DOI:
10.1108/LHT-07-2015-0070

Document Version
Peer reviewed version

Citation for published version (APA):
Blanke, T., Bryant, M., & Speck, R. (2015). Developing the collection graph. *LIBRARY HI TECH*, *33*(4), 610-623. https://doi.org/10.1108/LHT-07-2015-0070

Developing the Collection Graph

Journal:	*Library Hi Tech*
Manuscript ID	LHT-07-2015-0070.R1
Manuscript Type:	Original Article
Keywords:	Infrastructure, Research, Software, Archives, Collecting, Databases

1. Introduction

The European Commission continues to invest significant amounts of money into research infrastructures that bring together researchers across Europe and integrate their efforts (European Commission, 2015). While in the past, research infrastructures have been the sole domain of the sciences and have led to major scientific facilities such as large telescopes, infrastructures nowadays are also organised for the humanities and social sciences (Duşa et al., 2014). Their infrastructures are distinct from the large-scale scientific facilities as they are often distributed in nature and are mainly based upon an exploitation of the opportunities afforded by the digital transformation of research in the social sciences and humanities. For example, in 2010 the European Holocaust Research Infrastructure (EHRI, see http://www.ehri-project.eu) was funded to support research into the Holocaust (Blanke and Kristel, 2013) and (Speck et al., 2014). The project supplements significant national initiatives to develop and integrate the archival collections that document the event. Yad Vashem in Israel and the United States Holocaust Memorial Museum, for instance, have collected and copied many records on the Holocaust over the last decades. Both have also developed significant services to allow for remote access to these resources. Through EHRI, Europe has launched its own effort to integrate and provide access to Holocaust resources.

The Commission's definition of infrastructures as facilities, resources or services that support integrated research has proven to serve the sciences as well as the humanities (European Commission, 2015). For both, the promise of access to large amounts of data is a key motivation for research communities to join forces and to jointly develop facilities (Knobel, 2007). In the case of humanities research infrastructures much of the data for integration is not a product of the infrastructure itself. Instead, much of the relevant primary source materials is to be found in the traditional infrastructures of humanities research: collection holding institutions such as libraries, archives and museums (Speck and Links, 2013). As a consequence, humanities research infrastructures need to focus on the interrelated questions of how to integrate access to archives, connect knowledge and support the process of research for a particular community of researchers (Blanke and Hedges, 2013).

Archives and libraries have seen their own digital transformation that has started with the publication of online finding aids, catalogues, etc. but now increasingly includes digitised and/or born-digital collections (Duff et al., 2004). They have created their own dedicated portals and hubs to provide access to these collections. They have, for instance, collaborated in the context of the Europeana project to provide a single point of access to the holdings of cultural heritage institutions across Europe (Concordia et al., 2010), and they have been among the early adopters of semantic web-based access to catalogues and finding aids (Isaac and Haslhofer, 2013). Cultural heritage institutions have a strong tradition of fostering the democratisation of knowledge by facilitating public online access to their holdings.

Beyond the general public, the scholar has been an important client for both libraries and archives. Scholars are arguably among the main benefactors of the increasing online availability of the holdings of cultural heritage institutions. They also profit from new forms of access to digital material. While early digital libraries might have predominately focused on preparing online catalogues, very soon the vision of a digital library was expanded to one which includes working with digital materials in new ways. In one of the pioneer papers of the digital library movement Lagoze et al. (2005) argue that a digital library is not simply an online catalogue, but a place to meet and discuss as well as develop collaborations. In the reading rooms of traditional libraries and archives, scholars meet to work together, and this should not be different in the online world of digital libraries and archives.

A digital research environment is in this sense a continuation of the traditional one. However, a digital environment is often much more subject to change, as it is much easier to constantly add new material or update old one. EHRI had to react to this challenge and develop an integrated research environment that enables Holocaust researchers to work collaboratively on the information that EHRI provides. New modes of access to digital collections has invariably also lead to new research questions and to demands for faster access to these collections. This has become clear not just in the user requirement work that we have done in EHRI, but also in other work we have completed in the field of arts and humanities e-Research that has focused on different kind of user groups (Aschenbrenner et al., 2013) and (Blanke et al., 2010).

The requirement of a dynamically evolving research environment can be seen not only in the field of humanities research, but across several other disciplines where the focus is on interaction with research data. In fact, a focus on the ability to organise, research and analyse data and create exchanges around data has become one of the few general commonalities that binds together disciplines in various e-Science programmes. Indeed, in 2007, de Roure, one of the pioneers of the UK e-Science programme, announced a new e-Science paradigm, which concentrates on enabling the kind of interactions with data that research requires (de Roure, 2007). Libraries and archives, as important data and collection holding institutions, have therefore a key role to play in developing e-Research programmes.

