



**European Holocaust Research Infrastructure  
Preparatory Phase  
H2020-INFRADEV-2019-2  
GA no. 871060**

**Deliverable 7.4**

**Technical Report**

**Mike Bryant  
KCL / NIOD KNAW**

**Tobias Blanke  
KCL**

**Michael Levy  
USHMM**

**Start: December 2019 [M1]**

**Due: May 2022 [M30]**

**Actual: May 2022 [M30]**



**EHRI is funded by the European Union**

## Document Information

Project URL	<a href="https://www.ehri-project.eu">https://www.ehri-project.eu</a>
Document URL	<a href="https://www.ehri-project.eu/deliverables-ehri-pp-2019-2022">https://www.ehri-project.eu/deliverables-ehri-pp-2019-2022</a>
Deliverable	D7.4 Technical Report
Work Package	WP7
Lead Beneficiary	1 - KNAW
Relevant Milestones	MS3
Dissemination level	Public
Contact Person	Mike Bryant, <a href="mailto:michael.bryant@kcl.ac.uk">michael.bryant@kcl.ac.uk</a> , +44 (0)20 7848 4616
Abstract (for dissemination)	<p>This document provides an initial strategy for development of the EHRI-RI over the course of the implementation phase. We first describe the planned architecture of the EHRI-RI and review the current EHRI services which will, to a large extent, make up the RI's Central Hub. We then review the underlying systems infrastructure and how new technological developments will likely shape its evolution, and enumerate a set of additional services that will enhance the RI and its ability to operate effectively in a semi-distributed fashion.</p> <p>We then present a Service Integration Framework (SIF) which enumerates the RI's variety of interfaces for intra-service communication from a data publishing and a data capture standpoint, incorporating both current and future services. We then describe a number of exploratory and proof-of-concept applications that have been developed by EHRI to demonstrate potential uses for the proposed future RI services and how they will integrate in practice.</p>
Management Summary	<p>Recommendations within this document are contingent on various non-technical aspects that will shape the EHRI-RI, including its funding model, user-access model and organisational structure. The RI is not starting from a blank sheet but rather inheriting the</p>

existing legacy structure and practices of EHRI-1 and EHRI-2, and is also heavily contingent on the outcome of concurrent activities of EHRI-3.

The current approach to the provision of underlying technical systems — fully-managed IaaS services — is appropriate for the RI as it provides a high degree of flexibility and is straightforward to administer. In pessimistic funding scenarios dedicated servers may prove more cost-effective at the expense of flexibility and ease of administration. Over time and if funds are available, the RI may benefit from the adoption of more fully-managed database services to further lower administrative complexity. Despite the ongoing adoption of more containerised services it is unlikely EHRI-RI will shift to a fully-containerised infrastructure in the short to medium terms. The use of IaC will be expanded in order to increase visibility into infrastructure-wide configuration and provisioning, reduce key-person risk and harmonise automation techniques.

In addition to the services inherited from the EHRI projects 1-3, the RI will provide a number of additional services including: IIIF-compatible image support; a range of standardised geospatial APIs integrating with a repository for geospatial datasets; an expanded LOD framework; SSO support for platforms where IAM contributes significant administrative overhead; and the extension of existing internal metadata management tools for validation and XML conversion to external data providers.

The Service Integration Framework (SIF) enumerates the means by which the EHRI-RI will interact with RIs in the wider research community, EHRI's data provider partners, and downstream data consumers. As well as established EHRI data interfaces and those made available by the proposed new image, geospatial and LOD services, it incorporates ways in which the EHRI-RI could integrate with external RIs such as EOSC.

A range of proof-of-concept systems have been developed to explore the proposed new services including: the integration of a IIIF-compatible image server with the EHRI visualisations platform; setting up a test instance of the GeoNode geospatial data platform; the development of a client application for exploring EHRI's archival metadata powered by the EHRI Portal APIs; and integrating the Discourse discussion forum software with the portal via its Discourse Connect SSO protocol.

In appendix 1 we provide more detail about potential SSO topologies and the options for introducing SSO to EHRI's legacy platforms.

## History

<b>Version</b>	<b>Date</b>	<b>Reason</b>	<b>Revised by</b>
0.1	01/11/2021	Preliminary version	Mike Bryant
1.0	09/05/2022	General quality revisions Added section 2.3.6 on IaC cost-benefit analysis stemming from EHRI-PP MTR Replaced Figure 3 with a new and expanded version	Mike Bryant

## Table of contents

<b>1. Introduction</b>	<b>9</b>
<b>2. EHRI-RI Central hub</b>	<b>9</b>
2.1. Architecture	9
2.2. Central services	10
2.2.1. The EHRI Portal	10
2.2.2. Visualisations Platform	10
2.2.3. Document Blog	10
2.2.4. Digital Editions	10
2.2.5. Training Platform	11
2.2.6. Helpdesk	11
2.3. Technical systems	11
2.3.1. Server infrastructure	11
2.2.1.2. Fully managed / virtualised	11
2.2.1.1. Dedicated servers	12
2.2.1.3. Containerised systems	12
2.3.2. Storage	13
2.3.3. Databases	13
2.3.4. Automation and system administration	13
2.3.5. EHRI-RI system infrastructure: future directions	14
2.3.6. Cost-effectiveness considerations	15
2.4. Future EHRI-RI services	15
2.4.1. IIF support	16
2.4.2. Geospatial data	17
2.4.3. LOD services	17
2.4.4. Authentication and authorization services	17
2.4.5. EAD validation	18
2.4.6. Data conversion	18
<b>3. Service Integration Framework</b>	<b>20</b>
3.1. Data Publishing	21
3.1.1. Search API	21
3.1.2. GraphQL API	22

---

3.1.3. OAI-PMH server	22
3.1.4. EAD, EAC, and EAG	22
3.1.5. SKOS vocabularies	22
3.1.6. Blog RSS feed & REST API	22
3.1.7. Digital Editions REST API	22
3.1.8. SPARQL endpoint	22
3.1.9. Image APIs	23
3.1.10. Geospatial APIs	23
3.2. Data Capture	23
3.2.1. SKOS Vocabularies	23
3.2.2. Service Registry	23
3.2.3. Annotations and links	23
3.2.4. OAI-PMH harvesting	24
3.2.5. ResourceSync harvesting	24
3.2.6. Authentication/Identity attributes	25
<b>4. Proof of concepts</b>	<b>25</b>
4.1. IIIF Image API server	25
4.2. Portal search client	26
4.3. Discourse discussion forum	26
4.4. Geospatial Repository	27
<b>5. Concluding remarks</b>	<b>28</b>
<b>6. Appendix 1: Single Sign-On Configurations</b>	<b>31</b>
6.1. Third-party Identity Provider	31
6.2. RI Identity Provider	31
6.3. SSO Technologies	32
6.3.1. SAML	32
6.3.2. OpenID Connect	32
6.3.3. Discourse Connect	33
6.4. Implementation	33
6.4.1. The EHRI Portal	33
6.4.2. Wordpress	33
6.4.3. Omeka Classic	33

---

#### 6.4.4. Drupal

## Glossary

API	Application Programming Interface
AuthN	Authentication
AuthZ	Authorisation
CH	Central Hub
EAC	Encoded Archival Context
EAD	Encoded Archival Description
EAG	Encoded Archival Guide
EHRI	European Holocaust Research Infrastructure
EOSC	European Open Science Cloud
ISAD(G)	International Standard Archival Description (General)
ISAAR	International Standard Archival Authority Record
ISDIAH	International Standard Descriptions for Institutions with Archival Holdings
IaaS	Infrastructure as a Service
IaC	Infrastructure as Code
IAM	Identity & Access Management
IdP	Identity Provider
IIIF	International Image Interoperability Framework
LOD	Linked Open Data
NN	National Nodes
OAI-PMH	Open Archives Protocol for Metadata Harvesting
REST	Representational State Transfer
RI	Research Infrastructure
RP	Relying Party
SAML	Security Assertion Markup Language
SIF	Service Integration Framework
SKOS	Simple Knowledge Organisation System
SP	Service Provider
SSO	Single Sign-On
TEI	Text Encoding Initiative
VPS	Virtual Private Server
XML	eXtensible Markup Language



## 1. Introduction

Since its beginnings in 2010 and over the course of two subsequent funding phases EHRI's infrastructure has evolved and grown in scope, adapting to the changing needs of the project over time and with the changing technological landscape. The transition to a permanent Research Infrastructure (RI) will demand continued evolution in the services the EHRI-RI offers and the ways in which those services interact. The aim of this document is to explore how the RI can develop over the course of its implementation phase to better serve the needs of its users, to function more efficiently and effectively, and to align more closely with the broader digital research ecosystem.

The EHRI-PP deliverable D7.1: Survey of Technical Requirements<sup>1</sup> has reviewed various technical aspects of the EHRI-2 infrastructure with the development of the RI in mind and identified various areas for improvement. This document will not recapitulate these recommendations but instead take a more forward-looking approach to future development potential. While EHRI-2's existing services are the starting point for this document, the concurrent activities of the EHRI-3 project means that they are not standing still. Because EHRI-3 seeks to expand the scope of the infrastructure to make it more useful to researchers and other stakeholders, there will be significant overlaps between EHRI-3 activities and EHRI-RI services discussed below.

The following section will give an overview of the EHRI-RI architecture, discuss different approaches to the management and administration of the system resources on which it will run, and propose a range of additional services to broaden its capabilities and better meet current and future needs. Section three presents a Service Integration Framework (SIF) that offers a view of the proposed EHRI-RI infrastructure from the perspective of data flows and interfaces to data providers and higher-level e-Research infrastructures. Finally we describe a number of prototype services that have been developed to test service integration and offer some concluding remarks.

A preliminary draft of this report was previously made available in month 24.

## 2. EHRI-RI Central hub

### 2.1. Architecture

While the precise legal and administrative form of the final EHRI-RI is still being developed, on a technical level the infrastructure will be a distributed architecture consisting of a Central Hub (CH) providing services to consortia-administered National Nodes (NN). The services managed by the CH will include the central catalogue of transnational archival metadata (the EHRI Portal) and existing services inherited from EHRI-1, 2, and 3 project activities, as well as the additional future services enumerated below in section 2.4. On a technical level, CH and NN services will be hosted on different top-level domains (TLDs) and be capable of independent administration. Data transfer and functional interactions will be based on well-defined, publicly-documented interfaces. NNs and their constituent services may be granted privileged conditions of access — such as higher rate limits<sup>2</sup> — where non-trivial costs result from use of resources such as bandwidth and disk space.

---

1

<https://www.ehri-project.eu/sites/default/files/downloads/Deliverables/D7.1%20-%20Survey%20of%20Technical%20Requirements.pdf>

<sup>2</sup> For example, the number of times an API may be accessed in a given time-frame

## 2.2. Technical systems

Underlying the service infrastructure of the EHRI-RI will be a collection of lower-level compute, storage and networking resources. EHRI's IT architecture has evolved significantly since its inception and may continue to evolve throughout the preparatory and implementation periods in order to become more secure, cost-effective, or flexible. This section will briefly outline a range of approaches to infrastructure resourcing and outline their primary benefits and drawbacks.

