Towards Temporal URI Collections for Named Entities

Sergej Wildemann
L3S Research Center
Hannover, Germany
wildemann@L3S.de

Helge Holzmann
Internet Archive
San Francisco, CA, USA
helge@archive.org

ABSTRACT

Web archives represent crucial endeavors in preserving the Web from the past and provide a valuable resource for researchers of different disciplines. Due to their size, navigation in these collections is often limited to specifying an URI and the desired date. However, typical research questions often revolve around the evolution of entities instead of specific websites. Although full-text search often seems to be the first choice to look up web pages, while it provides a quick way to yield the best match with a keyword, its diversified ranking is not made for compiling reliable entity related collections. Further, it generally ignores the temporal relevance that is needed to find pages from the past, e.g., in web archives.

In this paper, we present a collection of ranked resource identifiers, characterizing named entities over time. For this purpose, different datasets were collected and evaluated by comparing each with a combination of others. Benchmarked against web search engines, our approach achieves a remarkable precision of 83.3 % and shows promising results for high-quality lookups and temporal collection building. To not only rely on existing datasets, we have implemented an interactive platform to get humans in the loop to expand the collection by contributing URIs, metadata and temporal information as well as to correct errors.

CCS CONCEPTS

• **Information systems** → *Retrieval models and ranking*; • **Applied computing** → *Digital libraries and archives.*

KEYWORDS

Web Archives, Temporal Information Retrieval, Collaborative Knowledge

ACM Reference Format:
Sergej Wildemann and Helge Holzmann. 2019. Towards Temporal URI Collections for Named Entities. In *Proceedings of Joint Conference on Digital Libraries (JCDL'19).* ACM, New York, NY, USA, 10 pages.

1 INTRODUCTION

In the search for information on the Web, search engines have become the natural starting point. However, as time goes on web pages are likely to change or delete their content[1], become harder to find [22] or even vanish from the search indexes[19]. Web archives like the *Internet Archive*[1] try to preserve this rapid evolution of the Web and document the digital history of humanity[14]. This creates a crucial resource for cultural[20], historic[25] and journalistic[7] analysis. But applying traditional search engine techniques to these multidimensional temporal collections has proven to be challenging[11, 26, 28]. Most resources on web archives are therefore only accessible when given the exact URI, which lowers discoverability and does not satisfy the demand of predominantly navigational intents[9].

The history of named entities like "Barack Obama" is characterized by a changing set of resources over each time period. Covering his life before the political career, in the campaign phase, as a president and the time afterwards. Knowing all relevant URIs in advance cannot be expected for every use case. A collection of links on important sources can for example be found in the references on his Wikipedia page, but covers only a fraction of possible topics and views[23]. To reliably retrace facts and analyze entity evolution, an easily accessible and extensive list of archived original sources is therefore needed.

Current approaches in entity recognition and information retrieval in web archives rely purely on existing data and provide no way for the user interact with the search results[16, 18, 21, 27]. Integrating the idea behind *Folksonomies* as social tagging systems would allow users to improve existing classifications[12], augment query results and influence ranking[17].

In this work, we create a collection of ranked and annotated URIs, characterizing named entities over time. Multiple diverse datasets are filtered, transformed, unified and integrated to exploit potential sources of URIs and corresponding metadata. Further comparison of each dataset with the combined results allows us to estimate their quality level for different entity types and adjust scoring accordingly. Our top results show a high precision with respect to standard web search engines and outperform currently deployed search solutions for web archives by differentiating between exact URIs and timeframes. Finally, we present the interactive web platform *Tempurion* to let users explore and extend this existing dataset. Potential users are given the opportunity to contribute by adding metadata or missing links. A voting mechanism is further incorporated into each aspect to distinguish the importance and influence ranking.

2 RELATED WORK

In the past, we made several attempts to provide an improved search for web archives. The first version of a system called *Tempas* matched the users query against tags supplied to URIs by the social bookmarking system *Delicious*[16]. This approach already showed a good recall in regard to the URIs that were clicked on analyzed query logs from AOL and MSN[17]. For ranking, we relied on the frequency of tags used for a given URI, without further evaluating the precision. Due to the limited time span of the dataset and its biased user base we later switched to analyzing the anchor texts in the web archive directly[18]. Using this method avoided a computationally intensive full-text index on large web archive collections and instead relied on the navigational character of anchor texts while showing promising results in empirical evaluations. In this work

[1] https://archive.org

JCDL'19, June 2019, Urbana-Champaign, Illinois USA
2019.

we want to shift our attention from the general-purpose character towards a topic focused collection for named entities.

User Intent in Web Archives: Information needs in traditional web search systems can be classified into being *informational*, *navigational* or *transactional* according to the taxonomy introduced by Broder[6]. While web archives contain content originating from the Web, studying query logs from the Portugese Web Archive, Costa and Silva[9] showed that users intents differ significantly. With the majority of queries being navigational instead of informational, users are more interested in exploring resources in the temporal dimension. Many navigational queries, but also nearly half of the information ones, referred to named entities. A later study[10] revealed a high preference for older documents. This was not reflected by the queries alone, as they rarely contained any temporal expressions or showed the use of date filters. They explained these observations by the use of traditional interfaces that were not evaluated for web archives[8].

Value of Folksonomies in IR:. Folksonomies provide a large set of informal classifications of resources in the form of tags by non-experts. The use of social bookmarks as a data source for information retrieval (IR) systems was explored by Heymann et al.[15]. Tags were observed to be relevant and objective, but were also contained in the resource content itself half of the times. These findings were also confirmed by Bischoff et al.[5], who further pointed at the additional information the other half adds. Beneficial for the use in search, tagging behavior was described as similar to the characteristics of a query for the respective resource. Aliakbary et al.[2] explored how tags could be used to classify new URIs for existing directories, stating an advantage against content based approaches. When dealing with a high amount of tags, Godoy and Amandi[12] highlighted the importance of preprocessing in order to reduce noise and improve classification accuracy.

3 DATASETS

Towards tackling the cold start problem in *Tempurion*, we sought for a substantial amount of data comprising of meaningful URIs related to named entities that would populate our database in the beginning. To further obtain validity times, describing tags as well as an initial ranking, a combination of multiple datasets was needed. This section describes the used data sources and their characteristics.

Wikipedia provides a well-known knowledge base for topics of all fields and thus, it covers most entities of public interest. Its English version contains millions of articles that can be obtained from regularly provided database dumps[2]. However, not all articles represent named entities and therefore, additional datasets are required to categorize them appropriately. Each article can also contain a section with external links that comprises of URIs to homepages or other meaningful content. Matching the criteria of this work, they were collected as well.

DBpedia continuously collects structured information from Wikipedia, publishes it as *Linked Data* and enables semantic queries of properties and their relations[4]. The latest obtainable dataset[3] describes over 6.6 million objects and classifies them in a consistent ontology (cp. Table 1). Because entries do not always reference specific Wikipedia articles or represent real world entities, it cannot be used independently to obtain a self-contained list of entities. Negative examples here include audio files from Wikipedia articles categorized under *Creative Work*.