EHRI's primary aim is to work with those libraries and archives that hold Holocaust-related material. We first mapped the landscape of collection holding institutions to better understand the organisations and practices that are involved (Blanke and Kristel, 2013). The history of the Holocaust and the Second World War had exacerbated a process commonly faced by many collection holding institutions, namely that important information is often dispersed, fragmented or even permanently lost (Speck et al., 2014). Most of the special collections of libraries and archives are incomplete and either intrinsically linked to, or even part of, collections in other institutions. This makes the integration and discovery of data particularly challenging. Thus, EHRI cannot foresee what kind of data we need to integrate next. On the contrary, we must expect the data from each collection holding institution to be unique in its characteristics and formats.

One possible solution that EHRI did not want to follow is to limit the amount of information we integrate by strictly imposing a common (meta-)data model on all institutions that provide us with data. Such a solution would be unacceptable to researchers as it reduces the amount of information they have access to. Our own user requirement work has clearly shown that researchers want to be able to make their own decisions on what information is relevant, and, therefore, that we had to keep information loss when integrating data to a minimum.

1.1. The EHRI technical architecture - Introducing the collection graph

We started to investigate a new kind of infrastructure that would allow us to focus on the content as a researcher would need it and take the data as we found it; heterogeneous and often incomplete but even in this form useful to researchers. Furthermore, it soon became apparent that the environment needs to support ad-hoc queries against the collected material, and especially queries of a kind we could not anticipate at the outset. Both requirements do not tally well with traditional data stores such as relational databases, which require a predefined understanding of the underlying data structure and redefinition of this structure in case there should there be a new type of connection.

Alternative semantic web approaches, while more compatible with our requirements, nevertheless rely mainly on triple stores queried via SPARQL [1] and are, therefore, also not flexible enough for us. In particular, such models have proven to be difficult to understand not just for researchers but often for developers, too. Recently, much research has gone into designing and developing more effective search and browse environments based on semantic web approaches. Our own research has shown (Blanke et al., 2012) that the graph structure of RDF [2] suits the typical ways of how humanities researchers explore sources. It supports the traditional types of browsing through connected sources. However, there are further disadvantages to the use of semantic web technologies such as triple stores that make them less suited to the data-driven research we would like to enable. We follow the critique of Sanderson (2013), who convincingly presented how RDF and triple stores pay too little attention to the data and too much to the structure. He recommended investigating new NoSQL [3] technologies instead.

Our work indeed shows that NoSQL technologies can provide a data infrastructure that supports the kind of dynamic research environment that we are seeking. They are excellent building blocks for a humanities research ecosystem. In this paper we concentrate on the part of the EHRI infrastructure that uses graph databases. Graph databases are a specific kind of NoSQL technology that have the additional benefit of allowing us to integrate all the many advantages of semantic web approaches for the publication and consumption of resources, as they can also function as a SPARQL endpoint (Blanke et al., 2013).

Graph databases have offered us a new approach that supports deep and rich investigation of data, and they seem a natural fit to our application domain of a research-led archival integration. They are a relatively old technology (Sadalage and Fowler, 2012) that have come to new prominence and achieved a new level of maturity within the NoSQL family of data stores. While most of the NoSQL databases are primarily concerned with the challenges of big data, graph databases address a different challenge traditional relational databases do not address well. They work well with large numbers of smaller records that are, however, heavily interconnected (Sadalage and Fowler, 2012). We have chosen graph databases rather than other types of NoSQL stores, as they can grow easily with any kind of new information that is added to them. According to Redmond et al. (2012), they scale best towards complexity or towards information that is not uniform. Secondly, they put relationships between information to the foreground. As discussed later on, the collections we are using in EHRI are full of complex relationships and thus naturally match the graph model.

Graph databases, however, pose the challenge that so far little work has been done with them in the field of research computing for history (at least when we started in 2010). This meant that we had to develop our own system and architecture from scratch, which increased our development risks. This paper will introduce the efforts by EHRI to create a flexible research environment using graph databases. It will present our attempt to use a novel set of graph technologies and methodologies to integrate material from diverse collection holding institutions about the Holocaust. In particular, the paper concentrates on the specific customisations we had to develop in the absence of existing solutions. In Section 2.1, we explain the serialisations of collections in the graph to provide for efficient processing. Because the EHRI infrastructure is highly distributed, we also had to invest a lot of effort into the development of a reliable distributed access control mechanism, which Section 2.2 describes. Finally, Section 2.3 analyses our user-facing work on a portal and a virtual research environment in order to discover, share and analyse Holocaust material.

