the ~/input directory

3.2 Open <http://localhost:8080> in a browser



3.3 Select your organization from the drop-down list



3.4 Select your files INPUT format



3.5 Select the transformation type

There are two types of transformation *Generic (default)* and *Specific*. Using the *Generic* one, you can transform your data files in the EAD 2002 format.


If you want to use the tool for transforming data in other formats, you should create your own xquery transformation schemes and add them to the `/xquery` folder. Then, they can be used when the *Specific transformation* type is selected.



3.6 Preview/edit the mapping config file

All mapping config files are stored as Google sheets. Depending on your access rights, you can view or edit them, directly in the EAD converter UI or by clicking the *View Google Spreadsheet* link.

If you need to use a custom mapping, you can add it to the `/mapping` folder and select it from the *Select local mapping file* drop-down list.



EUROPEAN HOLOCAUST
RESEARCH INFRASTRUCTURE

EAD CONVERSION TOOL

[DOCUMENTATION](#)

- 1
- 2
- 3
- 4
- 5
- 6

[VIEW GOOGLE SPREADSHEET](#)

target-path	target-node	source-node	value
/	ead	//doc	
/ead/	eadheader		
/ead/eadheader/	eadheader		
/ead/eadheader/profiledesc/	profiledesc		
/ead/eadheader/profiledesc/creation	creation	/str[@name="datetimemodified"]	"This EAD is created by EHRI on ", <dt
/ead/eadheader/archdesc/	archdesc		
/ead/eadheader/archdesc/did	did		
/ead/eadheader/archdesc/did/unitid	unitid	if (/str[@name="accession_number"]text() != /str[@name="id"]text()) then /str[@name="id"]text()	
/ead/eadheader/archdesc/did/unitid	unitid	/str[@name="accession_number"]	attribute label ("accession_number"), tr
/ead/eadheader/archdesc/did/unitid	unitid	/arr[@name="accession_number_add"]str	attribute label ("former_accession_nu
/ead/eadheader/archdesc/did/unitid	unitid	/arr[@name="rg_number"]str	attribute label ("recordgroup_number")
/ead/eadheader/archdesc/did/unitid	unitid	/arr[@name="subtitle"]str	attribute label ("subtitle"), text()
/ead/eadheader/archdesc/did/unitid	unitid	/arr[@name="title_alternate"]str	attribute label ("alternative"), text()

Select Local Mapping

[PREVIOUS STEP](#) [NEXT STEP](#)

3.7 Preview the input files



EUROPEAN HOLOCAUST
RESEARCH INFRASTRUCTURE

EAD CONVERSION TOOL

[DOCUMENTATION](#)

- 1
- 2
- 3
- 4
- 5
- 6

Organization: US-USHMM

Input Folder Content

Files
document.2016-11-22.xml

[PREVIOUS STEP](#) [START TRANSFORMATION](#)

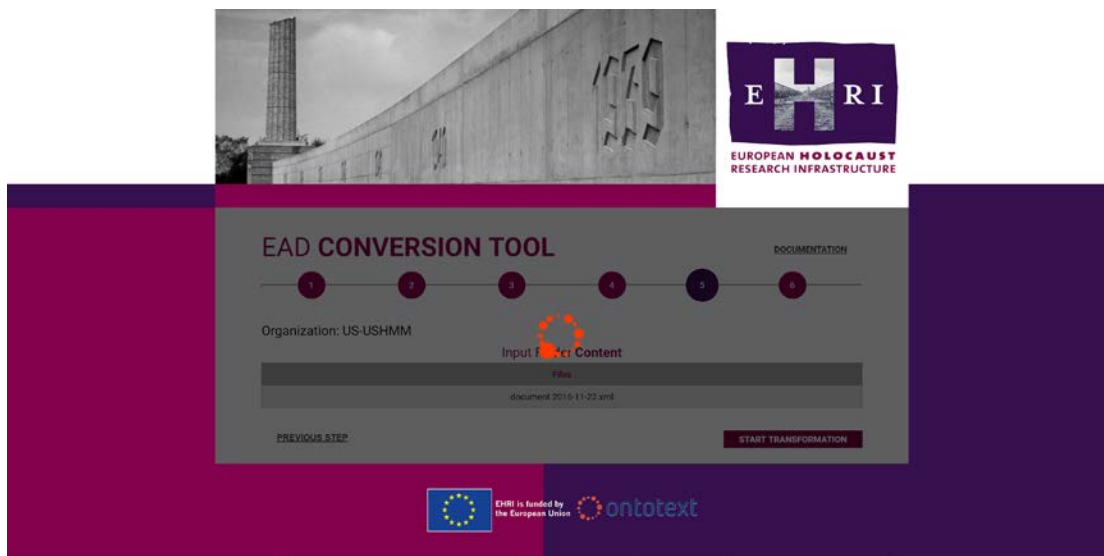


EHRI is funded by the European Union



3.8 Start transformation

To start the conversion, click the Start transformation button.