Table 1: Distribution of entity types in DBpedia

Type	Count	%
Person	1,500,000	22.7
Place	840,000	12.7
Creative Work	496,000	7.5
Organization	286,000	4.3
Other	2,378,000	36.0
N/A	1,100,000	16.7
	6,600,000	100

Delicious used to be a popular social bookmarking service that allowed users to manage, annotate and share their web bookmarks. The corresponding *SocialBM0311* dataset contains about 340 million bookmarks with metadata from nearly 2 mio. users up until march 2011[29]. Contained resources reflect the users' interests at the time when the bookmarks were created and therefore, are expected to be of high relevance. However, in previous work, we found a strong bias in this dataset introduced by the typical users' background towards computer and technology related topics, which limits its general applicability for temporal search on web archives[17]. The tags also require some preprocessing before further usage, because the same concepts can appear in different shapes (e.g., `star-trek` and `star.trek`). Of about 14.7 mio. distinct tags that can be found, 8.1 mio. were only used once.

Wikidata is an associated project of Wikipedia and provides a structured data knowledge base for collaboratively collected facts. The database dump from October 2018[4] contains facts of about 50 mio. articles/topics. Of main interest are the high-quality URIs of entity-related websites included in the facts, of which 3.3 % of the articles contain at least one, totaling around 1.7 mio. URIs. More can be obtained indirectly by combining specified profile names and identifiers for sites such as social media, which adds another 73 mio. URIs.

The **German Web Archive** from 1996 to 2013 consists of over 2 billion archived snapshots of web pages under the German `.de` domain. It was collected and provided to us by the *Internet Archive*. Extraction of entity link relations was done in different ways and led to the formation of two distinct datasets. Based on the previous work in Holzmann et al.[18], anchor texts in the collection were matched against entity names and the 10 most popular target URIs together with the respective years were obtained. This resulted in matches for 69.8 % of all entities and will be referred to as *GWA*. We further created a second dataset based on the German Web Archive with the same approach as before, but only with anchor texts on web pages referenced from German Wikipedia articles. The corresponding dataset is named *GWW* and covers 36.2 % of all considered entities.

[2] https://dumps.wikimedia.org/enwiki/20180901/

[3] https://wiki.dbpedia.org/develop/datasets/dbpedia-version-2016-10

[4] https://dumps.wikimedia.org/wikidatawiki/entities/20181008/

The **Wayback CDX** server[5] is part of the Internet Archive's *Wayback Machine* and serves the index that is used to look up captures. Wherever possible, we used it to fetch the first and last timestamp of every URI in our datasets in order to fill in the missing relevance timespans. This enables temporal filtering of our data, based on when a URI was archived, approximating their validity times, which can be refined by our users later on.

4 APPROACH

With a diverse set of multiple data sources, we now define a target database schema and describe the process of filtering, transforming and combining the datasets.

4.1 Target Schema

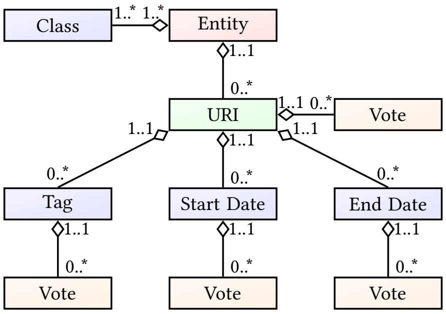


Figure 1: Target database properties and relations

Each of the presented datasets features different characteristics and contributes different properties, which become most useful when used in conjunction with the others. A common output schema is therefore created to define needed data classes and their relations for the resulting database (Figure 1). Our collection focuses on **named entities** that are classified in an ontology and belong to at least one of the following **classes**: *person*, *place*, *organization* or *creative work*. Each entity is further connected to a set of characterizing **URIs** annotated with tags and time periods, of which every single element can receive **votes** that indicate their relevance. Like in a folksonomy, the **tags** are used to describe and categorize linked web pages. The **start date** of a URI should correspond to the time when the respective content was published. Likewise, the **end date** marks the time of content deletion or other significant changes.

Corresponding to this design, we classified the properties of each dataset and define connection points for possible combinations in Table 2. A connection point means, a dataset is queried or filtered by this property based on data in another dataset. For instance, a list of entities from Wikipedia is used to obtain respective classes from DBpedia. The exact methodology applied will be described in the following sections.

4.2 Collection of Entities and URIs

Before starting to collect relevant URIs and associated metadata, a target list of possible named entities had to be compiled. Wikipedia contains articles spanning a vast number of topics and should therefore cover a large subset of entities of public interest. By combining the list of article titles with DBpedia's ontology we were able to extract 1.81 mio. named entities out of the total of 14.1 mio. titles.

Table 2: Dataset properties (●) and connection points (⋈)

Dataset	Entities	Classes	URIs	Tags	Dates
Wikipedia	●		●		
DBpedia	⋈	●			
Wikidata	⋈		●		
Delicious	⋈		●	● ⋈	●
GWA	⋈		●		●
GWW	⋈		●		●
Wayback CDX			⋈		●

A majority of the filtered Wikipedia articles contain links to external sources as references of the mentioned facts. Besides these, URIs can also appear in a separate "external links" section. According to Wikipedia's guidelines[6], this section should be kept short and only include links to further relevant information that cannot directly be added to the article for some reason. This results in 3,503,700 extracted URIs for 1,013,995 entities from the provided database dumps and by querying the Wikipedia API.

Additional URIs can be found on Wikidata, where its items directly correspond to Wikipedia articles. After filtering these items by our list of entities, 306,931 of them contain statements with a URI, resulting in a total of 320,074 URIs that are added to our database. Other statements provide identifiers for specific websites, like social media, and can be also transformed to URIs, when applying a suitable formatting template. Taking these into account, another 5,562,737 URIs can be assigned to 1,210,292 entities. Even though Wikidata also features temporal statements for a limited number of URIs that could potentially be used as start and end date, an examination of these dates showed that they often do not correspond to the validity of the URIs themselves but relate to information on the targets and were therefore not considered for extraction. Given that Wikidata provides a data source for Wikipedia, all links contained in both datasets were deleted from the Wikipedia dataset to avoid duplicated initial votes (see Section 4.3 below). This led to a removal of 1,061,740 URIs and 150,820 entities from the Wikipedia dataset.

Each entry in the Delicious dataset consists of user id, timestamp, URI and a list of tags. Without any other direct entity reference, the unordered collection of tags represents the only intuitive starting point. However, since user-supplied tags appear as a one-word lowercase term with possible interpunctuation, they cannot be directly mapped to our entity names without modification. For instance, possible matches for the entity "Star_Trek_(movie)" can be found encoded in different ways: "`startrek`", "`star-trek`", "`star_trek`" or as two separate tags "`star`" "`trek`". To successfully match multi-term entity titles the following approaches were considered:

(1) Splitting entity titles and matching against tags.
(2) Matching normalized entity titles against tag permutations.
(3) Direct matching of normalized entity titles and tags.

