Education for Information 39 (2023) 139–153
DOI 10.3233/EFI-230026
IOS Press

# An OER on digital historical research on European historical newspapers with the NewsEye platform

Cyrille Suire, Nicolas Sidère and Antoine Doucet*
*L3i Laboratory, Faculty of Science and Technology, La Rochelle Université, La Rochelle, France*

In this article, we introduce an Open Education Resource (OER) on digital historical research with historical newspapers,[1] intended to give students the means to understand the induced risks in working with large collections of digitised documents, as well as the keys to benefit from the advances of natural language processing over large multilingual collections of European historical newspapers. This resource exploits results of the NewsEye Horizon 2020 research and innovation project. It is part of a set of 7 OERs developed and shared within the Erasmus+ project Digital Methods Platform for Arts and Humanities (DiMPAH).

Keywords: Digital humanities, natural language processing, document analysis, historical newspapers

## 1. Introduction

Newspapers have been gathering information on cultural, political, and social events in a more comprehensive manner compared to other public records since the 17th century. They have documented billions of events, stories, and personal names every day, spanning across various languages and countries.[2]

In this article, we introduce an Open Education Resource (OER) intended to help digital humanities students and scholars to learn about the way large collections of digitised documents are automatically processed with natural language processing methods. This way, we aim to provide them with the ability to develop a good understanding of the opportunities, risks and limits that exploiting such collections implies. The OER is focused on historical newspapers, as they form a very good use case of documents that are numerous, and in very heterogeneous states of conservation.

In this section, we provide context on the analysis of historical, and the recent advances allowed by projects such as the recent Horizon 2020 NewsEye project[3] (Doucet et al., 2020). Section 2 describes the OER developed in this context, with

*Corresponding author: Antoine Doucet, L3i Laboratory, Faculty of Science and Technology, La Rochelle Université, Av. Michel Crépeau, 17042 La Rochelle, France. Tel.: +33 546456871; E-mail: antoine.doucet@univ-lr.fr.

[1] Available on DariahTeach: https://teach.dariah.eu/.

[2] https://www.newseye.eu/about/.

[3] https://www.newseye.eu/.

further details on its theoretical background in Section 3. Examples of what can be achieved with such tools are provided through digital humanities research use cases in Section 4, before we conclude and reflect in Section 5.

### *1.1. Context: Newspapers as big data of the past*

In many parts of the world and particularly in Europe, many efforts have been made to make this endless source of information available to all. Many countries have created digital libraries that make their collections available in digital form. In the United States, for example, the Library of Congress provides access to its online collection of historical newspapers, Chronicling America.[4] In France, the Bibliothèque Nationale de France (BNF) and its digital library Gallica[5] and Retronews[6] offer this service and provide access to over 5 million issues of newspapers and magazines. The Austrian National Library offers in its digital library access to 1400 press titles and more than 24 million pages through its dedicated newspaper platform ANNO.[7] New content is added regularly by all major digital libraries.

The volume of data available is thus very large. To navigate through this data, the websites of the major digital libraries offer search engines and filtering functions. It is possible to consult precise newspaper titles, on well-defined chronological periods or on the basis of keywords. These digital libraries also often offer the possibility to export the results and to access the raw texts.

The digitization of historical newspapers is therefore a great asset for research and for the digital humanities but the process that leads to the availability of text extracted from historical newspapers is a very complex one that requires many processing steps. At each stage of this process, scientific, methodological or technical problems may arise. In most cases, one does not know what process the data has undergone and is therefore faced with a "black box" problem, where automated outputs are difficult to understand and explain.

In order to be able to analyze historical newspapers while guaranteeing reliable results, it is essential to understand the nature of the processes undergone by the data and their impact on your research. In a 2013 article whose conclusions remain valid, Canadian historian Ian Milligan (2023), made the following observation:

> "It all seems so orderly and comprehensive. Instead of firing up the microfilm reader to navigate the Globe and Mail or the Toronto Star, one needs only to log into online newspaper databases. A keyword search, for a particular event, person, or cultural phenomenon, brings up a list of research findings. Previously impossible research projects can now be attempted. This process has fundamentally

[4] https://chroniclingamerica.loc.gov/.
[5] http://gallica.bnf.fr.
[6] http://www.retronews.fr.
[7] https://anno.onb.ac.at/.

> reshaped Canadian historical scholarship. We can see this in Canadian history dissertations. In 1998, a year with 67 dissertations, the Toronto Star was cited 74 times. However it was cited 753 times in 2010, a year with 69 dissertations."

The importance of newspapers as cultural heritage is thus irrefutable, but whilst some progress has been made concerning the digitisation of newspapers, it is their potential as data which opens up new possibilities for their exploration and analysis using digital methods. However, to truly unleash this potential, critical hurdles need to be passed, such as the masses of data that can cannot be grasped without automation, and the alternative risk of automated processes that cannot be to interpreted or that perform counterproductive operations on the collections. Unless these problems are solved, very large digitised collections can be made available but will nonetheless remain unpractical to access and analyse for digital humanities scholars and the general public.

### *1.2. The NewsEye project*

However, recent progress of dedicated projects such as NewsEye offer new ways not only to read, but also for access and in-depth analysis of large collections of European historical newspapers. NewsEye was a project funded by the European Commission through the Horizon 2020 programme (grant number 770299), running from May 2018 to January 2022. While the project formally ended, its results continued to be expanded upon, notably on its newspaper platform, the gateway where most of its results are showcased (Jean-Caurant & Doucet, 2020).