2. EHRI customisation to the graph

2.1. Serialising graphs and provenance

EHRI's standard-based metadata model is layered upon the graph database in such a way that single items (e.g. an archival institution or document) consists of multiple primitives (nodes, relationships, properties). A document, for example, may have many distinct descriptions, each of which may contain a set of date ranges referring to some aspect of the provenance of the material to which it refers. In a traditional relational database management system, the document, its descriptions and the multiple date ranges would each reside as rows in a separate table and be linked by foreign key references. In EHRI's graph they are represented by distinct nodes and linked by typed relationships. Managing this composite structure in such a way that multiple graph primitives can be created, updated and deleted as a unit is a key responsibility of the EHRI's persistence layer.

Just as relational database systems have the notion of cascading triggers on foreign key references, allowing child rows of one table to react to actions affecting their parent, EHRI's graph is aware of which relationships from one node to another represent a parent-child dependency. These parent-child relationships form tree structures (a sub-graph, but directed in an acyclic manner, internally named bundles), and the main unit on which EHRI's persistence layer operates as a graph (Angles, 2012). As is common when dealing with tree structures, several of these operations are recursive in nature.

Since persistence is based around operations on generic data, domain models within EHRI's graph have an attenuated role as compared to many object-oriented systems, where the model encapsulates both its data and functionality relating to its own life-cycle (e.g. the model object creates, manages, and deletes itself.) Instead, EHRI's domain models (representing items within the archival domain, such as institutions and their material) are primarily lightweight containers for metadata. This metadata typically relates to relationships (a single document is the parent to many descriptions) and constraints (e.g., a description must have a name and a language code property).

One important aspect of the persistence layer is that it handles the lifecycle of domain items in an idempotent manner; a characteristic which simplifies both data ingest and change auditing. As an example of this behaviour, consider a scenario where an item (represented by a data tree) is ingested into the system twice in immediate succession. Rather than creating two identical items, or creating one item and then updating it with the same data it already has, the persistence layer will compare the data in the second tree with that which it has previously stored, and, noticing that no differences would result, leaves the system unchanged.

Another area where the EHRI graph needed to connect specific functionality to the underlying data layer is entity serialisation, the process of putting data into an external format suitable for consumption by client applications. As with ingest, the principal unit of serialisation is a recursive tree-like data structure (the bundle). However, there is an important distinction in that outgoing bundles can also contain both non-hierarchical and cyclical relationships, because they form a sub-graph rather than purely a tree. These serialised sub-graphs allow the registry to provide both the domain entity itself (for example, a documentary unit item) and its context (parent items and institution to which it belongs).

The amount and type of context provided with serialised entities is defined on the domain level and includes rules designed to ensure that densely connected and mutually referential groups of graph nodes do not cause issues due to loops and runaway recursions. Serialisation of context is also restricted to relationships that are naturally constrained, in that they only flow upwards through hierarchies (from child to parent, parent to grandparent, etc.) and not downwards, where an unbounded number of child items could result in an unmanageable explosion in the quantity of data pulled in.

One of the key objectives of the EHRI graph is to maintain transparency with regards to where the data originated from, and how it has subsequently been administered. Traditionally the trust in the acquisition by archives is based on organised ‘provenance’ of an archive’s records. Graph models can deal particularly well with provenance requirements (Vicknair et al., 2010). Provenance is an important part of the research process in archives. A study by Duff and Johnson (2002) is an excellent source for understanding how historians work through archival holdings. One of the key methods Duff and Johnson investigate is the ‘provenance method’ that links subject requests to the context of the organisation that is being investigated. To keep the provenance and comply with the provenance method, we have decided in EHRI to take each digital information object as a different one from its canonical item depending on the context it appears in (Blanke et al., 2013) and (Bryant et al., 2015). This is our digital transformation of the traditional archival principle of ‘respect des fonds’ (Horsman, 1994). In this sense, the canonical item ‘Israel|Yad Vashem| Jan Karski ‘ is different from ‘A’s Research| Yad Vashem| Jan Karski’ or ‘C’s Research|Monograph B| Yad Vashem| Jan Karski’.

To keep the provenance is a challenge since changes to the EHRI graph can be triggered in many ways: via a harvested update to ingested third party data; via manual data curation; and via automated enrichment processes. Maintaining clarity as to the provenance of information within the online environment was therefore critical to the success of EHRI and doing so suggested it was necessary to tightly integrate bookkeeping and auditing mechanisms into the core of the system.

Whenever an item is created, deleted or otherwise modified within EHRI's graph a corresponding event record is kept, connected with the user who initiated the change, the item(s) being affected, and various other pieces of

environmental metadata. This event stream is modelled as a linked list, meaning that every event (a node within the graph) is connected to its temporal predecessor and successor. Traversing the event stream therefore leads from the most recent event back in time to the point when the EHRI registry was first created.

