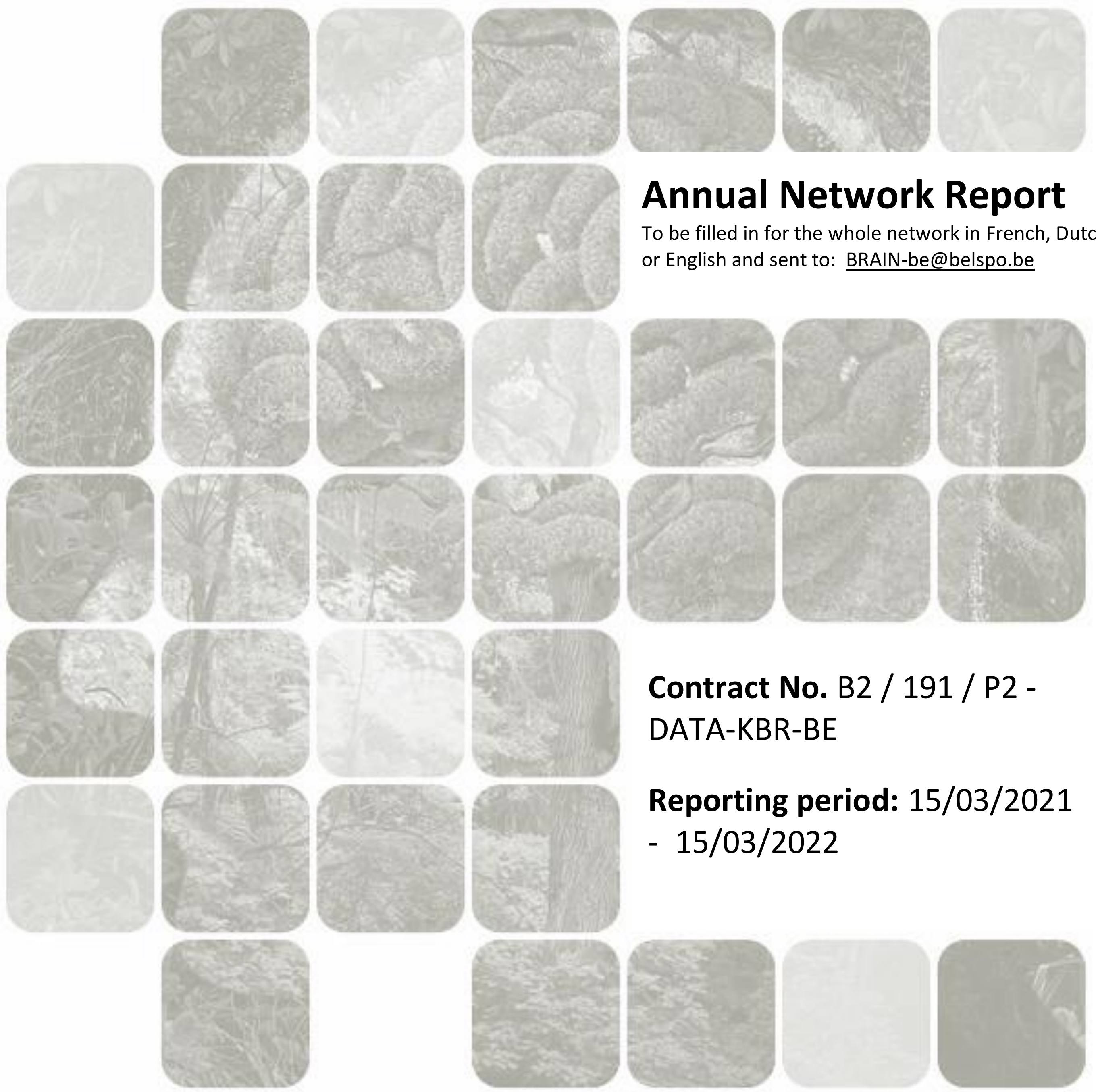
BRAIN-be 2.0

BELGIAN RESEARCH ACTION THROUGH INTERDISCIPLINARY NETWORKS - Phase 2



Annual Network Report

To be filled in for the whole network in French, Dutch or English and sent to: BRAIN-be@belspo.be

Contract No. B2 / 191 / P2 - DATA-KBR-BE

Reporting period: 15/03/2021 - 15/03/2022

The *Annual Network Report* (maximum 15 to 20 pages) is drawn up annually by the coordinator for the entire network and sent to the address BRAIN-be@belspo.be on the dates set in article 7.6 of annex I to the contract. It presents the state of progress and achievements of the research as well as forecasts for the following year. This information refers explicitly to the tasks and the project schedule defined in articles 2 and 3 of annex I. It also informs of any modification of the data included in the initial reports and gives the list of publications and missions carried out during the past year.

This template can be completed in French, Dutch or English.

NETWORK

COORDINATOR (PARTNER 1)

1. Frédéric Lemmers: KBR, Royal Library of Belgium

OTHER PARTNERS

2. Prof. dr. Christophe Verbruggen: Ghent Centre for Digital Humanities, Ghent University

3. Prof. dr. Steven Verstockt: Internet Technology and Data Science Lab, Ghent University

4. Prof. dr. Prof. Dirk Van Hulle: Antwerp Centre for Digital Humanities and Literary Criticism (ACDC), University of Antwerp

AUTHORS OF THIS REPORT

1. Sally Chambers: KBR, Royal Library of Belgium

2. Frédéric Lemmers: KBR, Royal Library of Belgium

3. Thuy-An Pham: KBR, Royal Library of Belgium

4. Dilawar Ali: Internet Technology and Data Science Lab (IDLab), Ghent University

5. Kenzo Milleville: Internet Technology and Data Science Lab (IDLab), Ghent University

6. Steven Verstockt: Internet Technology and Data Science Lab (IDLab), Ghent University

7. Wout Dillen (until 1st June 2021), Lamyk Bekius (from 1st June 2021): Antwerp Centre for Digital Humanities and Literary Criticism (ACDC), University of Antwerp

8. Julie M. Birkholz, KBR, Royal Library of Belgium and Ghent Centre for Digital Humanities, Ghent University

9. Antoine Jacquet (until September 2021), Brecht Deseure (from February 2022), KBR, Royal Library of Belgium and Université libre de Bruxelles, Sciences de l'information et de la communication Department

10. Tan Lu, KBR, Royal Library of Belgium and DIMA: Digital Mathematics research group, Vrije Universiteit Brussel (VUB).

11. Pieterjan De Potter: Ghent Centre for Digital Humanities, Ghent University

12. Vincent Ducatteeuw: Ghent Centre for Digital Humanities, Ghent University

PROJECT WEBSITE, SOCIAL NETWORKS …

- DATA-KBR-BE on the KBR website: https://www.kbr.be/en/projects/data-kbr-be
- Twitter: @kbrbe

TABLE OF CONTENTS

1. EXECUTIVE SUMMARY OF THIS REPORT

The DATA-KBR-BE: facilitating data-level access to KBR's Collections for Open Science project is financed by the Belgian Science Policy Office (Belspo) as part of the Belgian Research Action through Interdisciplinary Networks, BRAIN 2.0 programme. It is an interdisciplinary collaboration, led by KBR, Royal Library of Belgium, including cultural heritage experts, digital humanities researchers and data scientists.

The aim of DATA-KBR-BE is to optimise KBR's existing ICT infrastructure to stimulate sustainable data-level access to KBR's digitised and born-digital collections for digital humanities research. For this project, research teams at the universities of Ghent (GhentCDH and IDLab) and Antwerp (ACDC) work closely together with the digitisation, collections and ICT experts at KBR to co-design two interdisciplinary research scenarios. On the basis of these research scenarios, relevant thematic datasets from KBR's digitised historical newspaper collection, BelgicaPress are extracted for reuse and analysis using digital humanities methods.

This report provides an update on the achieved work, intermediary results, preliminary conclusions and recommendations for the second reporting period of the project (15.3.2021 - 15.3.2022). It also outlines the future prospects and planning for the next reporting period. The collaboration with the DATA-KBR-BE Scientific Advisory Board / Follow-up Committee, who provide scientific and guidance for the project, is also described. Furthermore, valorisation activities, any challenges including potential solutions, as well as any modifications to the project planning, e.g. personnel changes, since the initial report are also included.

In close collaboration with BELSPO, in February 2022 the duration of the DATA-KBR-BE project was extended for a period of 24 months to 15.3.2024. This report therefore reflects the mid-term achievements of the project.

2. ACHIEVED WORK

In order to achieve DATA-KBR-BE's overall objective of facilitating data-level access to KBR's digitised and born-digital collections for digital humanities research, the project is being managed in 5 work packages: *WP1: Co-designing Interdisciplinary Research Scenarios, WP2: Preparation of Datasets, WP3: Data access via data.kbr.be, WP4: Scientific exploitation and valorisation* and *WP5: Project Management and Communication*. An overview of the activities and the achievements in the second reporting period (15.3.2021 - 15.3.2022) per work package are outlined below:

WP1: Co-designing Interdisciplinary Research Scenarios - led by UGent

The aim of this work package is for the researchers in Ghent and Antwerp to work closely with the KBR's collection, digitisation and ICT experts to co-design *two interdisciplinary research scenarios* that can be used as *a basis for extracting relevant thematic datasets* in WP2. Building on the descriptions of the research scenarios described in the original proposal detailed descriptions of the research scenarios were prepared: 1) Collective Action Belgium, led by GhentCDH and 2) Feuilleton in Belgium, led by ACDC.

Since the original submission of the DATA-KBR-BE project proposal, the KBR had successfully been awarded two BELSPO FEDtWIN projects together with the universities Ghent and Brussels (ULB): the KBR Digital Research Lab led by Julie M. Birkholz (GhentCDH-KBR) and CAMille (Centre for Archives on the Media and Information, ULB-KBR) led by Antoine Jacquet (until September 2021) and Brecht Deseure (from February 2022). FEDtWIN projects are intended to build sustainable long-term research collaborations between Belgian Federal Scientific Institutions and Belgian universities. The DATA-KBR-BE project team agreed that close collaboration with both of these research labs would be very valuable for the project. Both FEDtWIN researchers were invited to join the DATA-KBR-BE project team. Additionally, it was agreed to design an *additional interdisciplinary research scenario* on the History of Belgian Journalism, led by ULB/KBR (CAMille).

Since the previous reporting period, a further FEDtWIN project, the KBR Data Science Lab, has been awarded to KBR. The KBR Data Science Lab, led by Tan Lu, in collaboration with the DIMA: Digital Mathematics research group at the Vrije Universiteit Brussel (VUB). Situated in the domain of Artificial Intelligence (AI) for cultural heritage, the KBR Data Science Lab is intended to serve as a research and development hub to bring together inspiration, expertise and resources for data intelligence in the cultural heritage sector. The Data Science Lab has two main goals: a) to facilitate both fundamental and applied research in areas such as mathematical modelling, image and natural language processing and b) to promote the implementation of relevant research outputs in digitisation workflows. The goals of the Data Science Lab align closely with those of DATA-KBR-BE. Tan Lu has been invited to participate in DATA-KBR-BE meetings and explore how his lab can potentially contribute to the interdisciplinary research scenarios.

In the second reporting period work has continued on the two original interdisciplinary research scenarios: 1) Collective Action Belgium, led by GhentCDH and 2) Feuilleton in Belgium, led by ACDC and, following a short hiatus due to staff changes, also on the additional research scenario, the History of Belgian Journalism. As reported in the first DATA-KBR-BE Annual Report, in the original project proposal it has been anticipated that the *interdisciplinary research scenarios* would be carried out in three steps: a) research scenario design, b) dataset preparation and c) scientific exploitation, which would occur sequentially. However, after the initial start-up phase of the project, it became clear that this process was more iterative rather than linear or sequential in nature. As the design of the research scenarios and the testing with the initial datasets had been carried out in the first reporting period, the focus on this second reporting period has been on preparation of the datasets and their initial analysis. These activities will therefore be reported in *WP2: Preparation of Datasets* and *WP4: Scientific exploitation and valorisation*.

