Project Number: **770299**

NewsEye

NewsEye: A Digital Investigator for Historical Newspapers

Research and Innovation Action
Call H2020-SC6-CULT-COOP-2016-2017

D7.12: Sustainability plan (c) (final)

Due date of deliverable: M45 (31 January 2022)

Actual submission date: 31 January 2022

Start date of project: 1 May 2018 **Duration**: 45 months

Partner organization name in charge of deliverable: BNF

Project co-funded by the European Commission within Horizon 2020		
Dissemination Level		
PU	Public	PU
PP	Restricted to other programme participants (including the Commission Services)	-
RE	Restricted to a group specified by the Consortium (including the Commission Services)	-
CO	Confidential, only for members of the Consortium (including the Commission Services)	-

Revision history

Document administrative information		
Project acronym:	NewsEye	
Project number	770299	
Deliverable number:	D7.12	
Deliverable full title:	Sustainability plan (c) (final)	
Deliverable short title	Sustainability plan (final)	
Document identifier:	NewsEye-T74-D712-SustainabilityPlan-Submitted-v3.0	
Lead partner short name:	BNF	
Report version:	V3.0	
Report preparation date:	31/01/2022	
Dissemination level:	PU	
Nature:	R	
Lead author:	Amanda Smith (BNF)	
Reviewers:	Sally Chambers, Minna Kaukonen (UH-NLF)	
Within input from:	With input from Marion Ansel (BNF), Sébastien Cretin (BNF), Antoine Doucet (ULR), Elisabeth Freyre (BNF), Axel Jean-Caurant (BNF), Max Kaiser (ONB), Minna Kaukonen (UH-NLF), Roger Labahn (UROS), Jean-Philippe Moreux (BNF), Günter Mühlberger (UIBK), Victor Musitelli (BNF), Eva Pfanzelter (UIBK) and Hannu Toivonen (UH).	
Status:	-	Draft
	-	Final
	X	Submitted

The NewsEye Consortium partner responsible for this deliverable has addressed all comments received, making changes as necessary. Changes to this document are detailed in the change log table below.

Change Log

Date	Version	Editor	Summary of changes made
22/12/2021	1.0	Amanda Smith (BNF)	First full draft, delivered to reviewers and project leader.
24/01/2022	2.0	Amanda Smith (BNF)	Final version, taking reviewer comments into accounts.
27/01/2022	2.1	Amanda Smith (BNF)	Clarifications and adjustments made in response to quality manager feedback.
31/01/2022	3.0	Antoine Doucet (ULR)	Minor adjustments and submission.

Executive summary

This report focuses on the sustainability of results after the NewsEye project's completion in January 2022. Following a presentation of the project and an introduction, Section 3 focuses on explaining the work undergone in the context of the report's first version, which was submitted in June 2020. This document set the goals for how sustainability was to be implemented during the project's duration. Thus, the section also defines the key sustainable results (KSRs) of the project, which were elaborated jointly with each work package (WP) leader.

Subsequently, Section 4 focuses on the second version of this report, which was submitted in April 2021. This document focused on how sustainability had been implemented since June 2020 and how sustainability could be ensured during the project's 9-month extension period, until January 2022, and beyond.

As will be detailed in Section 5 of this report, the extension period has allowed to further refine how sustainability can be ensured, especially concerning consortium libraries and in cooperation with READ-COOP.

Table of contents

1 About NewsEye

Newspapers collect information about cultural, political and social events in a more detailed and holistic way than any other public record. Since their beginnings in the 17th century, they have recorded billions of events, stories and names, in almost every language and every country, every day. Newspapers have, thus, always been an important medium for the dissemination of public and political opinions, literary works, essays and art. This wealth of information sets them at the centre stage for anyone interested in European cultural heritage. Since the late-20th century, millions of digitised newspaper pages have been available online and can be accessed individually, but their number makes the amount of information overwhelming and impossible to analyse for readers.

NewsEye: A Digital Investigator for Historical Newspapers, a research project which ran from May 2018 to January 2022, advanced the state of the art and introduced new concepts, methods and tools for digital humanities by providing enhanced access to historical newspapers for a wide range of users. With the tools and methods created by NewsEye, crucial user groups have been able to investigate views and perspectives of historical events and development and, as a consequence, the project has changed the way European digital heritage data is (re)searched, accessed, used and analysed.

The project's partners included the following institutions:

1. The National Library of Austria (ONB)
2. The National Library of Finland (UH-NLF)
3. The National Library of France (BNF)
4. The University of Helsinki (UH)
5. The University of Innsbruck (UIBK)
6. The University of La Rochelle (ULR)
7. The University of Montpellier (UPVM)
8. The University of Rostock (UROS)

2 Introduction

When the first Sustainability Plan deliverable was submitted in June 2020, the COVID-19 pandemic was causing lockdowns and shutdowns across the world. In light of this unprecedented situation, a 9-month extension was granted for the NewsEye project in early 2021. This extension focused on WP7, 'Demonstration, Dissemination, Outreach and Exploitation', which includes four tasks: 'T7.1 - Development of a NewsEye Demonstrator', 'T7.2 - Dissemination, communication and exploitation of project results', 'T7.3 - Project website and online presence' and 'T7.4 - Sustainability'. The National Library of France (BNF) assumed the role of lead beneficiary in charge of WP7 from M37 to M45 (May 2021 to January 2022).

It is within this revised context that this report, 'Sustainability plan (c) (final)' has been prepared. This deliverable is the last of the three deliverables, which have been prepared by BNF in the context of the Sustainability task. This report relates the state of play for the task as of month 45, as well as what is planned for after the project's end. As described in NewsEye's Description of the Action, the Sustainability plan consists of a '...sustainability strategy for the long-term access of tools and data generated by the project, to be planned in full detail at month 26, being implemented at month 36 and fully implemented at M45'. This report will, thus, detail how the sustainability task was managed both during the initial funding period and the extension period.

3 Sustainability Plan (a)

This section of the deliverable focuses on the first deliverable of the sustainability task, specifically with regard to the designation of key sustainable results (KSRs).

3.1 From outputs to key sustainable results

In order to create a coherent Sustainability Plan, the NewsEye project's tangible outputs were examined in detail by the BNF team. Considering a substantial volume of results were created by NewsEye consortium members, it was necessary to create a conclusive list of sustainable outputs, otherwise known as Key Sustainable Results (KSRs) The notion of KSRs was directly inspired by the ONB team's list of Key Exploitable Results (KERs), which were elaborated in the context of the Plan for the Exploitation and Dissemination of Results (PEDR). This continuation of methodological practices aims to create consistency between the deliverables of WP7, especially in terms of the future publicly-available deliverables D7.10 *Dissemination, communication and exploitation of results (d) (final)* (Task 7.2) and D7.9 *Sustainability plan (b) (final)* (Task 7.4). For the purposes of this report, the term **Key Sustainable Result** will refer to an outcome that holds the potential to be utilised after the NewsEye project's conclusion at the end of January 2022.

Although the NewsEye project members have generated a sizable number of outputs, not all of them will necessarily be sustained; most of these can be consulted via Zenodo and OpenAIRE NewsEye repositories. The first step in identifying NewsEye's KSRs involved the creation of a table containing the result description, output type, stakeholders likely to be interested and the target users; these KSRs were subsequently validated by the WP leaders. Their identification was crucial for facilitating wider reflection regarding the sustainability of NewsEye results. Likewise, a consideration of target users, the individuals or organisations which will be employing the results was especially important. Just as the NewsEye results have been created through the cooperation between the consortium researchers, the successful sustainability of the results ultimately will rely on concerted efforts to maintain them after the end of the project. This section of the report focuses on each KSR.

3.2 Work package 1: Data management

Work Package (WP) 1, which was led by the University of Innsbruck, aimed to 'provide a single access point for all tasks connected with the management of research data within the project'. The automated tools created and developed within this WP improved the accessibility of the data present within digitised newspaper pages and served as the foundation for subsequent WPs.

3.2.1 Tools to create ground truth (Transkribus)

The concept of *ground truth* played an important part in the NewsEye project as a launchpad for the development of the artificial intelligence (AI) systems present in most NewsEye tools. In this context, it can perhaps be best described as both 'data that is used to build AI systems', as well as the 'truth that underpins the knowledge in an AI system.'[1] The text recognition tool Transkribus, which was created in the context of the READ (Recognition and Enrichment of Archival Documents) project and is hosted by the Digitisation and Digital Preservation group at the University of Innsbruck, has played a pivotal part in the creation of ground truth for NewsEye. Since the NewsEye project has signed on to the Open Data Research Pilot (ODRP), its results are made available via open science tools such as Zenodo and GitHub. After the project's conclusion, this ground truth will be sustained through the Transkribus platform and the Transkribus GitHub page. These open-source sources should be of special interest to humanities researchers and information technology specialists working in the context of digital libraries, who will be able to use the Transkribus tool for their own projects.

