Ubiquitous Access to Digital Cultural Heritage

CHRISTIN SEIFERT, University of Passau
WERNER BAILER and THOMAS ORGEL, Joanneum Research
LOUIS GANTNER, Archeology and Museum Baselland
ROMAN KERN and HERMANN ZIAK, Know-Center GmbH
ALBIN PETIT, JÖRG SCHLÖTTERER, STEFAN ZWICKLBAUER, and MICHAEL GRANITZER, University of Passau

4

The digitization initiatives in the past decades have led to a tremendous increase in digitized objects in the cultural heritage domain. Although digitally available, these objects are often not easily accessible for interested users because of the distributed allocation of the content in different repositories and the variety in data structure and standards. When users search for cultural content, they first need to identify the specific repository and then need to know how to search within this platform (e.g., usage of specific vocabulary). The goal of the EEXCESS project is to design and implement an infrastructure that enables ubiquitous access to digital cultural heritage content. Cultural content should be made available in the channels that users habitually visit and be tailored to their current context without the need to manually search multiple portals or content repositories. To realize this goal, open-source software components and services have been developed that can either be used as an integrated infrastructure or as modular components suitable to be integrated in other products and services. The EEXCESS modules and components comprise (i) Web-based context detection, (ii) information retrieval-based, federated content aggregation, (iii) metadata definition and mapping, and (iv) a component responsible for privacy preservation. Various applications have been realized based on these components that bring cultural content to the user in content consumption and content creation scenarios. For example, content consumption is realized by a browser extension generating automatic search queries from the current page context and the focus paragraph and presenting related results aggregated from different data providers. A Google Docs add-on allows retrieval of relevant content aggregated from multiple data providers while collaboratively writing a document. These relevant resources then can be included in the current document either as citation, an image, or a link (with preview) without having to leave disrupt the current writing task for an explicit search in various content providers' portals.

CCS Concepts: • **Applied computing** → **Digital libraries and archives**; **Document searching**; *Document metadata;*

Additional Key Words and Phrases: Search aggregation, user context detection, metadata harmonization

ACM Reference Format:

Christin Seifert, Werner Bailer, Thomas Orgel, Louis Gantner, Roman Kern, Hermann Ziak, Albin Petit, Jörg Schlötterer, Stefan Zwicklbauer, and Michael Granitzer. 2017. Ubiquitous access to digital cultural heritage. ACM J. Comput. Cult. Herit. 10, 1, Article 4 (April 2017), 27 pages.
DOI: http://dx.doi.org/10.1145/3012284

The presented work was developed within the EEXCESS project funded by the European Union Seventh Framework Programme FP7/2007-2013 under grant agreement 600601.

Authors' addresses: C. Seifert, A. Petit, J. Schlötterer, S. Zwicklbauer, and M. Granitzer, University of Passau, Germany; emails: {christin.seifert, albin.petit, joerg.schloetterer, stefan.zwicklbauer, michael.granitzer}@uni-passau.de; L. Gantner, Archäologie und Museum Baselland, Amtshausgasse 7, CH-4410 Liestal; email: louis@gantner.ch; W. Bailer and T. Orgel, JOANNEUM RESEARCH—DIGITAL, Steyrergasse 17, 8010 Graz, Austria; emails: {werner.bailer, thomas.orgel}@joanneum.at; R. Kern and H. Ziak, Know-Center GmbH, Inffeldgasse 13/6, 8010 Graz, Austria; emails: rkern@know-center.at, hziak@know-center.at.

DOI: http://dx.doi.org/10.1145/3012284

1. INTRODUCTION

In the past three decades, GLAMs (Galleries, Libraries, Archives, and Museums) invested a lot of effort in various digitization activities within their institutions. A primary motivation of this digital wave was the need to introduce "object management systems" like databases (also termed *collection management systems* or *inventory systems*). Additionally, cultural heritage institutions aimed at digitizing parts of their physical collections using technologies like digital photography or scans (for paper documents, paintings, sculptures, photographs, maps, etc.) along with the describing metadata to provide easy access and at the same time preserve the object from physical destruction (preservation and access initiatives).

Growing technological capabilities and the availability of the Internet triggered the next phase in the year 2000. The GLAM institutions started to open up their collections for the World Wide Web with the goal to connect with peer institutions and users directly. The open access movement also included the cultural domain and therefore broadened the requirements for online collections in science and the humanities. The subsequent activities resulted in the deployment of different sorts of Web-access services like Web portals. The need for standardization was evident and in fact still remains a challenge. Although standards had already existed for many years for libraries, especially the museums and archives were and still are challenged with the adoption and introduction of standards for both museum management processes and metadata schemas (e.g., CIDOC CRM,[1] LIDO,[2] EDM[3]).

In summary, the digitization initiatives can be viewed as overlapping and interacting processes responding to internal and external requirements. This affects the domain of metadata management, technical infrastructure, software design, and rights management facing fast advances in technical evolution (Internet and social media). From a scientific point of view, the topics of digital curation, online management, provision, and exploitation of digital resources are embraced by the term and the research area of digital humanities, which covers both the traditional disciplines of the humanities and the latest developments in computing (data and text mining, visualizations, digital mapping, etc.).

1.1 Problem Setting

The current landscape for cultural heritage data can be characterized as follows. On the one hand, there are data silos in large and small institutions that have been made accessible through APIs and their own portals, such as the Public Library of America,[4] the University Libraries of Switzerland,[5] or Europe's aggregator of cultural content Europeana.[6] On the other hand, there are the culturally interested users, students, librarians, or researchers who want to access, search, and discover content of their interest within those data silos. These actors are characterized by their personal interest and current task. The distribution of content in various data silos requires these actors to (i) find the relevant access point to the data silo (if available) and (ii) search the content within the silo. General-purpose search engines (e.g., Google) realize a federated search through all (indexed) repositories and try to help users find the content, but they only partly address the problem for cultural content. First, those search engines are optimized for "mainstream content," whereas cultural content resides in the so-called long tail of the Internet [Barabási et al. 2000]. Second, not all content is available for indexing, due to licensing and rights management issues.

[1]http://www.cidoc-crm.org/.
[2]http://network.icom.museum/cidoc/working-groups/lido/what-is-lido/.
[3]http://pro.europeana.eu/page/edm-documentation.
[4]http://dp.la/.
[5]https://www.swissbib.ch/.
[6]http://europeana.eu/.

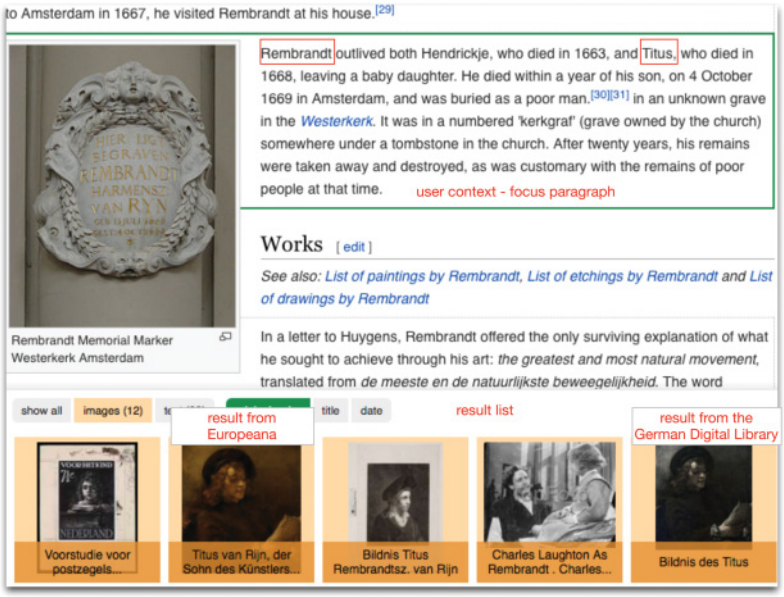

Fig. 1. Illustration of the proposed approach. Content is aggregated from different providers into a single result list and injected into the users' current digital context (the Wikipedia page of Rembrandt).

Thus, for a contextualized personalized discovery of cultural content, the following problems have to be solved. First, a single point of discovery is necessary that searches all related data silos, which in turn requires an intelligent aggregation of results and a harmonization of metadata. Second, users should be able to access the content from within their current task, and results should be contextualized toward the task and personalized based on user interests and knowledge.

1.2 Approach

To achieve the goal of ubiquitous access to cultural heritage, we propose the following conceptual idea: the single point of discovery is realized by a federated search and recommendation approach, accessing relevant data silos and integrating their content into a single result list (federated aggregation). Responses from all accessed data silos are harmonized with respect to their metadata (metadata harmonization). The request to the access point is injected into the user's current whereabouts (e.g., blogging frameworks (content injection)), following the principle of bringing the content to the user [Granitzer and Seifert 2016]. To ensure user acceptance, methods for individually retaining user privacy are employed while offering the full range of personalization (privacy preservation). Figure 1 illustrates this approach. When browsing a Web page (e.g., Wikipedia), relevant results aggregated from different data providers (e.g., the German digital library and Europeana) are presented in a single result list. The search is performed automatically based on the terms available on the page (in this example, "Rembrandt" and "Titus van Rijn").

In this article, we describe the EEXCESS[7] infrastructure for accessing cultural heritage content from the long tail of the Internet. This infrastructure comprises components for content aggregation from various sources and content injection into various client platforms. Single components communicate with standard Web technologies and well-defined APIs, and they can be used as stand-alone

[7]http://eexcess.eu.

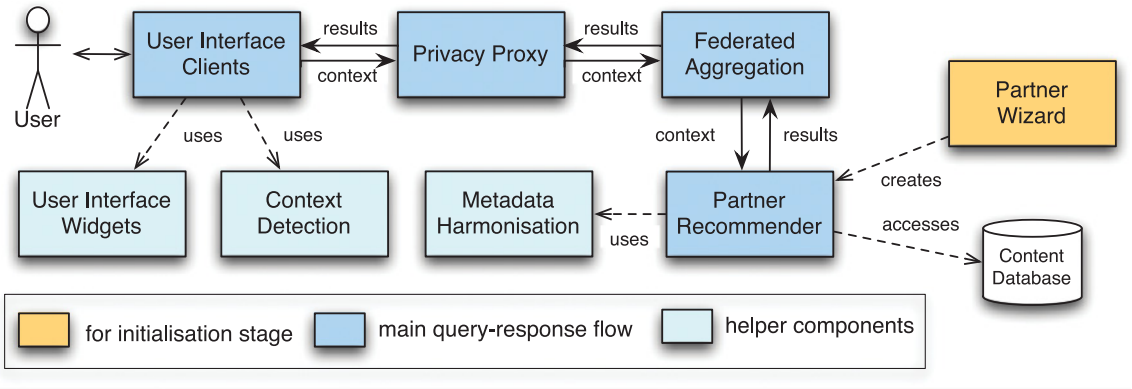

Fig. 2. Overview of the infrastructure outlining core components. The main query workflow is indicated with solid lines, and helper calls are indicated with dashed lines.

components with only minor modifications. The infrastructure and components are available as open source.[8]

The article is organized as follows. In Section 2, we describe the overall architecture outlining the single components and their interaction. Section 3 provides more details and evaluation results for single components. Example application scenarios with a focus on content providers and content consumers of cultural heritage content are described in Section 4. Related work is discussed in Section 5. Section 6 presents a summary and outlook on future work.