WP2: Preparation of Datasets - led by KBR

The aim of this task is a) to work with KBR's ICT team to extract the *thematic datasets* to support the research scenarios co-designed in WP1 and b) to document the various steps in the *data pipeline* to describe how the necessary data was extracted from KBR's ICT systems. This process will lead to the design of a *sustainable data extraction workflow* that will enable research-driven datasets to be extracted from the KBR's ICT infrastructure with minimal effort.

By the end of the first reporting period, two workshops (*DATA-KBR-BE Dataset 1 Workshop in March 2021* and *DATA-KBR-BE Dataset 1 - Follow-up Workshop, 22nd April 2021*) had taken place to prepare for the first data extraction. From a scientific perspective, it had been agreed that the datasets for the two interdisciplinary research scenarios should be somewhat comparable. For example, for each research scenario: 3 Dutch language and 3 French language titles were chosen, each from a specific year. An overview of each of these datasets is provided below:

Feuilleton in Belgium dataset: Six digitised newspapers from 1885. 1885 was selected as this was the first year where there were regularly at least three Dutch language (Het Handelsblad, Vooruit: socialistisch dagblad and De Koophandel) and three French language (Gazette de Charleroi, La Meuse and L'Echo du Parlement) newspapers. By 1885 the Feuilleton 'fashion' had become fairly standard. Furthermore 1885 was within the first century of the Belgian nation.

Collective Action Belgium dataset: Six digitised newspapers from 1913. 1913 was selected, as it was one of the years when a General Strike took place in Belgium, also the year of the World Fair in Ghent and it was the year with the most digital coverage of the newspapers. The dataset included he three Dutch language newspapers (Vooruit: socialistisch dagblad, Het Volk : antisocialistisch dagblad and Vaderland) and three comparative French language newspapers (Le Peuple : organe quotidien de la démocratie socialiste, 2) Le Vingtième Siècle and 3) La Meuse : journal de Liège et de la Province).

History of Belgian Journalism dataset: a different approach was taken for the *History of Belgian Journalism* research scenario as the research team wanted to focus on the entire run of the French-language Roman Catholic newspaper Le Vingtième Siècle (JB729) (1895 - 1940). This has been previously extracted for Antoine Jacquet, so no further data was needed for this research scenario as part of data extraction.

Following the sign-off of the list of titles and dates for DATA-KBR-BE Data Extraction 1 during the DATA-KBR-BE project team meeting in July 2021, the DATA-KBR-BE project team, in close liaison with KBR's ICT Team, prepared for the extraction of the data. A first step in this process was to decide on a method for securely transferring the extracted data to the research teams. KBR's ICT team proposed to scale-up KBR's existing 'SendFile' service (https://sendfile.kbr.be/) to enable larger volumes of data to be transferred. The resulting, was a new **KBR DataSend Service** (https://datasend.kbr.be/). Both services use 'Zend.to', a software package developed at the Electronics and Computer Science Department of the University of Southampton, UK.

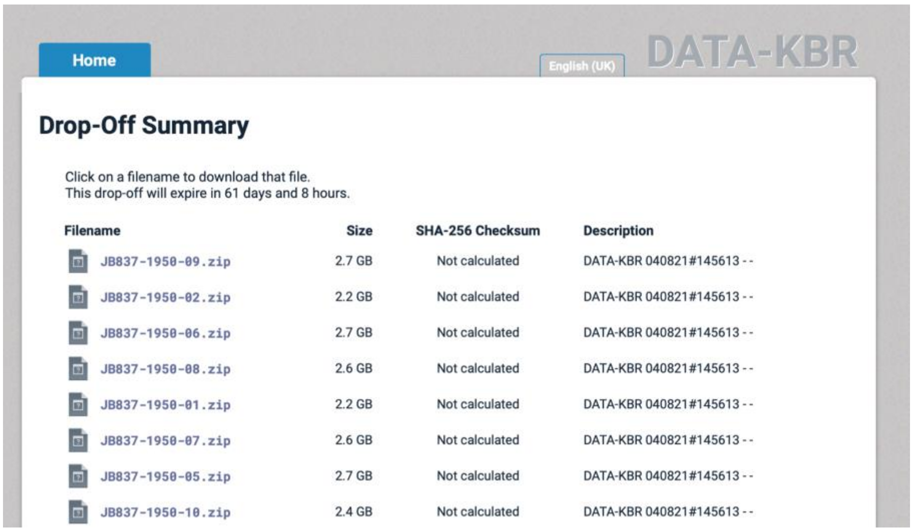
KBR's DataSend Service was developed by adapting a script used internally at KBR for sending digital files to Belspo's Long Term Preservation platform and instead connecting it to Zend.to's API. Currently, the service enables the KBR to transfer one year or one month of a newspaper to a researcher. For example, to transfer a year's worth of data, 12 zip files are created, one for each month. It is possible to download the whole year at once or download the files one month at a time.

Following the set-up of the 'beta' version of the service, it was agreed to undertake a phase of internal testing with the core DATA-KBR-BE project team before undertaking the 'real' transfer of DATA-KBR-BE Data Extraction 1. The aim of this initial test phase is to identify any issues prior to transferring the whole of Data Extraction 1. It was agreed that this testing would be undertaken with the colleagues from IDLab as well as the DATA-KBR-BE project coordinator. This initial phase of testing took place in August 2021 using one year of one newspaper title (Le Peuple) selected from the 'DATA-KBR-BE Data Extraction 1'.

To ensure that this test phase was as useful as possible, it was agreed to transfer Le Peuple from 1950 as this could also be used by the IDLab team to further train their image classification models. To document the evaluation process a Google document was created with a list of questions related to five themes to be answered by the team of testers: *1) Accessing KBR Send Data services?; 2) Downloading the Dataset; 3) Unzipping the dataset; 4) Length of validity of the Dataset for download and 5) Any other issues.* **The detailed results of this initial test phase can be found here, but are also summarised below:**

1) Accessing KBR Send Data services: An email from 'DATA-KBR' was sent to the testing team to notify them that the files were available for download. The email contained well-structured useful information about: the number of files being sent, the link for downloading those files, how long they would be available for, 'technical information' from Zend-to (i.e. Claim ID and Claim Password) and details about the specific files. The testing team did not experience any issues regarding the download of the files. However, it was noted that it would be interesting to know the total size of all the files in the whole 'dropoff' and that the Claim ID and Claim Password had not been needed to download the dataset. In the next interaction of the software, KBR's ICT Team proposed to investigate whether the data extraction script could be adapted to calculate the total (zipped) size of the dataset and to add it to the "Comments section".

2) Downloading the Dataset: For the initial test phase, it had been decided to use one year of one newspaper for testing purposes. This test dataset (as a Zip file) can be downloaded: a) as whole (ca. **29.9 GB**), or b) per month (ca. **2.5 GB**). The download process works by the user clicking on the download link in the email that they have received in step 1. After clicking on the download link, the user is taken to the 'DATA-KBR' website, where they are asked to prove that they are "not a robot" using the "reCAPTCHA" software. This is followed by a standard reCAPTCHA test. Once the user has successfully proved that they are not a robot, the user is provided with a "**Drop-Off Summary**" with a list of files that are available for download (see **Figure 1** below).



Drop-Off Summary

Click on a filename to download that file.
This drop-off will expire in 61 days and 8 hours.

Filename	Size	SHA-256 Checksum	Description
JB837-1950-09.zip	2.7 GB	Not calculated	DATA-KBR 040821#145613 - -
JB837-1950-02.zip	2.2 GB	Not calculated	DATA-KBR 040821#145613 - -
JB837-1950-06.zip	2.7 GB	Not calculated	DATA-KBR 040821#145613 - -
JB837-1950-08.zip	2.6 GB	Not calculated	DATA-KBR 040821#145613 - -
JB837-1950-01.zip	2.2 GB	Not calculated	DATA-KBR 040821#145613 - -
JB837-1950-07.zip	2.6 GB	Not calculated	DATA-KBR 040821#145613 - -
JB837-1950-05.zip	2.7 GB	Not calculated	DATA-KBR 040821#145613 - -
JB837-1950-10.zip	2.4 GB	Not calculated	DATA-KBR 040821#145613 - -

Figure 1. A screenshot of the 'Drop-Off Summary' from KBR's Send Data Service

The user is then required to click on the file they wish to download, after which the download starts automatically. Additionally, there is the option to "**download all files**" and "**download all files as a Zip"**.

The first of the testers (a computer scientist) selected the option to download the dataset as a whole. It was noted that the downloading process caused the testers laptop to slow down, but that the download was successfully completed. The limited amount of storage space on the tester's laptop was also noted (even though the downloaded files are around 29.8 GB, once they have been unzipped they expand to over twice the original size (61.2GB). It was estimated that the total download time was around 4 hours.

Another of the testing team (a computer scientist) noted that it would be useful to receive a direct link to download the files directly to the researcher's server (e.g. via WGet or Curl). Alternatively, details of how to structure a http request so that the files could be downloaded programmatically could be provided. KBR's ICT team noted that the "Zend.to" software offers an Autopickup API (https://zend.to/autopickup) for downloading files that had been dropped off to a server. This option also works with a username and password for additional security.

One of the other members of the testing team (a cultural heritage professional) first started by downloading the files per month. The zip file for January 1950 was 2.4GB. This tester did not experience any issues when downloading the files in this way. However, when the same tester tried the 'Download All Files as a Zip' option, this caused their laptop to slow down so significantly that it was no longer possible to use it. After about 30 minutes of waiting, when the user noted that "nothing much seemed to be happening", the user cancelled the download. They also tried the "Download All Files" option. This started individual downloads of all files at once, which was also too much for their laptop to handle, so the user cancelled this download too. However, the same tester persevered and during a second attempt, used the 'Download All Files as a Zip' option to successfully download the zip file (32 GB). However, the user also noted that she did not use her computer while the download was taking place. This user noted that it was no problem for her to save the zip file to her laptop as she had ca. 100 GB of space available. She also noted that she added the file to her Dropbox folder (a paid Dropbox account with 2TB of storage) which is synchronised with her laptop.

Additionally, there were a number of more general comments as part of the testing process. Firstly, it was also noted that Zend.to offers a **multilingual user interface** (including Dutch, French, German and English) which is very useful for the KBR context. Secondly, one of the user testers noted that it would be nice if the **files for download are listed in chronological order**, e.g. January 1950, February 1950, March 1950 etc., as this was not the case initially. The KBR ICT team looked into the possibility of this and noted that the files had been ordered by "file id". However, with a modification of the script, it was possible for the files for download to be displayed in chronological order. Finally, it was noted that it may be interesting to explore the possibility of i**ncluding e.g. KBR logo or branding on the download page**.

From this initial phase of user testing, it is clear that in general the process of downloading the datasets works quite well, however, there is room for improvement. These potential improvements can be summarised as follows:

File storage: it is important that the researchers have a place with enough storage capacity to store the files that they request for download. This is something that needs to be organised in advance of the file transfer. While some researchers, who are using smaller datasets (up to 10-15 GB) may have enough space on their laptop, this is unlikely to be the case for the majority of researchers. For researchers or research teams, it was proposed that a file sharing platform, such as NextCloud, which would probably be provided by the researcher's university could be used. Within DARIAH-BE (via CLARIAH-Flanders), a pilot project is currently underway to provide secure access to humanities research datasets via the Flemish SuperComputer. This would also facilitate computational analysis of the datasets (either using High Performance Computing Capacity or via lighter-weight solutions such as via Jupyter Notebooks). The Royal Danish Library's Cultural Heritage Cluster is a source of inspiration for Belgium.

