Large Scale Analysis of Semantic and Temporal Aspects in Cultural Heritage Collection's Search

Yasunobu Sumikawa
Tokyo Metropolitan University
Tokyo, Japan
sumikawa-yasunobu@tmu.ac.jp

Adam Jatowt
Kyoto University
Kyoto, Japan
adam@dl.kuis.kyoto-u.ac.jp

Antoine Doucet
University of La Rochelle
La Rochelle, France
antoine.doucet@univ-lr.fr

Jean-Philippe Moreux
National Library of France
Paris, France
jean-philippe.moreux@bnf.fr

ABSTRACT

Cultural Heritage (CH) collections store and represent numerous historical objects such as images and documents, which are often of unique type and of high cultural and historical value. While the maintained objects are typically well-understood and analyzed, relatively less is known about *what type of content is actually interesting to the searching users* and *how they find such content*. These kinds of analyses could help us to understand for what purposes users use cultural heritage collections and could lead to user intent classification and understanding. In this paper we report the results of a large-scale exploratory analysis based on a 15-month long snapshot of query logs generated at the online portal of the National Library of France. Besides understanding the nature of content search in cultural heritage collections and digital libraries, the results of our study can be used for designing content recommendation systems and could help to improve time-aware search applications.

KEYWORDS

Digital library, digital history, temporal query analysis

1 INTRODUCTION

A rapid increase of text content stored in digital archives is the result of widespread digitization and content curation initiatives aiming at preserving our heritage. Millions of newspaper articles, books, past snapshots of web pages or other document genres and objects are digitized and made accessible and searchable. Thanks to the efforts of digital archiving, the number of historical information one can access is then rapidly increasing. This offers opportunities for computational studies on explicit references to past events and opens up novel perspectives for the study of collective memories as well as the pursuit of public history. While computational collective memory studies have already been conducted on news articles [1, 6], Wikipedia [10, 11, 13, 15] and microblogs [5, 21], with regard to cultural heritage collections, few projects focused on user interest towards the type of stored content. In order to be able to successfully disseminate content of cultural heritage collections and to attract new users it is imperative to understand user needs, user experience, and search tactics. Query logs offer rich content towards this end.

The distinguishing feature of search in heritage collections and often also in other constrained collections is the high probability of a within-collection navigational search intent (also termed as known-item search), when one uses the well-known Broder's taxonomy [2]. In other words, rather than learning about some concept or answering some question as in typical information seeking process, users are more likely to wish to access certain digital artifacts or documents. Field (zone) based search, commonly associated with the choice over controlled vocabulary, is also typical for cultural heritage collections. For example, users input metadata information for required objects such as author names, creation dates, etc.

Nevertheless, many aspects of search withing CH collections are still unclear. Among others our study is guided by the following key questions:

(1) What do people search for in digital historical collections and what kinds of metadata filtering they use?
(2) What is the time horizon of such searches?
(3) How are collective memories expressed in search logs?

We approach these and other related questions by investigating portions of query logs generated from over the course of one year. As the underlying dataset we use request logs from the online portal to the digital library of the National Library of France (Bibliothèque nationale de France) (BnF) called Gallica which has been open to the public since 1997[1]. Gallica contains printed materials such as books, journals, newspapers, printed music, as well as other document types. It also serves graphic materials like engravings, maps, photographs and sound recordings. At present, Gallica holds 5M documents[2].

Based on the collected user interaction data, we investigate the distinguishing characteristics of queries from the semantic and temporal viewpoints. Among the key aspects we research are time horizons, concerned entities and the prevalence of different types of result filtering. Particularly, we are not aware of similar studies of search logs that would elucidate patterns of chronological result filtering that we perform in this work. As archival artifacts have natural temporal ordering, this kind of studies are quite important for understanding the intents and tactics of users.

Besides being a novel approach towards the general questions on semantic and temporal aspects of ways in which users search in cultural heritage collections as well as on how the past relates

[1] https://gallica.bnf.fr/accueil/en/content/accueil-en?mode=desktop
[2] https://gallica.bnf.fr/GallicaEnChiffres

to our lives, our investigations can be beneficial to several applications. First, the *archival appraisal* decisions can be supported thanks to the knowledge of user interest concerning past artifacts. It is well-known that archives have limited resources and funds for collecting, storing and managing objects, even, ones of digital character, and there must be a well-thought selection process conducted for deciding which objects should be included and for how long in archives. Next, the *construction of specialized content detection and recommendation systems* can be informed by the reported results. The objective of such systems would be to facilitate sharing of historical knowledge. Understanding the types of popular searched content and the context of sharing can be helpful for designing effective recommendation systems.

To sum up we make the following contributions in this paper:

(1) We undertake analysis of cultural heritage search based on a large scale data collected over a year.
(2) We propose to analyze the search log from entity-oriented and temporal viewpoints to uncover what types of and how the entities are searched for as well as how temporal filters are set by searchers.
(3) We outline novel research directions of analysis and mention potential applications that can better utilize archival contents for improving user experience.

The remainder of this paper is structured as follows. In the next section we present the related work. In Section 3 we describe the data collection and processing steps. Section 4 provides the findings of the semantic and temporal analysis. We then include additional discussions of limitations of our study in Section 5. Finally, the last section concludes the paper and outlines our future work.

2 RELATED WORK

2.1 Search Log Analysis in Digital Libraries

A lot of researches focused on analyzing query logs of search engines such as Web search engines [14]. User queries, session information, short-term or long-term search tasks as well as interaction patterns were studied among others.

Information access in digital libraries and constrained document collections such as scholarly repositories has been also subject of research interest. Sfakakis and Kapidakis [20] compared in 2002 the usage of a Digital Library with many different categories of collections concluding that the access points the users mostly refer to depend heavily on the type of content of the collection, the details of the existing metadata and the target user group. The authors have also found that most users tend to use simple query structures such as ones containing a single search term, and they tend to do very few and primitive operations to accomplish their requests. However, as users get more experienced, they reduce the number of operations in their sessions. Another work [4] presented the results of a large-scale case-study at the Royal Library of Belgium based on a data set of 83k queries from 29k visits over a year long period of the historical newspapers platform BelgicaPress (associated with the State Archives of Belgium). The authors investigated the application of simple text mining methods such as query clustering in cultural heritage settings. However, no other results were given except for few example data instances.