2. OVERALL ARCHITECTURE

The presented architecture is the basis for realizing (i) aggregation of various sources, (ii) provision of personalized content, and (iii) dissemination of content into various channels. In this section, we provide an overview of the architecture, the core components, and their interactions. Further, we detail the workflow from user context detection to the provision of related resources.

An overview of the architecture is shown in Figure 2, with the conceptual components being described in the following. An overview of the repository locations of individual components is provided in Table I.

The *user interface clients* component is the conceptual component with which end users directly interact (for more details, see Section 3.4). The component contains interface elements for explicit search for related resources and/or mechanisms for implicitly issuing automatically generated searches (making use of the context detection component). The user interface is also responsible for presenting the retrieved resources and might contain complex mechanisms for result interaction (e.g., advanced filtering tools). Available clients are, for example, a Google Docs plugin (see Section 4.4), an extension to the Chrome browser (see Section 4.3), a Wordpress plugin, and a plugin for the Moodle e-Learning platform.

The *context detection* component is responsible for observing the user behavior on the client and generating user interest profiles and search queries. This component is available as a library and can be easily integrated in any user interface client as long as this client supports standard Web technologies (HTML, JavaScript). Context detection is described in detail in Section 3.4.2.

[8]https://github.com/EEXCESS.

Table I. Client Applications and Component Overview

Component	Source Code URL[1]	Language	Comment
Client Applications			
Chrome extension	*GHE*/chrome-extension	HTML, CSS, JavaScript	available from Chrome Web store purl.org/eexcess/clients/chrome-extension
Google Docs plugin	*GHE*/gdocs-plugin	HTML, CSS, JavaScript	available from Google App Store purl.org/eexcess/clients/googledocs-plugin
Wordpress plugin	*GHE*/wordpress-plugin	HTML, CSS, JavaScript	available in Wordpress Store, plugin name "'EEXCESS"'
Moodle plugin	*GHE*/MoodleServerPlugin, *GHE*/MoodleAttoEditorPlugin	PHP	installable versions available from purl.org/eexcess/install/moodle-server-plugin and purl.org/eexcess/install/moodle-atto-editor-plugin
Components			
Context detection	*GHE*/c4	JavaScript	C4 library available via Bower[2]
User interface widgets	*GH*/visualization-widgets	HTML, CSS, JavaScript	Provides simple result lists and more complex visualizations as modules
Privacy protection	*GHE*/peas	JavaScript	Client components for privacy preservation
Privacy proxy	*GHE*/privacy-proxy	Java	Server components for privacy preservation
Federated aggregation	*GHE*/recommender	Java	Includes federation and single partner recommenders
PartnerWizard	*GHE*/EEXCESS/PartnerWizard	Java	Consists of a user interface and the partner recommender–generating subcomponent
Metadata model	eexcess.eu/schema/eexcess.owl	OWL	Metadata model (based on EDM and W3C PROV)
Metadata quality tools	*GHE*/data-quality	Java	Library to determine data quality metrics

[1]*GHE* is an abbreviation for the domain github.com/EEXCESS. [2]http://bower.io.

The *privacy proxy* is responsible for removing, hiding, and perturbing user-sensitive information. All communication between the client and server should go through the privacy proxy to ensure user privacy. This component is described in detail in Section 3.5.

The *federated aggregation* component is responsible for distributing the query to the *partner recommenders* (which access their own *content database*), retrieving the results from single partners, aggregating the result list, and delivering the results in a common format to the client components. This component is described in detail in Section 3.2.

Metadata harmonization contains the definition of a unified metadata model and a mapping component responsible for mapping a partner's metadata to the unified data model. This component is described in Section 3.1. *PartnerWizard* allows content providers without programming skills to add their system to the EEXCESS ecosystem by creating partner recommenders via a guided user interface. This component is used in the partner setup phase and is explained in more detail in Section 3.3.

The general workflow for injecting personalized, cultural content into the client is as follows. The client detects the user's information need based on the current user context (e.g., the edited document, the visited Web page, the browsing history). The client generates a search queryand a user interest profile, which is sent through the privacy proxy to the federated aggregation component. The privacy proxy ensures the user's unlinkability and indistinguishability. The federated aggregation component distributes the search query and the user profile to the partner recommenders, taking their availability and vocabulary into account. The partner recommenders retrieve relevant results from their content database and return them to the aggregator in a harmonized format making use of the metadata

harmonization component. The aggregator integrates the results into a single result list, reranking based on the user profile and criteria such as diversity. The final result list is then returned to the client and presented to the users.

3. COMPONENTS

This section describes the components in more detail, starting with the metadata integration (in Section 3.1) and federated aggregation component in Section 3.2. PartnerWizard, our solution for automatic integration is presented in Section 3.3. Section 3.4.1 then describes the user interface components, including the widgets for displaying content and the modules for detecting user context for automatic search. The solution on privacy protection in the architecture is described in Section 3.5.

3.1 Metadata Harmonization

EEXCESS uses a federated model to collect query results from different data providers (each in the format returned from the provider's API), to combine and rerank these results and provide them to a range of client applications. Clearly, combining and ranking search results can only be effectively implemented if at least a basic level of metadata harmonization between the data from different providers is achieved (e.g., mapping the title and year for each object in all sources to a unified metadata field). Thus, both a common metadata model and an approach for automatically transforming metadata from the format in which they were provided to the common model are needed. As EEXCESS aims to also make the long tail of niche content available, many records for a large set of assets will only be infrequently accessed. Thus, transforming all content in advance is not useful, but the transformation happens during the retrieval of resources. Caching would be possible but may be of limited use for the records that are only infrequently accessed and would require creating the centralized infrastructure to store and manage the cached data.

The requirements on metadata models and approaches from the EEXCESS use cases differ partly from those of content provision to cultural heritage portals such as Europeana, most notably in three aspects. First, the metadata comes from a potentially large pool of different providers, so a curated data provision step is not feasible. Second, the metadata of resources is not only automatically enriched from linked data sources but also linked with social media information. Third, the information collected from different providers is selected by taking the user context into account and is presented and processed in a way that is specific for each of the use cases. This means that the model needs to provide high flexibility to cover these diverse requirements as well as support for detailed provenance metadata due to the risk of incompleteness and inconsistency from automatic mapping and enrichment.

3.1.1 *Metadata Model.* Some ofthe models mentioned in Section 1 are candidates for the common metadata model. As EEXCESS deals with data from different domains, models that support cross-provider integration but are limited to a specific domain (e.g., museums only) are not considered. Thus, the CIDOC conceptual reference model (CRM) [ISO 21127 2014] and the Europeana data model (EDM) [Europeana Foundation 2015] are considered. CIDOC CRM is a quite comprehensive model to describe cultural heritage objects and their relations, and to integrate data models of different institutions. EDM was developed as model for data provision to Europeana, and several providers already support this model. Both models supports multiple views to one real-world resource, such as documentation of the object from different sources.

For EEXCESS, EDM was chosen because some metadata of our data providers are already in this format, and it is more basic, and thus it may be easier to define mappings. EDM lacks support for provenance metadata, so we complement it by using the W3C PROV ontology (PROV) [Lebo et al. 2013]. The EEXCESS tools extract metadata from the original resource and use it for enriching their description,

thus adding a new set of annotations. Following the linked data paradigm, EDM also defines a small set of core properties while leaving room to add domain-specific properties from appropriate metadata schemes. The information from all sources is in the EDM description, but it can still be separated by its source if needed. Although this is possible, it is also important to provide more details about the provenance of each of these annotations (if they have been created manually or automatically, their creation and modification dates, the organization creating the annotation, etc.). The core model of EDM does not include provenance information and needs to be complemented by another metadata format. The use of the PROV vocabulary enables describing the different annotations added to an object in more detail. A client application can thus make appropriate use of the available annotations and their provenance metadata.

3.1.2 *Mapping Definition.* As mentioned previously, the tools for metadata transformation must be applied automatically on the fly once such a mapping has been properly configured. We have therefore developed a tool with a focus on easy configurability of mappings, which can then be executed by an automatic service. A mapping can be defined from scratch, or a basic mapping created using PartnerWizard (see Section 3.3) can be used as a starting point for refinement.

Our metadata mapping approach is based on a mapping ontology representing mapping information to solve a mapping problem between a pair of metadata formats. Elements and attributes of the metadata formats involved in the mapping can be linked by using drag and drop. Their data types can be specified (if not defined in an XML schema of the format). In addition, a hierarchy of contexts can be defined to link metadata elements to the appropriate entity, such as to discriminate titles of a series, a book, and an article, which may be found in the source metadata.

All required mapping parameters are derived from this ontology, and mapping instructions are created. The mapping instructions are encoded into an XSL document [Michal 2007], which represents the transformation between an input XML document and an output document. The mapping configuration tool includes the mapping quality approaches described later so that users can obtain direct feedback on the applied mapping using a set of sample files. Further details on the mapping tool and approach can be found in Orgel et al. [2015].

3.1.3 *Quality Assessment.* To ensure the quality of the metadata processed and returned by the EEXCESS system, we need to address both the *source metadata quality* (i.e., the quality of records returned from a particular data provider) and the *mapping quality* (i.e., the completeness and fidelity of the metadata in the target common data model).[9] Metadata quality dimensions to be considered are completeness (the elements are provided and filled), accuracy (no syntactic errors), consistency (correct semantics and no logical errors), availability (referenced resources such as vocabularies can be accessed), and processability (structured and machine readable data). Further details on the implemented metrics and results on datasets from different providers can be found in Orgel et al. [2016].

For assessing the mapping quality of metadata, the dimensions completeness and consistency can be considered. The metadata formats used differ a lot among various data providers and require appropriate mappings. Generally, these mappings may not be lossless. If some source formats are limited in their expressiveness, some loss of information or imprecision in mapping would be unavoidable. The aim of mapping quality assessment is thus to quantify the loss of completeness and consistency of metadata documents resulting from mappings to provide feedback to the experts defining the mapping and to keep this loss as small as possible. Due to the scale of the problem, expert assessment of mappings for a range of formats and a significant number of metadata documents is not feasible. Thus, an

[9]Note that source metadata quality includes quality dimensions that support the findability of records but does not address the retrieval performance for specific queries.

automated method to assess the quality of mappings was implemented and integrated into the tool for configuring mappings described earlier. The method is based on round-trip mappings—that is, testing the mapping from the source format to another format by performing this mapping and mapping back to the source format so that the input and output of this process (both represented in the source format) can be compared. Two different variants of round-trip mappings are considered. The first variant considers only the internal intermediate conceptual representation of metadata properties, whereas the second variant also includes a specific target metadata format. In the first case, we would expect that input and output documents are identical if the mapping is correct and complete, whereas in the second case, the expected loss or imprecision between a pair of formats needs to be specified by an expert once per metadata format in the configuration tool. The mapping quality assessment tools provide feedback to a user defining a mapping and can be run directly in the tool after any modification of the mapping. A more detailed description of the implemented quality assessment methods and results on actual queries can be found in Höffernig et al. [2015].

