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

D7.1

Survey of Technical Requirements

Mike Bryant
KCL / NIOD KNAW

Tobias Blanke
KCL

Michael Levy
USHMM

Widia Mahabier
DANS KNAW

Start: December 2019 [M1]
Due: May 2020 [M6]
Actual: June 2020 [M7]

EHRI is funded by the European Union

Document Information

Project URL	https://www.ehri-project.eu
Document URL	https://www.ehri-project.eu/deliverables-ehri-pp-2019-2022
Deliverable	D7.1 Survey of Technical Requirements
Work Package	WP7
Lead Beneficiary	1 - KNAW
Relevant Milestones	MS1
Dissemination level	Public
Contact Person	Mike Bryant, michael.bryant@kcl.ac.uk, +44 (0)20 7848 4616
Abstract (for dissemination)	This report constitutes a review of the technical requirements for EHRI's metadata integration and dissemination activities, with the objective of informing development priorities for the future EHRI Research Infrastructure (EHRI-RI). In addition to a technical analysis of existing infrastructure we have sought to incorporate the perspectives of EHRI's transnational partners in determining where areas of opportunity, omission and unrealised potential exist.
Management Summary	(required if the deliverable exceeds more than 25 pages) N/A

Table of Contents

Glossary

API: Application Programming Interface

CHI: Collection Holding Institution

EAD: Encoded Archival Description

ECT: EAD Creation Tool

EOL: End-of-life

JVM: Java Virtual Machine

IAC: Infrastructure as Code

ISAD(G): International Standard for Archival Description (General)

LOD: Linked Open Data

MPT: Metadata Publishing Tool

OAI-PMH: Open Archives Initiative Protocol for Metadata Harvesting

OAI-RS: Open Archives Initiative Resource Sync

ODD: One Document Does-it-all

RDBMS: Relational Database Management System

RDF: Resource Description Framework

RNG: RelaxNG

SCM: Source Code Management

SLA: Service Level Agreement

SKOS: Simple Knowledge Organisation System

SWOT: Strength, Weakness, Opportunity, Threat [Analysis]

TEI: Text Encoding Initiative

XML: eXtensible Markup Language

1. Introduction

The EHRI infrastructure consists of range of administrative tools and services that facilitate the integration and dissemination of information about Holocaust-related collections held by EHRI partner institutions. This report constitutes a review of the technical requirements needed to serve these goals, with the objective of informing development priorities for the future EHRI Research Infrastructure (EHRI-RI). In addition to a technical analysis of existing infrastructure we have sought to incorporate the perspectives of EHRI's transnational partners in determining where areas of opportunity, omission and unrealised potential exist. Note that whilst EHRI's primary user-facing tools and services (such as the portal, document blog, and digital edition sites) are discussed in the context of technical requirements, an explicit review of their functionality is considered out of scope for this deliverable.

Section 2 will enumerate the basic technical requirements of the EHRI infrastructure as they exist today, and review the status of the various tools and services that are currently used to fulfill them, providing an assessment of their completeness, maintenance state, and possible areas of improvement. Section 3 reports on the results of a questionnaire circulated among EHRI partners with the intention of better understanding where gaps and omissions exist in the data infrastructure, and where priorities should lie for further technical development. Finally, we will offer some tentative conclusions as to how preparations for the future RI should proceed in order to build most effectively on the work of EHRI to date, with a supplementary SWOT (Strength, Weakness, Opportunity, Threat) matrix in Appendix 1.

2. Overview of data integration infrastructure and basic technical requirements

This section will enumerate the basic technical requirements of the EHRI infrastructure as of EHRI-2, review the tools and services that currently serve those requirements, and highlight areas where additional development or integration work could advance the project's goals.

2.1. Access to computational resources

EHRI's various tools and services run atop a commercial Virtual Private Server (VPS) platform currently procured from Digital Ocean, with additional storage services provided by Amazon Web Services (AWS). Datacenters are EU-based (located in the Netherlands and Germany respectively) as required to fulfill EHRI's obligations concerning data privacy.

As of 2020, EHRI's server infrastructure is administered using predominantly manual methods, i.e. servers are individually configured for specific purposes (e.g. serving websites, databases etc.). While this is the most straightforward and flexible way to manage a set of evolving resources — especially considering EHRI's lack of dedicated System Administration staff — it has disadvantages in terms of documentation, onboarding time for new staff, and disaster-readiness (the ability to quickly recover from data loss or security breaches.)

An alternative and more modern way to manage virtualised ("Cloud"-based) resources is so-called Infrastructure-as-Code (IAC), where configuration is described using a structured specification (the "code") which can then be instantiated on-demand using tools like Terraform. There are a number of advantages to this approach, but above all it allows infrastructure (or a description of it) to be specified in a central location and managed by versioning tools in much the same manner as application code, and in a format that is relatively self-documenting. To learn about the configuration, system administrators do not have to consult secondary documentation or look at the state of the servers themselves, since the specification is the “single source of truth.” Additionally, once a system has been specified, deploying additional identical or similar resources is substantially less effort since templates are readily available.

The disadvantages of an IAC approach include a higher bar to entry for general administration tasks (staff have to learn the intricacies of an IAC configuration tool in addition to system administration itself), more overhead for small changes (which have to be first checked in to the specification and then deployed), and the difficulty of moving from a traditional infrastructure to an IAC one, which, to avoid service outages, would entail a transition period where some resources were IAC-managed and others not.

2.2. Data storage and retrieval

At the core of the infrastructure is a database that stores metadata about archival collections and collection-holding institutions (CHIs), along with related entities such as controlled vocabularies and authority files. Additionally, the database contains information about the administration and provenance of these domain entities. Currently, this role is served by a Neo4j graph database, with an Apache Solr search engine used for free-text retrieval tasks.

2.2.1. Neo4j graph database

EHRI’s collection data is stored in a single instance of Neo4j, exposed via a bespoke REST-style JSON API, as it has been since EHRI-1. Overall, fears that this (at the time) somewhat unorthodox “NoSQL” technology would prove an unstable basis on which to build the infrastructure have not been realised; the database has been updated through several major-version changes while remaining stable and reliable in production. Moreover, over the course of EHRI-2 Neo4j has received many new features that make it a better persistent data store, such as additional schema integrity guarantees and constraints. Nevertheless, as EHRI transitions from a fixed-term research project to a long-term self-sustaining infrastructure it is worth reviewing the technical characteristics of Neo4j in EHRI’s context, and assessing whether it is still a sound choice as a primary database.

Several characteristics set Neo4j apart, for EHRI’s purposes, from more traditional relational database management systems (RDBMS) such as PostgreSQL or MySQL:

1. Its graph database design provided an efficient mechanism (and query syntax) for traversing tree structures with no fixed degree of nesting, such as the hierarchical collection descriptions common in the archival domain.
2. Being primarily “schema-free” it allowed modifications to the data schema — which occur very often during development — to be conducted on-the-fly, rather than requiring potentially complex migration efforts that could disrupt production systems and make integration of new data more difficult.
3. Its use of the Java Virtual Machine (JVM) and a highly extensible architecture meant that much of EHRI’s backend “book-keeping” systems (audit logging, data serialization, permissions and access control) could run efficiently as Java-based database plugins (similar to stored procedures, but with access to a wider software ecosystem and one that was already familiar to EHRI’s technical partners.)
4. It could be accessed through a programming abstraction layer (the Tinkerpop Blueprints stack), meaning that if Neo4j were to become drastically incompatible or otherwise unavailable we could switch to a competing graph database with relatively few changes to higher-level code.

As of 2020, these distinctions are still mostly relevant, but have undoubtedly become less so as alternative technologies have developed to answer similar requirements:

- While most RDBMS do support hierarchical queries in one way or another (either through recursive common table expressions or via nested sets, etc) doing so is still considerably more complex than via Neo4j's Cypher query language (or via imperitative Java code in a database plugin).
- Among RDBMS, PostgreSQL in particularly has gained much better support for schema-free data via its native JSON features (though these still involve various tradeoffs.) Additionally, as EHRI's infrastructure has matured and the pace of change slowed, the cost-benefit calculation to having fewer schema and data integrity features has shifted more to the cost side.
- Writing advanced functionality as database plugins is still supported by Neo4j but has been downplayed by the vendor in favour of simpler custom functions and procedures.
- The Blueprints abstraction layer is one part of EHRI's backend that is obsolete, and no longer provides a viable path to migrate from Neo4j to a comparable offering. Its primary purpose is now to serve as an Object Relational Mapping (ORM) layer and, although stable and well-tested, is unlikely to receive significant future updates or maintenance.

Other factors, such as the commercial status of Neo4j as a product, are also relevant concerns for long-term sustainability. Neo4j, as used by EHRI, is still open source and licensed under the GPL. However since it is developed primarily by a single (relatively small) company — Neo4j Inc. — it is arguably not as stable a basis for future growth as mature databases developed by the open source community, and development is less well roadmapped and predictable.