Server to server transfer: The possibility of directly transferring the data from 'server to server' transfer needs to be investigated, as for larger file transfers, the KBR Data Send web interface was seen as a bottleneck. For example, if a data platform is available at a researcher's university, the files could be directly transferred there. This would save significant time on upload and download. The possibility of programmatically transferring the data, e.g. via a HTTP request or potentially using the 'Autopickup API (https://zend.to/autopickup) could be explored. This would be important for larger datasets (> 15 GB).

Service usage: In the future it will be necessary to login to KBR Send Data so that the KBR can keep a record of who has downloaded which files, when etc. As part of the login procedure, the user will make a declaration that they are using the files for research purposes (in a similar way to has been implemented for My KBR). However, this functionality has not been implemented for this initial test phase.

3) Unzipping the dataset

Following the download process, the researchers need to unzip the transferred files. We asked the researchers to inform us if they experienced any difficulties with unzipping the files (e.g. my usual unzip programme did not open the files, some of the files were corrupted, expected files were missing, etc).

The computer science researcher did not experience any significant issues when unzipping the dataset. He used the standard unzipping programme provided in his Windows 10 installation. However, as there was less than 50GB of space on his laptop, he needed to manage the space on his computer carefully as he needed over 60GB of space (61.2GB) for the unzipped files. As mentioned in the previous section, for future datasets he would use the data sharing platform of his research group. Following a random check of the unzipped files, he noticed that there were several gaps in the dataset provided. These gaps were investigated by KBR's ICT team. It became clear that this was not a data transfer error, but that those editions were missing from KBR's collections. After further investigation, it became clear that the issues were missing as no newspaper had been published on those days, due to them being Sundays or Public Holidays.

For the second user tester, unzipping one month of digitised newspapers was not a problem. The researcher unzipped the files using 'UnArchiver for Mac'. The uncompressed version of the file was **5.33 GB**. Everything seemed to be correctly received. When unzipping the '**Download All Files as a Zip**' using 'UnArchiver for Mac', there were 12 Zip files inside. Like for the first user tester, the second user tester unzipped the Zip file for March 1950 (Zipped this was **2.9 GB** and uncompressed it was **6,34 GB**). There were also some days missing from the dataset. However, by checking BelgicaPress, it could be seen that the missing dates were also not available in BelgicaPress. For future iterations, it was noted that it may be useful to communicate if there are missing issues when transferring the data to the researcher.

4) Length of validity of the Dataset for download:

For this initial test phase, it was agreed that the datasets will be available for download for two months (62 days). We asked the user testers to tell us what they think of this validity period (e.g. too short, too long, just right). The testing team thought that 62 days were enough to download the dataset. However, it was noted that perhaps it could be considered too much time as DataSend was being used as a 'transitory' storage place, instead of ‘just’ a file transfer. However, this also means that the researcher can go back to the download page any time during the next two months period to download new files as they need them which was considered as a useful additional feature. Furthemore, the DataSend download page provides the user with an automatic count-down reminding them how much time they have left, which was also considered to be useful (e.g. This drop-off will expire in 5 days and 6 hours).

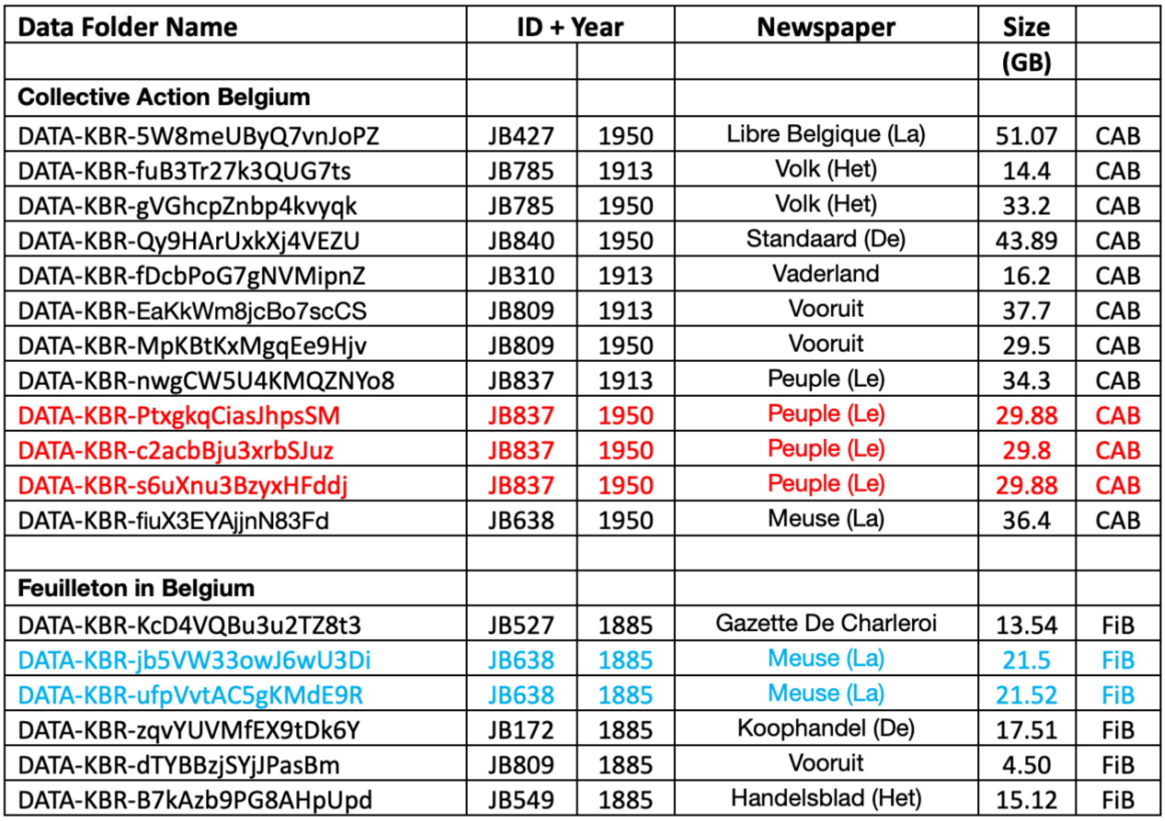
5) Any other issues

Finally, we asked the DataSend testers whether there are any other issues that they thought we would need to consider for the KBR Send Data service. In general, the testers felt that the whole process worked well and that they experienced no major issues when downloading and unzipping the dataset, especially for the smaller datasets, e.g. one month of newspapers of ca. 2.5 GB). Other things that were noted included:

- It was possible to download the same file twice (even if this is not necessarily considered as a problem)
- When opening up DataSend to ‘real’ users, then the security issues of the service would need to be considered. During the testing phase it would have been possible for DataSend’s download email to be forwarded to someone else so that they could download the files. Strengthening the security (perhaps using a login, rather than a Claim ID and Claim password would need to be further analysed and implemented by KBR’s ICT Department. During this testing phase the security of the files is dependent on the discretion of the DATA-KBR-BE team members.
- The ‘Zend.to’ software offers ‘Checksum calculation’ option, to check the authenticity of the files (i.e. that no changes have been made to the file between when they were uploaded by KBR and downloaded by the user). KBR’s ICT team had chosen not to activate this option as it would slow down the process. At this stage it is unlikely that calculating the checksum would be needed for DataSend.
- It was noted that the file names of the downloads include the ‘JB numbers’ (Journal Belgique, Belgian Newspaper) which are KBR’s identification system for each of the newspaper titles, e.g. JB837 = Le Peuple. It was suggested that maybe the list of ‘JB numbers’ could be extracted from KBR’s library catalogue and published on the DATA-KBR-BE website so that the users could easily check which files belong to which newspaper title. The KBR’s ICT team noted that this information is provided in the initial download email "The requested ZIP files for Project DATA-KBR-BE test-fase: Peuple (Le) [JB837 - 1950]", but could potentially also be published on the download page.
- For larger downloads (e.g. all the ‘DATA-KBR-BE Data Extraction 1’) the listing of the files may become confusing for the readers (even a list of 12 files is quite long when downloading one year of a newspaper). A potential solution would be to send the data in batches. However, for future iterations, perhaps a more sophisticated download page could be considered, e.g. with the files clusters by newspaper title, a dropoff page for each research scenario etc.

It is also important to know that in the future it will be necessary for users to login to KBR Send Data so that the KBR can keep a record of who has downloaded which files, when etc. As part of the login procedure, the user will make a declaration that they are using the files for research purposes (in a similar way to has been implemented for My KBR). However, this functionality has not been implemented for this initial test phase.

Transfer of Data Extraction 1 to DATA-KBR-BE Research Groups: With the initial teething issues ironed out as far as possible with the initial test phase, KBR’s ICT team prepared the KBR Data Extraction 1 for transfer to the research groups. An overview of the various datasets, created by Dilawar Ali (IDLab, UGent) can be found below (the final column of the table relates to which interdisciplinary research scenario the particular newspaper is for, e.g Collective Action Belgium (CAB) and Feuilleton in Belgium (FiB)).



Data Folder Name	ID + Year		Newspaper	Size	
				(GB)	
Collective Action Belgium					
DATA-KBR-5W8meUByQ7vnJoPZ	JB427	1950	Libre Belgique (La)	51.07	CAB
DATA-KBR-fuB3Tr27k3QUG7ts	JB785	1913	Volk (Het)	14.4	CAB
DATA-KBR-gVGhcpZnbp4kvyqk	JB785	1950	Volk (Het)	33.2	CAB
DATA-KBR-Qy9HArUxkXj4VEZU	JB840	1950	Standaard (De)	43.89	CAB
DATA-KBR-fDcbPoG7gNVMipnZ	JB310	1913	Vaderland	16.2	CAB
DATA-KBR-EaKkWm8jcBo7scCS	JB809	1913	Vooruit	37.7	CAB
DATA-KBR-MpKBtKxMgqEe9Hjv	JB809	1950	Vooruit	29.5	CAB
DATA-KBR-nwgCW5U4KMQZNYo8	JB837	1913	Peuple (Le)	34.3	CAB
DATA-KBR-PtxgkqCiasJhpsSM	JB837	1950	Peuple (Le)	29.88	CAB
DATA-KBR-c2acbBju3xrbSJuz	JB837	1950	Peuple (Le)	29.8	CAB
DATA-KBR-s6uXnu3BzyxHFddj	JB837	1950	Peuple (Le)	29.88	CAB
DATA-KBR-fiuX3EYAjjnN83Fd	JB638	1950	Meuse (La)	36.4	CAB
Feuilleton in Belgium					
DATA-KBR-KcD4VQBu3u2TZ8t3	JB527	1885	Gazette De Charleroi	13.54	FiB
DATA-KBR-jb5VW33owJ6wU3Di	JB638	1885	Meuse (La)	21.5	FiB
DATA-KBR-ufpVvtAC5gKMdE9R	JB638	1885	Meuse (La)	21.52	FiB
DATA-KBR-zqvYUVMfEX9tDk6Y	JB172	1885	Koophandel (De)	17.51	FiB
DATA-KBR-dTYBBzjSYjJPasBm	JB809	1885	Vooruit	4.50	FiB
DATA-KBR-B7kAzb9PG8AHpUpd	JB549	1885	Handelsblad (Het)	15.12	FiB

Table 1: DATA-KBR-BE Data Extraction 1: overview of newspapers, dates and file sizes.

