Building Cultural Heritage Reference Collections from Social Media through Pooling Strategies: The Case of 2020's Tensions Over Race and Heritage

DAVID OTERO, PATRICIA MARTIN-RODILLA, and JAVIER PARAPAR, Information Retrieval Lab, Centro de Investigación en Tecnoloxías da Información e as Comunicacións (CITIC), Universidade da Coruña, Spain

Social networks constitute a valuable source for documenting heritage constitution processes or obtaining a real-time snapshot of a cultural heritage research topic. Many heritage researchers use social networks as a social thermometer to study these processes, creating, for this purpose, collections that constitute born-digital archives potentially reusable, searchable, and of interest to other researchers or citizens. However, retrieval and archiving techniques used in social networks within heritage studies are still semi-manual, being a time-consuming task and hindering the reproducibility, evaluation, and open-up of the collections created. By combining Information Retrieval strategies with emerging archival techniques, some of these weaknesses can be left behind. Specifically, pooling is a well-known Information Retrieval method to extract a sample of documents from an entire document set (posts in case of social network's information), obtaining the most complete and unbiased set of relevant documents on a given topic. Using this approach, researchers could create a reference collection while avoiding annotating the entire corpus of documents or posts retrieved. This is especially useful in social media due to the large number of topics treated by the same user or in the same thread or post. We present a platform for applying pooling strategies combined with expert judgment to create cultural heritage reference collections from social networks in a customisable, reproducible, documented, and shareable way. The platform is validated by building a reference collection from a social network about the recent attacks on patrimonial entities motivated by anti-racist protests. This reference collection and the results obtained from its preliminary study are available for use. This real application has allowed us to validate the platform and the pooling strategies for creating reference collections in heritage studies from social networks.

CCS Concepts: • **Information systems → Information retrieval**; • **Applied computing → Arts and humanities**; • **Human-centered computing → Collaborative and social computing theory, concepts and paradigms;**

Additional Key Words and Phrases: Reference collections, social networks, pooling

This research has received financial support from: (i) Saving European Archaeology from the Digital Dark Age (SEADDA) 2019–2023 COST ACTION CA 18128; (ii) "Ministerio de Ciencia, Innovación y Universidades" of the Government of Spain and the ERDF (projects RTI2018-093336-B-C21 and RTI2018-093336-B-C22); (iii) Xunta de Galicia—"Consellería de Cultura, Educación e Universidade" (project GPC ED431B 2019/03); (iv) Xunta de Galicia—"Consellería de Cultura, Educación e Universidade" and the ERDF ("Centro Singular de Investigación de Galicia" accreditation ED431G 2019/01).

Authors' address: D. Otero, P. Martin-Rodilla, and J. Parapar, Information Retrieval Lab, Centro de Investigación en Tecnoloxías da Información e as Comunicacións (CITIC), Universidade da Coruña, Facultad de Informática, Campus de Elviña s/n, A Coruña, Spain, 15071; emails: {david.otero.freijeiro, patricia.martin.rodilla, javier.parapar}@udc.es.

1556-4673/2021/12-ART9 $15.00

https://doi.org/10.1145/3477604

ACM Reference format:
David Otero, Patricia Martin-Rodilla, and Javier Parapar. 2021. Building Cultural Heritage Reference Collections from Social Media through Pooling Strategies: The Case of 2020's Tensions Over Race and Heritage. *J. Comput. Cult. Herit.* 15, 1, Article 9 (December 2021), 13 pages.
https://doi.org/10.1145/3477604

1 INTRODUCTION

As an inherently social phenomenon, *"patrimonialization"* (understood as "a process by which a material or immaterial element becomes a constitutive part of a community's identity that imbues said element with meaning and significance" [43]) includes the processes of construction, identification, rejection, or destruction of heritage, among others. These processes are reflected in different areas of our social reality, including virtual communities, within the so-called social networks.

The study of social processes (in this case, concerning cultural heritage) has focused on the analysis of the information generated in social networks [22], where all kinds of people create content in various formats (comments and posts in textual format but also all kinds of multimedia or extended information such as photos, videos, links to related content or other types of content). This content is being studied from an archival perspective (with related conceptual, methodological, and disciplinary implications), where it is possible to find numerous examples of born-digital archives from social networks [39, 42, 47].

Resultant born-digital archives play an essential role for cultural heritage researchers in this context, since these archives constitute a primary source in the empirical study of *"patrimonialization"* processes. These archives could be potentially reusable, searchable, and, as a real-time snapshot of the process studied, of interest to other researchers or citizens [42].

However, and as will be seen throughout this article, retrieval and archiving techniques used in social networks within heritage studies have some problems: (i) they are still carried out manually or, (ii) carried out using computer techniques not adapted to the informational domain or the final purpose of the information, which is ultimately the creation of archives or collections as a reflection, preservation, and source for the study of social processes.

With the transdisciplinary perspective offered by Computational Archival Science [38], and after a review of the intersection between **Information Retrieval (IR)** and the creation of archives from social networks, we present a platform for the application of IR strategies (combined with the judgment of experts) for the creation of reference collections (as born-digital archives) in cultural heritage from social networks in a customisable, reproducible, documented, and shareable way.

In addition, the platform is evaluated through a real case study: the creation of a reference collection from Reddit [41] on the recent attacks on patrimonial entities motivated by anti-racist protests. This reference collection and the results obtained from its preliminary study are freely available for use and already constitutes a born-digital archive that will allow future researchers to have real information about citizen's motivations and comments, the different social sensitivities concerning these attacks on heritage entities, as well as references to places or heritage entities under attack throughout the protests.