### 2.2.1. Server infrastructure

At the scale of EHRI-RI, there are a number of different models for running an IT infrastructure in which cost is primarily traded against administrative overhead. The picture is not entirely straightforward, however, due to the multiple axes on which administrative overhead can weigh, taking into account the availability of expertise and experience of key staff. The three scenarios below present a simplified picture of how varying degrees of buy-in to “managed services” can be manifested, and their impact on costs.

#### 2.2.1.2. Fully managed / virtualised

A fully-managed infrastructure is the model under which the EHRI's services currently operate, using DigitalOcean<sup>3</sup> as the Infrastructure-as-a-Service (IaaS) provider. Fully-managed infrastructure uses a virtualisation layer atop physical hardware to create virtual machines on demand, allowing bare-metal servers to be divided up into many virtual private servers (VPSs) that are apportioned between multiple tenants. The key advantage to virtualised systems is their flexibility: servers can be created and destroyed rapidly and on-demand, billed by the minute, meaning that VPSs can be made available to precisely fit the clients' specific needs in terms of CPU, memory, storage or operating system for even very short-term needs. This flexibility makes it more practical to use a VPS-per-service model, significantly simplifying the deployment of services and making it much easier to scale them up by adding additional instances.

Disadvantages of fully-managed systems include less predictable performance, since the virtualisation layer sits between the server and the physical hardware and can be affected by other demands being made on it at any time, potentially by other tenants.<sup>4</sup> This can be particularly notable where access to storage (disk in/out) is concerned. While resource elasticity can make short-term use of resources cost-effective, for longer term situations prices for virtualised servers can be higher than dedicated alternatives. Bandwidth charges, common among IaaS providers, can also contribute towards price unpredictability, since a long- or short-term spike in the popularity of a service may substantially increase monthly billings.

#### 2.2.1.1. Dedicated servers

Dedicated servers which are rented by datacenters to a specific customer (a “single tenant”) can offer superior performance, security and — for some use-cases — be a cost-efficient way to run non-trivial IT systems. Without the overhead of virtualisation, it is possible to achieve performance that is consistent and aligns closely with the theoretical limitations of the hardware, without having to worry about the maintenance of the physical hardware itself (which typically resides in a datacenter.) For scenarios where security and regulatory compliance is a primary consideration, a dedicated server can also be the best option, since there is physical segregation from other datacenter users.

---

<sup>3</sup> <https://digitalocean.com>

<sup>4</sup> Leitner & Cito, 2016

In constrained financial scenarios a small number of dedicated servers *could* present EHRI-RI with the best value for money, i.e. the ability to obtain the highest specification compute, storage and memory resources for the least money.

A significant drawback of dedicated infrastructure is, however, the lack of flexibility and, as a consequence, additional administrative overhead. The use of a single dedicated server (or at most, a handful of servers) complicates the deployment and provisioning of services by requiring them to share the same operating system and networking environment (such as the IP address), and increases the chance of problems caused by incompatible dependencies between services. Running more services in a shared environment also increases the likelihood of extensive downtime due to upgrades and maintenance on the dedicated machine.

### **2.2.1.3. Containerised systems**

Taking virtualisation a step further are container orchestration systems. Container systems, such as Docker<sup>5</sup>, facilitate the packaging of services into discrete units that can access operating system-level resources directly but which otherwise run in a sandboxed manner. A separate network layer typically allows containers to communicate with each other in a segregated fashion. Containerisation has many benefits in terms of packaging services and their dependencies in a portable, reusable manner that can significantly simplify deployment, properties that have made them popular in development environments. For production use, the additional layer of abstraction added by the container runtime and the need to ensure storage persistence can add extra complexity and make monitoring services more difficult, sometimes outweighing the benefits of encapsulation.

In contrast to treating containerised services as a straightforward alternative to OS-level native services, container orchestration systems, such as Kubernetes<sup>6</sup>, take a higher-level approach to configuration and management of individual container units and in doing so can provide many benefits from an administration standpoint. These include highly centralised, declarative configuration and the ability to automatically scale up services, deploying more server instances as demand increases.

A practical drawback to the use of container orchestration systems is their relative newness, novelty, and rapid evolution, and the scarcity of expertise in their deployment and management that this entails. We expect that this will change as the technology stabilises and matures.

### **2.2.2. Storage**

Like servers, storage is available in variants that are more or less managed by the provider. In most IaaS systems, VPSs can be augmented with additional block storage on demand for performance-sensitive requirements such as databases. Fully managed Cloud-based storage, such as AWS's S3 and compatible services provide a range of convenience features (such as automatic versioning and object retention) that make them attractive for uses such as off-site backup and web content delivery, with a significantly lower burden of administration.

### **2.2.3. Databases**

Databases are also available in self-hosted or fully-managed varieties, with the latter providing a lower burden of administration (the provider taking care of routine backup, access management, and performance-tuning considerations) for a given hourly fee. At present,

---

<sup>5</sup> <https://docker.com>

<sup>6</sup> <https://kubernetes.io>

managed database solutions offered by EHRI's current IaaS provider (DigitalOcean) are not cost-efficient given the number of databases EHRI maintains and their characteristics such as size and workload, however this consideration may require reevaluation as the RI moves towards implementation.

#### **2.2.4. Automation and system administration**

While non-trivial IT infrastructures have always been subject to some degree of automated system administration, the shift towards fully virtualised IaaS systems has led to new tools becoming available to manage servers and their related configuration in more systematic and reproducible ways. Termed "Infrastructure as Code" (IaC), these tools typically allow managing IaaS systems via centralised, machine-readable definition files which, like other types of code, are subject to techniques such as version control and continuous integration.

While automation in general is often carried out to speed up operations and improve efficiency, a more significant advantage of systematic IaC is in codifying implicit knowledge and increasing the visibility of existing practices throughout an organisation. Since so much of system administration can be carried out interactively, by an administrator making changes to the configuration of running systems via remote consoles, this leads to a concentration of implicit knowledge about the state of the system that is time-consuming to adequately document and often difficult to reproduce consistently. Concentrations of implicit knowledge increase key person risk within an organisation, exposing it to potential difficulties resulting from staff turnover.

By ensuring that provisioning and configuration of systems occurs primarily through accessible, explicit, and repeatable mechanisms, an organisation can make its infrastructure more auditable and robust to disruption and reduce the likelihood of mistakes or misadministration. Moreover, the use of configuration languages such as YAML<sup>7</sup> that aim for a high level of human (as well as machine) readability further ensures that IaC has a self-documenting function, in contrast to actions carried out interactively.<sup>8</sup> A secondary benefit is that once defined, a certain set of configuration actions can easily be reused, or serve as a template for similar processes elsewhere.

The cost of introducing IaC primarily stems from the learning curve associated with tools like Terraform and Ansible, along with the addition of more points of failure and extra indirection between updating centralised configuration definitions and them being rolled out across the infrastructure. Some types of configuration can also be complex to centralise and orchestrate systematically, such as encryption certificates for peer-to-peer communication between resources. Mitigating these issues somewhat are two factors: firstly, IaC is becoming the norm throughout organisations of all sizes, driven by trends such as IaaS, DevOps, continuous integration systems, and containerisation.<sup>9</sup> Secondly, IaC does not in most cases preclude reverting to manual configuration when, for example, an emergency arises and quick fixes must be deployed without delay.

#### **2.2.5. EHRI-RI system infrastructure: future directions**

At present EHRI infrastructure runs on managed (virtual) servers but uses *predominantly* self-hosted databases, storage, and OS-level (non-containerised) services. With an increasing number of infrastructure vendors and OSS projects opting to push containerised solutions (primarily since they are easier to deliver in a cross-platform manner) it is likely that EHRI will adopt more containerisation in the future. Taking this to the logical extreme,

---

<sup>7</sup> <https://yaml.org>

<sup>8</sup> Cito et al. 2015 p397

<sup>9</sup> Erich et al. 2014

however, and moving to a full container orchestration system is unlikely in the short- and medium-terms, at least until the technology matures further.

In an optimistic funding scenario, and assuming legal considerations regarding data locality permit it, the EHRI-RI would likely adopt more fully-managed database and storage solutions in order to reduce the overhead involved in managing these systems.

EHRI has introduced IaC for provisioning of infrastructure components using an open-source orchestration tool called Terraform.<sup>10</sup> Definition files that define the type of servers EHRI uses, including their CPU, memory and disk specifications, exist alongside definitions for EHRI's web domains and the DNS settings that map them to server IP addresses, all on Github under version control. When changes are made, such as the addition of new servers or subdomains, Terraform will use the declarative definition files to determine the precise sequence of steps to take using the DigitalOcean's API. Whilst IaaS portability is not a primary concern, Terraform's support for alternative IaaS providers such as Linode, AWS and Microsoft Azure does provide the basis for a mixed or heterogeneous IaaS infrastructure if migration were to become necessary.<sup>11</sup>

IaC for resource *configuration* — as opposed to provisioning — is being introduced on an incremental basis using an open-source system administration tool called Ansible.<sup>12</sup> While non-managed configuration still exists for a number of EHRI's central services, the deployment and configuration of new services is fully IaC-oriented. As the number of EHRI's sites has grown with the introduction of (for example) the Digital Edition platform, a system where each installation comprises a multitude of separate services and plugins, this has reduced the overall administrative overhead. As more systems are migrated to IaC up to and into the RI implementation phase we expect to yield more benefits in terms of easier software updates, more consistency across systems, faster disaster recovery, and better knowledge sharing.

#### **2.2.6. Cost-effectiveness considerations**

While a financial analysis of infrastructure costs is outside the scope of this report, EHRI-PP's mid-term review raised questions about the costs and benefits of EHRI's general IaC approach that are worth discussing here - in particular it is worth discussing why staff resources are being directed towards IaC rather than services and systems more tangible to EHRI's user community.

As mentioned above, over its ten years of operation and three funded phases EHRI's activities have grown substantially, as new services have been introduced while legacy systems have — with few exceptions — been maintained in order to ensure continuity of access. The growth in the number of services and, as a consequence, the amount of complexity involved in maintaining the infrastructure as a whole has led to a corresponding increase in administrative overhead. Legacy services also acquire *technical debt*, a metaphor commonly used to refer to the ongoing need to refactor, rearchitect, and redesign software systems over time in order to ensure they can remain fit-for-purpose and maintainable (the tendency for mature software to become difficult to maintain, extend, and adapt is the *interest* that must be paid on the *debt*.)<sup>13</sup>

---

<sup>10</sup> <https://www.terraform.io>

<sup>11</sup> EHRI currently maintain some degree of heterogeneity in its IaaS systems by using AWS S3 for “off-site” storage, also managed through Terraform

<sup>12</sup> <https://www.ansible.com>

<sup>13</sup> Kruchten et al. 2012

Larger and more complex infrastructures are also more difficult to *secure* due to an increased surface area for attacks and greater scope for vulnerabilities, which can affect all aspects of the stack, from the hardware (managed by IaaS in EHRI's case), the operating system software and services, to end-user applications. Platforms such as Wordpress and Drupal, both of which are used by EHRI, are commonly affected by security vulnerabilities due to their widespread use and sprawling ecosystems of community-contributed code. An additional concern, alongside the risk of malicious attacks, is increased vulnerability to accidental damage caused by misconfiguration or mistakes leading to data loss, data leakage, or service downtime.

