Project Number: **770299**

NewsEye:
A Digital Investigator for Historical Newspapers

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

D1.10: Data collection and preservation (d) (final)

Due date of deliverable: M36 (30 April 2021)

Actual submission date: 15 April 2021

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

Partner organization name in charge of deliverable: UIBK-DEA

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:	D1.10	
Deliverable full title:	Data collection and preservation (d) (final)	
Deliverable short title:	Data collection and preservation	
Document identifier:	NewsEye-T12-D110-DataCollectionAndPreservation-d-Submitted-v3.0	
Lead partner short name:	UIBK-DEA	
Report version:	V3.0	
Report preparation date:	15.04.2021	
Dissemination level:	PU	
Nature:	Report	
Lead author:	Günter Mühlberger (UIBK-DEA)	
Co-authors:	Juha Rautiainen (UH-NLF), Jean-Philippe Moreux (BnF), Axel Jean-Caurant (ULR), Antoine Doucet (ULR), Max Kaiser (ONB), Florian Krull (UIBK-DEA), Günter Hackl (UIBK-IUI)	
Internal reviewers:	Johannes Michael (UROS), Sebastian Cretin (BnF)	
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
16.03.2021	1.0	Günter Mühlberger (UIBK-DEA), Günter Hackl (UIBK-IUI)	Draft version
06.04.2021	2.0	Günter Hackl (UIBK-IUI)	Internal reviews taken into account
15.04.2021	3.0	Günter Hackl (UIBK-IUI) and Antoine Doucet (ULR)	QM review integrated and final adjustments

Executive summary

This public deliverable introduces the cumulative work performed during 3 years in Task T1.2 on 'Data collection and preservation'. It has two main objectives:

- to manage the selection and collection process (further described in Sections 1 and 2) concerning the availability of data for further usage within NewsEye and
- to preserve and sustain this data not only for the duration of the project, but also to set the scene beyond the project period.

In period one (M1-M12) we focused on the availability of data, their properties and the ways they will be delivered to the NewsEye data repository. And in the second half of period one the actual collection process started. The main focus there was to organize the gathering of data for setting up the NewsEye Digital Library Demonstrator (NDLD).

In period two (M13-M24) we had to adjust the data preservation approach following the fact that the implementation of the NDLD has changed and no longer includes Samvera. Initially, preservation was to be handled through it. Now Zenodo is used for long-term data storage. The new approach is described in Section 3. Section 2.4 shows which tool we used for creating training data for semantic text enrichment and why we decided to use it.

In period three (M25-M36) we collected data for an English title of the Bibliothèque nationale de France (BnF) - 'The New York Herald'. This was then processed as usual with LA - layout analysis (WP2), ATR - automated text recognition (WP2), AS - article separation (and other structural elements) (WP2), NER - named entity recognition (WP3), NEL - named entity linking (WP3), SD - stance detection (WP3) and ED - event detection (WP3). The output is used as 'show case' to also address people who are not familiar with the four languages of the project. Moreover the preservation of the datasets and models was done in this last project year. The datasets were uploaded to Zenodo and the data models persist in the Transkribus platform.

Section 4 summarizes the task achievements in the three years of the project.

Note that this deliverable D1.10 'Data collection and preservation (d) (final)' is closely connected to the other public deliverable of WP1: D1.9 'Data models (d) (final)', as well as to the non-public deliverable D1.11 'Data generation (c) (final)'.

Contents

1 Data selection process

The NewsEye project follows a typical digitization workflow: Starting with simple image files and some general meta data these images are processed with automated text recognition (ATR), layout analysis (LA), article separation (AS) and named entity recognition (NER). Furthermore, named entities are linked with external sources - named entity linking (NEL) - and enriched with properties such as 'stance'. Finally, these data are used to set up powerful ways to access the collection not only by searching, but in an interactive and adaptive way taking into account the full wealth of the data. The simple image, together with some meta data is therefore the starting point of the whole workflow. In our case also text data from former optical character recognition (OCR) campaigns carried out by the participating libraries were available as a starting point and as a reference for comparing the results achieved in the project with the current state of the art. This section describes in more detail what data were available for the NewsEye project in the participating libraries and how we tackled the collection process itself.