With this double contribution (the platform and the resultant collection), the article is structured as follows: Section 2 offers an overview of the two areas of knowledge necessary to address the problem of creating reference collections (as born-digital archives) from social networks. In the first part, related work based on Computational Archival Science is explained in detail, focusing on existing works on born-digital archiving from social networks and the information retrieval problems. In the second part, we contextualise the Information Retrieval discipline, explaining how some selected techniques (i.e., pooling) constitute a sound basis for creating born-digital reference collections from social media to study cultural heritage social processes. Section 3 presents the **BeaverColNet platform** in depth, further documenting the complete methodology for creating a born-digital

archive using the platform. Section 4 details the complete process in a real case study, presenting the resultant collection to study racial tensions and their reflection in attacks on patrimonial entities in 2020. Section 5 critically discusses the approach, presents the conclusions, and Section 6 summarises future work lines.

2 BACKGROUND

2.1 Computational Archival Science, Born-Digital Archives, and Social Media

There is an intuitive connection, as knowledge areas, between Information Retrieval, which deals with the representation, storage, organisation of, and access to information items from a computational point of view [5], and the archival practices, understood as the knowledge area related to archives processing, access, analysis, storage, and long-term preservation [26]).

As a result of the identification, characterisation, and development of this kind of connection, the concept of Computational Archival Science has emerged. It is a "transdisciplinary field that integrates computational and archival theories, methods, and resources, both to support the creation and preservation of reliable and authentic records/archives and to address large-scale records/archives processing, analysis, storage, and access, with aim of improving efficiency, productivity and precision, in support of recordkeeping, appraisal, arrangement and description, preservation and access decisions, and engaging and undertaking research with archival material" [26].

A large volume of work and initiatives have been developed in recent years around the concept of Computational Archival Science from a transdisciplinary perspective, combining techniques, tools, and methodologies that expand disciplines and improve the treatment of large-scale records. Good examples are Computational Archival Science Workshops [55] or similar Big Data events in this domain [54]. Also, Computational Archival Science as a field (its definition and development) has required analysis and methodological and applied contributions by the disciplines that compose it, as can be seen in examples of projects with different sources: textual [49], cartographic-based [23, 49], or even from multimedia archives [18], as well as in recent compendia and theoretical studies on the discipline [23, 37].

This transdisciplinary connection becomes crucial when we manage born-digital archives, where each record or item presents an intrinsic digital nature (like a post within a social network). Thus, we can find completely born-digital archives: (i) seeing the web as a source for these archives [1, 42], and (ii) also from social networks, with approaches based on curating life shared experiences: a Google patent system [19], works on extracting personal archives from social networks [2], and more general approaches from massive social networks such as Youtube [14] or Facebook [40].

Additionally, this kind of works supports in some way the records continuum theory [50]. The upward theory establishes that any record has to be managed at an archival level without strict phases or protocols since each record's creation. This vision of archiving as a continuum flow is beneficial when we work with social network information. Social network retrieval (and the creation of reference collections from them), seeing as born-digital archiving, allows us to analyse and apply records continuum theory.

However, all these works present some problems in the retrieval techniques applied for building reference collections. We have detected the following:

- Some of them are retrieved manually or semi-manually [47], including automatic retrieval but with a second annotation phase in which experts have to annotate the entire collection. Although it is possible to adopt this approach in small projects or initiatives, these information retrieval techniques are extremely time-consuming and hinder the reproducibility, evaluation, and open-up of the collections created.
- Some of them are retrieved using computer techniques not adapted to the informational domain or the final purpose of the information [31, 53].
- Most of them are not dealing with privacy and security issues, such as anonymisation processes in the information retrieval workflow [28, 33], with also implications for cultural heritage archives [22].

- Most of them are one-case studies, in which the information retrieval techniques and software tools applied [6] and, in some cases, the resultant collections are not available for further use and applications [47].

In summary, there is a gap between IR techniques and the current work on building reference collections (as born-digital archives) from social networks. The following section contextualises IR techniques and our proposal for applying them in bridging this gap, solving some of the previous problems identified and allowing cultural heritage researchers to create reference collections from social networks.

2.2 Forming Collections With Pooling

Information Retrieval is a data-intensive discipline where strict evaluation procedures are needed to evaluate new models' development. The evaluation aspect has been extensively developed, with numerous improvements in recent years [11]. The most common approach of performing this evaluation is to use test collections that contain a set of documents (posts in social networks), topics representing the users' information needs, and judgments that capture the relevance relation between each document-topic pair [46]. To obtain this ground truth, a group of assessors judge which documents are relevant to each topic. Then it is possible to run the queries from the topics against a retrieval model to obtain a ranked list of documents (posts in social networks) and evaluate it using the relevance judgments from the collection. This paradigm is the *de facto* standard to perform IR evaluation [51].

Tests collections allow researchers to compare different retrieval methods effectively. However, building new benchmarks is difficult and expensive: manually assessing each document-topic pair's relevance is time-consuming. In the early days of IR, document sets were small, and judging the entire corpus was approachable. Nowadays, large-scale IR evaluation is prohibitively expensive due to the size of modern test collections. This is a general problem in IR that has also been identified in Cultural Heritage collections and archival environments [12]. Furthermore, judging the entire corpus would cause the assessors to spend their time assessing non-relevant documents, which is a waste of effort. This is the reason why these benchmarks are built using a process called **pooling**.

When building pooled test collections, assessors only judge a subset of the whole document corpus, thus significantly reducing the cost of obtaining new judgments. This is crucial to reduce the time-consuming tasks of judgement in social networks' retrieval environments. The subset of documents/posts that are judged is called a *pool*. The development of strategies (also known as adjudicating methods) to efficiently select which documents merit humans judgments is an area that has attended much attention in IR research [4, 7, 24, 25, 30].