While in general the file transfers went well, the testing team noted that there were 3 missing drop-offs in the dataset (La Meuse : journal de Liège et de la Province, JB638: 1913; Le Vingtième Siècle, JB729: 1913 and L'Écho du Parlement, JB92: 1885). These 3 missing drop-offs were rectified by KBR's ICT department. Additionally, some minor issues, including duplicate links were noticed in the emails as well as some small variations in file size (highlighted in red and blue on the table above).

Human error - expired downloads: The most significant challenge of the data transfer was however, a human one. As mentioned in the initial test phase, a length of validity of the download period of two months (62 days had been chosen). Even though the team was notified in advance that the DATA-KBR-BE Data Transfer was going to commence, only a few of the team downloaded the datasets before they expired. A number of possible causes for this were identified: **a) bystander effect:** the research team thought that someone else was taking care of the download, **b) deadline was too long:** perhaps with a shorter deadline, researchers may have downloaded the datasets straight away, rather than leaving it for 'later', which turned out to be too late. Additionally, the **lack of a shared platform** to store the data further exacerbated this problem.

Shared Data Platform: To resolve the issue of the lack of a shared platform, it was agreed to use an existing platform; **GhentCDH's Next Cloud environment** (https://data.ghentcdh.ugent.be) for sharing data for the research scenarios with the DATA-KBR-BE project team. NextCloud is an online and client-based sharing platform for securely sharing data, which made it a good solution for sharing data between the DATA-KBR-BE researchers, especially as some of the newspaper data that will be used in the project was published after 1918 and for this reason can only be used for research and education purposes (see: KBR's Copyright page).

To transfer the data to the GhentCDH NextCloud environment, Dilawar Ali (IDLab, UGent) acted as a '**data manager**' on behalf of all the research scenarios and downloaded all the files for all scenarios to a SSD (Solid-State Drive) hard drive for upload to the **GhentCDH NextCloud environment** by Pieterjan De Potter. A SSD hard drive was preferred due to decreased data transfer times. For future data extractions, direct transfer to the GhentCDH NextCloud environment could be considered. Additionally, if it is possible to implement the

server-to-server data transfer option in the future, then the expired download issue should disappear. For smaller data transfers being sent to individual users, expired downloads may be less of an issue as they may urgently need the data and therefore download it promptly. However, for future iterations, reminder emails of the expiry of the download could be automatically sent to the user a few days before the download deadline.

KBR's Data Infrastructure "Stock" and Future Data Extractions

KBR's SendData Service is currently based on KBR's existing data infrastructure. As a result, when additional data is needed, the KBR's SendData service will need to be updated to work with KBR's new data infrastructure ("Stock"). With the new 'Stock' system it will be easier to extract specific documents and download the related files. The 'stock' system will create a copy of files and add them to 'queue of downloads' for sending to the user for example, using a secure File Transfer Service such as KBR Send File or Belnet's FileSender. This process could be automated so that the 'dataset' is emailed automatically to the user without need for human interaction.

The first version of 'Stock' was released in Autumn 2021, at approximately the same time as the first Data Extraction was being transferred to the DATA-KBR-BE Research Groups. Following the initial release of the 'Stock' software and internal testing by KBR Staff, it was possible for KBR's ICT team to migrate KBR's digital archive of files to the new system. This work is currently in progress and is expected to be completed by late Spring 2022. Following the data migration it will be possible for selected departments at KBR (e.g. ICT and Digitisation) to use Stock's new functionality for data extraction. Once this new functionality is in place, the DATA-KBR-BE team will explore how it can be used to support data extraction for the project and more broadly, for example, for researchers requesting access to KBR data. Based on this experience, it will be interesting to explore how this '*corpus building functionality*' could be potentially extended to external KBR users, for example, via the BelgicaPress interface. This functionality could be the first step towards a 'dataset on demand' service, inspired by the KBR's 'digitisation on demand' service which has been particularly successful especially when due to the global Covid-19 pandemic, the KBR could not be open for readers.

At the time of writing, it is anticipated that further data extractions will not be needed immediately, as the transferred data will first need to be analysed as part of the Interdisciplinary Research Scenarios *(see: WP4: Scientific exploitation and valorisation)*.

WP3: Data access via data.kbr.be - led by KBR

The goal of WP3 is to: a) design, b) implement and c) test the usability of the new data.kbr.be data platform. The implementation of the *data.kbr.be platform* will include the Open Humanities datasets prepared in WP2, and the development of a *Digital Asset Registry* to inventorise KBR's digital (digitised and born-digital) collections as well as collections that are currently in KBR's digitisation pipeline. In this phase of the project (15.3.2021-15.3.2022) the majority of the activity has focussed on the data extraction and sharing (see: *WP2: Preparation of Datasets).* Now that the first Data Extraction is complete and the migration of the data to the KBR's new data infrastructure 'Stock' is underway, the DATA-KBR-BE team can turn their attention to designing the DATA-KBR-BE platform. This work builds further on the outcomes (see: workshop report) of the *DATA-KBR-BE Brainstorming Workshop* held in November 2020. A first concrete step in this process, is to organise a follow-up brainstorm with Thuy-An Pham from KBR's ICT department in Summer 2022 to further detail the technical and functional requirements for the DATA-KBR-BE platform (an initial overview of the notes from this brainstorm can be found here). It is intended that this brainstorm will iteratively lead to a **first version of the technical and functional requirements for the DATA-KBR-BE platform**. This document will then be used to undertake additional semi-structured conversations with key colleagues within KBR, e.g. ICT department, digitisation department, collection managers and KBR's digital data strategy team, to further iterate the document. Once mature enough, it will be also shared with the researchers in the DATA-KBR-BE project, including the Follow-up Committee and the FEDtWIN projects (e.g. KBR Digital Research Lab, CAMiLLE and the Data Science Lab).

WP4: Scientific exploitation and valorisation - led by UAntwerpen

The aim of WP4 is to *carry out the interdisciplinary research scenarios* co-designed in WP1 using the *thematic* *datasets* extracted in WP2 and published on *data.kbr.be* in WP3. Originally in the project proposal it was anticipated that these three steps: a) research scenario design, b) dataset preparation and c) scientific exploitation would occur sequentially. However, after the initial start-up phase of the project, this process was more iterative rather than linear or sequential. In this phase of the project (15.3.2021-15.3.2022) much of the work has been focused in *WP2: Preparation of Datasets,* however, some initial preparation for the analysis of these datasets has been carried out in WP4.

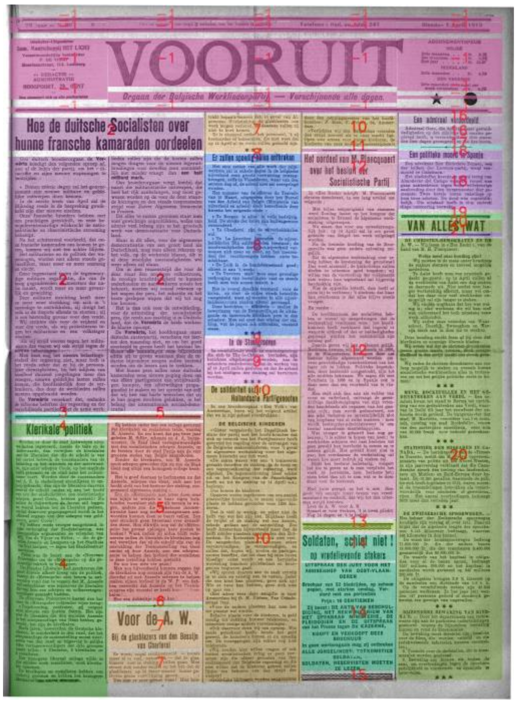
Following the first Data Extraction and sharing of the DATA-KBR-BE datasets from the via the **GhentCDH Next Cloud environment** (https://data.ghentcdh.ugent.be) (see *WP2: Preparation of Datasets),* a series of meetings took place with the interdisciplinary research scenarios to help guide the analysis process. Firstly in January 2022, individual meetings were held with the *Feuilleton in Belgium* and the *Collective Action Belgium* teams to understand: a) how would they like to analyse the data from Data Extraction 1 in order to carry out the research scenario, b) which specific research questions would they like to answer and c) to propose some next steps to start analysing the data. The purpose of these preparatory meetings was to provide a basis for understanding the **data science needs** of both research scenarios in order to agree on planning and next steps for a meeting in February 2022.

For **Collective Action Belgium**, the first challenge was to understand how to ***extract the articles*** from the dataset that are related to collective action, e.g. strikes, demonstrations and protests. The extraction of the relevant articles could be undertaken in a number of ways: for example, by performing keyword searches on the dataset using a range of different keywords (e.g. Werkstaking*; Algem*ne werkstaking*;
Gestaak; Staken; Stakers; Staaksters and Werkonderbreking) or alternatively using statistical methods such as topic modelling to determine which articles in the corpus are relevant for the research scenario. However, in both cases, the ***segmentation of the articles*** in the dataset would be crucial. Additionally, from a manual analysis of the dataset, the potential need to ***improve the OCR*** of a select part of the corpus was identified. It was agreed to start with the "*Vooruit"* from *April 1913* as the test dataset.

In comparison, the aim of the **Feuilleton in Belgium** research scenario is to investigate the prominent literary authors in the first century of the Belgian state. As a first step, it was proposed to use the *'Het Handelsblad'* from *January 1885* as the test dataset to initiate the analysis. Similarly to the Collective Action Belgium research scenario, ***article segmentation*** was key to the future analysis of the dataset. However, instead of needing to identify which of the articles were relevant for the research scenario, the article segmentation requirement was related to ***automatically identifying and extracting the feuilletons*** in the dataset as a specific type of article. Regarding the extraction of the feuilleton, a number of requirements were identified: a) *metadata*: for example provided in JSON format, both at the level of the newspaper title (e.g. title, date, page number(s)) and at the level of the feuilleton (e.g. title, author and sequence number (i.e. which instalment of the literary work, e.g. 3rd episode in the series) as well as the coordinates of the 'feuilleton zone', b) *full-text*: a transcript of the OCR'ed text, ideally in plain text (e.g. .txt file), and c) *image:* a 'cut out' or crop of the image of the feuilleton (e.g. as pdf).

The overall outcome of the February 2022 meeting was for Dilawar to run his article segmentation pipeline on a *test dataset* for both of the research scenarios: a) *Collective Action Belgium (Vooruit* from *April 1913*) and b) Feuilleton in Belgium (*Het Handelsblad* from *January 1885*) and then to present the results back to the team during the DATA-KBR-BE meeting in March 2022.

The DATA-KBR-BE team met again on 9th March 2022 for Dilawar to present the results of his work back to the team (see: Presentation). Firstly, for *Collective Action Belgium*, Dilawar used the **article segmentation pipeline** that he had developed earlier for *Le Peuple* from 1938 as the basis for undertaking article segmentation on the "*Vooruit"* from *April 1913*. Since his original iteration of the pipeline, he had already been working to improve it. However, when applying the (improved) 1938 model to the test dataset from 1913, this, understandably, led to the generation of new errors, e.g. the detection of some text blocks being missed and difficulties in recognising two articles in a single column.