In addition to maintaining an event stream for the system as a whole (the global stream), EHRI also maintains event streams for both users (those initiating an action) and individual items (multiple of which may be subjects of the same event in cases of batch operations). This enables the archive graph itself and client services to quickly answer questions such as “what has happened in the whole system?” or “which changes have individual users or items participated in?”.

Since events follow a well-defined temporal order, the linked list structure in which the EHRI graph maintains the stream puts the most up-to-date information closest to the subjects themselves in terms of the number of graph traversals needed to reach it. The ordered but potentially infinite and append-only nature of the EHRI event streams enable operations to provide efficient data access.

2.2. Role-based access control and permissions

Another key task of the EHRI research infrastructure is to manage a role-based access and permission system. This system was designed specifically to fit the requirements of a collaborative, transnational and multi-institutional research environment. As with the event/audition system discussed above, the access and permission controls are a vital and integral part of our responsibility to mediate the interaction between administrative staff from a wide range of backgrounds, the EHRI portal's public users (see Section 2.3 below) and the system’s metadata content. Experience gained with existing solutions around graph databases (Angles, 2012) pointed strongly to the advantages of building such responsibilities into the core of the system from the outset, rather than attempting to integrate them as an external component.

Graph databases are perfectly suited to traverse typical access control hierarchies in an efficient manner. Our implementation makes full use of the graph structures. Like the collection data itself and the organisational structures of many collection holding institutions, EHRI’s Access Control List/permission system is hierarchical and defined in terms of accessors, targets and scopes. The first component of this hierarchy revolves around what is termed an accessor, which is an entity to which permissions and access rights can be granted. In practice, an accessor is either an individual user or a group, to which both users and other groups can belong. Groups in the EHRI graph are effectively the same as roles within traditional role-based systems, but are thus named because they can serve other purposes in aggregating a set of users than that of the permission system.

Groups are intended to be able to model the structure of a collection holding institution through which roles and responsibilities derive, which is a key component of an institution's self-description and determines its authority. As such, they can form hierarchical structures, with sub-groups inheriting the attributes of their parent groups. Thus, we can model a specific archival institution, for instance, in the following manner:

The Archive
    Person A
    Archivists
        Person B
        Head Archivists
            Person C

In this case, persons A, B, and C are all members of the Archive group. Members B and C are members of the Archivists group, but only person C belongs to the Head Archivists group.

Up to this point we have discussed access control and permissions in the same terms, but the EHRI graph makes an important practical distinction between the two: access control is by default permissive, whereas permissions are by default restrictive. In other words, we assume that a given user can view a particular item unless instructed otherwise. By contrast, we assume that the same user cannot change an item unless we have been told that they can.

As with accessors, a target in the EHRI permission system is a polymorphic concept that can either be a content type or an individual item. A content type is a conceptual entity that represents all items of a given class: e.g., collection holding institutions, archival units, authority files, users and groups. A permission with a content type target applies uniformly to all items within that class. Conversely, a permission can apply to just one item.

Scopes provide the means to limit the granting of permission to within a particular hierarchical tree, in a manner analogous to locking a draw of a filing cabinet. While a permission with a content type target would normally apply to all items of that type, setting a scope limits it to just those items that are subordinate to the scope item.

Putting these three concepts together (accessors, targets and scopes), we can manage a wide range of permission-related scenarios demanded by EHRI's collaborative, transnational and cross-organisational structure. Typical examples from the everyday practices of the people working with the EHRI graph include, for instance, that staff in a given collection holding institution is enabled to manage and curate multiple types of data that apply to a particular country. In this case, the item representing the country is the permission scope, and the content types that can be subordinate to country items in EHRI's graph – archival institutions and, by extension, their archival unit metadata – the permission targets. Other examples are staff that manages the archival unit data within a particular institution, where the institution is the scope, or staff that looks after a single archival fonds and its children items, where the fonds is the scope.

To get a full picture of the flexibility of the system we need to introduce permission types. Permission types define granular actions such as creating, modifying or deleting. An important additional permission type is that of owning an item, which implies that items, which a given accessor has created, can also be modified and deleted by her. Owner permissions allow us to model common authority structures within physical collection holding institutions, whereby either an assigned collection manager has responsibility for managing and describing material within a particular domain, or a head collection manager has curation and editorial responsibilities that span domains.

