European Holocaust Research Infrastructure
H2020-INFRAIA-2019-1

GA no. 871111

Deliverable 9.6

Overview of Data Integration

Veerle Vanden Daelen, Herminio García González, Dorien Styven & Jarno Maertens, Kazerne Dossin

Michala Lônčíková, MUA

Angel Chorapchiev, Eliot Nidam & Eli Furman, Yad Vashem

Johannes Meerwald & Florine Miez, Institut für Zeitgeschichte

Rachel Pistol & Mike Bryant, KCL

Martin Posch, DSH

Éva Kovács and Fabio Rovigo, VWI

László Csősz, HJM

Mantas Šikšnianas, VGMJH

Giorgos Antoniou, AUTh

Laura Brazzo, CDEC

Rebecca Dillmeier, USHMM

Michał Czajka, ZIH

Start: September 2020 [M1]

Due: December 2024 [M52]

Actual: December 2024 [M52]

EHRI is funded by the European Union

Document Information

Project URL	www.ehri-project.eu
Document URL	https://www.ehri-project.eu/deliverables-ehri-3-2020-2024
Deliverable	D9.6 Overview of Data Integration
Work Package	WP9
Lead Beneficiary	8 KD
Relevant Milestones	MS4
Dissemination level	Public
Contact Person	Veerle Vanden Daelen, veerle.vandendaelen@kazernedossin.eu
Abstract (for dissemination)	At the end of EHRI-3, this report brings a quantitative and qualitative overview of data integration throughout the project and outlines updates in methodology for data identification and integration carried out throughout EHRI-3.
Management Summary	At the end of EHRI's third phase of funding, this report brings a quantitative and qualitative overview of data integration throughout the project and describes updates in methodology for data identification and integration carried out throughout EHRI-3. The deliverable gives an overview of the content of the Portal and outlines the methodology behind it on four levels: the content of the Portal, the archival and other standards used in the EHRI Portal, the technical developments, and the FAIR and sustainability aspects of the EHRI Portal data. It discusses reasons for the results in collection data integration and concludes with a look into the future.

Table of Contents

List of abbreviations

API	Application Programming Interface
AtoM	Access to Memory
AUTh	Aristotelio Panepistimio Thessalonikis
CDEC	Fondazione Centro di Dokumentazione Ebraica Contemporanea
CHI	Collection-Holding Institution
CPA	Content Provider Agreement
CSV	Comma-Separated Values
DAM	Digital Asset Management
DH	Digital Humanities
DMAOG	Data Mapping Access Objects Generator
DoA	Description of the Action
DOI	Digital Object Identifier
DP	Displaced Persons
DSH	Dokumentačné stredisko holokaustu
EAD	Encoded Archival Description
ECT	EAD Creation Tool
EHRI	European Holocaust Research Infrastructure
EHRI-PP	EHRI Preparatory Phase
EHRI-IP	EHRI Implementation Phase
EMT	Entity Matching Tool
ERIC	European Research Infrastructure Consortium
FAIR	Findable, Accessible, Interoperable and Reusable
GDPR	General Data Protection Regulation
GLAM	Galleries, Libraries, Archives, and Museums
GUI	Graphical User Interface
HJM	Magyarorszagi Zsido Hitkozsegek Szovetsege Tarsadalmi Szervezet
HTTP	HyperText Transfer Protocol
ICA	International Council on Archives
IfZ	Institut für Zeitgeschichte
IT	Information Technology
IHRA	International Holocaust Remembrance Alliance
ISAAR(CFP)	International Standard Archival Authority Record for Corporate Bodies, Persons and Families
ISAD(G)	General International Standard Archival Description
ISDIAH	International Standard for Describing Institutions with Archival Holdings
JSON	JavaScript Object Notation
KCL	King's College London
KD	Kazerne Dossin: Memorial, Museum and Research Centre on Holocaust and Human Rights
KG	Knowledge Graph
LIMIS	Lithuanian Integral Museum Information System
LLM	Large Language Model
LOD	Linked Open Data
MLC	Multi-Label Classification
MPT	Metadata Publishing Tool
MUA	Masarykuv Ustav a Arvhiv

OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
OWL	Web Ontology Language
PC	Personal Computer
PDF	Portable Document Format
PM	Person Month
RDF	Resource Description Framework
RI	Research Infrastructure
RiC	Records in Contexts
RML	RDF Mapping Language
SEO	Search Engine Optimisation
ShExML	Shapes Expressions Mapping Language
SKOS	Simple Knowledge Organisation System
TSV	Tab Separated Values
URL	Uniform Resource Locator
USHMM	United States Holocaust Memorial Museum
VGMJH	Vilna Gaon Museum of Jewish History
VMT	Vocabularies Matching Tool
VWI	Vienna Wiesenthal Institute for Holocaust Studies
W3C	World Wide Web Consortium
WP	Work Package
WPL	Work Package Leader
WWW	World Wide Web
XML	eXtensible Markup Language
XQuery	XML Query
XSLT	eXtensible Stylesheet Language Transformations
YV	Yad Vashem
ZIH	Żydowski Instytut Historyczny Im. Emanuela Ringelbluma

1 Introduction

At the end of EHRI's first phase, in March 2015, the EHRI Portal was launched at the project's final conference in Berlin. It provided information in 46 country reports, 1,853 collection holdings institutions (CHIs), and 152,691 collection descriptions. By the time of the conclusion of EHRI's third phase, the EHRI Portal will provide information in 65 country reports and at least 2,304 collection holding institutions and 386,257 archival descriptions. The EHRI Portal has clearly been a significantly growing resource, and the integration of Holocaust-relevant archival descriptions into one research portal, and furthering knowledge via this portal, is one of EHRI's main goals. As a central hub of information, EHRI offers researchers an overview in a centralised database of archival sources concerning the Holocaust that are physically dispersed across the globe.

To ensure high quality and high relevance of the information presented in the EHRI Portal, the EHRI project needs to keep the already integrated data up-to-date, as well as integrate new information from both already participating archives and from other institutions the project has not worked with yet. Ensuring a critical mass of high-quality data in the EHRI Portal is a prerequisite to attract and keep users as well as to create trans-national connections across the integrated data. Therefore, it was self-evident that during EHRI's third phase of funding, the project needed to continue investing time and energy in providing as much relevant content as possible and in creating as standardised and sustainable data exchange mechanisms (import and export) as possible[1]. The latter work built on developments of EHRI's first and second phases of funding. As such, tools that allow for more semi-automated (and thus more sustainable) data integration, developed during EHRI's second phase of funding – particularly the EAD Creation Tool (ECT)[2] and the Metadata Publishing Tool (MPT)[3] – were further fine-tuned and developed. The project continued to support both archives and EHRI staff themselves in integrating and updating metadata in the Portal in the most efficient and sustainable way. Content-wise, more local and regional approaches were followed in EHRI-3 which brought new insights on methodological challenges as well as new content to the Portal.

EHRI-3 especially targeted the many collection-holding institutions that need additional support as they only have limited IT expertise and resources in-house, and consequently require tailored assistance – be it in person or via digital means or other ways of communications – to implement the EHRI tools. Where digital tools and semi-automated data export and integration could not be implemented, manual data selection and integration, directly or indirectly, via the EHRI Portal admin site, was another possibility offered. Personal relationships play a key role in establishing such contacts and facilitate such work. Moreover, for many archives, the Holocaust-relevant collections are only a small part of their total holdings, and they therefore need expertise on such sources to adequately describe them, and thereby make them accessible to Holocaust researchers. To achieve both these ends, EHRI-3 WP9 implemented regional hubs, a (mobile) data integration lab and linking techniques that are all designed to provide collection-holding institutions with the required assistance.

The work in the WP focussed on ensuring that the right content was added and updated in the EHRI Portal in a manner that was as sustainable as possible. The first task was carried

[1] Veerle Vanden Daelen, Jennifer Edmond, Petra Links, Mike Priddy, Linda Reijnhoudt, et al.. Sustainable Digital Publishing of Archival Catalogues of Twentieth-Century History Archives. *"Open History: Sustainable digital publishing of archival catalogues of twentieth-century history archives"*, Dec 2015, Brussels, Belgium

[2] https://www.ehri-project.eu/sites/default/files/downloads/ehri_downloads/D10%201%20Collection%20Description%20Production%20Services.pdf

[3] https://documentation.ehri-project.eu/projects/rspub-gui/en/latest/rst/rsgui.about.html

out by all twelve Consortium partners active in the WP. Grouped under “Task 9.1 Development of regional hubs for data integration”, KD, YV, IfZ, CDEC, ZIH, USHMM, DSH, VWI, HJMA, MUA, AUTh and VGMJH all contributed information on sources which fall under their expertise. As Holocaust sources are widely dispersed and have been curated, described and preserved in many different national settings, having a large group of experts, supplemented with extra external local experts where possible, is of utmost importance. For the technical aspects to ensure that EHRI integrates and publishes its data as sustainably and openly as possible, KD and KCL were the two partners who via the EHRI data integration lab (Task 9.2) ensured that all targeted collection-holding institutions received the necessary support to export their data and use the EHRI data integration tools. While the lab certainly has worked as a mobile data integration lab, much support – if not even more – was given from a distance, via digital means. Even more than raw numbers, the quality of the descriptions delivered and the sustainability of the connections for data export have and will continue to determine the success of the EHRI Portal. Therefore, task 9.2 also included the maintenance and further development of data integration tools for the EHRI Portal, and even introduced new methods of metadata quality control for the Portal. The tasks on Re-contextualisation of archival materials (Task 9.3) and Thematic approaches (Task 9.4) further brought to the fore specific needs and requirements, both on a content and a standard- and IT-level. Synergies between the various WPs equally helped the EHRI Portal develop and grow.

This deliverable concludes the work of WP9 with an overview of the content of the Portal, the methodology behind it, the reasons for success and challenges, and a look into the future.