De Wilde *et al.* [8] analyzed search queries over 4 years' long period against the Historische Kranten corpus which contains over a million articles compiled from 41 Belgian newspapers published between 1818 and 1972 and written in Dutch, French, and English, that focus on the city of Ypres and its neighbourhood. Based on 10 top most popular queries they found that locations are especially common. Ceccarelli *et al.* [3] analyzed query log consisting of over 1 million queries issued during 6 months in order to find improvements useful for increasing the usability levels of the Europeana Portal. They found among others that about 20% users use filtering by type, the average query length is 1.86 terms and most of search sessions finish before 5 minutes. They also proposed to develop a query recommender system based on the analysis results. Lown *et al.* [16] through the search log analysis of NC State University's (NCSU) Libraries have observed that the default search box labeled "All" was used for 73.7 percent of searches while the other specialized sources (e.g., catalog labeled "Books & Media") were used less frequently. Same as in our case D'Alché-Buc *et al.* [7] have also studied Gallica search logs to analyze search patters in cultural heritage collections. They especially focused on classification of user sessions.

Our work differs in several aspects from these works. First, our focus is on entities included in queries and on the way in which users use temporal filters to constrain search results. We also investigate in detail the usage of metadata filters. Finally, unlike the other works we utilize large scale query log compiled over relatively long time period.

2.2 Named Entity Recognition

Finally, Named Entity Recognition (NER) and Classification (NEC) methods, especially ones related to query parsing, are also related to our work [12, 19, 22, 23]. The task consists of assigning an entity type such as a person, location, event or organization to entities identified from search queries. Many modern web search engines use information on named entities stored in Knowledge Graphs to be presented alongside organic result pages when queries are entity-centric [22, 23]. This means the need for recognizing and disambiguating entities mentioned in queries or entities strongly related to user queries [9, 22].

3 DATA COLLECTION

Dataset. As mentioned before, we use the interaction log from the online portal of the National Library of France (BnF) called Gallica [3]. We show the basic statistics of the dataset in Tab. 1 such as the numbers of records and the numbers of requests. We use records over the period of 15 months.

Table 1: Dataset statistics.

Duration	Jan. 2016 ~ Apr. 2017
Num. of records	2,844,553,550
Num. of requests	1,126,787,556
Num. of requests using temporal filters	35,040,848
Num. of requests using resource filters	333,689,345
Num. of requests using DC filters	26,075,373
Num. of requests using temporal and resource filters	5,665,126

[3] https://gallica.bnf.fr/accueil/en/content/accueil-en?mode=desktop

As it is impossible to extract parameters of POST methods from log file, all the interaction-related information we focused on was sent by GET methods, meaning that we analyzed content embedded in URLs. In Tab. 1, a record indicates any kinds of interactions between a user and the BnF server whereas a request is an interaction by GET method.

Parameters. We collected the following parameters from each request:

(1) *query*: it includes keywords input by users. This parameter also contains information about sorting, filtering and matching methods. These three conditions can be used in the advanced search settings of Gallica. As for the filtering, Gallica provides several ways, such as resource, temporal, language, theme, types of documents and so on.
(2) *location*: shows the name of the country from where the query is issued
(3) *timestamp*: represents a time when a user issues a query.

Preprocessing. After extraction of keywords from the *query* parameter, the following pre-processing steps were performed. First, all the words were lower-cased. Stopwords were then removed and the remaining words were subject to stemming. For stop-words removal and stemming process, we used Python NLTK functions designed for French language[4].

4 ANALYSIS

In this section, we first investigate characteristic features of requests focusing on locations and metadata based filtering. We then perform entity analysis for query words. Finally, we show results from the temporal analysis of queries based on temporal filters set by users. At the beginning of each subsection we list in rectangle boxes the guiding questions behind the analysis outlined in the subsection.

4.1 Spatial Analysis

Q. What is spatial distribution of originating requests?

Before we start query keywords analysis, we briefly show the spatial distribution of requests. Fig. 1 displays the top-10 places where users input requests with their corresponding rates. "NULL" indicates that there is no location information recorded at log. Naturally, France is the top country which is not surprising since data is hosted in the National Library of France. We can however observe that many European countries are included in this figure.

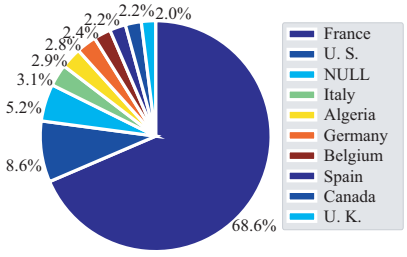

Figure 1: Top-10 countries from which requests originate.

[4]https://www.nltk.org/

4.2 Metadata based Filtering Analysis

Q. What kind of filters do searchers apply?
Q. What kinds of metadata and in what combinations are commonly used when searching based on Dublin Core?

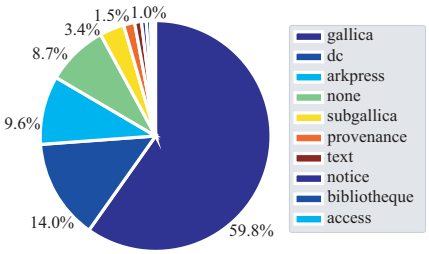

Figure 2: Top filters used with query words.

We now look at the distribution of filters used in queries shown in Fig. 2. Gallica's advanced search not only allows to choose the underlying resources, but also offers several options such as AND/OR retrieval, setting temporal filtering, specific languages used for describing digital items, and so on.

As it can be seen in Fig. 2, `dc` and `arkpress` were the main filters besides the default `gallica`. The Dublin Core (DC) schema is a small set of vocabulary commonly used to describe a wide range of digital resources (video, images, web pages, etc.). The schema describes metadata vocabularies that include sets of resource classes, type vocabularies, vocabulary encoding schemes and syntax encoding schemes. Its details are given in the official web site[5]. In Gallica settings, `arkpress` is used to perform queries within a newspapers title and `subgallica` is used to filter the results page (facets search).

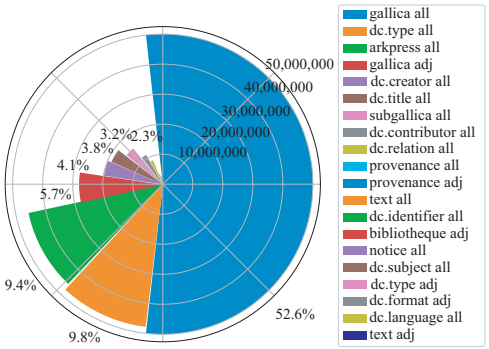

Figure 3: Top-20 filter–parameter combinations.

Fig. 3 plots more detailed ratios of metadata filtering styles. The advanced queries also include three options (`all`, `any` and `adj`) concerning the text matching way that Gallica enables. The first one indicates the case of matching of all words. The second one is used to find matches on, at least, one of the words. The third one is the exact/equal expression (i.e., a phrase like search which is

[5]http://dublincore.org/documents/dces/

typically available in search engines when using quotes). We can see that using `Gallica` with matching of all words occurs over 50% of the times. In addition, there are 3 kinds of selectors (`Gallica`, `dc` and `arkpress`) whose total ratio amounts to over 80%. In the top-20, it can be seen that the `all` parameter (i.e., matching all words) is the most common followed by the `adj` parameter (the exact/equal expression). However, we should keep in mind that a lot of these parameters, particularly `all`/`adj`, are tuned by the Gallica web app itself. Anyway, the results suggest that users are generally reluctant to use advance search options.