VOORUIT
Orgaan der Belgische Werkliedenpartij — Verschijnende alle dagen
Hoe de duitsche Socialisten over hunne fransche kamaraden oordeelen
Kleinkale politiek
Voor de A. W.
Soldaten, schiet niet!
VAN ALLES WAT

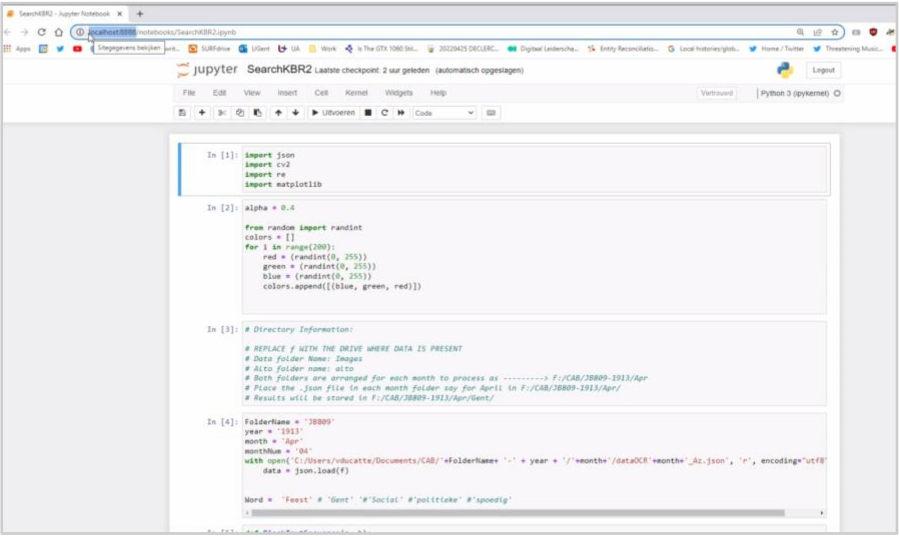
The addressing these new errors by, for example, refining the detection of text lines and using these text lines (instead of text blocks) for the automatic recognition of articles, led to further improvements in the article segmentation, such as the recognition of multiple news items in one column, the detection of text blocks which had been previously missed, as well improvements in the detection of links between images and text blocks. This improved article segmentation pipeline (image detection, line detection, layout detection and article detection) was then applied to the *Collective Action Belgium 'test dataset'* of the "*Vooruit"* from *April 1913*. **Figure 2** shows a visualisation of the article segmentation for the Vooruit from 1st April 1913.

Following the segmentation of the articles, the next step for the *Collective Action Belgium* research scenario is to identify which of the articles are related to collective action, e.g. strikes, demonstrations and protests etc. For this extraction of the article text and the related metadata was needed.

Figure 2. An example visualisation of article segmentation

The data extraction for *Collective Action Belgium* included: the *date* of the newspaper edition (e.g. 1st April 1913); which *file* the article could be found in (e.g. KB_JB809_1913-04-01_01-00001.jpg_img5.jpg); *type* of data extracted (e.g. header, article title, text, illustration); the *bounding box* (e.g. the coordinates of the article) and the *OCR* (e.g. the text which was been automatically generated using OCR software). This data was provided as a separate JSON file.

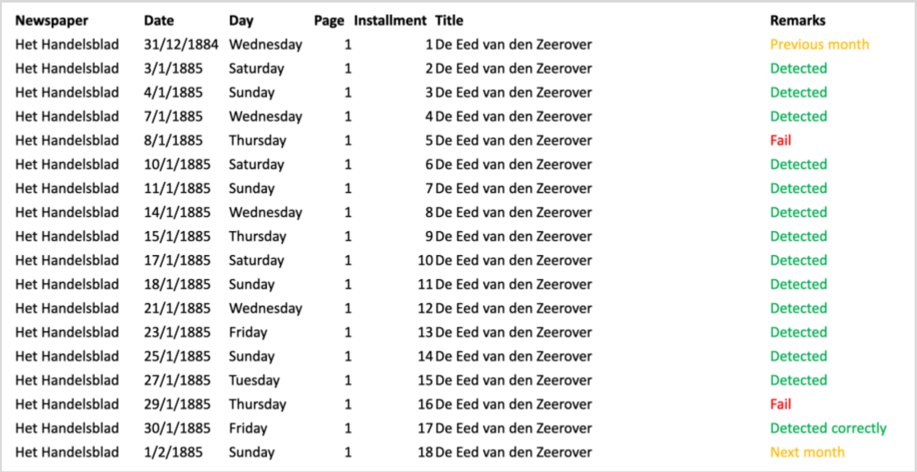
Finally, Dilawar Ali created a ***Jupyter Notebook*** to facilitate the searching of the test dataset to find the relevant articles related to collective action. This locally run notebook, which was uploaded to the GhentCDH NextCloud environment, enabled the JSON file with the OCRed text and relevant article metadata which had been extracted from the *Collective Action Belgium* test dataset to be uploaded and searched for keywords related to collective action (see: **Figure 3** below) to identify relevant articles. Although the *Collective Action Belgium* research team needed some technical assistance from the ID Lab team to get the notebook up and running, it did prove a useful method to explore the test dataset for relevant articles. In the future, the DATA-KBR-BE team could consider publishing standard notebooks for searching such datasets, for example, on [KBR's Github Repository.](#)



```
In [1]: import json
        import cv2
        import re
        import matplotlib

In [2]: alpha = 0.4

        from random import randint
        colors = []
        for i in range(200):
            red = (randint(0, 255))
            green = (randint(0, 255))
            blue = (randint(0, 255))
            colors.append([(blue, green, red)])

In [3]: # Directory Information:

        # REPLACE f WITH THE DRIVE WHERE DATA IS PRESENT
        # Data folder Name: Images
        # Alto folder name: alto
        # Both folders are arranged for each month to process as ---------> F:/CAB/JB809-1913/Apr
        # Place the .json file in each month folder say for April in F:/CAB/JB809-1913/Apr/
        # Results will be stored in F:/CAB/JB809-1913/Apr/Gent/

In [4]: FolderName = 'JB809'
        year = '1913'
        month = 'Apr'
        monthNum = '04'
        with open('C:/Users/vducatte/Documents/CAB/'+FolderName+ '-' + year + '/'+month+'/dataOCR'+month+'_Az.json', 'r', encoding="utf8'
            data = json.load(f)


        Word =  'Feest' # 'Gent' '#'Social' #'politieke' #'spoedig'
```

Figure 3. A Jupyter notebook to facilitate the identification of relevant articles from the test dataset

For the *Feuilleton in Belgium* a similar process was undertaken, this time with the test dataset of *Het Handelsblad* from *January 1885*. As the *Feuilleton in Belgium* research scenario requires the detection of a specific type of article, the literary supplement or feuilleton, when the improved algorithm from 1938, supplemented by the improvements from the *Collective Action Belgium* test from 1913 was applied to the test dataset from 1885, this again led to the generation of new errors, e.g. text blocks that are part of the feuilleton being detected as belonging to an article in the upper part of the page. However, for the test dataset (see **Figure 4** below) the majority of feuilletons were correctly detected, with only the feuilletons for 8th and 29th January missing.



Newspaper	Date	Day	Page	Installment	Title	Remarks
Het Handelsblad	31/12/1884	Wednesday	1	1	De Eed van den Zeerover	Previous month
Het Handelsblad	3/1/1885	Saturday	1	2	De Eed van den Zeerover	Detected
Het Handelsblad	4/1/1885	Sunday	1	3	De Eed van den Zeerover	Detected
Het Handelsblad	7/1/1885	Wednesday	1	4	De Eed van den Zeerover	Detected
Het Handelsblad	8/1/1885	Thursday	1	5	De Eed van den Zeerover	Fail
Het Handelsblad	10/1/1885	Saturday	1	6	De Eed van den Zeerover	Detected
Het Handelsblad	11/1/1885	Sunday	1	7	De Eed van den Zeerover	Detected
Het Handelsblad	14/1/1885	Wednesday	1	8	De Eed van den Zeerover	Detected
Het Handelsblad	15/1/1885	Thursday	1	9	De Eed van den Zeerover	Detected
Het Handelsblad	17/1/1885	Saturday	1	10	De Eed van den Zeerover	Detected
Het Handelsblad	18/1/1885	Sunday	1	11	De Eed van den Zeerover	Detected
Het Handelsblad	21/1/1885	Wednesday	1	12	De Eed van den Zeerover	Detected
Het Handelsblad	23/1/1885	Friday	1	13	De Eed van den Zeerover	Detected
Het Handelsblad	25/1/1885	Sunday	1	14	De Eed van den Zeerover	Detected
Het Handelsblad	27/1/1885	Tuesday	1	15	De Eed van den Zeerover	Detected
Het Handelsblad	29/1/1885	Thursday	1	16	De Eed van den Zeerover	Fail
Het Handelsblad	30/1/1885	Friday	1	17	De Eed van den Zeerover	Detected correctly
Het Handelsblad	1/2/1885	Sunday	1	18	De Eed van den Zeerover	Next month

Figure 4. Feuilleton detection on Het Handelsblad from January 1885

As for the previous research scenario, extraction of the article text and the related metadata was needed. This time, a separate text file (.txt) for each article including the article text was provided, as well as a JSON file with the related metadata. The related metadata was similar to that for *Collective Action Belgium*, but also included some additional data fields: "ID" (e.g. the number of the text block); the *date* of the newspaper edition (e.g. 3rd January 1885); which *file* the article could be found in (e.g. KB_JB549_1885-01-03_01-00001.jpg_img5.jpg); "articleID" (e.g. the identifier of the article which is composed of a number of text blocks); *type* of data extracted (e.g. Feuilleton); the *bounding box* (e.g. the coordinates of the article); "titleOCR" (e.g. containing the title of Feuilleton title extracted from the OCR) and the related "text" (e.g. containing the text of the Feuilleton extracted from the OCR); "AzureOCR" (e.g. the output of the reOCRing process using MS Azure) and "SequenceID" (e.g. relating to the *reading order* of the articles on the newspaper page). A Jupyter notebook is not needed for this research scenario.

The **results of the pre-processing of these test datasets** were uploaded to a new "research" folder which was added to the existing folders for each research scenario in the DATA-KBR-BE folder on the GhentCDH NextCloud platform. Even during the meeting, the first impressions from the *research scenario teams* were that this article segmentation was already very useful for the researchers in order for them to be able to start analysing the data.

Regarding the *History of Belgian Journalism* research scenario, Brecht Deseure took up his post in February 2022 as the new post-doctoral researcher for CAMille, the Centre for Archives on the Media and Information, ULB-KBR. A first meeting with Brecht took place in March 2022 to introduce him to the DATA-KBR-BE project, the achievements to date and the previous work that had been undertaken on the project by Antoine Jacquet. During this initial meeting, Brecht explained that regarding the use of digital methods, the previous work undertaken by Dilawar Ali on **automatic recognition of signatures** was particularly of interest. Currently, an existing *Dictionary of Belgian Journalism*, compiled by Pierre Van den Dungen, is being converted to a database. Within the CAMille project, the objective is to be able to provide links from the authors signatures identified in newspaper articles, to the biographical entries about the Belgian journalists in the database. Brecht already anticipated that **article segmentation** would also be of interest. It was agreed to set up a joint follow-up meeting with the whole CAMille team (Florence le Cam, Brecht Deseure and Sébastien De Valeriola) as well as colleagues from ID Lab and KBR Newspaper Curator Marc D'hoore to explore the collaboration possibilities further. Access to the DATA-KBR-BE data via the GhentCDH Next Cloud was provided to the *History of Belgian Journalism* team.

In the coming months, the research teams will continue to further explore and analyse the test datasets based on the needs of their research scenarios. Based on this analysis it will be determined whether further data from Data Extraction 1 will need to be pre-processed (e.g. running the article segmentation on the whole year or ReOCRing selected parts of the data) or whether additional data will need to be extracted.

DATA-KBR-BE: Stimulating Research Collaborations

In addition to two original research scenarios (*Collective Action Belgium* and the *Feuilleton in Belgium*), the additional research scenario (*History of Belgian Journalism*), the DATA-KBR-BE project has sparked the interest of further potential research collaborations.