Whilst IaC has the direct benefits discussed above, such as increased administrative efficiency and enhanced knowledge transfer, it is also one prong of a general strategy intended to reduce EHRI-RI's long-term susceptibility to technical debt and security vulnerabilities by making changes to the infrastructure easier to make (or, as the case may be, revert.) By dramatically reducing the time it takes to restore systems to a known working state from hours to minutes we can sleep easier taking the assumption that mistakes will inevitably be made, or even that data loss or malicious hacks will at some point occur. Efforts such as migrating to IaC, therefore, are a long-term investment in a more stable, resilient, and future-proof infrastructure, as befits EHRI in its transition from a time-limited project to a self-sustaining organisation.

### 2.3. Central services

The EHRI-RI Central Hub will be an evolution of EHRI's current infrastructure. At its core will be a number of existing services that have been developed throughout EHRI's three project phases:

#### 2.3.1. The EHRI Portal

The EHRI Portal<sup>14</sup> is a central repository of metadata about Holocaust-related archival collections, archival institutions, and information about the archival situation in countries of particular relevance to Holocaust researchers. For indexing purposes it also manages controlled vocabularies of Holocaust-related subject terms, ghettos and camps.

Information in the EHRI Portal is managed via a suite of administrative tools that facilitate the curation of data via both interactive and semi-automated methods, including the ingest of XML-encoded metadata obtained via regular harvesting of third-party sources. A range of different APIs are likewise available to search and retrieve collection metadata from the portal (see section 3.1. Data Publishing).

The portal also allows users to register for a free account providing them with additional functionality such as the ability to create a profile containing their research interests and affiliation, to contact other registered users, and to save items of interest. The portal's user accounts are also the basis for administrative functionality that provides EHRI staff with control over the collection holding catalogue metadata in a flexible role-based manner.

#### 2.3.2. Visualisations Platform

EHRI's visualisations platform<sup>15</sup>, based on the Omeka publishing platform and the Neatline geospatial plugin, facilitates the creation of interactive annotated maps or documents that can be embedded into other websites. It is most commonly used to provide interactive visual content for the EHRI Document Blog.

---

<sup>14</sup> <https://portal.ehri-project.eu/>

<sup>15</sup> <https://visualisations.ehri-project.eu/>

### 2.3.3. Document Blog

The EHRI Document Blog<sup>16</sup> is a Wordpress site with an EHRI-specific visual theme and a number of custom plugins tailored to EHRI's specific publishing requirements. Primarily a platform for narrative explorations of specific Holocaust-related topics, the Document Blog can also embed interactive media (such as annotated documents or maps created by the Visualisations Platform) and information drawn from the EHRI Portal via a plugin that connects to the portal's APIs.

### 2.3.4. Digital Editions

EHRI's Digital Editions platform, also based on Omeka, provides the means to create richly interactive sites from document scans and transcripts marked up using TEI. Incorporating images, maps and contextual narrative the documents connect users to authority files, subject terms and archival metadata on the EHRI Portal and other Holocaust-related resources on the web.

At the time of writing there are three public digital editions: *Begrenzte Flucht*<sup>17</sup>, *Diplomatic Reports*<sup>18</sup>, and *Early Holocaust Testimony*<sup>19</sup> — with more to come over the course of the EHRI-3 project.

### 2.3.5. Training Platform

EHRI's training site<sup>20</sup> is, like the project website, a Drupal-based CMS that hosts a number of self-guided courses on both Holocaust-related and research-focused topics. At the time of writing six courses are available on historical topics, plus an illustrated manual for using the EHRI Portal and an introduction to Cultural Analytics using the R programming language. As befits their focus on primary sources, the courses make heavy use of image-based material such as document scans.

### 2.3.6. Helpdesk

The EHRI helpdesk<sup>21</sup> serves as a first point of contact for enquiries, requests and feedback relating to EHRI's user-facing websites. Queries are retained in a ticketing system (based on the open source OSTicket software) and forwarded, when necessary, to relevant experts within the consortium.

## 2.4. Future EHRI-RI services

While EHRI's existing services will form the core of the EHRI-RI Central Hub, they will be further developed and augmented with new capabilities. This section enumerates a set of additional data services and systems that will enhance the experience of data consumers across the distributed infrastructure:

### 2.4.1. IIR support

While the EHRI Portal is focused primarily on archival *metadata*, a number of EHRI's digital activities, including the Digital Editions, Document Blog and training courses, are oriented to a significant degree around archival *content*. If it is assumed that some of the future activities of EHRI-RI National Nodes will also be similarly thematic and content-focused, and also

---

<sup>16</sup> <https://blog.ehri-project.eu/>

<sup>17</sup> <https://begrenzte-flucht.ehri-project.eu/>

<sup>18</sup> <https://diplomatic-reports.ehri-project.eu/>

<sup>19</sup> <https://early-testimony.ehri-project.eu/>

<sup>20</sup> <https://training.ehri-project.eu/>

<sup>21</sup> <https://helpdesk.ehri-project.eu/>

involve the delivery of scanned documents, photographs or audio/visual material, it is worth considering what value services offered by the Central Hub can bring in facilitating digital content delivery.

EHRI's current content-oriented platforms, while suitable in their specific roles, inevitably tend towards the siloing of the media they are tasked with managing. Images embedded in a blog post will be uploaded and managed by Wordpress, training example images by Drupal, and Digital Edition document scans by Omeka. Each of these platforms have different storage requirements and capabilities with regard to the deposit of multimedia material, the application of metadata to it, and its eventual dissemination.

While some degree of heterogeneity in the management of digital media across a range of services is an inevitable part of multi-platform web architectures, it invariably entails more fragmentation for the user seeking archival material. As EHRI as a whole expands into hosting more archival content the EHRI-RI can seek to limit such fragmentation, as well as better adhering to FAIR guidelines in media management, by providing central services for deposit and delivery of image, audio or video material.

EHRI-RI image services will be based around APIs that make up the International Image Interoperability Framework (IIIF, pronounced "triple-eye-eff")<sup>22</sup>. First proposed in 2011 as a collaboration between a multinational group of libraries, IIIF is today in widespread use — including by EHRI partners USHMM — for making image-based material such as manuscripts and archival documents accessible over the web.

Applications of IIIF APIs include making high-resolution uncompressed images of potentially very large size accessible over the web in an efficient manner via dynamic scaling and compression, allowing users to "deep zoom" into images and navigate around them. As well as dealing with single images it also caters to multi-image documents (such as, for example, books or manuscripts), allowing users to navigate, annotate, and link to component parts or the document as a whole. Newer versions of some IIIF APIs also have a degree of support for time-based audio-visual material.

The use of standardised APIs for advanced image-based functionality means that image data spread across distributed sources — for example, archival content hosted by EHRI data providers such as USHMM, or EHRI-RI National Nodes — can be mixed and combined to form composite presentations. Physically separated and dispersed Holocaust-related material, if available in a IIIF-compatible manner, could be recombined in a manner analogous to the "virtual collection finding aids" hosted by the EHRI Portal.<sup>23</sup>

IIIF consists of both backend and frontend systems, with the backends including both servers capable of rendering image data in response to remote web requests and services for creating IIIF manifests — parcels of metadata that describe a resource or composite resources. Frontend systems include "viewers" that consume compatible metadata and present a navigation interface to the user through which resources can be explored. A range of off-the-shelf, open-source tools are available for both backend and frontend requirements, including plugins for CMSs like Wordpress, Drupal, and Omeka. The combination of a centralised IIIF-compatible image server and viewer plugins for EHRI's existing content-management systems would allow image-based material to be hosted and disseminated in a more robust, interoperable, and reusable manner than is currently the case. We discuss a proof-of-concept implementation in section 4.1.

---

<sup>22</sup> <https://iiif.io>

<sup>23</sup> Bryant et al. 2015



### **2.4.2. Geospatial data**

Many of the themes explored on the EHRI Document Blog and Digital Editions are geographic and spatial in nature. To enhance the RI's geospatial capabilities and open them up to the wider Holocaust research community, a new EHRI-RI repository will be established to bring together spatial datasets and provide data services oriented around mapping and geolocated information. Authorised users will be able to upload spatial datasets (such as geolocated tabular data in standardised formats) that will be publicly findable and retrievable via a user-friendly web interface and a structured data REST-style API. Lower-level mapping APIs standardised by the Open Geospatial Consortium (OGC), such as WMS (Web Map Service) and WFS (Web Feature Service), will be supported to provide researchers with access to the data in industry-standard map visualisation and analysis tools such as ArcGIS or QGIS.<sup>24</sup>

### **2.4.3. LOD services**

While EHRI's data does connect to external datasets such as DBPedia<sup>25</sup>, the level of interlinking is limited to certain SKOS-format controlled vocabularies and there does not yet exist the means to run queries using protocols such as SPARQL which are tailored for Linked Open Data (LOD) applications.

The EHRI-RI will enhance the interoperability and reusability of its data by increasing its interconnectedness with external datasets such as DBPedia and Geonames, developing ways to express the more straightforward aspects of its archival data using existing high-level LOD ontologies such as schema.org<sup>26</sup>, and offering improved support for LOD querying.

### **2.4.4. Authentication and authorization services**

One of the most significant challenges in maintaining a distributed set of related IT services is dealing with user identities and associated authentication ("AuthN") and authorisation ("AuthZ") data in a manner that is scalable and able to be efficiently and securely managed. With all but very simple services having the notion of a "user", typically identified by a handle such as an email address and authenticated by a password, the deployment of multiple sites and services has the tendency to lead to a proliferation of such information in distinct databases. When a user logs into one site on the EHRI domain they might expect to be able to use their credentials with other sites on the same domain; likewise, when they change their password for one site they might expect this change to be propagated to other sites on the same domain, regardless of the specific platform being used. Such functionality is known variously as Identity and Access Management (IAM), or, when actualised across multiple services, Single Sign-On (SSO).

Over the course of its three funded phases the EHRI project has deployed a number of websites based on different platforms, including bespoke services like the EHRI Portal and those based on open source platforms such as Wordpress, Drupal and Omeka. Each of these systems incorporates a way of identifying users with various privilege levels — for example: contributors, editors and administrators — and corresponding authentication information. At the time of writing, however, there is no centralised IAM system in use, despite there being an overlap between the multiple sets of users. The reasons for this amount to the development costs of adding or enabling SSO functionality being judged to be greater than the overhead resulting from the management of duplicate accounts, given a lack of existing SSO support for some components.

---

<sup>24</sup> <https://www.qgis.org/en/site/>

<sup>25</sup> <https://www.dbpedia.org>

<sup>26</sup> <https://schema.org>

As the EHRI infrastructure continues to grow it is worth reassessing the value of SSO. For the EHRI-RI, having the possibility to employ SSO for both central services and the decentralised services offered by National Nodes, with the Central Hub acting as an identity provider (IdP), would significantly reduce the friction involved in managing authentication and authorization, and reduce the proliferation of potentially sensitive login information. Centralised sign-on would also permit the use of more secure login techniques, such as two-factor authentication (2FA), across the RI as a whole.

