Project Number: **770299**

NewsEye:

A Digital Investigator for Historical Newspapers

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

D6.9: Usability/Fit for research purpose test of tools and user interfaces (c) (final)

Due date of deliverable: M45 (31 January 2022)

Actual submission date: 19 January 2022

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

Partner organization name in charge of deliverable: UIBK-ICH

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:	D6.9	
Deliverable full title:	Usability/Fit for research purpose test of tools and user interfaces (c) (final)	
Deliverable short title:	Tool testing (final)	
Document identifier:	NewsEye-T61-D69-UsabilityTestOfTools-c-final-Submitted-v6.0	
Lead partner short name:	UIBK-ICH	
Report version:	V6.0	
Report preparation date:	19.01.2022	
Dissemination level:	PU	
Nature:	Report	
Lead author:	Eva Pfanzelter (UIBK-ICH); Sarah Oberbichler (UIBK-ICH)	
Co-authors:	Jani Marjanen (UH-DH); Nejma Omari (UPVM)	
Internal reviewers:	Hannu Toivonen (UH-CS); Minna Kaukonen (UH-NLF)	
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
29/01/2021	0.1	Eva Pfanzelter and Sarah Oberbichler (UIBK-ICH)	First draft
31/01/2021	1.0	Eva Pfanzelter (UIBK-ICH)	Proofread, included final comments from co-authors, submitted to internal reviewers
12/02/2021	2.0	Eva Pfanzelter and Sarah Oberbichler (UIBK-ICH)	Included comments from reviewers, changed some parts with the help of Hannu Toivonen, Lidia Pivovarova and Leo Leppänen (UH-CS), sent for quality management
26/02/2021	3.0	Eva Pfanzelter and Sarah Oberbichler (UIBK-ICH)	Included comments from quality management
14/12/2021	4.0	Eva Pfanzelter and Sarah Oberbichler (UIBK-ICH)	Draft update with the feedback on the work done during the project's extension period, sent to internal reviewers
21/12/2021	5.0	Eva Pfanzelter and Sarah Oberbichler (UIBK-ICH)	Final update including feedback from internal reviewers, sent for quality management
19/01/2022	6.0	Antoine Doucet (ULR)	Minor adjustments and submission

Executive summary

Within the project 'NewsEye: A Digital Investigator for Historical Newspapers', researchers from computer science and digital humanities collaborate with the three national libraries of Finland, France and Austria in order to develop methods and tools for effective exploration and exploitation of digital newspaper collections. To make these rich resources of cultural heritage better accessible by means of new technologies and 'big data' approaches, 'close' and 'distant reading' methods of digital humanities are being investigated and combined. The aim is to improve the ways researchers and experts, as well as the interested general public, study European cultural heritage.

This deliverable 'Usability/Fit for research purpose test of tools and user interfaces (c)' is the final and public report on Task T6.1 led by Eva Pfanzelter (UIBK-ICH), due at M34 and updated for M45. It deals with the testing of tools, methods and user interfaces by humanities researchers to ensure their usability/fit for research purpose. It is a result of the collaboration of the DH group on the mock-ups and prototypes, workshop/hackathon participation with the computer science groups, and the libraries providing extensive feedback on tools and methods. Members of the DH group in Innsbruck (UIBK-ICH), Helsinki (DH-UH), Montpellier (UPVM), and Vienna (UNIVIE) tested the methods and tools suggested and produced by computer scientists from the University of La Rochelle (led by Antoine Doucet, ULR), the University of Helsinki (led by Hannu Toivonen, UH-CS), the University of Innsbruck (led by Günter Mühlberger, UIBK-DEA), as well as mathematicians from the University of Rostock (led by Roger Labahn, UROS) in order to ensure their efficiency also when used in cases other than those analysed within the project.

The first part of the report describes the DH collaboration, demonstrating a mesh up of research disciplines, tasks and approaches. In the second part, the advances achieved within the project for the application of methods, tools and algorithms are reflected on. It can briefly be concluded that in all areas envisaged within the project, considerable progress was made. This especially applies to the tasks mentioned in WP 1 and WP 2 and to some tasks in WP 3, as well as to the Demonstrator mentioned in WP 7. The deeper the project team dug into approaches supporting both quantitative and qualitative interpretations (such as those mentioned in WPs 3 and 4), the more complex the challenges became. In order to help DH researchers and also a general audience to better use these tools (and also the data), a Personal Research Assistant (PRA) was developed in WP 5. It is seen as an aid to the user in analysis tasks and integrates the tools developed in WPs 3 and 4. Although the project delivered its results as intended, the PRA also shows where further research could prove especially useful for all disciplines involved in the project.

Contents

1 Collaboration between participating project groups for the advancement of tools, methods and approaches

Over the past years, a lively collaboration between the various teams working together in the NewsEye project ensued (detailed information about e.g. tools testing was distributed internally in the previous reports D6.3 and D6.8). Before the preparation of tools, the DH group conducted surveys on the libraries' interfaces and tried out various methods and tools in order to determine which applications seemed to be the most relevant for the different user groups addressed in the project. The analysis of surveys and existing methods was therefore the basis for the NewsEye team to engage in the development of specific sets of tools and algorithms. The DH groups' subsequent active participation in the evaluation processes of the developed tools will be summarized in the following subsections.

