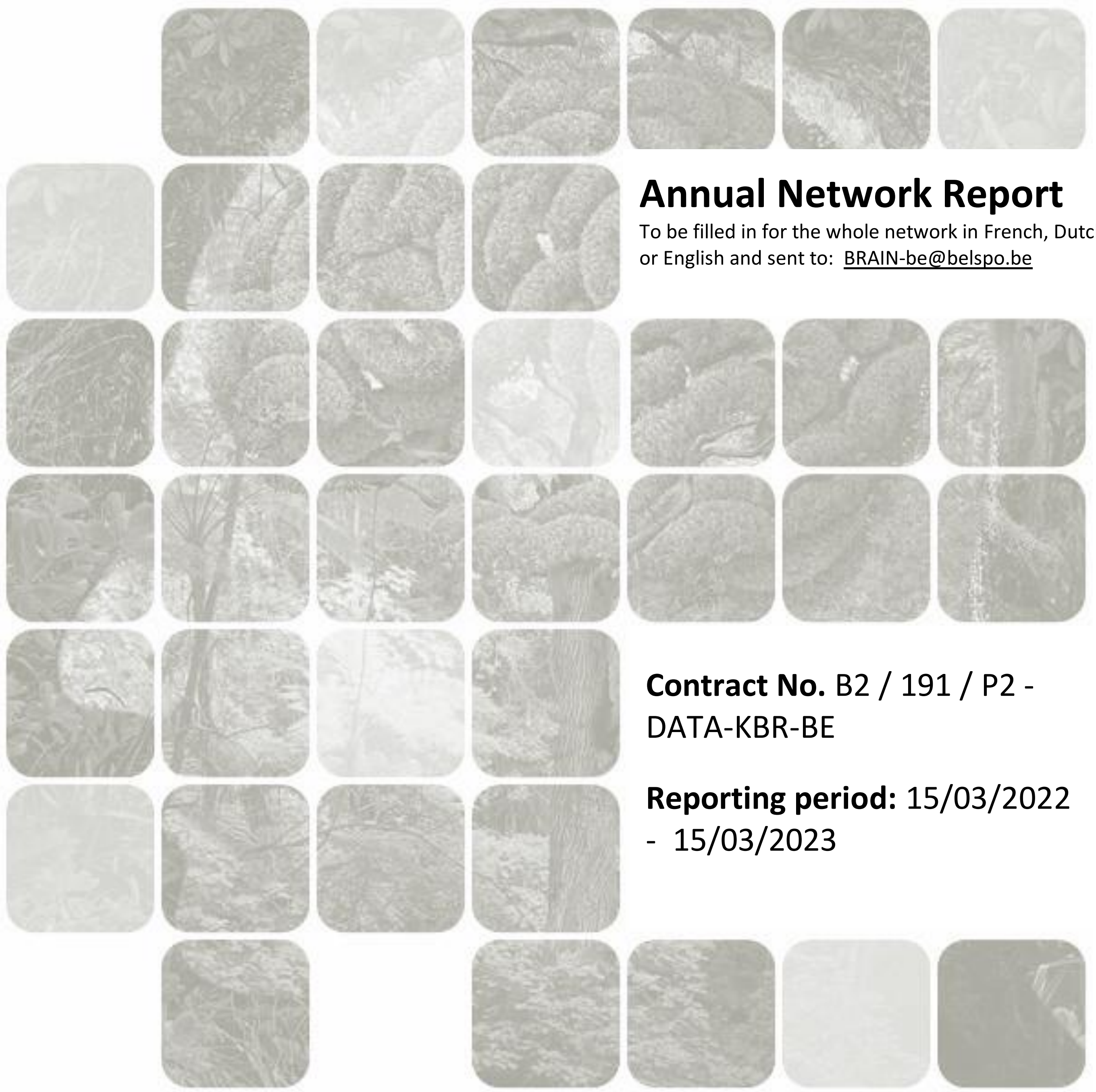
BRAIN-be 2.0

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: 15/03/2022 - 15/03/2023

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

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. 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. Alec Van den broek: Internet Technology and Data Science Lab (IDLab) and Ghent Centre for Digital Humanities, Ghent University
7. Steven Verstockt: Internet Technology and Data Science Lab (IDLab), Ghent University
8. Lamyk Bekius: Antwerp Centre for Digital Humanities and Literary Criticism (ACDC), University of Antwerp
9. Julie M. Birkholz, KBR, Royal Library of Belgium and Ghent Centre for Digital Humanities, Ghent University
10. Brecht Deseure, KBR, Royal Library of Belgium and Université libre de Bruxelles, Sciences de l'information et de la communication Department
11. Tan Lu, KBR, Royal Library of Belgium and DIMA: Digital Mathematics research group, Vrije Universiteit Brussel (VUB)
12. Pieterjan De Potter: Ghent Centre for Digital Humanities, Ghent University
13. Vincent Ducatteeuw: Ghent Centre for Digital Humanities, Ghent University

PROJECT WEBSITE, SOCIAL NETWORKS …

DATA-KBR-BE on the KBR website in **English**: https://www.kbr.be/en/projects/data-kbr-be; **Dutch**: https://www.kbr.be/nl/projecten/data-kbr-be/ and **French**: https://www.kbr.be/fr/projets/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. An additional research scenario led by ULB/KBR (CAMille, Centre for Archives on the Media and Information, ULB-KBR) was added during the project. 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 third reporting period of the project (15.3.2022 - 15.3.2023). It is useful to note that in February 2022 the duration of the DATA-KBR-BE project was extended for a period of 24 months to 15.3.2024. The report also outlines the future prospects and planning for the final reporting period.