Developing strategies to select which items merit human judgments from a pool of unlabelled items is a crucial topic well beyond Information Retrieval. Since it is a very mature area in this field, we propose to use it to build reference collections from social networks while limiting the effort needed to annotate the documents. We can bridge the gap between IR and the current work on building reference collections from social networks with this technique. We aim to reduce and alleviate some of the burdens of creating new collections, as explained in Section 2.1.

In this article, we apply pooling (which is extensively used in IR) to build archival reference collections. Using pooling, we can significantly reduce the number of posts annotated by the experts with respect to the whole collection size. A recent work presented Beaver [35], a platform to ease the building of new test collections for IR evaluation. This platform implements state-of-the-art pooling strategies to select the documents to annotate and simulate the rankings used to pool the documents. We adapt this platform to build (and release) reference collections. In the next section, we present the platform's workflow, its functionality, and the adaptations we have made to it.

3 BEAVERCOLNET PLATFORM

Beaver [34, 35] is a platform aimed to ease the process of building a new reference collection (test collection in the information retrieval approach) by providing simulated participant systems and state-of-the-art pooling

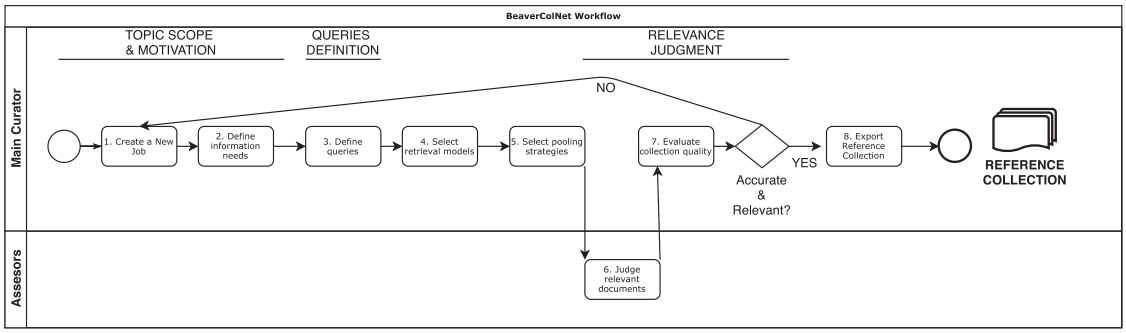

Fig. 1. BeaverColNet platform workflow.

Fig. 2. Creation of a new experiment: topic and queries.

strategies that significantly reduce the effort of obtaining new annotations. In the following paragraphs, we show the platform's workflow, and we explain the adaptations made to build reference collections for cultural heritage, specifically in our case study about the recent attacks on patrimonial entities motivated by anti-racist protests.

3.1 Workflow

The workflow of the platform is depicted in Figure 1. This figure shows that the functionality is divided into two different roles: Main Curator and Assessor. The system allows us to create, configure, and execute experiments, resulting in a different new collection. That is, each experiment represents the process of creating a different reference collection. We now explain the tasks to build a new collection that corresponds to each role.

3.1.1 Main Curator. The main curator is the person responsible for the whole experiment. This curator has to establish the main topic and scope of an experiment, representing the collection's information need. Once the topic is set, the first step to build the collection is to retrieve the posts that comprise the document set. The platform currently only works against the Reddit API to download posts from this social network. However, we plan to extend the document sources, e.g., other social networks. So, in this case, a document is a Reddit post. Reddit is a social network very suitable for our use case: to build a reference collection on the recent anti-racist protests and patrimonial attacks because of the threaded nature of its posts. The curator should derive a set of different query variants from the collection's main topic to obtain the posts. These queries are searched against the API to download the documents. Figure 2 shows an example of a set of queries for downloading documents.

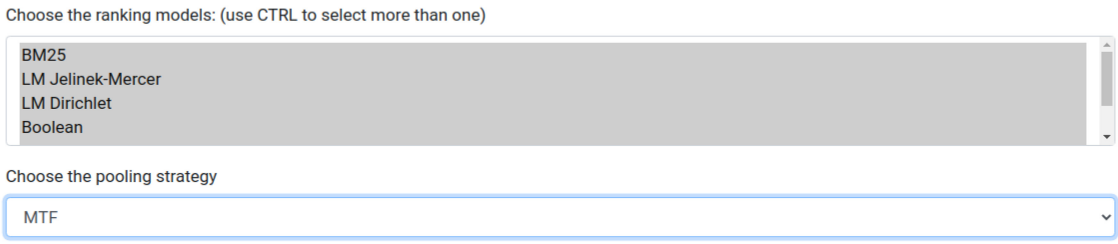

Fig. 3. Creation of a new experiment: ranking models and adjudicating methods.