1.1 Collections and datasets

First of all we have to define two important terms in the project, namely 'collections' and 'datasets'. We define a collection as the complete amount of digitized newspapers available in a specific library or library unit. In contrast, a dataset is a meaningful selection of newspaper pages and/or issues from a library collection. Datasets are foreseen to carry out an experiment or support specific investigations and research. E.g. all newspapers of the Austrian National Library form the ANNO *collection*, whereas all issues from the newspaper 'Neue Freie Presse' between 1850 and 1880 could serve as a *dataset* for a specific research topic.

It is not the objective of the NewsEye project to process all or even large amounts of the collections of digitized newspapers in the participating libraries, but an amount of pages which enables us:

- to support 'Digital Humanities' (DH) research in WP6
- to prove the quality of the developed tools
- to showcase the attractiveness of the final demonstrator to our target groups (libraries and their users, DH researchers, the public in general) and finally
- to demonstrate the scalability of the developed solutions in terms of quantity and quality. An important point in this respect is that wherever possible language independent methods are applied.

1.2 Decision making

The decision making process was guided by the above criteria. First of all, at the Steering Committee meeting of M6 in London a general decision was taken that around 500 000 pages per partner library would be selected as datasets and processed by the UIBK-DEA

team with ATR, LA and AS. These 500 000 pages per partner library allow DH researchers to use the NewsEye demonstrator with a meaningful amount and variety of data.

1.3 Collections by libraries

The following subsections list the available collections of each library with their quantity of pages.

1.3.1 BnF - Available collections

The BnF collection is twofold: 1.2M pages of OCRed content and 1.17M pages with article separation and (some) content classification.

Note: Other collections from the Gallica newspapers repository can be extracted if needed. A subset of this collection has article separation. In such case, the BnF will need the newspapers titles and time periods requested by the DH researchers.

- **OCR corpus**
  Presentation: https://api.bnf.fr/fr/documents-de-presse-numerises-en-mode-ocr-du-projet-europeana-newspapers [1]
  This collection includes French dailies from 19th and 20th centuries.
  Formats: METS (Metadata Encoding and Transmission Standard) and ALTO (Analyzed Layout and Text Object)
  Newspapers titles: Le Figaro, L'Echo de Paris, L'Univers, La Presse, L'Humanité, Le Constitutionnel, Le Petit Journal, Le Siècle, L'Action Française, L'Intransigeant, Le Temps, La Croix.
  Coverage: 1850-1944.
  Volume available: 1 287 500 pages / 275 000 issues.
  Language: French.

- **Optical layout recognition (OLR) corpus**
  Presentation: https://api.bnf.fr/fr/documents-de-presse-numerises-en-mode-article-du-projet-europeana-newspapers [2]
  This collection includes some of the French dailies from 19th and 20th centuries which have been digitized with article recognition (detailed description of the content of each issue: article, section, headings, captions... as well as the identification of advertisements and tables). The formats are METS (including the OLR logical structure) and ALTO.
  Newspapers titles: Le Gaulois, Le Journal des débats politiques et littéraires, Le Matin, Ouest Eclair (éditions de Caen, Rennes), Le Petit Journal illustré supplément du dimanche, Le Petit Parisien.
  Coverage: 1814-1944.
  Volume available: 790 000 pages / 147 000 issues.
  Language: French.

1.3.2 ONB - Available collections

The Austrian National Library (ONB) offers a collection of 14.5M pages for newspapers between 1845-1945, whereof 13.8M are OCRed.

OCR corpus:

- METS/ALTO (files produced with ABBYY FineReader) 10.2 million pages / 842 000 issues.
- METS/H-OCR (files from Austrian Books Online / Cooperation with Google) 1.425 million pages / 175 000 issues.
- ALTO (files produced in-house with Tesseract) 1.3 million pages / 90 000 issues.
- Images (still) without OCR (the number shrinks continuously) 0.347 million pages / 50 000 issues.
- Coverage: 1845-1945.
- Newspaper titles: Wiener Zeitung, Neue Freie Presse, Das Vaterland, Die Presse, Innsbrucker Nachrichten, Neues Fremdenblatt, Linzer Volksblatt, Wiener Sonn- und Montags-Zeitung, Wiener Salonblatt, Prager Tagblatt, Pester Lloyd, Reichspost, Neuigkeits-Welt-Blatt, Arbeiter Zeitung, Kronen-Zeitung and others.