This OER relies on the results of the NewsEye project, coordinated by La Rochelle University and involving several universities (University of La Rochelle, University of Helsinki, University of Innsbruck, University of Rostock, Université Paul-Valéry-Montpellier and University of Vienna) and the national libraries of Austria, Finland and France.

The key goal of the project was to develop tools and methods that would "change the way European digital heritage data is (re)searched, accessed, used and analysed". It focused on 1.5 million pages from selected issues of historical newspapers in five languages (Finnish, French, German, Swedish and English) from the late 19th to the mid 20th century. The main objective was to develop a set of tools and methods for the effective exploration and exploitation of newspaper collections by means of new technologies and big data approaches. The project aimed "to improve the users' capability to access, analyse and use the content contained in vast corpora of digitized historical newspapers".

NewsEye was a collaborative and interdisciplinary undertaking in digital humanities (Oberbichler et al., 2022), in which computer science research groups developed tools to analyse library collections in several languages, with the strong implication and feedback of humanities scholars in several fields, working on different research use cases on subcollections in different languages. The project functioned in a loop of feedback and improvement, using the diversity of the collections and humanities research topics as confirmation that the approaches could be generalised to any other language, discipline and research topic.

The main aim of NewsEye was to provide new and innovative tools and services to improve the accessibility, exploration, and analysis of historical newspapers. This was achieved by utilizing one of the biggest and most important digital collections of cultural heritage in Europe. The project aimed to have a wide-reaching impact and be used by a diverse range of users. The outcome of the project was a useful and cost-effective toolbox and demonstrator platform for assisting users of all types, available as open science through the project's Github repository[8] while public datasets and models were made available through Zenodo. The developed workflow was structured into four main layers, each of which provides advanced tools and techniques for:

- Text Recognition (Michael et al., 2019) and Article Separation (Michael et al., 2020), extracting the layout of newspapers (e.g. articles and graphical regions) from digitized newspapers and transforming the content to textual format, providing full articles through automatic layout analysis, text recognition and article separation.
- Semantic Text Enrichment, enhancing the utility of the newspaper collections by enriching the texts with higher-level semantic annotation using named entity recognition (NER) (Boros et al., 2022; Boros et al., 2020). Extracted named entities were linked to external references (such as the Wikipedia) across languages (Pontes et al., 2022; Pontes et al., 2019), with the goal to provide support for multilingual analysis and enable event detection through the discovery of patterns in textual content.
- Dynamic Text Analysis, offering tools to utilize the enriched data for in-depth analysis of user-selected newspaper content, facilitating interactive queries to discover different viewpoints (Hamdi et al., 2021), sub-topics (Zosa & Granroth-Wilding, 2019) or trends concerning the selected topic (Mutuvi et al., 2018; Zosa et al., 2021), named entity, newspaper, timeframe or other category, so as to provide insights into the newspaper collection in contextualized and comparative manners.
- Intelligent analysis and reporting ("Personalized Research Assistant" (Pivovarova et al., 2020)), offering an "intelligent" interface to the data and other tools, providing iterative analysis and reporting to the user in natural language. Users can authorize the Assistant to investigate a specific topic, time window, or newspaper on their behalf, and receive potentially interesting findings reported back to them in natural language. The Assistant uses multilingual natural language generation to produce textual descriptions of the results obtained by the Investigator, enabling reporting in multiple languages (Leppanen & Toivonen, 2021). The findings are reported in a transparent manner, allowing users to understand and verify them.

[8] https://github.com/NewsEye/.

The NewsEye consortium further involved experts whose role was to ensure (i) additional technical expertise in the above-mentioned aspects, (ii) access to and enrichment of digitized newspapers, (iii) insight and experience in using historical newspapers as a rich cultural heritage resource for the understanding of developments in society, economy and politics, (iv) use cases with the aim to address important humanities' research desiderata and gain experience and feedback to guide iterative development of the NewsEye demonstrator, and (v) strong dissemination and viable paths towards wider adoption and sustainability of the developed tools. All the results and outputs of the project are available on the project website, notably with data sets, publications and source code inventoried under its "Open Science" tab.[9]

## 2. OER summary

This OER is developed within the DiMPAH project (Digital Methods Platform for Arts and Humanities). DiMPAH is an Erasmus+ Strategic Partnerships for Higher Education (KA203), aiming to aggregate, connect and make widely available novel OERs on selected digital methods, apply these to interdisciplinary contexts and foster novel creative learning experiences by taking data from the past into future stories.[10]

Building on the NewsEye toolbox, platform and personal research assistant, this OER serves as an introduction to document analysis and understanding in the context of a historical document written in any language. This introduction focuses on highlighting the difficulties of using documents in a research context and the key points that the researcher must consider.

In terms of document analysis, the topics covered are the background of digitisation and the conversion process from scans (pictures) into structured (in the case of newspapers, mainly centred on articles) and interpretable contents (following optical character recognition which converts pictures into text). In terms of document understanding, in the context of 'noisy' documents (imperfect article separation and text recognition are unavoidable) that may be written in any language, the course covers the subsequent challenges and methods for event detection and semantic enrichment. One example is the extraction of named entities (persons, locations, organisations) and connecting them to a cross-lingual resource such as Wikidata using named entity linking and other text analytics techniques (like stance analysis, covered in another DiMPAH OER titled "Text Analysis: Linguistic Meets Data Science").