Table 2: Common combinations of resources in the same queries.

Rank	Resource	Resource	Num
1	dc	dc	6,450,642
2	dewey	dewey	3,252,938
3	dc	notice	700,619
4	notice	notice	267,614
5	gallica	gallica	396
6	gallica	arkpress	306
7	gallica	subgallica	66
8	gallica	dc	66
9	dc	subgallica	66
10	dc	jean	12

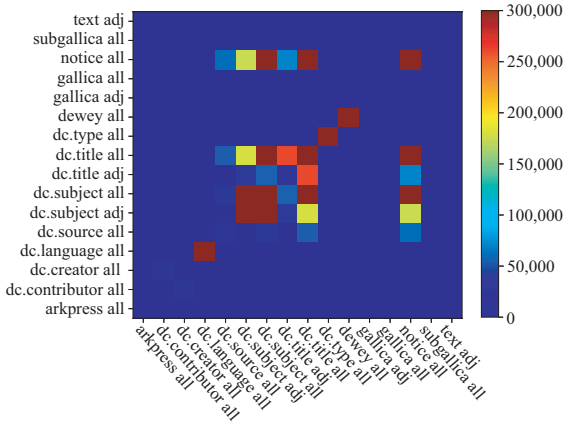

Figure 4: Numbers of filter–parameter combinations used together.

As in the advanced search option Gallica supports multiple search conditions for each query, we analyze combinations of filters and filters–parameters. First, Tab. 2 shows the numbers of co-used filters in queries. We can observe that `dc` and `notice` are commonly used together. `dewey` tends to be used separately. To better understand why some users use the same types of filters with different parameters, Fig. 4 plots numbers of co-used filters with their parameters. We can see that two kinds of parameters, `all` and `adj`, are the most common in this figure. `dc.subject all` and `dc.subject adj` tend to be used together. In addition, `notice all` and `dc.title all` are used with `dc.source all`, `dc.subject adj`, `dc.subject all`, `dc.title adj` and `dc.title all`.

Finally, Fig. 5 shows top-15 inputs in `dc.type`. We can see that the top-5 choices are related to periodical and books (periodical, monograph), manuscripts, images or maps. This result confirms the previously reported results of user survey and web analysis [18].

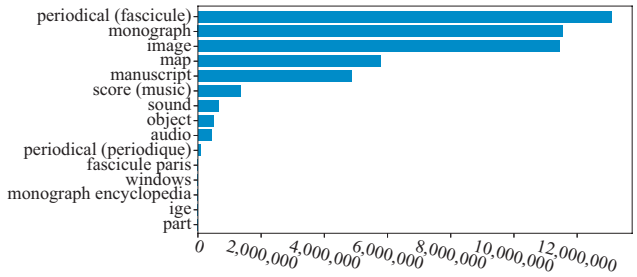

Figure 5: Top-15 data used in `dc.type` filed.

4.3 Entity Analysis

> Q. What is the distribution of searched entity types?
> Q. What entities are most popular?

In this section, we investigate what kind of entities and what kind of entity types searchers input in their queries.

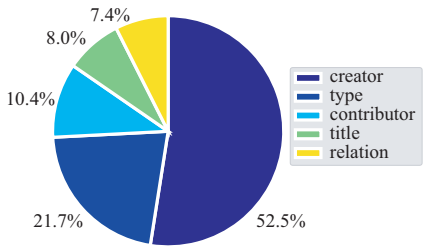

Figure 6: Top-5 metadata fields defined in `dc`

4.3.1 Entities in DC Fields. At first, we focus on analyzing what types of DC fields users tend to use. Fig. 6 shows the distribution of fields related to DC. Three fields, `dc.creator`, `dc.contributor` and `dc.title`, that can be regarded as entity representing ones are most prevalent. Approximately 71% queries that are subject to DC filter use at least one of them. Following, we list the top-20 entities used with these three fields in Figs. 7, 8 and 9, respectively. Note that the top-20 analysis provides only a partial picture as it reflects only the usage of the most common entities in DC fields. We leave the more extensive analysis of less common entities as future work.

We manually checked the nationalities, birth and death years as well as the occupations of the persons listed in Fig. 7. Among them, 10 are French, 3 are Portuguese[6] and 2 are Brazilians [7]. They all lived around 1800 and all were born before 1860. After counting

[6] One Portuguese (Ramalho Ortigão) taught French at a college in Porto, and another Portuguese José Feliciano de Castilho Barreto e Noronha studied medicine in France.
[7] Henrique de Beaurepaire-Rohan was traveller and explorer of French extraction. Pedro II of Brazil was elected to the French Academy of Sciences and died in Paris.

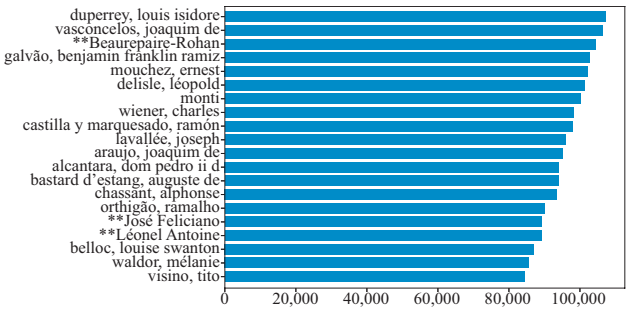

Figure 7: Top-20 inputs used with dc.creator field. "" mark indicates abbreviated names of entities are: **Beaurepaire-Rohan (beaurepaire rohan, henrique pedro carlos), **José Feliciano (noronha, josé feliciano de castilho de barreto e), **Léonel Antonie Feliciano (noronha, josé feliciano de castilho de barreto e)**

the number of occupations[8] we found that 8 persons worked as university faculty, 4 were politicians and 3 were army officers. There were also 2 journalists, 2 poets, 2 writers and 1 artist.

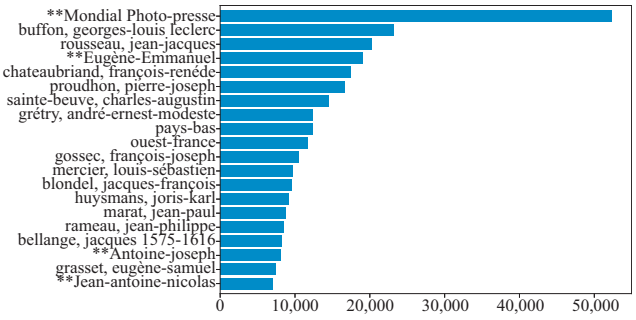

Figure 8: Top-20 inputs used with dc.contributor. "" mark indicates abbreviated names of entities are: **Mondial Photo-presse (agence de presse mondial photo-presse), **Eugène-Emmanuel (viollet-le-duc, eugène-emmanuel), **Antoine-joseph (dezallier d'argenville, antoine-joseph), **Jean-antoine-nicolas (condorcet, jean-antoine-nicolas de caritat)**

We did the same for Fig. 8 as for Fig. 7. In Fig. 8, 3 entity types are found: 2 media groups and 17 persons (the main nationality of persons was French: 14 French, 2 Swiss and 1 Belgium). Similar to the case of `dc.creator`, the 19th century is the main period of the lifetimes of these entities.