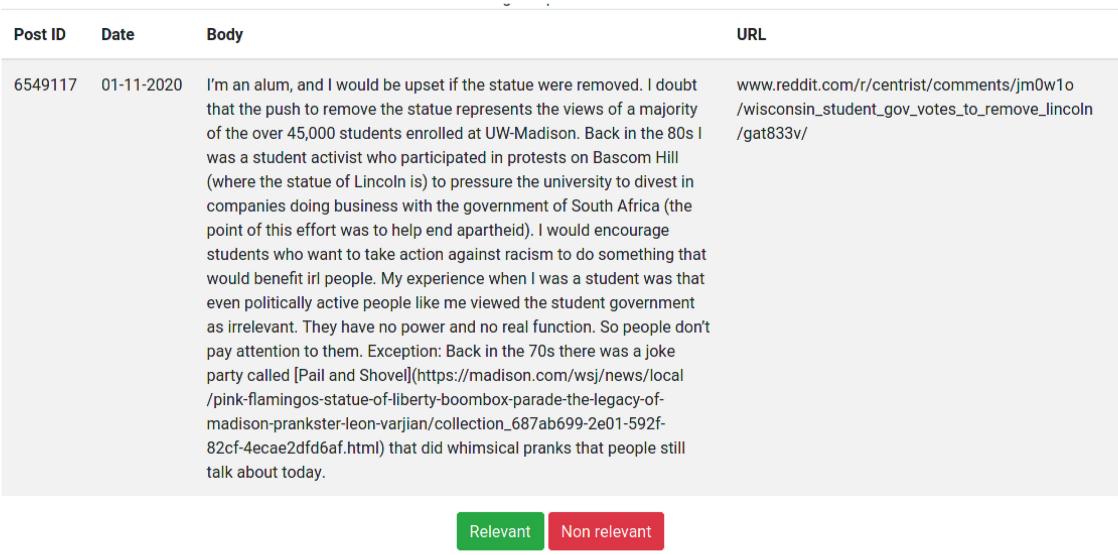

Fig. 4. BeaverColNet judging page.

The next step in the configuration of an experiment is to select the retrieval models and pooling methods used to pool the documents that merit human judgments. This selection is illustrated in Figure 3. Right now, the platform offers two adjudicating methods: DocID [52] and MTF [7].

Once the experiment is configured, the platform downloads the documents from Reddit. Then, the judging phase begins.

3.1.2 Assessors. The assessors may be one person or a group of persons that should know the domain and topic of the collection, so they will be able to judge the relevance of the downloaded documents that will be part of the final reference collection. In our case, assessors should be experts who work on cultural heritage attacks, for example. The primary and only work of the assessors is to judge the relevance of each downloaded documents (Reddit posts). This assessment should be according to the main topic and information needs of the experiment.

Figure 4 illustrates what is presented to the assessor. In this figure, the assessor sees the post being judged and two buttons to decide its relevance. We can see that, when judging a post, the platform also shows the date when it was published, the posts' content, and a link referring to the original Reddit page. Once all judgments are done, the pooling process is finished, and the collection can be exported and released for further use.

3.2 BeaverColNet Adaptation

BeaverColNet is an adaptation of the original platform, Beaver, for applying pooling strategies in building reference collections from social networks in cultural heritage. This implies the technological and domain adaptation of the platform and an effort to rethinking information retrieval workflow with archival and cultural heritage roles, goals, and possible applications. We have redesigned the platform's architecture at a technological and domain level to adapt it to build reference collections in cultural heritage. This implies changes in the retrieval workflow (how the platform retrieves posts from Reddit). It also implies modifications in the judging phase, since what is presented to the expert annotators is not the same as in the original platform. Finally, we have also changed how the collection is exported. Thus, as main curators and with the assessor's help, cultural heritage researchers can create now their own reference collections from social networks for studying *"patrimonialization"* processes and specific cultural heritage events.

4 A BEAVERCOLNET REAL CASE: A REFERENCE COLLECTION ON 2020'S TENSIONS OVER RACE AND HERITAGE

In this section, we explain the method we have followed to build a new reference collection by explaining the workflow illustrated in Figure 1. First, we explain our motivation and the scope of the collection. Then, we show how we have derived the queries used in the experiment. Next, we explain the judging process. Finally, we give some insights and information about the generated reference collection.

4.1 Topic Scope and Motivation

For illustrating the entire BeaverColNet workflow and as a validation of the platform, we have created a reference collection from Reddit about the recent attacks on patrimonial entities motivated by anti-racist protests.

During the year 2020, there have been many protests and attacks on heritage elements such as statues and commemorative plaques worldwide. All these revolts began due to the death of George Floyd in the United States. This event gave rise to a series of protests in which various cultural elements commemorating important figures in history were attacked under the motivation that the figures represented were racists and genocides.

This connection between heritage and racism has been widely studied previously, as well as its connection with the events that occurred in 2020, from different disciplinary perspectives, such as philosophy [3, 9, 44], history and anthropology [15, 20], sociological [48], or archaeological and heritage studies [29]. It is also possible to find analysis of the phenomenon within the so-called heritage in social conflict or social fractures in heritage, as detailed by anthropological studies in the area [8, 45].

Therefore, the patrimonial attacks that occurred in 2020 are events that reflect social and patrimonial processes of a very diverse nature that, in our opinion, require the existence of reference collections that allow the subsequent research on this type of issues from different methodologies, fields and contributions.

2020's protests and heritage attacks events also encouraged the appearance in social networks of people who comment and discuss their opinions on these issues. Thus, we want to build a new reference collection with these publications to serve as a social thermometer to study this "patrimonialization" process.

As we have previously shown, this topic is probably one of the current phenomena related to patrimonial entities that receive the most interest in public opinion and media and researchers in different disciplines (see References [3, 9, 20, 44, 48] among others). It has also been shown the critical role of social networks in influencing public opinion (and electoral processes in progress in USA), as a call for action instrument (both to attack cultural heritage entities and to manifest against attacks) and in the organisation of related platforms and collectives. The preservation of consistent information on social networks on 2020's racial tensions and heritage attacks and the possibility of his later study from different points of view seems to us a great motivation to validate and illustrate BeaverColNet application to a real case study.

Table 1. Lists of Terms for Both Groups

Conflicts related terms	Entity related terms
anti-black	monument
blacklives	removed
blacklivesmatter	removal
police brutality	statue
polive violence	tribute
abuse of authority	memorial
racism	plaque
racial bias	bust
anti-racism	take down
george floyd	beheaded
slavery	desecrated
slave	vandalized
-	vandalism
-	vandals
-	protests
-	protesters

Terms in the same row does not mean we have used them together.

