# Tempurion: A Collaborative Temporal URI Collection 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 preserve the history of the Web and help users to access resources that may not be discoverable anymore by traditional web search engines due to changes or deletion. Navigating these vast archives without knowing the exact URI of interest has proven to be challenging. When typical information needs revolve around named entities, the retrieval and temporal ranking of related archived resources cannot be solved by simple full-text searches.

In this paper, we demonstrate a web platform that provides an ordered and annotated collection of URIs that characterize named entities over specific time frames. To not only rely on existing datasets, we have implemented interactive mechanisms 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. Tempurion: A Collaborative Temporal URI Collection for Named Entities. In *Proceedings of Joint Conference on Digital Libraries (JCDL'19).* ACM, New York, NY, USA, 2 pages.

## 1 INTRODUCTION

While exploring the Web from the past in Web archives, a user's intent differs significantly from the one using search engines to browse the live Web[2]. With the majority of queries being navigational instead of informational, users are more interested in browsing resources in the temporal dimension. Many of the issued queries further refer to named entities and reveal an especially high preference for older documents[3]. Providing only direct access via known URIs or a basic full-text search does not satisfy this use case sufficiently[1].

In our previous works we explored several ways to emphasize the temporal dimension of these archives by providing improved retrieval methods as well as search interfaces. This included the usage of user generated tags from social bookmarking systems and the indexation of anchor texts as a surrogate of the target resources[4, 5].

In this paper, we demonstrate the web platform *Tempurion* which shifts the focus from a broad spectrum exploration tool for web archives towards an annotated and topic related URI collection for named entities. The underlying dataset is based upon the integration of multiple sources such as entity classifications from DBpedia, URIs and tags from Wikipedia, Wikidata, Delicious and the German Web Archive as well as temporal enrichments from the Internet Archive's CDX index. Potential users are encouraged to contribute to the collection by providing additional resources and metadata or influence the ranking of results by voting.

## 2 APPLICATION AREAS

The system should enrich information on entities by providing a list of temporal resources. Here, we want to present a motivation base by outlining possible usage scenarios.

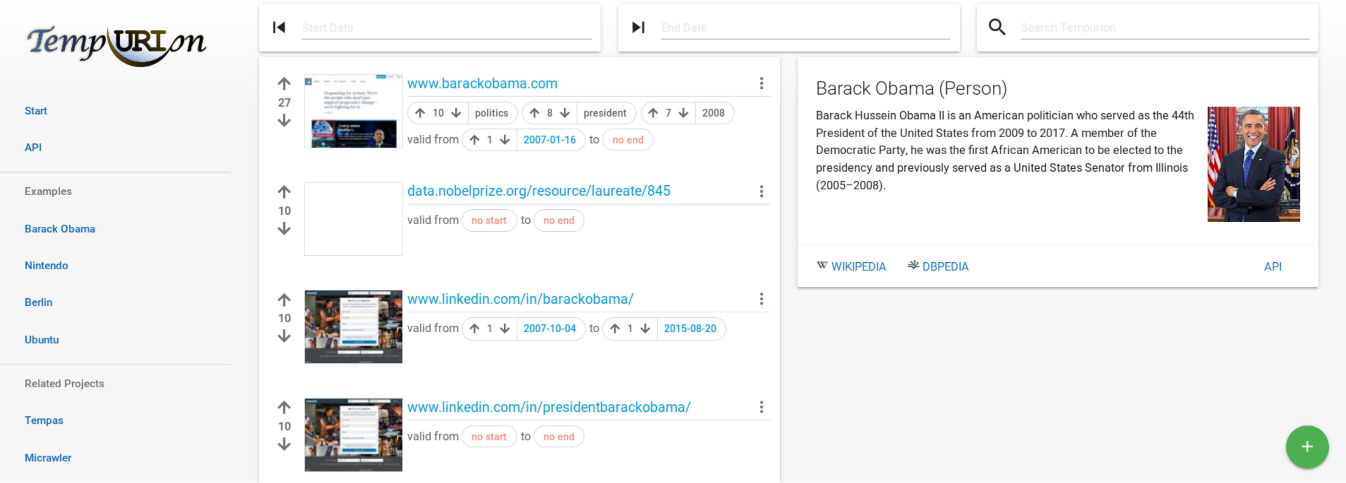
**Supporting Web Archives:** Frequent content changes and increasing amounts of data on the Web pose a particular challenge to its preservation by web archives. As a result, not all resources of an entity can be archived timely at their relevant date. On demand archive services like *archive.today*[1] or *Perma.cc*[2] provide snapshot mechanisms, but are not always easily explorable. The continuous provision of relevant and up-to-date URIs by users would result in a directive that is manageable compared to the dimensions of the web and that can be used to prioritize regular crawling targets. The same data can be used concurrently by "micro archives"[6] to capture digital representations of entities or even dynamically generate summary pages of a chosen time frame.