Newspaper articles, books, journals and reviews as well as a drama (True West) are among the most common inputs in `dc.title` field that are displayed in Fig. 9. Interestingly, locations Gare Saint-Lazare, Notre-Dame de Paris and Tierra del Fuego appear in the ranking, too. Similar to the results of the two metadata fields `creator` and `contributor`, the 19th century is the main period

[8] Several persons had more than one occupation at the same time.

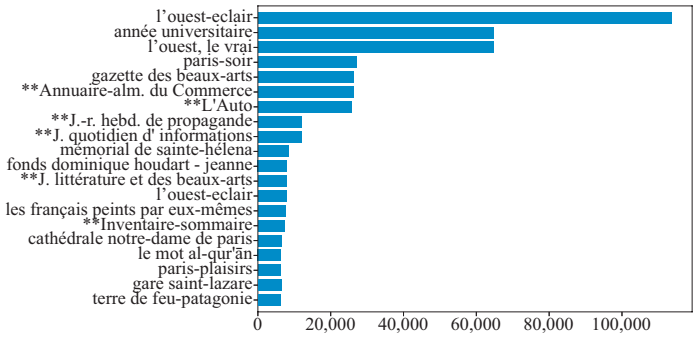

Figure 9: Top-20 inputs used with dc.title field. "" mark indicates abbreviated names of entities are: **Annuaire-alm. du Commerce (annuaire-almanach du commerce, de l'industrie), **L'Auto (l'auto-vélo: automobilisme, cyclisme), **J.-r. hebd. de propagande (le populaire: journal-revue hebd. de propagande), **J. quotidien d'informations (l'ouest, le vrai: journal quotidien d'informations), **J. littérature et des beaux-arts (l'artiste: journal de la littérature et des beaux-arts), ** (inventaire-sommaire des archives départementales antérieures)**

when the documents associated with most of these queries were published.

4.3.2 Entities in Text Input. Next, we extract entities from within free text inputs, i.e., from query keywords issued by users which are not related to any particular metadata field. First, we mapped all queries into DBpedia to extract entities from texts. We then determined their types using DBpedia in order to arrange them into 5 basic types: *places*, *persons*, *events*, *groups* and *others*. Fig. 10 shows the distribution of the entity types obtained from the free text inputs. Places is definitely the most popular entity type in the issued queries followed by persons. This is somehow natural as people often recall the past using the names of places where historical events happened as they may be interested in documents/objects related to the histories of locations they live at or are visiting.

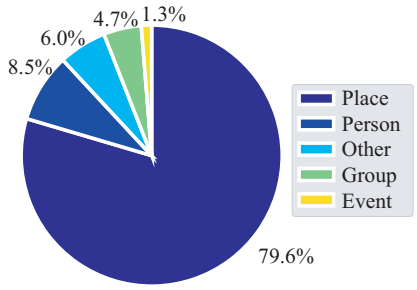

Figure 10: Ratio of types defined using DBPedia.

Tab. 3 lists also the top-10 entities per each type. Naturally, French entities occupy many top places (e.g., 4 French persons, 4 cities in France and at least 3 French groups). In the Person column, 2 singers (Larusso and Wallen) and 2 philosophers (Boethius

and Voltaire) are included. We can also see that there are 3 music groups in the Group column, which also suggests that musician profession is a common occupation of sought persons. Interestingly, after investigating the lifetimes we found that the musicians and music bands are rather present entities (their birth or union years being between 1978 to 2006). For the Group column, in addition to the 3 music groups there are many corporations (e.g., Renault, Michelin and Commer). Many of the corporations are characterized by rather long histories having been founded in the 19th or early 20th centuries. In the Place column, among others, there are 4 French cities and 3 cities (Laghouat, Mahanoro and Kabylie) whose history largely connected with France.

This automatic analysis mapping user queries to DBpedia shows a different trend from a detailed manual analysis of top-1,000 queries [17]. The manual analysis revealed that works (periodicals, monographs, etc.) are actually the most common queried entity type (38%) followed by persons (26%). As automatic mapping of user queries with work titles and person names is not an easy task, our automatic analysis may increase the ratio of places.

Table 3: Top-10 entities per type by DBpedia.

Rank	Persons	Place	Group
1	Boethius	Rambouillet	Renault
2	Polus	Maizy	Switchfoot
3	Larusso	Laghouat	Michelin
4	Wallen	Chile	Massai
5	Gabriel	Mayotte	Somalis
6	Etteilla	Mahanoro	Boer
7	Medea	Goa	Babyshambles
8	Taurinus	Kabylie	Ovni
9	Voltaire	Paris	Angelis
10	Ptahhotep	France	Animaux

Overall, from Figs. 7, 8, 9 and Tab. 3, we can observe that the most popular entities were ones active in the 19th century and that locations and persons are especially common. When analyzing persons and groups, we notice that there are many musicians/artists, writers and scientists as well as music bands or corporations.

4.4 Temporal Analysis

Q. What dates are commonly used for filtering results?
Q. How long are on average the filtering time spans?
Q. How similar are queries issued for different centuries?

Table 4: Temporal coverage of Gallica content.

Century	Num. of contents	Century	Num. of contents
2nd	759 (0.02%)	16th	39,239 (1.00%)
10th	1,010 (0.03%)	17th	81,781 (2.09%)
11th	1,128 (0.03%)	18th	229,609 (5.87%)
12th	2,150 (0.05%)	19th	1,603,188 (40.97%)
13th	3,324 (0.08%)	20th	1,821,245 (46.54%)
14th	4,900 (0.13%)	21st	112,706 (2.88%)
15th	11,903 (0.30%)		

We next analyze the way in which users filter the results by time. First, to provide background context, we start by showing the temporal span of content stored in Gallica in Tab. 4, from which we can observe that most of the content comes from the 19th and 20th centuries (40% and 46%, respectively).

Next, we collect all time constraints from parameters of requests. As it was shown in Tab. 1 there are over 35 million requests in a dataset for which a time filtering was applied. The time filtering of search results is implemented by a parameter `gallicapublication_data`. We then convert all the collected temporal expressions from the parameter into corresponding durations. For example, if there is a "1900 ≤ `gallicapublication_data` ≤ 1950" in a query's parameter, the temporal expression is recognized as [1900, 1950]. In case when no start year is given, we set AD 1 as the start year. On the other hand, when no end year is specified, we set the year when the query was posted as the end year.