[5] https://github.com/internetarchive/wayback/tree/master/wayback-cdx-server

[6] https://en.wikipedia.org/wiki/Wikipedia:External_links

The first two approaches have a drastically increased complexity, which resulted in much longer processing times. At the same time, a check of the matches revealed a high rate of false positives due to the possible formation of valid entity names from otherwise unrelated terms. Because of this and due to the fact that multi-term names in most cases also occur as a single tag by other users, only the third method is used in the next steps.

For this approach, entity titles and tags were normalized to share a common format. All punctuation and characters not in the English alphabet were removed and the resulting strings turned to lowercase. Entity disambiguation given in parentheses was ignored. As a downside of these transformations, multiple entities can share the same identifier, which leads to an increased number of possible matches when matching them to tags without a possibility of identifying the correct one. The term "Java" can for instance either match to the programming language, an island or a town in Georgia. To resolve some of these disambiguations, for every entity in the result set that provided further information in parenthesis, it was required that at least one other tag used in conjunction was present.

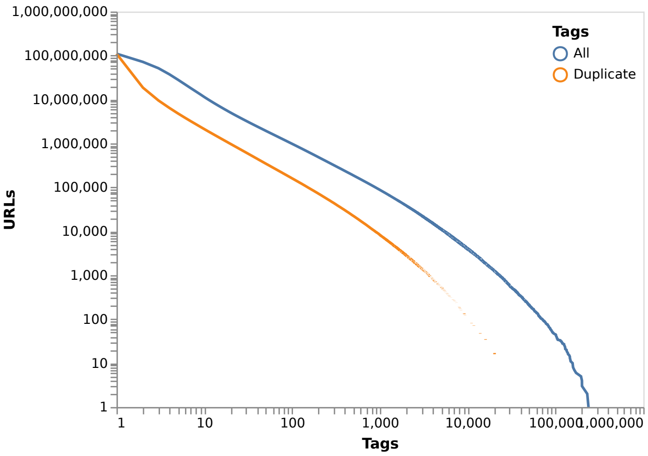


Figure 2: Overview of URIs in Delicious given a minimum number of overall or duplicate tags by users.

On the other hand, not every tag is useful and using all of them for the matching step would result in too many false positives. Some tags on Delicious like "`imported`", "`bookmark`" or "`rss`" were presumably generated by automated means and appear quite often. Therefore, to lower the noise level and improve accuracy, we only used tags that were given to one URI by at least 10 different users, resulting in around 2.1 mio. URIs that can be processed further (Figure 2). Finally, each tag-URI pair is sorted by relevance according to the number of bookmarks with this combination after normalization. Our assumption here is, a URI can be relevant for multiple entities, such as a news article about a political meeting. Based on this, we weight each possible match and encode this weighting within "votes" (see Section 4.3). Using this method, we extracted 2,326,440 URI assignments for 41,649 entities.

Further information about these matchings can be obtained by looking at the tags co-occurring with an entity in Delicious. This helps to describe the URI target and thus improves navigation in the result listings later. Web pages with political news might for example be tagged with "campaign", "election" or "satire". Therefore, for each matched tag, a list of all co-occurring tags is collected in every bookmark of the same URI. Tags that contain the matched one as a substring are removed to account for compound terms. This list is then sorted by the number of users and again given a normalized vote, like the entity matches. Out of these, in order to remove lower-quality results and reduce noise, we limited the number of additional tags to the most frequent five.

The timestamp of each bookmark can be used as an indicator of when a resource was important to the users, assuming that a website would not be bookmarked when it is not accessible anymore, or not annotated with a specific tag after the content has drastically changed. Thus, by collecting all timestamps for a matched tag and corresponding URIs would give us a date range for the relevance of a resource for an entity. However, considering that the dataset ends in 2011, estimating the end date for assignments near this limit would be speculative. Also, users tend to mainly create bookmarks for resources that are new on the Web[13] and fewer bookmarks for these web pages appear after this initial momentum. Thus, seeing less bookmarks does not automatically imply that a resource is not relevant for an entity anymore. We therefore concentrate our efforts on extracting the start dates. For every URI we select the date of the first appearance of any tag that is matched to an entity as the start date. This means that the same URI can have a different validity date for one entity before it starts to be relevant for another, accounting for websites that change their content over time.

The two datasets GWA and GWW which are based on the German Web Archive already contain assignments to entities due to the generation process (see above, Section 3) and do not need to be processed in that regard. Further, every URI in the datasets is augmented with a list of years in which it could be matched to a given entity. Start and end date of the assignment are reflected by the first and last year of this list. The latter is ignored if it is given as 2013 because this is the last year of this dataset. GWA being created without any filtering yielded 9,046,987 URIs for 1,266,241 entities whereas GWW only contributed roughly a third of that with 3,365,652 URIs on 656,415 entities.

Finally, over 1.6 million (91.3 %) of the 1.82 mio. identified named entities from Wikipedia are assigned at least one URI and thus covered by our datasets. On average, there are 12.54 URIs for each entity and 13.73, when only the covered entities are considered, which will be further discussed below. In total, this results in 22.75 mio. allocations, of which 1.27 % can be found in more than one dataset. These overlaps can be seen as an indicator that the corresponding assignments are of high confidence.

The majority of the entities are covered by the two datasets Wikidata and GWA, which both describe over 1.2 million entities (see Table 3). By GWW only half of the value is reached, although this is based on the same source data as GWA, only filtered differently. Here, an assessment of the quality of the assignments is particularly interesting, because at the same time the filter criteria can be evaluated. It is striking that the Delicious dataset only delivers results to 2.3 % of all entities. This value is probably due to the character of bookmarks that underlie them, which reflect the interests of the users and therefore cannot cover all possible topics. Secondly, our entity matching method could limit the number of assignments, especially for entities with disambiguation hints.