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 third reporting period (15.3.2022 - 15.3.2023) 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. In the third 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 as well as the additional research scenario, the History of Belgian Journalism, led by ULB/KBR (CAMille, Centre for Archives on the Media and Information, ULB-KBR). Furthermore, the DATA-KBR-BE project continued to work closely with the KBR Digital Research Lab and the KBR Data Science Lab. The focus of the work in this reporting period has been on further analysis of the extracted data including whether additional data needs to be extracted as the research scenarios evolve. An further issue that arose was that the quality of the textual layer of the digitised newspapers which had been extracted using Optical Character Recognition (OCR) was insufficient for the researchers to undertake the research as they had anticipated. These OCR quality challenges will be investigated further. As the design of the research scenarios are now reaching maturity, the activities related to further data extractions and data quality are 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.

As much of the activities in the previous reporting period focussed on the initial extraction of the data

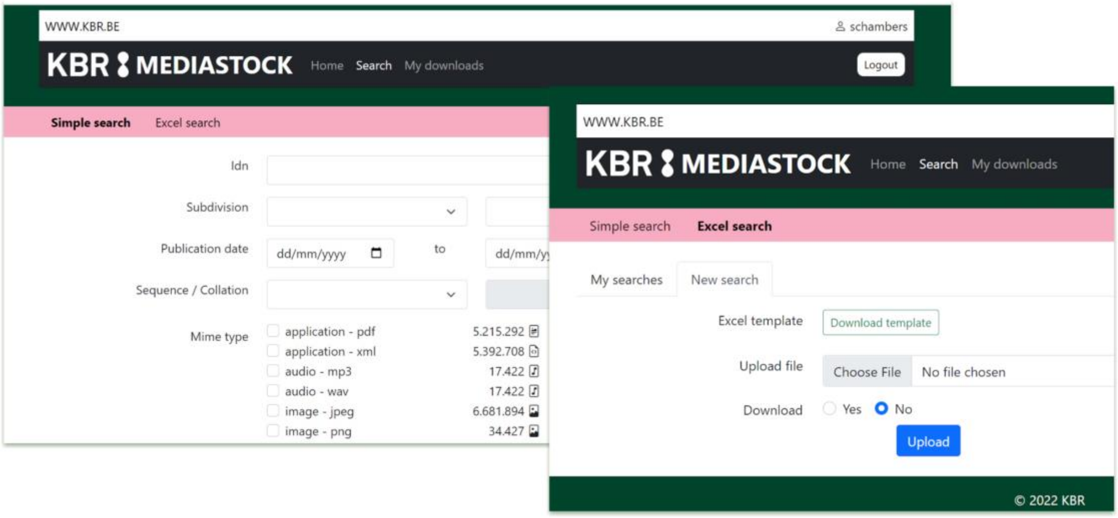
(DATA-KBR-BE Data Extraction 1), in this reporting period (15.3.2022 - 15.3.2023) the focus has been on further analysis of the extracted data including whether additional data needs to be extracted as the research scenarios evolve. The details of these additional **customised data extractions** are included in *WP4: Scientific exploitation and valorisation*, however **an overview of the methods and data extraction tools** used is provided in WP2.