This scenario is enabled in the following manner: Firstly, the collection manager role (the accessor) has create permissions (the type) for archival units (the content type target) within a given collection holding institution (the scope). This allows individuals belonging to the collection manager role to create new collection descriptions to which they will then automatically be assigned owner permission, meaning they can change (and if necessary delete) their own work, but not that of other individuals. Secondly, the head collection manager role has create, update and delete permissions for archival units within the collection holding institution. This allows individuals who belong to this role to have overall control to create, update and delete descriptions regardless of who created a given item. The ability for the target of permissions to be individual items in addition to item classes facilitates other common authorisation scenarios, such as allowing a user affiliated with a particular collection holding institution permission to modify its description (but not delete it, or create new archival units).

Furthermore, the access control and permission system is leveraged in several other ways. For instance, individual users have ownership permissions on their own profile, allowing them to update their personal information. User annotations (see next section) are private to the authoring user by default. However, individuals belonging to a moderators role can manage (or potentially update/delete) those annotations that users would like to make publically available.

The tree-like nature of permission scopes and the ability to nest groups within other groups means that the permission system must traverse two dimensions of hierarchy to calculate whether a given user has the authority to perform a given action and determine and aggregate the permission grants of any groups to which they might belong. If any of those grants are scoped to specific items, the users can explore whether those scope items exist within the permission scope ancestry of the subject item or content type. The ability of the graph database to traverse efficiently through these potentially unbounded hierarchies enables such complex operations to have minimal overhead, whilst remaining a generic system that is not tied to the semantics of any particular content type: the same system is used for managing all types of data within EHRI's graph, from collection metadata to user annotations.

2.3. External interfaces: Portal and VRE

The EHRI web portal (https://portal.ehri-project.eu) consists of two principal components: first, an interface through which researchers can explore EHRI's integrated data via free-text search and interconnected browsing; and second, a virtual research environment (VRE) where users can take notes, manage items of interest to them and explore and connect with other researchers on the site.

EHRI's data is structured in a hierarchical manner. The top level and intended entry-point for Holocaust research are the countries, which EHRI has detailed in its country reports (Bennett et al., n.d.). These reports summarise the body of knowledge on the Holocaust in particular countries. In addition to the report information, researchers can browse the collection holding institutions within each country that EHRI has determined hold Holocaust-related materials. From each institution, researchers can (for roughly 25% of institutions at the time of writing) proceed to a list of the collection descriptions held therein. Collection descriptions are similarly hierarchical, typically consisting of a top-level description with a varying number of child levels.

This country/institution/document hierarchy provides an overarching logical structure to EHRI's portal but it is not the only way to navigate the site. A universal free-text search facility returns results for any types of material matching a user's query, on the basis that the country reports and descriptions of institutions may be just as relevant in answering a researcher's question as descriptions of collection material. If a user has a more focused idea of what she is looking for, any of the hierarchical scopes discussed above (country, institution, material) can likewise be searched, allowing cases such as search document descriptions within this institution or search child items within this collection.

Where a specific item type is the target of a user's query, search results can be further narrowed via the application of facets to filter the data. Facets consist of coarse categories into which data can be clustered, such as language, source and level of detail and provide an important narrowing mechanism where a textual search query results in an overly broad set of matches.

An additional way for users to refine their search queries is by adding a field constraint, restricting matching to those where the query applies to a specific part of the target records. Fields available in the first release of the EHRI portal include the record identifier and title, as well as access point attributes such as persons, places and subject keywords. Since EHRI's collection holding institution database functions substantially as a directory, there is also a field specialised for searching addresses.

One aspect of the close integration between searching and hierarchical browsing of country, institution and collection metadata in the EHRI portal is that record-ordering is context-dependent; requiring nuanced behaviour and the application of certain heuristics. It was our experience that while record-

ordering was a significant component of the way in which individual collection holding institutions organised their material – a key part of the record context - it was in many cases not explicitly encoded in structured data received by EHRI.

Several scenarios were identified in which different ordering criteria are used:

1. If the user has provided a textual query we order records by relevance as determined by the search engine.
2. In the absence of a textual query, subordinate items (e.g. collection material belonging to a repository) are ordered by the local identifier as provided by the source institution, since this is the best proxy for their native ordering
3. Top-level views (e.g. all collection material) are otherwise ordered by the date of their last modification
4. In any situation, a user can opt to order results in a specific manner, the available options being the local identifier, title, date of last update and the record detail (a proxy for the amount of textual information present.)

One of the findings of EHRI's user requirements investigation was that Holocaust researchers originate from diverse backgrounds and possess a correspondingly wide set of working methods. The virtual research environment (VRE) features of the EHRI portal (Blanke et al., 2010) are therefore designed to complement a user's data exploration and research practices without attempting to impose upon them a particular workflow or set of tools. The VRE features take care of keeping track of relevant material as well as enabling the finding and contacting of other researchers.