The main innovation of the OER is to specifically address the qualitative analysis of historical documents, and provide keys to understanding how this is done and the subsequent limits. The course is therefore tailored for students with a background in

[9] https://www.newseye.eu/open-science/.

[10] https://lnu.se/en/research/research-projects/project-digital-methods-platform-for-arts-and-humanities/.

humanities, who will be given sufficient information to be fully aware of the limits of such methods. For example, since optical character recognition is imperfect, a search engine keyword query will not necessarily provide all the matching documents since keywords may be improperly recognised in a number of documents (Chiron et al., 2017). We believe that it is essential for DH students to get a good understanding of how the algorithms for document analysis function, in order to get an informed understanding of their capacities and limits.

The first part of the course provides background information on document analysis and understanding. In strong connection with other DiMPAH OERs ("Digital Methods in the Humanities: An Interactive Exploration and Guide for Students" and "Text Analysis: Linguistic Meets Data Science"), it will expand on the concrete use case of the analysis of historical documents. Humanities students will be given the keys to an informed understanding of the strengths and weaknesses of analysing historical documents, in particular newspapers. The materials for this part of the course have been developed from project start, in close collaboration with other OERs, so as to guarantee the pedagogical continuity across the DiMPAH OERs. Indeed, the other two mentioned OERs provide background on text analysis and digital methods in general, while this one focuses on the specific practical application to historical newspapers.

The second part of the course is built on the NewsEye toolbox, platform and personal research assistant. Two use cases (described in Section 4) are offered to students:

- Covid-19 and Spanish flu: A study of media and political discourses and
- Women in Pants: A study of the emergence of new clothing style for women in 19th et 20th centuries.

The two case studies have been designed to explain the impact of document processing steps on end-user research work. To demonstrate the most theoretical aspects, each case study is also associated with operating examples in the form of Jupyter Notebooks.

## 3. Theoretical content of the OER

The theoretical and technical works developed and presented in this OER are plenty. It is important that this section is only an overview. We present in particular details named entity recognition, one of the key tools of the OER.

### *3.1. Overview of document analysis and understanding*

Document understanding (DU) refers to a system's capability to automatically process documents, encompassing technologies that can comprehend and extract both text and meaning from diverse document types, which may include structured, semi-structured, and unstructured formats.

DU models receive documents as input and break down the pages of the documents into relevant parts, such as regions that correspond to specific tables or properties, typically utilizing optical character recognition (OCR) and some form of document layout analysis. These techniques utilize this information to comprehend the overall contents of the document, for instance, identifying that a particular bounding box or region represents an address or a news article.

Some examples of DU topics we developed in the OER are the following research tasks:

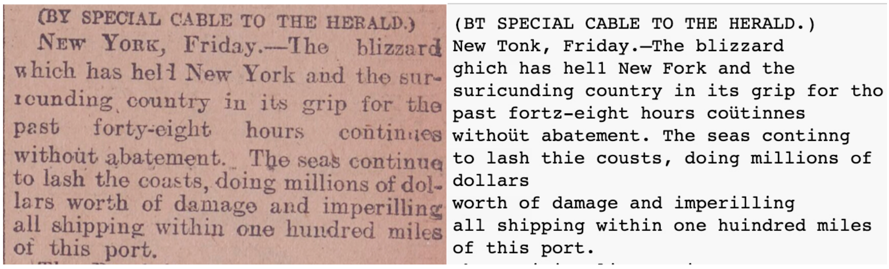
- Document Layout Analysis (DLA) is a module based on computer vision that analyzes the layout of a document by segmenting each page into discrete content regions. The purpose of this model is not only to differentiate between relevant and irrelevant regions, but also to classify the type of content it identifies.
- Optical Character Recognition – The purpose of this step is to extract text from images or PDF files by faithfully transcribing all written text present in the document.
- Information extraction (IE) is the process of converting unstructured text into a structured database, which contains selected information from the text. This step is crucial in making the text's information content usable for further processing. These models typically use the results of OCR or document layout analysis to understand and establish relationships between the information that is conveyed in the document. Generally specialized for a specific domain and task, these models provide the required structure to make a document machine-readable, rendering it useful for document understanding.
- Document Semantic Enrichment (DSE) – This process uses the extracted data from the previous step. These annotations are used to enrich the document metadata and to provide new types of visualizations in an information retrieval context.

### *3.2. Examples of document analysis and understanding tasks in the context of historical newspapers*

**Optical Character Recognition,** also called Automated Text Recognition (ATR) is the task of automatically transcribing the textual contents contained in images. This is a fundamental step in all digitization processes, as it provides the textual data to be further processed by Natural Language Processing tasks.

As seen in Fig. 1, documents that have been processed through OCR contain a significant number of errors, which are typically the result of the document's condition. This may be due to factors such as aging, inadequate storage conditions, or the use of low-quality printing materials. These errors can significantly reduce the efficiency of all downstream natural language processing tasks.

Moreover, in the specific case of digitized historical newspapers, a step related to OCR is necessary to fully exploit the potential of the data.