**Structuring the Web:** Instead of defining relationships of web pages according to their hyperlinks, the provided information enables us to semantically assign these resources to entities and enrich this relationship with metadata. This structure provides users with a way to navigate web archives without knowing a specific URI. In addition, the identification of documents in web archives that refer to the same entity allows for multiple analysis of these over time[8, 9]. For example, relationships between entities instead of individual texts can now be explored through shared resources. It could further group resources for the right entity even after name changes and would help towards tracking name evolution.

**Providing training data:** A meaningful and qualitative assessment of temporally relevant results for a search query represents a major problem in the area of *Temporal Information Retrieval*[7]. The construction of a database maintained by many users can create a reference dataset that can be further used for the training of advanced search methods.

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

[1] https://archive.today
[2] https://perma.cc


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 1: Entity result view in Tempurion**

## 3 SYSTEM OVERVIEW

A live version of Tempurion is accessible under:

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

The current dataset contains information to around 1.8 mio. entities with 13.9 mio. unique URIs and 20.6 million relations between them. Besides a search through entities and their related URIs, users are provided with the possibility to interact with all elements analogues to other folksonomies.

Exemplary, a result view for the entity "Barack Obama" is shown in Figure 1. The highest ranked results typically represent popular URIs that were found multiple times in the integrated datasets and often guide the user to homepages of these entities. To easier identify an entity and account for disambiguation, each listing shows an extract of the associated Wikipedia article and is categorized into one of the four main entity types person, organization, place or creative work.

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 if known from previous research. When trying to add already existing items like a specific URI or a tag results in an upvote for the respective element to prevent duplicates. In the case of a URI, the user is then additionally 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.

Public access to the underlying dataset is further provided in a machine-readable way via a RESTful API. We also created a more traditional search engine like view accessible under `/search-ui/`.

## 4 CONCLUSION

In this paper, we demonstrated an online system to retrieve temporal URI collections for named entities. Based on a combination of different datasets this system can improve the discoverability of resources in web archives and provides insights about the evolution of entities on the Web. Benchmarked against the search engine Bing, our approach achieves a remarkable precision of 83.3 % and shows promising results for high-quality lookups and temporal collection building. In the future, the differentiation of multiple languages could provide a better organization of the results and open this platform to a broader audience. This data will be constantly extended and evaluated by collaborative knowledge.

## ACKNOWLEDGMENTS

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

## REFERENCES

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

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

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

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

[5] Helge Holzmann, Wolfgang Nejdl, and Avishek Anand. 2016. On the Applicability of Delicious for Temporal Search on Web Archives. In *SIGIR*.

[6] Helge Holzmann and Mila Runnwerth. 2018. Micro Archives As Rich Digital Object Representations. In *WebSci*.

[7] Nattiya Kanhabua, Roi Blanco, Kjetil Nørvåg, et al. 2015. Temporal information retrieval. *Foundations and Trends® in Information Retrieval* 9, 2 (2015), 91–208.

[8] Marc Spaniol and Gerhard Weikum. 2012. Tracking entities in web archives: the LAWA project. In *WWW*.

[9] 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 *CIDR*.