Before she can make use of the portal's VRE features, a user has to create an EHRI account. We have endeavoured to make this as straightforward as possible by supporting common third-party authentication systems (such as Google, Yahoo, Facebook and OpenID) in addition to the typical local account. Every user requires a valid email address, which can be subsequently used to reset lost or forgotten passwords. Once they have created an account a user can opt to provide a photograph and details such as their location, interests and areas of research. These details are available to other registered EHRI users and are searchable and browsable. While a user's email address is not publically visible, an account by default enables a user to be contacted by other registered users via a messaging form.

The first feature intended to facilitate keeping track of relevant material in the EHRI portal is the ability for registered users to create notes on collection material. Notes can be created on both item descriptions and individual fields within a description and are visible to that user both in-context in the portal's browsing interface, and on the user's personalised notes page, where they can also be searched and exported in JSON, CSV or plain text format. Notes are, by default, private to the user who created them and invisible to others. If a user wishes to make a note publically visible, she can indicate this, either at the time of creation or later, adding it to a moderation queue. Members of EHRI's moderation group can then vote on notes submitted for publication.

Given the sensitivity surrounding Holocaust research it is of utmost importance to avoid the publication of inappropriate user-generated content, and only notes with a positive voting score become publically visible.

Another feature for keeping track of material is the ability to watch items by clicking the star icon that appears beside them in list views or on the item's detail page. In addition to adding the item to a searchable list that persists across browser sessions, watching also means that updates to the chosen items (such edits by administrators and the addition of public notes made by other users) will appear in the user's activity stream. The items in a user's watch list can also be exported in the same manner as notes.

In addition to messaging those who have created an account on the EHRI portal, it is also possible to follow other users and be followed in return. Following a user (the terminology is derived from the Twitter service) means subscribing to their activity in the portal's personalised activity stream, adding to the timeline notices when they add material to their watch list or create publically visible notes. It is intended, therefore, that the portal's personalised activity stream serves to facilitate serendipitous discovery by alerting users to the existence of material that is of interest to others.

The EHRI web service layer finally exposes the functionality of the EHRI graph via an HTTP interface; allowing interaction from clients in a manner that is programming-language agnostic. The web service layer constitutes our interface to the EHRI graph for other machines rather than human access through the portal. The interface to the service is based around the common Representational State Transfer (REST) pattern, using the HTTP verbs GET, POST, PUT and DELETE to manage content that is identified by Uniform Resource Identifiers (URIs) [4]. The general structure of this interface adheres to the following template:

URI	Function
{content-type}	POST to create a new item of a given type
{content-type}/{identifier}	GET, PUT, DELETE to fetch, update, and remove an item
{content-type}/{identifier}	POST to create a child item
{content-type}/list	List items of a given type
{content-type}/count	Count items of a given type

The full web service interface is detailed in http://ehri.github.io/docs/api/ehri-rest/ehri-extension/wsdocs/index.html. A significant aspect of the web service interface is its support for streaming responses that can handle potentially large amounts of data in an efficient, scalable manner.

As described above in Section 2.2, the EHRI graph maintains a global event log, plus temporally ordered logs for both item-specific events and the users whose actions trigger them. The continuous, ordered and unbounded nature of these event logs makes them natural candidates for access via streaming,

enabling large amounts of data to be transferred to clients in a scalable manner that does not need to rely on pagination. Such streams allow operations such as downloading the entire history of changes within the EHRI online system as a single archive.

The second area in which streaming patterns are a key feature of the EHRI architecture are those activities that involve accessing, exporting, or transforming large batches of data. One such activity is indexing the EHRI graph data. Apart from facilitating browsing of a graph, graph databases also offer technologies for more traditional searching. Graph databases work well with some advanced search techniques such as facetted searching and filtering. Our graph database tightly integrates with Solr [5] (Blanke et al., 2013). Any nodes and their properties can be indexed, which next to generic text-based searches can lead to very effective browsing possibilities for the underlying document space. Sub-graphs can be flexibly combined by using external indeces. Solr will also allow us to support more advanced means of access to facts in the documents and to enable deep and semantically meaningful access to the documents in the future.

Thus, EHRI's Solr-based search infrastructure is decoupled from the EHRI store for reasons of modularity and considered a secondary, expendable store. This means that it is important to quickly rebuild the search index in full in the event that it becomes de-syncronized, since this time is experienced as down-time by users of the EHRI portal. The EHRI store enables efficient bulk export of data in item type categories as single item streams, which are then transformed and indexed in a single transactional operation, with constant memory usage regardless of the stream size. Such an approach allows us to re-index approximately 65,000 domain objects (with rich metadata) per minute.