Table 3: Covered entities per type and dataset

Dataset	Person	Org.	Place	C.W.	∑
Wikipedia	400,070	117,260	157,066	188,739	863,135
Wikidata	537,364	164,562	327,144	251,748	1,280,818
Delicious	4,380	6,639	3,034	27,596	41,649
GWA	498,894	176,784	252,529	338,034	1,266,241
GWW	244,768	83,883	112,984	214,780	656,415

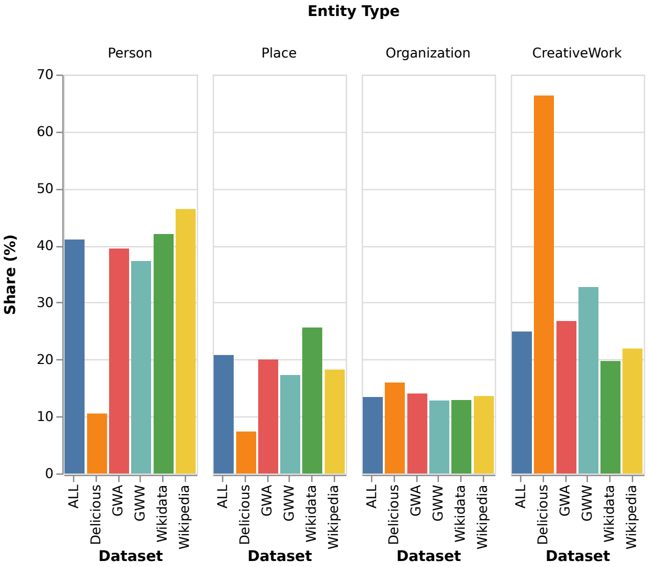


Figure 3: Distribution of entity types in the datasets

Delicious further features distinct preferences for entities of the type "Creative Work" (cp. Fig. 3). This coincides with the bias described in Section 3. However, one other cause of this distribution could also originate from the method used to associate tags with entities. For instance, person names are usually made up of several words, which increases the difficulty for assignments. The other datasets follow the overall entity type distribution without major derivation.

Table 4: Average URIs per entity type and dataset

Dataset	Person	Org.	Place	C.W.	All
Wikipedia	3.52	2.56	2.47	1.81	2.83
Wikidata	5.52	2.79	2.96	5.95	4.60
Delicious	22.15	51.43	63.53	61.43	55.86
GWA	6.54	7.16	6.52	8.49	7.14
GWW	4.43	5.32	4.63	6.10	5.13

Looking at the number URIs per entity and dataset (Table 4), the Delicious dataset again stands out while contributing 55.86 URIs on average. These are clearly too many to make sense and will to be addressed further after evaluating the quality of these assignments in Section 5.

4.3 Ranking

To rank the matches in our final collection, each obtained entity-URI combination from the datasets is assigned an initial vote based on our assumption of the quality of the respective dataset or specific factors like the number of users. Initial votes are assigned in the range from 1 to 10, where the latter reflects the highest confidence. We specifically choose a low upper limit of 10 to make future corrections easier for users, which were otherwise ruled out by the high number of generated votes. Further evaluation in Section 5 will reveal their true quality and allow additional adjustments.

Links from Wikipedia are split into two categories, the ones consisting of only domain names and the ones having a path. The former are often added when official websites exist for an article entity, represent resources of high relevance and thus are given a vote of 10. Other URIs, although hand-picked by the editors, often only refer to collections of information (e.g., links to the Internet Archive collection) and are given a lower vote of 5.

Direct URIs extracted from Wikidata are scarce with only 1.05 URIs per entity and most often seem to refer to official websites, thus receiving a vote of 10. Indirect ones are more abundant (e.g., 111 for "Barack Obama") and refer to common social media platforms or databases and therefore voted with 5.

Delicious provides a useful relevance measure in the form of the number of users who bookmarked a URI. For distinct URIs, the number of users U_i of a tag i is normalized with respect to the maximum number of users U_{max} of any tag for one resource. This results in a vote V_i for each matching entity where the upper limit is given by λ:

$$V_i = \left\lceil \frac{U_i}{U_{max}} * \lambda \right\rceil$$

The datasets GWA and GWW base on the same source material and provide no ranking indication for up to 10 URIs per entity. We chose a low vote of 3 for both with respect to the relatively high number of results.

4.4 Unification of URIs

The best results for an entity should be URIs that appear in most sources and have the highest ranking or confidence score within those sources. Given that the same resource can be accessed by slightly different variants of URIs, e.g., with different session parameters, a normalization step is performed in order to achieve a bigger overlap among the datasets. As an example, the following pair of URIs refer to the same resource:

(1) http://nytimes.com/2011/03/13/business/13coffee.html?ref=business
(2) https://www.nytimes.com/2011/03/13/business/13coffee.html

A URI consists of multiple, partially optional, segments: scheme, domain, path, query parameters and fragments. Because all of them can vary while still identifying the same resource, we need possible transformations for each one.

The **scheme** defines the protocol, which for web resources is either http or https. We assume that websites that use the latter, secured variant also provide a redirect to it when accessed via normal http, which is commonly the case. Therefore, all protocols are replaced by http only.

Domains used by websites are typically prefixed with the popular `www` subdomain, redirect to it if missing or accept either option. To unify these occurrences, we first collect a list of all domains in our datasets, which then can be used to check if a domain also exists with a `www` prefix and replace it accordingly.

Paths ending in index files like `"index.html"` rather than a directory name are handled equally in common web server configurations. Because both forms can exist for the besides same URI in the datasets, these path endings are removed.

Fragments provide a reference to a specific section within a web page and can generally be omitted without negative consequences.

Query parameters are given as key-value pairs and influence the requested documents, change their content or provide other information of the target web page. We found a number of occurrences, where values are left blank and similar URIs exist without the respective key. These and typical keys used for session management and tracking purposes were removed.

4.5 Adding Temporal Information

Not all datasets provide temporal information about their URIs, when they are to be considered relevant or even accessible. Approaches to estimate the creation time of a web resource by querying multiple sources were done[24] but do not scale well in our context with millions of URIs. Therefore, we added these by querying the Internet Archive's CDX server API (see Section 3) to determine an approximation of start and end dates.

Although captures of web pages in the Wayback Machine are not always consistent at defined intervals, it can be assumed that an approximately annual frequency is maintained[3]. Also, we found that in many cases the earliest time point in that a URI appears in the Wayback index is often not far from the publication date of the resource, which we use as an approximation of the start date in our initial dataset. On the other hand, retrieving the end date from this dataset turns out to be less reliable. Therefore, our heuristic is to assume a URI to be invalid or lost relevance if the last successful entry was before 2016, i.e., more than two years before the date of our database creation. Using this method, we augmented 8,107,941 URIs with additional temporal data.

5 EVALUATION

In the following, the different datasets are evaluated by exploiting their overlaps. Comparisons between individual datasets should provide information about the quality of URI assignments for different entity types and evaluate our approach as described before. Based on the findings in this evaluation, potential filters can be derived and rankings adjusted to improve the results in the final system.

After the removal of URIs from the Wikipedia dataset that were also present in Wikidata due to their project relationship, no further overlap can be found between these two. Therefore, these were combined and referred to as *Wiki* in this section.

5.1 Dataset Overlap

An assignment of a URI to an entity is considered of higher quality if it is supported by multiple datasets. This is the case for 13.4 % of all entities with URIs (see Table 5). However, a large proportion (79.5 %) of them feature only a single overlap, i.e., an identical URI in multiple datasets. The maximum value is achieved with 18 overlaps for a single entity, i.e., the visual programming language "Quartz Composer". At the same time, its assignments are largely dominated by the Delicious dataset, contributing 144 of 193 distinct URIs, which stresses again the strong bias of this dataset. Though, this example also shows the ability of our extraction method to maps tags with multi-term entities.