At this point, almost 10 years into EHRI's lifecycle, migrating to a new database technology would be a costly and time-consuming undertaking, and undoubtedly require the removal of some existing functionality from backend or frontend tools. Given the deliberate stratification of the technology stack however, and in particular the database-agnostic JSON-based API used for the large majority of data access from the frontend, it may be worth investigating the feasibility of a staggered migration of *some* database content from Neo4j into other media. Alongside this we should conduct an analysis of data access patterns to better ascertain how a migration from Neo4j to a pure open source RDBMS (mostly likely, PostgreSQL, since it is already used in EHRI's stack) could work, with a view to mirroring all or most Neo4j content in a different system for enhanced data redundancy.

In theory this could proceed in the following manner:

- develop a PostgreSQL schema able to accurately capture the data held in Neo4j
- develop export/import procedures from Neo4j to PostgreSQL
- run import/export procedures at regular intervals to keep the PostgreSQL mirror up-to-date

While the "live" data would continue to exist primarily in Neo4j, an RDBMS mirror would both provide a potentially useful backup, and the ability to analyse data in ways for which Neo4j may not be ideally suited (for example, time-based analysis via Window functions.)

To summarise: Neo4j continues to serve as a stable and reliable primary database and provides several distinct features that benefit EHRI given the nature of its data and how that data is commonly accessed. Nevertheless, in the interests of redundancy and future-proofing, it would be prudent to investigate additional databases (starting with PostgreSQL) that could serve as backups and auxiliary analysis tools, and as a potential replacement for Neo4j in the unlikely event that it becomes unsupported or unavailable.

2.2.2. Search engine

EHRI search engine still uses Apache Solr, as it has since the portal first went live. While search engine is primarily a frontend tool, used by the portal and the digital editions sites, it is also a component of the Search API. Additionally, whilst most mediation of the search index is performed via the portal's administrative interface, it is also possible for backend tools that update the data store (e.g. via collection ingest) to independently trigger full or partial reindexing.

The current version of Solr used by EHRI (6.2) is end-of-life (EOL), meaning it no longer receives most bug fixes. Since the search engine is not a public facing service (rather, public access is always mediated via a different frontend service) the risk from security bugs in Solr causing vulnerabilities in EHRI's systems is low. That said, it would serve the project well to migrate to a new and actively maintained version at some point prior to the establishment of the RI.

An additional enhancement of the search infrastructure that has the potential to improve the user's search experience is the expansion of automated query testing to use a larger and more representative dataset and a more comprehensive examples, drawn from real-life portal queries. This would provide the ability to better tune search parameters without the risk of large unintended changes in behaviour or other regressions.

2.2.3. Media storage

Since the EHRI portal is focussed on collection metadata rather than digital access to archival material (e.g. multimedia such as scans, audio or video) our file storage requirements are low. Nonetheless, the portal does make use of material such as institution branding, user profile images and various administrative log files, in addition to the more general need for backups that are "offsite" relative to our main Cloud provider (Digital Ocean.) At present this material resides on various AWS S3 buckets, as does the image and transcript data used by the EHRI Digital Editions sites. Were the future EHRI-RI to put more focus on archival *data* (rather than metadata) it would serve the project to incorporate this, and the existing file storage requirements, into an overall data management strategy, alongside that discussed below relating to EAD obtained from third-parties and its derivative data.

2.3. Archival data entry

While the majority (>85%) of catalogued collection descriptions in the EHRI portal have been ingested from metadata provided by EHRI partner CHIs, the remainder, created through manual data input by EHRI staff or external domain experts, represent some of the most detailed and comprehensive English-language material available overall. Moreover, the project's more recent focus on smaller and less well-known institutions means that manual data input will continue to be an important way of signposting the existence of relevant material in the future EHRI-RI. Other tasks where manual data input is the norm include the maintenance and upkeep of Holocaust country reports, the details of over 2,000 CHIs, and a large number of authority files and controlled vocabularies. Throughout EHRI-1 and EHRI-2, CHI records and collection descriptions have been entered directly into the database via the portal's administrative interface, using traditional web forms. At the end of EHRI-2 a more interactive interface was made available for the editing of controlled vocabularies based on SKOS structure.

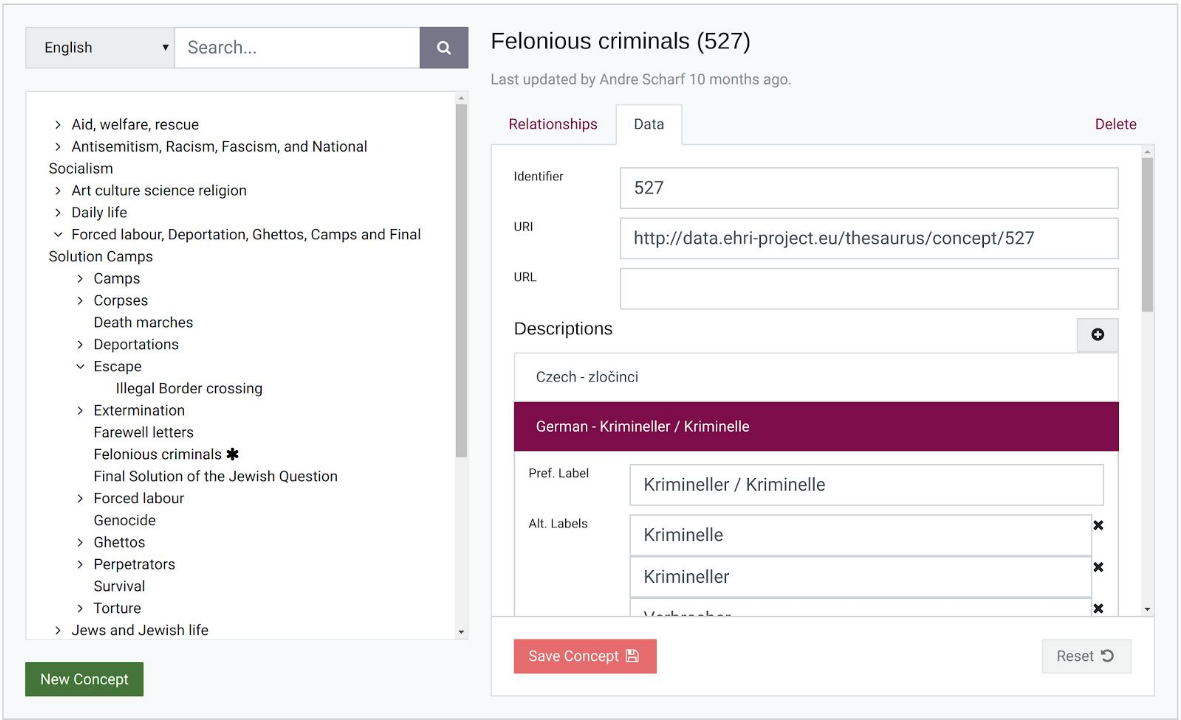
2.3.1. Management of archival entities

The portal backend incorporates tools for creating and editing hierarchical ISAD(G)-format collection descriptions, along with counterparts for ISDIAH (CHI) and ISAAR (authority file) data. These interfaces follow very traditional web-form-based design, and could potentially benefit from being updated to function in a more interactive manner, with better use of real-time validation (i.e. as-you-type hints and error messages.) While not an urgent priority, a renewed focus on the ergonomics of manual data input would be a good investment.

If attempted, a rethinking of the ISAD(G)-based data input interface would have considerable overlap with a tool that could facilitate the creation of EAD format descriptions, and could be useful elsewhere, not only within the consortium but outside it.

2.3.2. Managing controlled vocabularies

The portal backend incorporates a tool for editing thesauruses/vocabularies (figure 1) that have a graph-like structure (e.g. any non-root term can both belong to multiple categories and have multiple sub-categories.) While use of this tool is currently restricted to EHRI staff with administrative permission to edit vocabularies it could also potentially be useful to those outside the project if generalised into a generic editor for SKOS-like data.



Felonious criminals (527)

Last updated by Andre Scharf 10 months ago.