2 Content of the EHRI Portal: what has happened

Integration of new data in the EHRI Portal happens in two complementary ways: manually and (semi-)automatically. The former is the *de facto* method for updating and creating new CHI descriptions and it can also be used for the integration of archival descriptions. Nevertheless, as it will be explained later in this deliverable, this method, when used for archival descriptions, is limited in its reach and it cannot, therefore, scale sustainably. In these cases, where big datasets need to be ingested, a sustainable connection can be established guaranteeing future updates or a combination thereof is present, the (semi-)automatic approach is favoured. This is done by a combination of IT technologies (see Section 3.3) which ensures that incorporating a growing number of archival descriptions can be scaled adequately and sustained in the future phases of the project. Both methods have contributed to the expansion of the institutions and archival descriptions covered on the EHRI Portal whose statistics can be consulted in Table 1.

	Country reports	**CHIs**	**Archival descriptions**
Launch Portal, end EHRI-1, March 2015	46	1,853	152,691
End EHRI-2, May 2019	63	2,137	325,273
EHRI-3 (as of 18/11/2024)	63*	2,304	386,257

Table 1: Status of the EHRI Portal by the end of EHRI-1, EHRI-2, on 31 January 2022 and by the end of the EHRI-3 project. Archival descriptions comprise descriptions at all levels (e.g., fonds, series, sub-series, collection, folder, item, etc.). * Two further country reports will be uploaded before the end of EHRI-3, bringing the total of country reports to 65.

Nevertheless, these statistics cannot account for the growing effort of keeping both CHIs and archival descriptions updated for which a more detailed breakdown is needed. In this sense, Table 2 represents the amount of created and updated descriptions (for CHIs and collections descriptions) during the lifetime of the EHRI-3 project. Differences in totals between the two tables can be explained by the removal of some descriptions due to different circumstances (e.g., change on the holders rights) and a strong deduplication effort carried out during this project phase.

	Created	**Updated**
CHIs	151	1211
Archival descriptions (manual work)	951	438
Archival descriptions (semi-automatic work)	67,126	106,031

Table 2: Detailed statistics of the created and updated description on the EHRI Portal during the EHRI-3 project time. Archival descriptions comprise descriptions at all levels (e.g., fonds, series, sub-series, collection, folder, item, etc.). Statistics generated on 18/11/2024.

3 Methodology

This section outlines how the metadata was introduced into the EHRI Portal drawing the attention to the employed methodologies, encompassing the content, archival and IT perspectives.

3.1 Content of the EHRI Portal

At EHRI-3's start, the EHRI Portal contained relevant information from more than fifty countries, with archival descriptions expressed in more than twenty different languages and authored according to heterogeneous methodologies that reflect particular national histories, trajectories and traditions in regard to Holocaust documentation, research and memorialisation. This past experience suggested a clear need to work with experts that are familiar with these local contexts (the "distance factor"). At the same time, these experts need to be cognisant of the larger overarching multidisciplinary framework EHRI uses to merge data from such a wide variety of different origins and backgrounds. Indeed, addressing the individual and local requirements of cooperating institutions while nevertheless preserving the direct link to the centralised heart of the project is a huge challenge. To face this challenge, regional data integration hubs were organised (see Section 3.1.1), external local experts were engaged (see Section 3.1.2), content guidelines were adopted to direct EHRI's data identification and integration work into the EHRI Portal (see Section 3.1.3), new EHRI Country reports were authored, and all EHRI Country Reports updated (see Section 3.1.4). Additionally, two tasks focused on linking in the EHRI Portal and on thematic approaches (see Section 3.1.5).

3.1.1 The creation of EHRI's regional hubs for data integration

EHRI implemented and operated six regional hubs to advance its trans-national data identification and integration agenda. To address the individual and local requirements of cooperating institutions while nevertheless preserving the direct link to the centralised heart of the project, the idea of working with regional hubs emerged with the intention to provide both a top-down and a bottom-up approach to data identification and integration. The hub's experts acted as the local EHRI ambassadors, liaising between the central EHRI platform

and the local work, and were responsible for ensuring selection and integration of relevant information into the EHRI Portal.

The exact configuration of these hubs in terms of their geographical remits was first outlined in Month 6 of the project and further fine-tuned.[4] The configuration of the EHRI Regional Hubs takes into account the history of the Holocaust. The main focus is on the Axis and Nazi-occupied countries in Europe (including the former North-African colonies), followed by other Axis countries, Allied countries and so-called “migration countries” where the victims of Nazi persecution sought refuge before, during or after the Second World War.

The following hubs were formed (with the coordinating institutions indicated between brackets):

- Germany (IfZ)
- Baltic States, Eastern Europe and Russia (ZIH)
- Central Europe (VWI)
- Southern Europe (CDEC)
- Western and Northern Europe, including former North-African colonies (KD)
- Migration countries (KD)

Together with the EHRI Data Integration Lab, the EHRI regional hubs:

- Supported the CHIs in their cooperation with EHRI. The regional hubs allowed for CHIs to be able to use a local (native) language instead of English and to have a local contact point.
- Facilitated data integration (linguistical, historical and archival expertise).
- Worked towards a sustainable connection with CHIs for data sharing with EHRI.
- Supported the standardisation of the descriptions of CHIs and collections.

Within the various regional hubs, specific themes were taken into account, such as cross-border cooperation in data integration, with specific focus on so-called “borderlands”, areas which have switched countries during and after the Second World War with all the challenges on Holocaust documentation and research this entails. Other areas of focus included Holocaust-related ego documents as well as archives concerning the genocide of the Roma. All regional hubs have worked in close cooperation with the EHRI Data Integration Lab to bring in as many new descriptions and updates into the EHRI Portal in the most sustainable way possible. At the same time, the regional hubs also reached out to micro-archives, and have also worked in countries and with institutions where little was hitherto possible in a semi-automated way. In these cases, manual data integration in the EHRI Portal was done either by the regional hub members, by local experts, or by the institution.

It has to be noted that the regional hubs had different compositions (in number of consortium partners as well as in available PMs and expertise), and different leadership approaches and ways of working. This resulted in variations on the amount of work delivered, and in the type

[4] Veerle Vanden Daelen, Dorien Styven & Marek Fenners (KD), Anna Ullrich (IfZ), Éva Kovács, Mirjam Wilhelm & Marianne Windsperger (VWI), Emmanuelle Moscovitz, Sigal Arie-Erez & Zohar Neumann (YV), Giorgos Antoniou & Maria Pantazi (AUTh), Stanislovas Stasiulis & Šarūnė Sederevičiūtė (VGMJH), László Csősz (HJMA), Laura Brazzo (CDEC), Martin Posch (DSH), Michael Levy & Rebecca Dillmeier (USHMM), Michala Lônčíková (MUA), Mike Bryant & Rachel Pistol (KCL), *Deliverable 9.1 EHRI Regional Hubs Implemented.* Confidential Deliverable of the European Holocaust Research Infrastructure, H2020-INFRAIA-2019-1, GA no. 871111, February 2021; Veerle Vanden Daelen & Herminio García González (KD), Mike Bryant (KCL), Laura Brazzo (CDEC), Anna Ullrich (IfZ), Éva Kovács & Mirjam Wilhelm (VWI), Michał Czajka (ZIH), *Deliverable 9.3 Interim Report on Data Integration.* Confidential Deliverable of the European Holocaust Research Infrastructure, H2020-INFRAIA-2019-1, GA no. 871111, February 2022.

of data identification and integration that was carried out (as it reflected more local contexts and needs). The way work was organised very much depended on each regional hub's preferences and needs (such as the number of meetings, the organisation of workshops, etc.). These observations should be taken into account when considering how work will be organised once the EHRI-ERIC is established with its National Node structure (see Section 5).

3.1.2 Working with external local experts

In cases where survey work still needed on-site and/or predominantly or exclusively manual data integration work, local experts were hired. This was the case for micro-archival data integration, survey and data integration work on borderland regions or regions or institutions where specific overviews of Holocaust-related archival sources were not readily available. Local experts were also hired when it concerned expert specialists who had made overviews of Holocaust-related sources in their region or country of expertise, which were not or insufficiently accessible to researchers (because of being published on paper or in PDF only), especially when it concerned countries or regions which were not represented in the EHRI consortium. Finally, given the war in Ukraine, EHRI has given special attention to survey work on these threatened archives. In total, nine local experts were hired during EHRI-3 for data identification and integration work focussing on specific regions, countries or topics. Apart from their valuable data integration into the EHRI Portal, this outreach also extended EHRI's visibility and network as the local experts became part of the larger EHRI network.

3.1.3 Content guidelines for data integration into the EHRI Portal

In order to provide guidance and guidelines on which information to integrate into the EHRI Portal, EHRI adopted the IHRA Guidelines for Identifying Relevant Documentation for Holocaust Research, Education and Remembrance[5], which were launched on 23 March 2022. All consortium and external partners are invited to use these guidelines for their data identification and integration work. Using these guidelines equally provides a framework and understanding for users of the EHRI Portal on the content it provides.

3.1.4 New and updated country reports

During its third phase of funding, EHRI has worked on two new country reports (Turkey and Cyprus) and on updating 63 other country reports. Given the fact that sources on the genocide of the Roma are also integrated into the EHRI Portal, where relevant, the country reports include now a paragraph on the genocide of the Roma in the History section.

As a work in progress (both the country reports and all other information given in the EHRI Portal), the revised country reports' introduction explicitly acknowledges this and invites input by stating: "The content of the EHRI Portal is based on the best available information received and integration of new input is an ongoing process. If you have questions or comments about the country reports or the introduction, please contact info@ehri-project.eu." Where applicable, the lack of information is also explicitly stated in the country report with again the invitation to inform EHRI should one have such information: "There are no sufficient findings on the situation of Roma in [name country] during the Second World War. EHRI would be pleased to receive any information." During the revisions of the EHRI country reports, input and feedback was explicitly requested from the members of the Academic Working Group of the IHRA, all EHRI-3 and EHRI-IP consortium partners as well as representatives of countries considering setting up or having set up an EHRI National Node for their participation in the EHRI-ERIC.