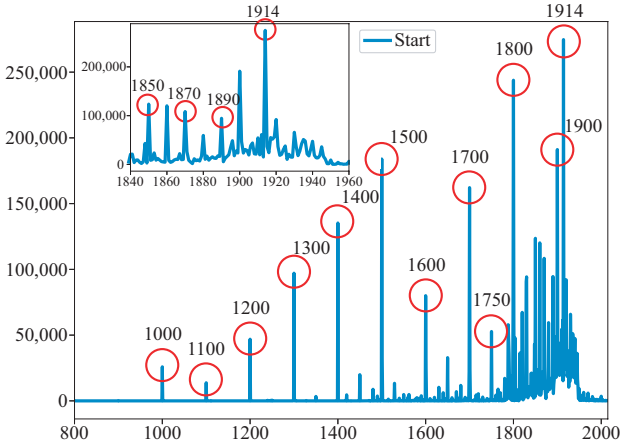

Figure 11: Distribution of start years in filters using `gallicapublication_data`.

4.4.1 Start Dates in Filtered Time Spans. First, Fig. 11 plots the distribution of start years of the `gallicapublication_data`. Looking at the figure (and also at the zoomed in plot in the inner graph), we can see that the number of start years is usually rapidly increasing towards the present, yet relatively fewer start dates are set at the last decades of the 20th century and onwards. In addition, several significant peaks are visible in Fig. 11. For example, the peaks are noticed for rounds dates from 1000 to 1900 which denote the start of new centuries. We can also see peaks on years starting new decades which are especially common in the 19th century (e.g., the zoomed plot in the inner graph shows such peaks from 1850 to 1900).

Also, it can be observed that on average non-round dates like 1938 or 1911 are more common in the 20th century than in any other century. Anyway, in general, the recent past is referred to more than the distant past. This is intuitive and correlates with the previous study conducted on news articles related to different countries [1]. Finally, we can see that there are two peaks on years 1750 and 1914, the latter being the most common start date in the entire dataset. To understand why users use often these two we collected the top-10 words used with these two filtering choices, and present them in Tab. 5.

Table 5: Top-10 words used with peaks of 1750 and 1914. We translated some of French words into English when the meaning was not immediately clear. "*" mark indicates words translated into English which are: *anatomy (anatomie), *National Library of France (Bibliothèque nationale de France), *Gallipoli Campaign (1915) (Dardanelles, expédition des 1915), *aviator (aviateur), *fatherhood (paternitè), *recognition (reconnaissance).

Rank	1750 (Freq.)	1914 (Freq.)
1	roquefort (25,152)	appartient à l'ensemble documentaire : abcdaire1 (94,529)
2	fromage (25,017)	*National Library of France (94,016)
3	thomery (7,914)	*Gallipoli Campaign (1915) (55,210)
4	dictionnaire (4,668)	*aviator (24,048)
5	tomery (4,041)	portrait (10,289)
6	cb32693529c_date (2,832)	cb34378481r_date (8,369)
7	*anatomy (2,772)	*fatherhood (4,749)
8	allcoll (1,216)	*recognition (4,749)
9	cb328131247_date (939)	officiel de la République française (3,262)
10	jean de l'ours (864)	1914–1918 (3,066)

In 1914, for example, we can see two WW1 related entities: Dardanelles and 1914–1918. The Dardanelles is a location in northwestern Turkey where the Britain–France military attacked the Ottoman Empire on 1915 in the event known as the Battle of Gallipoli. The second one (1914–1918) is related to a book whose author analyzed the causes, course and impact of the WW1.

Furthermore, "cb32693529c" and "cb328131247" are periodicals covering post 1750 periods, while "cb34378481r" is the ID of *Journal officiel de la République française*. "allcoll" is an internal keyword ("all collections") "fatherhood recognition" is probably a query performed by users in *Journal officiel de la République française* to find acknowledgement of paternity. Finally, "ABCdaire1" is a keyword which gives access to a thematic collection (childhood illustrated books).

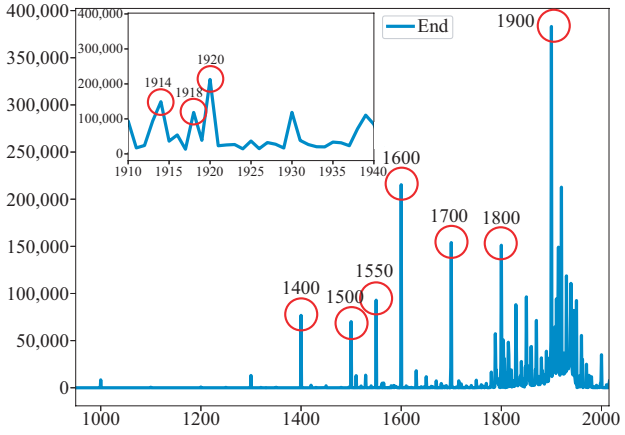

Figure 12: Distribution of end years in filters using `gallicapublication_data`.

4.4.2 End Dates in Filtered Time Spans. Next, Fig. 12 plots a distribution of end years of the `gallicapublication_data`. Similar to the previous distribution over start years, we see that the number of the set end years is usually increasing towards the present, especially they are common from the 19th century to the 20th century. In addition, several significant peaks are visible on years initiating new centuries within the period spanning from 1400 to 1900. Similar to Fig. 11 we observe that non-round dates in the 20th century (but not so much in the other centuries) are commonly used as end filtering constraints. Looking at the zoomed out plot in the inner graph, we can also notice 3 peaks on eventful years 1914, 1918 and 1920. Finally, somewhat interestingly, the dates within the 21st century (e.g., 2000, 2016 or 2017) are actually not so commonly used. Instead, 1900 is the most popular end date. This is understandable if we consider that Gallica mainly holds public domain material.

Table 6: Top-10 words used with peaks of 1918 and 1920. We translated some of French words into English when the meaning was not immediately clear. "*" mark indicates words translated into English. These are: *aviator (aviateur), *National Library of France (Bibliothèque nationale de France), *Chinese Brest (chinois brest), *Chinese (chinois), *news bulletin (bulletin presse), *Gallipoli Campaign (1915) (Dardanelles, expédition des 1915), *Chinese Dunkerque (Chinois Dunkerque), *car (voiture).