[1] Anderson, J. L., & Coveyduc, J. L. (2020). *Artificial intelligence for business: a roadmap for getting started with AI*. Wiley.

3.2.2 Text ground truth and article separation in four different languages

The text ground truth that has been created within NewsEye is available in four languages: German, Finnish, Swedish and French. This is also the case for the article separation ground truth which was created during the first two years of the project. The multilingualism present within these results reflects NewsEye's aim to 'produce tools that are operational on any newspaper collection written in any European language'. Article separation, which recognises the shape of different articles in a digitised newspaper page, greatly facilitates user queries.

3.2.3 Named entities and stance datasets in four different languages

Often referred to in the context of the '5 Ws of journalism' ('who', 'what', 'when', 'where' and 'why'), the term 'named entity' generally refers to specific proper nouns. The designation of these words through the creation of datasets greatly facilitates user queries, making it easier to search for specific terms; these datasets also pave the way for future research in historical newspapers. Likewise, the identification of named entity stances, whether positive, negative or neutral, contextualises the text found therein. The named entities and stance datasets produced by WP1 improved the data clarity of the digital newspaper pages included in the project and laid the groundwork for the stance recognition and named entity recognition tools created by WP3.

3.3 Work package 2: Text recognition and article separation

The objective of WP2, which was led by the University of Rostock, was to 'investigate, develop and implement methods, algorithms and tools for text recognition and article separation' (per the project's Description of Action). The tools created within this group, layout analysis, automatic text recognition and article separation, focused on making digitised newspaper text more easily searchable by enhancing digitised pages.

3.3.1 Layout analysis

The purpose of layout analysis is to detect baselines within digitised newspaper pages. In other words, the tool automatically recognises which parts of the images contain various types of content, including articles and photographs, for example. Layout analysis was especially important to the project because it laid the groundwork for further tools that enable enhanced searches of text, namely automatic text recognition. The layout analysis tool created by this WP improved the recognition quality of Optical Character Recognition (OCR) software, as it identified the layouts of historical newspaper pages.

3.3.2 Automatic text recognition and integration within Transkribus

Automatic text recognition tools extract lines of text from digitised newspaper pages. The automatic text recognition available specifically within Transkribus holds the potential to appeal to more stakeholders and users. The NewsEye text recognition engines are some of the most accurate in the world, running at a below one percent error rate. Before being integrated into an end-user application, this raw data should appeal to researchers both in the fields of Computer Science and Digital Humanities. Integration into a fully-functioning platform should result in increased user friendliness, making it more easily manipulated by Transkribus users and humanities researchers.

3.3.3 Article separation

Like today's newspapers, the pages of the written press in the nineteenth and twentieth centuries were typically divided into articles. By building upon the layout analysis tool, article separation tools recognise blocks in the text that belong to the same articles, making searching easier to carry out. Although article separation is not an entirely novel concept, this WP's innovative usage of multi-lingual methods contributed to the field of Computer Science research. NewsEye article separation is the first standard tool of its kind, which should enable easier and more intuitive access to historical

newspapers. The current iteration of this tool is available via the GitHub page of the University of Rostock's Computational Intelligence Technology Lab.

3.4 Work package 3: Semantic text enrichment

Building upon the work accomplished in WPs 1 and 2, WP3, which was led by the University of La Rochelle, focused on the semantic enrichment of individual documents and their contents. The tools created by the WP3 team facilitated research as they allowed users to sort through hundreds of thousands of digitised newspaper pages with greater ease.

3.4.1 Named entity recognition

Named entity recognition tools determine which words or phrases in the text refer to real-life entities. The NewsEye project focused on three types: names of people, locations and organisations. Currently, the tools can be freely accessed via the NewsEye GitHub page.

3.4.2 Named entity linking

Named entity linking, which is done automatically, takes named entity recognition a step further by connecting named entities with their respective Wikidata entries. Wikidata is a database within which a unique identifier is assigned for individual items. These identifiers transcend language barriers, allowing users to analyse data even if they do not speak a particular language. Even named entities which are spelled identically can be properly identified; for example, London, England (Q84) can be distinguished from London, Canada (Q92561). The named entity linking created by this WP has been crafted with historical newspapers in mind. The increased research clarity provided by these tools, which can be accessed via GitHub, should especially appeal to Computer Science researchers.

3.4.3 Stance detection

Stance detection refers to the recognition of the sentiment (positive, negative or neutral) applied to named entities referred to in the text. In theory, this information could result in further in-depth analysis of trends and evolving opinions over time. In practice, however, stance is often entirely subjective, as annotators may not interpret it in the same way. Regardless, the stance detection data and tools generated by WP3 were especially valuable to the project due to their innovativeness. In addition to integration into existing library systems, this technology could be especially interesting for computer scientists who wish to improve upon existing methods. Like other NewsEye tools, it is openly available via GitHub.

3.4.4 Event detection

Event detection tools focus on newspaper coverage of events. For example, they aim to recognise whether an event is described as breaking news, as well as how different pieces of information spread within a national or international context. This capability could especially be of use for Digital Humanities scholars who seek to analyse trends over time. Like the other tools within WP3, these are also available via GitHub.

3.4.5 Multi- and cross-lingual text enrichment in historical newspapers

Combining the tools from WPs 2 and 3, an important KSR is the ability is a complete high-performing workflow to process historical newspapers starting from digitised images and resulting in separate articles containing text in which named entities are recognised and linked, opinion relative to entity mentions is calculated and novel events are identified. An impactful aspect of this workflow, most extensively tested in Finnish, French, German and Swedish, is that it functions in any language, and further allows for the cross-lingual analysis of articles since everything that stems from WP3 is

represented in a language-independent fashion: whatever the writing language, named entities are linked to language-independent Wikidata identifiers, stance is represented by a real number and events are composed of markers (concepts and entities) also linked to Wikidata. This workflow should especially be of interest to national libraries and Digital Humanities scholars.

3.5 Work package 4: Dynamic text analysis

WP4, which was led by the University of Helsinki, focused on developing and implementing methods for contextualised and contrastive content analysis that is carried out dynamically per request. These methods, which include topic modelling and document linking, facilitated the investigation of document content and collections in a research context. It should be noted that UH-CS plans to extend the development of these methods in the context of future projects.

3.5.1 Topic modelling

Topic modelling uses machine learning to automatically analyse and organise document collections by topic. This analysis focuses on the statistical properties of document contents; in the context of the NewsEye project, these properties are the words present within digitised newspaper pages. Since it does not require much human oversight, topic modelling is especially helpful in the exploration of expansive collections. The NewsEye topic modelling source code is currently available via GitHub, although additional funding and computational resources from interested parties will be needed in order to keep the technology updated and accessible to potential users.

3.5.2 Document linking

The topic-based document linking present within the NewsEye project is dependent upon topic modelling. It works by linking documents with related topics, enabling and permitting users to find other similar documents or document sets. In terms of sustainability, document linking faces the same constraints as topic modelling.

3.6 Work package 5: Personal Research Assistant

The objective of WP5, which was also led by the University of Helsinki, was to develop methods and tools for automated, iterative analysis of corpus content and reporting of the results, culminating in the Personal Research Assistant that functions as the user's intelligent and transparent aid. There are three components to the Personal Research Assistant (PRA): the investigator, the explorer and the explainer. Building upon the dynamic analysis methods developed in WP4, these tools have been created in the form of web services. They have been designed with modularity in mind, meaning that an integration of the tools into other environments could be possible without excessive effort.

3.6.1 Personal research assistant investigator

The investigator constitutes the 'core component' of the PRA. It generates autonomous statistical analyses on data corpora with the purpose of finding interesting phenomena for users. Although the PRA investigator's interface was not finalised, and it could eventually be able to spontaneously adapt the way it explores data, instead of fully relying on pre-existing patterns. Although they are currently used for historical research, the tools that make up the PRA investigator could be implemented in the context of other intelligent personal assistants; in any case, this tool is openly accessible on the NewsEye GitHub page.

3.6.2 Personal research assistant reporter

The PRA reporter expresses the results produced by the investigator in the form of natural language textual reports. These reports focus on the computational path taken between the input of

PRA users and the generation of results by the investigator; furthermore, they also include a summarisation of analysed textual contents. The code for this tool can be found on GitHub.

3.6.3 Personal research assistant explainer

Going another step further, the PRA explainer describes to users how the investigator discovered the results reported by the explorer. This explanation is written in a transparent manner, which facilitates a greater understanding of research results; this KSR could especially be especially helpful for improving the accessibility of the PRA for users who are less familiar with its more technical components. As it is easily accessible online, future users could build upon this research to create other platforms.

3.7 Work package 6: Digital Humanities applications and uses

WP6, which was led by the University of Innsbruck, focused on the interaction and dialogue between humanities scholarship, technology and cultural heritage institutions. The results created by this WP truly reflected the interdisciplinarity at the heart of the NewsEye project. Deviating from the other WPs, WP6's KSRs took the form of deliverable reports and multimedia materials which will be sustained through digital and material means.