1.1 Testing of tools and methods

Several tool testing sessions (initially during face-to-face meetings, online since March 2020) took place and allowed the DH group to give feedback on the methods, tools and algorithms. Members of WP 6 collaborated with WP 2 (Text Recognition and Article Separation), WP 3 (Semantic Text Enrichment), WP 4 (Dynamic Text Analysis), WP 5 (Personal Research Assistant) and WP 7 (Demonstration, Dissemination, Outreach and Exploitation). The meetings took place as follows:

- 15 May 2019: Helsinki
- 20 November 2019: Montpellier
- 11 May 2020: online
- 24 September 2020: online
- 10 November 2020: online
- 17 December 2020: online
- 21 January 2021: online
- 19 March 2021: online
- 7 December 2021 (in conjunction with User Workshop): online

1.2 Annotations

Members of the DH group participated in the writing of annotation guidelines and carried out annotations according to these guidelines. This included annotations for the recognition and linking of named entities (dates, locations, person names, organizations, etc.), article separation and event detection in collaboration mainly with WP 1 (Data Management), WP 2 and WP 3. Although this was an ongoing process throughout the past years, the following summary of activities can be made:

- 2018 – 2021: continuous online discussions, exchanges on Slack, and web conferences
- collaboration in the named entity recognition (NER), linking (NEL) and stance working group
- participation in the article separation working group
- discussion and evaluation of guidelines for article separation as well as NER, NEL and stance annotations
- 2019 – May 2020: annotations for NE
- September 2020 – February 2021: annotations for event detection
- May 2021: annotations for Layout analysis

As a result, the DH team in Innsbruck created ground truth in article separation for 230 pages, annotated around 70 pages for NER/NEL and stance detection as well as about 50 articles for event detection in German newspapers. For Finnish-language and Swedish-language newspapers NER/NEL annotation was carried out by the UH-NLF team.

1.3 DH group team building and collaboration

In order to organize the collaborative work, several in-person meetings and internal workshops were organized by the DH group, and its members also participated actively in the bi-yearly steering committee meetings, which always included DH group meetings. These events, which were a good opportunity for getting to know each others' work and disciplines when developing tools, methods and algorithms, took place in different locations in-person and, since March 2020, online. To summarize:

- Steering committee meetings twice a year since April 2018:
  - May 2018: La Rochelle
  - November 2018: London
  - May 2019: Helsinki
  - November 2019: Montpellier
  - May 2020: e-Rostock (online)
  - November 2020: e-Vienna (online)
  - March 2021: Exeptional SC-meeting (online)
  - April 2021: e-Paris (online)
  - October 2021: e-La Rochelle (online)
- DH group internal meetings:
  - Monthly (since 2018) resp. bi-weekly (since March 2020): DH group online meetings
  - February 2019: DH group workshop Innsbruck
  - February 2020: DH group workshop Berlin
  - February 2020: DH group feedback workshop Vienna
- Work visits of DH group members:
  - May 2019: DH group members from Innsbruck, Vienna and Montpellier participated in the Helsinki DH Hackathon (#DHH19) organized by the DH and CS team members at the University of Helsinki
  - August–September 2019: Sarah Oberbichler from UIBK-ICH visited La Rochelle and Helsinki
  - March 2019–January 2020: Jani Marjanen from UH-DH visited the research group at UH-CS on a weekly basis
  - July–August 2020: Stefan Hechl from UIBK-ICH 'visited' the Rostock colleagues online on a regular basis to work together on layout segmentation and article separation
  - September–October 2020: Sarah Oberbichler from UIBK-ICH collaborated with the CS team in La Rochelle (tool development and documentary about NewsEye)
  - October 2020: Nejma Omari and CS colleagues from La Rochelle visited BNF for a documentary about NewsEye
- Collaboration for publications in different teams: The papers are publicly available on Zenodo: https://zenodo.org/communities/newseye/?page=1&size=20.
- Collaboration with WP 7 concerning the NewsEye platform: see Section 2.6 below

2 Advancement and application of methods, tools, and algorithms

2.1 WP 1 Data management

2.1.1 Selection of research data

WP 1 (led by Günter Mühlberger) specified data formats and data models in order to make them 'fit for purpose' for different research and application areas (e.g., research in computer science and digital humanities, end users, preservation-oriented applications in digital libraries). Although the members of the DH group were not directly involved in the management of the data, they pre-selected the data (based on their specific needs such as political orientation, periodicity, etc.) that was then given access to by the libraries (ONB, BNF, NLF) and processed by UIBK-DEA. The newspapers and time slices selected for further processing and research data within the project were:

Table 1. The newspapers and time slices selected for further processing

Title	Time Period
La Presse	1850-1890
Le Matin	1884-1944
La Fronde	1897-1929
Marie-Claire	1937-1944
L'Œuvre	1915-1944
Le Gaulois	1868-1900
Neue Freie Presse	1864-1873; 1895-1900; 1912-1922; 1934-1945
Illustrierte Kronen Zeitung	1912-1922; 1934-1945
Innsbrucker Nachrichten	1864-1873; 1895-1900; 1912-1922; 1934-1945
Arbeiter-Zeitung	1895-1900; 1912-1922; 1934-1945
Aura	1880-1897
Helsingin Sanomat	1903-1919
Päivälehti	1889-1905
Sanomia Turusta	1850-1904
Suometar	1850-1867
Uusi Aura	1898-1919
Uusi Suometar	1868-1919
Åbo Underrättelser	1850-1919
Hufvudstadsbladet	1863-1919
Västra Finland	1895-1919

NewsEye participated to the open research data pilot (ORDP) of the European commission, and details on data management and the FAIR usage of data within NewsEye can be found in the public Deliverable D1.12 'Data management plan', delivered at month 6 of the project. As stated there, the NewsEye contribution in this regard is that all tools, services and datasets developed within NewsEye are made available on the the project website, a GitHub repository, and the publications and datasets available on Zenodo, and to support sustainability beyond the project duration.

2.1.2 Metadata

An additional contribution made by the DH group made was to highlight the importance of metadata and to continuously demand the addition of metadata to the processed data. In an ideal setting, the requested metadata goes beyond the provided information on language, amount of pages/articles or newspaper ID, and concerns the analogue material (e.g., the state of the paper versions, publication frequency, missing issues, completeness of the digitized collection, information on the digitized newspaper, political orientation, changing editors and editorial offices, languages, etc.) as well as the digitization processes (e.g., utilized software and hardware, image resolution, text recognition accuracy, availability, ownership, re-use options, etc.). This kind of metadata is needed in DH studies because without it, digital source, tools, methods, interface, etc. criticism is difficult to perform and reliability on the digital data collections is diminished [1]. Transparency on OCR issues, digitization processes, the criteria for the selection of newspaper titles, and similar information can greatly enhance humanities scholars' research efforts. Although much of this metadata is not available for the newspaper data we were able to process within the NewsEye project, the awareness of the importance of the availability of metadata certainly grew over the duration of the project (for the results on usability/fitness, see [2]).

2.2 WP 2 Text Recognition and Article Separation

Source preparation, which contains all digitization steps including layout analysis and article separation, is the initial measure to provide digital access to cultural heritage material such as newspapers. The automated conversion of images of historical documents to electronic text is a first automated step in the workflow. Layout analysis and the highly important 'article separation', dividing the OCRed text into news units, is a second essential step in this process.

Within the NewsEye project, text recognition and article separation were implemented by the partners from the University of Innsbruck (led by Günter Mühlberger) and the University of Rostock (led by Roger Labahn). WP 2 investigated, developed and implemented methods, algorithms, and tools for automated text recognition (ATR) and article separation (AS).

2.2.1 Automated Text Recognition

From the perspective of the DH group and as a result of the various tools testing sessions, it is clear that text recognition errors for the data have been reduced to a point where their impact on search and analysis is minimal. As can be concluded from the publicly available deliverable on ATR (D2.5), the error rate for ATR in the NewsEye platform is now below 1 per cent. Since ATR has an effect on all further research steps, such low error rates are essential for humanities' research [2].

Despite this success, the NewsEye team has not stopped working on methods for further improvement and has continued investing ATR post-correction approaches. Computer scientists from the universities of La Rochelle and Helsinki are currently working on automated tools and methods that are able to eliminate residual ATR errors by either applying different spelling correction methods or using advanced neural methods along with word representations [3, 4, 5, 6]. Although these methods are not yet implemented in the NewsEye platform, the tools testing sessions showed that such approaches seem promising, not only for ATR noise, but also for spelling errors or variations that already exist in newspapers.

2.2.2 Article Separation

Work on article separation (AS) combining different models and algorithms by the CS colleagues from Rostock is still ongoing [7] (see also public Deliverable 2.7: Article Separation), and various approaches have been tried out in collaboration with Rostock, Helsinki, La Rochelle and Innsbruck by using topic models or word embeddings. As it became apparent in preparation of the corpora for tools, methods and automated analysis, AS is more important than previously thought, since skewed layout results (e.g. cut-off letters at the end of columns) influence both ATR and subsequent neural methods mentioned above. This has led to the development of manual options for merging articles, as can be read about in more detail in Section 2.6 on the NewsEye platform.

2.3 WP 3 Semantic Text Enrichment

In order to satisfy the needs for finding articles or managing collections, the NewsEye partners from La Rochelle (led by Antoine Doucet) have implemented methods that can provide finer-grained keywords (dates, locations, person names, events, etc.) than in multilingual named entity recognition [8] and linking [9] (see also public Deliverable 3.5: NE recognition and linking), stance detection (see public Deliverable 3.6: Stance Detection) or event detection [10] (see also public Deliverable 3.8: Event Detection).