Identifier: 527

URI: http://data.ehri-project.eu/thesaurus/concept/527

URL:

Pref. Label: Krimineller / Kriminelle

Alt. Labels: Kriminelle

Krimineller

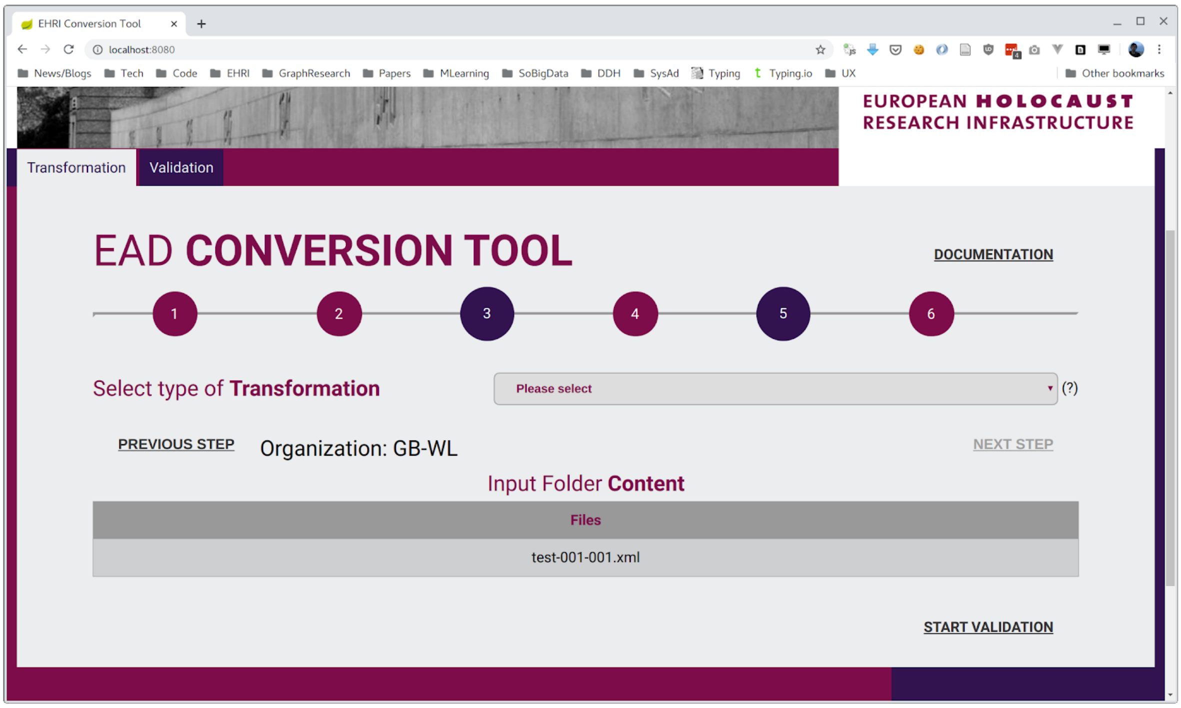
Figure 1: Thesaurus editing interface

2.4. Importing structured archival descriptions

As mentioned above, EHRI’s collection metadata is based on the ISAD(G) conceptual standard, for which the most common serialisation (transport) format is XML conforming to the Encoded Archival Description (EAD). Since EAD is too flexible a schema for many practical data integration purposes, EHRI uses a stricter subset as an ingest-ready target format, meaning that even schema-compliant EAD from third-parties will typically require some level of translation before ingest is possible. Since CHIs with relevant holdings frequently do not have the technical resources to adapt whatever their in-house cataloguing tools produce to match EHRI’s ingest format, a generic XML conversion and validation tool was needed. The EAD Creation Tool (ECT) was developed in EHRI-2 to meet these requirements.

2.4.1. EAD Creation Tool (aka “10.3 Ingest Tool”)

The ECT (on Github as the 10.3 Ingest Tool, figure 2)[1] is designed to take a set of XML files as input, transform them to EAD via a schema-specific mapping configuration, and validate the result according to EHRI’s standards. For these purposes it is a standalone desktop tool which utilises configuration stored in Google Sheets documents, a feature that facilitates remote collaborative development of mapping configuration files. While the ECT has an interactive web-based frontend (predominantly written in Javascript), the lack of authentication functionality and reliance on the user’s local computer for input and output data mean that in its current form it is not suitable for exposing as a public service hosted by EHRI. Instead it functions as a browser-powered desktop application.



Select type of Transformation: Please select (?)

Organization: GB-WL

Files
test-001-001.xml

Figure 2: the EAD Conversion Tool (ECT)

Schema-specific configuration files for mapping from institution-specific XML formats (possibly EAD, or something completely different) are written in tabular form, utilising the XPath language. While the ECT does not *necessarily* have to map from exactly one input file to one outfile file (multiple inputs can result in one output) this is the most straightforward usage, and the least complex to implement via XPath configuration files. In cases where multiple input files map to a single hierarchical EAD output other methods of conversion, including one-off scripts, might be preferred.

The ECT’s existing XML conversion and EAD validation functionality has the potential to be opened up to wider usage by integrating it into the portal’s administration interface, thus not requiring the user to download and run a Java web server on their own computer (something

[1] The ECT also contains functionality to ingest material into the EHRI portal but this is intended to be used by EHRI rather than data providers.

many users working at institutions with restrictive IT policies will be unable to do.) Since such functionality would need to deal with input files uploaded by current or potential data providers it would need to be designed alongside and consider aspects of the overall Data Management Plan (Deliverable 7.3).

2.5. Harvesting of externally-hosted metadata

Some EHRI partner institutions make their collection metadata available in a manner that is conducive to regular automatic retrieval based on established standards, and for EHRI this represents the best way to keep our knowledge of an institution’s holdings as up-to-date as possible. While workflows based on standards such as ResourceSync and OAI-PMH represent the best case scenario for persistent data integration, the opportunities for utilising it have, over the course of EHRI-1 and EHRI-2, been few and the technical hurdles high, not just for EHRI but especially for partners that are able to offer this. Over the course of EHRI-2 significant barriers were overcome by the creation of the Metadata Publishing Tool (MPT), allowing third-parties to more easily publish EAD for harvesting via the ResourceSync protocol, and, on EHRI’s side, the development and/or integration of tools that could ingest harvested material into the portal.

2.5.1. Data provider workflow

Figure 3 illustrates the *data provider’s* harvesting workflow for ResourceSync publishing, utilising EHRI’s EAD Creation Tool (ECT) and Metadata Publishing Tool (MPT). Data is first exported from the provider’s internal cataloguing system in a format that might be a flavour of EAD, but could also be XML of some arbitrary schema. The provider then uses the ECT to convert their data to EHRI-subset EAD, using a bespoke mapping configuration (typically developed by EHRI specialists based on an agreed-upon semantic mapping.) The ECT then validates the output EAD and the provider corrects any errors that may have been detected. Finally, the provider uses the MPT to create a ResourceSync manifest that can be harvested by EHRI.

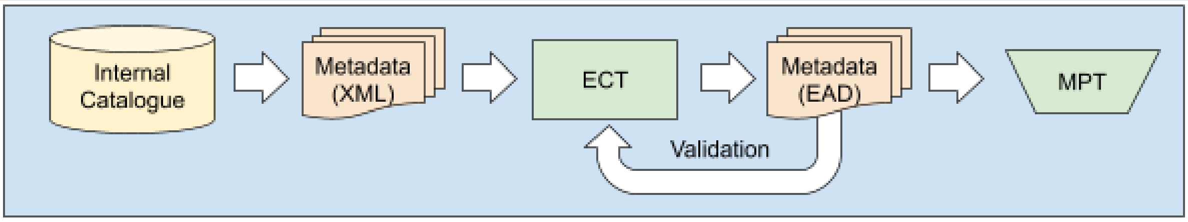


Figure 3: the ResourceSync harvesting process

2.5.2. Ingest workflow

Once a data provider has made valid EAD available in a harvestable format, the EHRI ingest workflow can commence (see figure 4.) For ResourceSync harvesting this begins with the EHRI RS Aggregator service synchronising to EHRI’s server the set of resources published by the provider. The ECT then (again) runs these resources through EHRI’s EAD validator to ensure suitability, and where necessary applies additional enrichments to the data required for ingestion, such as adding missing unit identifier tags. Finally, the ingest tool uses the data store’s web service API to ingest the final EAD into Neo4j.

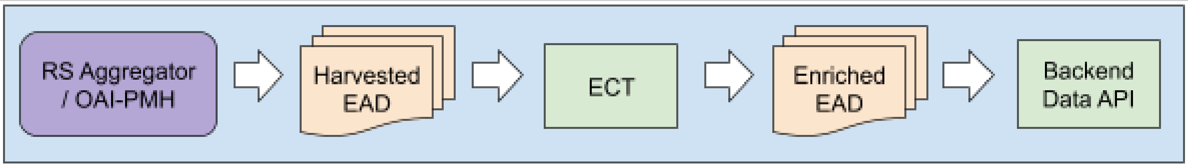


Figure 4: simplified post-harvesting ingest workflow

2.5.3. Metadata Publishing Tool

The Metadata Publishing Tool (MPT, figure 5) is a standalone desktop application that generates ResourceSync manifests from a set of input files, allowing institutions to package EAD in a form suitable for harvesting by EHRI on a repeatable basis. The MPT is cross-platform, with installers for both Windows and MacOS available. Since tools built using the PyQT framework are also usually compatible with Linux, creating an installer for this platform could be a worthwhile improvement for the future.