4.2 Queries Definition

To find posts relevant to this topic, we wanted to curate queries that included reference to a patrimonial entity and the citizen's opinion about it. For this reason, we created two lists of terms. The first list included terms related to patrimonial entities, such as *monument*, *memorial*, *plaque*, *bust*, *statue*, *tribute*, among others. The second list comprised useful terms to find people's opinion on the anti-racist protests and conflicts, such as *racism*, *slavery*, *slave*, *black lives*, *violence*, among others. The queries we have used contained terms from both lists, so that the posts that match the queries are related to our information need. In Table 1, we show the entire set of terms of each group.

We did initial research to find the terms and queries that retrieved the more precise and accurate comments on this topic. This initial research consisted of creating, with BeaverColNet, small experiments, one for each query, and see if the retrieved posts were relevant. Finally, the set of queries that retrieved the more relevant documents were the following: "statue racism," "monument racism," "police violence statue," "slave plaque," and "black lives monument." Thus, these are the five queries we have used to build this collection.

It is important to note that BeaverColNet allows us to replicate the entire collection building process and expand the set of terms that constitutes queries in a new experiment. Thus, it could be possible to use existing thesauri or similar linguistic resources to expand the queries done in the future.

4.3 Relevance Judgment

These selected queries resulted in 522 (the size of the pool) different post to judge. A group of three assessors made the judgments, where each one judged approximately the third part of the posts. As we said before, the

assessors should be familiar with the topic of the collection to provide standardised and high-quality judgments. In this case, all assessors have worked before in cultural heritage contexts and entities in conflict [17, 27]. Also, the criteria to decide the relevance of the posts were made a priori before the assessment process. This was made to mitigate any personal bias that can be introduced in the collection and provide the assessors with a simple guide on deciding about the relevance of the documents. Annotators, as seen in Figure 4, were presented with one post at a time. They only annotated as relevant posts that contained references to patrimonial entities, such as statues or plaques, which also contained the writer's opinion about the attacks and protests. This process took one week long.

Once the judgments were made, the platform downloaded the content of the threads that contained the relevant posts. The content of these threads comprises the final reference collection.

4.4 2020's Tensions over Race and Heritage Collection

We have applied the workflow shown in Figure 1 for the reference collection construction. Thus, the resultant reference collection contains Reddit information judged as positive by the assessors. The final reference collection consists of **296 Reddit** threads extracted from posts judged as positive. This means that, as a result of our querying and judgment processes, the posts judged as positive come from 296 different threads, with a total of **260,578 posts** in the reference collection (that gives us an average of 880,331 posts incorporated in the reference collection per thread). The posts were written between 25th May, 2020 (date of death of George Floyd) and 31st October, 2020. Note that it is necessary to judge only **522 posts** during the judgment phase, avoiding the complete annotation of the reference collection, for obtaining almost six months of Reddit activity and information about Lloyd's movement and anti-racism related protests.

The resultant reference collection is available at this link,[1] constituting the first reference collection, as far as we know, that preserve and give access for the study of the recent attacks on patrimonial entities motivated by anti-racist protests.

It is important to note that this high number of different threads ensures that BeaverColNet allows to retrieve from Reddit and incorporate to the reference collection posts that dealt with the issue of protests and their relationship with attacks on patrimonial entities, not only from explicit threads with that topic but also from many threads with different central topics. This is important to ensure coverage of the topic within the social network, not only recovering those conversational threads focused on the topic (which may reflect points of view that are excessively polarised or directed by associations or people directly implicated in the conflicts). In this way, we are also reaching threads with different topics than the main topic where the conversation has turned, dealing with protests at some point. These conversations may be less polarised and include citizens with more diverse profiles and responsibilities.

2020's Tensions over Race and Heritage Collection, as a real application, has allowed us to validate the platform and the pooling strategies for creating reference collections in heritage studies from social networks.

In addition to this reference collection, we also release the history published by a sample of users that participated in those threads. In total, from the 296 threads that comprise the main collection, we found more than 90,000 users participating in them. From these users, we sampled 1,400 of them and retrieved the whole history of posts published by each one. This resulted in a collection with 6,455,258 (including the content of the threads) different posts. We hope that this will help to research the sensitivities of those users with respect to patrimonial attacks and a more deep investigation into those users' profiles. In Table 2, we present a summary of both collections.

To avoid revealing private or personal information about the users, we have anonymised them by substituting their original Reddit username with a randomly generated identifier. We are aware that only hidding its original username in Reddit may not be sufficient to cloak their identification. However, we must note one important

[1] https://www.dc.fi.udc.es/~david/heritage.

Table 2. Summary of Released Collections

Collection	# posts
Only threads	260,578
Full collection	6,455,258

aspect here. All the data that we have crawled to create this collection is public data that was made available by Internet users. We have not gathered any private or personal information of the users. What we must point out is that if personal information of the users is retrieved from the use of this collection, it must be treated following practices that ensure their anonymity [21].

5 DISCUSSION AND CONCLUSIONS

In this article, we analyse the intersection between Information Retrieval and Computational Archival Science in the specific case of reference collections (as born-digital archives) from social networks. Social networks have become a real-time reflection of social processes, and several researchers use social network information for studying cultural heritage processes. From our particular interest in creating born-digital archives from social networks from a flexible paradigm and as a dynamic, continuum, and transdisciplinary archive, the article presents the BeaverColNet platform. BeaverColNet enables innovative pooling-based information retrieval strategies combined with expert judgment to create reference collections from social media. Besides, the platform is evaluated in a real case study on cultural heritage, with the creation of a born-digital reference collection from Reddit that retrieves, monitors, documents, preserves, and allows the evolutionary study of the phenomenon of attacks on heritage entities in the anti-racism protest of 2020 around the world.