3.7.1 Showcase case studies for the user interface

Deliverable D6.12, which was submitted at the end of April 2021, contains case studies which highlight the capabilities of the user interface and project website. It focuses on the following kinds of objects: screencasts, Jupyter Notebook documents, Twitter threads, case study descriptions and best practice examples. For the purposes of sustainability, it should be noted that this deliverable will be publicly consultable via the NewsEye CORDIS page. This will provide potential users and stakeholders with more practical ways to engage with other NewsEye project results by facilitating the accessibility of the outreach materials which will be discussed later on in this section.

3.7.2 Educational material for teachers, pupils and lay historians

D6.12 also provided an initial outline for possible usage of NewsEye material in the context of education by focusing on three user groups: pupils and teachers, university students and lay historians. An examination of national curricula within the NewsEye project consortium has shown that this material could be integrated into school lessons in at least Austria, France and Finland. A course taught at the University of Innsbruck by WP leader Eva Pfanzelter and at the University of Helsinki, Finland by Jani Marjanen, 'Sources and Auxiliary Sciences: Digital Sources and Analysis', has been developed, in which students are invited to use the NewsEye platform to improve their knowledge of digital sources. The plans for the sustainability of this educational material will be discussed in Section 4 of this report.

3.7.3 Contextualised case studies for academic use

In total, four reports focusing on contextualised case studies for academic use were submitted during the course of the project, during month 6 (D6.1), month 12 (D6.2), month 24 (D6.7) and month 36 (D6.10). The most recent version includes a list of journal papers and conference papers which have been published/presented or are expected to be in the near future. This collection of sources should provide scholars with an array of examples of how to use NewsEye within a research context.

3.7.4 Digital Humanities outreach materials

The final KSR of this work package includes the materials which have been or will be generated for the deliverables. These include, for example, but may not be limited to, publications, blog posts, podcasts and Jupyter Notebooks. Although they do not all share the same stakeholders and target

audiences, they all are aimed at circulating NewsEye results outside of the consortium. Furthermore, educational materials which were developed by the Digital Humanities team have been made available for educators to download from the project website.

3.8 Work Package 7: Demonstration, Dissemination, Outreach and Exploitation

Generally speaking, WP7, which was led by the Austrian National Library during the initial funding period, focused on demonstrating the tools created by other work packages in a library environment, as well as spreading research results and sharing them with the different user groups. For the purpose of sustainability planning, this report will focus on the WP's objective to develop and implement the NewsEye Demonstrator as a user interface to the enrichment and analysis tools developed in WP3 and WP4, and for the Personal Research Assistant (WP5).

3.8.1 NewsEye demonstrator

The NewsEye demonstrator is the interface and platform through which users interact with the digitised collections by benefiting from all the tools of the NewsEye tool chain. As of this writing, two versions of the demonstrator are publicly accessible. The demonstrator will be sustained on the servers of La Rochelle University and its components may likely be used for future research projects.

4 Sustainability Plan (b)

This section of the deliverable focuses on the lessons learned from the second deliverable of the Sustainability task.

4.1 Interviews with professionals

The content of the second version of the Sustainability plan was upgraded by a series of interviews carried out between BNF and eight professionals from external organisations:

- Antoine Isaac, Research and Development Manager (Europeana)
- Yves Maurer, Technical Lead (The National Library of Luxembourg)
- Clemens Neudecker, Research Advisor to the Directorate General (Berlin State Library)
- Tomasz Parkoła, Head of Department (Poznań Supercomputing and Networking Center)
- Laurent Romary, Senior Researcher (The French Institute for Research in Computer Science and Automation)
- James Smithies, Director (King's Digital Lab, King's College of London)
- Deborah Thomas, Chief of the Serial and Government Publications Division (The Library of Congress)
- Nathan Yarasavage, Digital Projects Specialist (The Library of Congress)

An important concern raised during these discussions was the importance of being able to explain the NewsEye results in a clear and understandable way to those external to the project. Although the list of key sustainable results (KSRs) developed in the context of the first Sustainability plan was sent to interviewees prior to each meeting, discussing the results in detail remained challenging. As explained later in this report, the need for standardised documentation became apparent, which led to the preparation of technical specification documents.

Another point raised several times during these discussions was the importance of choosing which results to sustain after the end of the project. As a result of this consideration, BNF delegated the choice of which project results to prioritise for sustainability to Work Package (WP) leaders, who are particularly knowledgeable about the outcomes of their respective WPs.

4.2 Legal considerations

A meeting was held in February 2021 with the BNF's legal service with the purpose of clarifying the terms of results ownership laid out in the project's Consortium Agreement and Grant Agreement. The Consortium Agreement specifies that 'Results are owned by the Party that generates them' and, more specifically, that 'Each new software is the property of the Party that generates it, as regards to scientific human, material and financial inputs'.

Although sustainability is not explicitly mentioned in the Grant Agreement, it should be noted that each beneficiary of results holds certain responsibilities for a period of four years after the project's end. In particular, Article 28 states that 'Each beneficiary must — up to four years after the period set out in Article 3 [the duration of the action] — take measures aiming to ensure 'exploitation' of its results (either directly or indirectly, in particular through transfer or licensing'. As this 'exploitation' will occur after the project's end, this obligation ultimately falls under the category of sustainability, meaning that each project partner is currently making efforts to fulfil their contractual obligations.

4.3 Sustainability Working Group (SWG)

Following a decision made during the NewsEye Consortium's biennial steering committee in May 2020, a Sustainability Working Group for national libraries was organised. The ultimate goal of this working group was to decide which project results can be integrated into existing platforms of the three national libraries in the consortium (BNF, ONB and UH-NLF). Colleagues from the Royal Library of Belgium (KBR) were also invited to join in order to offer a perspective as a potential future user of NewsEye services.

The first meeting of this working group, which took place in July 2020, focused on the possible usage of KSRs in the context of national libraries. From this meeting, it quite quickly became clear that the list of KSRs was not suited for external presentation or consumption in part because their organisation by WP only makes sense to those very familiar with the project. After the second meeting of the working group in September 2020, colleagues from BNF, KBR and UH-NLF were able to instead offer a tentative list of the KSRs which could potentially be of interest to their respective institutions.

During a bilateral meeting between colleagues from BNF and KBR in November 2020, it became even more apparent that it would be crucial to create documentation outlining the project results which could be exploited and/or sustained externally. These technical specification documents, which will be explained in the next section, will allow the members of the Sustainability Working Group (SWG) to present NewsEye project results to their institutions. This facilitated further negotiations, especially during late Spring and Summer of 2021 when the libraries were defining their budgets for 2022. The next SWG meeting took place during the first week of May 2021, which resulted in a clearer vision of how negotiations would proceed during the coming months. A full synopsis of the group's meetings and decisions is included in Section 4.

4.4 Technical specification documents

The idea of preparing technical specification documents arose from the need for BNF to be able to present NewsEye results in a transparent manner to potential organisations which could play a part in sustaining the results. After the BNF team developed a tentative template for this documentation, a prototype was completed by ULR for WP3 in February 2021.

Thereafter, each WP leader was contacted and asked to decide which results would be included and to fill in the documents for their corresponding WP. These documents are presented in Appendix A of this report.

The technical specification documents played a role in exploitation and sustainability negotiations going forward, particularly when it came to preparing meetings with external organisations. The inclusion of crucial information such as formats and publicly available deliverables provided a basic

level of understanding of results which could be built upon in the context of bilateral or multilateral exchanges.

4.5 READ-COOP working group

As mentioned in Section 3, the European cooperative society READ-COOP proposed to exploit and sustain some NewsEye results during and after the project's extension. Following a decision made during the NewsEye Consortium's biennial steering committee meeting in November 2020, a first meeting was organised in December 2020 in order to examine this possibility in detail with representatives from ULR, ONB, UROS, UH-CS (University of Helsinki-Department of Computer Science), BNF and UIBK. This initial discussion laid the groundwork for the creation of a READ-COOP working group chaired by BNF, which held its first meeting in February 2021.

In preparation for the aforementioned meeting, Günter Mühlberger, who is Chair of the READ-COOP Board of Directors, prepared a short paper detailing which roles the cooperative society could play in the larger scope of the exploitation and sustainability of NewsEye results. This internal document, which was presented to the READ-COOP working group in February 2021, has since become a living document that has been reviewed and revised by several members of the consortium.

BNF arranged a meeting between ULR and READ-COOP in March 2021, which focused on the first steps for integrating Named Entity Recognition (NER) tools into Transkribus, as well as the possibility of sustaining and enhancing the NewsEye Demonstrator platform through READ-COOP. Other tools which may be sustained through READ-COOP include Article Separation (AS) and Layout Analysis (LA). Various bilateral and multilateral negotiations regarding which results would be sustained and in which manner continued to take place, the details of which follow in Section 5.

