A Collaborative Named Entity Focused URI Collection to Explore Web Archives

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

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

ABSTRACT

Vast amounts of data are stored by Web archives in order to preserve the history of digital mankind. But without ways to easily navigate and access resources of interest, their potential cannot be fully exploited. When full-text indexes seem unfeasible due to the size and additional temporal dimension, topic focused collections could provide structure and a starting point for many research questions. Here, we present a collaborative Web platform to collect and annotate URIs that characterize named entities over specific time frames. Initial data is provided by aggregating and evaluating multiple datasets and further enrichment with metadata.

1 INTRODUCTION

Navigating Web archives like the Internet Archive's *Wayback Machine*[1] can be difficult without knowing the exact URI and date of interest. To improve access, a site search based on anchor texts was implemented there[1], which guides the users to relevant domains for the entered keywords while ignoring the exact path and date.

In the past, 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[2, 3].

Here, we present the Web platform *Tempurion* (see Fig. 1) 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. Public access to the underlying dataset is further provided in a machine-readable way via a RESTful API. A live version is accessible under:

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

2 CONCLUSION

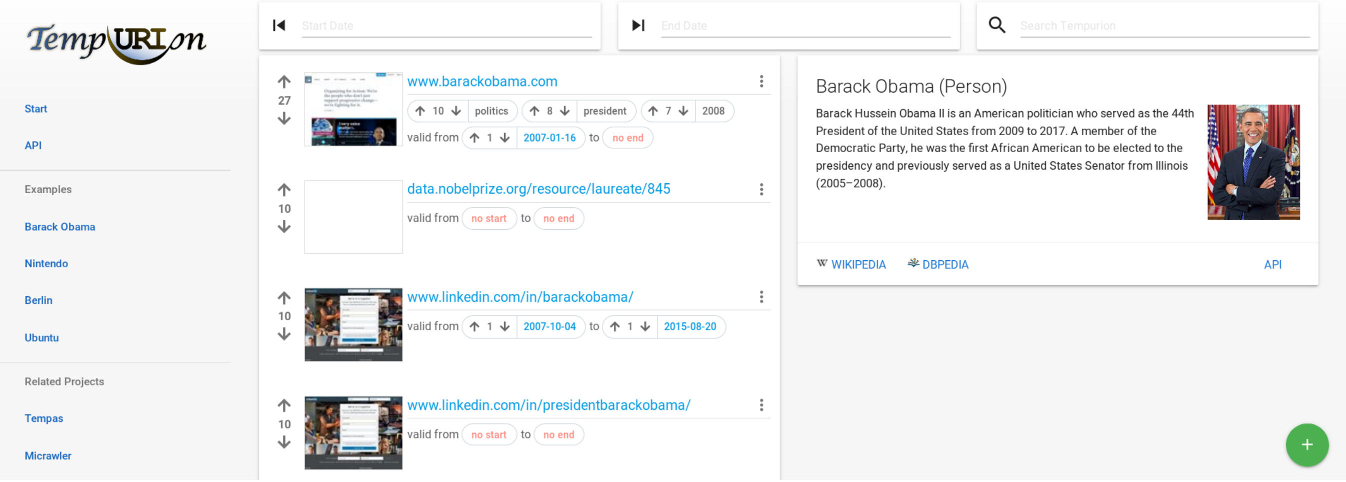
As we built this platform and collect initial data, we are interested in feedback from and collaboration with other researchers to enrich the collection, improve usability and integrate ideas to add more structure.

REFERENCES

[1] Vinay Goel. 2016. Beta Wayback Machine – Now with Site Search. Retrieved April 3, 2019 from https://blog.archive.org/2016/10/24/beta-wayback-machine-now-with-site-search/

[2] Helge Holzmann and Avishek Anand. 2016. Tempas: Temporal Archive Search Based on Tags. In *WWW, Companion Volume*.

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



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

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

[1] https://web.archive.org