The resulting collection consists of more than 260,000 relevant posts with all kinds of opinions, visions, and attitudes towards the racial conflict and the heritage involved, in different areas of the world and by very different people's profiles. This constitutes a born-digital archive about the attacks suffered by patrimonial entities and the activity in social networks generated around these attacks and the anti-racist riots in 2020. Information retrieval in real-time and the possibility that BeaverColNet offers to carry out different retrieval processes that increase the collection over time will allow us to retrieve the activity in the Reddit social network about these attacks and their relationship with the riots as a testimony of these cultural heritage events through time.

Making a critical analysis of the work carried out, we want to highlight two points. First, the work carried out so far takes the Reddit social network as a source of information for the collections, although the methodology and the BeaverColNet platform can be generalised to other social networks. Therefore, we should have in mind that the resulting collections, including the one presented here, have the implicit bias of the social network itself, requiring analysis of demographic profiles and even treatment regarding their balance to try to minimise the implicit bias. Several authors have recently dealt with this topic from IR [16, 36], with different methods for adjusting results and minimising bias [10, 32]. Therefore, we recommend that both in the use of the collection and of the platform and methodology, these issues should have taken into account from the very beginning of the archival project and from a transdisciplinary perspective, an approach that is naturally taken in Computational Archival Science [26, 37].

Second, it is important to note that, although we believe that BeaverColNet platform solves some of the problems detected regarding the application of information retrieval techniques (see Section 2) on this kind of works, it does not intend to automate the task of creating reference collections (as born-digital archives) from social networks. We are aware that the improvements in the technological approach that BeaverColNet offers must be applied from a transdisciplinary paradigm [13], where experts in the domain (in this case Cultural Heritage) and experts in archival science can design their archiving workflows. In these contexts, BeaverColNet applies

innovative information retrieval techniques to the creation of these collections in these contexts, avoiding full expert annotation of collections, manual approaches impossible to maintain at a certain volume of data, and facilitating reproducibility and availability of the collections created.

6 FUTURE STEPS

As a first step, we plan to enrich the reference collection on tensions between racism and heritage in two directions, already allowed at a technological level by BeaverColNet:

(1) To enrich the collection with new data for the diachronic study of the conflicts and their relationship with heritage. This implies the preparation of new experiments in BeaverColNet in the coming months for capturing with temporal criteria, for example, the impact on the theme of current events (COVID-19 disease, for example). The new data sets will be added to the reference collection and will be freely available for use.
(2) To enrich the collection with information about the profiles of users of the social network whose posts are in our reference collection. To do this, we will add to the published reference collection all the posts in all the subreddits of each user who is involved in the current collection. This will allow us to have a personal history of activity on the social network of each of the users who expressed their opinion about racial tensions and attacks on heritage: what other topics interest them and what do they post, their interactions registry, expertise in the social network, and so on. This personal information is of complementary value for studies focused on people: What kinds of people have posted about tensions over race and heritage? What else can we know about them? The inclusion of user profile's information in the collection also will require work on privacy issues and bias correction of the retrieval results, following innovative methods [10, 21, 36].

We plan, together with researchers in aspects of racism and heritage connections within the work with the collection itself, to use natural language processing and text mining techniques on the collection to identify and computationally extract heritage sites or entities. This kind of analysis allows us to elaborate maps on where the attacks have been, which heritage entities or historical events involve, and so on.

In the future, we also plan to test the platform in various case studies in collaboration with researchers and institutions in cultural heritage. As we have seen in the article, intelligent pooling strategies that heavily reduce the assessor's work make this process a task less expensive in terms of computational resources, time consumed by the main curator, and assessor's and better availability of the resultant reference collections. This will also allow us to carry out empirical studies with all types of users of the platform, going more deeply in measuring the retrieval accuracy and quality of the resulting collections from the point of view of its main researchers and users, as well as satisfaction, usability, and detection of possible BeaverColNet's improvements.

REFERENCES

[1] Samer Abdallah, Emmanouil Benetos, Nicolas Gold, Steven Hargreaves, Tillman Weyde, and Daniel Wolff. 2017. The digital music lab: A big data infrastructure for digital musicology. *J. Comput. Cultur. Herit.* 10, 1 (2017), 1–21.
[2] Amelia Acker and Jed R. Brubaker. 2014. Death, memorialization, and social media: A platform perspective for personal archives. *Archivaria* 77 (2014), 1–23.
[3] Jason Arday. 2020. It's the end of the world as we know it: Racism as a global killer of black people and their emancipatory freedoms. *Edu. Philos. Theory* (2020), 1–3. https://doi.org/10.1080/00131857.2020.1782722
[4] Javed A. Aslam, Virgil Pavlu, and Robert Savell. 2003. A unified model for metasearch, pooling, and system evaluation. In *Proceedings of the 12th ACM International Conference on Information and Knowledge Management*. Association for Computing Machinery, New York, NY, 484–491. https://doi.org/10.1145/956863.956953
[5] Ricardo A. Baeza-Yates and Berthier Ribeiro-Neto. 1999. *Modern Information Retrieval*. Addison-Wesley Longman Publishing.
[6] Tobias Blanke, Michael Bryant, and Mark Hedges. 2013. Back to our data-experiments with nosql technologies in the humanities. In *Proceedings of the IEEE International Conference on Big Data*. IEEE, 17–20.