5 https://holocaustremembrance.com/wp-content/uploads/2023/08/IHRA-Guidelines-for-Identifying-Relevant-Documentation-for-Holocaust-Research-Education-and-Remembrance.pdf

The revisions of the country reports brought to the fore that the strict format by which the EHRI country reports are structured is something that can only be overseen by a central editorial board. Very often experts only read the country report of their expertise and suggest changes, additions or restructuring that diverge from the format as outlined in the Introduction to the EHRI Country Reports on Holocaust History and Archives[6]. At the same time, for this central board to successfully be able to update 65, and potentially more in the future, country reports, local experts are an absolute necessity. The combination of National Nodes and a Central hub in the EHRI-ERIC could accommodate these needs for the future.

3.1.5 Linking and thematic approaches

WP9 further investigated current practices and future possibilities on two specific topics for the EHRI Portal, namely the interconnectivity of descriptions within the EHRI Portal, by linking descriptions and by looking at thematic approaches.

Task 9.3 concerned the re-contextualisation of archival materials and was carried out by YV, KD, VGMJH, IfZ, USHMM and KCL. As Holocaust-related archives are among the most copied archives worldwide and are accessible for research in both original and copied form, there is often more than one description for an archival document, be it described as an analogue original or as a digital (or other) copy. Collection-holding institutions and projects that have copied such archives often mention the source of acquisition, but they do not have the capacity to keep this information up-to-date. Building upon preliminary methodological work undertaken in EHRI-2, EHRI-3 advanced the re-contextualisation of archival materials in the Portal by further linking copies and originals and looking into methodologies to do so. As descriptions of copied or otherwise strongly related source materials were often made in diverse languages and according to diverse descriptive paradigms, explicitly linking such information in the EHRI Portal greatly enhances its value for both researchers and archivists. The Task group focussed on analysing the ongoing work on linking original-copy collections (and potentially otherwise related collections) and how EHRI standards and EHRI Portal have implemented linking, and how this methodology is received by data providers and end users. Clearly, this re-contextualisation of archival materials within the EHRI Portal needs to receive further attention and is one the most challenging, but equally interesting and rewarding topics to engage with, bringing to the fore the many variations within the information and formatting of different descriptions, sometimes even within one and the same institution. Successful continuation of work on this topic includes both content (archivists) and IT specialists[7].

Task 9.4 “Thematic Approaches” was carried out by YV, KD, IfZ, MUA and USHMM and resulted in Deliverable 9.5.[8] This deliverable summarises the findings of experts whom EHRI asked to assess the EHRI Portal for four specific transnational research themes: Borderlands; Jewish Allied Soldiers (including chaplains and physicians in DP camps); Roma; and Current Research Trends and New Approaches to the Holocaust. The analysis is based on the data collected from a survey distributed to specialists. The goal was to examine the ways in which the EHRI Portal currently supports research, to analyse the usability of the Portal and to find ways to further enrich its productivity in the future. These expert assessments have provided important input and feedback concerning both the content and the usability of the Portal with regard to tracing trans-national themes, and thus inform

[6] https://www.ehri-project.eu/country-reports

[7] Eli Furman & Hillel Solomon (YV), Herminio García González, Dorien Styven & Veerle Vanden Daelen (KD), Johannes Meerwald & Anna Ullrich (IfZ), Mantas Šikšnianas (VGMJH), Joel Lee (USHMM), *Deliverable 9.4 Linking.* Confidential Deliverable of the European Holocaust Research Infrastructure, H2020-INFRAIA-2019-1, GA no. 871111, August 2023.

[8] Eliot Nidam Orvieto (YV), Michala Lônčíková (MUA), Johannes Meerwald & Anna Ullrich (IfZ), Herminio García González & Veerle Vanden Daelen (KD), *Deliverable 9.5 Thematic Approaches.* Confidential Deliverable of the European Holocaust Research Infrastructure, H2020-INFRAIA-2019-1, GA no. 871111, August 2023.

EHRI's future development work. Suggestions to further develop the EHRI Vocabularies, indicate contexts in which descriptions on the EHRI Portal were written as well as narrowing query results came to the fore. Apart from the Thematic Approaches task, specific thematic choices were also made within regional hubs, the "Central European" hub, led by VWI, for example, gave special attention to the topic of the genocide of the Roma and ego documents and testimony collections as Holocaust-relevant sources. They collected information on both CHIs and collections on both themes for integration into the EHRI Portal.

3.2 Archival and other standards used in the EHRI Portal

3.2.1 Revision and rewriting of the standards and guidelines

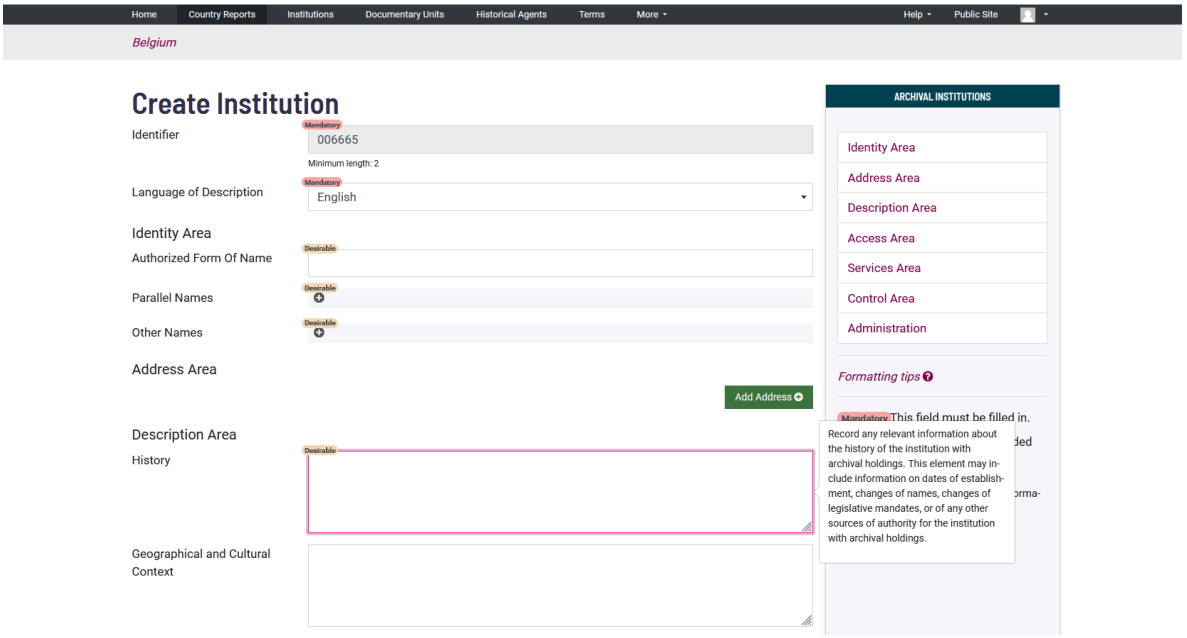

Figure 1: Screenshot of the administrative module of the EHRI Portal showing how when editing a description a tooltip appears to give more information about the field in focus.

Apart from the content guidelines for identification and integration of content into the EHRI Portal, EHRI was also often asked about the EHRI Portal Data Model. This led to a revision of the EHRI standards and guidelines as outlined by EHRI-1 and which had remained a non-public document[9]. Apart from the fact that the EHRI-1 standards and guidelines were not publicly available, the data model standards and guidelines were also drafted in archivist's jargon, explaining how EHRI applied standards of the ICA such as ISAD(G) or ISDIAH, for example. Especially for non-archivists contributing to or using the EHRI Portal, this wording was daunting. Moreover, EHRI-developed data models, such as the format for the EHRI Country Reports, did not yet figure in these standards and guidelines. During the revision, the very strict metadata formats for describing archival institutions, archival holdings and authorities were also reconsidered. The basis idea for the EHRI Portal Data Model has been to encourage as many as possible to join and work with EHRI, and therefore it limits the number of "mandatory" fields, while at the same time it encourages all to provide as much information as possible by indicating fields which are considered "desirable". The EHRI Portal Data Model[10] is published on the EHRI Portal and all explanations of mandatory and

[9] Veerle Vanden Daelen (Ceges-Soma, WP leader), Giles Bennett, Dieter Pohl & Pascal Trees (IfZ), Michal Czajka (ZIH), Judith Levin (YV) & Reto Speck (KCL), *Deliverable D.15.6 Final report.* Confidential Report of the European Holocaust Research Infrastructure (Theme [INFRA-2010-1.1.4] GA no. 261873), December 2014.

[10] https://portal.ehri-project.eu/help/datamodel

desirable fields are also available via pop-up windows in the EHRI admin pages where manual input is being entered (see Figure 1). Both formats also allow for possible updates in the future in a smoother way by including a management module for this data model in the admin site of the EHRI Portal.

3.2.2 Communication

In order to communicate as openly as possible about the EHRI Portal content and the way the information is structured and standardised within the Portal, the revision of the Introduction to the EHRI Country Reports on Holocaust History and Archives includes references both to the content guidelines and the Data Model used for the EHRI Portal. This allows for both data providers and EHRI Portal users to have access to this information. It may also be of interest to other RIs.