Rank	1918 (Freq.)	1920 (Freq.)
1	*aviator (24,048)	*National Library of France (50,014)
2	*National Library of France (24,044)	*Gallipoli Campaign (1915) (49,876)
3	*Chinese Brest (21,057)	portrait (15,168)
4	*Chinese (11,862)	boutiques parisiennes (14,649)
5	cb32732912f_date (9,149)	*Chinese (13,569)
6	quotidien presse étrangère (8,879)	*Chinese Saint Yorre (11,682)
7	*news bulletin (7,121)	*Chinese Dunkerque (9,459)
8	cb34355551z_date (4,917)	navigation (8,556)
9	cb34378481r_date (4,318)	cb34519208g_date (8,097)
10	marquis de givenchy (3,366)	*car (6,480)

Tab. 6 shows the top-10 words used with the two peak years 1918 and 1920. For example, as 1918 is the end year of WWI, we can notice few war related words (aviateur that means pilot, china and press) are ranked in the top-10 words on this year.

4.4.3 Lengths of Filtered Time Spans. We then analyze how long time periods users tend to use for result filtering. Fig. 13 plots the distribution of the lengths from the start to end years of the `gallicapublication_data`. The numbers of set periods tend to decrease along with their duration, meaning that relatively short time spans are more commonly used than long durations. Indeed, many users set filters spanning a single year. A single year is actually an interval being associated with the highest number of queries (the interval of 100 years being the 2nd most common). Similar to the above two results presented in Figs. 11 and 12, there are peaks on 'round lengths' such as 100, 200, 300, and to lesser extent on 500 and 1,000. Interestingly, we can see that four peaks occur on 1788, 1800, 1829 and 1850. The reason of the occurrence of such long-spans is because users sometimes do not input start or end years; they just input only one of them as the filtering date.

As we looked at start and end years in the above analyses, we now check which start and end dates in particular are often used together.

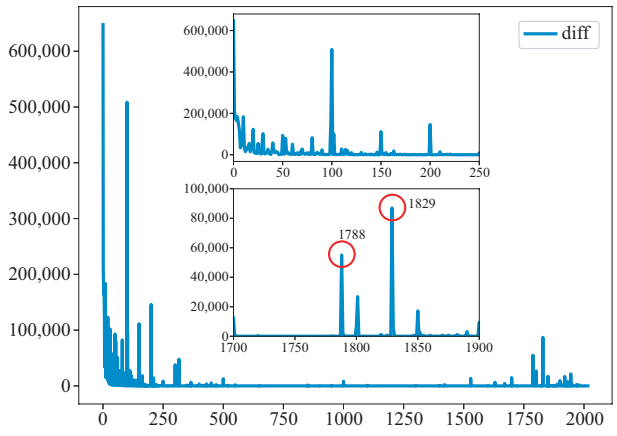

Figure 13: Distribution of lengths of chosen time scopes based on `gallicapublication_data`. The x axis represents the lengths in years whereas the y axis represents the numbers of queries associated with a given time duration.

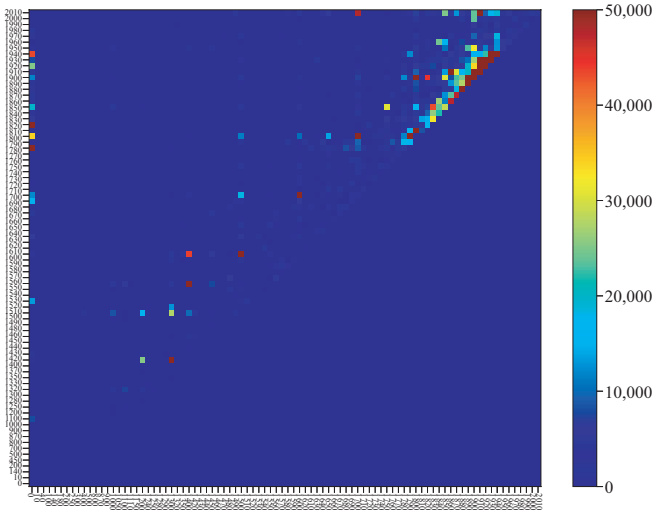

Figure 14: Correlation between decades representing the start and end years of `gallicapublication_data`.

For this we map all years from `gallicapublication_data` into decade-level granularity, e.g., 1988 is mapped into 1980s. Fig. 14 shows the numbers of co-used start and end years mapped into decade-level granularity. In this figure, x and y axes indicate the start and end decades, respectively. In addition, Fig. 15 presents the selection of the most commonly used decades in which only the top-20 most common decades are shown either as a starting or ending decade.

We can observe three straight lines: a) from 1780s/1790s to 1930s/1940s, b) first column of 0s, and c) the most upper row of 2010s. The first line indicates that users set filter durations to be relatively short; i.e., users tend to set years within 30 years from 1820s to 1840s, 40 ~ 60 years from 1850s to 1920s. After reexamining Figs. 7 and 8, we believe that users may search for works of writers, artists, scientists (and other similar occupations) who were alive from 1820s to 1920s. The remainder two lines mean that searchers sometimes set the start date as years on decade 0s or the end date as years falling into 2010s.

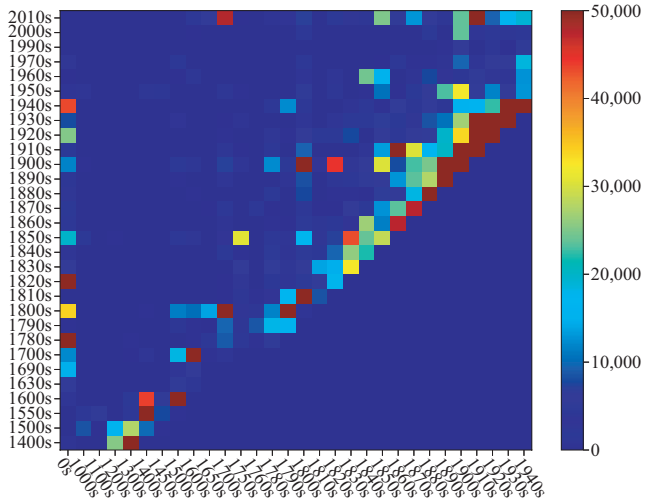

Figure 15: Top-20 decades representing the start and end years of `gallicapublication_data`.

4.4.4 Time Filter Combination Analysis. Next, we analyze the combined time signals. Fig. 16 plots three lines corresponding to `dc.creator`, `dc.contributor` and `gallicapublication_data`. To plot a line of `gallicapublication_data`, we collected years used with the filtering conditions. On the other hand, to plot the other two lines, we used years input in two fields: `dc.creator` and `dc.contributor`. After collecting all the temporal inputs, we converted them to probability distributions over their corresponding timespans using year level granularity. For a given time input (e.g., 1960s) with t_b denoting its start year (1960) and t_e indicating its end year (1969) we set the probability distribution with zero values for $t < t_b$ and for $t > t_e$ (e.g., before 1960 and after 1969) and with non-zero values for $t_b \leq t \leq t_e$ that sum to 1 (e.g., 1/10 for each year from 1960 to 1969). We then combined for every year all the computed probability distributions based on all the time signals. The formal definition of the probability distribution for a year y is given in Eq. 1.

$$S(y) = \sum_{[t_b,t_e]\in T} \delta(y,[t_b,t_e]) * \frac{1}{t_e - t_b + 1} \quad (1)$$

where T includes all the time inputs, and the function δ returns 1 if the first argument is included in the second argument; otherwise, it returns 0.