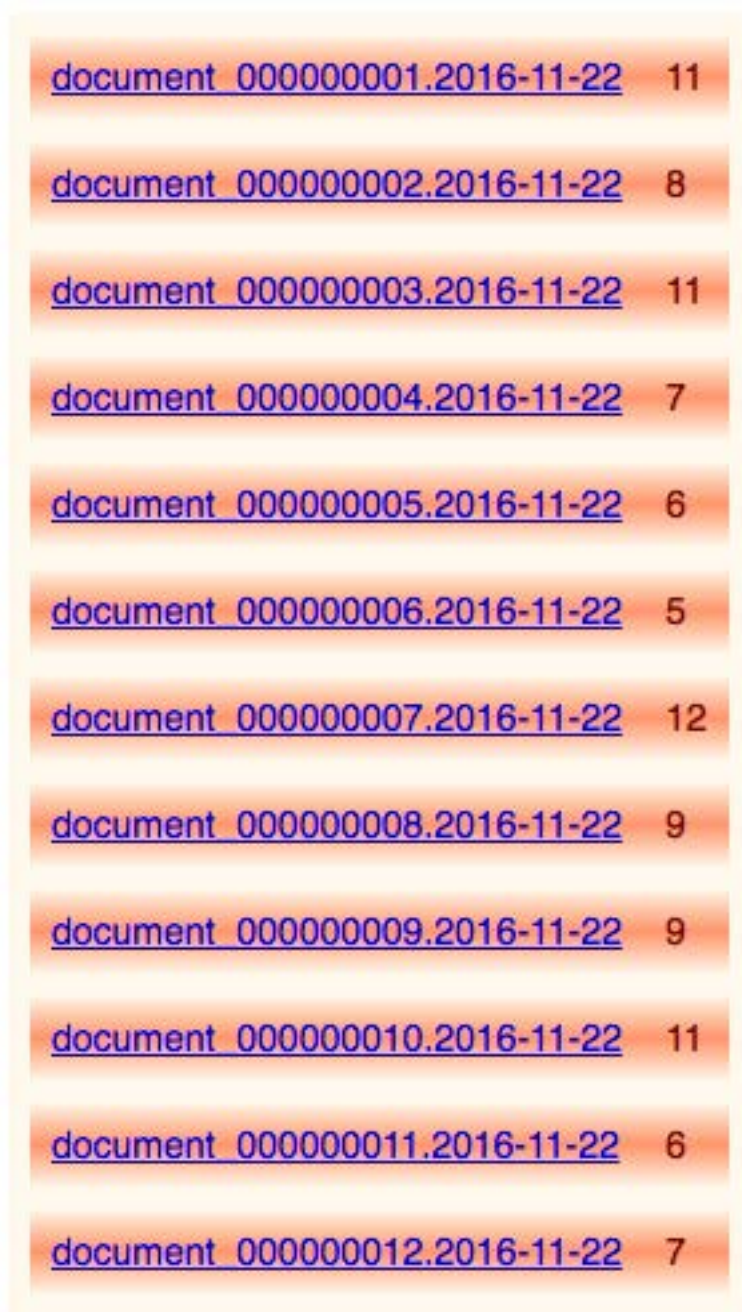
3.9 Explore the conversion report

The transformed EAD files are added to the /ead folder. The Conversion report shows the number of EAD files created, as well as the number of inconsistencies (errors) found in each of them.



3.10 Exploring the EAD validation inconsistencies

1. Go to the `~/output/<timestamp>/html` folder and click the `index.html` file. It lists all files containing errors.

A screenshot of a web browser displaying a list of 12 documents. Each document name is underlined and followed by a number representing the error count. The documents are listed in descending order of error count.

document_000000001.2016-11-22	11
document_000000002.2016-11-22	8
document_000000003.2016-11-22	11
document_000000004.2016-11-22	7
document_000000005.2016-11-22	6
document_000000006.2016-11-22	5
document_000000007.2016-11-22	12
document_000000008.2016-11-22	9
document_000000009.2016-11-22	9
document_000000010.2016-11-22	11
document_000000011.2016-11-22	6
document_000000012.2016-11-22	7

2. Click a file name to check its errors.

Each EAD generated file is presented in a user-friendly HTML format. The navigation menu on the left lists all XML elements that do not comply with the EAD 2002 standard.

Profile Description Archival Description Date of the Unit Date of the Unit Conditions Governing Access Acquisition Information Arrangement Biography or History Biography or History Scope and Content Conditions Governing Use	<p>Encoded Archival Description</p> <p>EAD Header</p> <p>Profile Description [ERROR] element "profiledesc" not allowed yet; missing required element "eadid"</p> <p>Creation</p> <p>This EAD is created by EHRI on 2017-02-02+02:00 based on the JSON file provided by USHMM on TODO: find out where to get this . This JSON file is constructed on a Catalog Record that was last modified on 2016-11-17 11:12:18 .</p> <p>Archival Description [ERROR] element "archdesc" missing required attribute "level"</p> <p>Descriptive Identification</p> <p>ID of the Unit</p> <p>im515021</p> <p>ID of the Unit</p> <p>Label accession_number</p> <p>2004.273.1</p>
---	---

3. Click the EAD element to see its errors and correct them.

For example, the picture below shows that the “Profile Description” element is not allowed, because there is a missing “eadid” element. In order to correct this error, you must add a “eadid” element to your XML input file. Depending on the validation errors, you can correct them in the input file, the mapping configuration, or the source code.

Profile Description Archival Description Date of the Unit Date of the Unit Conditions Governing Access Acquisition Information Arrangement Biography or History Biography or History Scope and Content Conditions Governing Use	<p>Profile Description [ERROR] element "profiledesc" not allowed yet; missing required element "eadid"</p> <p>Creation</p> <p>This EAD is created by EHRI on 2017-02-02+02:00 based on the JSON file provided by USHMM on TODO: find out where to get this . This JSON file is constructed on a Catalog Record that was last modified on 2016-11-17 11:12:18 .</p> <p>Archival Description [ERROR] element "archdesc" missing required attribute "level"</p> <p>Descriptive Identification</p> <p>ID of the Unit</p> <p>im515021</p> <p>ID of the Unit</p> <p>Label accession_number</p> <p>2004.273.1</p>
---	---

3.11 Validate the corrections

To validate the corrections, repeat the whole procedure and check the conversion report again.