3.2 Federated Aggregation

The task of the federated aggregation component is to take a user's query as input and create a single result list containing information from all data providers. This task is also called *aggregated vertical search in an uncooperative setting* [Lu and Callan 2005]. Therefore, the user's query needs to be adapted to the specific data providers (i.e., their vocabulary, their query language), and the results from the individual data providers need to be aggregated into a single result list. The three main challenges of this task are the following: (i) the queries themselves may be heterogeneous (e.g., may vary in length), (ii) the data providers' behavior can only be indirectly steered, and (iii) the results returned from the data providers vary greatly in terms of content and metadata. Given the context of the EEXCESS project, the final aggregated result list should contain results that are related to the user's information need, are diverse in nature, and incorporate serendipitous elements. In addition, the whole process should return results in a short period of time to keep the time the user waits for results as short as possible (i.e., the latency should be low).

3.2.1 *Source Selection.* The first part of the federated recommender is the so-called source selection module. Source selection is the task to find the set of matching data providers for any given query [Shokouhi and Si 2011]. This is motivated by the insight that not all data providers are equally suited to handle a user's information need. Therefore, we provides algorithms that try to predict which data providers are particularly well suited for a given user query.

The core of the source selection module is the category mapping function. This function takes a set of terms as input and produces a weighted set of categories. The input might be generated from a user's query or a document, as it is retrieved from one of the data providers. As soon as a data provider joins the EEXCESS system, the content of the data provider is probed. The module synthesizes a set of automatically generated queries, resembling ambiguous user queries, which are then sent to the newly registered data provider. The returned results are analyzed by invoking the category mapping function. After an amount of processed probing queries, a profile for the data provider can be inferred.

When a query is sent to the federated aggregation component, the category mapping function is applied to compute the categories for the query. Next, the profiles of all registered data providers are matched against the query category. If the overlap between a data provider profile and the query category is too low, the data provider is completely excluded. This reduces the necessary resource consumption and helps achieve a lower latency. In addition to the source selection based on categories, a similar technique is also applied on the language of the query, the age span, and other criteria.

3.2.2 *Query Processing.* The federation operates in a so-called uncooperative setting—that is, the data providers are effectively treated as black boxes. The only way to steer the behavior of the individual data providers is via the issued queries. Therefore, each query is rewritten individually for each data provider to achieve optimal results. The federation component already provides a wide variety of different query generation strategies, ranging from disjunction queries to queries featuring data provider–specific syntax elements.

In addition, the queries themselves may be heterogeneous. Therefore, we provide functionality to deal with both short and long queries. For short queries, additional related query terms are added to the original query. This technique is known as query expansion. After thorough evaluation [Ziak and Kern 2015], we settled for a local query expansion technique based on pseudorelevance feedback using an existing knowledge base (KB). Our query expansion technique turned out not only to help in short queries but also was shown to increase the diversity [Rubien et al. 2015].

For long queries, the federated aggregation component foresees techniques to split the query into coherent smaller queries. These smaller queries are individually issued to the data providers, and the results are combined. The most promising approach is based on a combination of query segmentation [Hagen et al. 2012] and the use of DBPedia to detect coherent concepts.

3.2.3 *Result List Aggregation.* Once results from all relevant data providers are retrieved, they need to be combined into a single, ranked result list. At first, available metadata from the results are used to filter out unwanted content (e.g., to remove results in languages not understood by the user). Next, the results are scanned from near duplicates to avoid a situation where a user is confronted with a set of virtually identical results. For each of the data providers' result lists, the overlap with the original query is computed. This allows the ranking of individual data providers according to how well they matched the user's information need. The computed weight is used to control the likelihood of a data provider's result to be added to the final, aggregated result list. This way, a weighted round-robin scheme is implemented. To achieve a degree of diversity within the search result, results generated via our query expansion method can be additionally mixed into the result list. The same technique is optionally applied to trigger a serendipity effect. If available, the query is additionally expanded using the long-term interest of a user. The aggregated result list is finally passed to the calling component.

3.2.4 *Runtime Behavior.* To assess the scalability of the federation, one first needs to identify the main bottlenecks. Therefore, we conducted several experiments to measure the runtime behavior of the federation component and its subcomponents. In this experiment, we always submitted the same number of queries but with different degrees of parallelism. In Figure 3, an overview is given for a range of increasingly more parallel calls to the federation component. As expected, the total runtime decreases with more requests being conducted in parallel. At about 2,000 parallel calls, some of the queries took longer than a predefined threshold, and thus the respective data provider is ignored in the returned result list. Overall, one can state that the federation component itself is not the main bottleneck and instead the runtime behavior is dominated by the individual data providers.

3.3 Data Provider Setup

For the success of EEXCESS, the number and diversity of data providers is an important criteria. Many organizations provide an API to make their data available but do not have the technical expertise in-house to implement the necessary software components to connect to the EEXCESS system. We have thus developed a tool called *PartnerWizard*, which automatically creates the recommender component for a data provider that connects to the federation component and thus integrates with the EEXCESS framework. PartnerWizard guides the user through a few easy and interactive configuration steps. The

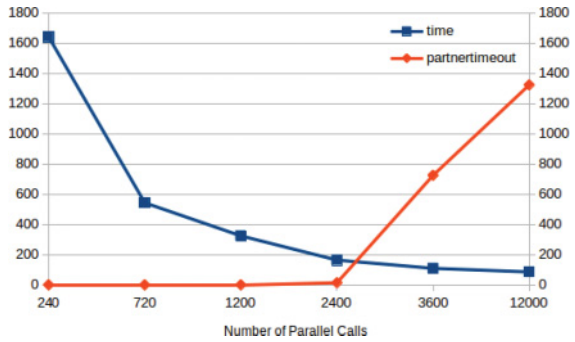

Fig. 3. Scalability behavior of the federation with three partners (the right axis shows the number of time-outs, and the left axis shows the total time in seconds). At about 2,000 parallel calls, the first time-outs occur. Overall, the scalability of the whole system is dominated by the latency of the data providers.

user specifies the URL of the endpoint of the data provider's API, as well as the XPath of the root of a record and to basic fields needed for metadata mapping. All parameters can be immediately tested with sample queries to see how changes to the configuration impact the handling of an actual response from the data provider's system.

Recommender query generation is a further step in PartnerWizard, which enables the user to optimize the search results for the data provider. This is also a Web-based process, where different types of query processing strategies with multiple search terms and different strategies for combining search results are compared. The user selects the preferred result for some example queries, which then adjusts the source code in the recommender accordingly.

PartnerWizard was implemented using Apache Maven.[10] In particular, the archetype feature of Maven is used to create a template that is turned into the source code of the recommender for the data provider. As Maven fetches all required dependencies, the ready-to-use Java classes can be built automatically. Along with the implementation, unit tests are generated from the example data entered by the user during the configuration process.

3.4 User Interface Clients

The user interface clients component is the conceptual end-user interface, which can be instantiated by various applications. Several instantiations were already implemented within EEXCESS, such as the Google Chrome extension (see Section 4.3). However, we do not aim to develop a client for each and every use case. Instead, we provide a modular platform, which allows the quick creation of new and easy integration into existing client solutions. This platform consists of two major parts: modules for context detection and query extraction (see Section 3.4.2) and widgets to display results and interact with them (see Section 3.4.1).

3.4.1 *User Interface Widgets.* The widgets to display and interact with the results are self-contained Web pages, which can be easily included via an iframe. With this architecture, developers do not need to care about configuring the widgets, but only need to implement the interface to communicate with them. The communication takes places via the Web Messaging API.[11] Most importantly, a developer needs to implement the interface to pass retrieved results to the widget. As the interface is consistent across all widgets, they are easily interchangeable once the interface is implemented.

[10]https://maven.apache.org/.
[11]http://www.w3.org/TR/webmessaging.

Fig. 4. Screenshot of the filter area of FacetScape. Selection filters are applied for English and Swedish in the language facet.

In addition to a basic search result view, as depicted in the screenshot shown later in Figure 7, more advanced tools and alternative result views are available. Among them are the visualization dashboard [Tschinkel et al. 2015], providing several views on and filter possibilities of the results, and FacetScape [Seifert et al. 2014], a widget for result space exploration. We briefly describe the latter as an example for an alternative or additional widget to the basic search result view. Figure 4 shows a screenshot of the filter area of FacetScape for a particular result set. In the screenshot, Swedish and English have been chosen in the *LANGUAGE* facet, hence only results in those languages are displayed. The selection has an impact on the available selection filters in other facets. Filters that are not available in the current selection are displayed with less opacity. For example, *Swissbib* in the *PROVIDER* facet is grayed out, as no results in English or Swedish are available from this provider, as opposed to *Europeana*, for which 23 results are available (indicated by the superscript number). When hovering over a selection filter, the amount of results that match this selection in the current set is displayed along with the total amount of results in the current set and the amount of results that would be removed or added by this filter. This situation is depicted for the *IMAGE* filter in the *MEDIATYPE* facet in the screenshot: 18 results in the currently selected set of 38 results are images, and when applying this filter, 20 results would be removed from the current set.

3.4.2 *User Context Detection and Query Generation.* Detecting the user's current context and generating a query out of it are the first steps on the way to retrieve relevant results. We provide modules for these steps, packaged in a Bower[12] repository called *C4* and adhering to asynchronous module definition.[13] This way, any desired modules can be loaded on demand. For example, in the Google Docs add-on (see Section 4.4), the context is defined by the editor's typed text and the context detection module does not apply.

3.4.2.1 *Context Detection*. In a Web setting, the observable user context encompasses the Web pages visited and additional information like the user's location. We focus on the visited Web pages and hence account for the textual content of Web documents in the context detection for query generation. Information such as the user's location can be used to populate the user profile, yet the user is in control of which information of this profile is disclosed (see Section 3.5). The textual context of visited

[12]http://bower.io.
[13]https://github.com/amdjs/amdjs-api/blob/master/AMD.md.