3.2.3 A tool for quality control within the EHRI Portal

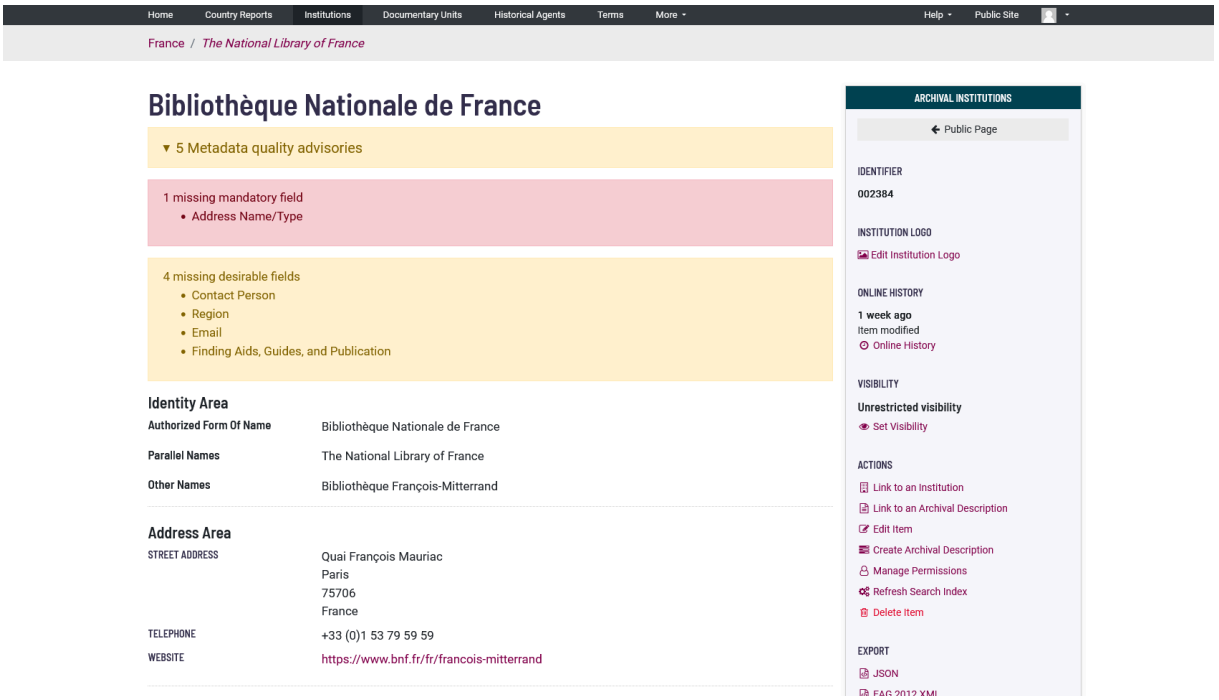

Figure 2: Screenshot of the administrative module of the EHRI Portal showing how the metadata quality advisories are shown to the user guiding data quality tasks.

The revision of the EHRI standards and guidelines also brought to the fore the need for having quality control features directly available within the work space for manual data integration, namely the admin pages of the EHRI Portal. Therefore, the EHRI admin pages do not only include pop-up windows with explanations to each mandatory and desirable field; once a description is being saved on the EHRI Portal a banner on top of the page will inform the person working on the description about any possible mandatory or desirable fields missing. It does so by stating “[number] of metadata quality advisories” with the possibility to open up a more precise listing of missing mandatory fields, indicated with a red background, and missing desirable fields, indicated with an orange/yellow background (see Figure 2). Prior to this, people working manually in the EHRI Portal did not have the possibility to receive this quality control within the EHRI Portal, but rather had to compare a separate list of mandatory and desirable fields with the EHRI Portal content. The provision of this direct quality control within the EHRI Portal has been very well received. It will also be very useful when EHRI further moves towards working with National Nodes. Metadata quality (based on the centrally-defined data model) can also be audited Portal-wide, via a tool that can scan all

items of specific types and report those with missing mandatory or desirable fields, allowing cleanup actions to be taken on an individual item or database level.

3.3 Technical development

3.3.1 Adaptation of the tools to the EHRI Portal environment

At the conclusion of EHR-2, data integration infrastructure consisted of a suite of discrete tools for the following purposes:

- ECT: a web-based, standalone GUI tool for converting arbitrary XML to EAD using an XQuery-based tabular mapping system integrated with Google Sheets.
- MPT: a Python-based GUI tool for creating ResourceSync manifests allowing data providers to make EAD created with the ECT available online for harvesting by EHRI.
- ResourceSync Aggregator: a command-line Java-based tool for harvesting data from ResourceSync endpoints.
- OAI-PMH harvester: a command-line shell script for harvesting OAI-PMH endpoints.

The ECT and MPT were standalone tools because an important design criteria was to empower data providers *themselves* to design crosswalks, convert their own data, and publish it for EHRI to harvest. One of the shortcomings of the EHRI-2 data integration system, however, was the difficulty in tracking crosswalks and metadata assets in an end-to-end manner, and dealing with multiple versions of metadata assets as the conversion process was iteratively developed. Additionally, because several steps relied on using command-line tools on a remote server environment, the overall process was complex and relatively challenging to administer, particularly when a large number of institutions were involved.

In order to try and make the system more holistic and easier-to-use we opted to integrate the ECT's crosswalk development environment, conversion process, and various harvesting components into the EHRI Portal, with a new web-based user interface. The first phase of development occurred during the EHRI-PP project in summer 2020 and proceeded with the following milestones:

- Conversion and ingest of a single uploaded dataset per institution, via the EHRI Portal administration pages.
- Batch validation for source or converted EAD files.
- Multiple datasets per institution, for processing groups of XML files in different ways.
- Addition of ResourceSync or OAI-PMH harvesting for datasets.
- Versioning of ingested files, accessible via the original ingest event in the EHRI Portal.
- UI for interactive development of XQuery mappings or XSLT transformations.
- Addition of per-institution co-reference tables for mapping access points between institution-specific and EHRI keyword thesauruses.
- Content snapshots, for removing stale material after ingest operations.
- Import logs, for tracking the number of files created, updated or unchanged by ingest operations per-institution.

While the core technologies involved are the same (XQuery mappings, XSLT, ResourceSync, OAI-PMH) the current data integration environment has a number of capabilities not available in EHRI-2:

- Interactive crosswalk development, where the output of a transformation can be previewed in real-time, prior to running a conversion.
- Ability to chain individual XQuery or XSLT transformation steps in pipelines of arbitrary length.
- Versioning of harvested or uploaded files on Cloud storage.

- Management of stale material, e.g. deletion of archival metadata no longer published by a data provider.
- Management of coreferences for remote/local access points.
- Visualisation of historical import logs.
- Ability to view the source file of a given archival unit from a particular ingest operation.

Overall, the biggest change to the general system was that whereas the EHRI-2 approach was tailored towards the harvesting of valid EAD XML, with conversion taking place on the data provider's side, in EHRI-3 we favoured the harvesting of material in any XML-based format, and performing the conversion to EAD on our own systems. A trade-off made in doing this was that the standalone conversion tools – particularly the ECT – would not benefit from further development. This was a pragmatic choice, based on the assessment that very few institutions had the technical capacity to design crosswalks, run the ECT, and publish EAD files.

Additional adaptations of the ingest pipeline to cater for specific institutional peculiarities are described in Section 3.3.3.

3.3.2 Setting up a new workflow

Given the incorporation of the data integration tools to the EHRI Portal webpage and the foundation of a data integration lab, it was necessary to set up a new workflow for data integration which could take into account this new working procedure and streamline the process of attending to the received requests.

The first step of this workflow is defined by a questionnaire[11] to be filled by any institution which would like to file a request to the data integration lab in order to explore the possibility of automatically ingesting their archival descriptions in the EHRI Portal. The questionnaire collects a series of questions aiming to get a better knowledge of the technical stack with which the institution is working, its data exchange possibilities and its technical expertise.

Upon its completion, the answers are automatically transferred to a kanban board (in the form of a Trello board) with the following states: **New**, **To Do**, **In Progress**, **In Stage** and **In Production** which serves as a coordination tool for the members of the data integration lab.

Once one of these answers is received, a member of the data integration lab is assigned it and makes contact with the designated representative from the institution, which typically results in scheduling a virtual meeting. This meeting serves to gather more detailed information on the institution, such as: the materials they hold and the possible necessity for some filtering, the archival level of the descriptions, examining in more detail the technical capabilities and start the design of the data ingestion process, and presenting the CPA. The latter is a legal requirement needed for the interchange of data between the providing institution and the EHRI project, which has constituted one of the most significant roadblocks to establishing a data integration workflow with an institution due to, amongst other things, different legislations and need for clarifications. In this regard, the WP9 has counted on the collaboration of the WP3 and WP4 which assisted in the explanation and signature processes with the institutions. All the collected information is entered into the collaborative tool and the task is moved to the **To Do** category from which all the needed details are clear and the practical set up can be started.

When a data integration lab member is ready to take on one of the tasks, they must move it to the category **In Progress** to denote that this task has been assigned and thus should not be taken by another member. This step involves the creation of the data ingestion process from the acquisition of the files, their conversion to EAD, and their final ingestion into the EHRI Portal's database. All the process is done in a staging instance of the EHRI Portal to

[11] https://forms.gle/YFCEhJzkEkSVT7yt9

which the providing organisation's representative has access to verify that the descriptions are represented according to their needs, and no information has been omitted or misplaced. As soon as the process is finished, and the representative has been asked to verify the data, the task is moved to **In Staging**. If the representative asks for changes then the task will be moved backwards to the **In Progress** state and the changes made, upon which it can be moved again to **In Staging**. This iterative process is repeated for as long as the representative asks for changes.

After the final approval by the institutions' representative the data ingestion workflow is recreated in the production Portal and the institution is notified about this. Finally, the task is moved to the **In Production** state and the technical resources used for this process are widely shared in the designated GitHub repository (see Section 3.4.4).

3.3.3 Ongoing adaptations to cover more institutions' technical stacks

During the course of this project, some cases required some adaptations on the data integration platform to be fully operational or to enhance its sustainability. In this section, we describe some of the added features in more detail.