Looking at the lines shown in Fig. 16, we can say that present data is more popular than past data as all the lines are increasing towards the present time. Indeed, Figs. 7 and 8 show that most common persons are related to the 19th century.

4.4.5 Query vs. Time Span Analysis. Finally, Fig. 17 shows the similarities between query terms associated with different centuries. We calculated the similarities by Jaccard Coefficient computed by

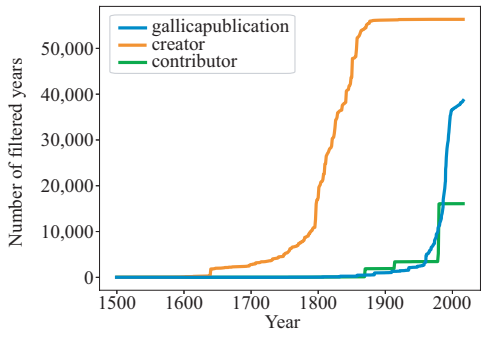

Figure 16: Distribution of time references in filters with `gallicapublication_data` and ones in texts of `dc.creator` and `dc.contributor` fields.

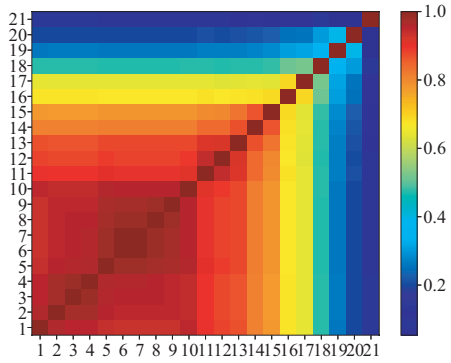

Figure 17: Jaccard similarity between centuries based on their associated queries. The x and y axes represent centuries associated with queries that have `gallicapublication_data` filtering.

first measuring how many common words queries related to any two centuries are shared and then by normalizing the obtained number by the sizes of the two word sets. The more the two centuries have common words, the more similar they are. To plot the figure, we first collected queries that have start and end years as the `gallicapublication_data` filtering. We then determined which centuries are associated with each query. In particular, we counted how many years of each century are included within the interval delimited by the start and the end years. The centuries having the largest overlap were then assigned to the query. The algorithm is based on the following equation:

$$Century(d) \quad = \quad freq(\{map(y)|y \in d\}) \tag{2}$$

where *freq* returns elements that are the most frequent elements in the list of given arguments and *map* is a function taking a year and returning its century. For example, to a query with the start date of 1690 and the end date of 1950, the 18th and 19th centuries are assigned.

From Fig. 17, we can see that century-to-century similarities tend to decrease towards the present. This result indicates that users may tend to retrieve similar data from very distant past whereas the queries associated with the present tend to be more diverse. This could be due to the larger size and more diversity of content from more recent history. Finally, we list top-20 common queries for different centuries in Tab. 7 which were collected from `dc.creator`, `dc.contributor`, `gallica` and free text input. To save space, Tab. 7 shows the data for only the top-4 centuries.

5 LIMITATIONS

Input Words vs. Clicks In this work, we have not investigated how the queries were input. Some of the queries may originate as a result of users clicking on recommended items displayed on the front page of the Gallica portal (this could explain why some queries seem bit 'artificial', e.g., being too long or having numbers mixed with characters). "ABCdaire1" (cf. Tab. 5) query is such an example of prewired queries in Gallica. Furthermore, Gallica provides various APIs which could be the reasons for some of such queries. We left their determination and analysis as a future work.

Data Collection Time Span. We note that the data collection we used is a portion (15 months) of all logs published in Gallica. Subsequent work should investigate portions of data collected over different time frames in order to verify which of the results are specific to the particular time frames of data generation and which are of general character. Once we can obtain data generated in different durations, we plan to perform such comparative investigation.

Different Languages. As Gallica is an online portal, anyone can use it. In this paper, we implicitly made an assumption that all queries are in French and hence we have applied tools dedicated to French language. While the large majority of queries are indeed in French, naturally, there are ones in other languages, too. Further exploration should then involve different languages (e.g., English, Italian) that Gallica supports as well as the cross-comparison of the obtained results. Also, geographical analysis is another interesting exploration.

6 CONCLUSIONS & FUTURE WORK

Cultural heritage collections are commonly used for storing, preserving as well as finding and learning information related to the past. This wealth of data is then source of interaction and search from many ordinary users.

In this paper we have studied how users search content in cultural heritage collections based on data collected from Gallica portal of the French National Library. The data we used is of large scale spanning 15 months. We mainly concentrated on entity analysis looking into the types of entities in queries and on the way in which temporal filters are associated with search queries. We have also analyzed which metadata filters and their combinations are commonly used.

Our analysis provides observations that can complement existing investigations of CH collection search thanks to the novel focus on semantic (entity) and temporal angles. It can also lead towards search experience improvement. For example, the examination of searched entities suggests which entity types tend to be popular, and so, the results related to these kinds of entities could be then recommended to visitors. We have also observed several temporal landmarks in temporal filtering (e.g., ones every 100, 50 or 10 years depending on a century). This might indicate that users first view a big picture concerning the results of their queries and likely do

Table 7: Top-20 queries of `dc.creator`, `dc.contributor` and free text. We translated some of French words into English when the meaning was not immediately clear. "*" mark indicates words translated into English. These are: *tale (conte), *Peru (Pérou), *impiety (impiété), *America (Amérique), *Algeria (Algérie), *housewife (ménagère), *laundres (blanchisseuse), *wash house (lavoir), *sex education (éducation sexuelle), *Comoros (Comores), *peyoye (peyotl).