Web pages can be subdivided into five levels of granularity (from fine-grain to coarse): *terms*, *phrases*, *paragraphs*, *pages*, and *sessions* [Schlötterer 2015]. To generate queries and present results according to the current context, we focus on the paragraph and phrase level. The session level is utilized to personalize the generated queries (see Section 3.4.2.3). Browser events such as mouse movements or scroll position yield only limited accuracy in determining the phrase currently read by the user [Hauger et al. 2011]. Hence, we rely on explicit user interaction on the phrase level (i.e. a text selection), which is a strong indicator for reading focus [Hauger et al. 2011]. Treating terms as single-term phrases, the term level is also covered by the phrase-level context detection (which would also cover the page level when a user selects the whole page content, but this is unlikely to happen). On the paragraph level, we first separate actual text passages from navigational menus, advertisements, and so forth, then afterward identify the currently focused paragraph. To minimize the computational effort due to limited resources, we apply a heuristic based on a fixed-length threshold of DOM text nodes for the extraction of paragraphs. To determine the focused paragraph, we first limit the set of candidate paragraphs to those currently in the viewport. Afterward, the focused paragraph is determined by scroll position and highlighted. If the user does not agree with this selection, she can change it with a simple click on another paragraph. This paragraph will then be marked as focused until it leaves the viewport (due to scrolling).

3.4.2.2 *Query Generation*. We cannot influence the behavior of different providers' search engines integrated via the federation component. Therefore, we treat the search engines as black boxes and focus on the query side of retrieval. Since the context in our setting is defined by natural language text (e.g., a paragraph on the Web site), keywords provide a compact representation of the paragraph and hence make up natural candidates to construct a query for related results. Keyword extraction algorithms with reasonable performance are readily available [Mihalcea and Tarau 2004; Rose et al. 2010]. However, greater than 71% of (user-generated) search queries contain named entities [Guo et al. 2009], and Wikipedia page titles (i.e., named entities) have been shown to be beneficial to query segmentation [Hagen et al. 2012], a task for query optimization. In addition, named entity extraction can be seen as some kind of keyword extraction task, as the original text is represented by a smaller set of terms. For these reasons, we base our query construction approach [Schlötterer et al. 2016] on named entities.

In pretests with different digital library repositories, we discovered that repositories that perform the search over metadata of repository objects return a very broad range of results. Moreover, we observed that results related to only one keyword in the query suppressed results related to more keywords in the query. We assume that in those repositories, the keywords are combined via Boolean OR queries. The problem hereby is that in most cases, the results, which were triggered by a single keyword only, did not fit the main topic of the text from which the query was constructed very well and hence were quite unrelated. To overcome this problem, we aimed for a solution that incorporates the overall topic and ensures it is well represented in the query, which resulted in a Boolean query of the following form:

("main topic") AND ("keyword 1" OR "keyword 2" OR . . .)

(Named) entities located within a paragraph form the base for an underlying query. To extract these entities from a textual document, we perform a (named) entity annotation (NEA) that relies on two important subtasks: (named) entity recognition and (named) entity disambiguation. Entity recognition forms the first step of creating entity annotations. It identifies proper nouns (in the following denoted as surface forms) that can be linked to a semantic meaning. The task of entity disambiguation establishes links between identified surface forms and entities within a KB and faces the problem of semantic ambiguity [Zwicklbauer et al. 2016a, 2016b].

To provide a robust NEA in terms of reliability and performance, we apply the named entity recognition and named entity disambiguation system DBpedia Spotlight.[14] DBpedia Spotlight is one of the first semantic approaches (2011) and constitutes an entity-centric approach that is based on DBpedia. And based on a vector-space representation of entities and using the cosine similarity, this approach has a publicly available Web service. The service is able to recognize and disambiguate English and German language entities as determined in the request. Furthermore, we detect dates in documents and treat them like normal entities.

To determine the main topic of a paragraph, we use Doc2Vec [Le and Mikolov 2014] as a topic detection method. Generally based on Word2Vec, Doc2Vec produces a vector given a sentence or document. Hence, we use the entire input paragraph and infer a representative vector given a Doc2Vec model created on the Wikipedia corpus. We compare this vector with the vectors of the Wikipedia pages (entities) by computing the cosine similarity. The Wikipedia page (entity) with the highest similarity to the input paragraph represents the main topic. To significantly improve performance, we reduce the target entity set to those entities that have been annotated in the given paragraph.

3.4.2.3 *Query Adaption and Personalization*. The automatically generated queries are also displayed to the user and can be modified by the user. Possible modifications include filtering the keywords for locations or persons, adding or deleting keywords, and editing the main topic. Keywords can be added either manually via textual input or via a text selection within the page. Similarly, the main topic can be set via a text selection within the page or via drag and drop of a keyword (manual editing is also possible).

In particular, large paragraphs lead to a large amount of extracted keywords (named entities). To counter this fact, we subdivide large paragraphs into smaller subparagraphs and generate separate queries for each subparagraph. The subquery with the highest overlap with the user profile is selected as the query to be sent to the federated aggregation component. We took this approach because a subparagraph usually covers a particular aspect of the topic of the whole paragraph, and by the highest overlap with the user profile, we deem this aspect to be the most interesting to the user. The user profile overlap is based on categories associated to the entities in the current and previously executed queries. The associated categories are provided by the category assignments in Wikipedia. For example (among others), the category *Women in technology* is assigned to the entity *Ada Lovelace*. The categories of all previously executed queries for which the results have been viewed are stored in the user profile.

3.4.2.4 *Evaluation and Results*. The performance of context detection and query generation was evaluated in a user study with 77 participants. Each of the participants had to perform four tasks. In the first three tasks, users were free to choose a Wikipedia page from a predefined set of featured articles, whereas in the last task, the page was predefined and the same for all. Apart from this, the procedure was the same in every task. First, participants were instructed to navigate to a particular section within this page. Second, they had to check whether the paragraph identified as the currently focused paragraph by our context detection mechanism was correct. In case of an incorrect identification, they had to change the focused paragraph either by clicking on another extracted paragraph or by selecting the relevant piece of text. The first option applied when the extraction of paragraphs was correct but the identification of the focused paragraph failed. The second option applied when the extraction failed already. After a potential correction of the focused paragraph, a query was generated automatically and participants had to provide relevance feedback for the results of this query. Then they were instructed to adapt the query and rate the results of the modified query until one of the following criteria was met: the participant was fully satisfied with the results, the participant did not

[14]https://github.com/dbpedia-spotlight/dbpedia-spotlight/wiki.

Table II. Evaluation Results of the User Study for Context Detection and Query Generation

	Precision	Recall	F1-Score	Extracted Paragraph Fit	Focused Paragraph Fit	Main Topic Fit
Automatic Query	0.29	0.25	0.24	84%	65%	83%
User Query	0.33	0.33	0.31			

believe that the search engine was able to deliver any results on the topic, or the time was up (around 8 minutes per task).

The evaluation was carried out on a cleaned dataset (e.g., queries that could not doubtlessly be assigned to a task were removed), resulting in 558 queries executed by 69 users in 228 tasks and 6,985 relevance ratings for the results. The evaluation results are depicted in Table II.

The reported precision, recall, and F1-scores are macro averaged over all queries (hence, the F1-score is below the value obtained from the average precision and recall scores). Clearly, we cannot measure the true recall value, as we do not have ground truth relevance feedback for the whole collection. Instead, we approximate the recall with all positive results retrieved via all queries executed in the context of a particular Wikipedia page. For the evaluation of user queries, we took the best query a user was able to formulate for a paragraph. This means that if the initial automatic query scored better than all subsequent modifications by the user, we took the initial query. As can be seen from the table, performance generally is quite low, with the user-generated queries performing slightly better than the automatic queries.

Participants modified 16% of the extracted paragraphs—that is, 84% of the paragraphs were extracted correctly or meaningful from a user perspective. From the correctly extracted paragraphs, the focused paragraph was identified correctly in 65% of the cases. These results are quite promising, particularly since they are based on a simple heuristic. However, they are only valid for Wikipedia pages, and we cannot draw any conclusions for other Web pages.

We consider the chosen main topic as suitable when it is not changed by the user or a change does not lead to an improvement. In the main topic evaluation, we removed all cases where we could make a definitive statement about the reasons the main topic was not changed. This comprises a task in which no stopping criterion was provided, or the task was stopped due to time limitations. According to this evaluation, the main topic was appropriately chosen in 83% of the automatic queries.

3.4.2.5 *Outlook*. Given the promising results of paragraph extraction and detection on Wikipedia, we plan to evaluate the approach via a crowd-sourcing experiment on a larger variety of Web pages and integrate further improvements. Even though the main topic choice is suitable with a high accuracy already, there is still room for improvement. In particular, we discovered that the suggested main topic is less subject to modifications when it is identical to the topic of the page. Moreover, in 82% of the queries where a main topic modification resulted in an improvement, the originally suggested main topic was different from the page topic. These findings suggest that the main topic extraction should be based on the whole page instead of the focused paragraph. Regarding the keywords (named entities in addition to the main topic) of the query generation process, we are researching optimizations to filter the extracted keywords to improve query quality and add additional keywords in addition to the named entities.

3.5 Privacy Preservation

Sending users' personal information to the federated aggregation component raises serious privacy issues. For that reason, we introduce mechanisms to help users control the privacy preservation of their data. First, we design a user interface to enable users to define their own privacy policy (i.e., which information they agree to share with the federation component). Second, we design a

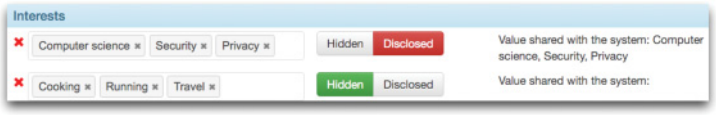

Fig. 5. Cutout of the user interface for defining the privacy settings.

privacy-preserving protocol between the client and the federation component to ensure that users' queries cannot be used to profile any users.

3.5.1 *Privacy Policy.* A user can explicitly specify which information from her user profile she agrees to pass to the system. Indeed, when a query is issued, it is expanded with the context of the user contained in her user profile (e.g., age, language, interests). Consequently, if the user chooses to not disclose specific information, this information is not added to the query. To avoid confusion for nonexpert users, we created a simple interface as depicted in Figure 5. For each attribute contained in the user profile (e.g., age, language), a user specifies if she wants to hide or disclose the information to the system. Other attributes enable a more precise configuration (e.g., for location, the user can choose to share her country, her city, or nothing). By default, all attributes are not disclosed.

3.5.2 *Query Protection.* The privacy-preserving protocol between users and the federated aggregation component consists of hiding the user identity and masking the user query. A user query contains two pieces of information: her identity (via her IP address) and her interests (via the content of the query). To successfully protect user privacy, these two data need to be separated. This is the role of the privacy proxy. This entity is composed of two servers: the receiver and the issuer. The receiver knows the identity of the requester (with her IP address) without learning her interests, whereas the issuer accesses queries without knowing their provenance.

The privacy-preserving protocol works as follows. The receiver receives a message from the user, which contains an encrypted request (which can only be decrypted by the issuer). The receiver then forwards this message to the issuer. The issuer deciphers it and forwards the query to the federated aggregation component. Upon receiving a response, the issuer ciphers the results and forwards them to the receiver. Finally, the receiver forwards the message to the user, who retrieves the results by deciphering the message.