1.3.3 NLF - Available collections

The National Library of Finland (NLF) is able to provide 2.7M pages of OCRed newspaper content from 1771 to 1910/20. A collection of one title Uusi Suometar published in 1869-1918 has improved OCR.

1.4 Selection of datasets

After the SC meeting of M6, detailed decisions were taken by the DH groups as to which newspaper titles (language, period, etc.) should belong to the fully processed datasets. The size of these specific datasets depended on several factors such as the research questions to be answered, the tools necessary to process the datasets, the improvements which can likely be made in comparison of what is already available (e.g. quality of the raw text), the readiness of new tools but also along the criteria of 'different and varying layout structures', which require dedicated training data.

The description of the chosen datasets as well as the ongoing data generation process can be found in Deliverable D1.11: Data generation (b). Since this is not a public deliverable the chosen datasets per library are summarized here as well; The datasets from the BnF, NLF, and ONB collections have the following properties: time span 1850-1950, four to seven daily newspaper titles per language with about 0.5 million pages for each library.

1.4.1 Dataset for ONB

The DH-groups in Austria (Innsbruck and Vienna) have selected four newspapers with different time-lines from the ONB collection, namely:

- *Neue freie Presse* with timelines '1864-1873', '1895-1900', '1911-1922' and '1933-1939', text type Kurrent, 3 columns
- *Illustrierte Kronen Zeitung* with timelines '1911-1922' and '1933-1939', changing column layout
- *Innsbrucker Nachrichten* (Mittags-Zeitung) with timelines '1864-1873', '1895-1900', '1911-1922' and '1933-1939', text type Kurrent, 1 column at the beginning up to 4 columns
- *Arbeiter-Zeitung* with timelines '1895-1900', '1911-1922' and '1933-1939', text type Kurrent and Latin, 2 to 4 columns

The chosen newspapers differ in their political orientation and publication period. The 'Neue Freie Presse' and the 'Illustrierte Kronen Zeitung' are supraregional newspapers. While the 'Neue Freie Presse' was the leading newspaper of the Habsburg Monarchy and one of the highest-circulation newspapers before WWI, the 'Illustrierte Kronen Zeitung' is a tabloid newspaper that did not appear before the 20th century. The 'Innsbrucker Nachrichten' was selected as regional, independent, democratic daily newspaper for Tyrol and Vorarlberg which was the NSDAP Gau-Tirol party organ between 1938 and 1945. On the other hand, the 'Arbeiter-Zeitung' was an organ of the Austrian Social Democracy in the interwar period.

1.4.2 Dataset for NLF

The Digital Humanities team of the University of Helsinki (UH-DH) and NLF selected primarily papers that 1) have been printed for a long period and thus constituted long time series, 2) were deemed as important in public discourse from the second half of the 19th century onward and 3) represent different regions. The following newspaper titles are listed separately by language:

- Finnish dataset
  - 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
- Swedish dataset

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

1.4.3 Dataset for BNF

The dataset selected from the BNF collections, including the web links, can be found in Gallica.

- La Presse : from 1850 to 1890 (https://gallica.BnF.fr/ark:/12148/cb34448033b/date)
- Le Matin : from 1884 to 1944 (https://gallica.BnF.fr/ark:/12148/cb328123058/date)
- La Fronde : from 1897 to 1929 (https://gallica.bnf.fr/ark:/12148/cb327788531/date)
- Marie-Claire : from 1937 to 1944 (https://gallica.bnf.fr/ark:/12148/cb343488519/date)
- L'OEuvre : from 1915 to 1944 (https://gallica.bnf.fr/ark:/12148/cb34429265b/date)
- Le Gaulois : from 1868 to 1900 (http://gallica.bnf.fr/ark:/12148/cb32779904b/date)

The project team of the University Paul-Valéry Montpellier (UPVM) and BnF have selected the above newspapers because of their importance for the French press. The newspapers differ in their political orientation, publication period and readership.