Firstly, thanks to the participation of the DATA-KBR-BE project coordinator in a Belgian meeting of the European Holocaust Research Infrastructure (EHRI) in September 2021, in the framework of another project, a conversation started about DATA-KBR-BE and BelgicaPress with a researcher from the Zentrum für Ostbelgische Geschichte (ZOG). From the conversation, it became clear that the ZOG had a collection of around **50 German-language digitised historical newspapers titles from East Belgium**, that they were looking to valorise further. Recognising the opportunity, a follow-up meeting was organised with another of the ZOG team members to discuss the collection further.

As a result of the initial meeting, a larger meeting of the ZOG team, together with the DATA-KBR-BE project coordinator and colleagues from KBR's Digitisation team was organised in January 2022 to explore a longer-term collaboration between ZOG and KBR. An outcome of the meeting was to pilot the integration of three of ZOG's digitised newspapers (*Eupener Zeitung, Eupener Nachrichten and Grenz-Echo*) into Belgica Press. As a first step, ZOG provided KBR with a test dataset with some sample data from the three newspaper titles which was analysed by Peter Catrie (Digitisation Expert for Newspapers, KBR) to assess its suitability for inclusion in BelgicaPress. Following which, it was agreed to integrate the three newspaper titles on a step-by-step basis into BelgicaPress. It was agreed to start with the Grenz-Echo, both because it is a newspaper title that is still published today and as there are already several (different) editions of this newspaper in BelgicaPress, followed by the other two titles in the coming months. It was agreed to organise online meetings every 2 months to track progress.

By March 2022, several test samples of the Grenzo Echo were already online in BelgicaPress (e.g. 9th June 1928; 1st January 1929; 13th July 1929 and 23rd August 1930). The potential of scaling up this ingestion to ZOG's full digitised newspaper collection can already be anticipated. Furthermore it was agreed to explore the possibility of including short descriptions of the newspaper titles to provide additional contextual information for researchers. Finally, the possibility of including an **additional research scenario related to the ZOG's East German newspapers** was explored. This additional research scenario will be discussed further following the ingestion of the newspaper data into BelgicaPress.

Additionally, **KBR receives requests from individual researchers for digitised files,** for example, from KBR's digitised newspaper collections. One such request came from a Ghent University PhD student (Elias Degruyter). For his PhD, Elias is researching media discourse and nation building in Flanders (1944-1962). He was interested in exploring the possibilities of using digital methods to analyse some of the Flemish historical newspapers (*De Standaard*, *Het Laatste Nieuws*, *Vooruit* en *De Rode Vaan*) which have already been digitised. Currently, KBR generally digitised historical newspapers until 1950. To date, such requests are dealt with on a case-by-case basis. In the context of the DATA-KBR-BE project, it was agreed to **explore how a standardised and sustainable workflow could be developed for providing such data to researchers**. In the coming months, the DATA-KBR-BE team agreed to work with Elias - who is already connected with the Ghent Centre for Digital Humanities - to provide access to this newspaper data and use it as a case study to design a workflow for this type of '**datasets on demand**'.

Finally, later in 2022, once the analysis of Data Extraction 1 has progressed further, it will be important to discuss how to valorise the DATA-KBR-BE research scenarios further. As reported in the previous Annual Report, it was agreed that the DATA-KBR-BE team would prepare at least **one interdisciplinary, peer-reviewed journal article** such as DSH: Digital Scholarship in the Humanities. If possible, we would like to additionally prepare one peer-reviewed journal article per research scenario. Based on data science work undertaken by Dilawar Ali (IDLab, UGent) and for partial fulfilment of his PhD, in January 2022, Dilawar together with other members of the DATA-KBR-BE team submitted an article entitled: "*Computer Vision and Machine Learning Approaches for Metadata Enrichment to Improve Searchability of Historical Newspaper Collections*" to a special edition of the Journal of Documentation on "Artificial Intelligence for Cultural Heritage Materials", coordinated by the AEOLIAN (Artificial Intelligence for Cultural Organisations) network (see: Call for Papers). The outcome of the review process is expected later in Summer 2022.

Other potential methods for valorising the DATA-KBR-BE research scenarios include: *blog posts* (e.g. to be published on the KBR website); the *DATA-KBR-BE Hackathon* which is scheduled to take place towards the end of the project and the *publication of derived datasets* from the research scenarios in SODHA, Social Sciences and Digital Humanities Archives, data repository. Since the last report, KBR now has a SODHA sub-repository (https://www.sodha.be/dataverse/kbr), which could be used for this purpose. Finally, as mentioned earlier, the DATA-KBR-BE team could also consider *publishing standard notebooks for searching DATA-KBR-BE datasets*, for example, on KBR's Github Repository.

Finally, the NewspAIper demonstrator, which uses Le Peuple from 1938, demonstrates the potential of: a) article segmentation, b) linking text recognition with open data, c) finding similar images across the collection and d) initial ideas for an interactive filter which could be implemented within BelgicaPress. This demonstrator continues to be used to showcase the results of the data science research undertaken in the framework of DATA-KBR-BE. Having access to the NewspAIper demonstrator makes it easier to show to potentially research collaborators the potential of DATA-KBR-BE's research to provide '**Newspapers as Data**'.

The demonstrator was presented at the international NewsEye Conference, What's Past is Prologue: the NewsEye International Conference in March 2021 and a demonstration session is planned at the DH Benelux Conference in June 2022.

WP5: Project Management and Communication - led by KBR (M1-M24, July 2020 - June 2022)

The aim of WP5 is to ensure the timely management and monitoring of the project, including liaison with BELSPO, organisation of Follow-up Committee meetings and communication activities. The DATA-KBR-BE project team continued to meet regularly, however, during the second reporting period, there has been an increased need for more specific meetings with various team members, e.g. meetings regarding the data extraction workflow; research-scenario specific meetings and meetings to discuss the development of the DATA-KBR-BE platform. During this reporting period, a meeting of the DATA-KBR-BE Scientific Advisory Board ('Follow-up Committee') was organised in June 2021 via Zoom, with an additional update being provided in December 2021. Further details related to the Follow-up Committee can be found in *Section 6*. The DATA-KBR-BE communication activities are detailed in *Section 7*. As noted in the previous report, the possibility of extending the duration of the DATA-KBR-BE project was being explored. This project extension took place in February 2022. Further details regarding this are provided in *Section 8*.

3. INTERMEDIARY RESULTS

This section provides a detailed description of the deliverables completed in this reporting period:

WP1: Co-designing Interdisciplinary Research Scenarios

Following the initial delivery of *D1.1 Report describing Co-Designed Interdisciplinary Research Scenarios* in the first reporting period (December 2020), in the second reporting period, work has continued on the two original interdisciplinary research scenarios: 1) Collective Action Belgium, led by GhentCDH and 2) Feuilleton in Belgium, led by ACDC and, following a short hiatus due to staff changes, also on the additional research scenario, the History of Belgian Journalism. As the design of the research scenarios and the testing with the initial datasets had been carried out in the first reporting period, the focus for this reporting period was on the preparation of the datasets and their initial analysis. The results of these activities are reported in *WP2: Preparation of Datasets* and *WP4: Scientific exploitation and valorisation*.

WP2: Preparation of Datasets

There are two deliverables foreseen in WP1: *D2.1 Extraction of thematic datasets to support research scenarios (M9, M12, M16)* and *D2.2 Sustainable data extraction workflow design (M24)*.

D2.1 Extraction of thematic datasets to support research scenarios

During the second reporting period, following the sign-off of the list of titles and dates for DATA-KBR-BE Data Extraction 1 in July 2021, the DATA-KBR-BE project team, in close liaison with KBR's ICT Team, prepared for the extraction of the data. This included a number of steps. The first step in the process was to decide on **a method for securely transferring the extracted data to the research teams.** It was agreed to scale-up KBR's existing 'SendFile' service (https://sendfile.kbr.be/) to enable larger volumes of data to be transferred. The resulting, was a new **KBR DataSend Service** (https://datasend.kbr.be/). Following the set-up of the 'beta' version of the service, a phase of internal testing with the core DATA-KBR-BE project team before undertaking the 'real' transfer of DATA-KBR-BE Data Extraction 1. The results of this **initial testing phase**, which took place in August 2021 are detailed in *Section 2* above, however, the **key results** are detailed here.

Firstly, in general **the process of downloading the datasets works quite well**, however, there is room for improvement. The key improvements are detailed as follows: **1) File storage:** it is important that the researchers have a place with enough storage capacity to store the files that they request for download, **2) Server to server transfer**: the possibility of directly transferring the data from 'server to server' transfer needs to be investigated, as for larger file transfers, the KBR Data Send web interface was seen as a bottleneck, **3) Service usage:** in the future it will be necessary to login to KBR Send Data so that the KBR can keep a record of who has downloaded which files, when etc. As part of the login procedure, **the user will make a declaration that they are using the files for research purposes** (in a similar way to has been implemented for My KBR). This functionality would need to be implemented if the KBR DataSend Service is launched as a live service. On the basis of this initial testing phase, it was possible to make a number of improvements before carrying out the 'real' data transfer.

Following the initial testing phase, the **Transfer of Data Extraction 1 to DATA-KBR-BE Research Groups** was undertaken in September 2021. In general the file transfers went well, with minor issues such as missing drop offs, duplicate links in the emails and some small variations in file size. However, the most significant challenge of the data transfer was a human one: **the download time had expired before the prepared datasets had been downloaded**. A number of possible causes for this were identified: **a) bystander effect:** the research team thought that someone else was taking care of the download, **b) deadline was too long:** perhaps with a shorter deadline, researchers may have downloaded the datasets straight away, rather than leaving it for 'later', which turned out to be too late. Additionally, the **lack of a shared platform** to store the data further exacerbated this problem.

To resolve the issue of the lack of a **data sharing platform**, it was agreed to use an existing platform; **GhentCDH's Next Cloud environment** (https://data.ghentcdh.ugent.be) for sharing data for the research scenarios with the DATA-KBR-BE project team. All the files for **Data Extraction one** were transferred by Dilawar Ali (IDLab, UGent) to GhentCDH's Next Cloud environment and **all the DATA-KBR-BE project team were given access to the data** on this platform. It is anticipated that further data extractions will not be needed immediately, as the transferred data will first need to be analysed as part of the Interdisciplinary Research Scenarios *(see: WP4: Scientific exploitation and valorisation)*. It may be that the researchers will have sufficient data with Data Extraction 1 or will require **customised data extractions** as their research progresses.

D2.2 Sustainable data extraction workflow design.

The extraction of DATA-KBR-BE Data Extraction 1 has been very useful to start documenting the technical and practical details of the data extraction to prepare for the design of the **sustainable data extraction workflow.** However, it is important to note that the **data extraction workflow** that has been implemented is based on the **KBR's existing data infrastructure.** As a result, when additional data is needed, **the data extraction workflow will need to be updated to work with KBR's new data infrastructure ("Stock")**. Following the initial release of the 'Stock' infrastructure in Autumn 2021, KBR's ICT team is in the process of migrating KBR's digital archive of files to the new data infrastructure. This work is currently in progress and is expected to be completed by late Spring 2022. Following the data migration it will be possible for selected departments at KBR (e.g. ICT and Digitisation) to use **Stock's new functionality for data extraction**. Once this new functionality is in place, the DATA-KBR-BE team will explore how it can be used to support data extraction for the project and more broadly, for example, for researchers requesting access to KBR data. This new functionality for data extraction will be used as a basis for creating the **Sustainable data extraction workflow design (D2.2).**

WP3: Data access via data.kbr.be

There are three deliverables foreseen in WP3: *D3.1 Design of the KBR Open Data Platform: data.kbr.be*; *D3.2 Implementation of data.kbr.be including dataset publication* and *D3.3 Digital Asset Registry: inventory of KBR's Digital Collections.* In this reporting period, the focus of our activities has been on D3.1.