Nevertheless, hiding the user identity is not enough to correctly protect users (as was shown in Peddinti and Saxena [2014] and Petit et al. [2016]). That is why we reinforce the user protection by masking the query with multiple fake queries. A message is then composed of the original query and k fake queries. Generating plausible fake queries is a difficult task. We perform this operation by reusing queries already sent to the system. This is made possible by the issuer, which aggregates all queries it receives in a group profile. Queries are aggregated such that publishing the group profile does not leak information about individuals (i.e., publishing real past queries) but contains enough data to estimate real queries. As a consequence, fake queries generated by the Chrome extension seem realistic, as they are semantically coherent and their topics fit with the users in the system. However, sending fake queries has an impact on accuracy. The federation aggregation component answers results about all queries (and not only the original one). For that reason, we also implement a filtering algorithm to only retrieve results corresponding to the initial request. Finally, the impact on the accuracy is relatively small, as for 95% of queries more than 80% of the expected results are correctly returned. Further details about the privacy-preserving protocol and the obfuscation algorithm are accessible in Petit et al. [2015].

4. EXAMPLE APPLICATIONS

In this section. we present examples of how the presented infrastructure and components can be used to distribute cultural content to interested users and how content providers can be connected to the services. Section 4.1 describes the process for integrating cultural objects from small museums in Switzerland. The section also describes an internal process and approach to establish multilateral readiness for the technical integration of the data. A fully automatic integration using the PartnerWizard component is described in Section 4.2 for content providers with available search API. Two content-related processes can be differentiated on the Web, namely *content consumption* (e.g., reading online news, researching information) and *content creation* (e.g., authoring Web pages or creating blog entries) [Granitzer et al. 2013]. We present two client applications supporting these processes: a Chrome extension for content consumption in Section 4.3 and an add-on for Google Docs for content creation in Section 4.4.

4.1 Groundwork in a Cultural Heritage Institution

The following section describes the content-related work conducted in both a mid-size and in about 20 small-size museums on their journey to the World Wide Web. The state museum of the canton Baselland (AMBL) curates collections in the fields of natural history, ethnology, industrial history, archeology, arts, and historical photography. Out of a total of around 2 million objects, records of about 400,000 objects have been migrated into a single system. Furthermore, the state museum was leading a regional project (KIM.bl)[15] with the goal to establish a network among 45 small and large museums to harmonize and centralize their inventories using a central collection management system. The challenge to aggregate and exploit the heterogeneous mass of objects by the provision of one API has been met with the setup of a project that included technical and content-related subtracks. Figure 6 illustrates both the technical and the content-related tracks on a high level.

4.1.1 *Integration Process.* The process of content integration was performed in four stages: (i) target format definition, (ii) collection of metadata requirements, (iii) collection qualification and mitigation activities, and (iv) iterative quality checks. In the following, these four stages will be described in more detail. Lessons learned during the process will be outlined afterward.

Based on discussions with many partner institutions from a network across Europe, the decision had been taken to adhere to the LIDO XML harvesting schema as the target format because this schema is intended for the delivery of metadata for use in various online services. In addition, the strength of LIDO lies in its ability to support the full range of descriptive information on museum objects across all collection domains, such as art, architecture, cultural history, history of technology, and natural history.

In a next step, a list of metadata requirements was derived from user scenarios and best practices. The scenarios tailored for meeting the target channels in focus: the Museum Portal and EEXCESS. The analysis of the requirements translated into a list of necessary metadata fields that in turn were compared to the LIDO schema. All information had been aggregated into a flat "master metadata matrix," which served as our lookup table to track progress during the whole project.

Subsequently, every collection in scope was screened and assessed for compliance with the matrix. Missing pieces were marked and translated into mitigation activities. Individual quality control activities per collection were identified and labeled with a timeline. The activities ranged from simple to complex IT-driven data migration and completion routines. In other cases, profound discussions with subject matter experts were conducted and resulted in labor-intensive handwork. Where

[15]http://www.kgportal.ch.

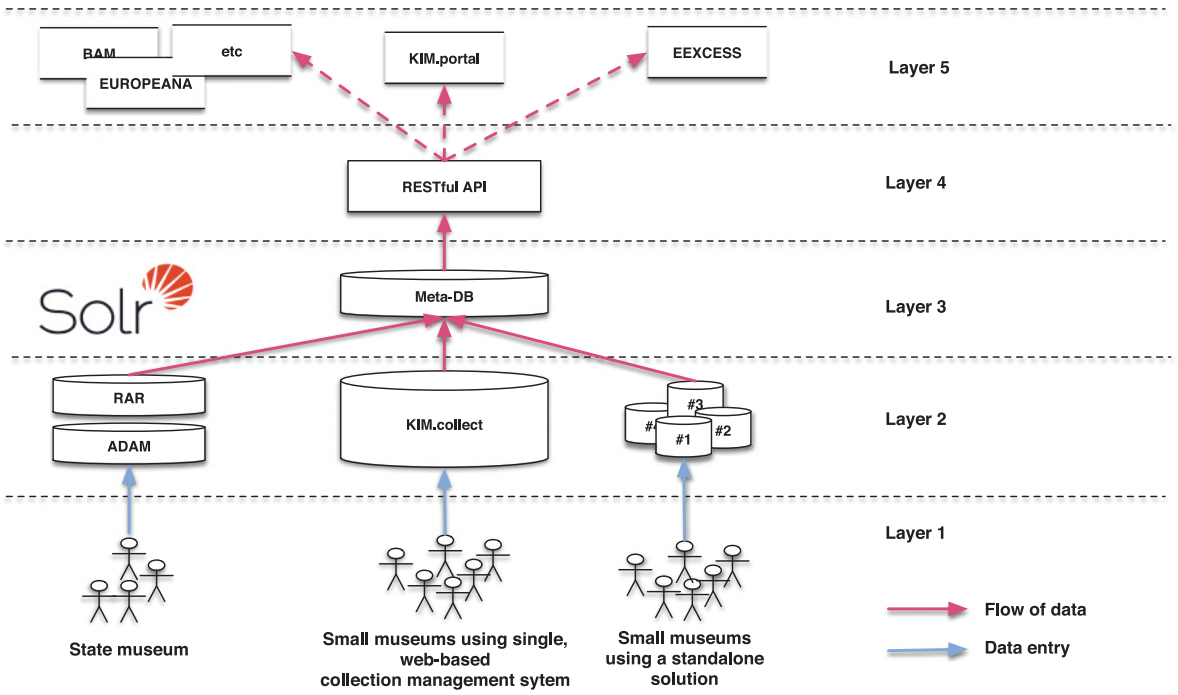

Fig. 6. Overview of the process for integrating the KIM.bl collection into EEXCESS. Curators in the museums manage their collections in their own museum management systems (Layer 1 = Data entry). The systems (Layer 2) provide their data (metadata and media files) using a dedicated export and import routine into the central metadata layer (Solr Server; Layer 3). A restful API on top of Layer 3, Layer 4, enables querying by different exploitation channels (Layer 5, e.g., EEXCESS, portal, Europeana) using XSLT routines.

necessary, external staff had to be hired to complete and harmonize the metadata. A separate track included the completion of photographs.

Collections that qualified for testing purposes were migrated to the test environment. In joint collaboration with the partner teams from the project, the results were continuously assessed. That way, missing pieces and gaps could be approached early in the project.

4.1.2 *Summary and Lessons Learned.* While implementing the outlined process, three main aspects turned out to be a persisting challenge: (i) engagement and awareness by curators, (ii) rights management and licensing, and (iii) technical implementation. To complete and harmonize the metadata per collection, the engagement of subject matter experts is crucial. The establishment of a continuous dialogue with the experts ensures that they understand the goals of the projects and contribute valuable input. The provision of metadata to a publicly available channel like EEXCESS asks for careful considerations different from the established internal curators' work (from unstructured to structured information, transformation of time and periods, structured geoinformation, introduction of controlled vocabularies, etc.). With regard to the workload for screening and complete metadata, these tasks are very labor intensive and often cannot be processed using routines. The involvement of specialized and skilled staff needs to be included in the calculation (time and budget). Going public requires a broad

discussion on rights management and licensing. The digital turn in GLAMs asks for a digital strategy of the institution. More information can be found on the Web page of the open GLAM initiative.[16]

In addition to these content-related tasks, it is important to state that a cultural heritage institution needs to have technical experts at its disposal. A museum management system typically does not feature the functionalities necessary to go online. A robust and flexible IT solution must be envisaged, realized, and maintained to interface with a project like EEXCESS. This issue was emphasized in a white paper targeted at potential new content providers.[17]

4.2 EEXCESS Provider Integration Wizard

The manual process described previously ensures the highest quality of data contribution to the EEXCESS framework. However, several steps in this process can be supported by (semi)automatic tools. This is especially useful in cases where the required technical expertise is not available, which is often the case in smaller institutions. The following three components facilitate the data provider integration process.

PartnerWizard, introduced in Section 3.3, enables the basic integration of a data provider—that is, ensuring that queries are correctly fed into their API, defining mapping of basic metadata fields, and optimizing the way queries are constructed. All configuration is done in a Web application and does not require detailed technical knowledge or programming skills. PartnerWizard was developed based on manually integrated data providers (as described earlier), and the result of automatic integration was validated against those from manual integration. In addition, eight data providers have so far been integrated with PartnerWizard as a starting point. Although PartnerWizard is able to connect the API and provide basic mapping, it cannot leverage the full potential of the provided metadata if mappings require expert knowledge or adjustments to the data provided by the API are necessary.

The mapping created by PartnerWizard covers the basic fields required by the EEXCESS framework. The APIs of many data providers return additional metadata, often in specific structures, that can improve the findability of assets. The mapping configuration tool described in Section 3.1.2 provides the functionality to refine and extend the mapping for a data provider. Using this tool requires more skills than PartnerWizard, particularly concerning the metadata model of the data provider. The tool can provide previews of transformed metadata records so that the impact of changes to the mapping can be directly observed. An updated mapping can then be used by the recommender component for the respective data provider.

The quality assessment tools described in Section 3.1.3 run as a background process in the system and provide valuable information for the data provider concerning the source metadata quality (i.e., the data returned from the provider's API) and the mapping to the EEXCESS model (i.e., the quality of the mapping definition). The assessment results can be used to improve the quality of the data and thus the quality of the recommendations for the consumer. In some cases, even simple adjustments may have large impact on the quality (e.g., choice of default values, omitting empty fields, mapping additional fields). Addressing some recommendations may require changes to the data provider's search API but has the potential to not only improve the quality of the data provided to EEXCESS but also to other consumers of the API.

4.3 Client Application for Content Consumption

In the previous two sections, we discussed possibilities for integrating new data sources into the presented infrastructure, either manually or automatically. In this and the next section, we present

[16]http://openglam.org/.
[17]http://eexcess.eu/content-providers/.