[7] Gordon V. Cormack, Christopher R. Palmer, and Charles L. A. Clarke. 1998. Efficient construction of large test collections. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, New York, NY, 282–289. https://doi.org/10.1145/290941.291009
[8] J. Cortés-Vázquez, G. Jimenez-Esquinas, and C. Sanchez-Carretero. 2017. Heritage and participatory governance: An analysis of political strategies and social fractures in Spain. *Anthropol. Today* 33, 1 (2017), 15–18. https://doi.org/10.1111/1467-8322.12324
[9] Adam Davidson-Harden. 2020. "I can't breathe": Praxis, parrhesia and the current historical moment. *Edu. Philos. Theory* (2020), 1–5. https://doi.org/10.1080/00131857.2020.1779580
[10] Michael D. Ekstrand, Robin Burke, and Fernando Diaz. 2019. Fairness and discrimination in retrieval and recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, New York, NY, 1403–1404. https://doi.org/10.1145/3331184.3331380
[11] Nicola Ferro and Carol Peters. 2019. *Information Retrieval Evaluation in a Changing World: Lessons Learned from 20 Years of CLEF*. Vol. 41. Springer.
[12] Jacob Foley, Paul Kwan, and Mitchell Welch. 2017. A web-based infrastructure for the assisted annotation of heritage collections. *J. Comput. Cultur. Herit.* 10, 3 (July 2017), 25 pages. https://doi.org/10.1145/3012287
[13] Weston Paul Gabriele, Emmanuela Carbé, and Primo Baldini. 2017. If bits are not enough: Preservation practices of the original contest for born digital literary archives. *Bibliothecae.it* 6, 1 (2017), 154–177.
[14] Robert Gehl. 2009. YouTube as archive: Who will curate this digital wunderkammer? *Int. J. Cultur. Studies* 12, 1 (2009), 43–60.
[15] Mandy Tompkins Gibson and Gabriel A. Reich. 2017. Confederate monuments: Heritage, racism, anachronism, and who gets to decide? *Soc. Edu.* 81, 6 (2017), 356–361. Retrieved from https://www.ingentaconnect.com/content/ncss/se/2017/00000081/00000006/art00007.
[16] Gisoo Gomroki, Hassan Behzadi, Rahmatolloah Fattahi, and Javad Salehi Fadardi. 2021. Identifying effective cognitive biases in information retrieval. *J. Info. Sci.* https://doi.org/10.1177/01655515211001777
[17] Cesar Gonzalez-Perez, Patricia Martín-Rodilla, Cesar Parcero-Oubiña, Pastor Fábrega-Álvarez, and Alejandro Güimil-Fariña. 2012. Extending an abstract reference model for transdisciplinary work in cultural heritage. In *Metadata and Semantics Research*. Springer, Berlin, 190–201.
[18] Hoda Hamouda, Jessica Bushey, Victoria Lemieux, James Stewart, Corinne Rogers, James Cameron, Ken Thibodeau, and Chen Feng. 2019. Extending the scope of computational archival science: A case study on leveraging archival and engineering approaches to develop a framework to detect and prevent "Fake Video." In *Proceedings of the IEEE International Conference on Big Data (BigData'19)*. 3087–3097. https://doi.org/10.1109/BigData47090.2019.9006170
[19] Jonathan Hull. 2008. System and method for creating online social-networks and historical archives based on shared life experiences. U.S. Patent App. 11/894,525.
[20] Jelani Ince, Fabio Rojas, and Clayton A. Davis. 2017. The social media response to black lives matter: how twitter users interact with black lives matter through hashtag use. *Ethnic Racial Studies* 40, 11 (2017), 1814–1830. https://doi.org/10.1080/01419870.2017.1334931
[21] Maryam Kiabod, Mohammad Naderi Dehkordi, and Behrang Barekatain. 2019. TSRAM: A time-saving k-degree anonymization method in social network. *Expert Syst. Appl.* 125 (2019), 378–396. https://doi.org/10.1016/j.eswa.2019.01.059
[22] Matthew Kirschenbaum, Richard Ovenden, Gabriela Redwine, and Rachel Donahue. 2010. Digital forensics and born-digital content in cultural heritage collections. https://www.clir.org/wp-content/uploads/sites/6/pub149.pd.
[23] Myeong Lee, Yuheng Zhang, Shiyun Chen, Edel Spencer, Jhon Dela Cruz, Hyeonggi Hong, and Richard Marciano. 2017. Heuristics for assessing computational archival science (CAS) research: The case of the human face of big data project. In *Proceedings of the IEEE International Conference on Big Data (BigData'17)*. 2262–2270. https://doi.org/10.1109/BigData.2017.8258179
[24] Dan Li and Evangelos Kanoulas. 2017. Active sampling for large-scale information retrieval evaluation. In *Proceedings of the ACM on Conference on Information and Knowledge Management*. Association for Computing Machinery, New York, NY, 49–58. https://doi.org/10.1145/3132847.3133015
[25] David E. Losada, Javier Parapar, and Álvaro Barreiro. 2017. Multi-armed bandits for adjudicating documents in pooling-based evaluation of information retrieval systems. *Info. Process. Manage.* 53, 5 (Sept. 2017), 1005–1025. https://doi.org/10.1016/j.ipm.2017.04.005
[26] Richard Marciano, Victoria Lemieux, Mark Hedges, Maria Esteva, William Underwood, Michael Kurtz, and Mark Conrad. 2018. Archival records and training in the age of big data. In *Re-Envisioning the MLS: Perspectives on the Future of Library and Information Science Education*. Emerald Publishing Limited.
[27] Patricia Martín-Rodilla, Marcia L. Hattori, and Cesar Gonzalez-Perez. 2019. Assisting forensic identification through unsupervised information extraction of free text autopsy reports: The disappearances cases during the brazilian military dictatorship. *Information* 10, 7 (2019). https://doi.org/10.3390/info10070231
[28] Jasmine McNealy. 2011. The privacy implications of digital preservation: Social media archives and the social networks theory of privacy. *Elon L. Rev.* 3 (2011).
[29] Lynn Meskell. 2002. Negative heritage and past mastering in archaeology. *Anthropol. Quart.* 75, 3 (2002), 557–574.
[30] Alistair Moffat, William Webber, and Justin Zobel. 2007. Strategic system comparisons via targeted relevance judgments. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, New York, NY, 375–382. https://doi.org/10.1145/1277741.1277806