However, it is important to note that, in some cases, the number of possible overlaps is limited by dataset characteristics. For instance, there are at most 10 links per entity in the GWA and GWW datasets due to their generation process, which decreases further after normalization and unification. In addition, as presented above, Delicious covers only 41,649 entities, of which not all are covered by other datasets as well.

Table 5: Entities with number of URIs from multiple sources

Overlap	Person	Org.	Place	C.W.	$\sum$
1	51,336	48966	27,542	48,919	176,763
2	7,577	8,351	4,292	12,108	32,328
3	1,566	2,281	1,157	3,819	8,823
4	386	699	315	1,361	2,761
≥ 5	173	290	236	1066	1765
	61,038	60,587	33,542	67,273	222,440

Overall, the spread of overlapping entities does not correlate with the previously studied distribution (see Fig. 3). Organizations take a considerably larger share, while places are underrepresented. The Creative Work type is the largest contributor, possibly suggesting that Delicious has a high overlap ratio there due to its bias. This will be examined in more detail in the following Section 5.2.

For an overview of the scope and overlaps of the datasets is given by Figure 4. Here, the number of URI-entity mappings of each dataset together with the value of overlaps with one or more others is shown. The highest number of overlaps can be seen between GWA and GWW, resulting from their common origin. The fact that this is a relatively small fraction though, shows that they are very different. In conjunction with Wiki, GWA achieves a significantly higher overlap than GWW both in absolute and relative terms. Additionally, a large part of the overlaps of GWW and Wiki can also be found in GWA, so that only 4,113 additional assignments from the former are added. At the same time, the direct links from Wikidata account for about 66 % of these overlaps.

Delicious has more in common with GWA than with the GWW dataset, where the latter only adds 1,153 URIs. Between Delicious and GWW there is the least commonality when looking at the combination of two datasets. When also considering the relative amount of overlaps, even less similarity occurs with Wiki despite being our second largest dataset. Thus, Delicious overall appears

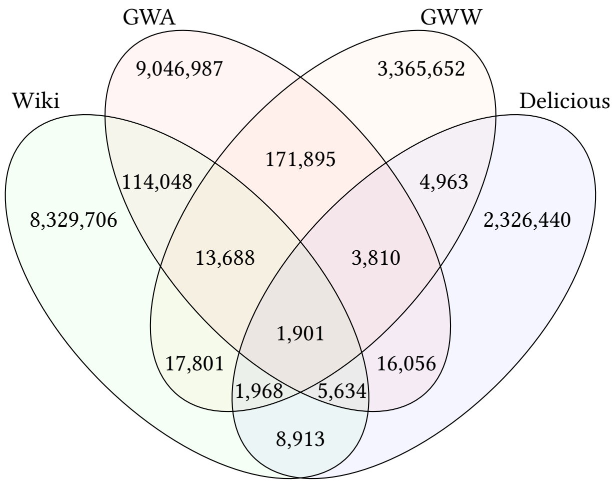


Figure 4: Number of URI-assignments to entities per dataset with overlap to others.

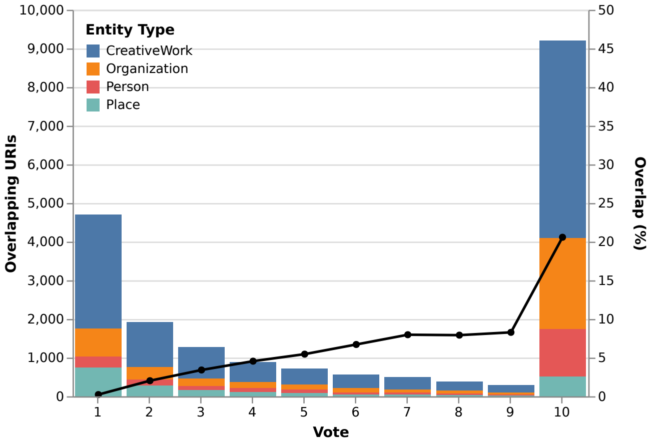


Figure 5: Overlap of URIs in the Delicious dataset by vote. The relative amount is depicted by the line plot.

to have the worst results given its size. Finally, the overlap of all records results in 1,901 URLs for 1,884 entities.

5.2 Ranking of Delicious

Despite the small number of entities described, the Delicious dataset has by far the most URI assignments per entity. Therefore, it is worthwhile to look at it in more detail. In Section 4.3, each matching of a tag to a URI was given a vote, to reflect its quality proportional to the number of users of a tag. However, with a total share of almost 90 % (2,093,599 of 2,326,440 URIs), the lowest voted assignments constitute the majority of all assignments. So it is to be expected that a higher vote should probably provide much better results.

In Figure 2, the number of overlapping URIs is associated with their respective votes. Most of the overlaps occur for entities of the type "Creative Work". This was to be expected from the described breakdown of entity types in the dataset and is not much different for the remaining types. Although many of the URIs with a vote of 1 appear in other datasets as well, they represent a complete exception with a 0.22 % share. URIs with the highest vote of 10 contribute to the largest number of overlaps. With a 20.57 % share of all URIs in this vote, they generally appear to be of reasonable quality.

The likelihood of an overlap already drops significantly to 8.27 % with a vote of 9 and declines further. At the same time there are relatively few overlapping URIs per vote in the range between 2 and 9, with the larger part being in the lower ranks. Splitting this interval, votes over 5 achieve an overlap rate of over 6.7 % for a total of 35,891 URIs. For lower votes, down to 2, only an average of 2.7 % overlap by constituting a larger share of 152,249 URIs. This distribution shows that our initial quality assessment is appropriate and accurate results are to be expected at votes of 5 and up. Nevertheless, it should be noted that the overlap fraction only serves as a quality indicator and other URIs may still be relevant too.

5.3 Quality of Ranking

To study the quality of a dataset's results, we employed a different kind of evaluation. Here, it is of interest on which ranks of the sorted result sets for entities the URIs of a dataset exhibit overlaps. Higher ranks for matches should also represent more important results. It should be noted that the order of ranks is strongly influenced by what initial vote was given to the individual datasets in advance. However, since such applications cases lack suitable evaluation criteria, only a subjective assessment could be made.

The first measure to use is the *Mean Reciprocal Rank* (MRR), which is meant to evaluate results for different queries. It results from the average reciprocal rank; the multiplicative inverse of the rank of the first correct answer in a series of queries. The set of queries (Q) here consists of all URIs a dataset contains for an entity. For each URI i, the rank (rank_i) in the list of results without this dataset is then determined. If no rank can be determined because the URI does not exist in the other datasets, its Reciprocal rank is 0. This results in:

$$\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_\text{i}}$$

For each entity of an entity type in which one dataset overlaps another, a corresponding MRR can now be calculated (see Figure 6). In the following we will determine which dataset provides the best results per entity type.