Launched in 1836, *La Presse* was one of the first major popular newspapers. It opposed political parties' newspapers intended for low readership. *Le Gaulois* was also one of the biggest sales success of its time. Monarchistic at its beginnings, Bonapartistic and republican thereafter, the newspaper took a conservative and legitimistic turn when it was redeemed in 1879. Despite its social aspects, it had a certain political power, being read by the nobility and the bourgeoisie.

Le Matin was also a great success with a circulation of more than one million copies around 1914. Favorable to moderate Republicans, it was opposed to the Boulangist movement and socialist ideas. After the war, it became a conservative title. As for *L'Œuvre*, which became a daily newspaper in 1915, it was originally committed to the left-wing party before the German occupation. The researchers chose this newspaper because it employed many female journalists, which is a quite rare phenomenon for the time. In fact, as UPVM works on the topic 'gender', *La Fronde* (1897-1929), a feminist daily newspaper of the late 19th century in which many women are involved in the intellectual, literary and political world was selected as well. The woman's weekly magazine *Marie-Claire* (from 1937 to 1944) also seemed very interesting to study editorial poetics and media issues.

2 Collecting data

This section shows the possibilities of collecting data from the libraries with the focus on data formats and data access, hosting and exchanging data in the NDLD and enriching data with the tools of the NewsEye project.

2.1 Data flow

The NDLD contains all data produced in the project and make them available in different ways and via different channels to several user groups.

Figure 1 details the data flow in the project and illustrates how data are collected, exchanged and preserved within NewsEye.

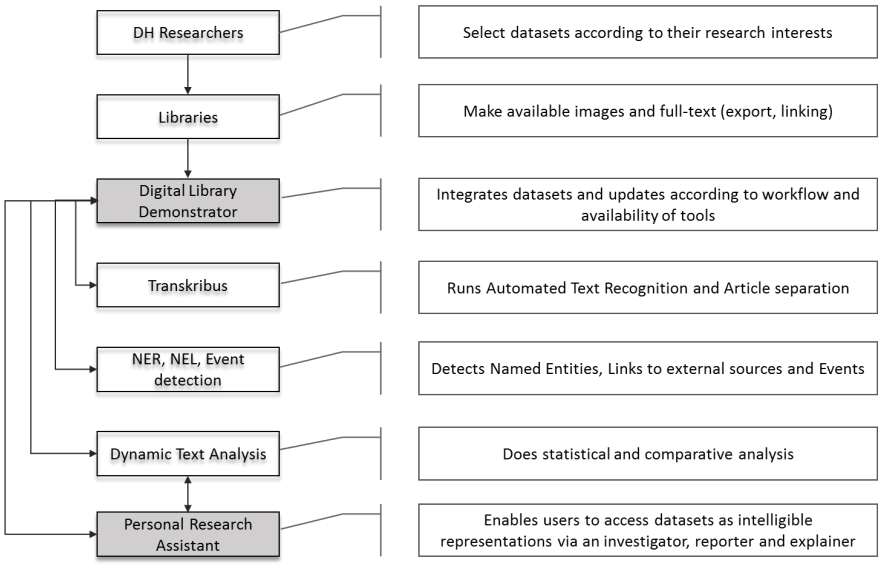

Figure 1: Data flow in NewsEye

2.1.1 BnF

The BnF newspaper collections for NewsEye may have slightly different formats (all variants of ALTO and METS XML formats):

- BnF collection: ALTO LoC + METS (document manifest) + CCS METS (for the OLR logical structure). A JSON export for the OCRed text is also available: https://api.bnf.fr/fr/texte-des-documents-de-presse-du-projet-europeana-newspapers-xixe-xxe-siecles [3]
- BnF OCR collection (on request): ALTO BnF (variation of ALTO LoC) + BnF METS (document manifest)
- BnF OLR collection (on request): ALTO BnF (variation of ALTO LoC) + BnF METS (document manifest) + BnF METS (for the OLR logical structure)

For all these documents, access to the scanned images can be performed thanks to IIIF Image API. Other Gallica APIs (see api.bnf.fr) may also be used to extract (if needed) information that is not reported in the documents manifest (like newspapers calendar, see http://api.bnf.fr/api-document-de-gallica for a description of the Issues API).