D3.1 Design of the KBR Open Data Platform: data.kbr.be

The goal of WP3 is to: a) design, b) implement and c) test the usability of the new data.kbr.be data platform. The implementation of the *data.kbr.be platform* will include the Open Humanities datasets prepared in WP2, and the development of a *Digital Asset Registry* to inventorise KBR's digital (digitised and born-digital) collections as well as collections that are currently in KBR's digitisation pipeline.

In this reporting period, the majority of the activity has focussed on the data extraction and sharing (see: *WP2: Preparation of Datasets).* However, building further on the outcomes (see: workshop report) of the *DATA-KBR-BE Brainstorming Workshop* held in November 2020, a follow-up brainstorm with Thuy-An Pham from KBR's ICT department is planned for Summer 2022. The aim of this brainstorm is to further detail the technical and functional requirements for the DATA-KBR-BE platform (an initial overview of the notes from this brainstorm can be found here). It is intended that this brainstorm will iteratively lead to a **first version of the technical and functional requirements for the DATA-KBR-BE platform**.

WP4: Scientific exploitation and valorisation

There are two deliverables foreseen in *WP4: D4.1 Publication of Open Datasets in a Trusted Digital Repository* and *D4.2 Report of the High-Profile Hackathon.* In this reporting period, the focus of our activities has been on D4.1.

D4.1 Publication of Open Datasets in a Trusted Digital Repository

Following the first Data Extraction and sharing of the DATA-KBR-BE datasets from the via the **GhentCDH Next Cloud environment** (https://data.ghentcdh.ugent.be) (see *WP2: Preparation of Datasets),* a series of meetings took place with the interdisciplinary research scenarios to help guide the analysis process. The purpose of these meetings was to understand the **data science needs** of the research scenarios.

For **Collective Action Belgium**, the key data science needs are: a) how to ***extract the articles*** from the dataset that are related to collective action, e.g. strikes, demonstrations and protests, b) ***segmentation of the articles*** and c) ***improving the OCR***. For the **Feuilleton in Belgium**, the key data science needs are: a) ***article segmentation***, b) ***automatic identification and extracting the feuilletons*** as a specific type of article, including the *metadata*; *full-text* and the *image* of the feuilleton. For the **History of Belgian Journalism**, the key data science needs are: a) **automatic recognition of signatures** and b) **article segmentation**.

The data science team (UGent, IDLab), presented the results (see: Presentation) of their initial analysis of the test dataset using the **article segmentation pipeline**. The **results of the pre-processing of these test datasets** were uploaded to a new "research" folder which was added to the existing folders for each research scenario in the DATA-KBR-BE folder on the GhentCDH NextCloud platform. The first impressions from the *research scenario teams* were that **article segmentation** in particular was very useful for the researchers to start analysing the data.

Additionally, a ***Jupyter Notebook*** to facilitate the searching of the test dataset to find the relevant articles was developed. In the future, the DATA-KBR-BE team could consider publishing standard notebooks for searching such datasets, for example, on KBR's Github Repository.

In the coming months, the research teams will continue to further explore and analyse the test datasets based on the needs of their research scenarios. Based on this analysis it will be determined whether further data from Data Extraction 1 will need to be pre-processed (e.g. running the article segmentation on the whole year or ReOCRing selected parts of the data) or whether additional data will need to be extracted.

WP5: Project Management and Communication
There were originally two deliverables foreseen in WP5*: D5.1 Annual Report* and *D5.2 Final Report.* However, due to the agreed extension of the duration of the project for a period of 24 months to 15.3.2024, two additional Annual Reports were requested for the periods: *15.03.2021-15.3.2022* and *15.03.2022-15.3.2023*. This report provides the activities and achievements for 15.03.2021-15.3.2022*.*

During the second reporting period, the DATA-KBR-BE project team continued to meet regularly, however, there has been an increased need for more specific meetings with various team members, e.g. for the data extraction workflow; research-scenario specific meetings and meetings to discuss the development of the DATA-KBR-BE platform. During this reporting period, a meeting of the DATA-KBR-BE Scientific Advisory Board ('Follow-up Committee') was organised in June 2021 via Zoom, with an additional update being provided in December 2021. Further details related to the Follow-up Committee can be found in *Section 6*. The DATA-KBR-BE communication activities are detailed in *Section 7*. The project extension took place in February 2022. Further details regarding this are provided in *Section 8*.

4. PRELIMINARY CONCLUSIONS AND RECOMMENDATIONS

The overall aim of the DATA-KBR-BE project is to facilitate data-level access to KBR's digitised and born-digital collections for digital humanities research, through the optimisation of KBR's existing ICT infrastructure. During this second reporting period, the project has already undertaken some core activities (see Section 2) and achieved some significant results (see Section 3). On the basis of these activities and results, the following preliminary conclusions and related recommendations can be drawn:

Stimulating and facilitating academic research using KBR's collections
As identified at the end of the first reporting period, through DATA-KBR-BE, KBR has been able to take its first concrete steps towards ***extending KBR's services for researchers.*** This has not only been evidenced by the research collaboration within the project consortium through the two original research scenarios (*Collective Action Belgium* and the *Feuilleton in Belgium*) and supplemented by the inclusion of the additional research scenario (*History of Belgian Journalism*), but the DATA-KBR-BE project has additionally sparked the interest of further research collaborations *beyond* the project consortium.

These additional collaborations so far include the Zentrum für Ostbelgische Geschichte (ZOG), where a pilot project is taking place to evaluate the feasibility of the integration of ca. **50 German-language digitised historical newspapers titles from East Belgium** into BelgicaPress and exploring the possibility of including an **additional research scenario related to the ZOG's East German newspapers**.

Additionally, **KBR receives requests from individual researchers for digitised files,** for example, from KBR's digitised newspaper collections. As a result, it has been agreed to **explore how a standardised and sustainable workflow could be developed for providing such data to researchers**, which would be a first step towards developing a '**datasets on demand**' service.

Data preparation as an initial step towards corpus building
In the first reporting period, we identified that the design and implementation of the interdisciplinary research scenarios is an iterative process rather than a linear series of steps. However, in the second reporting period, the complexity and also iterative nature of data preparation also came to light. It became evident that corpus development is also an ongoing process, which is grounded in the historical context of the period being studied and iteratively develops through exploration. It is often quite 'organic', rather than methodical and is seldom documented. This process starts with the process of selecting which newspaper titles should be selected for extraction, which quickly raised challenges related to **data transfer** and the need for a **data sharing platform**, which were key conclusions from this phase of the project. Additionally, once the data has been extracted, the initial exploration and analysis of the extracted data opens up a further series of requirements and challenges, such as **article segmentation** and **OCR quality** or a more advanced features

such of **recognising specific types of articles, e.g. feuilletons** or **automatic recognition of author signatures**. These discoveries indicate the complexity when a cultural heritage institution, such as KBR, embarks on the process of providing their **collections as data.**

Increased understanding of what 'collections as data' means for KBR
During the first reporting phase, the importance of distinguishing between '**cultural heritage datasets**' as the 'raw' or 'unprocessed' cultural heritage data or ***primary resources*** for humanities research prior to analysis (such as the thematic datasets that are extracted in WP2) and '**humanities research data**' which are the ***derived datasets*** resulting from analysing the 'cultural heritage datasets' using digital humanities methods (such as the interdisciplinary research scenarios carried out in *WP4 - Scientific exploitation and valorisation*) became clear. It is the 'cultural heritage datasets' that will be published on the data.kbr.be platform, whereas the 'humanities research data' will be published in a Trusted Digital Repository such as SODHA. Already the project team is starting to capture a list of potential datasets which could be published on the DATA-KBR-BE Platform towards the end of the project.

5. FUTURE PROSPECTS AND PLANNING

WP1: Co-designing Interdisciplinary Research Scenarios. As the design of the research scenarios and the testing with the initial datasets had been carried out in the first reporting period, the focus for this reporting period was on the preparation of the datasets and their initial analysis. In the following reporting period, the analysis of the datasets will continue and resulting outcomes will be valorised (see: *WP4: Scientific exploitation and valorisation*). As needed, additional data or *customised data extractions* will be undertaken (see: *WP2: Preparation of Datasets*).
WP2: Preparation of Datasets. By the end of the second reporting period, it became clear that further data extractions will not be needed immediately as a significant amount of time will be needed for the analysis of the datasets as part of the Interdisciplinary Research Scenarios *(see: WP4: Scientific exploitation and valorisation)*. Indeed, it may be that the researchers will have sufficient data with Data Extraction 1 or will require **customised data extractions** as their research progresses. Additionally, the DATA-KBR-BE Data Extraction 1 has been very useful to start documenting the technical and practical details of the data extraction to prepare for the design of the **sustainable data extraction workflow.** However, it is important to note that the **data extraction workflow** that has been implemented is based on the **KBR's existing data infrastructure.** As a result, when additional data is needed, **the data extraction workflow will need to be updated to work with KBR's new data infrastructure ("Stock")**.

WP3: Data access via data.kbr.be. A key focus of the work in the following reporting period will be to further detail the **technical and functional requirements** for the DATA-KBR-BE platform (see: working document). Not only will this prepare for the implementation of the platform, but it will also help the project team to understand which skills the DATA-KBR-BE Data Scientist will require so that a job vacancy can be prepared for their recruitment.

WP4: Scientific exploitation and valorisation The analysis of DATA-KBR-BE Data Extraction 1 by each of the interdisciplinary research teams will continue in the following reporting period. Once sufficient results have been achieved, it will be important to consider how to valorise this work. As previously agreed, the DATA-KBR-BE team would publish at least *one interdisciplinary, peer-reviewed journal article* for the project and if possible, one peer-reviewed journal article per research scenario. Additionally, other potential ways of valorising the results of the DATA-KBR-BE research scenarios will need to be explored, such as *blog posts*, e.g. to be published on the KBR website. The resulting *derived datasets* will need to be published in a trusted digital repository such as SODHA. The DATA-KBR-BE team could also consider publishing *Jupyter notebooks related to the research scenarios*, for example, on KBR's Github Repository.

WP5: Project Management and Communication. During the following reporting period, the second additional *Annual Report* for the period *15.03.2022-15.3.2023* will need to be prepared, along with *D5.2 Final Report* towards the end of the project. Additionally, it will be important to continue specific meetings with various team members, e.g. research-scenario specific meetings and meetings to discuss the development of the DATA-KBR-BE platform, as well as to organise a meeting of the Scientific Advisory Board ('Follow-up Committee'). The recruitment of the KBR Data Scientist will be a key activity for the next reporting period.

6. FOLLOW-UP COMMITTEE

DATA-KBR-BE Scientific Advisory Board ('Follow-up Committee'), Thursday 10th June 2021

The second DATA-KBR-BE Scientific Advisory Board ('Follow-up Committee') took place online on *Thursday 10th June 2021 (10:00 - 12:00 CET via Zoom).* The agenda for the meeting can be found here. Following a round of introductions, a presentation of the achievements of the project since the last meeting was provided. After which the NewspAIper Demonstrator was presented. The goal of the demonstrator is to improve the searchability of historical newspaper collections through automatic metadata enrichment. In the second half of the meeting, the current questions that the project team are tackling were discussed, such as the design of the *sustainable data extraction workflow*. This was followed by a presentation of the proposed next steps and a discussion.

Similar to the first meeting, the feedback at advice of the Scientific Advisory Board was again both positive and valuable. The presentations and the demonstrators sparked a lively discussion between the DATA-KBR-BE project team and the Scientific Advisory Board around a number of issues.