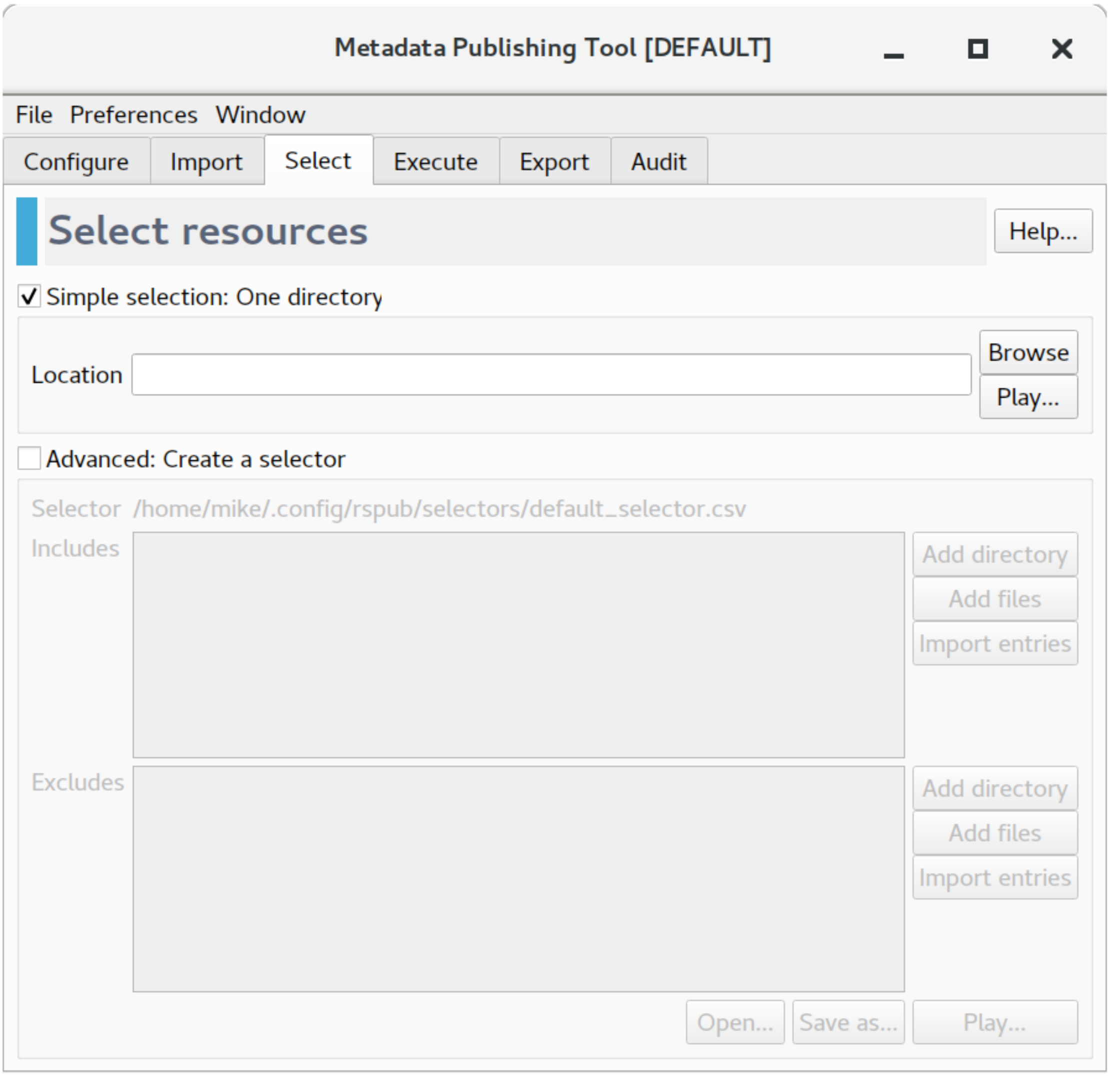


Figure 5: the Metadata Publishing Tool (MPT)

2.5.4. OAI-PMH

Two institutions were set up to expose data via OAI-PMH:

- CDEC (OAI link)
- NIOD (OAI link blocked to normal traffic)

Data is fetched from these endpoints using the Shell OAI Harvester, a command-line tool configured via files located on the server and run on-demand.

A special case concerning OAI-PMH was selective harvesting from endpoints such as the UK's Archives Hub aggregator, which hosted metadata sourced from several different CHIs. To handle these cases a tool was developed to harvest only documents belonging to specific institutions based on a list of preselected item identifiers. Since this configuration is only accessible given access to the server it cannot be changed by data specialists or institution-affiliated staff without the help of an IT administrator.

2.5.5. OAI ResourceSync

Six institutions were set up to expose data via OAI ResourceSync:

- AJA (via USHMM)
- Cegesoma
- Fortunof (via USHMM)
- Kazerne Dossin
- USHMM
- Yad Vashem

Data was fetched via an EHRI tool called the RS Aggregator, a persistent service that repeatedly checks and if necessary updates the set of resources available at all of the configured endpoints. As with OAI-PMH, endpoint configuration requires administrative access to the server.

While both the OAI-PMH and OAI-RS harvesters work effectively, there are a number of ways that they could be made easier to manage and monitor. In particular, providing a web interface through which OAI-PMH and OAI-RS endpoints could be configured, tested (to ensure validity) and monitored (e.g. for frequency of updates to harvested data) would considerably reduce the difficulty involved in performing harvesting operations. For OAI-PMH harvesting from aggregator sites, web-accessible configuration for selective harvesting would additionally be beneficial.

2.6. Ingest data management

At present, input data in the form of raw exports from data providers or EAD that has resulted from a conversion process resides on individual EHRI production or staging servers, from whence ingest typically takes place.[2] Since a typical non-harvested import from just one data provider might involve several back-and-forth revisions, and several stages at which manual or programmatic modifications to files might take place, tracking the provenance of a given import-ready EAD file represents a considerable challenge, and to compound this problem there is not a one-to-one relationship between hierarchical EAD input files and the archival entities that result on the portal. While the provenance of an item in the EHRI portal is usually

[2] The EHRI portal's administration interface does include an interface that allows direct import from files uploaded by the user, but for EHRI-2 this was only employed in a minority of cases.

discoverable from its audit log, this is not a given, and overall the process of tracking material from data provider to portal could benefit from considerable streamlining.

Hitherto, the project has not used a centralised file repository software for the purpose of tracking raw, intermediate, or import-ready data files, instead employing file sharing services such as B2Drop and Google Drive, with convention-based folder structures and naming. Notwithstanding the advantages of a simple convention-based model, for the future RI a more formal approach should be adopted, based on a central repository and an interface that enforces suitable structure and change auditing. The EHRI Data Management Plan (D7.3) will explore these issues further.

2.7. Review and verification

Given the inevitable semantic and linguistic complexities of transnational data integration and the wide range of cataloguing styles found among EHRI partner institutions, it is typically the case that the integration process is an iterative one incorporating feedback from the data providers themselves. In practical terms, this requires seeing their collection descriptions in the context of the EHRI portal prior to their being available in the *portal proper*. For this purpose EHRI has hitherto used a completely separate instance of the portal and its administrative tools, hosted at a different domain, known as the *staging environment*, in which material in the process of integration can be tested and feedback incorporated prior to final release.

2.7.1. Staging Environment

Given that the staging environment is essentially a copy of the production portal database, it was necessary to synchronise it with production data at semi-regular intervals so as to prevent too much drift occurring. This synchronisation was done manually rather than automatically, since at any given time there might be material on staging that needed to be critically evaluated by EHRI staff or staff from EHRI partner institutions and a database refresh would cause this to be overwritten.

This integration workflow had significant shortcomings from the perspective of sustainability and contributed an unwelcome degree of complexity into the overall data pipeline since operations would have to be duplicated between two servers. Since data integration was a concurrent, distributed process it was difficult to know when a refresh of staging was possible, since potentially many reviews of in-progress data still needed to be signed off by external visitors. This led to the staging environment becoming undesirably out-of-sync with the production environment.

Solving the problem of managing in-review and/or work-in-progress material should be a priority for the development of the RI, and two approaches are possible:

1. improve the synchronisation process between staging and production databases so it becomes an automatic process
2. perform review and testing of material directly in the production environment

While both options are complex, option 2) seems the most tractable at this stage since there already exist mechanisms to restrict visibility of material under certain conditions. It would, however, still necessitate a significant engineering effort to cleanly integrate a review pipeline into the portal administrative interface with one of the main issues being maintaining the usefulness and clarity of the audit log in situations where trial-and-error data imports were taking place.

2.8. Data publishing and export

EHRI is not just an aggregator of information but a creator of it, and in addition to that produced by the consortium itself there is also the possibility for users of the portal to contribute both private and publicly visible data in the form of annotations. It is necessary therefore for the EHRI infrastructure to incorporate the means to export metadata from the database in standardised ways, and to extract information in structured form via Application Programming Interfaces (APIs.) As of the conclusion of EHRI-2 these mechanisms exist at several levels. Collection, CHI, and authority files can be exported in XML as EAD, EAG, and EAC respectively, whilst controlled vocabularies such as the EHRI thesaurus are exportable for administrative users as SKOS RDF. Two distinct APIs exist for search and structured retrieval respectively (using the JSON format) and an OAI-PMH compatible endpoint exists for harvesting collection data in EAD or Dublin Core XML formats. Finally, there exists the facility to publish arbitrary datasets, dynamically generated from the database, in spreadsheet-compatible tabular formats.