2.1.2 NLF

OpenURL links to the page image based on the data information on the URL parameters. The available parameters can be seen from the example below:

- genre (journal, no need to change)
- date (YYYY-MM-DD)
- ISSN (the identifier of the newspaper or journal)
- spage (page number)

http://digi.kansalliskirjasto.fi/openurl/query.html?genre=journal&date=1888-01-03&issn=0355-6913&spage=2

Also a list of all newspapers or journals can be retrieved in JSON:

- https://digi.kansalliskirjasto.fi/api/newspaper/titles?language=fi
- https://digi.kansalliskirjasto.fi/api/journal/titles?language=fi

OAI-PMH returns the basic binding level metadata. Offered formats are listed on http://digi.kansalliskirjasto.fi/interfaces/OAI-PMH?verb=ListMetadataFormats [4]

2.1.3 ONB

ONB is working on a IIIF interface for its digital collections. The current preferred output format are METS/ALTO packages which are formed according to the EU Newspaper project: ENMAP format as defined in the EU Newspaper project.

2.2 NewsEye Digital Library Demonstrator – NDLD (ULR)

The main demonstrator in NewsEye serves as the central data host for all data in the project. Datasets are provided by the libraries, enriched by Transkribus and several other tools, analyzed with dynamic text analysis methods and made available to users via the NDLD and the Personal Research Assistant as a distinct part of the final demonstrator.

The internal data representation within the NDLD is based on Solr and the images are fetched directly via IIIF. The suggestion how to represent data of historical newspapers was discussed and agreed at the NewsEye consortium meeting in London and is also described in Deliverable D1.9 Data models.

2.3 Transkribus (UIBK-DEA, UROS)

Transkribus is a research and service platform which was mainly developed in the H2020 Project READ[1]. It provides state-of-the-art tools for LA and ATR developed by UROS. Within NewsEye the Transkribus platform was enhanced and augmented with new features and tools for processing historical newspapers and here especially article separation (AS). According to the progress made by UROS, AS will be integrated into Transkribus as well.

Transkribus is designed to enable users such as scholars, librarians or archivists to carry out all steps of the digitization, text recognition and text augmentation workflow on their own. The success of this approach can be seen by the following figures:

- Over 53 000 users registered in the platform.
- There are around 2 700 'active' users per month (unique logins).
- Users uploaded 20 million images to the platform so far.
- About 300 neural nets ('models') were trained by users themselves per month based on thousands of ground truth pages.

According to this user-centred approach the focus of READ was therefore to develop a rich expert client (JAVA-SWT) which provides all the features which are necessary

- to upload documents to the platform
- to create training data for text recognition
- to train neural networks (ATR/HTR engine) and produce models on 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.

In terms of data exchange we have to distinguish two main use cases:

- Training data
  Since all three tools from UROS (LA, ATR and AS) are at least partly based on supervised machine learning, training data (ground truth - GT) needs to be generated before the recognition process. This can be done directly in Transkribus.
- Processing
  Once training data and tools are available, the selected datasets can be processed. This can be done as a batch process on the Transkribus platform or by utilizing the Transkribus API. The API is already available but needs to be adapted to the specific

[1] https://read.transkribus.eu/

needs within the project. Uploading issues into Transkribus via IIIF was already implemented in Y2. What is missing is a job workflow that allows to select which jobs should be executed one after the other. There is one special use case for this: The researcher selects one newspaper issue in the NDLD and wants to have all possible results which were implemented during the NewsEye project. From Transkribus she will get LA, ATR and AS. The newspaper issue will be uploaded to Transkribus and one job after the other is executed and finally the enriched PAGE XML files will be delivered back.

Once the datasets are processed in Transkribus they are made available to the NDLD via the API. Several exchange formats are available for this purpose such as XML (PAGE) or JSON. It was part of the work in Y2 to adapt the interface of Transkribus in a way that data can be directly exchanged with the NDLD.