The key challenges involved in assessing the feasibility of SSO functionality are:

- weighing the benefits of delivering an SSO solution against the engineering and administrative costs
- selecting the most appropriate technical solution from a range of possibilities, including possible hybrid solutions that mix two or more solutions in different AuthZ and AuthN contexts

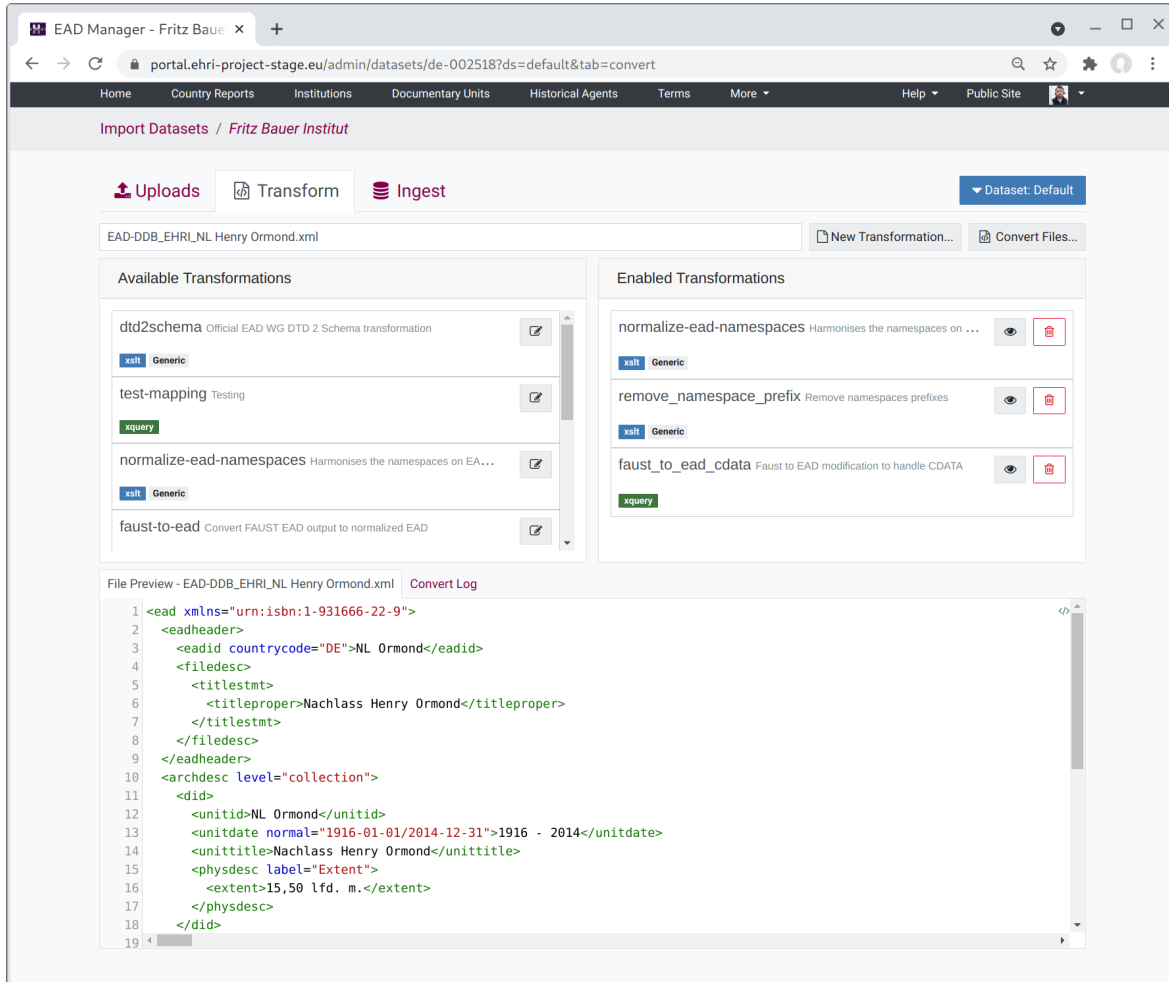
Appendix 1 contains a brief discussion of the available SSO technologies and potential configurations for an EHRI-RI SSO system.

#### **2.4.5. EAD validation**

Data providers with an account on the EHRI Portal will be able to interactively validate (using a web interface) archival metadata in EAD format according to EHRI-RI's specific guidelines. These validation rules will be a superset of general EAD rules, i.e., the rules are more stringent in certain particulars. This ensures that metadata which passes validation will be fully EAD schema-compliant and will also more easily reusable in other contexts.

#### **2.4.6. Data conversion**

Data providers will be able to use EHRI-RI's services for building and testing metadata crosswalks and running batch conversion using their own input data. EHRI-RI data conversion services will provide the ability to create chainable conversion pipelines consisting of many discrete XSLT or XQuery-based conversion steps (see Figure 1 for an example.) Users will be able to develop data transformations in an interactive manner with a user interface that provides live feedback as to the state of the input and output data given the active transformations (see Figure 2.)



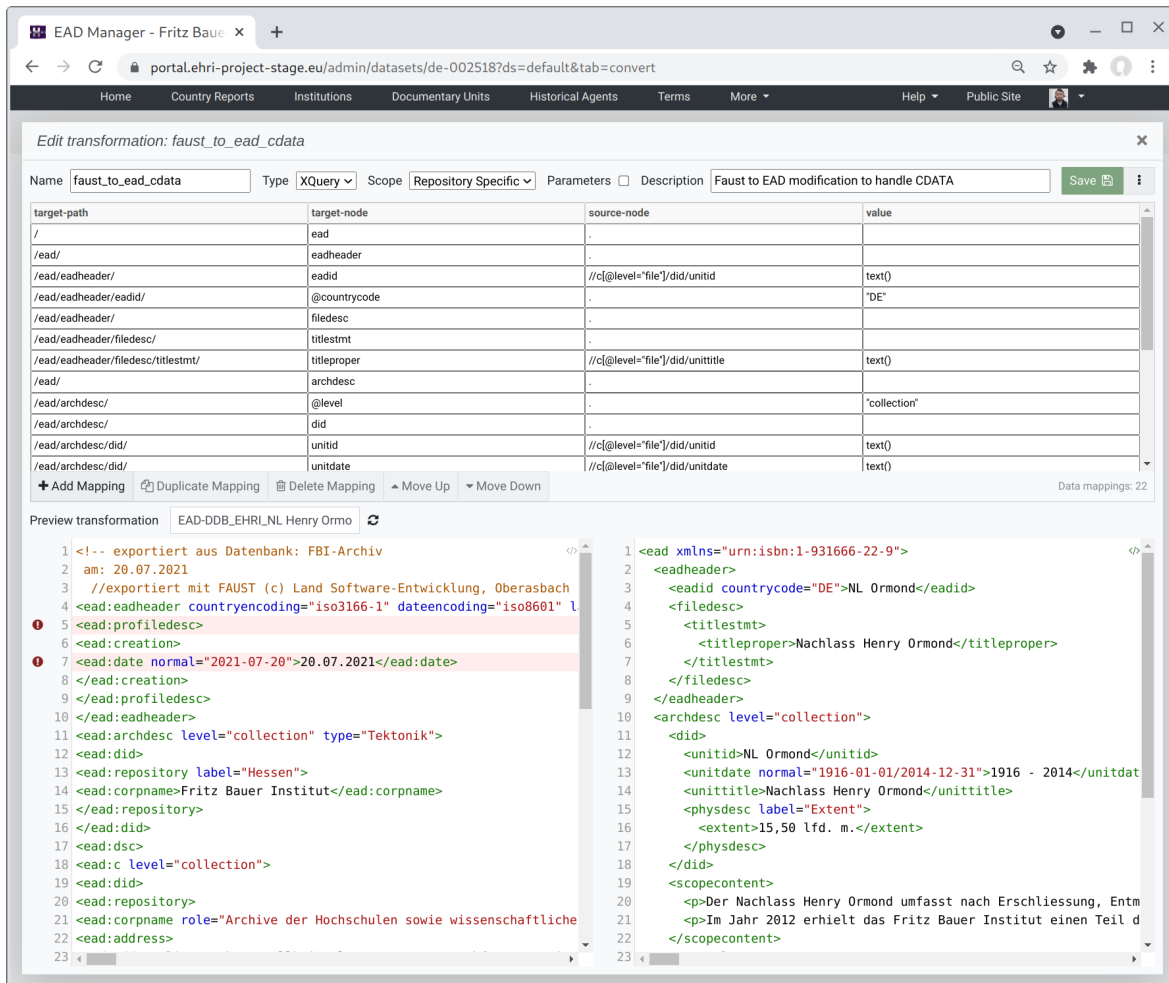
The screenshot shows the EAD Manager interface for a dataset named 'EAD-DDB\_EHRLNL Henry Ormond.xml'. The interface is divided into several sections:

- Navigation:** Home, Country Reports, Institutions, Documentary Units, Historical Agents, Terms, More, Help, Public Site.
- Actions:** Uploads, Transform, Ingest. A 'Dataset: Default' dropdown is also present.
- Available Transformations:**
  - dtd2schema:** Official EAD WG DTD 2 Schema transformation (xslt, Generic)
  - test-mapping:** Testing (xquery)
  - normalize-ead-namespaces:** Harmonises the namespaces on EA... (xslt, Generic)
  - faust-to-ead:** Convert FAUST EAD output to normalized EAD (xslt, Generic)
- Enabled Transformations:**
  - normalize-ead-namespaces:** Harmonises the namespaces on ... (xslt, Generic)
  - remove\_namespace\_prefix:** Remove namespaces prefixes (xslt, Generic)
  - faust\_to\_ead\_cdata:** Faust to EAD modification to handle CDATA (xquery)
- File Preview:** EAD-DDB\_EHRLNL Henry Ormond.xml. The XML content is displayed with line numbers 1 through 19. The XML structure includes:
 

```

1 <ead xmlns="urn:isbn:1-931666-22-9">
2 <eadheader>
3 <eadid countrycode="DE">NL Ormond</eadid>
4 <filedesc>
5 <titlestmt>
6 <titleproper>Nachlass Henry Ormond</titleproper>
7 </titlestmt>
8 </filedesc>
9 </eadheader>
10 <archdesc level="collection">
11 <did>
12 <unitid>NL Ormond</unitid>
13 <unitdate normal="1916-01-01/2014-12-31">1916 - 2014</unitdate>
14 <unittitle>Nachlass Henry Ormond</unittitle>
15 <physdesc label="Extent">
16 <extent>15,50 lfd. m.</extent>
17 </physdesc>
18 </did>
19 </archdesc>
      
```

Figure 1: the EHRI-RI data transformation interface will allow multiple discrete metadata transformations to be run in a pipeline manner in order to convert from one format to another.