The DH team contributed to the development of NER, NEL, stance and event detection by providing annotations in all languages (German, French, Finnish, Swedish) and continuously giving feedback to the computer scientists on how their tools performed. In doing so, they followed the annotation guidelines written by the CS colleagues and adapted them for NewsEye needs. These annotations were needed in order to evaluate and train the methods to maximize their quality and to be as appropriate as possible for humanities researchers. The collaboration has been very rewarding and the computer science team in La Rochelle has made competition-winning progress in multilingual NER and NEL [11]. From the perspective of DH scholars, the tools developed for NER and NEL have proven useful for humanities research, especially when combined with specific keyword searches or corpus building functions.

2.3.1 Named Entity Recognition and Linking

Multilingual named entity recognition and linking extracts relevant information such as dates, locations, person names, organisations, etc. from documents. In the NewsEye platform, the tool is based on various statistical and probabilistic natural language processing (NLP) techniques that rely on the latest advancements in the field of artificial intelligence (e.g. deep neural networks, language models) (for more information, see public Deliverable D3.5: Named Entity Recognition and Linking). Even though the automated recognition and especially the automated linking of named entities is still error-prone [12, 13], the tool promises to be a real step forward. The DH group engaged in various experiments with their colleagues in order to find out how to make the most of the multilingual newspaper material available, e.g. they tried out several ways to find common topics in the multilingual material. Although these experiments were not always successful, the usability of NER/NEL for multilingual research questions was proven [11], highlighting the importance of NEL for humanities research in different language settings, such as identifying outstanding personalities as well as places and organizations that have a particular importance within the researched topic. To use the identified NE for further occurrence searches

turned out to be very successful. The same applies to linking of NE to Wikidata, which provided some necessary contextualization.

2.3.2 Event Detection

An event detection tool extracts events and the participants of these events. An event can be a public health occurrence, a cultural event, an act of terrorism or crime, etc. Event information usually contains locations, dates, organizations, persons, or consequences as well as further named entities. Even though the event detection tool is not yet implemented in the NewsEye platform and the annotation process is still running, some research papers [14, 15, 16] (see also report D3.7: Event Detection) and preliminary experiments have shown that event detection supports humanities' research for the creation of individual collections or by supporting the search process. At the moment, the promise of event detection for historical scholarship is still undefined. Once detection of events provides more fine-grained information for different types of events, it should be very reliable in producing networks between events, persons and organizations.

2.3.3 Stance Detection

WP 3 also worked on methods for stance detection, an automatic classification of textual content into one of these three classes: positive, negative, and neutral [17] (see also report D3.6: Stance Detection). Even though the DH group was skeptical about the reliability and functionality of stance detection, the final results proved useful for humanities research to a certain extent. Stance detection gives an overall picture regarding the positive, negative or neutral use of place names, person names or organisations, which can be seen as a hint for further research or particular (changing) moods. For example, analysis of a dataset on return migration showed that Brazil in particular was often described negatively, which is consistent with the discourse in newspapers. At the same time, positive, negative and neutral attributions are not that easy to make, e.g. censorship regimes and press control force journalists to use irony in their reporting, which in itself is tricky to identify by human readers and, as of today, almost impossible to trace with algorithms. Also, stance detection can only determine explicit expressions, while implicit expressions remain hidden. In order to make the method usable for research, the DH team has drawn attention to these and many other issues and asked for help files and understandable visualizations that explain the results to the users in a transparent way.

2.4 WP 4 Dynamic Text Analysis

Many historians use language data in newspapers as an entry point for studying historical processes that the newspapers reported on by using e.g. frequency analysis [18], but quite a few historians are also interested in studying the discourse in its own right, meaning that there is a renewed interest in language as an indicator of historical change. Such studies move from using interfaces and algorithmic methods to finding relevant sources for producing representations of changes in past discourses. For instance, a topic model can be used to cluster similar documents and produce a subcorpus for closer study [19], but it can also be used as an indicator for something that is studied in the data.

Within the NewsEye project, WP 4 (led by Mark Granroth-Wilding) developed methods and provided tools for the analysis and exploration of historical newspapers for above-mentioned research purposes.

Implementing the methods developed in WP 4 in the platform, which contains 1.5 million newspaper pages, has shown that new approaches that work well with small datasets need to be adapted for the huge amount of data available in the platform. It soon became evident that this demanded compromise on various levels. Several tool testing sessions helped address issues regarding functionality, understanding, and transparency of the analysis visible in the platform. The DH group highlighted the importance of supporting interface, tool, methods, and algorithm criticism, which is essential for reliable humanities research. As a result, the NewsEye platform and the tools existing therein were improved step by step over the past months. Both the platform and the tools are designed in order to support researchers in their critical engagement with the data, metadata, and methods.