2.4 Named entity recognition/linking and event detection (ULR)

Based on the raw full-text which comes either initially from the ingested files of the libraries or represents an improved full-text coming from Transkribus, more data layers can be added to the documents. Named entity recognition and linking as well as event detection are the main processes here and are carried out in WP3 (semantic text enrichment). Since ULR is responsible both for the NDLD as well as this task we assume that the implementation of the interface between the two systems is straightforward.

2.4.1 Training data for semantic text enrichment

The interface for creating training data for NER and NEL, stance and event detection is distinct. It can for instance be Transkribus or the semantic annotation platform 'Inception'[2]. Both enable users to define their own named entities, or to correct those which were supplied by the NER engine. For the decision which tool should be used all the pros and cons had been discussed at the beginning of Y2 and the decision was made in favor of Transkribus.

2.4.2 Transkribus vs. Inception

The decision which tool to use for creating named entity GT was made in the first half of Y2. The main reason for using Transkribus and not Inception is clearly that the expertise in the Newseye project team - more precisely the UIBK-DEA team in Innsbruck - is much higher for Transkribus. And as UIBK-DEA is responsible for managing the task of creating NER, NEL and stance GT it was obvious to use and extend Transkribus for the different subtasks like user guidance, tutorials, development and data exchange. Moreover the already used and newly required data formats are well known and easier to integrate as with a foreign application. Since the final format of both tools is ConLL and IOB, collecting, exchanging and using training data from partner projects like *impresso*[3] is easy.

[2] https://inception-project.github.io/
[3] https://impresso-project.ch/

2.5 Dynamic text analysis (UH)

As soon as full-text was available in the NDLD it enabled a first processing by dynamic text analysis tools. During the course of the project the datasets within the NDLD were updated according to the progress made. E.g. text recognition was improved, articles were better separated, and named entities added and linked to external sources.

2.6 Personal research assistant (UH)

The personal research assistant sits on top of the data pyramid and takes benefit from all annotations provided by the several tools. It is be able to access single articles with accurate full-text and tagged persons and places, in some cases even with a link to an external database.

3 Preservation

According to the NewsEye work plan, it was foreseen to set up a Git repository in order to manage the collection and preservation of all data in the NewsEye project. Then, at the kick-off meeting in La Rochelle, ULR proposed to use the Samvera repository software to cope with this task (see Deliverable D3.1). Therefore the use of a Git repository became obsolete and it was planned to fully replace it by the NDLD. But during the implementation of the NDLD it became clear that this plan was not feasible with the given resources and time. Moreover, using Samvera has some overhead which can be omitted with starting more or less from scratch. This means using some basic parts of the Samvera architecture and skip others. But this also implies that the preservation feature could not be used as initially thought. Therefore the new strategy for the preservation task was changed into uploading all generated project data to Zenodo[4]. Here are some important notes from the Zenodo policies:

- Retention period: Items are retained for the lifetime of the repository. This is currently the lifetime of the host laboratory CERN, which currently has an experimental programme defined for the next 20 years at least.
- File preservation: Data files and metadata are backed up nightly and replicated into multiple copies in the online system.
- Fixity and authenticity: All data files are stored along with a MD5 checksum of the file content. Files are regularly checked against their checksums to assure that file content remains constant.
- Succession plans: In case of closure of the repository, best efforts will be made to integrate all content into suitable alternative institutional and/or subject based repositories.

We have clearly stated in the 'Description of Action' that the focus of NewsEye is not on preservation but on exploring new methods and tools to improve access to digitized

[4] https://zenodo.org/

newspapers. However with the choice to upload all generated data we believe that this aspect is from a technological view now covered in the best way. Although Fedora has the preservation feature included, running the servers with the developed prototypes (like NDLD with all the interfaces) is in most cases not possible after the project ends. But an online storage service with a strong focus on durability is much more reliable in this regard.

4 Achievements

The fundamental work for this deliverable was already done in the first year. In the second year we had to adapt the preservation strategy and proof the data flow while collecting all the required project data.