With the exception of EHRI's list of ghettos that has been published on Wikidata (see Cooey, 2018), an area of weakness in EHRI's data publishing efforts concerns linked open data compatible with the semantic web. While the project does use the https://data.ehri-project.eu subdomain for raw RDF (Resource Description Format) files and authority lists, this is something of a relic of EHRI-1 and the files themselves are not kept in-sync with the data on the portal. Data such as controlled vocabularies that are compatible with the SKOS (Simple Knowledge Organisation System) format that is available on the portal is not yet exportable as RDF by non-administrative users, yet alone queryable via SPARQL or other standardised means. Additionally, while recommendations were developed in EHRI-2 for the use of Uniform Resource Identifiers (URIs) across the project's various archival entities these have mostly not yet been adopted, and those URIs in use (e.g. in SKOS data) are not resolvable, meaning a URI does not map to a web address (URL.)

These shortcomings hinder the ability for EHRI's partners (and others in the archival or Holocaust-research communities) to make use of EHRI's controlled vocabularies and authority sets. In the medium term some steps should be taken to improve the situation:

- deprecate existing "data" subdomain and access to raw RDF dumps, which are out-of-date
- expose existing SKOS-compatible data as RDF via the portal, using a URI schema as recommended in EHRI-2 WP11 report on standards
- make URIs resolvable via appropriate HTTP redirects, e.g. when an attempt is made to visit a URI, re-map it to the relevant URL for a vocabulary or concept item

For querying EHRI data via SPARQL the adoption of a suitable triplestore (such as Ontotext's GraphDB or an open-source alternative) would most likely be required, along with a periodic synchronisation with the reference data in Neo4j (similar to that proposed above for an SQL mirror.)

2.9. Documentation

EHRI maintains a dedicated documentation website (https://documentation.ehri-project.eu/) based on the Sphinx platform and hosted by Read the Docs. Its primary purposes, to date, have been for technical documentation aimed at developers of EHRI software, and documentation for portal data entry and other administrative functions.

Documentation for current and potential EHRI partners and data providers, by contrast, is currently quite fragmented and limited in scope. While tools such as the ECT and MPT do have their own user manuals, higher-level guides are missing. For example, it is difficult for those interested in sharing data with EHRI to learn information such as the type of data EHRI accepts (other than that it is Holocaust-related); the formats and specification that we use; whether technical assistence is available; what tools can be used to facilitate or assist data conversions; and what types of services EHRI can harvest. At present, most of this information is available in EHRI-2 project deliverables which — while technically public — are difficult to find and not written for external audiences.

An improved set of documentation for data providers should aim to cover at least the following areas:

What EHRI does with collection metadata: How metadata is presented on the portal and via its APIs, and how EHRI facilitates the wider availability of metadata in multiple formats to researchers and other stakeholders.

How EHRI accepts collection metadata: This should discuss the target format for ingest into the EHRI database, namely our own subset of EHRI 2002. We should not only try to outline the “profile” of an EHRI-format EAD, which elaborates and clarifies where necessary the definition of the various fields (where inconsistencies have previously been observed) but highlight where EHRI’s superset is stricter in terms of validation than the public schema. We should endeavour to explain the various modes of validation available: for example, the RelaxNG schema.

Data Security: We should document the fact that the EAD provided to EHRI is not made publically available; rather that EAD available via the portal is re-exported from the database. We should briefly outline where data provided to EHRI is stored (e.g. on our servers or via 3rd party Cloud-based systems.)

One-off vs. repeatable ingests: We should document the limitations with regard to currency of metadata provided to EHRI, in the absence of a repeatable connection via one of the currently-supported methods, OAI-PMH or ResourceSync. We should additionally document the requirements for usable OAI-PMH or ResourceSync endpoints, such as supported metadata formats and HTTP accessibility.

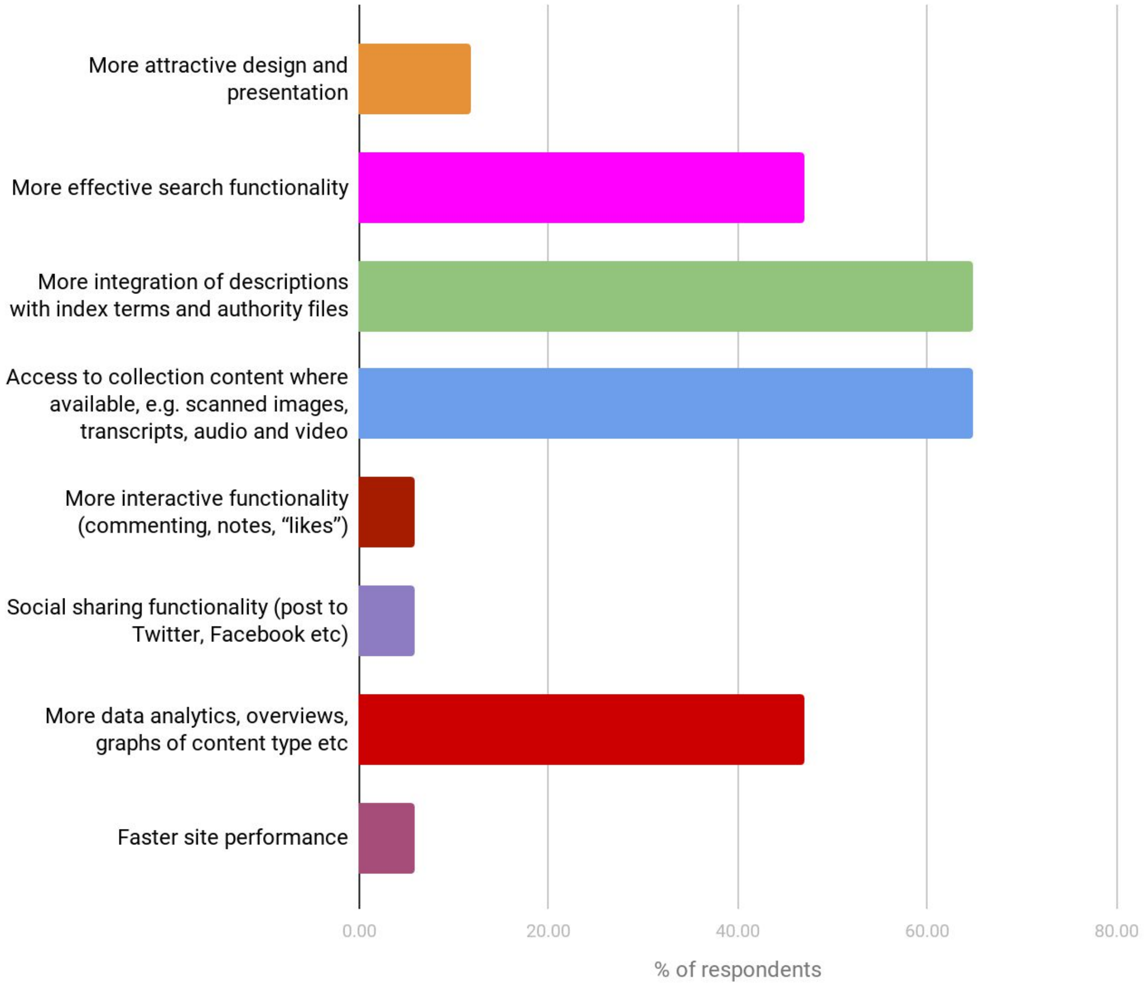
3. Survey

As part of this review of technical requirements we circulated a questionnaire among EHRI-PP and EHRI-3 partners that was intended to solicit feedback from a wide range of perspectives in a less formal manner than previous surveys conducted by the project in previous phases. Given the relatively small pool of possible respondents and their transnational, multidisciplinary nature, questions were intended to be broad and begin from a basic assumption of familiarity with EHRI’s outputs (e.g. its various websites), but include more detailed technical questions for those with the appropriate backgrounds and knowledge. Finally we included an optional open-ended question intended for additional ideas and feedback.

Responses were encouraged from individuals rather than institutions, so in some cases there were multiple responses per EHRI partner. In total there were 17 responses, with 15 respondents volunteering their workplace as an archival institution.

Q1. What in your view would most enhance the experience of using the EHRI portal (choose up to three answers)?