As during the first meeting, the **legal and ethical considerations** regarding the publication of 'collections as data' were raised. Issues such as copyright, privacy concerns (e.g. related both to Named Entity Recognition, as well as the right to be forgotten), European legislation on open data and Text and Data Mining (TDM) exceptions for research were discussed. The use of the "MyKBR" account, including an online declaration that the datasets will be used for academic use only, to enable secure access to datasets was congratulated. This delegates the responsibility for the use of the content to the researcher. In the future, the possibility of linking the My KBR account to the Belnet Federation to provide a single login for access to the data will be explored. It was suggested that it could be interesting to **organise a round table with historians, cultural heritage experts and legal experts** to discuss these complex issues.

The NewspAIper Demonstrator sparked much interest. In particular the **article similarity** functionality and whether it had been tested with researchers. The Impresso project's **recommender system** was mentioned as a related example. This looks for similar articles based on a number of factors (e.g. similar titles, articles, entities etc.) and can be fine-tuned if it is applied to a specific sub-corpus selected by the researchers. From the feedback from researchers in Impresso, this functionality seems useful for historians. With regard to **Named Entity Linking**, the question of the multilingual strategy was raised. For example, the source newspaper in the NewspAIper Demonstrator is the French-language newspaper *Le Peuple*, however, the named entities that have been extracted are currently linked to an English-language wikipedia page. It was suggested that it may be better to link to a French language page, i.e. the language of the source text. The question of which **viewer** to use was explored. The viewer used in the Ghent Altarpiece project was mentioned (https://iipimage.sourceforge.io) as was the implementation of a IIIF viewer (based on Cantaloupe, see: https://cantaloupe-project.github.io) at the National Library of Luxembourg (see: a Newspaper example: https://persist.lu/ark:70795/q84pj5).

The idea of setting up a **'dataset on demand' service** was discussed. Here the National Library of the Netherlands' Data Services were mentioned as a best-practice example. The service had originally started by providing API access to the data. For in-copyright material, researchers needed to request an **API key** and to **sign an agreement.** However, some challenges in using an API-based service arose, due to the level of digital skills of the humanities researchers who wanted to access the data. A **'harvesting tool'** was set up, so that KBR staff could extract the datasets they needed. This harvesting tool could be compared to a *'dataset on demand' service*. It was noted that when a library puts considerable effort into making such datasets available, it is important that they are published for download by other users. Furthermore the KB has been experimenting with **basic Jupyter notebooks for users to create subsets of the data themselves**. It will be interesting to see how many researchers use them. Currently, KB receives around 2-3 dataset requests per month. Sometimes such requests need to be declined for copyright reasons. The importance of **ensuring that your 'dataset on demand' service is scalable** was stressed**.**

A further topic of discussion was the importance of using **interoperable or uniform data formats** across different projects. For example, NewsEye and Impresso have been working on ensuring the interoperability of data formats used in both projects and would be happy to share this work if it would be of interest. In addition to the use of the International Image Interoperability Framework (IIIF), it was noted that the Impresso team had also been looking into the Distributed Text Services (DTS) specification, which defines an API for working with collections of text as machine readable data. It is also intended to help publishers of text collections make their data Findable, Accessible, Interoperable and Reusable (FAIR). DTS would also be of interest to DATA-KBR-BE. The topic of **article level persistent identifiers (PiDs)** was also discussed. Within the context of Delpher, such PiDs have proved challenging, as automatic article segmentation is not always perfect. This can cause problems when the PiDs need to be updated. Additionally, the **level of granularity of the PiDs** can also be tricky. The challenge of integrating data that has been enriched through a research project such as DATA-KBR-BE was also raised. It was suggested that Linked Open Data (LOD) may offer solutions for this.

Although another meeting of the Scientific Advisory Board was not organised in this reporting period, the first DATA-KBR-BE Annual Report, along with presentation outline of DATA-KBR-BE's latest achievements and next steps was shared with the Follow-up Committee in **December 2021**.

7. VALORISATION ACTIVITIES

7.1 PUBLICATIONS

Our objective is to both valorise the project as a whole, particularly within the Belgian digital cultural heritage community, as well as the scientific outputs of the project, particularly in relation to the research scenarios. Our aim is to prepare at least one interdisciplinary, peer-reviewed journal article as the DATA-KBR-BE team. Potentially for DSH: Digital Scholarship in the Humanities. If possible, we would like to additionally prepare one peer-reviewed journal article per research scenario.

Publications to date

Chambers, S. and Lemmers, F. (2021). Inspiratie: 'Collections as Data'. *META: Tijdschrift voor Bibliotheek en Archief,* 2021(3), 36-37. https://www.vvbad.be/meta/meta-nummer-20213/collections-data

Based on data science work undertaken by Dilawar Ali (IDLab, UGent) and for partial fulfilment of his PhD, in January 2022, Dilawar together with other members of the DATA-KBR-BE team submitted an article entitled: "*Computer Vision and Machine Learning Approaches for Metadata Enrichment to Improve Searchability of Historical Newspaper Collections*" to a special edition of the Journal of Documentation on "Artificial Intelligence for Cultural Heritage Materials", coordinated by the AEOLIAN (Artificial Intelligence for Cultural Organisations) network (see: Call for Papers). The outcome of the review process is expected in Summer 2022.

7.2 PARTICIPATION/ORGANISATION OF SEMINARS (NATIONAL/INTERNATIONAL)

The DATA-KBR-BE project team intends to participate in, and where appropriate organise both national and international workshops, conferences and other events relevant to the project. A list of events that the project team participated in during this first reporting period can be found below:

Chambers, S., Lemmers, F., Pham, T-A., Birkholz, J.M., Jacquet, A., Dillen, W., Ali, D. and Verstockt, S. (2021). *Collections as Data: interdisciplinary experiments with KBR's digitised historical newspapers: a Belgian case study*. Presentation at: DH Benelux 2021: 'The Humanities in a Digital World', 2-4 June 2021, Leiden (Online)

Presented DATA-KBR-BE as part of the *"DH related research projects @ KBR"* coordinated by Julie M. Birkholz at the Research Seminar organised by KU Leuven Artes Libraries Presentation, 8 June 2021.
Ducatteeuw, V., Chambers, S., Birkholz, J. M. and Verbruggen, C. (2021) Studying Collective Action in Belgian Socialist Newspapers with digital approaches (1885-1940). Presentation at 51st Conference of the International Association of Labour History Institutions (IALHI), Organised by the Swiss Social Archive, Zurich, Switzerland, 8-10 September 2021

Chambers, S. and Lemmers, F. (2021). Experimenting with 'Collections as Data' at KBR: an interdisciplinary collaboration. Presentation at: ADOCHS Online Study Day: Image and Data Processing in the Cultural Heritage Sector, 14th September 2021 (Online).

Chambers, S. and Lemmers, F. (2021). Experimenting with Collections as Data: exploring sustainable workflows to facilitate corpus building in the Digital Humanities. Presentation at: International Conference "Cultural Heritage in the Digital Dimension", 20-22 October 2021. (Online)

Chambers, S. (2021). Experimenting with Collections as Data in Europe: a Belgian Experience. University of Texas at Austin, Department of Germanic Studies, DHLunch@GS Fall 2021, 16th November 2021

Chambers, S. (2021). Opening up Collections as Data: the International GLAM Labs Experience. Presentation at: Austrian National Library (ONB) Labs Symposium 2021: Openness, 24 November 2021 (Online).

Birkholz, J. M. and Chambers, S. (2021) Towards a Digital Data Strategy for KBR: a cross-departmental collaboration. Presentation for the Belgian Association for Documentation (ABD-BVD), 25th November 2021

Upcoming: Chambers, S. (2022) DATA-KBR-BE: Data-level access to digitised collections for digital humanities research. Presentation at: CLARIN and Libraries Workshop, National Library of the Netherlands, 9-10 May 2022.

Upcoming: Ali, D., Milleville, K., Van den broeck, A., & Verstockt, S. (2022). NewspAIper : AI-based metadata enrichment of historical newspaper collections. Demonstration. DH Benelux 2022 - ReMIX: Creation and alteration in DH, Esch-sur-Alzette, Luxembourg, 1-3 June 2022.

7.3 SUPPORT TO DECISION MAKING

The DATA-KBR-BE project plays a crucial role within the KBR regarding exploring new ways of providing access to its digitised and born-digital collections for the research community, and in particular digital humanities researchers. In particular, the DATA-KBR-BE team contributes to the development of **KBR's Digital Data Strategy**, as part of which 'Collections as Data' has been identified as one of the core data types and to the development of **KBR's Research Strategy**. Both of these activities directly contribute to the KBR's Action Plan for 2022-2024.

As mentioned in the first reporting period, the DATA-KBR-BE team is investigating opportunities to continue the work of DATA-KBR-BE, both in Belgium, e.g. via other Belspo funding programmes such as ESFRI-

FED and BRAIN or European Funding, e.g. Horizon Europe. During this reporting period (2021-2022), the DATA-KBR-BE project team contributed to two project proposals. Firstly, in September 2021, Julie M. Birkholz in collaboration with Sally Chambers, Thuy-An Pham and Xavier Delor, submitted a project proposal to the Belspo, ESFRI-FED Programme for the ***KBR Virtual Lab: e-infrastructure for facilitating access and research of KBR's collections as data*** as a Belgian contribution to DARIAH, the Digital Research Infrastructure for the Arts and Humanities. The project was selected for financing in December 2021 and is expected to start in 2022. Secondly, in October 2021, KBR - coordinated by Sally Chambers - contributed to the Horizon Europe proposal ***NewsData: Newspapers as Data: sustainable solutions for widening access to Europe's news heritage*** led by the University of La Rochelle (4M€, 12 Partners). The outcome of the proposal is expected in April 2022.

7.4 OTHER

n/a

8. ENCOUNTERED PROBLEMS AND SOLUTIONS

Encountered problems/obstacles, implemented and/or considered solutions, if any.

During the second reporting period DATA-KBR-BE did not encounter any challenges which significantly affected the progress of the project. The main adjustment for the project team was the extension of the duration of the project.

As noted during the first reporting period, the DATA-KBR-BE project team had initiated a discussion with the BELSPO project officer regarding the possibility of extending the duration of the project. This possible extension, without additional funding, would further strengthen the institutional embedding of the outcomes of DATA-KBR-BE into the day-to-day work of the KBR, and to take full advantage of the implementation of the KBR's 'Stock' infrastructure, the first phase of which was expected to be rolled out in late 2021. Furthermore, this would also provide the researchers more time to analyse the extracted datasets and to recruit the KBR Data Scientist. Such an extension would be made possible because the DATA-KBR-BE project coordinator had been appointed in a part-time (50% FTE), rather than a full-time position. In early 2022, the DATA-KBR-BE project team formally applied to BELSPO for an extension of the duration of the project for a period of 24 months to 15.3.2024. This extension was approved in February 2022. The next step will be for the project team to adjust the scheduling of the project in light of the extension.

9. MODIFICATIONS COMPARED TO THE PREVIOUS REPORT

9.1 PERSONNEL

Please note: this section of the report has not been published as it contains personal information.

9.2 COMPOSITION OF THE FOLLOW-UP COMMITTEE

The members of the DATA-KBR-BE Scientific Advisory Board ('Follow-up Committee') are:

- Carlo Blum, National Library of Luxembourg
- Estelle Bunout, Centre for Contemporary History, Potsdam
- Steven Claeyssens, National Library of the Netherlands
- Wout Dillen, University of Borås, Sweden (from 1 June 2021)
- Ann Dooms, Vrije Universiteit Brussel
- Maud Ehrmann, École polytechnique fédérale de Lausanne (EPFL)
- Aurore François, Université catholique de Louvain

10. REMARKS AND SUGGESTIONS

No further remarks.