4.1 Period M1-M12

All tasks foreseen in T1.2 'Data collection and preservation' have been considered in Y1:

- we overviewed the available data from the participating libraries;
- we organized the selection process among DH groups and libraries in order to get meaningful data for testing the NewsEye tools and for presenting the results to the public via the NDLD and
- we clarified the preservation aspects of the NewsEye project

And in Y1 we were in contact with representatives from the *impresso* project and trained several ATR models from the Neue Zürcher Zeitung (NZZ). Such cooperation can be intensified also as part of WP7 Demonstration, Dissemination, Outreach and Exploitation.

4.2 Period M13-M24

In Y2 we were mainly engaged with generating training data for WP3 (semantic text enrichment), collecting and preparing the 1.5 million pages of the partner libraries, and finally running the layout and text recognition on these pages. During this process we also verified the designed data flow of this deliverable. Although minor adjustments were necessary during the practical work, the invented data flow has not been substantially amended.

In addition, we had to adjust the preservation strategy because the implementation of the NDLD has changed. Therefore - as explained in Section 3 - we now upload all generated data to Zenodo to fulfill the conservation goals of the project.

The designed data flow can be used internally in the project but also offered to associated partners and moreover also to external groups which are not project partners, increasing the project impact and dissemination. Opening the NewsEye workflow for these projects undoubtedly has an impact in terms of exploitation for WP7.

4.3 Period M25-M36

As already mentioned in the previous section we planned to offer project partners and external institutions the developed data flow, data models, tools and workflows as services in small pilot projects. This will naturally have positive impact on the exploitation of the project. But due to the Covid-19 crisis, with limited opportunities to learn about, meet and approach other projects, these plans were deferred to the project extension phase.

On the other hand the results of the NewsEye project already had a positive impact for the READ-COOP. The public model of the german newspapers from the National Library of Austria (ONB) is very well suited as a general model for other newspapers as well. For example READ-COOP won the public tender to recognise all the newspapers in https://zeitpunkt.nrw/ including the DFG-sponsored recognition of the 'Kölnische Zeitung' project. Also two projects with the Vienna City Library [5] could be realized successfully. Last but not least the re-OCRing of 2 million pages for the project partner NLF was started in the year 2020 using their own NewsEye models.

Via the Transkribus REST API, it is now possible to upload a newspaper issue via IIIF (this was already implemented in Y2) and start LA, ATR and deliver the output via the REST interface. This way the NewsEye demonstrator can directly communicate with the Transkribus API without manual intervention. The same data flow can be arranged for further tasks like named entity recognition, event detection and so on. Moreover it is planned to integrate the AS tool of UROS as well. This will happen when the research is finished and a stable version of the tool is available.

'The New York Herald' was collected and processed successfully in the third year of the project and is now available in the NDLD as a show case for a wider international audience.

For the preservation task we uploaded the datasets to Zenodo. While the datasets can be downloaded by external research groups to use them for their own training algorithms, the trained models themselves are sustained in the Transkribus platform. Those are available as public models for all users of Transkribus. At the moment, over 53 000 Transkribus users can benefit from this NewsEye development. The models and their properties are described in more details on the Transkribus website https://readcoop.eu/transkribus/public-models/. The ATR datasets on Zenodo for all libraries and the four project languages can be found following the 'references' below:

- NewsEye / READ OCR training dataset from Austrian Newspapers (19th C.) [5]
- NewsEye / READ OCR training dataset from French Newspapers (18th, 19th, early 20th C.) [6]
- NewsEye / READ OCR training dataset from Finnish Newspapers (18th, 19th, early 20th C.) [7]
- NewsEye / READ OCR training dataset from Swedish Newspapers (18th, 19th, early 20th C.) [8]

And the AS datasets can be found here:

[5] https://www.wienbibliothek.at/english

- NewsEye / READ AS training dataset from French Newspapers (19th, early 20th C.) [9]
- NewsEye / READ AS training dataset from Finnish Newspapers (19th C.) [10]
- NewsEye / READ AS training dataset from Austrian Newspapers (19th, early 20th C.) [11]

For the NER, NEL and stance GT data, a resource paper introducing the dataset for all languages was accepted for publication in the ACM proceedings of the SIGIR 2021 conference[6], [12], effectively increasing the visibility of the dataset uploaded on Zenodo [13].