In general, it can be stated that GWA performs the worst in all categories and thus, often contributes seemingly irrelevant URIs. Best results are achieved for the "Organization" type. Much better rankings are seen for the similar dataset GWW, which performs best for the entity type "Person". The use of a different filter mechanism to extract URIs from the common web archive in order to generate this dataset proofs successful. At Delicious, the distribution of values for persons and organizations is similar and at a high level. On the other hand, for "Place" it achieves the worst overall values where other dataset perform equally poor. Wiki compares favorably with a good distribution for all categories except for people. This could be because of many indirect links from Wikidata for this category, which besides social media websites refer to lesser known

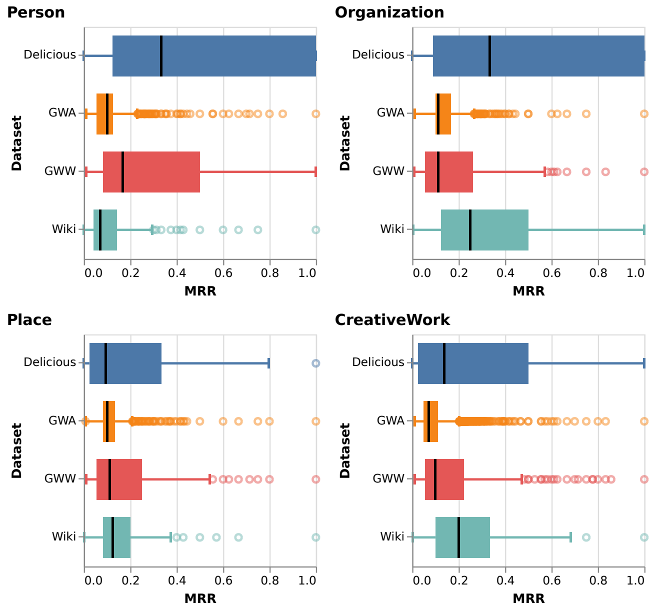


Figure 6: Mean reciprocal rank distribution of datasets by entity type. (Higher values are better)

directories. It can be also seen that results for "Organization" and "Creative Work" are much higher up front.

Overall, this confirms the assumption that Wiki provides one of the higher qualitative sources. Also, the URIs from GWW are better selected from the underlying web archive than in GWA. Despite many assignments of URIs to entities, Delicious shines especially for the type "Creative Work" for which it includes the most URIs.

5.4 Examining Overlapping URIs

If not all URIs of the datasets are considered, but only those with an overlap, the quality of these can be assessed and tells us which source is responsible for a majority of the best-placed results. This can be determined by the *Discounted Cumulative Gain* (DCG), which specifies a value for the usefulness of a result depending on its rank in the result list. The usefulness of a URI at rank r is related to the relevance of the URIs ($rel_i \in \{0, 1\}$) up until this rank, where 1 stands for high relevance or overlap:

$$\mathrm{DCG_r} = \sum_{i=1}^{r} \frac{2^{rel_i} - 1}{\log_2(i+1)}$$

Here the value is formed over all URIs of a dataset for an entity and r therefore equals their total number. URIs without overlap in the result list, i.e., results not in the considered dataset, do not need to be included due to the numerator in the equation. Since entities have different number of URIs, a normalization must be applied using the *normalized DCG* (nDCG):

$$\mathrm{nDCG_r} = \frac{\mathrm{DCG_r}}{\mathrm{IDCG_r}}$$

where

$$\mathrm{IDCG_r} = \sum_{i=1}^{|REL|} \frac{2^{rel_i} - 1}{\log_2(i+1)}$$

The *Ideal DCG* (IDCG) represents the maximum achievable DCG-value. This is done by defining the list of relevant URIs up to rank r as REL. Because our ranking is based on votes, multiple URIs can have the same rank. However, the ideal ranking in this case would be that all URIs share the first rank. Accordingly, the logarithm in the denominator always gives the value 1, as does the numerator. In simple terms, the IDCG here can thus be described as $|REL|$, i.e., the sum of all URI overlaps. Compared to the previously presented MRR measure, the IDCG would differ only in the denominator, if just the overlapping URIs where included for the MRR calculation. Here, the values are then a bit higher in comparison:

$$\mathrm{IDCG_r} = \frac{\mathrm{DCG_r}}{\mathrm{IDCG_r}} = \frac{\mathrm{DCG_r}}{|REL|} = \frac{1}{|REL|} \sum_{i=1}^{r} \frac{2^{rel_i} - 1}{\log_2(i+1)}$$
$$= \frac{1}{|REL|} \sum_{i=1}^{r} \frac{rel_i}{\log_2(i+1)}$$

Analogous to the previous section, the distribution of the IDCG-values are shown in Figure 7. It immediately stands out that Wiki contains the best quality URIs with overlap for all entity types. At the same time, there are many outliers. They are created, for example, by overlapping with low-rated Delicious URIs or in cases where all records yield results and only GWA or GWW are hit. In addition, the low overlapping fraction with Delicious from Figure 4 and the low number of described entities from the latter must be taken into account (see Table 5). For a large part of the included entities there are only entries from GWA and GWW, which all occupy first place in the result list due to the identical votes.

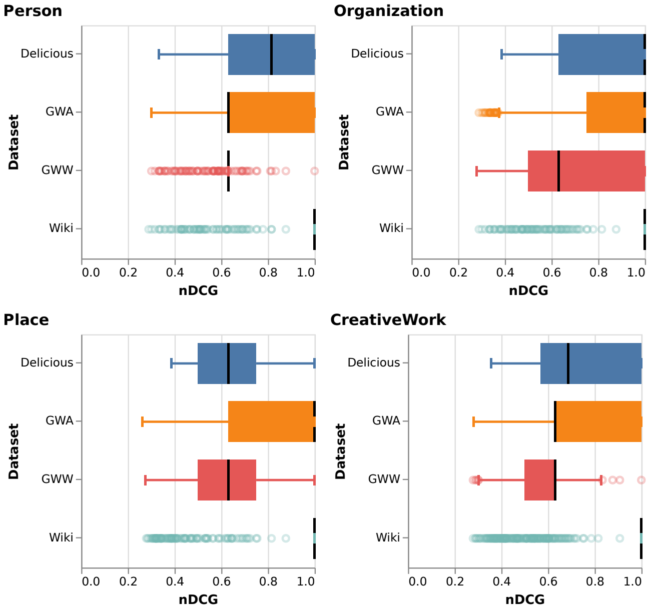


Figure 7: Normalized discounted cumulative gain distribution of datasets by entity type

In contrast to the generally low quality of the URIs from GWA in the MRR comparison, the results here are significantly better. Overlapping URIs are of high relevance for places and organizations right after those in Wiki. Furthermore, no significant weaknesses for the other types were obvious compared to the other datasets. In all cases, higher values are achieved than with GWW, which is particularly surprising after results with the MRR measure. The latter has the lowest values here and can only present satisfactory URIs for organizations. Thus, although GWW has better results overall, GWA contains the more relevant URIs.