5 Sustainability during the extension period and beyond

During the project's extension period by month 45, BNF assumed the role of WP7 leader; these additional nine months permitted a greater understanding of how NewsEye results could be sustained. As such, efforts regarding the planning and implementation of sustainability after the end of the extension period can be divided into two parts:

- Defining a more highly involved role of the consortium libraries
- Defining a plan for how certain results could be sustained through cooperation with READ-COOP.

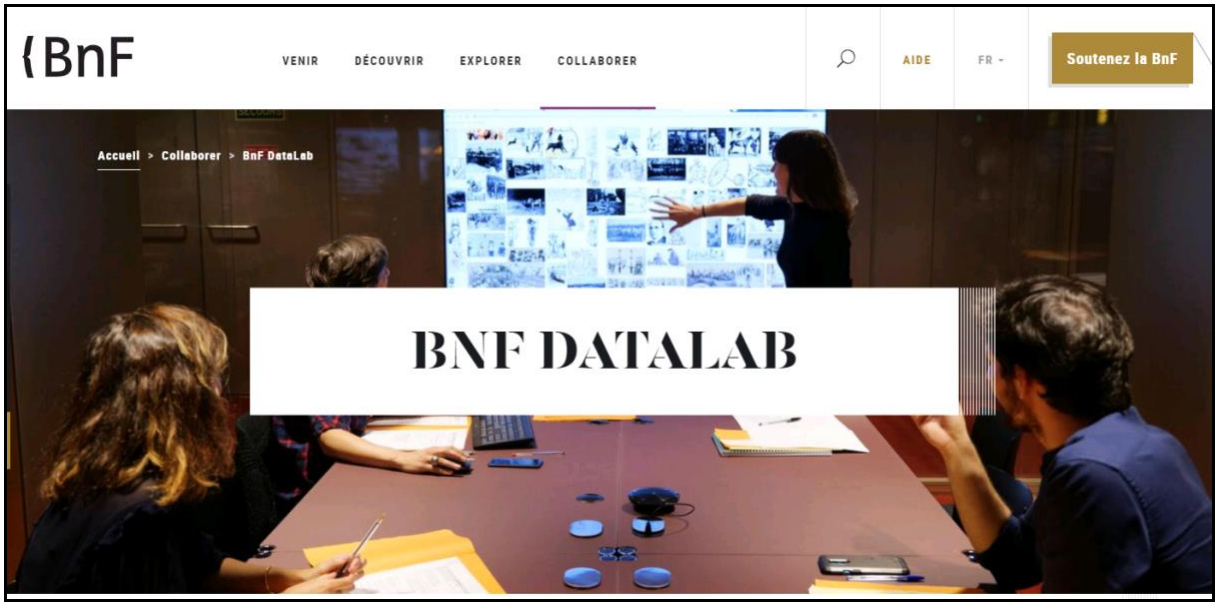
5.1 Role of consortium libraries

Following up on the work of the Sustainability Working Group accomplished during the initial funding period, Sustainability task leader BNF worked on confirming the roles of the three consortium libraries in the sustainability of project results. Considering the fact that NewsEye was a project focused on research and innovation, the idea to integrate the NewsEye Demonstrator into experimental data lab environments was explored in detail.

1. The National Library of France (BNF)

At an institutional level, negotiations with BNF, ULR and the BnF DataLab were initiated in 2020 with the aim of defining how NewsEye results could be exploited or sustained through this channel. Opened in October 2021, the DataLab is a service for researchers who wish to work on the digital collections of BNF. It is currently composed of a physical space on the François Mitterrand site in the 13th *arrondissement* of Paris where researchers can work and be advised by specialists, and an online toolbox for researchers will be made accessible during the coming months. It is in this context that the second version of the NewsEye Demonstrator will be shared with researchers working with

digitised newspapers. Elaborated during the project extension, this version of the demonstrator has been streamlined and omits the experimental elements of the PRA which were less viable. A document in French (included in Appendix B and written by Axel Jean-Caurant of ULR) details the components and usages for the demonstrator and will be translated into English in order to be reused in the contexts of other European institutions. Integration is expected to be carried out in the BnF DataLab environment during the coming months.



The BnF DataLab is expected to attract researchers from both France and abroad, and this additional dissemination channel will contribute to the sustainability of the NewsEye demonstrator.

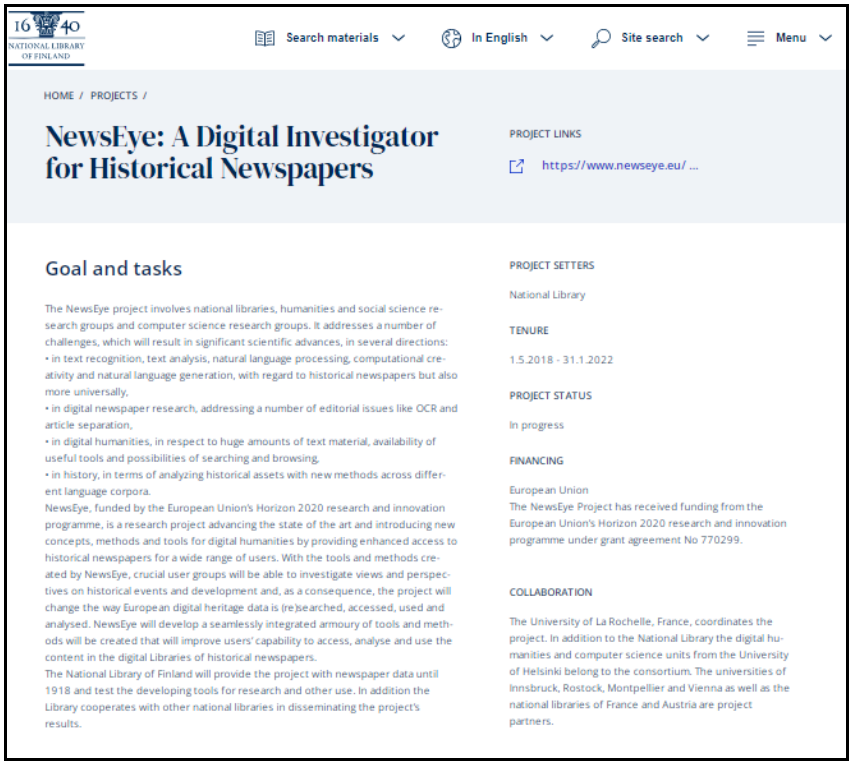
2. The National Library of Austria (ONB)

A series of negotiations led by BNF with ONB and ULR regarding the integration of the NewsEye Demonstrator into the ONB Labs environment was carried out throughout the extension period. At the time of writing, the creation of a dedicated instance of the NewsEye Demonstrator is planned for the coming months, and tests are being currently carried out by the labs team in order to confirm the feasibility of such an implementation. In this case, the demonstrator will be presented as an experimental platform which users will be able to access through ONB. This work coincides with the recent redevelopment of the ANNO (Austrian National Library digitised historical newspapers collection) website.

The current version of ANNO (January 2022)

3. The National Library of Finland (UH-NLF)

Although NLF does not currently have an experimental data lab space like its fellow consortium member libraries, the institution is very keen to remain involved in sustaining the project's results. It is expected that the demonstrator may be used in the context of future research projects. Also, most of the digitised newspaper pages which were provided for the project will remain in the demonstrator, although this will be limited to newspapers published until the end of the 1910s due to copyright restrictions. The library has also shared information about the project on a dedicated page of NLF website.



This screenshot from January 2022 shows how UH-NLF presents the project within the context of the institution's research activities.

5.2 Role of READ-COOP

As mentioned earlier in this report, READ-COOP will play an important role in the sustainability of the NewsEye project results. READ-COOP SCE (European Cooperative Society) with limited liability was established on 1 July 2019 to sustain and further develop the Transkribus platform. Transkribus was developed within the Horizon 2020-funded READ (Recognition and Enrichment of Archival Documents) project by a consortium of leading research groups from all over Europe, headed by the University of Innsbruck. At the time of writing, there are 102 members from 26 countries in the cooperative, including the following NewsEye project partners: the University of Vienna, the University of Rostock and the University of Helsinki.

Following the negotiations between the company and ULR, a dedicated API is currently being elaborated with the goal of integrating Named Entity Recognition (NER) data into Transkribus, which will both make the NER tool accessible to a wide audience and ensure that it remains accessible. READ-COOP is also working closely with UROS in improving Article Separation methods.

6 Conclusions

The sustainability task has formed an integral part of the NewsEye project, and the extension period has permitted more time for exploring how the NewsEye results can be sustained in a more concrete manner. It should be noted that the NewsEye GitHub and Zenodo pages, as well as both versions of the demonstrator, will be sustained by La Rochelle University for the foreseeable future of at least four years.

In addition, many of the results created during the project's duration may be reused in the context of future projects regarding digitised historical or contemporary newspapers. It is clear that facilitating access to these resources will remain an objective for cultural heritage institutions, both in Europe and elsewhere, for the foreseeable future. Thus, NewsEye results will be able to contribute to the digital transformation of European culture heritage and Digital Humanities research in 2022 and beyond.