**EAD Manager - Fritz Bauer**

portal.ehri-project-stage.eu/admin/datasets/de-0025187ds=default&tab=convert

Home Country Reports Institutions Documentary Units Historical Agents Terms More Help Public Site

**Edit transformation: faust\_to\_ead\_cdata**

Name:  Type: XQuery Scope: Repository Specific Parameters: Description: Faust to EAD modification to handle CDATA Save

target-path	target-node	source-node	value
/	ead	.	
/ead/	eadheader	.	
/ead/eadheader/	eadid	//c[@level="file"]/did/unitid	text()
/ead/eadheader/eadid/	@countrycode	.	'DE'
/ead/eadheader/	filedesc	.	
/ead/eadheader/filedesc/	titlestmt	.	
/ead/eadheader/filedesc/titlestmt/	titleproper	//c[@level="file"]/did/unittitle	text()
/ead/	archdesc	.	
/ead/archdesc/	@level	.	'collection'
/ead/archdesc/	did	.	
/ead/archdesc/did/	unitid	//c[@level="file"]/did/unitid	text()
/ead/archdesc/did/	unitdate	//c[@level="file"]/did/unitdate	text()

Data mappings: 22

Preview transformation: EAD-DDB\_EHRI\_NL Henry Ormo

```

1 <!-- exportiert aus Datenbank: FBI-Archiv
2 am: 20.07.2021
3 //exportiert mit FAUST (c) Land Software-Entwicklung, Oberasbach
4 <ead:eadheader countryencoding="iso3166-1" dateencoding="iso8601" l
5 <ead:profiledesc>
6 <ead:creation>
7 <ead:date normal="2021-07-20">20.07.2021</ead:date>
8 </ead:creation>
9 </ead:profiledesc>
10 </ead:eadheader>
11 <ead:archdesc level="collection" type="Tektonik">
12 <ead:did>
13 <ead:repository label="Hessen">
14 <ead:corpname>Fritz Bauer Institut</ead:corpname>
15 </ead:repository>
16 </ead:did>
17 <ead:dsc>
18 <ead:c level="collection">
19 <ead:did>
20 <ead:repository>
21 <ead:corpname role="Archive der Hochschulen sowie wissenschaftliche
22 <ead:address>
23
  
```

```

1 <ead xmlns="urn:isbn:1-931666-22-9">
2 <eadheader>
3 <eadid countrycode="DE">NL Ormond</eadid>
4 <filedesc>
5 <titlestmt>
6 <titleproper>Nachlass Henry Ormond</titleproper>
7 </titlestmt>
8 </filedesc>
9 </eadheader>
10 <archdesc level="collection">
11 <did>
12 <unitid>NL Ormond</unitid>
13 <unitdate normal="1916-01-01/2014-12-31">1916 - 2014</unitdate>
14 <unittitle>Nachlass Henry Ormond</unittitle>
15 <physdesc label="Extent">
16 <extent>15,50 lfd. m.</extent>
17 </physdesc>
18 </did>
19 <scopecontent>
20 <p>Der Nachlass Henry Ormond umfasst nach Erschliessung, Entm
21 <p>Im Jahr 2012 erhielt das Fritz Bauer Institut einen Teil d
22 </scopecontent>
23
  
```

Figure 2: the EHRI-RI data transformation editor will allow interactive editing of metadata crosswalks.

### 3. Service Integration Framework

This section will outline a Service Integration Framework (SIF): a set of data pathways into and out of the planned EHRI-RI that will enable it to connect with providers of Holocaust-related information, to a wider ecosystem of Linked Open Data (LOD), and to the wider EU research community. We will first consider outgoing pathways (publishing) and then incoming (data capture).

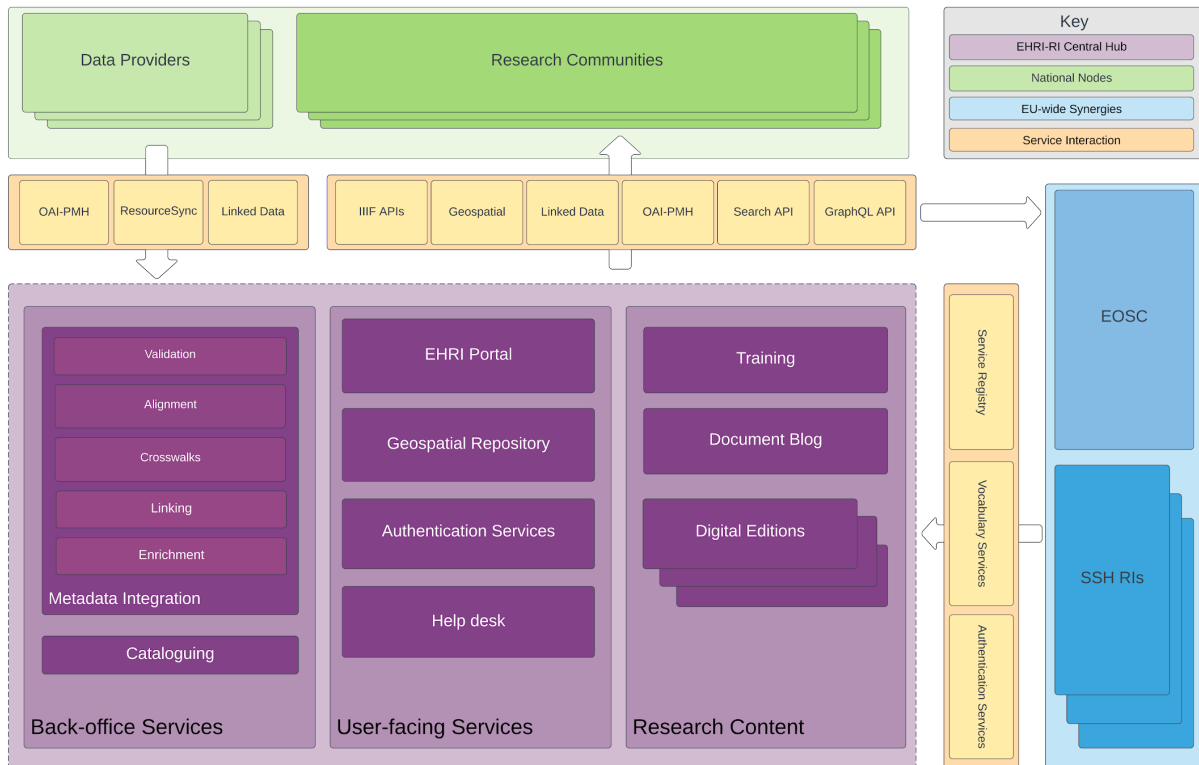


Figure 3: A high-level overview of the EHRI-RI SIF (Service Integration Framework), showing the data interactions between the EHRI-RI Central Hub, National Nodes, and the wider EU research ecosystem.

### 3.1. Data Publishing

The RI Central Hub will continue to support the existing structured data interfaces offered by EHRI in its current guise, augmenting them with additional Linked Data services to better integrate EHRI with large open datasets such as DBpedia and Geonames, as well as the wider semantic web.

The availability of heterogeneous forms of structured data stems from the range of use-cases at which the various interfaces are targeted.

#### 3.1.1. Search API

The EHRI-RI Search API will conform to the JSON:API<sup>27</sup> specification and will provide a succinct way to search the RI's archival metadata using REST-style URL queries. When used without a search query it will default to listing all items available in a paginated manner, and will also serve to retrieve single items and items nested in a given hierarchical scope (e.g. items belonging to a particular archival collection) using their IDs. While the principal use-case will be ad-hoc data retrieval using command-line HTTP tools such as Curl, the Search API will also enable integration with Wordpress and Omeka via plugins for those systems, providing the means to embed metadata from the EHRI Portal into external CMS-powered sites.

<sup>27</sup> <https://jsonapi.org>

An OpenAPI<sup>28</sup>-compatible interface specification will additionally be provided to enable third-parties to analyse the Search API's capabilities in a machine-readable manner and to facilitate automatically-generated interactive documentation websites.

### **3.1.2. GraphQL API**

GraphQL is a schema-driven domain-specific language originally invented by Facebook and now in widespread use around the web. The EHRI-RI GraphQL API will be an interface focused purely on data retrieval, with a considerably greater scope than the Search API and the ability to traverse networks of interrelated objects with fewer round-trip API requests. Used in combination with the Search API, it will grant access to most data accessible on the EHRI Portal website, but in a structured, programmatic manner. The GraphQL API will also be used by the Digital Editions platform to fetch structured data about EHRI authorities and subject terms.

### **3.1.3. OAI-PMH server**

OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) is a well-established system for making object metadata accessible via an XML-over-HTTP interface. The EHRI-RI will provide an OAI-PMH interface to grant third-party aggregators the means to harvest metadata about archival collections in the EHRI Portal in either Dublin Core or EAD format. EHRI-RI's OAI-PMH implementation will support the full range of capabilities defined by the standard, including time-based querying and enumeration of previously deleted items.

### **3.1.4. EAD, EAC, and EAG**

For third-parties wishing to retrieve standardised XML representations of EHRI-RI domain entities, URL-based access to EAD, EAC and EAG data will be provided for archival descriptions, authority files and archival institutions respectively.

### **3.1.5. SKOS vocabularies**

EHRI-RI's controlled vocabularies will be modelled as SKOS concept schemes and will be available to download as RDF (Resource Description Format) triples in either RDF+XML or Turtle format. SKOS data can be commonly imported into other archival data management systems such as Access to Memory (AtoM).

### **3.1.6. Blog RSS feed & REST API**

Programmatic access to summaries and metadata about posts on the EHRI-RI Document Blog will be available via an RSS (Really Simple Syndication) feed, a well-widely supported XML-based format. Full data for blog posts will be secondarily available via JSON-based REST API.

### **3.1.7. Digital Editions REST API**

The Digital Editions platform will incorporate a REST-style API to facilitate JSON-based programmatic access to document metadata, image data, transcripts, and search functionality.

### **3.1.8. SPARQL endpoint**

Providing the ability to run distributed LOD queries on EHRI-RI's database of archival metadata, the SPARQL endpoint will at the outset support a limited subset of primarily structural data attributes that can be aligned with common LOD ontologies such as SKOS and schema.org, as discussed above. If and when ontologies focused on the archival domain

---

<sup>28</sup> <https://www.openapis.org/>

such as Records-in-Context<sup>29</sup> become more mature and widely used the number of supported data attributes may increase.

### **3.1.9. Image APIs**

A IIIF image API server, as discussed above, will provide the means to serve preservation-quality scanned material in web-friendly formats, providing on-the-fly image navigation and manipulation capabilities via an HTTP-based interface.

EHRI-RI's Omeka-based visualisation and Digital Edition platforms will be extended to support the generation of IIIF Presentation API manifests, capable of being consumed by any IIIF-compatible viewer.

### **3.1.10. Geospatial APIs**

Access to data in the planned EHRI-RI geospatial repository will be available via several distinct API:

- a REST API for retrieving metadata about individual datasets
- a WFS API for retrieving dataset content as GeoJSON or other applicable vector-based formats
- a WMS API for retrieving tile-based map raster data

## **3.2. Data Capture**

As an infrastructure dealing with relatively sensitive historical subjects, and which is subject to a range of national and EU-based privacy laws, the pathways through which data enters EHRI's services must be well considered. Hitherto, this has meant that only EHRI staff, via a hierarchical role-based permission system that defines scopes of responsibility mirroring the country/institution/collection structure of the portal — have the ability to create, modify or delete archival metadata.