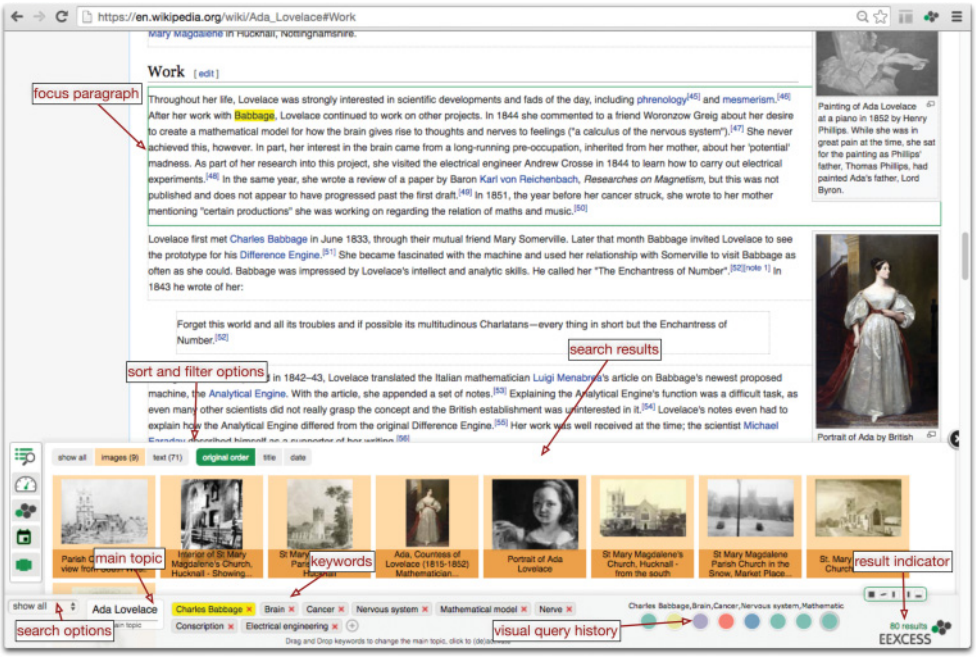

Fig. 7. Screenshot of the Chrome extension. When hovering over a keyword at the bottom, the respective term in the focus paragraph is highlighted, indicating the keyword context (yellow highlights).

example applications for distributing content to interested users for consumption purposes (e.g., using related cultural resources as information consumption) and for content creation purposes (e.g., integrating cultural resources into newly created content).

For consuming cultural content related to the current (textual) user context, we developed an extension for the Chrome browser. The Chrome extension allows one to access cultural content that is relevant to the current Web page. While browsing the Web or reading news articles or blog entries, additional resources available in the content provider's databases are automatically retrieved and suggested for further investigation.

When installed and activated,[18] the extension is visible as a light gray bar at the bottom of every Web page, as shown in Figure 7. The paragraph automatically detected as active is outlined in green within the Web page. The extracted keywords are visible at the bottom of the page, together with the main topic of the paragraph. Additional search tools, such as a search for persons only, are available (bottom left). The main topic and the keywords are automatically sent as a query to the federated aggregation component, and the number of retrieved results is indicated at the bottom right. Keywords can be manually added, modified, or deleted. A keyword can become the main topic by simply dragging it on the main topic field. When the result indicator is clicked, the search result list becomes visible. The search result view can be moved and changed in size. Search results can be sorted and filtered according to different metadata values (e.g., the media type of the result). Additionally, the search history is visualized (bottom right) and can be used to revisit previous searches.

To collect feedback about the general usability from a broader audience, we generated a survey using the SUS scale [Brooke 1996]. Answers were collected on a 5-point Likert scale (1 = strongly

[18]https://purl.org/eexcess/clients/chrome-extension.

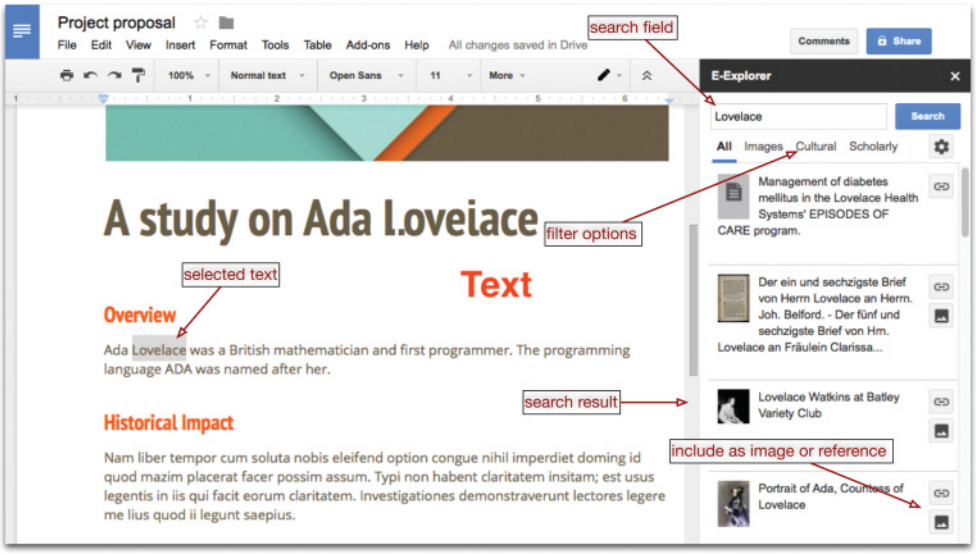

Fig. 8. Screenshot of the Google Docs plugin.

agree and 5 = strongly disagree). An open question was added to the standard questionnaire asking for comments and possible improvements. The survey was prepared in English and German and was sent to project-internal mailing lists with a potential reach of 1,231 people. Twenty-six responses were received between November 24, 2015 and December 31, 2015. The overall SUS score was 65.6, which is slightly below the average score of 68. The received qualitative feedback can be summarized as follows. The relevance of the results should be improved and certain parts (i.e., some visualizations included in the extension but not described in this article) of the user interface were too complex. Further, users were disappointed because they received only responses from specialized providers. Note that this evaluation only considers usability issues of the current system with a fixed number of providers. We have not yet tested how user satisfaction changes with the number of providers.

4.4 Client Application for Content Creation

In the content creation scenario, additional cultural resources from the content providers are retrieved while users are creating their own (textual) content. To showcase this scenario, we implemented an add-on for the online word processor Google Docs.[19] The core idea of the Google Docs add-on is to have a collaborative way of writing documents and—while writing—a simple way to include related images and citations. Real-time online collaboration is available in Google Docs per se; the EEXCESS add-on extends Google Docs with cultural resources that can be inserted with one click.

The EEXCESS Google Docs add-on can be downloaded from the App-Store[20] and needs to be started for each document separately from the menu *Add-ons* → *E-Explorer* → *Start*. When activated, the add-on is visible on the right side of the currently edited document (Figure 8). To get results, a phrase in the document can be highlighted (either by double clicking or keyboard selection) and is automatically added to the search field. The add-on searches all connected content providers via the federation component and displays the results as a list on the sidebar. Details of the results can be accessed by

[19]https://docs.google.com.
[20]https://purl.org/eexcess/clients/googledocs-plugin.

clicking on the list entry. This brings the user to a page with more detailed information about the digital object. Resources can be included in the document either as a reference or an image (two icons on the right side of a result in Figure 8). The add-on was evaluated with 25 students of a seminar at the Munich School of Philosophy using a questionnaire. The questionnaire asked about general usability, quality, and variety of results and the perceived level of privacy. Students reported an overall good usability and overall relevance—for example, "[..] gives the user a bigger freedom of choice due to its ability to connect various articles regarding a certain topic." With respect to privacy, students reported differing viewpoints: "Compared to regular search services via sites like Google, I would say that EEXCESS is a possibility to find information from certain sources you trust, insofar as you can choose which providers you want to see" and "Where is my personal data stored? How can I access it? Can I access it?"

In summary, the Google Docs add-on is a usable tool providing access to cultural resources when collaboratively writing documents. Although the overall infrastructure ensures user privacy, it is not transparent what other stakeholders (in this case Google) can infer about the user from the usage of the add-on.

5. RELATED WORK AND DISCUSSION

In this section, we discuss works, projects, and initiatives related to the presented infrastructure and single components with respect to (i) federated search and recommendation, (ii) metadata mapping, (iii) metadata quality assessment, (iv) privacy preservation, and (v) just-in-time retrieval.

A closely related initiative to the presented infrastructure is Europeana,[21] the largest aggregator for cultural content in Europe providing access to approximately 44 million objects (beginning in 2016). Although Europeana focuses on content aggregation, EEXCESS also provides content dissemination components and relates cultural content with scientific content by integrating scientific data repositories as data providers. EEXCESS includes Europeana as a data source and provides access to approximately 130 million objects, including scientific content.

5.1 Federated Search and Recommender Systems

Combining multiple search results into a single consolidated search result is often referred to as metasearch. Typically, the focus of such techniques has been on Web search, with the goal to improve quality and reduce unwanted content [Dwork et al. 2001], where results of multiple Web search engines are combined. In such a scenario, the results are expected to be homogeneous, sharing the same characteristics and metadata. If more heterogeneous results are to be combined, the term *vertical search* is commonly used, particularly for systems that display multiple search results side by side. For example, image and text search are both triggered to react on a specific information need. The term *aggregated search* is used to indicate that multiple search results are combined into a single one [Murdock and Lalmas 2008; Kopliku et al. 2014].

According to the literature [Shokouhi and Si 2011; Lu and Callan 2005], the three main challenges that aggregated search faces are (i) the selection of the appropriate sources, (ii) the so-called collection representation problem (i.e. inferring the key characteristics of a KB while keeping the effort minimal), and (iii) the aggregation of the results returned by the different sources.

An important aspect in an uncooperative aggregated setting (where the underlying search engines cannot be modified) is the query processing aspect. Here, problems may arise due to short queries, which have to be expanded with related terms for better results [Montgomery et al. 2004]. At the

[21]http://europeana.eu.

opposite end of the spectrum are queries consisting of many terms with distinct concepts, that need to be split into segments, which themselves are coherent [Hagen et al. 2012].

There is a strong connection between recommender systems [Ricci et al. 2011] and information retrieval. One type of recommender systems is the so-called content-based recommender systems [Lops et al. 2011]. The key aspect is here that the recommendations (i.e., search results) are provided without users having to explicitly state their information need.

5.2 Metadata Mapping and Quality Assessment

There are several initiatives in the cultural heritage domain to provide services for mapping between library and museum metadata formats (e.g., the OCLC crosswalk service[22]). MINT[23] is a recent framework facilitating metadata mapping and aggregation of cultural heritage information from heterogeneous sources and performing metadata mapping. This framework was used in the Athena, Carare, and EUscreen projects (all completed), as well as in recent projects such as LoCloud, AthenaPlus, and EUscreenXL to prepare cultural heritage metadata for ingestion into Europeana. The PrestoPRIME project[24] has developed a metadata mapping service for audiovisual metadata [Höffernig et al. 2010], including support for the EDM. Most of these approaches are designed for the interactive definition of metadata mappings, which is the typical scenario when ingesting data into a cultural heritage portal. Such a workflow is not feasible in EEXCESS, as metadata transformation needs to be done on the fly.