2.4.1 Topic Modeling and Document Linking

The DH group started with high expectations concerning topic modeling. A large amount of research data was transferred to the computer science colleagues in Innsbruck, Rostock, La Rochelle and Helsinki. The latter tried out several existing mono- and multilingual topic modeling methods [19, 20] (see also public Deliverable 4.5: Analysis of data in a given context). After training latent Dirichlet allocation (LDA) and dynamic topic models (DTM) in German, Finnish and French over the entire corpora, it soon became clear that the trained topics (20, 30 and 50 per language) did not make (enough) sense when DH researchers tried to find answers to specific research questions. Since creating topic models 'on the fly' was not feasible, the solution is a temporary workaround. Depending on the search or dataset of the user, the tool in the platform now points to the most salient topics within the previously trained topics for each language. The DH groups' request to link to articles that belong to a specific topic (needed to better understand the topics) through visualizations was implemented. Also, after several testing sessions, the number of topics was increased to 100, which made the topics more meaningful for DH use.

It can be concluded that the last and improved version of the topic modeling tool in the NewsEye platform does a great job of supporting humanities research, albeit in a different way than topic models would do 'on the fly'. Again, this has to be explained to users. Pre-trained models that represent the entire corpus per language (in 100 topics) help DH researchers get an overview of rather general topics in their datasets, such as 'war' or 'finances', but miss the very specific discourse that is only relevant for this specific dataset. It is also important to note that a topic model trained on the whole dataset includes variance in the types of topics it produces. Some of the topics are of the types mentioned above, but the probabilistic models of LDA and DTM also produce topics that have another logic to them. Some topics consist almost entirely of collections of words that do not correspond to general ideas about what a topic is, but are rather based on other data-specific features. For instance, depending on which lists for stopwords are used, one or even more topics will consist only of stopwords, while other topics will contain words that serve a similar function in sections of newspapers (such as names of towns or prices of goods). They may be seen as topics, but the probabilistic model sometimes brings them together because of other reasons than semantic similarity. Explaining this feature of topic models helps historians to better interpret the topics and use the tool.

The document linking, a feature to find documents that are similar to the given dataset, is closely connected to topic modeling and can be performed in the experiment section of the work space in the Demonstrator. Document linking calculates and compares the topic distribution of the given corpus and articles (of the whole collection) with a similar topic distribution and will be automatically extracted and

presented as clickable links (the amount of links can be chosen before running the task). Depending on the research question and especially on the homogeneity of the dataset, this tool can sometimes be more helpful, sometimes less.

2.4.2 Dynamic Frequency Analysis Tools

Another feature thought of to support the work of the DH group was to allow them to experiment with various frequency analysis functionalities. These tools allow the user to extract facets (newspaper titles), words and bi-grams, to find new, related keywords and to generate time series. While methods to achieve satisfactory search results including simple keyword searches and frequency graphs can help establish a general idea about a research theme, more sophisticated techniques such as bi-gram searches or keyword suggestions (based on word embeddings) can increase search outcomes and accuracy significantly. Frequency analyses are quite 'simple' methods to explore a dataset or research request. However, if not explained and contextualized, results can be misinterpreted. In addition, comparing frequency results with context information, the DH group was able to point to bugs in the data and/or tools that needed to be fixed.

2.5 WP 5 Personal Research Assistant

The Personal Research Assistant (PRA), developed at the University of Helsinki (led by Hannu Toivonen), is thought to function as the user's intelligent and transparent aid in analysis tasks. The PRA integrates the tools developed in WPs 3 and 4 and makes them available to the Demonstrator (cf. next section) in a unified manner, and also offers the user automated ways of using them. The PRA also synthesizes the results in natural language, making them more understandable to the user.

The PRA has an investigative layer (the 'Investigator') to design queries and to analyze their results, a 'Reporter' layer to communicate the findings in natural language and, finally, an 'Explainer', to explain the process and the findings in a transparent manner.

2.5.1 Investigator

In the user interface, the Investigator (see public Deliverable D5.6) is invoked by starting an automated 'Experiment' on a user-defined dataset or saved search. This is done in the Workspace, on the Datasets or Saved Searches page, by clicking on the 'Experiment' link next to the dataset/search. The Investigator then starts several 'tasks' as part of the experiment, applying various analysis tools on the dataset or saved search. In other words, the Investigator tries to discover interesting phenomena in the respective documents, rather than answering a specific question asked by the user. The user interface then illustrates the experiment (the set of tasks) as a workflow diagram. The actual results are available as 'Reports', see below.

Currently there is also a second way to use the tool set. The 'Experiment' section in the Workspace allows the user to create its own experiments by using specific datasets or search and applying one or more tools. The actual results are also termed 'Reports'.

The set of tools available is currently limited, as integration of all tools via the Demonstrator is ongoing during the time of testing. The Investigator is also able to split the dataset in various ways and compare results obtained on the different parts in order to identify interesting subsets, but this functionality is not fully integrated into the Demonstrator. The Investigator has been designed in a way that allows experiments started by Investigator to be continued by a user, and the other way around, but this feature will likely not be implemented during the project.