Appendix A: Technical specification documents

Transkribus tools for creating and exporting ground truth data and for training and publishing recognition models

Description:

The Transkribus platform has been set up as part of the Horizon 2020 Project READ which was coordinated by the University of Innsbruck (DEA) team. Although originally developed for handwritten text recognition (HTR), it turned out that the tools and methods are also suitable to produce excellent results for printed material. The focus of READ was to develop a rich expert-client (JAVA-SWT) which provides all the features which are necessary in order:

- To upload documents to the platform.
- To create training data.
- To train neural networks (ATR/HTR engines) and produce models on the basis of the training data.
- To measure results in an objective and standardized way.
- To export data in various formats.
- To make all data and services also available via the Transkribus API.

As part of the NewsEye project, the Transkribus platform was enhanced and augmented with new features and tools for processing historical newspapers mainly towards the inclusion of specific ground truth (GT) tools capable to create training data for all tasks in the project like layout analysis (LA), automatic text recognition (ATR), article separation (AS) and named entity recognition (NER), named entity linking (NEL) and stance detection. All the training data can be exported, e.g. the named entity training data in IOB (inside–outside–beginning) format, article information is either included directly in the PAGE XML or is contained in METS/ALTO or METS/PAGE.

In addition to the possibility of creating GT data, the trained models of the project for LA and ATR (for all four project languages: French, German, Finnish and Swedish) are publicly available in the Transkribus platform either as they were trained originally or as they were used in other large models like ‘Transkribus print 0.3’, which is a general public model for ‘German printed newspaper pages’. With this approach, all NewsEye models can therefore also serve as base models for further models.

The training data itself from ATR and AS is deposited in Zenodo.

Formats:

- PAGE XML
- IOB
- Trained models (internal Transkribus format)
- Training data sets on Zenodo

Requirements for sustainability:

As long as Transkribus has a solid financial basis, the tools will be available.
The data will still be accessible indefinitely online via GitHub and Zenodo.

Associated partner:

University of Innsbruck, Digitisation and Digital Preservation (UIBK-DEA)

Associated Work Package

WP1: Data generation

Article separation toolbox

Description:

This toolbox aims to separate newspaper articles. It relies on recognized text baselines and automatically recognised text. The tools are comprised of:

- The clustering of baselines to text blocks;
- The detection of headings;
- The detection of separators and their use for correcting the baseline clustering;
- The text block similarity computation based on word embeddings;
- The text block similarity computation based on BERT language models;
- The Graph Neural Network based estimation of confidences for merging text blocks in articles;
- The clustering of text blocks to articles based on three algorithms.

Most of the above-mentioned components are Machine Learning-based and the according tools are contained for:

- Training;
- Validation;
- Model and hyperparameter choice;
- Testing and visualisation;
- Application.

For details of the overall workflow, see the Article Separation Workflow diagram in the public project deliverable D2.7.

Format(s):

Publicly released source code

Requirements for sustainability:

Open-source code released on GitHub and permanently available:

- Python tools to separate articles, mainly for historical newspaper pages.
- Python tool to measure the performance of an Article Separation algorithm.
- Python utility tools: PAGE-XML, read&write, parse, plot …

Associated partner:

University of Rostock (UROS)

Associated Work Package:

WP2: Text recognition and article separation

Named entity recognition and linking tools

Description:

Named entities are among the most relevant information that can help to thoroughly index digital documents and easily retrieve them. However, most digitised documents are indexed through a noisy version produced by an optical character recognition (OCR) system. The noisy version contains numerous OCR errors that change the content of these documents and naturally make their access more difficult in digital libraries. Unlike contemporary data that have a large number of NER and NEL resources and tools, historical documents face the problem of lacking annotated resources. Contemporary resources are not suitable to build accurate tools over historical data because of variations in orthographic and grammatical rules, not to mention the fact that the names of persons, organisations and places are significantly changing over time.

Named entity recognition (NER) is a natural language processing (NLP) task that aims to locate important names and proper names in a given text and to categorise them into a set of predefined classes. Typical NER tag sets define three classes for named entity labelling: persons, locations and organisations. In the NewsEye project, in addition to these classes, NER targets a class including human products (like news articles) and specifies a subtype for the class person when it corresponds to the author of an article. In the context of newspapers, it is indeed very useful to be able to differentiate the person(s) mentioned in an article from the person(s) who wrote and signed the article.

Named Entity Linking (NEL) is the task of recognising and disambiguating named entities to a Knowledge Base (KB). NEL is a challenging task because named entities may have multiple surface forms, such as its full name, partial names, aliases, abbreviations, and alternate spellings. Given a KB containing a set of named entities and a set of documents, the goal of named entity linking is to map each named entity in these documents to its corresponding named entity in a KB, e.g., Wikidata. Wikidata is a free and open KB that can be read and edited by both humans and machines. This KB acts as central storage for the structured data of its Wikimedia sister projects including Wikipedia, Wikivoyage, Wiktionary, Wikisource, and others. In a nutshell, NEL aims to retrieve the Ground Truth entities in a KB referred to in a document by locating mentions, and for each mention accurately disambiguating the referent entity.

Format:

The format used to train new models and the output format of running existing models is identical. We use the IOB (inside, outside, beginning) format which is a tab separated format where the first column (A) represents the tokens of the text, the second (B) is the NER annotation (with **B**egin, **I**nside and **E**nd tags for multi words entities) and the third (c) is the NEL annotation (with a Wikidata ID or NIL if the link could not be found).

	A	B	C
1	stället	O	_
2	för	O	_
3	hofrättsassessorn	O	_
4	C	B-PER	NIL
5	.	I-PER	NIL
6	W	I-PER	NIL
7	.	I-PER	NIL
8	Sulin	E-PER	NIL
9	,	O	_
10	som	O	_
11			
12			
13	Hugo	B-PER	Q14271401
14	Neuman	E-PER	Q14271401
15	.	O	_
16			
17	Åbo	O	_
18	,	O	_
19	Slottsgatan	O	_

Requirements for sustainability:

NER: https://github.com/NewsEye/Named-Entity-Recognition
NEL: https://github.com/NewsEye/Named-Entity-Linking/

Associated partner:

University of La Rochelle (ULR)

Associated Work Package

WP3: Semantic text enrichment

Implementations of topic modelling methods

Description:

This work concerns the analysis of potentially large and noisy document sets by statistical methods, as well as various ways of using such analyses. The main focus has been on *topic modelling*: an approach to unsupervised analysis of document sets on the basis of the words the documents use. It has included the development of a number of novel topic modelling methods, extensions of existing methods and new implementations of existing methods where suitable implementations did not previously exist.

Format:

Publicly released source code

Requirements for sustainability:

This open-source code has been released on GitHub and remains permanently available.

- Multilingual Dynamic Topic Model
- Polylingual Topic Model Python implementation
- Scripts for training topic models with (open-source) Gensim toolkit
- Scripts for training Dynamic Topic Models with Gensim for Disappearing Discourses research

Associated partner:

University of Helsinki (UH-CS)

Associated Work Package:

WP4: Dynamic text analysis

Implementation of methods to compare topics between contexts

Description:

This work addresses the comparison of contrasting contexts, such as time periods, countries or publications. One approach we have adopted is to carry out analyses of the documents associated with the given contexts using topic models and compare the sets in terms of the detected topics. Code for a range of methods for performing such analyses is released publicly on GitHub for future use and to serve as the basis for development of similar methods in future.

Format:

Publicly released source code

Requirements for sustainability:

The open-source code has been released via GitHub and will be permanently available.

Associated partner:

University of Helsinki (UH-CS)

Associated Work Package:

WP4: Dynamic text analysis

Cross-lingual news linking

Description:

Part of this work involved development of methods to find related news articles in a large corpus, potentially spanning multiple languages – *cross-lingual news linking*. Several methods were compared experimentally in the publication A Comparison of Unsupervised Methods for Ad-hoc Cross-Lingual Document Retrieval, and the code that implements all of the methods concerned and executes the experimental comparison was released publicly.

Format(s):

Publicly released source code

Requirements for sustainability:

The open-source code has been released via GitHub and will be permanently available.

Associated partner:

University of Helsinki (UH-CS)

Associated Work Package:

WP4 Dynamic text analysis

Implementation of training and visualisation of diachronic embeddings

Description:

As part of this work and in collaboration with Digital Humanities researchers, we implemented methods to train *diachronic word embeddings* – statistical representations of words based on their usage observed in a corpus, explicitly modelling changes over time. We have released code for training such representations, analysing them using clustering and visualising them in order to support further analysis and research.

This code was released to support the following publications:

- Word Clustering for Historical Newspapers Analysis
- Clustering ideological terms in historical newspaper data with diachronic word embeddings

Format:

Publicly released source code

Requirements for sustainability:

The open-source code has been released via GitHub and will be permanently available.