3. Conclusion

This article has presented the novel approach of the European Holocaust Research Infrastructure to comply with the needs of its research community. EHRI's work on integrating dispersed research collections from a range of collection holding institutions has inspired the research community to new thinking and topics of enquiry. EHRI has managed to develop a research infrastructure and a digital platform that allows researchers to investigate trails of research materials across collection holding institutions, connect their evidence and discover new material. In order to achieve these aims, EHRI had to develop an innovative digital environment that corresponds to the fast changing needs of the research communities and allows for dynamic enquiry into collection material. This article has presented our graph approach to achieve these aims.

Research infrastructures are best understood as complex ecosystems and networks that need to consider evolving user needs and dynamic changes to the research material they contain. We have chosen to implement the EHRI integrated information resource using graph databases, and to experiment with new additions that customise graph databases for our specific needs. With their emphasis on relationships, graph databases are particularly well suited for historical research in particular and humanities research in general. In this article, we have concentrated on those parts of the infrastructure that we had to add to the generic deployment of graph databases.

First we discussed how we serialised the collections we were given by the partner institutions into our graph. We had to take various steps to make their retrieval more efficient. We needed to serialise tree structures in order to allow for a better ingest and easier consumption by clients. A key challenge we had to address was how to manage these serialisations in such a way that multiple graph primitives can be created, updated and deleted as a unit. The EHRI graph database made the serialisation of key materials in collections holding institutions easy and also helped with maintaining an order whenever an item is created, updated or deleted. We implemented a linked list of event records that kept track of all these events. We have shown how this has helped us to keep important provenance information in EHRI.

The graph database not only efficiently supports the implementation of the EHRI collection environment, but also the management of this environment through a comprehensive implementation of access control mechanisms. Based on past experiences, we decided to implement this business logic directly into the graph, as this will work better in highly distributed curation environments, where several editors and authors work on the holdings simultaneously. This required a complex access permission system that reflects not just the organisational structure of EHRI but also the ones of its participating institutions. We used the graph structure to implement specific requirements for accessors, targets and scopes.

Finally, we presented our user-facing work and the EHRI portal with its VRE functionalities that uses graphs to allow for effective browsing of the EHRI collections as well as searching based on full-text indeces. We implemented typical search facilities such as facetted browsing. In particular, we let users define scopes within the EHRI graph to search, for instance, sub-graphs linked to a particular institutions. The VRE uses the graph to link annotations and other user work to EHRI material. Researchers can add notes to EHRI graphs. The EHRI environment further enables the publication of these annotations through various standard export mechanisms, while users can follow other users' updates in the graph.

EHRI has just received another four years of funding through the European Union's Horizon 2020 program, which we will use to continue our work exploring the potential of graph databases and to investigate further the social platform for EHRI in order to support communication between researchers and between collection specialists and researcher. We will enhance especially the VRE capabilities of EHRI and provide stronger external interfaces to our EHRI graph.

4. References

Angles, R. A. (2012), "A comparison of current graph database models", in *Proceedings 2012 IEEE 28th International Conference on Data Engineering Workshops (ICDEW 2012)*, IEEE, Los Alamitos, pp. 171-7.

Aschenbrenner, A., Blanke, T., Fritze, C., Pempe, W. (2013), "Data-Driven Research in the Humanities — the DARIAH Research Infrastructure", in Atkninson, M. et al. (eds.), *The Data Bonanza: Improving Knowledge Discovery in Science, Engineering, and Business,* Wiley, Hoboken, pp. 417-30.

Blanke, T., Candella, L., Hedges, M., Priddy, M., Simeoni, F. (2010),"Deploying general-purpose virtual research environments for humanities research", *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences,* Vol. 368**,** pp. 3813-28.

Blanke, T., Bodard, G., Bryant, M., Dunn, S., Hedges, M., Jackson, M. and Scott, D, (2012), "Linked data for humanities research—The SPQR experiment", in *Proceedings of the 2012 6th IEEE International Conference on Digital Ecosystems Technologies (DEST)*, IEEE, pp. 1-6.

Blanke, T., Bryant, M., Hedges, M. (2013), "Back to our data—Experiments with NoSQL technologies in the Humanities", *Proceedings of the 2013 IEEE International Conference on Big Data*, IEEE, pp. 17-20.

Blanke, T., Hedges, M. (2013), "Scholarly primitives: Building institutional infrastructure for humanities e-Science", *Future Generation Computer Systems,* Vol. 29**,** pp. 654-61.

Blanke, T., Kristel, C. (2013), "Integrating Holocaust Research", *International Journal of Humanities and Arts Computing,* Vol. 7**,** pp. 41-57.

Bennett, G. et al. (eds.) (n.d.), "Country Reports on Holocaust History and Archives, and the Data Identification and Integration Work of the European Holocaust Research Infrastructure (EHRI)*"*, available at http://ehri-project.eu/country-reports (accessed 10 September 2015).