Data Extraction from KBR's Data Infrastructure

As noted in the previous reporting period, the data extracted for Data Extraction 1 provided a substantial amount of data for the research teams to analyse as part of their *interdisciplinary research scenarios.* Unlike anticipated in the proposal writing phase, large data extractions like the first one were less likely to be needed, but rather more **customised data extractions** as the research scenarios developed. For example, in the case of *Collective Action Belgium*, the research team thought that it would be interesting to also explore the articles connected to strikes and demonstrations from the Vooruit from 1912 in addition to 1913 which had been extracted initially. For the *Feuilleton in Belgium* it was noticed that the first feuilleton to appear in *Het Handesblad* for 1885, "De Eed van den Zeeroover", which was published on 3rd January 1885, was actually the second instalment (1e vervolg.) of this feuilleton. The first instalment was published on 31st December 1884. To publish a digital scholarly edition of this particular feuilleton, a data extraction of *Het Handelsblad* for December 1884 would therefore be needed.

In order for KBR staff to be able to extract specific data for researchers based on their evolving research questions, a **data extraction tool** is needed. With the migration to KBR's new data infrastructure, the infrastructural foundations for developing such a tool were in place. In Summer 2022, **KBR's data extraction tool** was launched for use by KBR's ICT and Digitisation Departments.

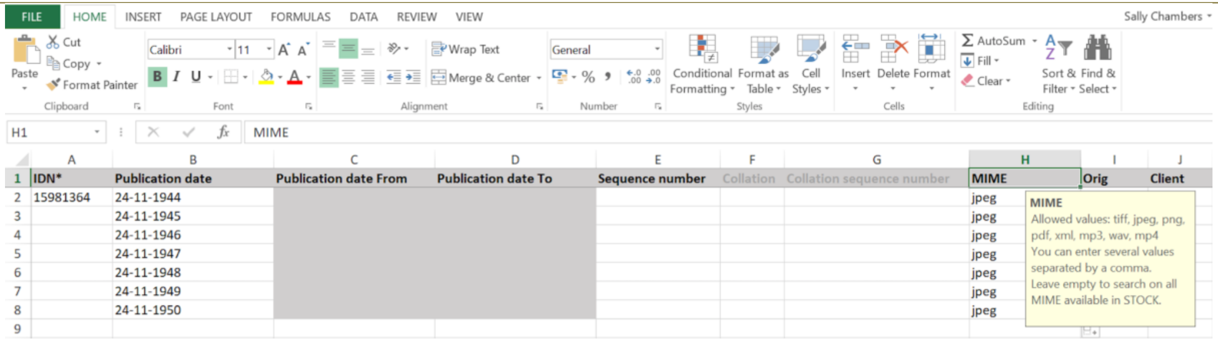
KBR's data extraction tool has two key options (see **Fig. 1** below): a) a *simple search* which is intended for downloading specific files between two publication dates and b) an *excel search* which is intended for requesting the extraction of multiple files or datasets from KBR's data infrastructure for download. It is important to note that it is not currently possible to download files from KBR's publicly available digital platforms such as BelgicaPress.



Mime type	
□ application - pdf	5.215.292
□ application - xml	5.392.708
□ audio - mp3	17.422
□ audio - wav	17.422
□ image - jpeg	6.681.894
□ image - png	34.427

Fig. 1 - The "simple search" and "Excel search" interfaces of KBR's Data Extraction Tool

The **Excel search** is particularly useful for the DATA-KBR-BE project as it can be used to request the additional data needed by the interdisciplinary research scenario teams. The Excel template is downloaded by KBR staff. The bibliographical information related to the required documents is added to the Excel template, e.g. KBR's *IDentification Number (IDN)* of the document required, the *publication dates*, together with the *required file formats* (e.g. pdf, XML, jpeg, TIFF etc.). For example, here is a completed data request template for the JPEG files from *Vooruit* from the 24th November for the years 1944-1950 (see **Fig. 2** below).