### **3.2.1. SKOS Vocabularies**

As mentioned above, EHRI currently maintains a number of SKOS-format controlled vocabularies pertaining to Holocaust-related subject terms, ghettos, and camps. EHRI-RI will endeavour to establish a community-driven process for the maintenance and curation of controlled vocabularies to promote their uptake, and thus greater interoperability, of indexed Holocaust-related collections. This process will be driven by existing upstream digital infrastructure, such as DARIAH Vocab Services<sup>30</sup>, along with community-managed resources like Geonames<sup>31</sup> and Wikidata<sup>32</sup>, with changes fed back to EHRI-RI services via LOD harvesting.

### **3.2.2. Service Registry**

EHRI-RI services will be catalogued in an online service registry hosted by the European Open Science Cloud (EOSC) Marketplace.<sup>33</sup> EHRI-RI will retrieve data from the EOSC API to generate an EHRI-specific service registry hosted on an EHRI-RI domain.

### **3.2.3. Annotations and links**

Selected archival metadata on the EHRI Portal, including country reports, institutions, and archival descriptions, will be able to be annotated by registered users. Annotations — or

---

<sup>29</sup> Such as RiC-O: <https://www.ica.org/en/records-in-contexts-ontology>

<sup>30</sup> <https://vocabs.dariah.eu/en/>

<sup>31</sup> <https://www.geonames.org>

<sup>32</sup> <https://www.wikidata.org>

<sup>33</sup> <https://marketplace.eosc-portal.eu/providers/ehri>

notes — will be visible to the user who created them but not publicly visible unless 1) the user opts to make the note public, and 2) a moderator “promotes” it. While programmatic access for retrieval of textual annotations will be enabled by the GraphQL API, the *creation* of textual annotations via this method is not currently planned due to predicted low demand.

EHRI-RI staff will additionally be able to annotate archival material in the EHRI Portal by creating links between related items (such as those which share a common provenance) or by coreferencing the index terms of multiple CHIs to common vocabularies. Like textual annotations, link annotations will be visible on the EHRI Portal site and will be retrievable programmatically using the GraphQL API. Unlike textual annotations, however, creation will be limited to EHRI staff, rather than general users of the portal.

The provision of a public user interface (and/or API) for creating link annotations could, if suitably controlled and moderated, increase the interconnectivity of archival data across institutions by allowing users to suggest suitable access points or connections, and will be considered in the implementation phase.

#### **3.2.4. OAI-PMH harvesting**

With the EHRI Portal being a central part of the EHRI-RI, the ability to harvest and integrate data from third party sources in a sustainable manner will be a major factor in the RI's success as an effective virtual observatory of Holocaust-relevant material. To date, EHRI has used a standards-based approach to harvesting, employing two methods sponsored by the Open Archives Initiative (OAI).

The OAI's Protocol for Metadata Harvesting (OAI-PMH) is a well-established system, dating from 2001, which provides a HTTP-based server specification and XML-based schema that can serve as a container for various metadata formats such as EAD and Dublin Core. EHRI-RI's harvesting tools will support harvesting data in any XML-based format that allows crosswalks to be developed for conversion into EAD.

#### **3.2.5. ResourceSync harvesting**

The OAI's newer harvesting system, ResourceSync, is based on the Sitemaps protocol<sup>34</sup> and provides the means to describe arbitrary file sets on a web server, incorporating an efficient mechanism for delivering partial and incremental changes.

The EHRI Metadata Publishing Tool (MPT), developed in EHRI-2, provides institutions with a way to automatically generate ResourceSync manifests from a set of metadata files (typically in some XML-based format such as Dublin Core or EAD.) The manifests, together with the catalogue metadata, can then be placed on a web server and be harvested by a ResourceSync-capable crawler.

EHRI-RI will support a subset of the capabilities offered by the ResourceSync specification for syncing archival metadata as discrete (non packaged) files linked via sitemaps. Advanced capabilities, such as incremental changesets and support for compressed archive formats, will be considered if demand proves sufficient.

Which protocol — OAI-PMH or ResourceSync — an institution supports, if any, depends on their internal IT systems and level of IT support. Some cataloguing and web-publishing tools, such as AtoM<sup>35</sup> include at least partial support for OAI-PMH that can be enabled relatively straightforwardly by an institution if they use it to publish a public holdings catalogue, but the protocol can be complex to implement in other environments. ResourceSync, by contrast,

---

<sup>34</sup> <https://www.sitemaps.org>

<sup>35</sup> <https://accesstomemory.org>



assumes little about the IT environment but *does* require some degree of IT support to publish material in a web-accessible manner, as well as producing standards-compliant metadata exports.

### **3.2.6. Authentication/Identity attributes**

For authenticating users via external Identity Providers (IdPs), the exchange of certain fixed data attributes will be transferrable to EHRI-RI services that employ distributed authentication mechanisms such as OpenID Connect and/or Shibboleth. For typical authentication purposes a minimum set of attributes would consist of just a unique identifier and an email address, but might also extend to information such as a display name and profile image. The aforementioned AuthN methods also allow Service Providers (SPs) to request more extensive data, including custom attributes, from IdPs with a user's consent, a mechanism which could allow upstream RIs such as DARIAH or EUDAT to integrate more closely with EHRI-RI SPs.<sup>36</sup> For example, if an EHRI-RI SP had cause to know a user's institutional affiliation this information could be requested from an upstream federated IdP when authenticating via a SAML implementation.

## **4. Proof of concepts**

In order to explore the feasibility of the new services discussed above, as well as ways in which existing services can be extended, a number of proof-of-concept systems have been developed or integrated with current EHRI tools.

### **4.1. IIIF Image API server**

Cantaloupe<sup>37</sup> is an open-source image server compatible with the IIIF Image API. Given access to a pool of image files — in EHRI's case using an AWS S3-compatible managed storage system — it provides an HTTP-base interface to a wide range of image navigation and manipulation operations, including pan, zoom, rotation.

For IIIF Presentation API support, the IIIF Toolkit Omeka plugin was installed on EHRI's visualisation platform and configured to use a Cantaloupe instance hosted on EHRI's servers. Used together, they provide users with a more efficient and powerful way to navigate image-based content on the platform, using the Mirador viewer (see Figure 4.) The IIIF Toolkit also generates public manifest URLs, allowing this content to be displayed by compatible viewers hosted elsewhere, making them easily embedded and shared elsewhere on the web.

---

<sup>36</sup> For more examples see the [DARIAH AAI SAML attribute list](#).

<sup>37</sup> <https://cantaloupe-project.github.io>

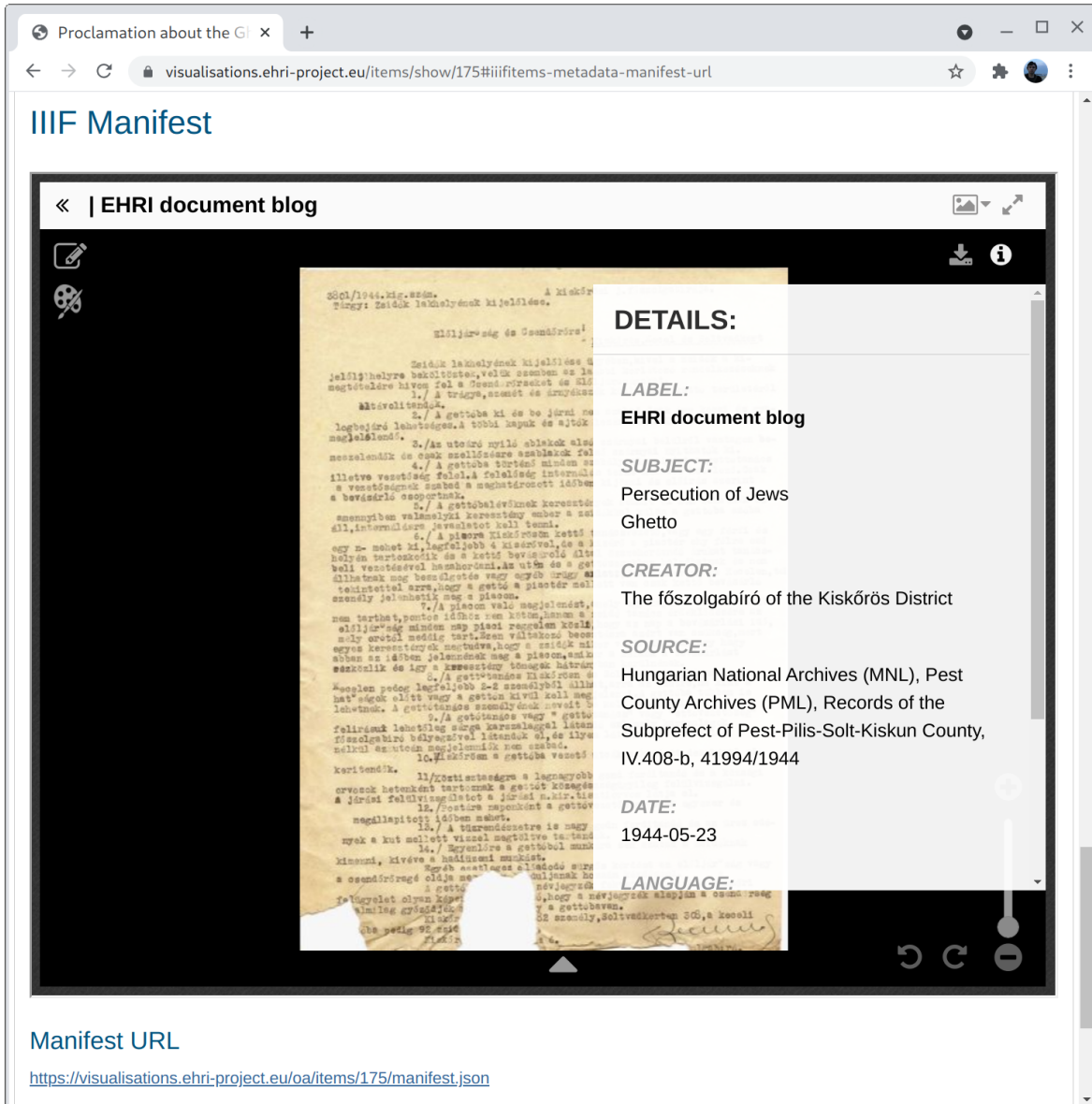


Figure 4: the Mirador content viewer showing integration of the IIF Image and Presentation APIs on the EHRI Visualisations platform.

## 4.2. Portal search client

EHRI has developed a prototype client for its Search API which is intended to allow archival institutions (or “micro-archives”) to publish a browsable catalogue of their holdings as a standalone site, with data pulled from the EHRI Portal’s database. This serves a use-case where an archival institution (or individual with archival holdings) uses the EHRI Portal directly as its backend cataloguing tool, whilst still being able to run a public-facing website of its own - removing much of the complexity from maintaining their own database.

## 4.3. Discourse discussion forum

EHRI has installed and configured a test instance of the Discourse discussion software that allows the creation of public or role-restricted topics (see figure 5). Discourse is a free and open-source forum software that can be used as either a hosted (SaaS) or self-hosted

system. For moderation and administration purposes it includes an extensive role-based access and permission system and offers a wide range of authentication mechanisms.