URLSet harvesting: Since EHRI-2, the two preferred methods for data harvesting were the two Open Archives specifications: OAI-PMH[12] and ResourceSync[13]. This caused, however, a big burden on institutions that had to implement one of them in order to supply a sustainable connection to the EHRI Portal. With the move of the data integration tools to the EHRI Portal, a new files upload option appeared which sought to cover any alternative cases in which these protocols did not exist in the organisations' software or an export is the only viable option. However, the absence of an in-between approach capable of dealing with, amongst others, existing APIs, export options in the organisations' websites or similar HTTP-enabled technologies; created a very big limitation when establishing sustainable connections due to our system's inability to handle these possibilities and having to rely on unsustainable options involving *ad-hoc* file transfer. Hence, this motivated the creation of the URLSet harvester, where the implementation involves a set of web URLs directly downloaded in the EHRI Portal as input files for the data integration process. This has enabled the establishment of sustainable connections with a lot of institutions that formerly would have been supported using other means. The main drawback related to this new option is the need for creating the mentioned list of URLs which in case of updates to the number of files needs to be recreated. However, at the same time, this has also enabled a new filtering method in the EHRI Portal – bearing in mind that the EHRI Portal is only focused on Holocaust-relevant material and that many archives hold more collections than just Holocaust ones – without which it would have been much more difficult to establish a filtered and sustainable connection with the provider.

Batch import: In the vast majority of cases the sets of descriptions coming from the institutions only rose up – as a maximum – to a few thousands of descriptions. Nevertheless, some very specific cases coming from big archives delivered far bigger datasets. Without arguably being considered "big data", this substantial amount of data posed some problems to our existing infrastructure in the form of bottlenecks, server crashes and downtime. In order to avoid the server crashes caused by ingesting large quantities of data, the batch import option was added to the ingestion options. In essence it allows to define a batch number that will be used to split the input descriptions in batches of that size and ingested in the portal in sequential transactions. This allows us to import big sets of descriptions in a timely fashion without affecting our users.

[12] https://www.openarchives.org/OAI/openarchivesprotocol.html
[13] https://www.openarchives.org/rs/toc

3.3.4 Special cases

Even though for most of the cases our infrastructure was able to deliver a working solution for integrating the received data, there are some special cases that needed an external adaptation or bypass to correctly being integrated in the platform. This is mainly caused by the technical limitation of solely supporting XML files for the data integration process. While XML can arguably be classified as the *de facto* standard in the GLAM sector for data interexchange, for some different reasons, some providers privileged other formats like Excel and JSON. Therefore the following two cases were set up as a test case upon which future support for these formats can be built on the data integration tools available within EHRI.

Importing Excel files: For different reasons, ranging from technical limitations to familiarity with spreadsheets, some of the received requests could only contemplate providing an export in the Microsoft Excel format. Therefore, it was necessary to design a repeatable workflow to convert these files to an XML counterpart, and ideally to an EAD output. Practically, a generic workflow using Open Refine was put in place allowing for cleaning and normalising the data and afterwards exporting it to an EAD file by means of a generic EAD template. This has also led to the publication of an entry in the EHRI Document Blog[14] further explaining the workflow and how to apply it to other cases.

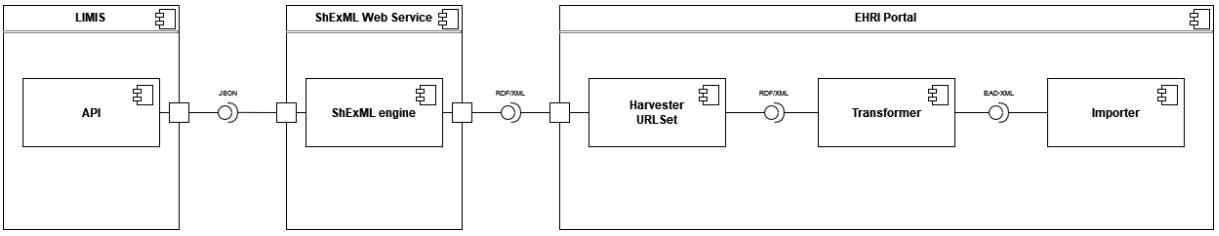

Figure 3: Diagram of the architecture implemented to transform the received JSON files into EAD using the ShExML engine and then the normal EHRI Portal data transformation workflow.

Converting JSON files with ShExML: More and more cultural heritage institutions are favouring the JSON format over the previously ubiquitous XML format. This tendency, already noticeable for some years in other fields, is attributable to its less strict schema and reduced verbosity (which in the context of data exchange and the so-called “post-PC” era is translated to less use of bandwidth). At the same time, declarative mapping rules are being proposed as a more flexible, adaptable and reusable way of integrating heterogeneous data sources under a single representation, superseding *ad-hoc* approaches.[15] ShExML, developed and maintained by one of the authors of this deliverable and which can be classified under the realm of declarative mapping rules, allows for the integration of JSON, XML, CSV and relational databases under a single RDF representation.[16] Therefore, we created a test case for integrating the VGMJH collections sourced from the Lithuanian national aggregator LIMIS[17] represented in the JSON format. For that purpose an input ShExML file was developed and the ShExML engine was embedded into a new EHRI web service which is later called from within the previously-introduced new URLSet harvesting option. This workflow results in the acquisition of files under the RDF/XML format by the

[14] Herminio Garcia González (2022). What Can I Do With This Messy Spreadsheet? Converting from Excel Sheets to Fully Compliant EAD-XML files. Version 1.0.0. EHRI. [Training module]. https://blog.ehri-project.eu/2022/04/25/converting-from-excel-to-ead-xml/

[15] Van Assche, D., Delva, T., Haesendonck, G., Heyvaert, P., De Meester, B., & Dimou, A. (2023). Declarative RDF graph generation from heterogeneous (semi-) structured data: A systematic literature review. *Journal of Web Semantics*, *75*, 100753.

[16] García-González, H., Boneva, I., Staworko, S., Labra-Gayo, J. E., & Lovelle, J. M. C. (2020). ShExML: improving the usability of heterogeneous data mapping languages for first-time users. *PeerJ Computer Science*, *6*, e318.

[17] https://www.limis.lt/

EHRI Portal. Those RDF/XML files can be later on converted to EAD following the usual workflow and tools. A graphical representation of this process can be found in Figure 3.

3.3.5 Consultancy

During several meetings with the prospective metadata-providing institutions, assistance was requested about different technical aspects as this is something that, in many cases, cultural heritage institutions really lack. While these activities do not strictly fall under the DoA of this WP, it was deemed fair, under the wider scope of the EHRI project, and productive from EHRI's perspective, to offer some consultancy on these technological aspects in combination with the main data integration pursuit of this WP. Moreover, this had a two-fold benefit: 1) just offering the integration of the institutions' metadata on the EHRI Portal under the premise of gaining visibility is *per se* of great value, but for some institutions this is not enough, so the addition of technical consultancy made it more attractive to them; 2) in some cases there was a lack of technical capacity to export the data – or even present it in a semi-structured and therefore machine-processable format – which by introducing them to different technical solutions makes getting data into the EHRI Portal possible. Cases under the realm of 1) included – but are not limited to – SEO, presentation of different systems for archival collection management, API best practices, etc. For solving the issues under 2) two main lines of action were explored which are described below.

AtoM as an open source archival descriptions management system: In some cases, institutions do not have access to fully-fledged archival management systems due to a variety of reasons, ranging from lack of funding, lack of access to specialised companies and/or training. This scenario forces archives to rely on solutions like Word and Excel documents or Access databases to store their descriptions. While there are many proprietary solutions and companies offering these kinds of services in the market, we ought to offer as open as possible solutions, in line with the open source advocacy of EU-funded projects.

Therefore, the selected open source tool was AtoM[18] which is based around the ICA standards, is web based (offering a combined solution for managing and exposing the descriptions), and supports multiple formats for export and import, avoiding vendor lock-in. In addition, as it is open source, there is no need to pay a licence to use it and the only ongoing cost is related to the running and management of the server infrastructure making it a very cost-effective solution for small institutions.

Based on this software a course of training material was developed[19] and offered to institutions who may need it, along with the possibility to offer this training in-person. Even though this training had a limited uptake during the lifetime of this project, one EHRI partner, the Elie Wiesel National Institute for the Study of the Holocaust in Romania has benefited from it and is implementing an instance of AtoM to manage their archive and effectively deliver their descriptions to the EHRI Portal.

Data Mapping Access Objects Generator: Sometimes, there is some technical expertise available for the cultural heritage institutions but the funding to set up a full system falls short. This was, for example, the case at Kazerne Dossin, which for various reasons never allocated the necessary resources to make their library visible to the public despite having the information structurally collected in an Access database. In order to solve this problem, research was done in a solution that would alleviate the development time while at the same time would be flexible enough to cover a big range of different sets of data.

[18] https://www.accesstomemory.org

[19] https://docs.google.com/presentation/d/1vSReg_2RwzEJ0zmPef8Cy250AC4Fr2VEXcmlKuQgN4M/edit?usp=sharing

This process culminated in the creation of the DMAOG library[20] which is able to create a data-access layer automatically from a given RDF file or a set of mapping rules in ShExML or RML[21]. After this, the developer only needs to develop the front-end code while all the data access concerns are encapsulated by DMAOG. In practice, this has led to the implementation of the Kazerne Dossin's library website[22] as a first use case to prove its reliability. In the future this webpage can be adapted to other institutions in a much faster fashion than developing a normal application as only the front-end code needs to be adapted while, on its turn, the back-end code is adapted automatically by DMAOG.

This tool is fully based on semantic technologies, making the data compatible with LOD formats and helping institutions to make their data more FAIR, promoting the realisation of initiatives such as the European Heritage Cloud[23]. The findings of this research have been gathered in a paper which is currently under review[24].