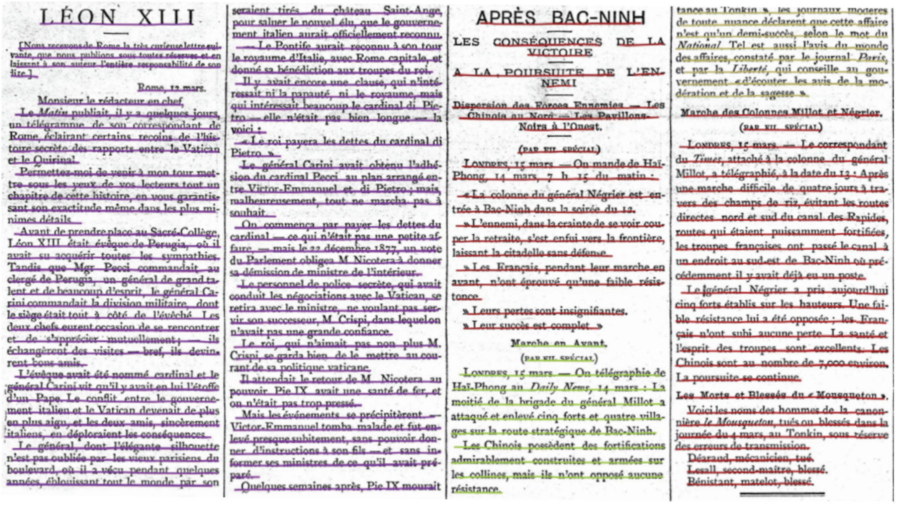


(BY SPECIAL CABLE TO THE HERALD.)

NEW YORK, Friday.—The blizzard which has held New York and the surrounding country in its grip for the past forty-eight hours continues without abatement. The seas continue to lash the coasts, doing millions of dollars worth of damage and imperilling all shipping within one hundred miles of this port.

```
(BT SPECIAL CABLE TO THE HERALD.)
New Tonk, Friday.—The blizzard
ghich has hell New Fork and the
suricunding country in its grip for tho
past fortz-eight hours coütinnes
withoüt abatement. The seas continng
to lash thie cousts, doing millions of
dollars
worth of damage and imperilling
all shipping within one huindred miles
of this port.
```

Fig. 1. A digitized crop from NY Herald and its OCRed version (including errors).



## LÉON XIII

[Nous recevons de Rome la très curieuse lettre suivante, que nous publions sous toutes réserves et en laissant à son auteur l'entière responsabilité de son dire.]

Rome, 12 mars.

Monsieur le rédacteur en chef,

Le *Matin* publiait, il y a quelques jours, un télégramme de son correspondant de Rome, éclairant certains recoins de l'histoire secrète des rapports entre le Vatican et le Quirinal.

Permettez-moi de venir à mon tour mettre sous les yeux de vos lecteurs tout un chapitre de cette histoire, en vous garantissant son exactitude même dans les plus minimes détails.

Avant de prendre place au Sacré-Collège, Léon XIII était évêque de Perugia, où il avait su acquérir toutes les sympathies. Tandis que Mgr Pecci commandait au clergé de Perugia, un général de grand talent et de beaucoup d'esprit, le général Carini commandait la division militaire, dont le siège était tout à côté de l'évêché. Les deux chefs eurent occasion de se rencontrer et de s'apprécier mutuellement; — ils échangèrent des visites — bref, ils devinrent bons amis.

L'évêque avait été nommé cardinal et le général Carini vit qu'il y avait en lui l'étoffe d'un Pape. Le conflit entre le gouvernement italien et le Vatican devenait de plus en plus aigu, et les deux amis, sincèrement italiens, en déploraient les conséquences.

Le général, dont l'élégante silhouette n'est pas oubliée par les vieux parisiens du boulevard, où il a vécu pendant quelques années, éblouissant tout le monde par son seraient tirés du château Saint-Ange, pour saluer le nouvel élu, que le gouvernement italien aurait officiellement reconnu. — Le Pontife aurait reconnu à son tour le royaume d'Italie, avec Rome capitale, et donné sa bénédiction aux troupes du roi.

Il y avait encore une clause, qui n'intéressait ni la papauté, ni le royaume, mais qui intéressait beaucoup le cardinal di Pietro — elle n'était pas bien longue — la voici :

« Le roi payera les dettes du cardinal di Pietro. »

Le général Carini avait obtenu l'adhésion du cardinal Pecci au plan arrangé entre Victor-Emmanuel et di Pietro ; mais, malheureusement, tout ne marcha pas à souhait.

On commença par payer les dettes du cardinal — ce qui n'était pas une petite affaire — mais le 22 décembre 1877, un vote du Parlement obligea M Nicotera à donner sa démission de ministre de l'intérieur.

Le personnel de police secrète, qui avait conduit les négociations avec le Vatican, se retira avec le ministre, ne voulant pas servir son successeur, M Crispi, dans lequel on n'avait pas une grande confiance.

Le roi, qui n'aimait pas non plus M. Crispi, se garda bien de le mettre au courant de sa politique vaticane.

Il attendait le retour de M. Nicotera au pouvoir. Pie IX avait une santé de fer, et on n'était pas trop pressé.

Mais les événements se précipitèrent — Victor-Emmanuel tomba malade et fut enlevé presque subitement, sans pouvoir donner d'instructions à son fils — et sans informer ses ministres de ce qu'il avait préparé.

Quelques semaines après, Pie IX mourait

## APRÈS BAC-NINH

LES CONSÉQUENCES DE LA VICTOIRE

A LA POURSUITE DE L'ENNEMI

Dispersion des Forces Ennemies — Les Chinois au Nord — Les Pavillons-Noirs à l'Ouest.

(PAR FIL SPÉCIAL)

LONDRES, 15 mars — On mande de Haï-Phong, 14 mars, 7 h. 15 du matin :

« La colonne du général Négrier est entrée à Bac-Ninh dans la soirée du 12.