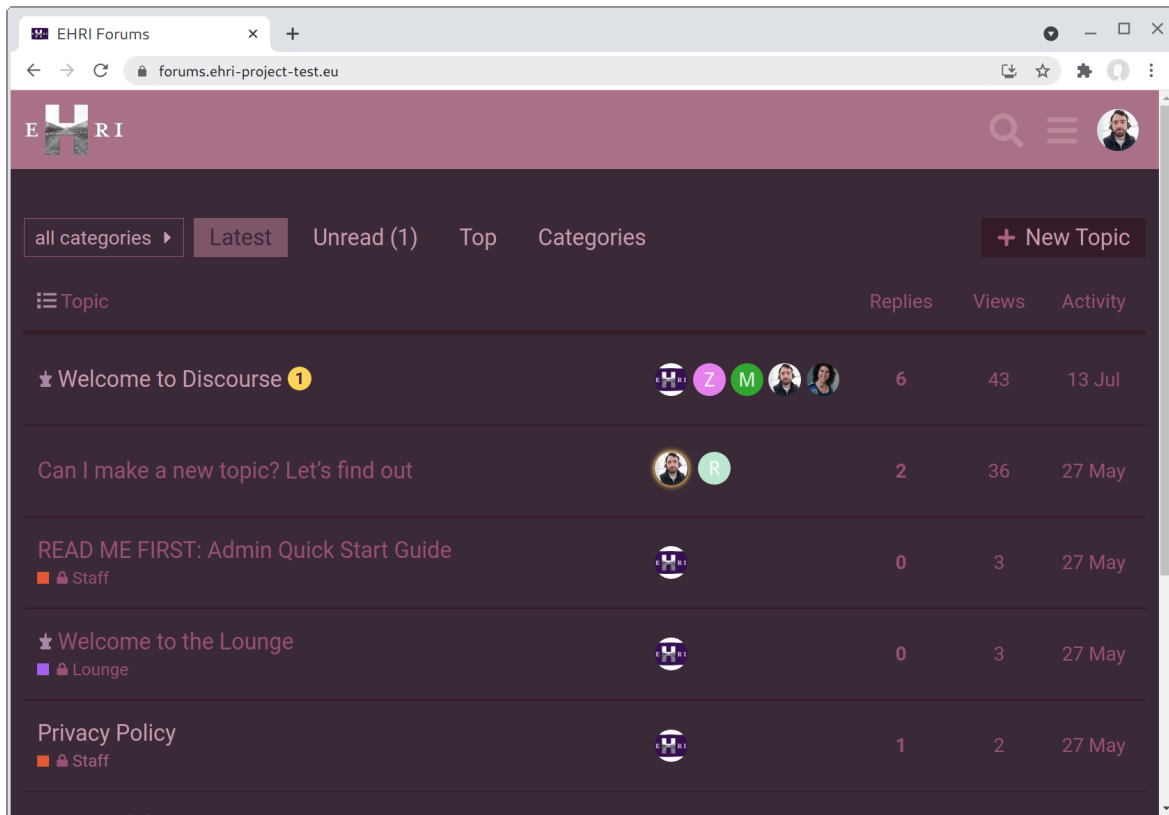


Figure 5: EHRI's prototype Discourse discussion board system

In order to better fit into existing environments Discourse also provides a mechanism, called Discourse Connect, whereby it can be integrated into an existing authentication system on another website. Discourse Connect can be implemented relatively straightforwardly for both the identity provider and the client because the protocol allows only a limited amount of data to be transferred between them, and assumes a shared administration environment where, for example, group identifiers can be coordinated.

In order to test a simple form of SSO using Discourse as the service provider (SP), an implementation of the Discourse Connect protocol has been integrated into the EHRI Portal. This allows Discourse Connect clients — including but not limited to Discourse itself — to use the EHRI Portal as an identity provider (IdP). This means that if EHRI were to create a discussion forum for Holocaust-related topics users could access it via their existing EHRI Portal accounts, and new users would be required to sign up to the portal to access the discussion board.

#### 4.4. Geospatial Repository

Currently being tested for use in EHRI-3, an instance of the GeoNode<sup>38</sup> geospatial content management system (CMS) has been configured for use on EHRI's systems (Figure 6.) The geospatial repository will serve as a database of historical datasets with a spatial dimension

<sup>38</sup> <https://geonode.org>

and with some relation to Holocaust geographies. Datasets will be available to researchers via a web frontend in addition to structured data APIs and will be provided with sufficient metadata to accord with the FAIR (Findable, Accessable, Interoperable, Reusable) guidelines. The repository may also allow the creation of custom maps combining one or more separate layers which may be embedded in other presentational contexts.

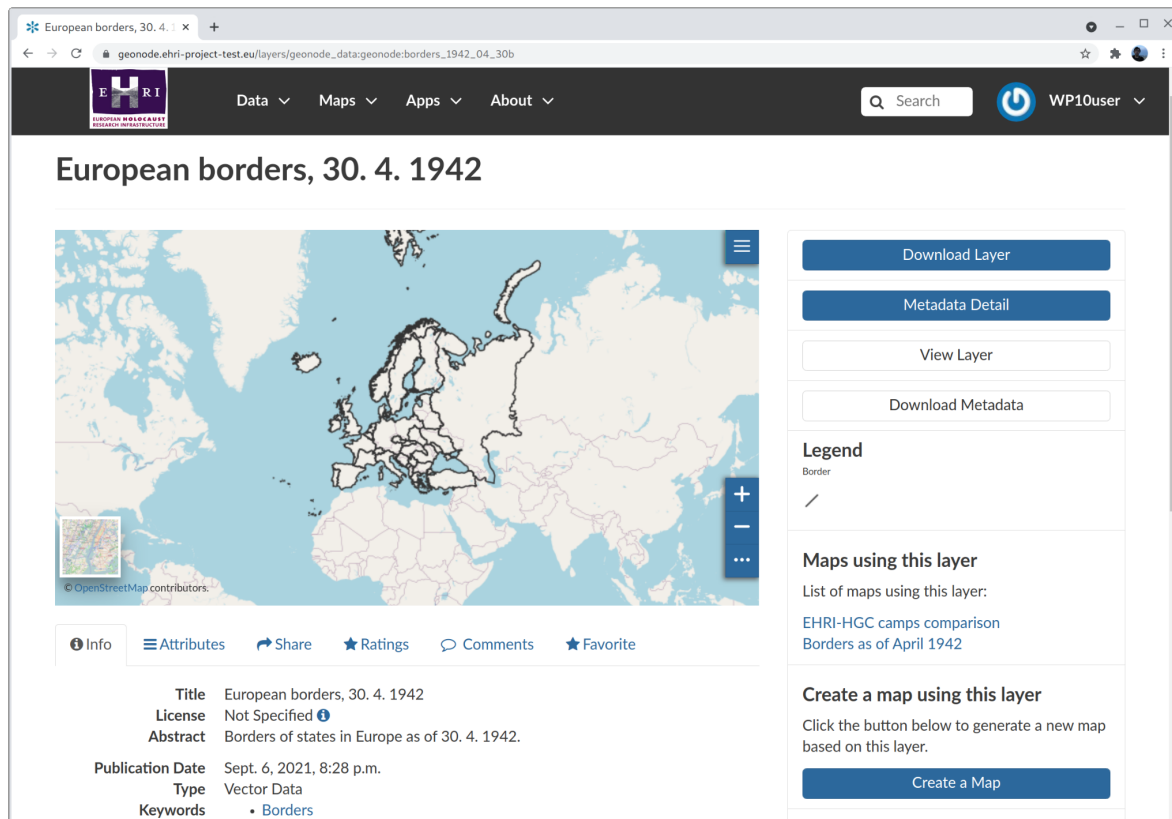


Figure 6: an instance of the Geonode geospatial CMS

## 5. Concluding remarks

This document provides an initial implementation strategy for the EHRI-RI. While the EHRI-RI Central Hub will play a key role in the provision of data services, including current services such as the EHRI Portal, there will be a degree of decentralisation too, with some existing services potentially being adopted and maintained by National Nodes. For the Central Hub and National Nodes to integrate we have presented above a Service Integration Framework (SIF), enumerating the ways in which structured information can be consumed by both EHRI-RI nodes and third-party entities such as upstream and downstream RIs.

While many components of the SIF exist today, others do not. The addition of services that support IIIF APIs for publishing image-based content will improve the functionality and FAIR-compliance of existing content-based tools such as the Digital Editions, as well as new use-cases such as those relating to micro-archival content currently being explored in EHRI-3. New services providing dedicated tools for the management and publication of geospatial datasets, also being explored in EHRI-3, can facilitate the creation of content that explores geotemporal and geospatial themes. Increasing the alignment of EHRI's data with

other LOD datasets and providing LOD querying endpoints will provide new integration points for tools both within the EHRI-RI and outside of it.

One of the more challenging aspects of administering a distributed or partially distributed infrastructure is managing the digital identities of users as they move between related systems. Since the access mode for the large majority of EHRI's online services is free and unrestricted, and thus does not require the enforcement of an identity or affiliation, the problem of identity management is relatively attenuated. Services that may not allow anonymous use however — such as the prototype Discourse discussion forum described in section 4 — require the coordination of user accounts via SSO. The Discourse Connect implementation that uses the EHRI Portal as the identity provider (IdP) is one approach to this that presents a relatively low technical barrier, but we have also noted how the EHRI Portal could itself be integrated as a service provider (SP) to an upstream RI such as DARIAH-AAI using alternate SSO technologies such as SAML.

While the technical development of the EHRI-RI starts from the infrastructure as of EHRI-2, it will ultimately build on concurrent EHRI-PP activities that define the RIs organisational structure, funding model, user access model, and research strategy. The ongoing EHRI-3 project likewise continues to expand the boundaries of EHRI's activities and the RIs eventual scope. The strategy outlined here will be subject to an ongoing process of reconsideration and reevaluation up to and throughout the implementation phase, taking on board lessons learned inside and outside of the project and keeping pace with technological innovation within the RI domain.

## References

Bryant M., Reijnhoudt L., Speck R., Clerice T., Blanke T. (2015) The EHRI Project - Virtual Collections Revisited. In: Aiello L., McFarland D. (eds) Social Informatics. SocInfo 2014. Lecture Notes in Computer Science, vol 8852. Springer, Cham.  
DOI:[https://doi.org/10.1007/978-3-319-15168-7\\_37](https://doi.org/10.1007/978-3-319-15168-7_37)

Jürgen Cito, Philipp Leitner, Thomas Fritz, and Harald C. Gall. 2015. The making of cloud applications: an empirical study on software development for the cloud. In Proceedings of the 2015 10th Joint Meeting on Foundations of Software Engineering (ESEC/FSE 2015). Association for Computing Machinery, New York, NY, USA, 393–403.  
DOI:<https://doi.org/10.1145/2786805.2786826>

Erich, F., Amrit, C. and Daneva, M., 2014. Report: Devops literature review. University of Twente, Tech. Rep.

P. Kruchten, R. L. Nord and I. Ozkaya, "Technical Debt: From Metaphor to Theory and Practice," in IEEE Software, vol. 29, no. 6, pp. 18-21, Nov.-Dec. 2012,  
DOI:<https://doi.org/10.1109/MS.2012.167>.