3.4 FAIR and sustainability

The work carried out in this work package can be seen as an effort to make the metadata of Holocaust archives around the world more FAIR[25] by means of a centralised integration on the EHRI Portal. While this remains true, it would be neglectful not to acknowledge that more and more aggregators are coming to the fore – in many cases, with overlapping scopes – making interoperability a totally new challenge. Moreover, archives are faced with an increasing challenge of being present on all these platforms, having to adapt, in many cases, different techniques for this and at the same time being forced to deliver their data multiple times. This situation, in the context of an already overloaded Cultural Heritage sector, foretells the necessity for a change of paradigm. Therefore, in the context of the EHRI project, this WP has investigated different technological solutions that can contribute to the future alleviation of this growing problem and a better interconnectedness of the field.

3.4.1 Linked Open Data technologies

The Semantic Web was proposed as a new form of representation in the world-wide web that, by extending it, would enable the interlinking – and navigation – of entities by means of standards and shared vocabularies.[26] Its realisation would potentially allow machines to understand and navigate the contents exposed on the web as humans have been doing from its very beginning and even make inferences based on the ground truth and a set of axioms (defined beforehand). While this vision has existed already more than 20 years, the underlying technologies have only been gaining traction during the last years in – amongst other fields – the GLAM sector, given their ability to unambiguously identify different entities

[20] https://github.com/herminiogg/dmaog
[21] Dimou, A., Vander Sande, M., Colpaert, P., Verborgh, R., Mannens, E., & Van de Walle, R. (2014). RML: A generic language for integrated RDF mappings of heterogeneous data. *Ldow*, *1184*.
[22] https://bibliotheek.kazernedossin.eu/
[23] https://research-and-innovation.ec.europa.eu/research-area/social-sciences-and-humanities/cultural-heritage-and-cultural-and-creative-industries-ccis/cultural-heritage-cloud_en
[24] García-González, H., Bryant, M., & Vanden Daelen, V. (Under review). "Stop writing repetitive code!" – Scaffolding a semantic data access layer to abstract developers from semantic technologies and boost their productivity.
[25] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, *3*(1), 1-9.
[26] Berners-Lee, T., Hendler, J., & Lassila, O. (2001). Web Semantic. *Scientific American*, *284*(5), 34-43.

and interlink them across datasets, better supporting historical research[27] using digital humanities techniques[28].

This recent shift is also reflected in the new standardisation efforts that the ICA is undertaking in the form of the new RiC Conceptual Model which seeks to supersede and merge the standards in use for more than 20 years in the archival community (e.g., ISAD(G), ISAAR, ISDIAH, etc.). For doing so, apart from the conceptual model, the ICA also launched the RiC Ontology developed in OWL, one of the standards proposed by the W3C for the Semantic Web[29]. After more than 10 years of work on this new standard, the ICA released earlier this year the first stable version of RiC (v1.0).

As the EHRI Portal has followed from its inception the ICA standards for archival representation, it was deemed necessary to explore this new standard, moreover when the underlying technologies can suppose an advancement for the data integration activities developed under this WP given its interlinking (across different databases) capabilities. To this effect the data in the EHRI Portal was aligned and transformed to RiC and a test platform has been set up which has been dubbed as the EHRI Knowledge Graph (EHRI-KG).[30] This platform is at the moment of writing in testing mode but funding has been secured[31] to make it production-ready in the course of the next two years.

Rather than replace the EHRI Portal as it exists today, we envision the enrichment of it with new data emerging from the EHRI-KG. For example, it is possible to interconnect on-the-fly data from some of our partners (as it has been demonstrated with CDEC), enriching the current capabilities of the EHRI Portal. It can also ameliorate the challenges of EHRI data integration, as less data would need to be integrated into the EHRI Portal if it can be queried from the data providers, ensuring that it is fully up to date. Finally, as mentioned at the beginning of this section, an increasing number of aggregators are appearing for which these technologies can bring some coherence given the possibility to link between different aggregators and data providers in a seamless manner.

This work has also led to the publication of a conference paper[32] in the International Semantic Web Conference 2023 where it was nominated as one of the best “in-use” papers.

3.4.2 Thematically connecting the archival descriptions

Since the EHRI-1 phase, EHRI has elaborated and maintained three SKOS-format vocabularies. They have served as multilingual taxonomies of subject headings, camps and ghettos, whose terms can be linked to the access points used in archival descriptions. Ultimately, this has enabled a multilingual and thematic search across the archival descriptions held in the EHRI Portal irrespective of the language in which they are represented. However, the coverage of these vocabularies from within the archival descriptions is not as high as it would be desirable due to the challenges of accurately assigning access points at scale, or co-referencing in-house vocabularies used by data providers. In order to improve this coverage and thus make this thematic search more useful

[27] Meroño-Peñuela, A., Ashkpour, A., Van Erp, M., Mandemakers, K., Breure, L., Scharnhorst, A., ... & Van Harmelen, F. (2015). Semantic technologies for historical research: A survey. *Semantic Web*, *6*(6), 539-564.

[28] Hyvönen, E. (2020). Using the Semantic Web in digital humanities: Shift from data publishing to data-analysis and serendipitous knowledge discovery. *Semantic Web*, *11*(1), 187-193.

[29] https://www.w3.org/2001/sw/wiki/Main_Page

[30] https://lod.ehri-project-test.eu/

[31] https://oscars-project.eu/projects/ehri-kg-european-holocaust-research-infrastructure-knowledge-graph

[32] García-González, H., Bryant, M. (2023). The Holocaust Archival Material Knowledge Graph. In: Payne, T.R., *et al.* The Semantic Web – ISWC 2023. ISWC 2023. Lecture Notes in Computer Science, vol 14266. Springer, Cham. https://doi.org/10.1007/978-3-031-47243-5_20

for researchers, a series of semi-automatic and automatic approaches have been tested during the duration of the EHRI-3.

The VMT[33] was introduced as an adaptation of the already existing EMT[34] which allowed for matching text entries against normalised entities from the EHRI Portal (i.e., people, corporate bodies, terms and archival Institutions) and places from the existing GeoNames vocabulary. However, the EMT presented certain limitations when using its results for linking archival description access points to the EHRI vocabularies: 1) the matches are limited to exact or phonetic matches, limiting the possibility for more fuzzy matches, for example, cross-lingual transference based on the similarity of terms in different languages; 2) the outputs can be only exported in a bespoke CSV format that cannot be easily imported in the EHRI Portal.

Due to these reasons, the VMT was developed based on the existing EMT code adapting its matching system, its interface and the export capabilities. The matching system was completely changed and made generic to link to terms on any existing SKOS vocabulary using a set of approximate-string-matching algorithms which will calculate a confidence score for all the labels of a term in the given vocabulary. This algorithm has been enclosed in the reusable library label2thesuarus which is available as open source on GitHub[35]. On the interface level, an additional text area has been added in order to ask the user for the list of vocabularies to be analysed. The candidates selection mechanism has been maintained in order to let users decide which results are best suited for their problem and thus supporting a semi-automatic workflow. Finally, an additional export option has been incorporated that enables the user to export a TSV file in the format used by the EHRI Portal for the coreference table.

At the same time, an adaptation was made in the coreference table section of the data integration tools of the EHRI Portal to allow for importing the TSV exports of the VMT as new entries of it. Hitherto, it was only possible to create a coreference table from the access points manually linked through the manual archival descriptions creation interface on the EHRI Portal making the whole process for automatic-ingested descriptions long and tedious. This, ultimately, lets the data integration lab receive the matching results from content experts of the institutions, which provide descriptions to the Portal and incorporate these new links between their access points and the EHRI vocabularies in a seamless manner.

Institution	Found matches	New coreferences	Created links
Kazerne Dossin (n=2)	41	28	552
Fritz Bauer Institut	10	10	10
Státní okresní archiv Zlín	7	7	11
Institut für Zeitgeschichte–Archiv	12	2	445
Wiener Library Tel Aviv	15	15	417

Table 3: Results obtained by content experts after using the VMT using the access points of the archival descriptions provided by their institutions as the input.

[33] https://vmt.ehri-project-stage.eu/
[34] https://emt.ehri-project.eu/
[35] https://github.com/herminiogg/label2thesaurus

In order to prove this workflow, a small test study was carried out which involved 6 representatives of 5 different institutions represented on the Portal. Table 3 collects the results obtained by the participants showing very mixed results, evidencing that, while it is a useful tool that improves the existing workflow, it is still far from perfect and more research has to be done both on the algorithms and interface side. This seems to be supported by the participant comments which, in general, found the tool useful, but some of its results were considered as naive and/or not very intelligent by the participants. In addition, a lack of access points prevents the use of this tool for which further methods need to be studied and perhaps integrated into the VMT.

3.4.3 AI approaches and LLMs

In order to test the possibilities of accurately matching access points to archival descriptions at scale, EHRI investigated a range of machine-learning and AI approaches. Since the task was essentially one of MLC, with the EHRI vocabularies as large (but not extreme) label sets, these tests employed an existing MLC tool called Annif, developed by the National Library of Finland. Annif provides a framework within which a variety of MLC approaches and algorithms can be orchestrated and evaluated side-by-side, and can be extended with additional “backends” for the integration of new classifiers. Out-of-the-box, Annif includes a set of classifiers incorporating both lexical (string matching or rule-based) and statistical-associative (supervised machine-learning) approaches.

In addition to Annif’s classifiers, EHRI also wanted to trial two approaches to MLC employing LLMs. The first of these used an LLM that was fine-tuned using examples of text from archival descriptions that were assigned EHRI keywords. The second was an off-the-shelf LLM trained only on general-purpose material, a so-called “zero-shot” approach (because the LLM had not been trained on any EHRI-specific material). In both cases, the starting points for the tests were models that had been trained on a wide range of multilingual material, as close as possible to the set of languages used across EHRI’s data providers. Custom Annif backends were created for the fine-tuned and zero-shot LLM tools respectively, allowing them to be evaluated in an identical environment to Annif’s native backends.