Most of the existing literature on metadata quality considers the metadata of single or multiple records of a collection, such as our source metadata. However, as EEXCESS accesses metadata from a range of different sources, there is no single application profile that can be checked against. Bruce and Hillmann [2004] defined the following measures for quality: completeness, accuracy, provenance, conformance, logical consistency and coherence, timeliness, and accessibility. A taxonomy of 22 measures for information quality was proposed by Stvilia et al. [2007], grouped into three categories: intrinsic, relational/contextual, and reputational information quality. An approach that attempts automation was proposed by Bellini and Nesi [2013]. The authors start from viewing metadata quality as the fitness for use for a specific purpose and propose three metrics: completeness, accuracy, and consistency. Debattista et al. [2014] proposed an extensible framework for assessing quality of linked open data called *Luzzu*. One important contribution of their work is a data quality ontology (daQ), which is also used an input to W3Cs recent work on this topic. Trippel et al. [2014] (and similarly Reiche et al. [2014]) proposed a quality assessment framework using similar criteria as earlier works, but they calculated a single score over all of these criteria. The work of Gavrilis et al. [2015] proposed an assessment framework called *MQEM* but included a set of concrete metrics that yield numeric values. Another aspect of metadata quality is the use of controlled vocabularies (e.g., using the eight classes of criteria proposed in Dröge [2012]).

5.3 Privacy Preservation for Search

The main solutions to query a content provider or a recommender system in a privacy-preserving way can be classified into two categories: (i) protocols to ensure *unlinkability* between requesters and their queries and (ii) systems guaranteeing *indistinguishability* of user interests. Unlinkability solutions hide a user's identity to prevent the distant server from identifying and analyzing queries issued by a single user. Basic techniques consist of sending queries through a proxy [Shapiro 1986] or a VPN server [Seid and Lespagnol 1998]. Unfortunately, these mechanisms only shift the privacy problem

[22]http://www.oclc.org/research/activities/xwalk.html.
[23]http://mint.image.ece.ntua.gr/redmine/projects/mint/wiki/Wiki.
[24]http://www.prestoprime.org/.

from the search engine to the relay (which is able to extract a user's interests). To solve this issue, anonymous protocols (e.g., Goldschlag et al. [1999], Dingledine et al. [2004], and Ben Mokhtar et al. [2013]) ensure, with cryptographic primitives, that all third parties in the system cannot access both the user identity and the content of her query. However, most of these solutions rely on a heavy cryptographic protocol orall-to-all communication that makes them impractical for querying a recommender system. Castellà-Roca et al. [2009] and Lindell and Waisbard [2010] proposed a fully decentralized architecture in which users exchanged their queries and sent them on behalf of each other. But similarly to anonymous protocols, this solution produces significant overhead (more network traffic and higher latency).

Indistinguishability solutions consist of making all analysis of user queries inaccurate. For instance, TrackMeNot [Toubiana et al. 2011] periodically sends fake queries to obfuscate the user profile created by the distant server. In Murugesan and Clifton [2009] and Domingo-Ferrer et al. [2009], the user query is directly obfuscated by sending the initial query with k extra fake queries. As a result, the distant servers cannot distinguish which query (among the $k + 1$) is the correct one. Furthermore, Query Scrambler (presented in Arampatzis et al. [2013]) protects users by generating and issuing queries related to the initial query. As the initial query is not sent, Query Scrambler reduces the leak of information, but the accuracy of the recommendation is highly impacted. A recent approach, Dispa [Juarez and Torra 2015], reconciles personalization with user protection. It consists of making the distant server constructing multiple user profiles for a single user. Each user profile contains partial but accurate information about the user. As the distant server does not know that all small user profiles are related to the same user, this method reduces the amount of disclosed information. However, this solution relies on the strong assumption that the distant server only uses cookies to identify users (and not other elements, e.g., the IP address or the HTTP header).

Several privacy attacks [Peddinti and Saxena 2014; Gervais et al. 2014; Petit et al. 2016] against unlinkability and indistinguishability solutions have been published in the literature. However, most of these privacy solutions are not robust against such attacks, and therefore we proposed a new approach to unlinkability and indistinguishability.

5.4 Just-in-Time Retrieval

Proactive retrieval of resources relevant to the current user context and unobtrusive presentation of the retrieved results was first made popular by Rhodes as just-in-time retrieval [Rhodes 2000]. Rhodes' research was continued under the topic of zero effort queries [Allan et al. 2012] with special emphasis on mobile applications [Lee and Sumiya 2009]. Zero effort queries require minimal and ideally no effort from the user in expressing her information need. Although earlier work [Rhodes and Maes 2000; Lieberman 1997; Budzik and Hammond 1999] focused on document retrieval, a wide variety of content is taken into account in more recent work [Shokouhi and Guo 2015]. However, these systems either treat the retrieval system as an integral part of the application or focus on domain-specific sets of information needs. EEXCESS is agnostic of the underlying retrieval system and can be applied to any search system whose contents are searchable via a REST-API.

6. SUMMARY

In this article, we presented an infrastructure for aggregating and distributing cultural content into various channels. Conceptually, the infrastructure realizes a personalized, privacy-preserving just-in-time retrieval of cultural content supported by federated aggregation of different sources and metadata harmonization. The single components communicate with standard Web technologies and well-defined APIs, and thus can be used as stand-alone components with only minor modifications. The EEXCESS

infrastructure and all components are available as open source.[25] The developed components comprise (i) metadata harmonization, (ii) partner recommenders and a federated aggregation component, (iii) privacy preservation, (iv) user interface clients including presentation widgets and context detection, and (v) a component for automatic inclusion of new data providers.

The metadata harmonization component includes tools for creating, maintaining, and applying metadata mappings, as well as performing metadata quality assessment. These components are not restricted to data provision to EEXCESS use cases but can also efficiently support data providers in related tasks (e.g., data provision to Europeana or other cultural heritage portals). Together with PartnerWizard, they facilitate data provision from GLAM institutions.

The federated aggregation component is responsible for retrieving a single result list given a user's query as input querying partner sources in the process. Necessary steps within this component to create highly relevant and diverse result lists are the selection of appropriate sources (given the query), the query reformulation to adapt queries to single data providers, and result list aggregation. Although these processes had been expected to be computationally intensive, they have been found to be quite efficient in practice.

The privacy preservation component tackles potential privacy issues that occur when a query is sent to the federated aggregation component. Mechanisms for ensuring user privacy are a user interface to enable users to define their own privacy policy and a privacy-preserving protocol between the client and the federated aggregation component to ensure that users' queries cannot be used for user profiling.

The user interface client component comprises a module-based client architecture of user interface widgets and a context detection library. All client components can be easily integrated into client applications for presenting cultural content given the current Web-based user context.

The integration of client components has been exemplified in two applications: the Google Docs add-on for supporting content creation processes and the Chrome extension for supporting content consumption processes. The Google Docs add-on allows one to include related images and citations while collaboratively writing a document. The Chrome extension can be used to discover related cultural content while browsing the Internet. Evaluation results indicate an already overall satisfying usability. However, the following areas for improvement have been identified in a user study: (i) increase the result relevance by improving query generation and result ranking, (ii) offer a user-friendly guideline on scope of use of the Chrome extension, and (iii) improve the integration of the user interface components and decrease the complexity for first-time users.

The result relevance and ranking was targeted since the presented user evaluation. Further, a screencast was prepared and is distributed along with the extension to communicate the purpose of the module. For complex user interface parts (e.g., the visualizations that have not been described here), a tutorial was implemented to guide users through the functions of the respective module.

The client modules were also successfully implemented in other scenarios (e.g., as a Wordpress plugin [Seifert et al. 2015] and as a plugin for the Moodle e-Learning platform). Another example of the adaption of components from the EEXCESS framework is the integration of a search visualization client in the *imdas pro*[26] collection management software.

REFERENCES

James Allan, Bruce Croft, Alistair Moffat, and Mark Sanderson. 2012. Frontiers, challenges, and opportunities for information retrieval: Report from SWIRL 2012. *SIGIR Forum* 46, 1, 2–32.

Avi Arampatzis, Pavlos S. Efraimidis, and George Drosatos. 2013. A query scrambler for search privacy on the Internet. *Information Retrieval* 16, 6, 657–679.

[25]http://github.com/eexcess.

[26]http://culture.joanneum.at/.

Albert-Lásló Barabási, Réka Albert, and Hawoong Jeong. 2000. Scale-free characteristics of random networks: The topology of the World-Wide Web. *Physica A: Statistical Mechanics and Its Applications* 281, 1–4, 69–77.

Emanuele Bellini and Paolo Nesi. 2013. Metadata quality assessment tool for open access cultural heritage institutional repositories. In *Information Technologies for Performing Arts, Media Access, and Entertainment*. Lecture Notes in Computer Science, Vol. 7990. Springer, 90–103.

Sonia Ben Mokhtar, Gautier Berthou, Amadou Diarra, Vivien Quéma, and Ali Shoker. 2013. RAC: A freerider-resilient, scalable, anonymous communication protocol. In *Proceedings of the IEEE International Conference on Distributed Computing Systems (ICDCS'13)*. 520–529.

John Brooke. 1996. SUS: A 'quick and dirty' usability scale. In *Usability Evaluation in Industry*, P. W. Jordan, B. Weerdmeester, A. Thomas, and I. L. Mclelland (Eds.). Taylor & Francis, London, England, 189–194.

Thomas R. Bruce and Diane I. Hillmann. 2004. *The Continuum of Metadata Quality: Defining, Expressing, Exploiting*. ALA Editions, Chicago, IL, 238–256.

Jay Budzik and Kristian Hammond. 1999. Watson: Anticipating and contextualizing information needs. In *Proceedings of the Annual Meeting of the American Society for Information Science*. 727–740.

Jordi Castellà-Roca, Alexandre Viejo, and Jordi Herrera-Joancomartí. 2009. Preserving user's privacy in Web search engines. *Computer Communications* 32, 13, 1541–1551.

J. Debattista, S. Londoo, C. Lange, and S. Auer. 2014. LUZZU—a framework for linked data quality assessment. arXiv:1412.3750. http://arxiv.org/abs/1412.3750

Roger Dingledine, Nick Mathewson, and Paul Syverson. 2004. Tor: The second-generation onion router. In *Proceedings of the 13th Conference on USENIX Security Symposium, Volume 13 (SSYM'04)*. 21.

Josep Domingo-Ferrer, Agusti Solanas, and Jordi Castellà-Roca. 2009. h(k)-Private information retrieval from privacy-uncooperative queryable databases. *Online Information Review* 33, 4, 720–744.

Evelyn Dröge. 2012. *Criteria for Vocabulary Evaluation and Comparison*. Technical Report. Humboldt-Universität zu Berlin.