Delicious contributes fairly qualitative URIs across all types. For "People" and "Creative Work" the second-best results can be achieved. This again confirms the usefulness of a completely user generated URI collection.

5.5 Comparison with Web Search Engines

Although the content of interest for a user might already be vanished from their index, web search engines represent the commonly used way to navigate the Web. Possibly satisfying the same information need, we considered them adequate to compare against. For this, we used the search engine *Bing*[7] and gathered returned URIs from the first result page for a sample of 50,000 entities having the highest number of overlapping URIs in our dataset. Entity titles were transformed to not contain any special characters in order to simulate a user's query. On average, these queries yielded 9.98 URIs per entity.

As a measure of precision, our system provides at least one identical URI for 83.29 % of the queried entities, while their average position on Bing is 2.26 with a median of 1. This result is achieved by URIs on the top position of our system, mostly formed by overlapping sources. Querying in reverse by measuring the fraction of Bing URIs contained in our dataset, we can observe that 116,261 of 499,024 URIs are present and we thus get a recall of 23.29 %.

5.6 Conclusions

Finally, possible improvements for the overall dataset can be derived from all results. Overall, Delicious performs reasonably well, but contains a large subset of seemingly poor URIs. Here a filtering according to the given votes would be a possibility for an increase in quality. The best results are observable from matches with a vote of 10. In order not to reduce the dataset too much, URIs could also be included down to vote 5, since the overlap stays above 5 %. However, lower rated ones could enrich other records with meta-information when overlapped and thus still be useful.

Wiki is a well-maintained dataset that includes qualitative URIs. The poorer overall score results largely from the many "indirect" URIs that point to more unknown web pages and directories, and therefore, for the most part, do not appear in any other records. Since a grading of the votes has already been made for this, no further measures are recognizable.

The quota of qualitative URIs is higher in GWW than in GWA, but the quality of the individual URIs is the other way around. Because fewer "false positives" are more desirable for the purpose, a slight increase in the vote for GWW might be beneficial. Additional filter mechanisms would seem to be desirable to decrease the dataset sizes, but were not done here due to unclear criteria.

The temporal aspect was left out of the evaluation. Only the Delicious dataset provides this information with reasonable accuracy. Although enrichment via the Wayback CDX service can complement missing information, it lacks appropriate evaluation methods and ground truth.

6 SYSTEM OVERVIEW

To explore the processed datasets from the previous section, we provide the web platform *Tempurion*. Besides a search through entities and their related URIs, users are provided with the possibility to interact with all elements analogues to other folksonomies. Public access to the underlying dataset is further provided in a machine-readable way via a RESTful API. A live version is accessible under:

http://tempurion.l3s.uni-hannover.de

Exemplary, a result view for the entity "Barack Obama" is shown in Figure 8. To easier identify an entity and provide disambiguation, each listing shows an extract of the associated Wikipedia article and is categorized into one of the four main entity types.

Every URI in the result set is accompanied by a vote as well as metadata like tags and possible start and end dates, each having a visible vote counter. The attached dates on the URIs link to the respective snapshot on the Internet Archive. A user is able to influence the ranking of all elements by in- or decreasing the respective vote counter once. Furthermore, missing information can be supplemented. When trying to add already existing items like a specific URI or a tag results in an upvote for the respective element. In the case of a URI, the user is then directly redirected to the result entry to encourage further interaction.

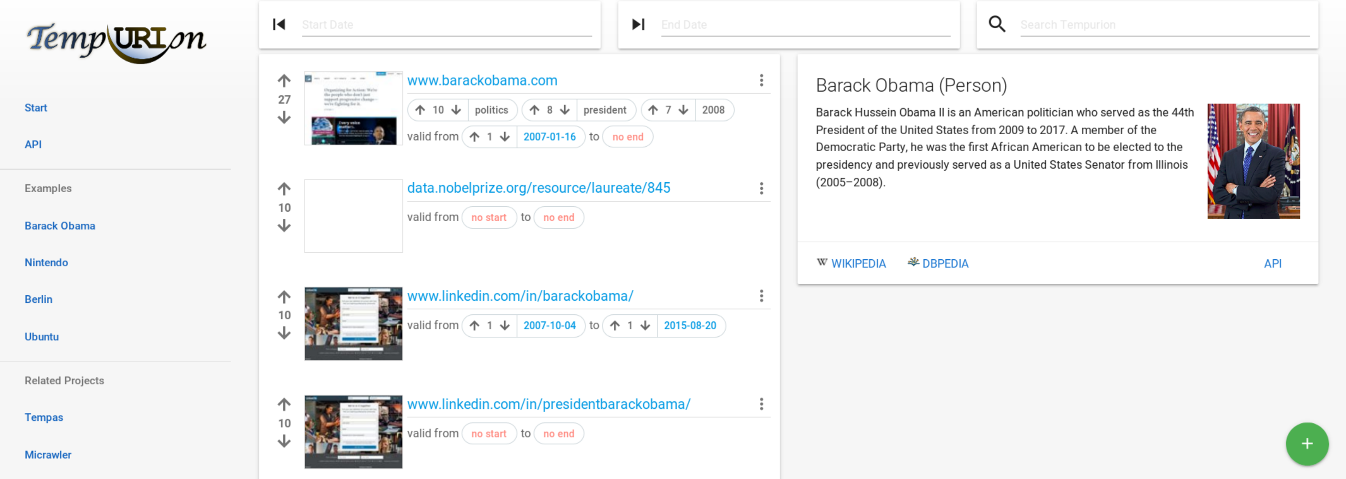
To also navigate in the temporal dimension, the result set can be filtered to a desired time period by selecting a start and end date. One of the boundaries can also be left out, i.e., to look at all results that were relevant until 2010. Within the period, a URI must have been valid, that is, the beginning is before the selected end time and the end after the start time. Since each URI can have multiple start and end times that have been voted differently, only the most highly rated entry will be used for filtering.

7 CONCLUSION AND OUTLOOK

In this paper, we presented a collaborative temporal URI collection for named entities and showed how a combination of different datasets can be unified and integrated to identify relevant results. By structuring and classifying URIs, this enables users to better explore web archives and gain new insights into the evolution of entities. While not every dataset is equally eligible, weaknesses and strengths were identified to derive filter criteria and adjust rankings. Using the tags of a social bookmarking service in order to match entities to relevant resources proofed to be a valuable approach and has validated the findings of other works.