2.5.2 Reporter

The Reporter part of the PRA (see public Deliverable D5.6) is involved as follows. The page that presents the results of an experiment contains two buttons: 'Report' and 'Explain'. The 'Report' button opens a pop-up window with text produced by the Reporter to describe the key results of the experiment as a whole, i.e. over all its tasks. The current version produces reports in English and Finnish, while the final version will also be able to produce texts in French and German.

A usability issue is that the generated texts are not tailored to the readers' technological or scientific backgrounds: The resulting texts refer at places to methods (e.g. topic modeling) and uses terminology (e.g. TF-IDF) in a way that can be difficult to understand for many users. Another issue is that the texts are not very fluent, and a bulleted list of results could be considered a more readable format. The Reporter can also produce a summary of the textual contents of a user-defined dataset. At the time of testing, the results did not seem too descriptive of the actual contents.

In addition to summarizing a complete experiment using the 'Report' button, the Reporter can also produce a description of a single task, also displaying some of the results in a graphical form (e.g. with a bar or line diagram). This is done by selecting the task in the workflow diagram of the experiment and then clicking on 'View result' on the left. This feature seemed particularly helpful for DH researchers when trying to understand the experiments in detail, or when using parts of the results for further investigation.

2.5.3 Explainer

The Explainer (see public Deliverable D5.8) is a tool that is intended to describe, in natural language, what steps the Investigator carried out in the automated mode, as well as its reasoning for doing those steps. It is supposed to complement the Reporter and its results. An explanation of the steps taken as part of an Investigation can be viewed by clicking the 'Explain' button when viewing an Experiment. The functionality aims at giving the user a better understanding of how the results were reached, and a possibility to continue with further experiments. At the time of testing, the Explainer was only partially integrated with the Investigator, but it was already possible to see what this feature will ultimately present. Again, the DH team retains the Explainer as a valuable help, especially when they retrace the steps of their own searches and the PRA's automated experiments. Knowledge of the research steps and algorithms is important for source and tool critique, which in turn is essential for the work of DH researchers. The Explainer is, at present, only able to produce text in English, but is to support Finnish, German and French by the end of the project.

2.6 WP 7 Demonstration, Dissemination, Outreach and Exploitation

Task T7.1 of WP 7 is consists in the development of the NewsEye platform, thought of as an experimental interface (see public Deliverable D7.8). It gives access to the enriched datasets prepared in WPs 1 to 2, the analysis tools developed in WPs 3 and 4, and the functions made available by WP 5. It allows users to experiment with the tools and to prove the workings of the algorithms by trying out individual research questions or by investigating historical topics. The platform was created by the University of La Rochelle [21] and engineered by Axel Jean-Caurant. The features in the platform come from the needs expressed by the DH group and were improved with each tool testing and feedback session.

The request of the DH group for a personal workspace, a function that creates a sample of the results as well as the possibility to create and manage datasets, was implemented and improved over time. Frequency functionalities, keyword suggestions, as well as faceted access to information were implemented, and some of the tools mentioned above were added. Overall, the DH group can summarize that impressive improvements can be seen in the platform as it is accessible today. The approaches are certainly new, and the interface is more user-driven than many existing digital newspaper interfaces. High potential is seen in the dataset functions and the possibility for users to interact with the datasets (e.g. the possibility to manually change the given borders of automated separated articles), as well as in the experiment section (e.g. choosing which tool should be used with what dataset or query). What also became clear while improving the platform was that there actually exists no good way to meaningfully help users to get the most out of such an interface. Although screen casts, help files, best practices, etc. were added to the platform, it remains a challenge for further projects to find good ways and intuitive solutions for help and self-explanatory systems on all levels of interfaces. The PRA is intended as an investigative and explanatory system that can further such an intuitive approach. Even as it is today, the platform could become a model for interfaces of historical newspapers. It could set an example for how an interface for historical newspapers can support source, tool, and corpus criticism. In order to achieve this, all groups working together in NewsEye agree that although the project has delivered its results, more software engineering than is possible within the limits of the project would need to be done.

The project extension made it possible to re-think user-friendliness and openness. Therefore, a second user interface, the "(NewsEye) exploitation platform" was created. While the NewsEye platform as the platform of the project was kept in its original state (`https://platform.newseye.eu`), the exploitation platform (`https://platform2.newseye.eu`) can be seen as a first step towards a production-ready tool, integrating the most mature features of the NewsEye platform. Some of the complexity has been reduced by offering three clearly separated sections: Search, Datasets, and Experiments. All three sections are kept in a simple but intuitive way. While the Search and Dataset sections remain similar to the ones in the NewsEye platform, the Experiment section is the one that has changed the most. It allows users to construct dataset analysis workflows in a flexible manner, including both pre-processing steps and text analysis algorithms like the creation of bi-grams. At the time of writing, it is a proof of concept with only a couple of tools, but the tools run smoothly and the single steps are easy to follow.