C. Dwork, E, Kumar, M. Naor, and D. Sivakumar. 2001. Rank aggregation methods for the Web. In *Proceedings of the 10th International Conference on World Wide Web.* 613–622. DOI:http://dx.doi.org/10.1145/371920.372165

Europeana Foundation. 2015. *Definition of the Europeana Data Model*. Technical Report. Europeana Foundation. http://pro.europeana.eu/page/edm-documentation.

D. Gavrilis, D.-N. Makri, L. Papachristopoulos, S. Angelis, K. Kravvaritis, C. Papatheodorou, and P. Constantopoulos. 2015. Measuring quality in metadata repositories. In *Research and Advanced Technology for Digital Libraries*. Lecture Notes in Computer Science, Vol. 9316. Springer, 56–67.

Arthur Gervais, Reza Shokri, Adish Singla, Srdjan Capkun, and Vincent Lenders. 2014. Quantifying Web-search privacy. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*. ACM, New York, NY, 966–977.

David Goldschlag, Michael Reed, and Paul Syverson. 1999. Onion routing. *Communications of the ACM* 42, 2, 39–41.

Michael Granitzer and Christin Seifert. 2016. Taking cultural and scientific content to users through the EEXCESS project. *D-Lib Magazine* 22, 3–4, 1. DOI:http://dx.doi.org/10.1045/march2016-contents.

Michael Granitzer, Christin Seifert, Silvia Russegger, and Klaus Tochtermann. 2013. Unfolding cultural, educational and scientific long-tail content in the Web. In *Late-Breaking Results, Project Papers, and Workshop Proceedings of the 21st Conference on User Modeling, Adaptation, and Personalization*. http://ceur-ws.org/Vol-997/umap2013_project_1.pdf.

Jiafeng Guo, Gu Xu, Xueqi Cheng, and Hang Li. 2009. Named entity recognition in query. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'09)*. ACM, New York, NY, 267–274. DOI:http://dx.doi.org/10.1145/1571941.1571989

Matthias Hagen, Martin Potthast, Anna Beyer, and Benno Stein. 2012. Towards optimum query segmentation: In doubt without. In *Proceedings of the International Conference on Information and Knowledge Management (CIKM'12)*. ACM, New York, NY, 1015–1024.

David Hauger, Alexandros Paramythis, and Stephan Weibelzahl. 2011. Using browser interaction data to determine page reading behavior. In *Proceedings of the 19th International Conference on User Modeling, Adaptation, and Personalization (UMAP'11)*. 147–158. http://dl.acm.org/citation.cfm?id=2021855.2021869

Martin Höffernig, Werner Bailer, Günter Nagler, and Helmut Mülner. 2010. Mapping audiovisual metadata formats using formal semantics. In *Semantic Multimedia*. Lecture Notes in Computer Science, Vol. 6725. Springer, 80–94.

Martin Höffernig, Thomas Orgel, Silvia Russegger, and Werner Bailer. 2015. Assessing quality in automated metadata aggregation and mapping services. In *Proceedings of the Workshop on Cloud-Based Services for Digital Libraries*.

ISO 21127. 2014. ISO 21127:2014: Information and documentation—a reference ontology for the interchange of cultural heritage information. Retrieved February 20, 2017, from http://www.iso.org/iso/catalogue_detail?csnumber=57832.

Marc Juarez and Vicenc Torra. 2015. DisPA: An intelligent agent for private Web search. In *Advanced Research in Data Privacy*. Vol. 567. Springer, 389–405.

Arlind Kopliku, Karen Pinel-Sauvagnat, and Mohand Boughanem. 2014. Aggregated search: A new information retrieval paradigm. *ACM Computing Surveys* 46, 3, 41.

Quoc V. Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML'14)*. 1188–1196.

Timothy Lebo, Satya Sahoo, and Deborah McGuinness (Eds.). 2013. PROV-O: The PROV Ontology. Retrieved February 20, 2017, from http://www.w3.org/TR/prov-o/.

Ryong Lee and Kazutoshi Sumiya. 2009. Zero-effort search and integration model for augmented Web applications. In *Proceedings of the 9th International Conference on Web Engineering (ICWE'09)*. 330–339.

Henry Lieberman. 1997. Autonomous interface agents. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI'97)*. ACM, New York, NY, 67–74.

Yehuda Lindell and Erez Waisbard. 2010. Private Web search with malicious adversaries. In *Proceedings of the 10th International Conference on Privacy Enhancing Technologies (PETS'10)*. 220–235.

Pasquale Lops, Marco De Gemmis, and Giovanni Semeraro. 2011. Content-based recommender systems: State of the art and trends. In *Recommender Systems Handbook*. Springer, 73–105.

Jie Lu and Jamie Callan. 2005. Federated search of text-based digital libraries in hierarchical peer-to-peer networks. In *Advances in Information Retrieval*. Springer, 52–66.

Kay Michal. 2007. XSL Transformations (XSLT) Version 2.0. W3C Recommendation. Retrieved February 20, 2017, from http://www.w3.org/TR/2007/REC-xslt20-20070123/.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into texts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Jesse Montgomery, Luo Si, Jamie Callan, and David A. Evans. 2004. Effect of varying number of documents in blind feedback: Analysis of the 2003 NRRC RIA workshop "bf_numdocs" experiment suite. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'04)*. ACM, New York, NY, 476–477.

Vanessa Murdock and Mounia Lalmas. 2008. Workshop on aggregated search. *ACM SIGIR Forum* 42, 2, 80.

Mummoorthy Murugesan and Chris Clifton. 2009. Providing privacy through plausibly deniable search. In *Proceedings of the 2009 SIAM International Conference on Data Mining*. 768–779.

Thomas Orgel, Werner Bailer, Martin Höffernig, Werner Preininger, and Silvia Russegger. 2016. *Integration and Enrichment Services Final Prototype*. Technical Report. EEXCESS Deliverable 4.4. EEXCESS.

Thomas Orgel, Martin Höffernig, Werner Bailer, and Silvia Russegger. 2015. A metadata model and mapping approach for facilitating access to heterogeneous cultural heritage assets. *International Journal on Digital Libraries* 15, 2–4, 189–207.

Sai Teja Peddinti and Nitesh Saxena. 2014. Web search query privacy: Evaluating query obfuscation and anonymizing networks. *Journal of Computer Security* 22, 1, 155–199.

Albin Petit, Thomas Cerqueus, Antoine Boutet, Sonia Ben Mokhtar, David Coquil, Lionel Brunie, and Harald Kosch. 2016. *SimAttack: Private Web Search Under Fire*. Technical Report. Institut National des Sciences Appliquées de Lyon ; Universität Passau. https://hal.inria.fr/hal-01289861

Albin Petit, Thomas Cerqueus, Sonia Ben Mokhtar, Lionel Brunie, and Harald Kosch. 2015. PEAS: Private, efficient and accurate Web search. In *Proceedings of the 2015 IEEE Trustcom/BigDataSE/ISPA Conference*, Vol. 1. IEEE, Los Alamitos, CA, 571–580.

K. J. Reiche, I. Schieferdecker, and E. Höfig. 2014. Assessment and visualization of metadata quality for open government data. In *Proceedings of the International Conference for E-Democracy and Open Government*.

B. J. Rhodes. 2000. *Just-In-Time Information Retrieval*. Ph.D. Dissertation. Massachusetts Institute of Technology, Cambridge, MA.

B. J. Rhodes and P. Maes. 2000. Just-in-time information retrieval agents. *IBM Systems Journal* 39, 3–4, 685–704.

Francesco Ricci, Lior Rokach, and Bracha Shapira. 2011. *Introduction to Recommender Systems Handbook*. Springer.

Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. 2010. *Automatic Keyword Extraction from Individual Documents*. John Wiley & Sons. DOI:http://dx.doi.org/10.1002/9780470689646.ch1

Raoul Rubien, Hermann Ziak, and Roman Kern. 2015. Efficient search result diversification via query expansion using knowledge bases. In *Proceedings of 12th International Workshop on Text-Based Information Retrieval (TIR'15)*.

Jörg Schlötterer. 2015. From context to query. In *Proceedings of the ACM Symposium on Applied Computing (SAC'15)*. ACM, New York, NY, 1108–1109.

Jörg Schlötterer, Christin Seifert, and Michael Granitzer. 2016. Supporting Web surfers in finding related material in digital library repositories. In *Proceedings of the International Conference on Theory and Practice of Digital Libraries (TPDL'16)*.

H. A. Seid and A. L. Lespagnol. 1998. Virtual private network. US Patent 5,768,271.

C. Seifert, J. Jurgovsky, and M. Granitzer. 2014. FacetScape: A visualization for exploring the search space. In *Proceedings of the 2014 18th International Conference on Information Visualization (IV'14)*. 94–101.

Christin Seifert, Nils Witt, Sebastian Bayerl, and Michael Granitzer. 2015. Digital library content in the social Web: Resource usage and content injection. *IEEE STCN Newsletter* 3, 1. https://sites.google.com/a/ieee.net/stc-social-networking/e-letter/stcsn-e-letter-vol-3-no-1/.

Marc Shapiro. 1986. Structure and encapsulation in distributed systems: The proxy principle. In *Proceedings of the 2013 IEEE 6th International Conference on Distributed Computing Systems (ICDCS'86)*. 198–204.

Milad Shokouhi and Qi Guo. 2015. From queries to cards: Re-ranking proactive card recommendations based on reactive search history. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'15)*. ACM, New York, NY, 695–704.

Milad Shokouhi and Luo Si. 2011. Federated search. *Foundations and Trends in Information Retrieval* 5, 1, 1–102.

B. Stvilia, L. Gasser, and M. Twidale. 2007. A framework for information quality assessment. *Journal of the American Society for Information Science and Technology* 58, 12, 1720–1733.

Vincent Toubiana, Lakshminarayanan Subramanian, and Helen Nissenbaum. 2011. Trackmenot: Enhancing the privacy of Web search. arXiv:1109.4677.

T. Trippel, D. Broeder, M. Durco, and O. Ohren. 2014. Towards automatic quality assessment of component metadata. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*.

Gerwald Tschinkel, Cecialia di Sciascio, Belgin Mutlu, and Vedran Sabol. 2015. The recommendation dashboard: A system to visualise and organise recommendations. In *Proceedings of the International Conference on Information Visualisation (IV'15)*. 241–244.

Hermann Ziak and Roman Kern. 2015. Evaluation of pseudo relevance feedback techniques for cross vertical aggregated search. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. Lecture Notes in Computer Science, Vol. 9283. Springer, 91–102.

Stefan Zwicklbauer, Christin Seifert, and Michael Granitzer. 2016a. Robust and collective entity disambiguation through semantic embeddings. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'16)*.

Stefan Zwicklbauer, Christin Seifert, and Michael Granitzer. 2016b. DoSeR—a knowledge-base-agnostic framework for disambiguating entities using semantic embeddings. In *Proceedings of the European Semantic Web Conference (ESWC'16)*.

Received April 2016; revised August 2016; accepted October 2016