» L'ennemi, dans la crainte de se voir couper la retraite, s'est enfui vers la frontière, laissant la citadelle sans défense.

» Les Français, pendant leur marche en avant, n'ont éprouvé qu'une faible résistonce.

» Leurs pertes sont insignifiantes.

» Leur succès est complet. »

Marche en Avant.

(PAR FIL SPÉCIAL)

LONDRES, 15 mars — On télégraphie de Haï-Phong au *Daily News*, 14 mars : La moitié de la brigade du général Millot a attaqué et enlevé cinq forts et quatre villages sur la route stratégique de Bac-Ninh.

Les Chinois possèdent des fortifications admirablement construites et armées sur les collines, mais ils n'ont opposé aucune résistance.

tance au Tonkin », les journaux modérés de toute nuance déclarent que cette affaire n'est qu'un demi-succès, selon le mot du *National*. Tel est aussi l'avis du monde des affaires, constaté par le journal *Paris*, et par la *Liberté*, qui conseille au gouvernement « d'écouter les avis de la modération et de la sagesse ».

Marche des Colonnes Millot et Négrier.

(PAR FIL SPÉCIAL)

LONDRES, 15 mars. — Le correspondant du *Times*, attaché à la colonne du général Millot, a télégraphié, à la date du 13 : Après une marche difficile de quatre jours à travers des champs de riz, évitant les routes directes nord et sud du canal des Rapides, routes qui étaient puissamment fortifiées, les troupes françaises ont passé le canal à un endroit au sud-est de Bac-Ninh où précédemment il y avait déjà eu un poste.

Le général Négrier a pris aujourd'hui cinq forts établis sur les hauteurs. Une faible résistance lui a été opposée ; les Français n'ont subi aucune perte. La santé et l'esprit des troupes sont excellents. Les Chinois sont au nombre de 7,000 environ. La poursuite se continue.

Les Morts et Blessés du « Mousqueton ».

Voici les noms des hommes de la canonnière *le Mousqueton*, tués ou blessés dans la journée du 4 mars, au Tonkin, sous réserve des erreurs de transmission.

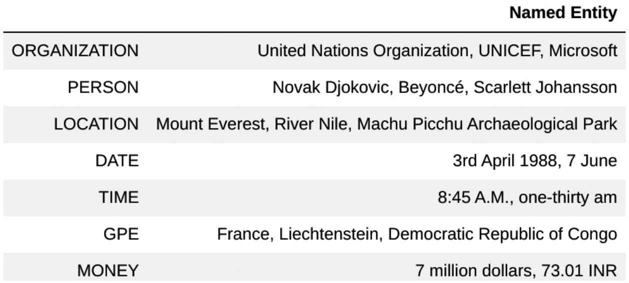
Déarand, mécanicien, tué.
Lesall, second-maître, blessé.
Bénistant, matelot, blessé.

Fig. 2. Example of an article separation output.

**Article Separation (AS)** refers to the process of identifying the sections of a newspaper page that contain information relevant to a single article, which is a crucial step for natural language processing applications in the information retrieval pipeline. However, the AS task can be divided into two parts: block recognition and subsequent clustering.

The accurate performance of this step plays a crucial role in our ability to query not only entire newspaper pages, but also individual articles. This makes it considerably easier to construct a corpus for further analysis. An example of Article Separation is given in Fig. 2.

**Semantic Text Enrichment** involves analyzing documents and augmenting their contents with semantic metadata. In the case of historical newspapers, this typically involves focusing on named entities, which are proper nouns that refer to specific entities such as people, organizations, locations, and dates. Some additional examples of named entities are provided in Fig. 3.



| | Named Entity |
|---|---|
| ORGANIZATION | United Nations Organization, UNICEF, Microsoft |
| PERSON | Novak Djokovic, Beyoncé, Scarlett Johansson |
| LOCATION | Mount Everest, River Nile, Machu Picchu Archaeological Park |
| DATE | 3rd April 1988, 7 June |
| TIME | 8:45 A.M., one-thirty am |
| GPE | France, Liechtenstein, Democratic Republic of Congo |
| MONEY | 7 million dollars, 73.01 INR |

Fig. 3. Examples of several types of named entities.

To be more specific, Semantic Text Enrichment aims to identify these entities within documents, disambiguate them to a knowledge base, and associate them with the stance in which they are mentioned.

Named Entity Recognition is one of the most important and most studied tasks in NLP. It has many industrial and scientific applications. It can clearly facilitate information retrieval in historical newspapers and is one the most important topic developed in the OER.

**Named entity recognition.** Named entities are typically proper nouns that refer to specific entities such as people, organizations, locations, dates, and so on. Let us use the following example sentence: “Mount Everest is the tallest mountain above sea level”. “Mount Everest” should be detected as a named entity of type location. Other examples of named entities can be found in Fig. 3.

The definition of a named entity (NE) in a text is subjective and varies depending on the desired information to be extracted. However, the most commonly used set of named entity classes consists of three basic entity types: person (PER), location (LOC), and organization (ORG). These three entity types are collectively referred to as enamex, a term that originated from the MUC-6 competition (Grishman & Sundheim, 1996).