3 Summary

The vision of DH researchers was that at the end of the project, the DH team would be able to write about common themes and topics contained in the newspaper corpora from France, Finland and Austria collected in the NewsEye platform by being given automatically created, overlapping multilingual topics enriched with statistical analysis and large amounts of metadata. All of this should be available for

download and further processing, and be clearly and understandably documented and explained, while the algorithms would hint at events missed by the researchers and give clear answers to the prevailing stance. Even though the NewsEye members from all disciplines have very clear ideas about where they want to go, it has become clear that this vision would not be entirely realized within the time frame of the project. This became especially clear for the PRA, which was intended as the tool bringing all research steps together and explaining them to the user in a meaningful way. To show its potential properly, the preceding steps, methods and tools need to be even further advanced than they are today. Even though we were able to make impressive improvements in ATR as well as AS, all the tools building on somewhat noisy datasets return even noisier, incorrect or inexplicable results the further we advance in the envisaged working pipeline. Therefore, tools that build on noisy outcomes of the preceding analysis will return even noisier results. All these findings, however, have also led to a good understanding of what supports or hinders research on historical newspapers. In particular, the project extension has made it possible to work on a new platform that concentrates on the integration of those tools that have proven as being essential for the work with historical newspapers. This new "(NewsEye) exploitation platform" also takes the demands for more user-friendliness and the reduction of complexity into account.

A key factor in the NewsEye project was the interdisciplinary collaboration of all involved WPs. In our understanding, the collaboration proved most successful where the DH groups' contextualization or background knowledge pointed to faulty data, skewed results or meaningless analysis and where, as a consequence, the colleagues from CS adapted the tools and algorithms accordingly, or tried out other methods that better met the DH groups' expectations. To summarize, interdisciplinary collaboration is what can bring about progress to all the fields involved here. A collective effort was made to come closer to ideal research settings. As it turned out, noteworthy progress was made, the main improvements of which can be read about above. Interdisciplinary collaboration has also led to joint publications where researchers from all groups used digital methods to answer historical research questions. Not all of experiments were so successful that they were publishable, but they have contributed to mutual understanding and respect for the possibilities, but also limitations, of each other's disciplines. While computer scientist colleagues became aware of how important it is for humanities scholars to be able to track results, to create subcorpora tailored for individual research, and to evaluate their specific usage of tools, methods and algorithms critically, the DH group learned to have more realistic expectations regarding the possibilities of automation as it exists today. Digital tools cannot – and definitely should not – replace the meticulous and manual work of researchers, but they can be a valuable support, as could be shown by the developments of the project.

References

[1] Eva Pfanzelter, Sarah Oberbichler, Jani Marjanen, Pierre-Carl Langlais, and Stefan Hechl. "Digital interfaces of historical newspapers: opportunities, restrictions and recommendations". In: *Journal of Data Mining & Digital Humanities* HistoInformatics (Jan. 11, 2021). URL: https://jdmdh.episciences.org/7069/pdf.

[2] Sarah Oberbichler and Eva Pfanzelter. "Tracing Discourses in Digital Newspaper Collections: A Contribution to Digital Hermeneutics while Investigating 'Return Migration' in Historical Press Coverage". In: *Digitised Newspapers – A New Eldorado for Historians? Tools, Methodology, Epistemology, and the Changing Practices of Writing History in the Context of Historical Newspapers Mass Digitization*. 2021.

[3] Mika Hämäläinen and Simon Hengchen. “From the Paft to the Fiiture: a Fully Automatic NMT and Word Embeddings Method for OCR Post-Correction”. In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing*. RANLP 2019. Varna, Bulgaria: INCOMA Ltd., Sept. 2019, pp. 431–436. DOI: 10.26615/978-954-452-056-4_051. URL: https://www.aclweb.org/anthology/R19-1051.

[4] Vinh-Nam Huynh, Ahmed Hamdi, and Antoine Doucet. “When to Use OCR Post-correction for Named Entity Recognition?” In: *22nd International Conference on Asia-Pacific Digital Libraries, ICADL 2020*. Nov. 2020, pp. 33–42. DOI: 10.1007/978-3-030-64452-9_3.

[5] Thi Tuyet Hai Nguyen, Adam Jatowt, Nhu-Van Nguyen, Mickael Coustaty, and Antoine Doucet. “Neural Machine Translation with BERT for Post-OCR Error Detection and Correction”. In: *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*. JCDL ’20. Virtual Event, China: Association for Computing Machinery, 2020, pp. 333–336. ISBN: 9781450375856. DOI: 10.1145/3383583.3398605. URL: https://doi.org/10.1145/3383583.3398605.

[6] Thi-Tuyet-Hai Nguyen, Adam Jatowt, Mickaël Coustaty, Nhu-Van Nguyen, and Antoine Doucet. “Deep Statistical Analysis of OCR Errors for Effective Post-OCR Processing”. In: *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. Champaign, France: IEEE, June 2019, pp. 29–38. DOI: 10.1109/jcdl.2019.00015.

[7] Max Weidemann. *NewsEye/Article-Separation*. Nov. 22, 2019. URL: https://github.com/NewsEye/Article-Separation.

