

Accessing and Investigating Large Collections of Historical Newspapers with the NewsEye Platform

Axel Jean-Caurant, Antoine Doucet

▶ **To cite this version:**

Axel Jean-Caurant, Antoine Doucet. Accessing and Investigating Large Collections of Historical Newspapers with the NewsEye Platform. JCDL '20: The ACM/IEEE Joint Conference on Digital Libraries in 2020, Aug 2020, Virtual Event, China. pp.531-532, 10.1145/3383583.3398627 . hal-03026938

HAL Id: hal-03026938
https://hal.science/hal-03026938

Submitted on 7 Nov 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Accessing and Investigating Large Collections of Historical Newspapers with the NewsEye Platform

Axel Jean-Caurant
axel.jean-caurant@univ-lr.fr
University of La Rochelle, L3i laboratory
La Rochelle

Antoine Doucet
antoine.doucet@univ-lr.fr
University of La Rochelle, L3i laboratory
La Rochelle

ABSTRACT

The NewsEye project demonstrator is a proof of concept of a digital platform dedicated to historical newspapers, intended to show benefits for researchers and the general public. This platform presently hosts newspapers from partner libraries in four different languages (Finnish, Swedish, German and French) providing users with various analysis tools as well as allowing them to manage their research in an interactive way. The platform gives access to these enriched data sets, and additionally interfaces with analysis tools developed in the NewsEye project, letting users experiment with tools specifically developed for investigating historical newspapers.

CCS CONCEPTS

• **Information systems** → *Search interfaces.*

KEYWORDS

digital library, historical newspapers, user interface

ACM Reference Format:
Axel Jean-Caurant and Antoine Doucet. 2020. Accessing and Investigating Large Collections of Historical Newspapers with the NewsEye Platform. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020 (JCDL '20), August 1–5, 2020, Virtual Event, China.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3383583.3398627

1 INTRODUCTION

In the last decades, millions of newspaper pages have been digitized by European libraries. This massive amount of data is made available to researchers as well as the general public through digital portals managed by each of the libraries. This large number of access points can be an issue for users who have to repeat their search as many times as the number of digital portals. Efforts have already been made to gather all this data in a single digital library. The Europeana project is one example of such a digital library. Several millions newspaper pages were made available through the Europeana portal (along with manuscripts, Art pieces and other digitized objects). However, the general aspect of such a website does not provide specialized users with powerful search and analysis tools. The NewsEye project [2] aims at solving this by providing users with enhanced newspaper data (article separation, named entity linking, topic models, ...) as well as powerful analysis tools that can produce relevant results from multilingual data.

2 THE NEWSEYE PIPELINE

The main goals of the NewsEye project are to provide researchers with new means of exploring and exploiting large amounts of newspaper materials by offering tools and methods that combine close reading and distant reading. Numerous challenges need to be resolved in order to provide digital humanities (DH) scholars with data usable for their research. Thanks to the three European libraries which are partners of this project (National Library of France, National Library of Finland and National Library of Austria), datasets have been made available in four different languages: Finnish, Swedish, German and French. These datasets are processed to extract structured information from the images of newspaper pages provided by the partner libraries. A processing pipeline has been set up to produce and analyse enhanced data, starting from digitised documents (i.e. pictures of newspaper pages). The first step is to run OCR on the images to recover textual and layout information. Next, individual articles are tagged to provide logical structure adequate to historical newspapers. Then, named entity mentions are extracted and linked to a knowledge base (Wikidata is being used). Named entities are crucial for information access as they are the main entry point of user search [1, 3]. Moreover, linking the entities to a crosslingual knowledge base is a good way to connect the documents across linguistic borders.

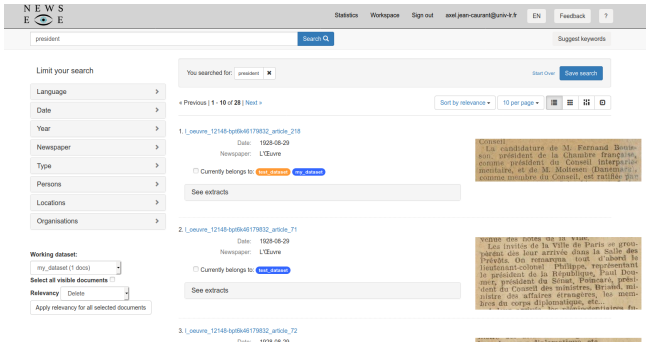


Figure 1: The search engine results page of the website.

3 THE NEWSEYE DEMONSTRATOR

A public website [1] was created to showcase the different outputs of the project, including its data or tools. The demonstrator presented

[1] https://www.newseye.eu, visited on 4 April 2020

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
JCDL '20, August 1–5, 2020, Virtual Event, China
© 2020 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-7585-6/20/06.
https://doi.org/10.1145/3383583.3398627

hereby is available through a direct link[2]. The features of this web interface stem from the needs expressed by the DH scholars of the project. For instance, one such need is the availability of a personal workspace. Users indeed can create their own datasets which gather documents relevant to a particular research question they are interested in. The benefit of using datasets is that it allows users to save their research process to be continued in a different session. Moreover, datasets are at the base of various processing that can be performed within and through the platform.