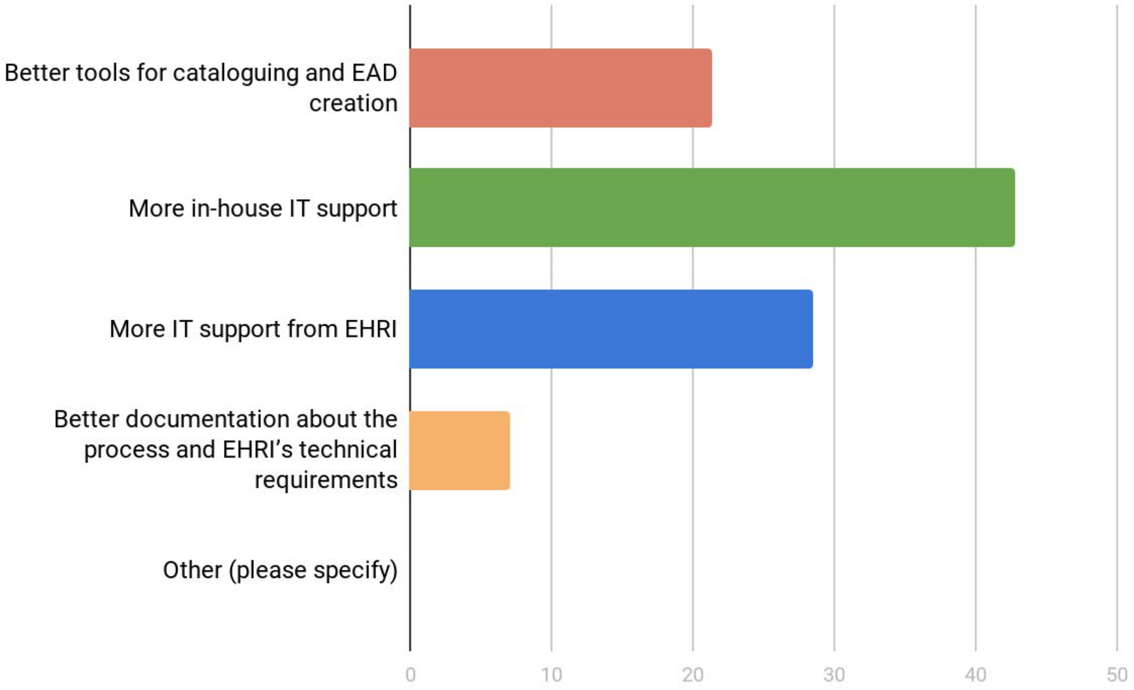
This question was set up so that most individual EHRI contributors and associates could easily lend their opinion, based on their own experience with the EHRI portal, one of its key technical outputs.



The topmost voted answers revealed a preference for more multimedia material such as scanned content, and better integration of index terms. Slightly less frequently requested, but still popular, were improvements to search functionality and more tools for aggregate data analysis. The responses did not reveal much demand for more "interactive" web functionality such as social sharing buttons.

Q2. In your view, what would most improve the process of sharing metadata about your institution's collection holdings with EHRI?

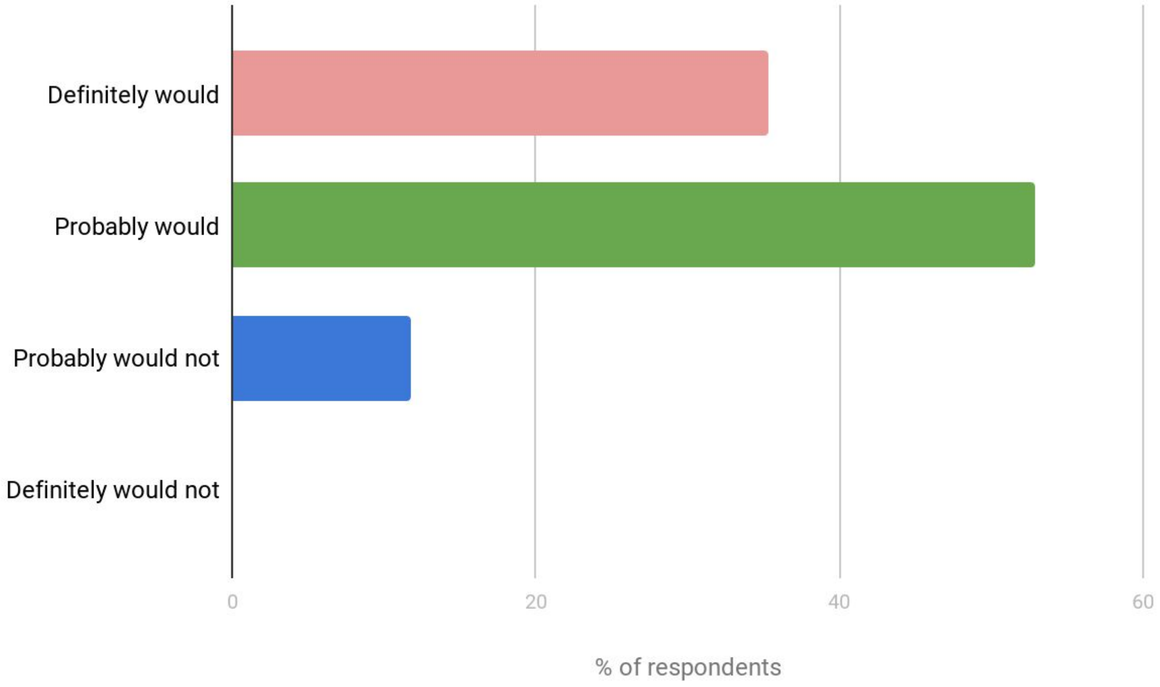
This question was intended to better understand where the primary friction points existed in sharing metadata with EHRI.



In-house IT support is, according to the responses, the biggest single sticking-point for data sharing, a factor not directly within EHRI's control. *However*, the second- and third-ranked points — better tools for cataloguing and EAD creation and more IT support from EHRI — are factors where the project can make a difference, and account for a plurality of responses.

Q3. In EHRI-2 we developed a tool for converting from generic XML-format metadata to the Encoded Archival Description (EAD) format. Would this be useful to your institution if made available as a public service?

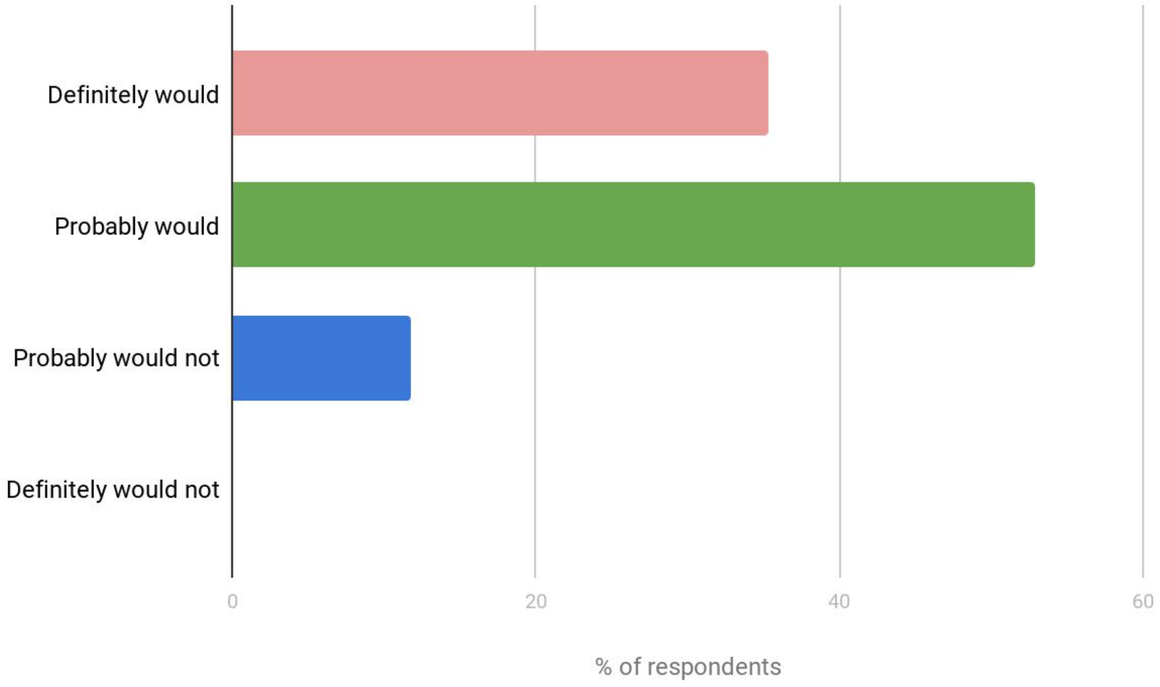
This question was intended to gauge interest in the further development of ECT (EAD Creation Tool) conversion functionality.



Together, 88% of respondents answered that this EAD conversion functionality would probably or definitely be useful to them or their institution.

Q4. In EHRI-2 we developed a tool for validating EAD files according to EHRI's standards. Would this be useful to your institution if made available as a public service?