[8] Emanuela Boroş, Ahmed Hamdi, Elvys Linhares Pontes, Luis Adrián Cabrera-Diego, Jose G. Moreno, Nicolas Sidere, and Antoine Doucet. “Alleviating Digitization Errors in Named Entity Recognition for Historical Documents”. In: *Proceedings of the 24th Conference on Computational Natural Language Learning*. CoNLL 2020. Online: Association for Computational Linguistics, Nov. 2020, pp. 431–441. URL: https://www.aclweb.org/anthology/2020.conll-1.35.

[9] Elvys Linhares Pontes, Luis Adrián Cabrera-Diego, José G. Moreno, Emanuela Boros, Elvys Linhares Pontes, Ahmed Hamdi, Nicolas Sidère, Mickaël Coustaty, and Antoine Doucet. “Entity Linking for Historical Documents: Challenges and Solutions”. In: *22nd International Conference on Asia-Pacific Digital Libraries, ICADL 2020*. Vol. 12504. Lecture Notes in Computer Science. Springer, Nov. 2020, pp. 215–231. DOI: 10.1007/978-3-030-64452-9_19.

[10] Nhu Khoa Nguyen, Emanuela Boroş, Gaël Lejeune, and Antoine Doucet. “Impact Analysis of Document Digitization on Event Extraction”. In: *4th Workshop on Natural Language for Artificial Intelligence (NL4AI 2020) co-located with the 19th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2020)*. Vol. 2735. Proceedings of the 4th Workshop on Natural Language for Artificial Intelligence (NL4AI 2020). Virtual, Italy, Nov. 2020, pp. 17–28. URL: https://hal.archives-ouvertes.fr/hal-03026148.

[11] Emanuela Boroş, Elvys Linhares Pontes, Luis Adrián Cabrera-Diego, Ahmed Hamdi, Nicolas Sidere, and Antoine Doucet. “Robust Named Entity Recognition and Linking on Historical Multilingual Documents”. In: Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum. Thessaloniki, Nov. 2, 2020.

[12] Elvys Linhares Pontes, Ahmed Hamdi, Nicolas Sidère, and Antoine Doucet. “Impact of OCR Quality on Named Entity Linking”. In: *International Conference on Asia-Pacific Digital Libraries 2019*. Kuala Lumpur, Malaysia, Nov. 2019. DOI: 10.1007/978-3-030-34058-2_11.

[13] Ahmed Hamdi, Axel Jean-Caurant, Nicolas Sidère, Mickaël Coustaty, and Antoine Doucet. "Assessing and Minimizing the Impact of OCR Quality on Named Entity Recognition". In: *Digital Libraries for Open Knowledge 24th International Conference on Theory and Practice of Digital Libraries, TPDL 2020, Lyon, France, August 25–27, 2020, Proceedings*. Aug. 2020, pp. 87–101. DOI: 10.1007/978-3-030-54956-5_7.

[14] Emanuela Boroş. "Neural Methods for Event Extraction". PhD thesis. Université Paris Saclay, Sept. 27, 2018.

[15] Rachele Sprugnoli. "Event Detection and Classification for the Digital Humanities". PhD thesis. University of Trento, Apr. 24, 2018. 213 pp. URL: http://eprints-phd.biblio.unitn.it/2865/.

[16] Olaf Berg. "Capturing Displaced Persons' Agency by Modelling Their Life Events: A Mixed Method Digital Humanities Approach". In: *Historical Social Research / Historische Sozialforschung* 45.4 (2020), pp. 263–289. URL: http://www.jstor.org/stable/26956101.

[17] Thi Tuyet Hai Nguyen. *NewsEye/Stance-Detection*. Apr. 23, 2020. URL: https://github.com/NewsEye/Stance-Detection.

[18] Jani Marjanen. *What's the frequency, Kenneth?* NewsEye Blog. June 25, 2019. URL: https://www.newseye.eu/blog/news/what-s-the-frequency-kenneth/.

[19] Elaine Zosa, Simon Hengchen, Jani Marjanen, Lidia Pivovarova, and Mikko Tolonen. "Disappearing Discourses: Avoiding Anachronisms and Teleology with Data-Driven Methods in Studying Digital Newspaper Collections". In: Digital Humanities in the Nordic Countries DHN 2020. Riga, 2020. URL: https://researchportal.helsinki.fi/en/publications/disappearing-discourses-avoiding-anachronisms-and-teleology-with-.

[20] Elaine Zosa and Mark Granroth-Wilding. "Multilingual Dynamic Topic Model". In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*. RANLP 2019. Varna, Bulgaria: INCOMA Ltd., Sept. 2019, pp. 1388–1396. DOI: 10.26615/978-954-452-056-4_159. URL: https://www.aclweb.org/anthology/R19-1159.

[21] Axel Jean-Caurant and Antoine Doucet. "Accessing and Investigating Large Collections of Historical Newspapers with the NewsEye Platform". In: *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*. JCDL '20. New York, NY, USA: Association for Computing Machinery, Aug. 1, 2020, pp. 531–532. ISBN: 978-1-4503-7585-6. DOI: 10.1145/3383583.3398627. URL: https://doi.org/10.1145/3383583.3398627.