The datasets used for training and evaluation of both Annif and LLM-based classifiers was extracted from the EHRI Portal and consisted of a single text (created by concatenating relevant fields of archival descriptions) and the subject headings assigned to that descriptions from the EHRI Terms vocabulary. This dataset was limited in size due to the limited coverage of labels assigned, with only 30% of descriptions with access points having EHRI Terms subject headers (and only 75% of descriptions having access points at all). The dataset furthermore had some class imbalance issues due to the majority of assigned labels having come from co-referencing third-party catalogues (and therefore favouring more general/generic terms) and the minority deriving from EHRI’s manual cataloguing (favouring more specific terms). Nonetheless, we felt that despite these issues the dataset was robust and representative enough to give us a good indication of the potential of different MLC approaches. An iterative stratification method was used to divide the dataset into training and evaluation portions that contained representative examples of labels and text in particular languages.

Two types of evaluation were used. A quantitative evaluation, using the scores derived from Annif’s in-built evaluation framework, and a qualitative approach where the outputs from the best-performing Annif model, fine-tuned LLM, and zero-shot LLM were adjudicated by three judges in a blind setting, alongside the labels used as the “gold-standard” ground truth. In the quantitative evaluation, Annif’s Neural Network ensemble classifier (aggregating scores from several distinct machine-learning models) performed the best, closely followed by EHRI’s fine-tuned LLM-based model. Quantitatively, the zero-shot LLM fared very poorly by comparison. The qualitative evaluation broadly agreed with these conclusions but judged the zero-shot model to fare much more strongly, particularly with assigning a larger number of

labels of greater specificity. The general takeaway from these tests is that statistical-associative (non-LLM machine-learning) approaches are most suitable for classification at scale, but LLMs have significant potential at suggesting good labels for machine-assisted human cataloguing, and other data-augmentation tasks.

This work was carried out in collaboration with other WPs and the findings of this experiment were presented at the DH Benelux 2024 conference.[36]

3.4.4 *Practising what we preach*: making the data integration workflow FAIR

Establishing a sustainable connection between the EHRI Portal and a provider institution entails a lot of effort which needs to be done all over again for setting up another connection between the same provider institution and another aggregator. This duplication of efforts goes against the own cultural heritage ecosystem and foremost against the provider whose resources – as discussed earlier – are normally limited. While the technologies described in Section 3.4.1 can suppose an alleviation of these efforts due to the realisation of a federated cultural heritage cloud the needed technological shift will not be achieved in an immediate future. Therefore, streamlining the data integration efforts can prove beneficial for the whole community where others can build upon our findings.

This is why we have opted in this project phase to release the documentation and the code generated by the data integration lab under an open source repository[37]. Moreover, it is a transparency effort towards our providers which can see how the integration on the EHRI Portal is achieved and they can reuse the workflow and its resources for their endeavours in other platforms. A permanent version of this repository with the specific version achieved during the EHRI-3 lifetime can be accessed on Zenodo through the following DOI: 10.5281/zenodo.14162555[38].

4 Reasons for the results on collection data integration

This section analyses the reasons for the results on the data integration into the EHRI Portal and the front-end use of the EHRI Portal, which is linked both to its content and user-friendliness.

On the content-provider side, apart from very enthusiastic cooperation, which is clearly evidenced in the high numbers presented for the EHRI Portal, we also wish to analyse aspects that slowed down or prevented successful data integration into the EHRI Portal. One of the largest obstacles to institutions or individuals providing content to the EHRI-Portal was the need to complete the aforementioned CPA. Many institutions (and persons) have no legal support to whom they could present this CPA and were therefore very reluctant to sign such an agreement in the absence of authoritative legal guidance. Thanks to support from WP3 (Impact, innovation and sustainability), and workshops in cooperation with WP4 (Localisation and Capacity Building), this was partly solved, but it remains a point of attention.

A second important aspect was that some archives, for various reasons, were not ready or able to share their information with EHRI. The reasons varied greatly and included, for example, the degree of openness of the institutions, the impact of GDPR legislation, understaffing and the precarious physical conditions in which the archives were operating. Very often the metadata available was a key reason for non-sharing. Many institutions considered their metadata as not yet ready to share (for example no metadata available yet on a collection) or uncertainty as to their institution's technical and archival skills to be able to

[36] Dermentzi, M., Bryant, M., Rovigo, F., & García-González, H. (2024, June 3). Multilingual Automated Subject Indexing: a comparative study of LLMs vs alternative approaches in the context of the EHRI project. DH Benelux 2024, Leuven, Belgium. https://doi.org/10.5281/zenodo.11457688
[37] https://github.com/EHRI/DataIntegrationLabResources
[38] https://doi.org/10.5281/zenodo.14162555

meet both the content and technical quality requirements to publish information on the EHRI Portal. Other institutions were busy with internally reorganising their metadata publishing and/or DAM system, and institutions in transition were typically disinclined to share metadata from their legacy outgoing system.

Apart from great variations in archival information system quality in the various institutions, what EHRI noticed all along is that many institutions, big and small, are in a way held hostage by the external companies providing them with their cataloguing and DAM systems. All too often institutions do not own the rights or knowledge to share or export their own (meta-)data. When requesting such to external companies, prohibitive costs were often involved. Time, people and budget remain issues when requesting institutions to work with an RI like EHRI, emphasising the need for RIs to take this further into account. Thus, providing sufficient incentives for metadata providers shall be kept in the spotlight in future EHRI data integration endeavours.

Language and linguistic barriers equally continue to play a role on multiple levels. First there is the historical element: because of the war and the post-war histories and changes of borders and even countries, documents in languages which are not the official language of a current country nevertheless ended up being preserved there. A typical example are the Slovak-Hungarian borderlands, or looted and otherwise displaced archives. This often causes a backlog in the opening and accessibility of the archives by the lack of adequate metadata. Another linguistic challenge is that English is the main work language in EHRI. While many efforts have been made to reach out in the local languages, the fact that most of EHRI's communication and its offer are foreseen in English may be a reason for certain regions to join less easily, both as a content-provider and as a user.

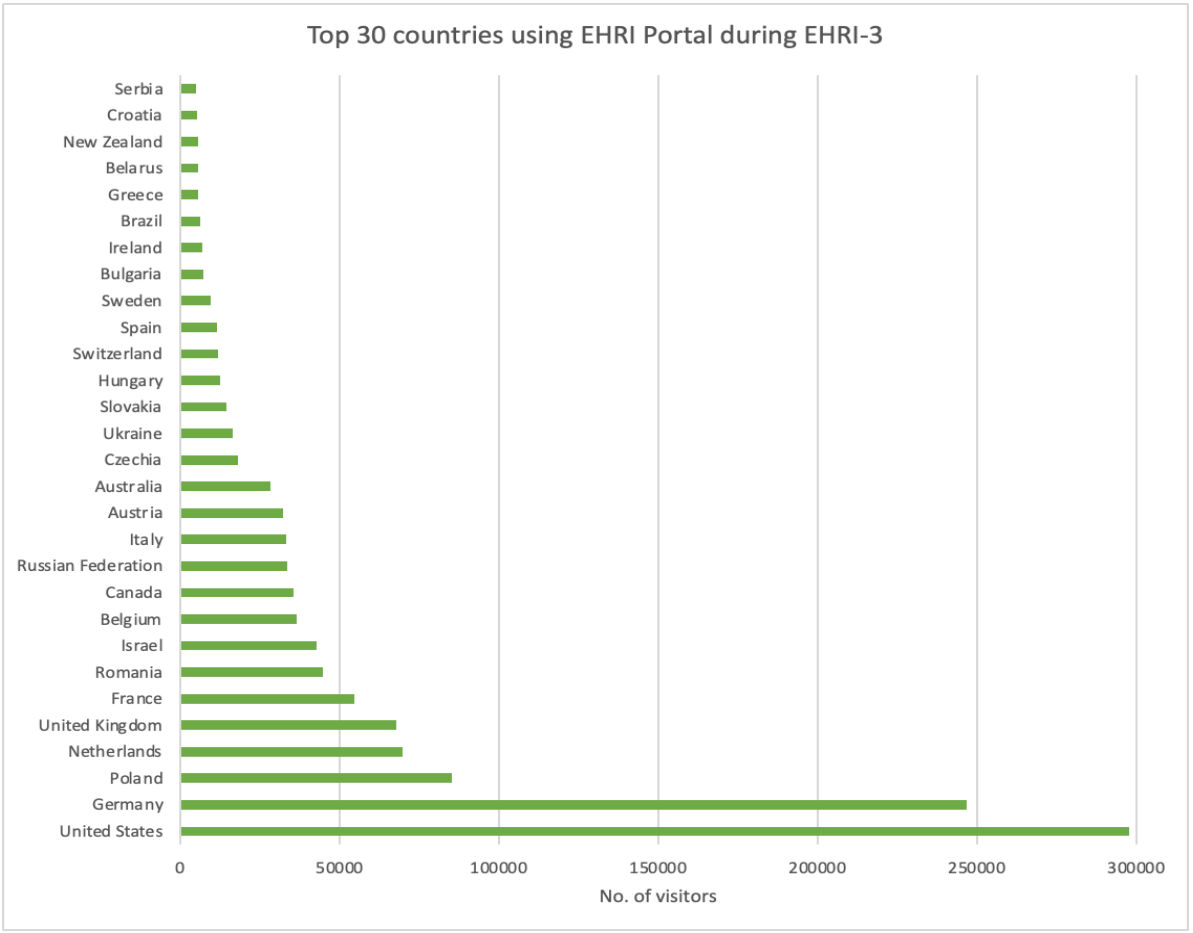

Figure 4: Visitors to the EHRI Portal by country from September 1st 2020 to October 31st 2024

There has not been any targeted communication strategy from EHRI-3's Communication WP (WP2 Dissemination) towards possible CHIs or aggregators to share information on their organisation and the metadata it manages with the EHRI Portal. The potential outreach carried out by the regional hubs and the data integration lab was deemed enough to reach the objectives. Even when tailoring a specific communication strategy in this regard, the possible answers would be difficult to be answered given the limited capacity of the regional hubs and the data integration lab. However, this could be a strategy envisaged for the future. Especially targeted efforts could happen within EHRI national nodes or towards countries which either have a low use-rate of the EHRI Portal (see Figure 4) or less participating data providing organisations than elsewhere. Looking in the EHRI fellowship applications could also be helpful in this regard, as well as a more general analysis of the attention for Holocaust research and archives in a certain country or region.