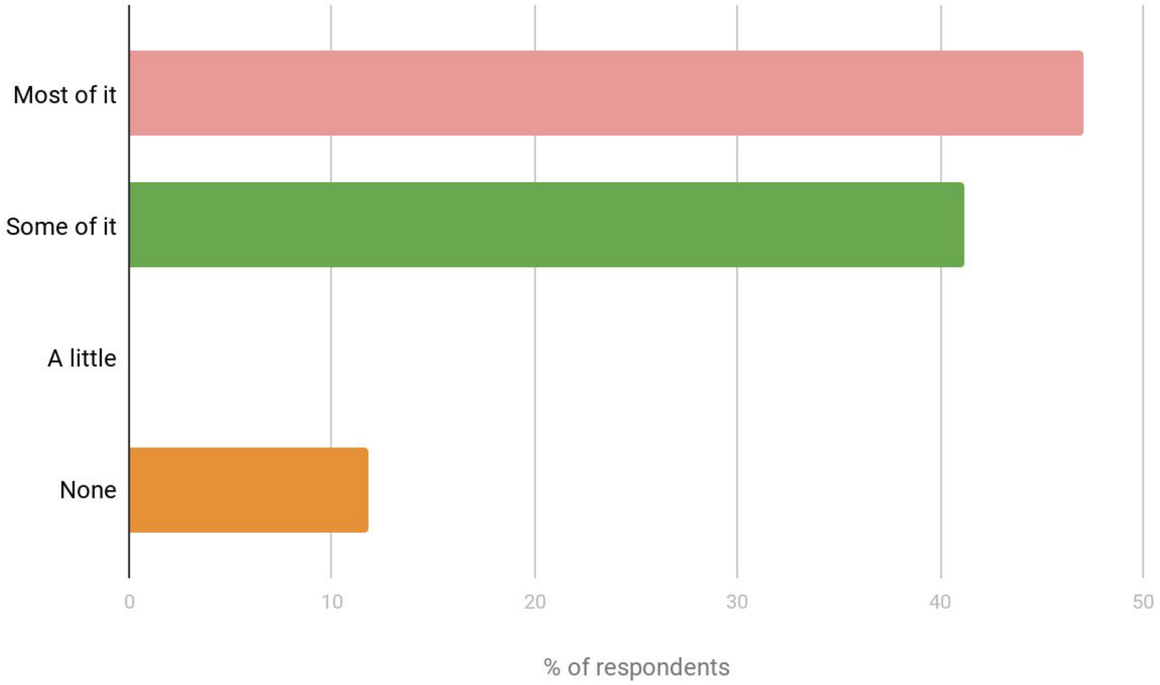
This question was intended to gauge interest in further development of the ECT's EAD validation functionality.



Once again, 88% of respondents answered that this EAD validation functionality would probably or definitely be useful to them or their institution.

Q5. Does your institution host collection- or item-level metadata about its holdings online (excluding the EHRI portal)?

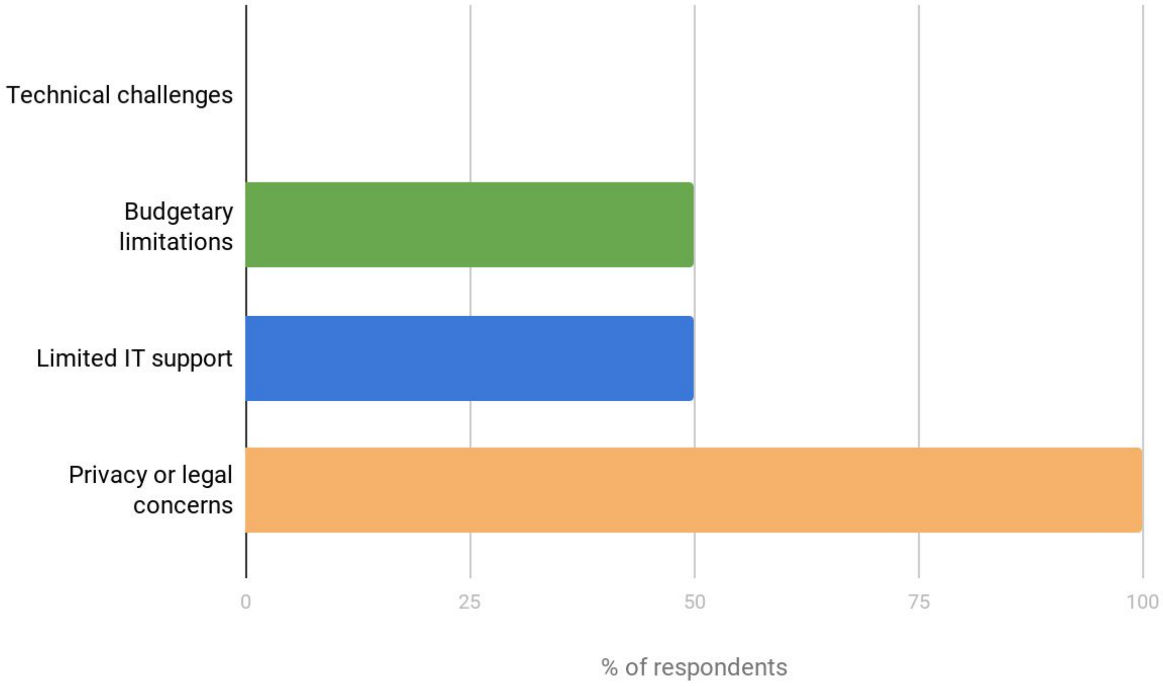
This question was intended to better find out about the existing degree of data sharing among EHRI's partner institutions.



In total 88% of respondents answered that their institution hosted some or most of their collection data online. The remainder answered that none was available. This figure is considerably greater than the number of Holocaust-related archives in general that make comprehensive metadata available digitally.

Q6. If your institution does not make at least some collection-level information available online, what are the main reasons for this (check all that apply)?

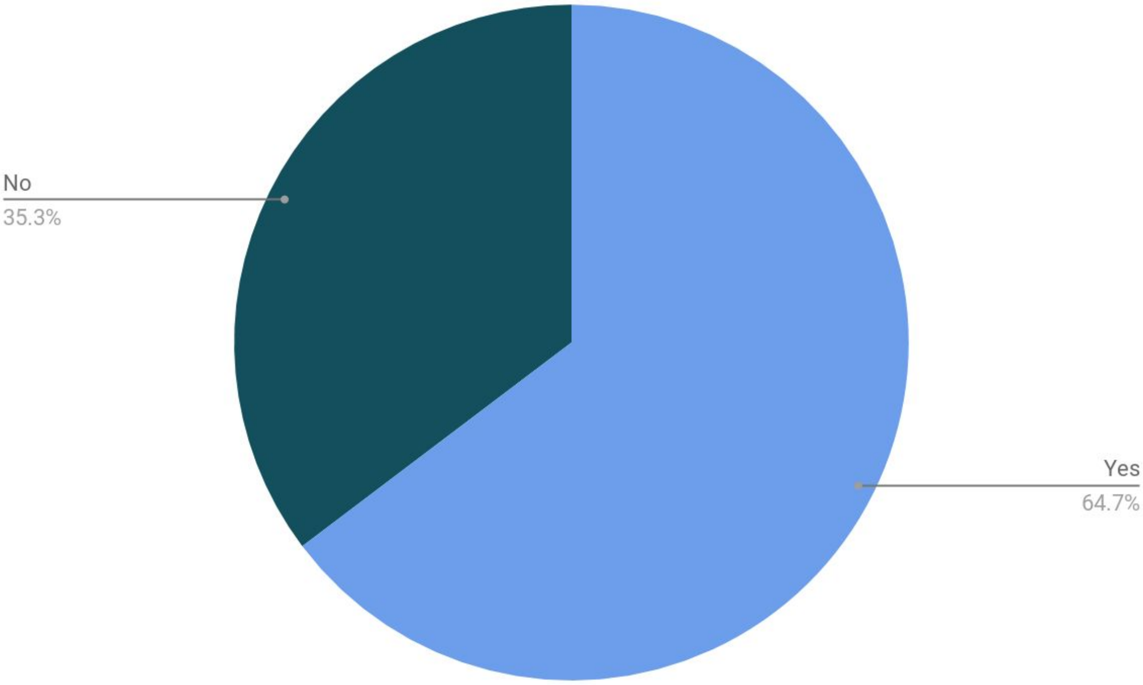
This question was intended to better understand the reason why institutions do not make their collection metadata available publicly via the web.



Responses to this question appear to reaffirm that the primary limitations to metadata publication involve staffing or legal issues, as opposed to technical challenges. Note however that due to the results of Q5 above there were few responses to this question.

Q7. Does your institution make use of controlled vocabularies in the cataloguing of collection material?