Bryant, M., Reijnhoudt, L., Speck, R., Clérice, T., Blanke, T. (2015), "The EHRI Project: Virtual Collections Revisited", *Springer Lecture Notes in Computer Science (LNCS)*, vol. 8852, pp. 294-303)

Concordia, C., Gradmann, S., Siebinga, S. (2010), "Not just another portal, not just another digital library: A portrait of Europeana as an application program interface", *IFLA journal,* Vol. 36**,** pp. 61-9.

De Roure, D. (2007), "The New e-Science", paper presented at IEEE e-Science Conference, Bangalore, available at: http://www.slideshare.net/dder/the-new-science-bangalore-edition (accessed 10 September 2015).

Duff, W., Craig, B., Cherry, J. (2004), "Historians' use of archival sources: Promises and pitfalls of the digital age", *The Public Historian,* Vol. 26**,** pp. 7-22.

Duff, W., Johnson, C. A. (2002), "Accidentally found on purpose: information-seeking behavior of historians in archives", *The Library Quarterly***,** Vol. 72, pp. 472-496.

Duşa, A., Nelle, D., Stock, G., Wagner, G. G. (2014), *Facing the Future: European Research Infrastructures for the Humanities and Social Sciences*, Scivero, Berlin.

European Commission (2015), "Research Infrastructures", available at http://ec.europa.eu/research/infrastructures/index_en.cfm (accessed 10 September 2015).

Horsman, P. (2013), "Taming the elephant: An orthodox approach to the principal of provenance", Abukhanfusa, K and Sybeck, J. (eds.), *The principle of provenance. Report from the first Stockholm conference on archival theory and the principle of provenance*, Swedish National Archives, Stockholm, pp 51-63.

Isaac, A., Haslhofer, B. (2013), "Europeana linked open data–data.europeana.eu", *Semantic Web,* Vol. 4**,** pp. 291-297.

Knobel, C. (2007), "Understanding infrastructure: Dynamics, tensions, and design", available at http://escience.caltech.edu/workshop/UnderstandingInfrastructure2007.pdf (accessed 10 September 2015).,

Lagoze, C., Krafft, D. B., Payette, S., Jesuroga, S. (2005), "What is a digital library anymore, anyway", *D-Lib Magazine,* Vol. 11**,** n.p..

Redmond, E., Wilson, J. R., Carter, J. (2012), *Seven databases in seven weeks: A Guide to modern databases and the NoSQL movement,* Pragmatic Bookshelf, Dallas.

Sadalage, P. J., Fowler, M. (2012), *NoSQL distilled: a brief guide to the emerging world of polyglot persistence*, Pearson Education.

Sanderson, R. (2013), "RDF: Resource Description Failures and Linked Data Letdowns", Paper presented at CNI Spring Forum, 4-5 April 2013, San Antonio, available at http://journalofdigitalhumanities.org/2-3/rdf-resource-description-failures-and-linked-data-letdowns/ (accessed 10 September 2015).

Speck, R., Blanke, T., Kristel, C., Frankl, M., Rodriguez, K. & Vanden Daelen, V. (2014), "The Past and the Future of Holocaust Research: From Disparate Sources to an Integrated European Holocaust Research Infrastructure", Rapp, A. et al. (eds.), *Evolution der Informationsinfrastruktur: Forschung und Entwicklung als Kooperation von Bibliothek und Fachwissenschaft*, Verlag Werner Hülsbusch, Glückstadt, pp. 157-77.

Speck, R., Links, P. (2013), "The Missing Voice: Archivists and Infrastructures for Humanities Research", *International Journal of Humanities and Arts Computing,* Vol. 7**,** pp. 128-46.

Vicknair, C., Macias, M., Zhao, Z., Nan, X., Chen, Y., Wilkins, D. A (2010), "Comparison of a graph database and a relational database: a data provenance perspective", in *Proceedings of the 48th annual Southeast Regional Conference*, ACM, p. 42.

5. Endnotes

[1] SPARQL stands for SPARQL Protocol and RDF Query Language. It is a semantic query that enables the retrieval and manipulation of data stored in Resource Description Framework (RDF) format.
[2] RDF stands for Resource Description Framework. RDF encompasses a variety of specifications of the World Wild Web Consortium (W3C), and provides a general method for the conceptual description and modelling of web resources.
[3] „NoSQL“ or „non relational“ stands for methods of storing and retrieving data that is not modelled according to the tabular relations that are used in relational database systems.
[4] In practice, EHRIs resource identifiers are Internationalised Resource Identifiers (IRIs) because they can contain more than just US-ASCII characters.
[5] Solr is an open source search platform developed in the Apache Lucene project.