An interesting question is to what degree in-person meetings have played a role in institutions and individuals engaging with EHRI's Portal, both as a content provider and a user. In the outline of EHRI-3's project proposal, it was strongly believed that EHRI physically had to reach out and meet people to engage them and their institutions. Starting up the project in the heat of the Covid-19 pandemic, including lockdowns and severe travel restrictions, this most certainly posed huge challenges for the team to get started, and forced the team to revise its strategy of having an "EHRI mobile data integration lab". Surely, EHRI travelled and organised many in-person meetings, which often led to successful data integration into the EHRI Portal. However, successful data integration also happened remotely, and not all in-person visits paid off. A lesson learned, also considering our ecological footprint, is to evaluate very carefully which travels are worthwhile. A thorough preparation and exploration of all online possibilities are a prerequisite.

At the same time, EHRI has strongly contributed to strengthening regional cooperation in data integration and in the identification of regional and thematic aspects of Holocaust studies, as was clearly evidenced by the work of the Central-European Hub, both within the cooperation with the consortium partners and beyond. For Ukraine, EHRI has strongly supported continuation of data identification and integration in this country hit by war and destruction. As such, even remotely, EHRI's work has been able to make a significant contribution.

Multiple synergies between various WPs work and EHRI online services have further strengthened the EHRI Portal. New additions to the EHRI Document Blog[39], Online Courses[40] and Online Editions[41], the EHRI Podcast series developed in EHRI-3[42], and most recently the EHRI Geospatial Repository[43] have continued to support, challenge and make contributions to the EHRI Portal, its content and functionalities. Various other EHRI events, both analogue and online, such as the EHRI workshops, seminars, webinars and fellowships have actively used and/or enriched the EHRI Portal. The EHRI Webinar "The EHRI Knowledge Graph, New Possibilities for the EHRI Portal's Data"[44] of 27 November 2024 and the EHRI Online Course Aligning Holocaust data with Open Research and FAIR data principles[45] are particularly interesting to (potential) EHRI Portal content providers. The focus by WP10 on "Thematic layers across collections" as well as by WP11 on "Connecting micro-archival communities and standards" have furthermore helped us to understand complexities of both thematic layers and micro-archives. The work of WP4 "Localisation and capacity building" provided further introductions to (or information concerning) possible data providers to the

[39] https://blog.ehri-project.eu/
[40] https://www.ehri-project.eu/ehri-online-courses
[41] https://www.ehri-project.eu/ehri-online-editions
[42] https://www.ehri-project.eu/ehri-podcast-for-the-living-and-the-dead
[43] https://geodata.ehri-project.eu/geonetwork/srv/eng/catalog.search#/home
[44] https://www.ehri-project.eu/next-ehri-webinar-27-november-ehri-knowledge-graph-new-possibilities
[45] https://openplato.eu/blocks/catalog/detail.php?id=78&catalogauthuser=1

EHRI Portal, offering online and in-person support to engage with them. All these synergies allowed WP9 to better understand specific needs, challenges and expectations upon engaging with EHRI.

All practices and experiences throughout EHRI-3's data identification and integration work for the EHRI Portal continue to give evidence of the increased benefit of having both manual and IT-supported work. According to this experience, the combination of content experts (mostly archivists and historians) and digital humanists shall remain in further phases of EHRI. Moreover, technical, archival and content standards and guidelines are an absolute *conditio sine qua non* for such a large and increasingly distributed infrastructure. The more quality control that can be exerted at the time of data ingestion and during manual integration, the more coherent and cohesive data can be offered. Ensuring all these aspects will help continue the steady growth in users that the EHRI Portal has seen during the EHRI-3 lifespan (see Figure 5).

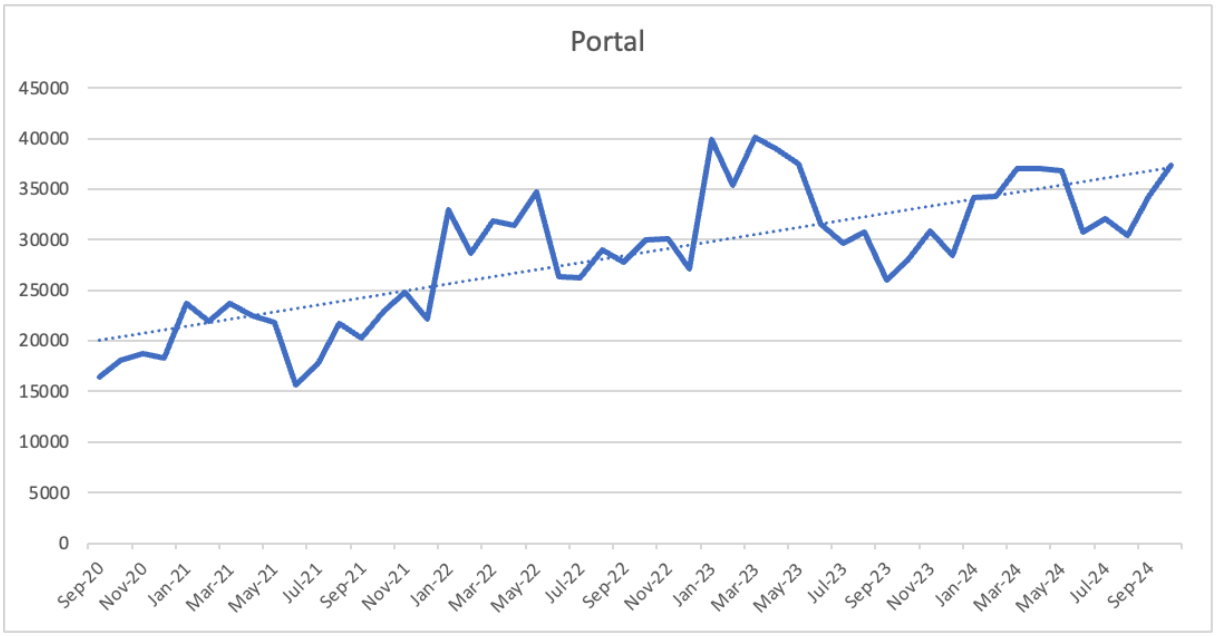

Figure 5: Visitors to the EHRI Portal during EHRI-3.

5 Looking at the future

Working with regional hubs in EHRI-3 has been a very useful experience which brings elements to the fore that help structure and organise the work on data identification and integration into the EHRI Portal once EHRI has become an ERIC and will start working with National Nodes and a Central Hub. In many ways, the regional hubs can be seen as a stepping stone toward the National Nodes (as well as to the potential Working Groups within the EHRI-ERIC). The decentralisation from the WPL to the regional hub leaders strengthened local and regional cooperation and the ability to assist and communicate in local languages, which has been evaluated as very positive. At the same time, the decentralised way of working also led to variations in the interpretation of the format and content of the EHRI Portal. One concrete suggestion to take into account for EHRI's further phases would be to also allow for parallel descriptions in local languages of both the EHRI Country Reports and descriptions of CHIs (at this moment this is only possible on archival descriptions), thereby letting archives and EHRI reach further local audiences. Furthermore, this has to be implemented alongside a wider languages-coverage policy in which the EHRI Portal interface gets translated into more local languages (in addition to the Czech and Spanish translations included during this funding phase).

For the EHRI Portal Country Reports, the suggestion to EHRI-ERIC would be to ensure that the Central Hub instals a Country Reports Revision Board to evaluate on a regular basis the content of the EHRI Country Reports History, Archival situation, and EHRI Research Summary sections. For the EHRI Research Extensive sections, it is advised to let the National Nodes update these should they wish so. If not, this could equally be done by the Country Reports Revision Board or the Extensive Report could be archived (with clear date-indication on when the last update was made).

For the identification and integration of new data or the updating of existing data, the adopted IHRA Guidelines for Identifying Relevant Documentation for Holocaust Research, Education and Remembrance[46] provide indications on what types of sources to include from a content perspective. Additionally, the guidelines for manually working within the EHRI Portal have been added to the EHRI Documentation[47], which equally includes the EHRI Portal Front- and Back-End Technical Documentation. Moreover, the EHRI Portal Data Model, which describes the standards for writing Country Reports and for describing Collection-Holding Institutions and Holocaust-relevant Collections can be found on a dedicated EHRI Portal page[48], but is also available when manually working on the admin pages of the EHRI Portal. These are very concrete tools for all involved in EHRI to ensure that relevant content is being added in a manner that meets EHRI's high quality standards. However, it would still be advisable to monitor the incoming content, both on a content and standards level and it is important to bring this into the Central Hub level in EHRI-ERIC. These suggestions could already be further fleshed out by the current EHRI-IP project, in order to make for a smooth transition of the EHRI Portal from EHRI-3 to EHRI-ERIC.

[46] https://holocaustremembrance.com/resources/guidelines-archival-documentation
[47] https://documentation.ehri-project.eu/en/latest/index.html
[48] https://portal.ehri-project.eu/help/datamodel