[31] Devon Mordell. 2019. Critical questions for archives as (big) data. *Archivaria* 87, 87 (2019), 140–161.
[32] Fred Morstatter and Huan Liu. 2017. Discovering, assessing, and mitigating data bias in social media. *Online Soc. Netw. Media* 1 (2017), 1–13. https://doi.org/10.1016/j.osnem.2017.01.001
[33] Benedicta Obodoruku. 2016. Social networking: Information sharing, archiving and privacy. In *Proceedings of the 24th BOBCATSSS Conference Proceedings and Abstracts*.
[34] David Otero. 2019. *Plataforma Para la Etiquetación Asistida de Casos de Riesgo Temprano en Internet*. Bachelor's Thesis. Univesity of A Coruna. Advisor(s) Daniel Valcarce and Javier Parapar. Retrieved from http://hdl.handle.net/2183/24557.
[35] David Otero, Javier Parapar, and Álvaro Barreiro. 2020. Beaver: Efficiently building test collections for novel tasks. In *Proceedings of the Joint Conference of the Information Retrieval Communities in Europe (CIRCLE'20)*. CEUR-WS.org.
[36] Jahna Otterbacher. 2018. Addressing social bias in information retrieval. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. Springer International Publishing, Cham, 121–127.
[37] Nathaniel Payne. 2018. Stirring the cauldron: Redefining computational archival science (CAS) for the big data domain. In *Proceedings of the IEEE International Conference on Big Data (Big Data)*. IEEE, 2743–2752.
[38] Computational Archival Science (CAS) Portal. 2020. Retrieved from https://ai-collaboratory.net/cas/.
[39] Jennifer Pybus. 2013. Social networks and cultural workers: Towards an archive for the prosumer. *J. Cultur. Econ.* 6, 2 (2013), 137–152.
[40] Jennifer Pybus. 2015. *Accumulating Affect: Social Networks and their Archives of Feelings*. MIT Press, 234–249.
[41] Reddit. 2020. Retrieved from https://www.reddit.com.
[42] Thorsten Ries and Gabor Palko. 2019. Born-digital archives. *Int. J. Dig. Human.* 1, 1 (2019), 1–11.
[43] Pilar Rivero, Inaki Navarro, and Borja Aso. 2020. Educommunication Web 2.0 for heritage: A view from Spanish museums. In *Handbook of Research on Citizenship and Heritage Education*, Emilio Jose Delgado-Algarra and Jose Maria Cuenca-Lopez (Eds.). IGI Global, Hershey, PA, 450–471.
[44] Nubras Samayeen, Adrian Wong, and Cameron McCarthy. 2020. Space to breathe: George Floyd, BLM plaza, and the monumentalization of divided American Urban landscapes. *Edu. Philos. Theory* (2020), 1–11. https://doi.org/10.1080/00131857.2020.1795980
[45] Cristina Sánchez-Carretero. 2013. Significance and social value of cultural heritage: Analyzing the fractures of heritage. *Sci. Technol. Conserv. Cultur. Herit.* (2013), 387–392. https://doi.org/10.1201/b15577-90
[46] Mark Sanderson. 2010. Test collection based evaluation of information retrieval systems. *Found. Trends. Info. Retriev.* 4, 4 (June 2010), 247–375. https://doi.org/10.1561/1500000009
[47] Günther Schefbeck, Dimitris Spiliotopoulos, and Thomas Risse. 2012. The recent challenge in web archiving: Archiving the social web. In *Proceedings of the International Council on Archives Congress, Brisbane, Australia*. 1–5.
[48] Stephan A. Schwartz. 2020. Police brutality and racism in America. *EXPLORE1* 6, 5 (2020), 280–282. https://doi.org/10.1016/j.explore.2020.06.010
[49] Hrvoje Stančić. 2018. Computational archival science. Moderna Arhivistika. Časopis arhivske teorije in prakse. *Journal of Archival Theory and Practice* 1, 2 (2018), 323–329.
[50] Franklyn Herbert Upward. 2005. The records continuum. In *Archives: Recordkeeping in Society*. Centre for Information Studies, 197–222.
[51] Ellen M. Voorhees. 2002. The philosophy of information retrieval evaluation. In *Proceedings of the Evaluation of Cross-Language Information Retrieval Systems: Second Workshop of the Cross-Language Evaluation Forum*. Springer, Berlin, 355–370. https://doi.org/10.1007/3-540-45691-0_34
[52] Ellen M. Voorhees and Donna K. Harman. 2005. *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press.
[53] Jane Winters and Andrew Prescott. 2019. Negotiating the born-digital: A problem of search. *Arch. Manuscripts* 47, 3 (2019), 391–403.
[54] IEEE Big Data 2018 Workshop. 2020. Retrieved from https://saaers.wordpress.com/2019/01/22.
[55] Computational Archival Science Workshops. 2020. Retrieved from https://dcicblog.umd.edu/cas/international-cas-workshop/.

Received November 2020; revised May 2021; accepted July 2021