Associated partner:

University of Helsinki

Associated Work Package:

WP4: Dynamic text analysis

Personal Research Assistant (PRA)

Description:

Historical newspapers collect information about cultural, political and social events in a more detailed way than any other public record. At the same time, analysing the wealth of information in the newspaper archives has traditionally been difficult and time-consuming. The NewsEye project develops methods and tools for effective exploration and exploitation of historical newspaper archives.

The core concept of NewsEye is a set of tools and methods, from text recognition to automated exploration of texts, that improve the users' capability to access, analyse and use the content of historical newspapers, stored in digital libraries.

The Personal Research Assistant carries out automated, iterative analysis of corpus content and reports on the results, functioning as the user's intelligent and transparent aid. The Personal Research Assistant consists of three primary components (Investigator, Reporter and Explainer) and a Controller component. The Investigator component conducts experiments on datasets, attempting to identify interesting phenomena from the historical newspaper corpora. The Reporter component then proceeds to describe the most salient of the phenomena identified by the Investigator in natural language, highlighting to the user the aspects believed to be most interesting. Finally, the Explainer component provides brief descriptions of what steps were taken by the Investigator during an Experiment, and why, so that a user might inspect, replicate and modify the types of experiments conducted by the Investigator. The Controller component ensures interfaces between the Personal Research Assistant and other parts of the NewsEye Platform, namely the Demonstrator, APIs developed in WP4 and the database that contains pre-processed project data. The Controller also provides interfaces between the components of the Personal Research Assistant itself: Investigator, Reporter and Explainer. Thus these parts are independent from each other and could be potentially reused without the other parts.

Format(s):

The three main components are structured as, and can be deployed as, web applications that provide and communicate through JSON APIs that are documented in Deliverables D5.6, D5.7 and D5.8.

The software is available as standard source code repositories, mainly in the Python 3 programming language.

Requirements for sustainability:

The source code repositories for the Personal Research Assistant components are available at the following URLs:

- Investigator and Controller: https://github.com/NewsEye/Newseye-WP5-Investigator
- Reporter: https://github.com/NewsEye/Newseye-WP5-Reporter
- Explainer: https://github.com/NewsEye/Newseye-WP5-Explainer

Associated partner:

University of Helsinki (UH-CS)

Associated Work Package:

WP5: Personal Research Assistant

Appendix B: Introduction to the NewsEye Platform for Research and Analysis of Historical Newspapers for the BNF DataLab

Introduction à la plateforme NewsEye pour la recherche et l'analyse de la presse ancienne

Financé par le programme de recherche et d'innovation Horizon 2020 de l'Union européenne, NewsEye est un projet de recherche visant à faire progresser les connaissances actuelles et à introduire de nouveaux concepts, méthodes et outils pour les humanités numériques en offrant un meilleur accès à la presse ancienne à un large éventail d'utilisateurs.

Les journaux recueillent des informations sur les événements culturels, politiques et sociaux d'une manière plus détaillée que toute autre archive publique. Depuis leur apparition au XVIIe siècle, ils enregistrent des quantités phénoménales d'événements, d'histoires et de noms, dans presque toutes les langues, dans tous les pays et quasi-quotidiennement. Les journaux constituent depuis toujours un moyen incontournable pour diffuser les opinions publiques et politiques, les œuvres littéraires, les essais et l'art : une richesse thématique qui les rend incontournables pour quiconque s'intéresse au patrimoine culturel européen.

Les dernières décennies ont vu la numérisation et la mise à disposition en ligne de dizaines de millions de pages de journaux émanant des bibliothèques européennes, alors même que les bibliothèques nationales intensifient leurs efforts de numérisation pour les années à venir du fait de la forte demande pour accéder à la presse ancienne. Si cette ressource culturelle et historique suscite un intérêt général auprès du grand public, elle revêt une importance capitale pour de nombreux chercheurs en sciences humaines.

La plateforme NewsEye (https://platform2.newseye.eu) a été développée pour mettre à disposition du grand public les résultats de recherche de ce projet.

1. Données disponibles

A. Journaux accessibles

Grâce à cette plateforme, il est possible de rechercher de l'information dans la vingtaine de titres disponibles, en cinq langues différentes. Ces titres de presse ont été fournis par les trois bibliothèques partenaires du projet : la Bibliothèque Nationale de France (BNF), la Bibliothèque Nationale d'Autriche (ONB) et la Bibliothèque Nationale de Finlande (NLF). En voici la liste :

BNF (français)

- La Presse : de 1850 à 1890 (Lien)
- Le Matin : de 1884 à 1944 (Lien)
- La Fronde : de 1897 à 1929 (Lien)

- Marie-Claire : de 1937 à 1944 (Lien)
- L'Œuvre : de 1915 à 1944 (Lien)
- Le Gaulois : de 1868 à 1900 (Lien)

ONB (allemand)

- Neue freie Presse : 1864-1873, 1895-1900, 1911-1922 et 1933-1939
- Illustrierte Kronen Zeitung : 1911-1922 et 1933-1939
- Innsbrucker Nachrichten (Mittags-Zeitung) : 1864-1873, 1895-1900, 1911-1922 et 1933-1939
- Arbeiter-Zeitung : 1895-1900, 1911-1922 et 1933-1939

NLF (finnois et suèdois)

- Sanomia Turusta : 1850-1900
- Aura : 1880-1896
- Uusi Aura : 1897-1918
- Suometar : 1847-1866
- Uusi Suometar : 1869-1918
- Päivälehti : 1889-1904
- Helsingin Sanomat : 1904-1918

- Åbo Underrättelser : 1824-1827 et 1829-1918
- Västra Finland : 1895-1918
- Hufvudstadsbladet : 1864-1918

Ces titres de presse ont tous été analysés par la chaîne de traitement du projet NewsEye. Cette dernière consiste en différentes étapes d'analyse d'image et d'extraction d'information :

- Analyse de l'agencement des pages
- Reconnaissance automatique du texte
- Extraction des articles

Il est important de noter que seul le journal L'Œuvre a été traité par l'algorithme final d'extraction des articles. Ainsi, la qualité des articles des autres journaux sera bien moindre que celle du journal L'Œuvre.

B. Entités nommées

Ensuite, un enrichissement sémantique des articles est effectué grâce à l'analyse des entités nommées. Ces dernières représentent un morceau d'information textuelle particulièrement porteur de sens. On peut leur attribuer de nombreux types mais dans notre cas, nous nous intéressons particulièrement aux *personnes*, *lieux*, *organisations* et *créateurs de contenu*. Le processus d'analyse est constitué des étapes suivantes.

1. Identification des mentions : il s'agit ici de repérer dans le texte des articles les mots ou suites de mots qui sont considérés comme des entités nommées.
2. Classification de ces mentions : une fois les mentions identifiées, il faut les classer dans les types prédéfinis (*personnes*, *lieux*, *organisations* et *créateurs de contenu*).
3. Mise en relation avec une base de connaissance : cette étape tente de supprimer certaines ambiguïtés en associant les mentions à un identifiant Wikidata. Ainsi, si deux personnes différentes portent le même nom, nous sommes en capacité de les différencier. Le deuxième avantage est lié à l'aspect multilingue de la collection

disponible. Des mentions de langue différente peuvent ainsi être liées, créant un lien sémantique important entre deux articles de langue différente.

4. Analyse de la polarité : nous tentons ici d'analyser le point de vue (positif, neutre ou négatif) d'un article envers les différentes mentions d'entités nommées qu'il contient.

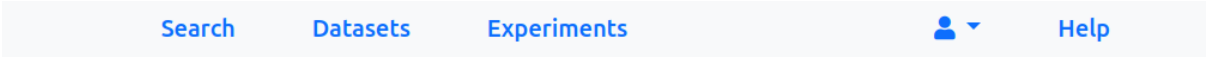
2. Premiers pas

A. Inscription et connexion à la plateforme

La première étape avant de pouvoir utiliser la plateforme est de se créer un compte. C'est gratuit et très facile. Connectez-vous d'abord à l'adresse suivante à l'aide d'un navigateur internet récent (https://platform.newseye.eu). Puis cliquez sur le lien "inscription" visible sous le formulaire de connexion. Remplissez le formulaire avec votre email et un mot de passe. Une fois cette étape réalisée, vous serez automatiquement connecté. La prochaine fois que vous voudrez vous connecter, il vous suffira de rentrer votre email et votre mot de passe.

B. Aperçu de la page d'accueil

Une fois connecté, vous êtes automatiquement redirigé vers la page d'accueil de la plateforme. Elle est composée de plusieurs éléments. Parlons d'abord de la barre de navigation, qui sera visible sur toutes les pages de la plateforme.



Les différents liens permettent d'accéder à différentes fonctionnalités de la plateforme. L'interface de recherche est disponible en cliquant sur ***Search***. Vous pouvez également accéder à vos jeux de données personnels (***Datasets***) ou à l'assistant d'analyse de données (***Experiments***). L'icône utilisateur vous permet de vous déconnecter ou de modifier votre mot de passe. Finalement, l'icône ***Help*** vous permet de regarder des vidéos d'aide concernant l'utilisation de la plateforme.