In our context, NER is used for **exploring historical documents** Tools like NER can be extremely valuable to researchers, historians, or librarians for adding structure to the volumes of unstructured data and for improving access to the historical digitized collections. For instance, by conducting a basic keyword search, historians can quickly determine whether a collection has useful material for their research, which can save them a considerable amount of time searching through archives and reviewing pages. NER can identify person names and locations, which are frequently prominent in news articles where people play a crucial role in the events reported. For example, the EU’s digital platform for cultural heritage, Europeana,[11] is using NER to make historical newspapers searchable. The detection of entities can be considered as a first step in the exploration of data collections. This task is of course useful for the analysis of digitised historical newspapers, but is also crucial in many other cases, such as:

[11] https://www.europeana.eu/.

- Classifying content for news providers
- Automating customer support
- Extracting valuable information from medical documents
- Aiding risk assessment for financial institutions
- Easing the research process

*Text classification*

Text classification is the process of categorizing text into pre-defined groups. By using natural language processing, text classifiers can automatically analyse text and then assign a set of given categories based on the research question. This automated classification of text into predefined categories is an important method for managing and processing a large number of newspaper clippings. This also applies to subcorpora for a specific research topic. The aim of this task is to train a model using previously created training/test corpus and to use this model to get an overview of the category distribution throughout a collection. Another goal is to export categorized data for further analysis. This makes it possible to examine, for example, the advertisement about a specific topic.

*Event extraction*

Event extraction is a type of information extraction (IE) that involves extracting specific information from texts about certain incidents. The task involves two subtasks: event detection (ED) and event argument extraction. ED involves extracting critical information, such as a keyword, a sentence, a phrase, or a text span that evokes the event being discussed in the text. For instance, a news article could mention a recent outbreak of an epidemic or the appointment of a new president. In such cases, the events to be identified are typically denoted by the name of the epidemic or the term "election." Event argument extraction focuses on extracting more detailed information about the event, such as the location, participants, and other related details. Event extraction aims to answer the 5W1H questions (who did what, when, where, why, and how), which can describe the presence of events in an article. For instance, the location of the epidemic even, the name of the president or the country of the election would need to be detected in event argument extraction.

Events are a natural structuring concept, as they tie together time, space, and participants, making them an important consideration when dealing with historical texts.

*Stance detection*

Stance detection is a subfield of natural language processing that aims to determine whether the author of a piece of text is in favor of, against, or neutral towards a target entity (such as a person or organization) mentioned or implied in the text.

In NewsEye, this task involves analyzing news articles to determine whether they have a positive, negative, or neutral stance towards a named entity mentioned in the article. However, there are challenges in accurately determining stance, such as correctly identifying named entities referred to by pronouns and dealing with errors and inconsistencies (ATR) in historical texts caused by automated text recognition and spelling differences. These factors can affect the performance of contemporary polarity lexicons used to determine the sentiment of historical words.

## 4. Use cases

To practice the notions, methods and tools developed in the OER, we have chosen to develop two case studies, the media and political discourse of the Spanish flu pandemic on the one hand and the emergence of new clothing styles for women in the 19th and 20th centuries on the other. These two case studies allow us to experiment with the notions of corpus creation, data validation, data visualization and corpus analysis with different research objectives and methodologies. In both cases and at each step of the research process, a critical perspective is proposed to the students on the research of technical biases they may face.

### *4.1. Covid-19 and Spanish flu: A study of media and political discourses*

During the Covid 19 pandemic, many people, politicians, journalists or commentators have compared this event to the Spanish flu epidemic that took place between 1918 and 1920. Although the political, social and sanitary contexts are quite different, an analysis of the historical newspapers reveals surprising similarities between these two events.

The analysis of digitized historical newspapers we propose highlights these similarities between the two crises in terms of their political management and media treatment. More specifically, the analysis and methodology proposed in the OER allow students to observe, for example, the similarity of the debates on:

- the severity of the virus at the beginning of the epidemic
- the preventive measures recommended to the population to slow down the development of the disease
- the proposed remedies (wearing a mask, medication, and so on).

To highlight these similarities, this case study draws on the different methods studied in the OER. Before looking at media discourse, students must first construct a consistent corpus of data and export it using the NewsEye platform or another digitized historical newspapers digital library. Since one of the objectives of the OER is to demonstrate methods that can be applied in all languages, the work on this case study can be carried out in any European language, as long as the historical newspapers data are available. Secondly, after data collection, it is necessary to measure the quality of raw data and visualize its distribution over time. The objective of these two steps is to make the students aware of the of the heavy research work involved in creating a reliable corpus. Once the corpus is established, students will have the opportunity to apply the different pre-processing steps (tokenization, stemming, POS tagging, etc.) and analysis (N-grams, Topic Modelling) useful to the use case and developed in the OER.

### *4.2. Women in Pants: A study of the emergence of new clothing style for women in 19th et 20th centuries*

Pants have been a symbol of masculinity since the late 18th century and are considered the most significant gender marker in the Western world for the past two centuries. Gender has been a crucial topic in Europe during the 19th and 20th centuries. Although newspapers have contributed to reinforcing gender discrimination and prejudices experienced by women, they have also played a crucial role in breaking down certain barriers. Women used newspapers to gain entry into literary and journalistic circles, as well as to fight for their right to vote and dress as they please. The OER proposes that learners study this theme in digitized newspapers and observe the emergence of a new style of dress for women, its success and the criticism it has generated.

Unlike the previous use case which is limited in time, the study of this subject must be carried out over a long period of time. For France, as an example, we find evidence of debates in historical newspapers between 1850 and 1945. The analysis also implies different methodologies and tools. The development of the wearing of pants for women is indeed linked to other social phenomena, such as the use of the bicycle. It is therefore necessary to compare the data and master the data visualization tools.