Philipp Leitner and Jürgen Cito. 2016. Patterns in the Chaos—A Study of Performance Variation and Predictability in Public IaaS Clouds. ACM Trans. Internet Technol. 16, 3, Article 15 (August 2016), 23 pages. DOI:<https://doi.org/10.1145/2885497>

## 6. Appendix 1: Single Sign-On Configurations

Notwithstanding EHRI's existing authentication solutions, there are several different distinct scenarios in which single sign-on technologies could be employed within a *hypothetical* distributed RI:

### 6.1. Third-party Identity Provider

The infrastructure could rely on a third-party authentication solution, such as, in this example, DARIAH's Authentication and Authorization Infrastructure (DARIAH-AAI)<sup>39</sup>, as the sole identity provider (IdP). Attempting to log in an RI tool (a "service provider", or SP) would redirect the user to DARIAH AAI where they could authenticate in one of two ways: using their own DARIAH account (termed "self-service"), or their credentials from a home institution within the DARIAH network (Figure 7.) An example of an infrastructure using this "externalised" authentication model can be seen in the CENDARI project.<sup>40</sup>

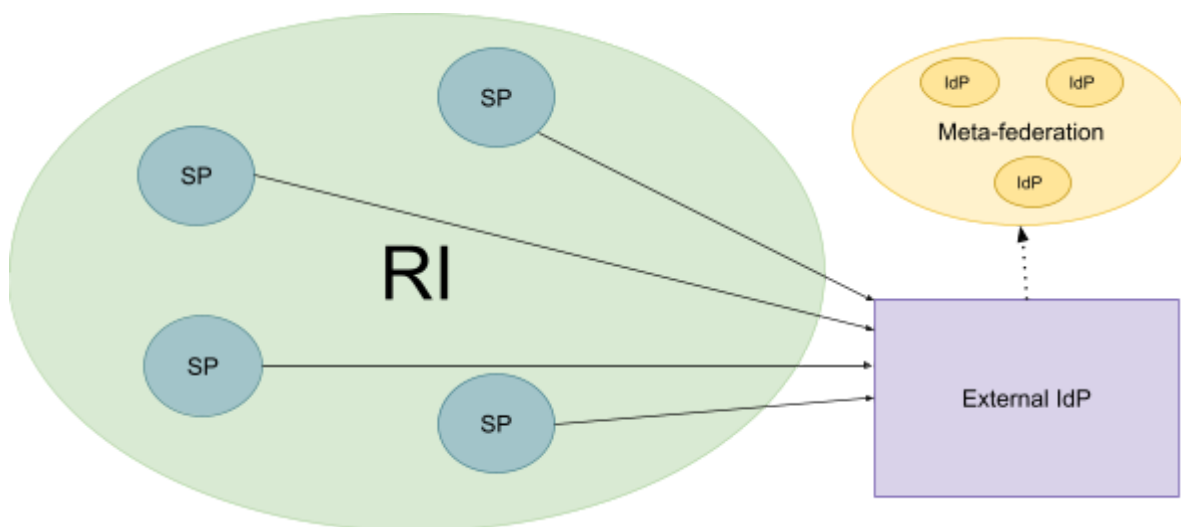


Figure 7: Third-party external identity provider provides common identity services for all RI nodes.

### 6.2. RI Identity Provider

Another scenario is where the infrastructure manages its own IdP centrally. Users attempting to log in to an SP would be redirected to the RI's own IdP system. The RI IdP could use its own login system, or defer to a federated IdP, such as DARIAH AAI or another decentralised authentication protocol, to authenticate the user before passing their attributes (username, display name, group membership, etc) back to the SP (Figure 8).

For the EHRI-RI, the use of an externalised IdP would provide a benefit in that divesting responsibility for IAM to a third-party could reduce administrative overhead and responsibility for maintaining critical security-centred systems. Conversely, there would be costs associated with migrating legacy EHRI accounts, reduced flexibility to adapt and integrate

<sup>39</sup> <https://wiki.de.dariah.eu/display/publicde/DARIAH+AAI+Documentation>

<sup>40</sup> See: the CENDARI Infrastructure: <https://dl.acm.org/doi/abs/10.1145/3092906>

new systems into the RI, and potentially other opportunity costs associated with loss of control over a core part of the user experience for EHRI-RI services.

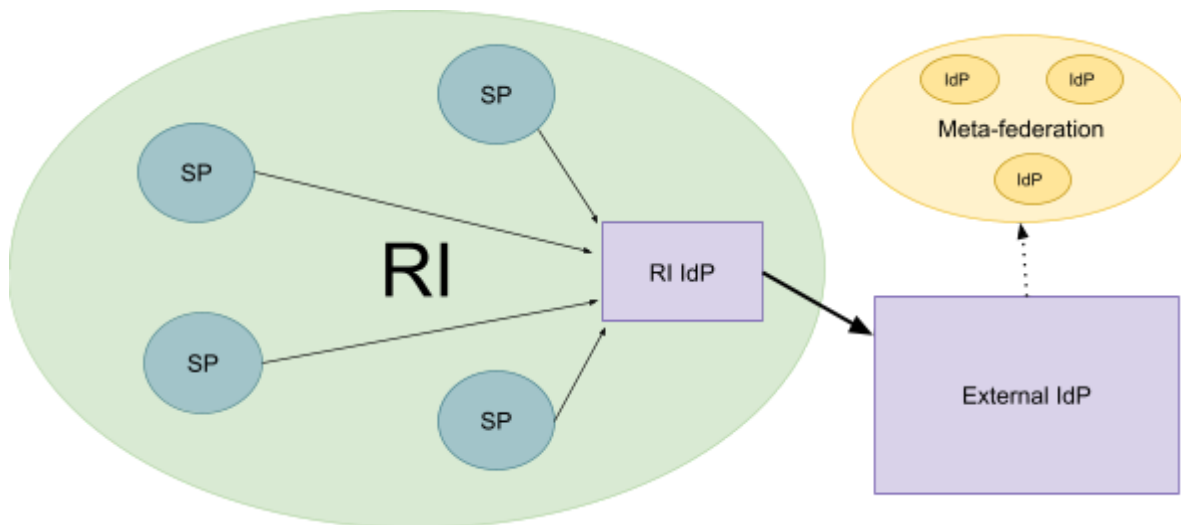


Figure 8: Internal identity provider serves RI nodes but can defer to an external IdP.

### 6.3. SSO Technologies

A number of technologies have been developed to facilitate single sign-on functionality in different scenarios, though only two — SAML and OpenID Connect — are current standards and have significant adoption.

#### 6.3.1. SAML

Widely-used within the enterprise and higher-education spaces where a key use-case involves verifying that a user belongs to a particular home institution (or enterprise), SAML (Security Assertions Markup Language) is an XML-based protocol for exchanging authentication and authorisation information between multiple domains. SAML-based SSO implementations include Shibboleth<sup>41</sup> and SimpleSAMLphp<sup>42</sup>.

#### 6.3.2. OpenID Connect

OpenID Connect (OIDC) is at its core a flavour of OAuth 2.0<sup>43</sup> that has been further standardised for authentication purposes. It specifies the HTTP request/response flow and payload configuration for redirecting a user from a client application (the Relying Party, or RP) to an identity provider (the OpenID Provider, OP) and returning a digitally-signed token that can be used by the RP to fetch scoped user identity attributes. It has many commonalities with SAML but uses the building blocks of OAuth 2.0, such as JSON web tokens, and direct RP to OP (back channel) communication instead of routing all communication via the user's browser.

The EHRI Portal currently implements OpenID Connect as an RP to allow users to sign up or log in using their existing Google, Microsoft or Facebook identities. In a scenario where the

<sup>41</sup> <https://www.shibboleth.net/products/>

<sup>42</sup> <https://simplesamlphp.org/>

<sup>43</sup> <https://oauth.net/2/>



Portal became the identity provider for EHRI-RI services it would be required to extend the existing implementation with OP (IdP) functionality.

### **6.3.3. Discourse Connect**

Discourse Connect is not a generalised SSO system but is mentioned here as *one example* of an intra-application protocol with a considerably narrower scope than the SAML or OIDC, being tailored for specific authentication scenarios with a small, predefined set of data being exchanged between two applications within the same sphere of trust. It is also not standardised by any consortium or body and therefore offers something of a moving target, but from the perspective of a potential EHRI-RI SSO solution has the benefit of much greater simplicity and a smaller “surface area” than more complex, expansive alternatives. Developed to allow third-party applications to integrate with the open-source Discourse discussion board software as either IdP or RP, Discourse Connect defines the mechanism where by a user attempting to log in to the forum is redirected to an IdP and their identity attributes exchanged with Discourse.

## **6.4. Implementation**

The technical feasibility of implementing RI-wide SSO depends on several factors:

- existing support on EHRI platforms for SSO technologies
- feasibility of adding SSO support for platforms where it does not already exist
- whether SSO support is extended to all platforms or just a subset, as needed

Current platforms that will be part of the EHRI-RI Central Hub include:

### **6.4.1. The EHRI Portal**

The EHRI Portal currently employs an OIDC RP implementation that allows users to authenticate via a fixed set of identity providers including Google, Facebook and Microsoft. The portal does not yet have an OIDC IdP implementation. We have discussed above the prototype Discourse Connect IdP implementation, which allows users of the Discourse forum software to log in using their EHRI Portal accounts.

### **6.4.2. Wordpress**

Wordpress has an active ecosystem which includes plugins providing support for SAML-based SSO (including Shibboleth) and generic OIDC. An official Wordpress Discourse plugin provides Discourse Connect functionality but currently does not support IdPs other than the configured Discourse instance. This means that the existing plugin can not be used to implement Wordpress SSO with, for example, the EHRI Portal as the IdP.

### **6.4.3. Omeka Classic**

Currently neither Omeka Classic (the basis of EHRI’s Digital Editions) or the newer Omeka S support SSO via SAML or OIDC, nor are there (non-commercial) plugins available that provide such functionality. Implementing SSO with the EHRI-RI would therefore require the development of this functionality from scratch.

### **6.4.4. Drupal**

The Drupal Content Management System (CMS), used by both the EHRI project website and the current training platform, has a number of modules available that support SAML and OIDC as both IdP and SP.

There is no one-size-fits-all SSO solution for EHRI's existing platforms. Whilst SAML-based approaches lend themselves to integrating with upstream RIs such as DARIAH-AAI, the complexity and breadth of the protocols means that there is limited support in terms of libraries, up-to-date expertise and documentation to aid custom integrations of the type required for intra-RI use (e.g. between EHRI-RI's own services.) OIDC, or for authorisation scenarios OAuth 2.0, by contrast, has significant adoption among 3rd party commercial IdPs such as Google, Github, and Facebook, is more straightforward to implement than SAML, and has more extensive tooling and library support. In cases where custom implementation is required, bespoke intra-application protocols such as Discourse Connect are much more narrowly scoped and therefore present the lowest barriers to realisation, both for IdP and RP.

As a result, a heterogeneous SSO environment is the most likely future approach, where internal EHRI platforms (such as the Digital Editions) that require a from-scratch RP SSO solution leverage a compact and narrowly-scoped system like Discourse Connect, and others use off-the-shelf OIDC RP implementations with the EHRI Portal as an IdP. It is also likely that for some internal systems with a low IAM overhead, particularly those in operation prior to the establishment of the EHRI-ERIC, the cost of implementing, configuring and maintaining an SSO implementation will outweigh the benefits.