3. Chercher de l'information dans les collections

A. Recherche textuelle

L'interface de recherche est constituée des éléments suivants.

La première chose à choisir est le type de recherche que vous voulez réaliser : recherche exacte ou approximative. Vous pouvez sélectionner le type de recherche souhaitée en cliquant sur l'un des deux boutons à gauche de la barre de recherche. Par défaut, la recherche exacte est sélectionnée mais vous pouvez la changer pour une recherche approximative (***Stemmed search***).

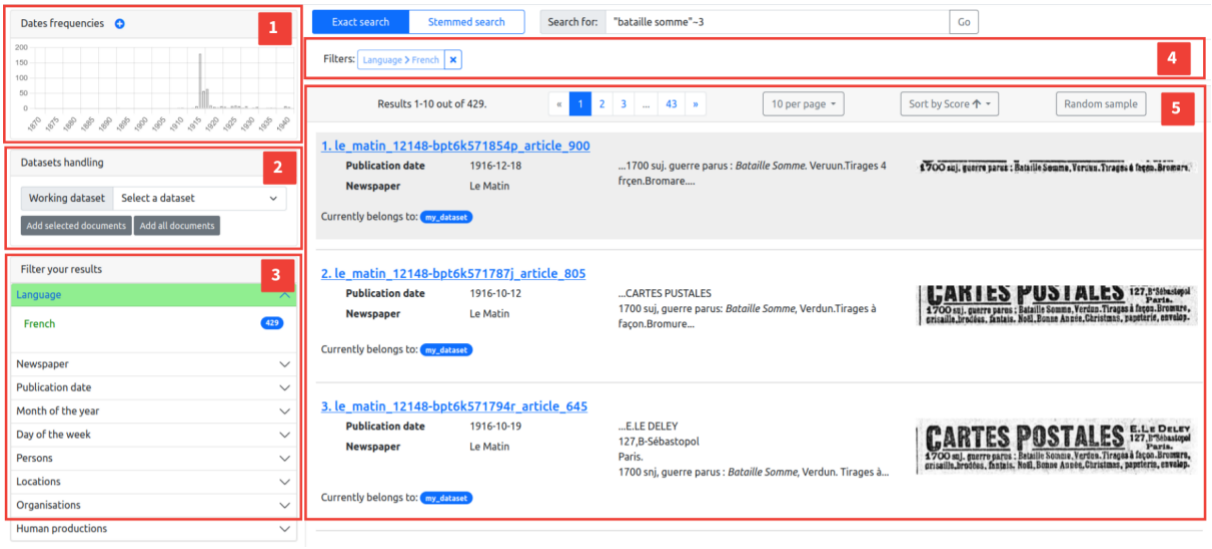
Recherche exacte : le système tentera de trouver des documents qui contiennent exactement les mêmes termes que ceux que vous indiquerez dans la barre de recherche.

Ce type de recherche est utile pour trouver des documents de manière plus précise. En revanche, il est possible que certains documents pertinents ne soient pas renvoyés par le système à cause des erreurs de reconnaissance du texte qui peuvent exister.

Recherche approximative : ici, le système va identifier des documents qui contiennent le radical de vos termes de recherche. Ainsi, si vous cherchez le terme ***université***, le système renverra des documents qui contiennent des termes comme ***universités***, ***universitaire***, mais aussi ***universel***. Si ce mode de fonctionnement permet de trouver plus de documents potentiellement pertinents, des documents non pertinents peuvent donc aussi être renvoyés.

Dans le champ de recherche, vous pouvez indiquer un ou plusieurs termes qui vous intéressent. Vous pouvez également indiquer au système l'éloignement de ces termes dans les documents renvoyés. Pour ce faire, il vous faut mettre ces termes entre guillemets. Ainsi la requête ***"bataille somme"~10*** va renvoyer des documents qui contiennent les mots ***bataille*** et ***somme*** espacés au maximum de 10 mots. Vous pouvez omettre ***~10*** pour rechercher deux termes qui doivent se trouver l'un à côté de l'autre.

B. Liste de résultats



Les documents renvoyés par le système après une recherche sont affichés sous forme de liste (**5**). Vous pouvez ici choisir le nombre de résultats affichés par page et naviguer entre les différentes pages. Vous pouvez aussi choisir l'ordre de tri des résultats (par ordre de date de publication ou par pertinence par rapport à votre requête). Finalement, cliquer sur le bouton ***Random sample*** permet d'afficher 10 articles pris au hasard parmi les résultats de recherche. Cette fonctionnalité permet d'avoir un aperçu rapide des résultats et donne ainsi des indications sur la qualité des mots-clés choisis par l'utilisateur par rapport à son besoin. Cliquer sur le titre ou l'image d'un document permet d'accéder à l'outil de consultation (voir partie 3.C).

Le graphique situé en **1** représente la distribution temporelle des résultats renvoyés par le système après une requête. Il permet d'avoir un aperçu rapide de l'étendue temporelle des documents correspondant à votre requête. Les paramètres de cette requête sont résumés dans le cadre **4**.

Le cadre **2** permet de gérer l'appartenance des documents à un jeu de données préalablement créé (voir section 4.A). Il est possible de sélectionner les articles parmi les résultats de recherche en cliquant dessus. Un cadre apparaît autour des articles sélectionnés de cette manière. Il faut ensuite choisir le jeu de données auquel les articles seront ajoutés en le sélectionnant dans la liste. Finalement, cliquer sur le bouton ***Add selected documents*** va ajouter les documents au jeu de données. Une autre possibilité est d'ajouter l'ensemble des résultats de recherche à un jeu de données. Pour faire cela, il est nécessaire de cliquer sur le bouton ***Add all documents***.

Finalement, la zone **3** permet de filtrer la liste de résultats selon différents aspects.
Langue : la plateforme comporte des documents en 5 langues différentes. Ce paramètre permet de filtrer la liste de résultats pour se concentrer sur un langage particulier.
Date : cliquer sur le "from" ou "to" permet d'établir les bornes d'un intervalle temporel qui va filtrer la liste de résultats.
Month/Day : ces paramètres permettent de filtrer les résultats selon le mois de l'année ou le jour de la semaine de la date de publication. Cela peut par exemple servir à concentrer les résultats sur les éditions du Samedi, peu importe le mois ou l'année.
Journal : ce paramètre vous permet de filtrer les résultats selon le journal duquel ils proviennent.
Persons/Locations/Organisations/Human Productions : les filtres d'entités nommées permettent de limiter les résultats à ceux qui comportent une mention d'entité nommée qui a été associée à une base de connaissance.

C. Lecteur de document

Après avoir cliqué sur un des résultats présentés après une recherche, vous arriverez sur une nouvelle page comportant le lecteur de document.

Le lecteur situé en **3** vous permet de naviguer à travers les pages du journal grâce à votre souris. Vous pouvez sélectionner ou désélectionner des articles en cliquant dessus. Lorsqu'un article est sélectionné, sa transcription automatique apparaît en **6**.

Des informations sur le numéro que vous êtes en train de visualiser sont présentées en **4**. Vous pouvez y retrouver la date de publication du numéro, le journal duquel il provient ainsi que le nombre de pages qui le compose.

La zone **1** permet de gérer l'appartenance du numéro ou de ses articles à un jeu de données précédemment créé. Pour ce faire, vous pouvez sélectionner le jeu de données auquel vous voulez ajouter des articles avec le sélecteur "working dataset". Une fois fait, vous pouvez ajouter le numéro dans son ensemble ou l'article sélectionné dans ce jeu de données.

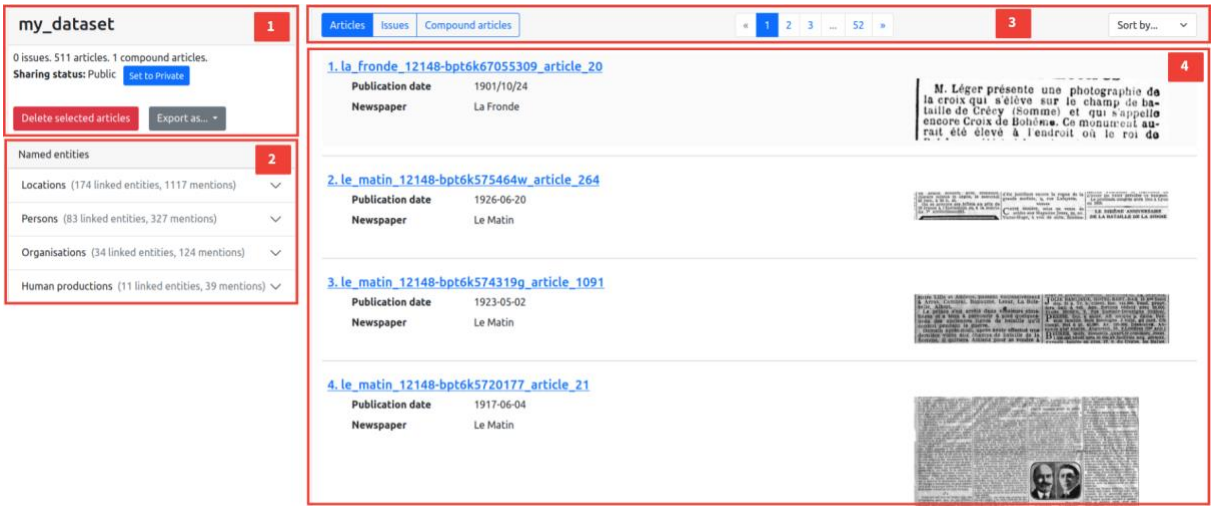
La zone **5** permet de construire des articles composés, plus d'informations à ce sujet sont présentées plus tard dans ce documents.