More specifically, the students working on this use case will thus be led to study and reflect on the following issues:

- Refining a corpus by finding new keywords, in our case, going beyond the search with only “woman” and “pants”
- Identify the context of use of these words, for example the use of the keyword “bicycle” with the keyword “pants”, which informs us about the context in which women wore pants
- Observe the evolution of the use of these keywords to determine when their use is most intense in the sources.

This use case is also more strongly related to the country and the language of the analyzed newspapers. The issue of the use case is therefore more about identification of associated keywords. The students will thus find a different interest from the previous use case, more focused on document discovery and corpus creation than on information extraction and analysis methods.

## 5. Conclusion and future stories for Europe

This OER is intended to give students means to grasp large collections of historical documents in any language, to discover how to use analysis tools to simplify their tasks, and to find new ways to discover knowledge and develop research. The course is also intended to circumvent the black-box effect that is often inherent to the automation of tasks in large collections, giving students the tools to better understand what happens from input to output.

Historical newspapers are mirrors to the past, describing historical events with a contemporary point of view. However, unlike current news, they are analysed and manually annotated with hindsight, through the validation of historians. In principle, this implies that the quality of this annotated data, provided for the training of artificial intelligence systems, has stronger value than that based on contemporary data.

This also implies that the lessons of the past can be applied the present in a systematic way, through the use of historical big data as a way to learn (also in the sense of machine learning) what to do in the present. A typical application of this is to support policy makers, by helping them for instance benefit from crises of the past in order to manage crises of the present and anticipate those of the future.

In practice, the two use cases described in this OER do echo with contemporary issues. Regular discussions emerge in different groups on how some of their members should or should not dress. The current rhetoric of such public debate always shares a lot with the one encountered with in the “women in pants”. Even more strikingly, as the Covid-19 pandemic recently affected us all, the way the Covid-19 pandemic was handled is incredibly similar to the way the Spanish flu was handled, over a century earlier: initial dismissal of any danger, ridiculing the outlook of face masks deemed useless, arguable miracle cures, promise that the post-pandemic society will be radically better... Several popularising articles were published on this topic, in English,[12] in French,[13] and in German.[14]

These references from the NewsEye project are provided in three languages as an illustration that all the approaches presented in this OER are applicable to any language, and are already implemented in the platform in English, Finnish, French, German, and Swedish. This is a critical element in truly raising awareness on European cultural heritage.

## Acknowledgments

This work was co-funded by the Erasmus+ Program of the European Union, project Digital Methods Platform for Arts and Humanities (2020-1-SE01-KA203-0778789). The European Commission’s support for the production of this publication does not constitute an endorsement of the contents, which reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

[12] https://www.newseye.eu/blog/news/spanish-flu-covid19/?no_cache=1&cHash=41d0c05bd1aab345fc210ab354e725b3.

[13] https://theconversation.com/covid-19-et-grippe-espagnole-quand-la-presse-du-xx-siecle-rappelle-celle-de-2020-137035.

[14] https://www.newseye.eu/blog/news/epidemie-spanische-grippe/?no_cache=1&cHash=696111c9d3bea158e54e3e82ab798bb7.

## References

Boros, E., González-Gallardo, C., Giamphy, E., Hamdi, A., Moreno, J.G., & Doucet, A. (2022). Knowledge-based contexts for historical named entity recognition & linking. In G. Faggioli, N. Ferro, A. Hanbury, & M. Potthast (Eds.), *Proceedings of the Working Notes of CLEF 2022 – Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th-to-8th, 2022*. CEUR-WS.org. pp. 1064-1078. http://ceur-ws.org/Vol-3180/paper-84.pdf.

Boros, E., Hamdi, A., Pontes, E.L., Cabrera-Diego, L.A., Moreno, J.G., Sidere, N., & Doucet, A. (2020). Alleviating digitization errors in named entity recognition for historical documents. In R. Fernández, & T. Linzen (Eds.), *Proceedings of the 24th Conference on Computational Natural Language Learning, Conll 2020, Online, November 19–20, 2020*. Association for Computational Linguistics. pp. 431-441. doi: 10.18653/v1/2020.conll-1.35.

Chiron, G., Doucet, A., Coustaty, M., Visani, M., & Moreux, J.-P. (2017). Impact of ocr errors on the use of digital libraries: Towards a better access to information. *Proceedings of the 17th ACM/IEEE Joint Conference on Digital Libraries*, pp. 249-252.

Doucet, A., Gasteiner, M., Granroth-Wilding, M., Kaiser, M., Kaukonen, M., Labahn, R., Moreux, J., Mühlberger, G., Pfanzelter, E., Therenty, M., Toivonen, H., & Tolonen, M. (2020). Newseye: A digital investigator for historical newspapers. In L. Estill, & J. Guiliano (Eds.), *15th Annual International Conference of the Alliance of Digital Humanities Organizations, DH 2020, Ottawa, Canada, July 20–25, 2020, Conference Abstracts*. https://dh2020.adho.org/wp-content/uploads/2020/07/721%5C_NewsEyeAdigitalinvestigatorforhistoricalnewspapers.html.

Grishman, R., & Sundheim, B. (1996). Message Understanding Conference – 6: A brief history. *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*. https://aclanthology.org/C96-1079.