Rank	17th	18th	19th	20th
1	*tale (471,900)	*impiety (649,200)	*Algeria (1,291,398)	*Algeria (1,510,692)
2	boussac (314,700)	*tale (471,900)	ancien régime (1,215,000)	*Comoros (489,600)
3	*America (270,300)	guinguette (368,640)	*housewife (690,300)	madagascar (489,600)
4	petite fleur bleue (260,700)	boussac (314,700)	*laundress (690,300)	avignon (444,960)
5	alpilles (251,100)	tunnel (272,400)	*wash house (690,300)	département d'alger (311,922)
6	tristan l'hermite (245,210)	*America (270,300)	*sex education (303,300)	province d'alger (311,922)
7	Luther (194,760)	petite fleur bleue (260,700)	boussac (294,219)	province de constantine (288,672)
8	*Peru (182,100)	alpilles (251,100)	tunnel (272,400)	département de constantine (288,672)
9	*impiety (170,415)	buffon, georges-louis leclerc (202,504)	département d'alger (266,643)	*peyoye (261,429)
10	buffon, georges-louis leclerc (160,202)	thomery (199,359)	province d'alger (266643)	afrique occidentale française (256534)

not have sufficient knowledge on how to setup the correct dates. It might also be a user tactic to avoid returning of too large result list. [9] Hence, for clear and concrete queries, it might be helpful to automatically support setting temporal filtering. For example, given a name of a searched entity such as an artist, the search engine could recommend filtering years when he/she mainly worked or when remained active.

Future work will investigate in more detail *user behavior: how many times and what words the same user inputs, how the users change parameters (e.g., temporal filtering) during search sessions, what are characteristics of failed/successful search sessions* and so on. Due to the time and space constraints, we have skipped the user (session)-focused analysis in this work; however, they are interesting directions and should be explored to better understand user search patters for potential improvement of user experience, creating recommendation/support systems, and for other objectives. Finally, we wish to undertake the study of how external events affect the choice of search keywords in CH search as well as we will investigate in detail how user queries correlate with the content provided in the collections and with the users' knowledge of this content.

Acknowledgments. This work has been supported by the European Union's Horizon 2020 research and innovation programme under grant 770299 (NewsEye), in part by MEXT Grant-in-Aid (17K12792, 17H01828, 18H03243) and by MIC SCOPE (171507010).

REFERENCES

[1] C.-m. Au Yeung and A. Jatowt. 2011. Studying How the Past is Remembered: Towards Computational History Through Large Scale Text Mining. CIKM '11, Glasgow, Scotland, UK, 1231–1240.

[2] A. Broder. 2002. A taxonomy of web search. In *ACM Sigir forum*, Vol. 36. ACM, 3–10.

[3] D. Ceccarelli, S. Gordea, C. Lucchese, F. M. Nardini, and G. Tolomei. 2011. Improving europeana search experience using query logs. Springer, 384–395.

[4] A. Chardonnens, E. Rizza, M. Coeckelbergs, and S. v. Hooland. 2017. Mining User Queries with Information Extraction Methods and Linked Data. *CoRR* abs/1709.07782 (2017). arXiv:1709.07782 http://arxiv.org/abs/1709.07782

[5] F. Clavert, B. Majerus, and N. Beaupré. 2015. #ww1. Twitter, the Centenary of the First World War and the Historian.

[6] J. Cook, A. D. Sarma, A. Fabrikant, and A. Tomkins. 2012. Your Two Weeks of Fame and Your Grandmother's. WWW '12, Lyon, France, 919–928.

[7] F. D.-Buc, E. Bermès, A. L. M.-Rieux, C. Prieur, V. Beaudouin, P. Chevallier, A. Nouvellet, and F. Roueff. 2017. *Analysis of Gallica and Data BnF logs and Modelling of Behaviour Patterns: Presentation of the Main Results.* Technical Report. [Research Report] Bibliothèque nationale de France (Paris); Télécom ParisTech.

[9] However, we need to keep in mind that these may be exploratory queries without a particular search intent hence this phenomenon needs to be studied more in future.

[8] M. D. Wilde and S. Hengchen. 2017. Semantic enrichment of a multilingual archive with linked open data. *Digital Humanities Quarterly* (2017).

[9] G. Demartini, C. S. Firan, M. Georgescu, T. Iofciu, R. Krestel, and W. Nejdl. 2009. An Architecture for Finding Entities on the Web. In *2009 Latin American Web Congress, Joint LA-WEB/CLIHC Conference, Merida, Yucatan, Mexico, 9-11 November 2009*. 230–237.

[10] M. Ferron and P. Massa. 2011. Collective Memory Building in Wikipedia: The Case of North African Uprisings. WikiSym '11, Mountain View, California, USA, 114–123.

[11] R. G.-Gavilanes, A. Mollgaard, M. Tsvetkova, and T. Yasseri. 2017. The memory remains: Understanding collective memory in the digital age. *Science Advances* 3, 4 (2017).

[12] J. Guo, G. Xu, X. Cheng, and H. Li. 2009. Named Entity Recognition in Query. SIGIR '09, New York, NY, USA, 267–274.

[13] A. Jatowt, D. Kawai, and K. Tanaka. 2016. Digital History Meets Wikipedia: Analyzing Historical Persons in Wikipedia. JCDL '16, Newark, New Jersey, USA, 17–26.

[14] D. Jiang, J. Pei, and H. Li. 2013. Mining search and browse logs for web search: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)* 4, 4 (2013), 57.

[15] N. Kanhabua, T. N. Nguyen, and C. Niederée. 2014. What Triggers Human Remembering of Events?: A Large-scale Analysis of Catalysts for Collective Memory in Wikipedia. JCDL '14, London, United Kingdom, 341–350.

[16] C. Lown, T. Sierra, and J. Boyer. 2013. How users search the library from a single search box. *College & Research Libraries* 74, 3 (2013), 227–241.

[17] J.-P. Moreux and G. Chiron. 2017. *Image Retrieval in Digital Libraries - A Large Scale Multicollection Experimentation of Machine Learning techniques.* Technical Report. IFLA News Media Section, Dresde, Germany.

[18] Bibliothèque nationale de France. 2017. *Enquête auprès des usagers de la bibliothèque numérique Gallica.* Technical Report. https://multimedia-ext.bnf.fr/pdf/mettre_en_ligne_patrimoine_enquete.pdf

[19] M. Paşca. 2007. Weakly-supervised discovery of named entities using web search queries. ACM, CIKM'07, 683–690.

[20] M. Sfakakis and S. Kapidakis. 2002. User behavior tendencies on data collections in a digital library. Springer, TPDL'02, 550–559.

[21] Y. Sumikawa, A. Jatowt, and M. Düring. 2018. Digital History Meets Microblogging: Analyzing Collective Memories in Twitter. JCDL '18, New York, NY, USA, 213–222.

[22] A. Tonon, M. Catasta, G. Demartini, P. C.-Mauroux, and K. Aberer. 2013. TRank: Ranking Entity Types Using the Web of Data. In *The Semantic Web - ISWC 2013 - 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21-25, 2013, Proceedings, Part I.* 640–656.

[23] A. Tonon, V. Felder, D. E. Difallah, and P. C.-Mauroux. 2016. VoldemortKG: Mapping schema.org and Web Entities to Linked Open Data. In *The Semantic Web - ISWC 2016 - 15th International Semantic Web Conference, Kobe, Japan, October 17-21, 2016, Proceedings, Part II.* 220–228.