4. Utilisation avancée

A. Jeux de données

Une des nouveautés de la plateforme NewsEye est la possibilité pour les utilisateurs de créer différents jeux de données dans lesquels ils peuvent ajouter des articles ou numéros entiers, provenant de différentes recherches. Ces jeux de données rassemblent des documents qui traitent d'une problématique commune selon votre besoin. Vous pouvez ajouter des documents à un jeu de données depuis la page des résultats de recherche (partie 3.B, cadre n°**2**) ou depuis le lecteur de document (partie 3.C, cadre n°**1**).

Vous pouvez créer un jeu de données ou accéder à l'ensemble de vos jeux de données en cliquant sur l'icône ***Datasets*** du menu dans la partie supérieure de la page. Sur cette page vous verrez aussi les jeux de données publics, partagés par d'autres utilisateurs. Vous pouvez importer ces jeux de données et les observer ou les manipuler comme s'ils étaient les vôtres. Lorsque vous cliquez sur le titre d'un de vos jeux de données, vous accédez à la page suivante.



Les documents présents dans votre jeu de données sont présentés sous forme de liste, de la même manière que pour les résultats d'une recherche (cadre n°**4**). Vous pouvez naviguer entre les pages, choisir quel type de document afficher et trier ces documents dans le cadre n°**3**. Le cadre n°**1** comporte des informations sur votre jeu de données ainsi que différents outils. C'est ici que vous pouvez notamment choisir de rendre votre jeu de données public. Ici, vous pouvez aussi exporter votre jeu de données en différents formats. Le format ZIP est le plus facile à prendre en main, et comporte un fichier par document de votre jeu de données. Ce fichier contient uniquement la transcription automatique du document. Le format JSON est plus avancé est vous permet de récupérer non seulement le

texte associé à chaque article mais aussi différentes métadonnées, les entités nommées et un lien vers une image de cet article.

Le cadre n°**2** vous permet de voir quelles entités nommées sont présentes dans votre jeu de données. Les mentions qui n'ont pas pu être liées à une base de connaissance (voir partie 1.B) sont regroupées sous le titre ***Unlinked***.

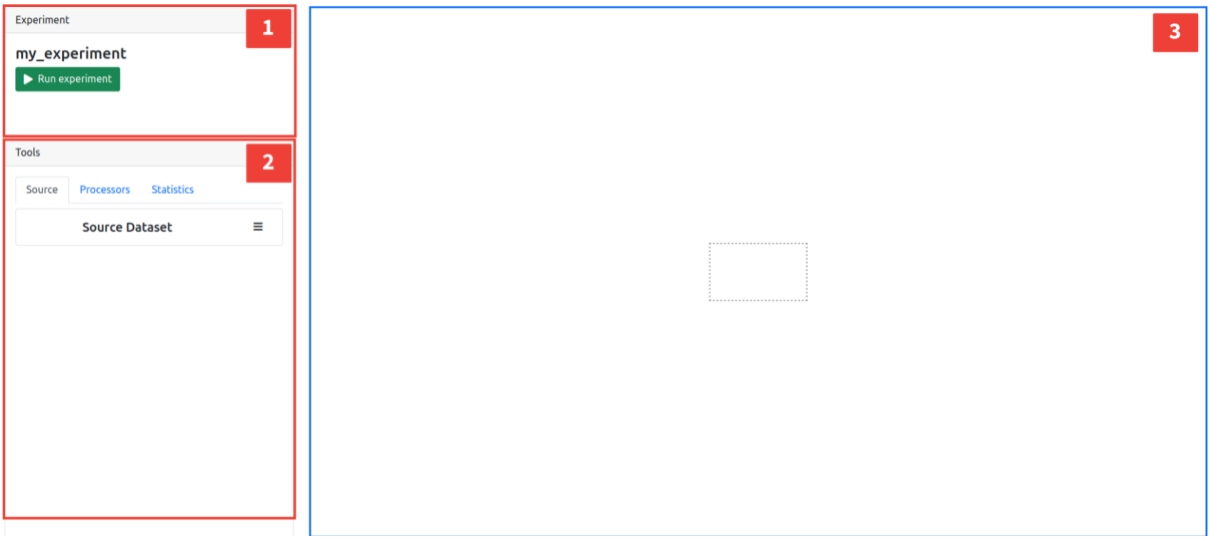
B. Articles composés

Comme vous pourrez le constater, la séparation automatique des articles marche plus ou moins bien selon les documents. Or, il peut être intéressant de vouloir ajouter à un jeu de données un article composé de plusieurs blocs de textes (d'où l'appellation d'article composé). Vous pouvez réaliser cette opération depuis le lecteur de document, en activant ce mode (partie 3.C, cadre n° **5**). Une fois activé, vous n'avez qu'à cliquer sur les différents blocs de texte que vous voulez fusionner. Vous pouvez modifier l'ordre de ces blocs dans le cadre n°**5**. Une fois terminé, vous pourrez obtenir un aperçu de votre article composé et lui donner un nom.

Pour ajouter un article composé à un jeu de données préalablement créé, il faut sélectionner celui-ci dans la liste située en haut à gauche de la page. Un nouveau bouton apparaît alors dans le cadre n°**1** et vous permet d'ajouter cet article composé à votre jeu de données.

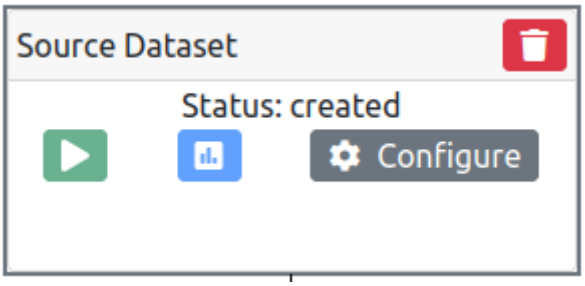
C. Espace expérimental

Ce site contient aussi un espace expérimental en cours de création qui permet d'analyser un jeu de données précédemment créé. Cet espace est accessible en cliquant sur le lien ***Experiments*** dans la barre de navigation (voir section 2.B). Une fois sur cette page, vous pouvez créer une nouvelle expérience et lui donner un titre. Il suffit ensuite de cliquer sur la nouvelle entrée apparue dans la liste pour accéder à la page de l'expérience.



La zone **1** contient le nom de l'expérience actuelle ainsi qu'un bouton pour exécuter les différents composants de l'expérience. La zone **2** est l'endroit où vous trouverez l'ensemble des outils disponibles. Pour l'instant très réduit, leur nombre est amené à augmenter. Pour ajouter un outil à l'expérience, il suffit de le glisser-déposer dans la zone **3**.

Un emplacement en surbrillance apparaîtra à l'endroit où il est possible d'utiliser cet outil. Une fois placé, voilà à quoi ressemble un outil.



Avant toute chose, chaque outil doit être configuré en cliquant sur le bouton ***Configure***. Une fenêtre apparaît alors avec les différents paramètres possibles de l'outil. Le bouton vert permet d'exécuter l'outil, le bouton bleu permet d'afficher ses résultats (une fois qu'il a été exécuté) et le bouton rouge permet de le supprimer de l'expérience.

Le premier outil, essentiel, est la source de données (***Source Dataset***) qui correspond à un jeu de données précédemment créé. Un autre outil disponible est le pré-traitement des textes du jeu de données. Plusieurs paramètres sont disponibles. Il est ainsi possible de :

- passer les textes en minuscule
- enlever la ponctuation
- enlever les chiffres
- enlever les mots vides (mots peu porteurs de sens comme ***le***, ***la***, ***les***, etc.

Le dernier outil disponible, pour le moment, est l'extraction de n-grammes. Ces derniers correspondent à toutes les séquences de n mots que l'on peut trouver dans les textes d'un jeu de données. Cet outil calcule donc ces n-grammes et compte leur fréquence.

Il est ainsi possible d'enchaîner ces différents outils pour créer une chaîne de traitement permettant d'analyser le contenu d'un jeu de données. Encore une fois, cet espace expérimental est incomplet, et de nouveaux outils d'analyse apparaîtront dans le futur.