Hamdi, A., Pontes, E.L., Boros, E., Nguyen, T.T.H., Hackl, G., Moreno, J.G., & Doucet, A. (2021). A multilingual dataset for named entity recognition, entity linking and stance detection in historical newspapers. In F. Diaz, C. Shah, T. Suel, P. Castells, R. Jones, & T. Sakai (Eds.), *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11–15, 2021*. ACM. pp. 2328-2334. doi: 10.1145/3404835.3463255.

Jean-Caurant, A., & Doucet, A. (2020). Accessing and investigating large collections of historical newspapers with the newseye platform. In R. Huang, D. Wu, G. Marchionini, D. He, S.J. Cunningham, & P. Hansen (Eds.), *JCDL '20: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020, Virtual Event, China, August 1–5, 2020*. ACM. pp. 531-532. doi: 10.1145/3383583.3398627.

Leppänen, L., & Toivonen, H. (2021). A baseline document planning method for automated journalism. In S. Dobnik, & L. Øvrelid (Eds.), *Proceedings of the 23rd Nordic Conference on Computational Linguistics, Nodalida 2021, Reykjavik, Iceland (Online), May 31–June 2, 2021*. Linköping University Electronic Press, Sweden. pp. 101-111. https://aclanthology.org/2021.nodalida-main.11/.

Michael, J., Labahn, R., Grüning, T., & Zöllner, J. (2019). Evaluating sequenceto-sequence models for handwritten text recognition. *2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20–25, 2019*, pp. 1286-1293. doi: 10.1109/ICDAR.2019.00208.

Michael, J., Weidemann, M., Laasch, B., & Labahn, R. (2020). ICPR 2020 competition on text block segmentation on a newseye dataset. In A.D. Bimbo, R. Cucchiara, S. Sclaroff, G.M. Farinella, T. Mei, M. Bertini, H.J. Escalante, & R. Vezzani (Eds.), *Pattern Recognition. ICPR International Workshops and Challenges – Virtual Event, January 10–15, 2021, Proceedings, Part VIII*. Springer. pp. 405-418. doi: 10.1007/978-3-030-68793-9_30.

Milligan, I. (2013). Illusionary order: Online databases, optical character recognition, and canadian history, 1997–2010. *The Canadian Historical Review*, *94*, 540-569. doi: 10.3138/chr.694.

Mutuvi, S., Doucet, A., Odeo, M., & Jatowt, A. (2018). Evaluating the impact of OCR errors on topic modeling. In M. Dobreva, A. Hinze, & M. Zumer (Eds.), *Maturity and Innovation in Digital Libraries – 20th International Conference on Asia-Pacific Digital Libraries, ICADL 2018, Hamilton, New Zealand, November 19–22, 2018, Proceedings*. Springer. pp. 3-14. doi: 10.1007/978-3-030-04257-8_1.

Oberbichler, S., Boros, E., Doucet, A., Marjanen, J., Pfanzelter, E., Rautiainen, J., Toivonen, H., & Tolonen, M. (2022). Integrated interdisciplinary workflows for research on historical newspapers: Perspectives from humanities scholars, computer scientists, and librarians. *Journal of the Association for Information Science Technology (JASIST)*, *73*(2), 225-239. doi: 10.1002/asi.24565.

Pivovarova, L., Jean-Caurant, A., Avikainen, J., Al-Najjar, K., Granroth-Wilding, M., Leppänen, L., Zosa, E., & Toivonen, H. (2020). Personal research assistant for online exploration of historical news. In J.M. Jose, E. Yilmaz, J. Magalhães, P. Castells, N. Ferro, M.J. Silva, & F. Martins (Eds.), *Advances in Information Retrieval – 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II*. Springer. pp. 481-485. doi: 10.1007/978-3-030-45442-5_62.

Pontes, E.L., Cabrera-Diego, L.A., Moreno, J.G., Boros, E., Hamdi, A., Doucet, A., Sidere, N., & Coustaty, M. (2022). MELHISSA: A multilingual entity linking architecture for historical press articles. *International Journal Digital Libraries*, *23*(2), 133-160. doi: 10.1007/s00799-021-00319-6.

Pontes, E.L., Hamdi, A., Sidere, N., & Doucet, A. (2019). Impact of OCR quality on named entity linking. In A. Jatowt, A. Maeda, & S.Y. Syn (Eds.), *Digital Libraries at the Crossroads of Digital Information for the Future – 21St International Conference on Asia-Pacific Digital Libraries, ICADL 2019, Kuala Lumpur, Malaysia, November 4–7, 2019, Proceedings*. Springer. pp. 102-115. doi: 10.1007/978-3-030-34058-2_11.

Zosa, E., & Granroth-Wilding, M. (2019). Multilingual dynamic topic model. In R. Mitkov, & G. Angelova (Eds.), *Proceedings of the International Conference on Recent Advances in Natural Language Processing, RANLP 2019, Varna, Bulgaria, September 2–4, 2019*. INCOMA Ltd. pp. 1388-1396. doi: 10.26615/978-954-452-056-4_159.

Zosa, E., Mutuvi, S., Granroth-Wilding, M., & Doucet, A. (2021). Evaluating the robustness of embedding-based topic models to OCR noise. In H. Ke, C.S. Lee, & K. Sugiyama (Eds.), *Towards Open and Trustworthy Digital Societies – 23rd International Conference on Asia-Pacific Digital Libraries, ICADL 2021, Virtual Event, December 1–3, 2021, Proceedings*. Springer. pp. 392-400. doi: 10.1007/978-3-030-91669-5_30.