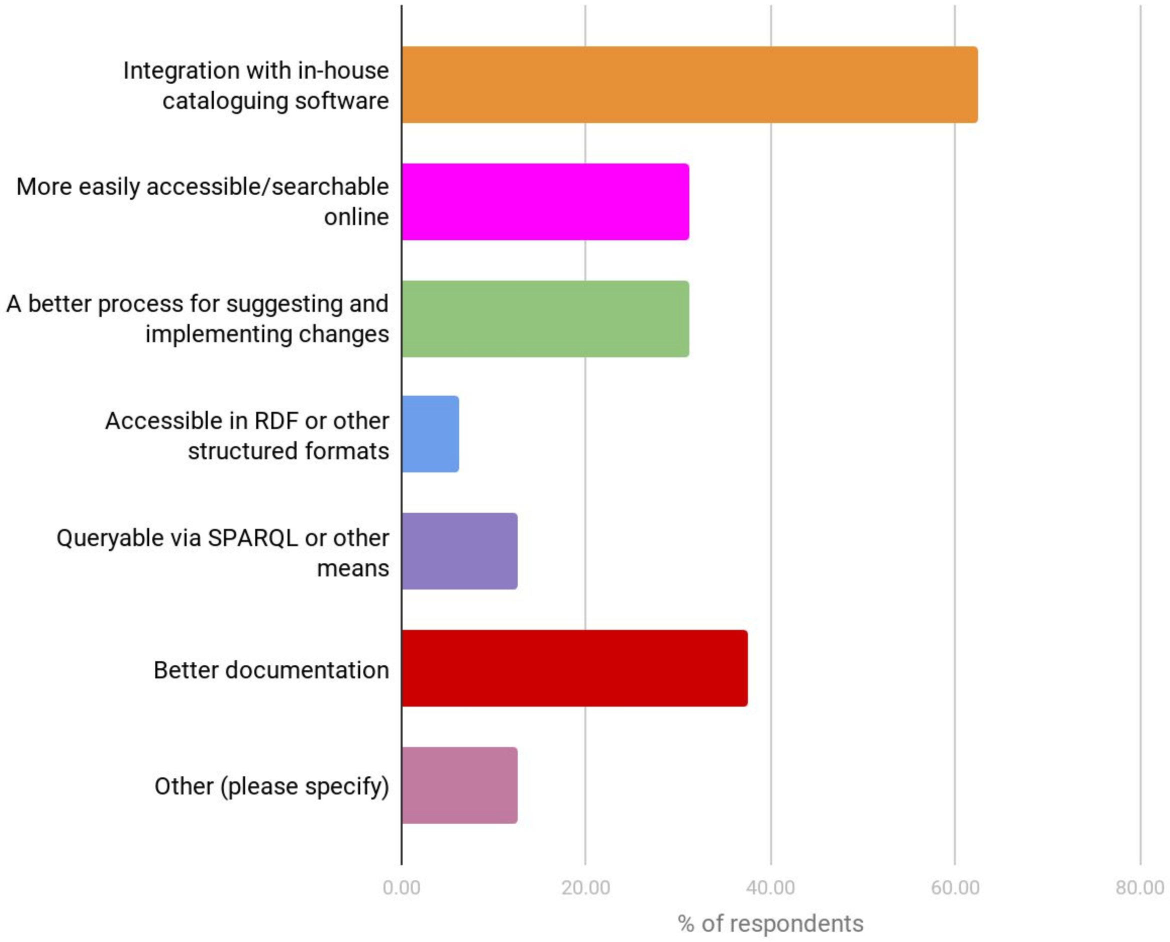
This question was intended to ascertain which, if any, controlled vocabularies were used by EHRI consortium member institutions.



Here, other than the Library of Congress Subject Headings (LCSH), most respondents noted that their institution used an in-house lexicon or thesaurus.

Q8. The EHRI portal hosts a number of controlled vocabularies including the EHRI subject term thesaurus and authority files about persons and corporate bodies. Would any of the following make it more likely that your institution could make use of them for cataloguing collections (check all that apply)?

The intention of this question was to better understand what EHRI could do to increase external usage of its controlled vocabularies.



One respondents used the “other” answer to comment on the difficulty of bringing data in-house, and how the feasibility of this would depend on the institutional need for data exchange. Another answered that it would depend on control over EHRI’s authority file entries.

Q9. In your view, what additional digital services should EHRI offer?

This final question was intended as a catch-all, to solicit additional ideas and comments about EHRI’s digital activities. Most respondents omitted this open-ended question or reaffirmed previous answers. Among other suggestions were:

- automatic translation for English-language material on the portal
- recognition of text in images (OCR)

4. Conclusions

This section summarises the above review of EHRI's technical requirements, along with the results of our consortium questionnaire.

Access to computational resources: Infrastructure-as-Code (IAC) should be considered as a more maintainable and self-documenting approach to Virtual Private Server (VPS) provisional and management of other Cloud-based resources.

Data storage and retrieval: While Neo4j still works well as our primary database, EHRI's shift from fixed-term research project to self-sustaining RI has, along with parallel technological developments, slightly undermined its advantages in a cost-benefit analysis relative to open-source RDBMS solutions such as PostgreSQL. While a wholesale database migration would be infeasible given projected development resources, creating a (non-real-time) RDBMS mirror of EHRI's Neo4j data would provide an escape-hatch if migration later became necessary and have other benefits in the meantime, such as additional data redundancy and access to a mature analytical toolset ecosystem.

EHRI's search engine should be equipped with more comprehensive automated query testing tools in order to facilitate easier (and safer) parameter tuning. Apache Solr should also be upgraded to a non-EOL version.

EHRI's overall data management strategy should incorporate the existing media storage needs of the portal, along with expanded requirements for managing limited amounts of archival media, as requested by survey respondents.

Archival data entry: The administrative interface that currently facilitates manual entry of collection and institution descriptions would benefit from updating to a more interactive style, and work in this area could potentially benefit other ISAD(G) or EAD-creation tools. EHRI's existing editor for SKOS-compatible data could likewise be of use outside the project if generalised to support other storage media (or browser-only client-side storage.) The survey supports the need for more open-source tools that support archival data creation.

Import of existing structure collection descriptions: The tools developed in EHRI-2 constitute a complete pipeline from a data provider's internal cataloguing system to EHRI's metadata store. While some tools, such as the MPT, require little or no additional work, others, such as the ECT and harvesting tools, would benefit from more integration with the portal administrative interface. The survey supports the view that making the ECT's validation and conversion functionality available as EHRI-hosted web-based services would be of potential benefit to data providers.

Harvesting of externally-hosted metadata: EHRI's tools for harvesting ResourceSync and OAI-PMH endpoints could benefit from further integration with portal administrative tools to facilitate easier testing, monitoring and configuration of harvesting activities. More generally, both the experience of EHRI-2 and the survey suggest that the hurdles for EHRI partner institutions to provide harvestable resources remain high, with the primary limiting factors being non-technical in nature. In light of this, the future RI should not limit its approach to data collection to harvesting alone.

Ingest data management: A more centralised approach to the management of data such as EAD files from third-parties (as well as harvested material) should be devised as part of the EHRI data management plan.

Review and verification: The system for hosting test imports for review and approval by data providers is complex to manage at present and alternatives should be investigated that serve this purpose in a more maintainable manner.

Data publishing and export: While EHRI publishes and facilitates the export of structured data in several forms, more work is needed in publishing *linked* open data. The survey also supports the need for EHRI to make its controlled vocabularies easier to use and adopt by third parties, a situation that could be incrementally addressed through improved documentation, a more transparent change process, and a more thorough and integrated approach to supporting RDF-friendly formats.

Documentation: While the documentation site covers several technical and administrative aspects of the portal and standalone tools like the ECT and MPT, we are currently lacking high-level documentation aimed at current and potential data providers, explaining the how’s and why’s of sharing data with EHRI. The survey also indicates the need for more documentation about the controlled vocabularies in particular.

———

Overall, surveying EHRI’s consortium members did not, in our view, reveal significant gaps in the existing infrastructure for metadata aggregation, though automated translation and text recognition (OCR) were suggested as possible additional services that could be provided to partner institutions in the future to aid with their cataloguing activities. It has also reaffirmed the need for a heterogeneous approach to data integration for future integration activities. In preparing for the future RI, therefore, an approach to the development of technical services based around consolidation and improved service integration, building on the tools developed over EHRI-1 and EHRI-2, is likely most appropriate one.

Implementation of the activities discussed above will take place both in the context of EHRI-PP and the forthcoming implementation phase, with high-priority tasks targeted for the former. Tasks prioritisation itself will be conducted in the coming weeks in cooperation with relevant stakeholders.

Appendix 1: SWOT Analysis

Internal	***Strengths*** ➔ Graph-based data store remains a stable and reliable platform with some unique and advantageous capabilities ➔ The MPT provides current and potential EHRI partners a way to publish data in harvestable form and requires little or no additional development or maintenance ➔ EAD validation and conversion functionality has been extensively prototyped ➔ Harvesting functionality has been successfully developed and tested with EHRI partner CHIs	***Weaknesses*** ➔ ECT requires additional integration with EHRI's web-based tools to be of practical use to current or potential data providers ➔ Harvesting functionality is complex to manage, test, and monitor, and requires knowledge of server administration ➔ Open data is not sufficiently "linked" through consistent URIs and other semantic web best-practices ➔ Validation/staging workflow is overly complex and time-consuming to administer ➔ Documentation for current or potential data providers is insufficient
External	***Opportunities*** ➔ Increased conformance to archival metadata standards among current or potential partner CHIs ➔ Increased use of standards-compliant cataloguing tools (e.g. AtoM) by partner CHIs makes harvestable data more common ➔ Maturing of new archival standards such as Records-in-Context facilitates new LOD capabilities ➔ Cloud-based infrastructure costs go down, allowing us to use more resources for the same budget ➔ Synergies with other data infrastructure projects relating to archival metadata integration	***Threats*** ➔ Current Cloud VPS providers cease providing required EU-based services or increase prices to a level we cannot sustain ➔ Neo4j becomes unavailable or future versions incompatible with the current backend data store ➔ New staff struggle to become familiar with infrastructure components due to their uncommon or unorthodox nature ➔ New 3rd party data becomes available for ingest does not match expected characteristics ➔ CHIs do not have the technical capabilities to use EHRI's tools ➔ New privacy regulations make archival data integration on EHRI's scale infeasible