References

[1] BnF. *Digital documents of the newspapers collections processed during the Europeana Newspapers project with text recognition (OCR, optical character recognition).* URL: https://api.bnf.fr/fr/documents-de-presse-numerises-en-mode-ocr-du-projet-europeana-newspapers.

[2] BnF. *Digital documents of the press collections processed during the Europeana Newspapers project with OLR (optical layout recognition).* URL: https://api.bnf.fr/fr/documents-de-presse-numerises-en-mode-article-du-projet-europeana-newspapers.

[3] BnF. *Text of the newspapers collections processed during the Europeana Newspapers project.* URL: https://api.bnf.fr/fr/texte-des-documents-de-presse-du-projet-europeana-newspapers-xixe-xxe-siecles.

[4] NLF. *Basic binding level metadata.* URL: https://digi.kansalliskirjasto.fi/interfaces/OAI-PMH?verb=ListMetadataFormats.

[5] Guenter Muehlberger and Guenter Hackl. *NewsEye / READ OCR training dataset from Austrian Newspapers (19th C.)* Zenodo, Sept. 2019. DOI: 10.5281/zenodo.3387369. URL: https://doi.org/10.5281/zenodo.3387369.

[6] Guenter Muehlberger and Guenter Hackl. *NewsEye / READ OCR training dataset from French Newspapers (18th, 19th, early 20th C.)* Zenodo, Nov. 2020. DOI: 10.5281/zenodo.4293602. URL: https://doi.org/10.5281/zenodo.4293602.

[7] Guenter Muehlberger and Guenter Hackl. *NewsEye / READ OCR training dataset from Finnish Newspapers (18th, 19th, early 20th C.)* Version 1. Zenodo, Mar. 2021. DOI: 10.5281/zenodo.4599472. URL: https://doi.org/10.5281/zenodo.4599472.

[8] Guenter Muehlberger and Guenter Hackl. *NewsEye / READ OCR training dataset from Swedish Newspapers (18th, 19th, early 20th C.)* Version 1. Zenodo, Mar. 2021. DOI: 10.5281/zenodo.4599624. URL: https://doi.org/10.5281/zenodo.4599624.

[9] Guenter Muehlberger and Guenter Hackl. *NewsEye / READ AS training dataset from French Newspapers (19th, early 20th C.)* Version 1. Zenodo, Mar. 2021. DOI: 10.5281/zenodo.4600636. URL: https://doi.org/10.5281/zenodo.4600636.

[6] https://sigir.org/sigir2021/call-for-resource-papers/

[10] Guenter Muehlberger and Guenter Hackl. *NewsEye / READ AS training dataset from Finnish Newspapers (19th C.)* Version 1. Zenodo, Mar. 2021. DOI: 10.5281/zenodo.4600746. URL: https://doi.org/10.5281/zenodo.4600746.

[11] Guenter Muehlberger and Guenter Hackl. *NewsEye / READ AS training dataset from Austrian Newspapers (19th, early 20th C.)* Version 1. Zenodo, Apr. 2021. DOI: 10.5281/zenodo.4693413. URL: https://doi.org/10.5281/zenodo.4693413.

[12] Ahmed Hamdi, Elvys Linhares Pontes, Emanuela Boros, Thi Tuyet Haï Nguyen, Guenter Hackl, Jose G. Moreno, and Antoine Doucet. "A Multilingual Dataset for Named Entity Recognition, Entity Linking and Stance Detection in Historical Newspapers". In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '21. Virtual Event: Association for Computing Machinery, 2021, pp. 1–7.

[13] Ahmed Hamdi, Elvys Linhares Pontes, Emanuela Boros, Thi Tuyet Hai Nguyen, Guenter Hackl, Jose G. Moreno, and Antoine Doucet. *Multilingual Dataset for Named Entity Recognition, Entity Linking and Stance Detection in Historical Newspapers*. Version V1.0. Zenodo, Mar. 2021. DOI: 10.5281/zenodo.4573313. URL: https://doi.org/10.5281/zenodo.4573313.