3.1 Search engine and user workspace

One of the contribution of the project is to improve access to digitized newspapers. This is first achieved by allowing users to search for information in a similar way as they can in most digital libraries. The search interface is visible on Figure 1. The left side of this interface provides facets, which allow users to filter the results against various bibliographic metadata (language, date, ...) as well as named entities. On this page, users can also see if the documents returned by the search belong to previously created datasets. If this is the case, a coloured tag is signaling it. The colour of this tag depends on the relevance of the document with respect to the dataset. As a matter of fact, when users want to add a document to a dataset, they have to select its degree of relevance ("not relevant", "somewhat relevant", "relevant", "very relevant"). These relevance values can be used by various processing tools to leverage the importance of documents in a dataset. Users also have the possibility of modifying and exploiting their created datasets by various means. They can merge datasets, split an existing dataset or even export the content of a dataset to be used with tools external to the platform.

3.2 Topic modelling tools

The web interface allows users to exploit topic models trained on the data available. On the one hand, these models can be visualised interactively by the user to understand the content of the detected topics. On the other hand, these models can be queried manually by the user to get information on a previously created dataset. As of now, users have the possibility to inspect the topic distribution of a previously created dataset to get a better insight on its content. From these topic distributions, users also have the possibility to discover new documents with similar topic distributions. This is a way to obtain potentially relevant documents without going through a new search.

3.3 Personal Research Assistant

One of the main innovations of the NewsEye project is the Personal Research Assistant (PRA). This tool is meant to help users navigate through the large number of newspapers available, by automatically building new searches and producing detailed reports on its findings. It is composed of three components. The *investigator* is in charge of building new searches automatically, the *reporter* generates natural language reports on those findings while the *explainer* mitigates blackbox effects by detailing how the investigator arrived to its conclusions. The capability for autonomous investigation can be a serious advantage in historical analysis, where it is important to stay objective when studying a particular research question. The various tools that can be used on a dataset are integrated in the platform and presented as "tasks". Because the results of these tools may take time to compute, the personal workspace allows to list the tasks that are running. When the tasks are completed, users can access their results. Figure 2 presents the results of a dataset description task. To avoid a black-box effect, a dedicated component is being developed to explain the different steps and suggestions made by the PRA during an autonomous analysis task.

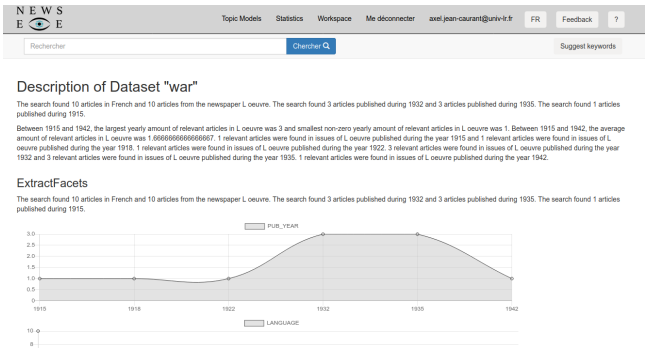


Figure 2: Several tools are used by the PRA to describe a dataset. The automatically generated reports allow users to get an overview of the content of a dataset.

4 CONCLUSION AND FUTURE WORKS

This paper presented the current state of the interface developed to showcase the different outputs of the NewsEye project. Its key objective is to facilitate research over historical newspapers for DH scholars and the general public. In the near future, we plan to improve the ergonomics of this demonstrator as well as to increase the number of documents available. As it is, the PRA can already be used to gain insights on the content of a document collection. In the future, several analysis tools will be added, as well as autonomous functionalities, allowing the PRA to find new documents on its own, and to make relevant suggestions to the end user. Another improvement will concern the way analysis results are visualised. Each of the tools, which can both be accessed directly by users or through the PRA, will need tailored visualisation interfaces to increase their impact.

ACKNOWLEDGMENTS

This work has been supported by the European Union Horizon 2020 research and innovation program under grants 770299 (NewsEye).

REFERENCES

[1] Guillaume Chiron, Antoine Doucet, Mickaël Coustaty, Muriel Visani, and Jean-Philippe Moreux. 2017. Impact of OCR errors on the use of digital libraries: towards a better access to information. In *Proceedings of the 17th ACM/IEEE Joint Conference on Digital Libraries*. IEEE Press, 249–252.

[2] Antoine Doucet, Martin Gasteiner, Mark Granroth-Widling, Max Kaiser, Minna Kaukonen, Roger Labahn, Jean-Philippe Moreux, Guenter Muehlberger, Eva Pfanzelter, Marie-Eve Therenty, Hannu Toivonen, and Mikko Tolonen. 2020. NewsEye: A Digital Investigator for Historical Newspapers. In *Digital Humanities 2020, DH 2020, Conference Abstracts, Ottawa, Canada, July 22-24, 2020*. Alliance of Digital Humanities Organizations (ADHO).

[3] Paul Gooding. 2014. Exploring Usage of Digital Newspaper Archives through Web Log Analysis: A Case Study of Welsh Newspapers Online. (2014).

[2]https://platform.newseye.eu, visited on 4 April 2020