Currently, not all entities are yet provided with respective resources and no language awareness is applied. Further datasets are also needed to extend the covered time-frames and incorporate more recent URIs. While initial temporal annotations can be derived from web archives and social bookmarks, verification of them proofed to be problematic. In order to improve the quality of the

[7] https://bing.com


TempURIon
Start
API
Examples
Barack Obama
Nintendo
Berlin
Ubuntu
Related Projects
Tempas
Micrawler
Start Date
End Date
Search Tempurion
27
www.barackobama.com
10 politics
8 president
7 2008
valid from 1 2007-01-16 to no end
10
data.nobelprize.org/resource/laureate/845
valid from no start to no end
10
www.linkedin.com/in/barackobama/
valid from 1 2007-10-04 to 1 2015-08-20
10
www.linkedin.com/in/presidentbarackobama/
valid from no start to no end
WIKIPEDIA
DBPEDIA
API
+

Barack Obama (Person)

Barack Hussein Obama II is an American politician who served as the 44th President of the United States from 2009 to 2017. A member of the Democratic Party, he was the first African American to be elected to the presidency and previously served as a United States Senator from Illinois (2005–2008).

Figure 8: Entity result view in Tempurion

results from prepared data, more filters and evaluation criteria have to be developed. In the future, this data will be constantly extended and evaluated by collaborative knowledge.

From a user’s perspective, the corresponding platform represents another entry point for the search in web archives. Furthermore, the existing data provides an interesting resource for further investigation of entity evolution and relationships based on their appearances on the Web.

ACKNOWLEDGMENTS

The work was partially funded by the European Commission for the ERC Advanced Grant ALEXANDRIA (No. 339233).

REFERENCES

[1] Teru Agata, Yosuke Miyata, Emi Ishita, Atsushi Ikeuchi, and Shuichi Ueda. 2014. Life span of web pages: A survey of 10 million pages collected in 2001. In *IEEE/ACM Joint Conference on Digital Libraries*.

[2] Sadegh Aliakbary, Hassan Abolhassani, Hossein Rahmani, and Behrooz Nobakht. 2009. Web page classification using social tags. In *International Conference on Computational Science and Engineering*.

[3] Internet Archive. 2018. Worldwide Web Crawls. Retrieved November 18, 2018 from https://archive.org/details/widecrawl

[4] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary G. Ives. 2007. DBpedia: A Nucleus for a Web of Open Data. In *International Semantic Web Conference*.

[5] Kerstin Bischoff, Claudiu S Firan, Wolfgang Nejdl, and Raluca Paiu. 2008. Can all tags be used for search?. In *ACM Conference on Information and Knowledge Management*.

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

[7] Sarah Cohen, Chengkai Li, Jun Yang, and Cong Yu. 2011. Computational Journalism: A Call to Arms to Database Researchers. In *Conference on Innovative Data Systems Research*.

[8] Miguel Costa and Mário Silva. 2009. Towards information retrieval evaluation over web archives. In *SIGIR Workshop on the Future of IR Evaluation*.

[9] Miguel Costa and Mário J Silva. 2010. Understanding the information needs of web archive users. In *International Web Archiving Workshop*.

[10] Miguel Costa and Mário J. Silva. 2011. Characterizing Search Behavior in Web Archives. In *Workshop on Linked Data on the Web*.

[11] Miguel Costa and Mário J. Silva. 2012. Evaluating Web Archive Search Systems. In *International Conference on Web Information Systems Engineering*. Springer.

[12] Daniela Godoy and Analía Amandi. 2010. Exploiting the Social Capital of Folksonomies for Web Page Classification. In *IFIP Conference on e-Business, e-Services, and e-Society*.

[13] Scott A Golder and Bernardo A Huberman. 2006. Usage patterns of collaborative tagging systems. *Journal of information science* 32, 2 (2006), 198–208.

[14] Daniel Gomes and Miguel Costa. 2014. The importance of web archives for humanities. *International Journal of Humanities and Arts Computing* 8, 1 (2014), 106–123.

[15] Paul Heymann, Georgia Koutrika, and Hector Garcia-Molina. 2008. Can social bookmarking improve web search?. In *International Conference on Web Search and Data Mining*.

[16] Helge Holzmann and Avishek Anand. 2016. Tempas: Temporal Archive Search Based on Tags. In *International Conference on World Wide Web, Companion Vol.*

[17] Helge Holzmann, Wolfgang Nejdl, and Avishek Anand. 2016. On the Applicability of Delicious for Temporal Search on Web Archives. In *International ACM SIGIR conference on Research and Development in Information Retrieval*.

[18] Helge Holzmann, Wolfgang Nejdl, and Avishek Anand. 2017. Exploring Web Archives Through Temporal Anchor Texts. In *ACM Conference on Web Science*.

[19] Judit Bar Ilan. 1999. Search Engine Results over Time: A Case Study on Search Engine Stability. *Cybermetrics: International Journal of Scientometrics, Informetrics and Bibliometrics* 2/3, 1 (1999).

[20] Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, et al. 2011. Quantitative analysis of culture using millions of digitized books. *science* 331, 6014 (2011), 176–182.

[21] Tu Ngoc Nguyen, Nattiya Kanhabua, Claudia Niederée, and Xiaofei Zhu. 2015. A Time-aware Random Walk Model for Finding Important Documents in Web Archives. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[22] Alexander C Nwala, Michele C Weigle, and Michael L Nelson. 2018. Scraping SERPs for Archival Seeds: It Matters When You Start. In *ACM/IEEE Joint Conference on Digital Libraries*.

[23] Cindy Royal and Deepina Kapila. 2009. What’s on Wikipedia, and what’s not...? Assessing completeness of information. *Social Science Computer Review* 27, 1 (2009), 138–148.

[24] Hany M Salah Eldeen and Michael L Nelson. 2013. Carbon dating the web: estimating the age of web resources. In *International Conference on World Wide Web*. 1075–1082.

[25] Susan Schreibman, Ray Siemens, and John Unsworth. 2008. *A companion to digital humanities*. John Wiley & Sons.

[26] Michael Stack. 2006. Full Text Search of Web Archive Collections. Retrieved October 25, 2018 from http://archive-access.sourceforge.net/projects/nutch/iwaw/iwaw-wacsearch.pdf

[27] Nina Tahmasebi, Gerhard Gossen, Nattiya Kanhabua, Helge Holzmann, and Thomas Risse. 2012. NEER: An Unsupervised Method for Named Entity Evolution Recognition. In *International Conference on Computational Linguistics*.

[28] Gerhard Weikum, Nikos Ntarmos, Marc Spaniol, Peter Triantafillou, András A. Benczúr, Scott Kirkpatrick, Philippe Rigaux, and Mark Williamson. 2011. Longitudinal Analytics on Web Archive Data: It’s About Time!. In *Conference on Innovative Data Systems*.

[29] Arkaitz Zubiaga, Victor Fresno, Raquel Martinez, and Alberto Perez Garcia-Plaza. 2013. Harnessing Folksonomies to Produce a Social Classification of Resources. *IEEE Transactions on Knowledge and Data Engineering* 25, 8 (2013).