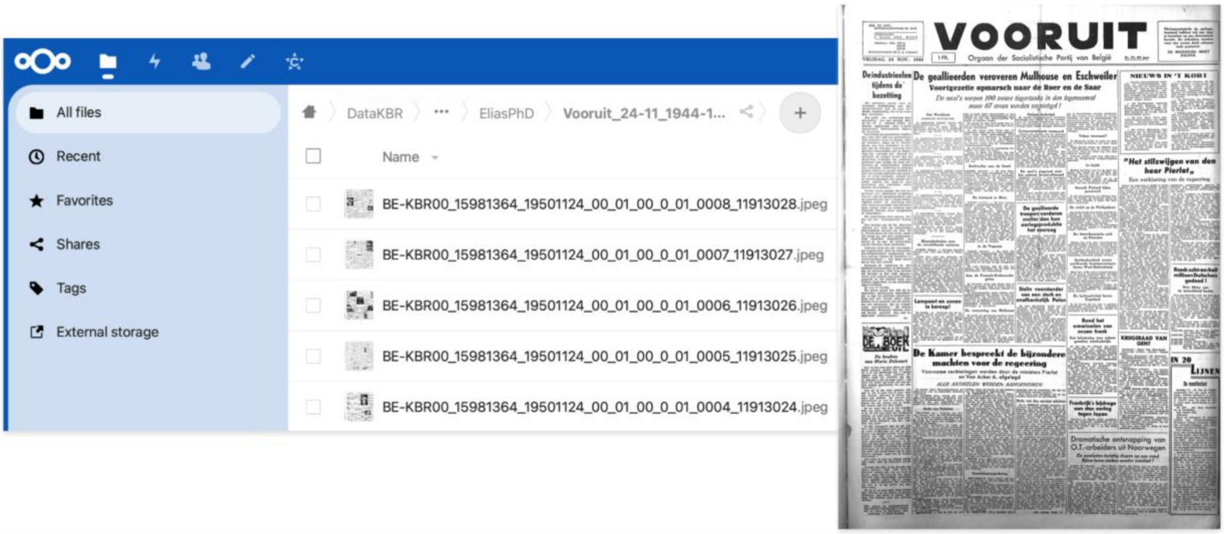
H1 MIME

	A	B	C	D	E	F	G	H	I	J
1	IDN*	Publication date	Publication date From	Publication date To	Sequence number	Collation	Collation sequence number	MIME	Orig	Client
2	15981364	24-11-1944						jpeg		
3		24-11-1945						jpeg		
4		24-11-1946						jpeg		
5		24-11-1947						jpeg		
6		24-11-1948						jpeg		
7		24-11-1949						jpeg		
8		24-11-1950						jpeg		
9										

MIME
Allowed values: tiff, jpeg, png, pdf, xml, mp3, wav, mp4
You can enter several values separated by a comma.
Leave empty to search on all MIME available in STOCK.

Fig. 2 - An example "data request template" for the Vooruit newspaper from 1944-1950

Once the request has been submitted, the data is automatically extracted from KBR's data infrastructure. This request can take several minutes to potentially several hours depending on the volume of the data requested and the number of other requests KBR's data infrastructure needs to process. Once the requested data is ready for download, the KBR member of staff receives an email including a link to the files on KBR's file system. Within the context of the DATA-KBR-BE project, these files are then downloaded from the KBR's file system and uploaded to the **GhentCDH Next Cloud environment** (https://data.ghentcdh.ugent.be) so that they can be accessed by the researcher (see: **Fig. 3**) below.



BE-KBR00_15981364_19501124_00_01_00_0_01_0008_11913028.jpeg
BE-KBR00_15981364_19501124_00_01_00_0_01_0007_11913027.jpeg
BE-KBR00_15981364_19501124_00_01_00_0_01_0006_11913026.jpeg
BE-KBR00_15981364_19501124_00_01_00_0_01_0005_11913025.jpeg
BE-KBR00_15981364_19501124_00_01_00_0_01_0004_11913024.jpeg

Fig. 3. An example data extraction of the Vooruit newspaper from 1944-1950 uploaded to the GhentCDH Next Cloud Environment

Towards a KBR data extraction service for researchers

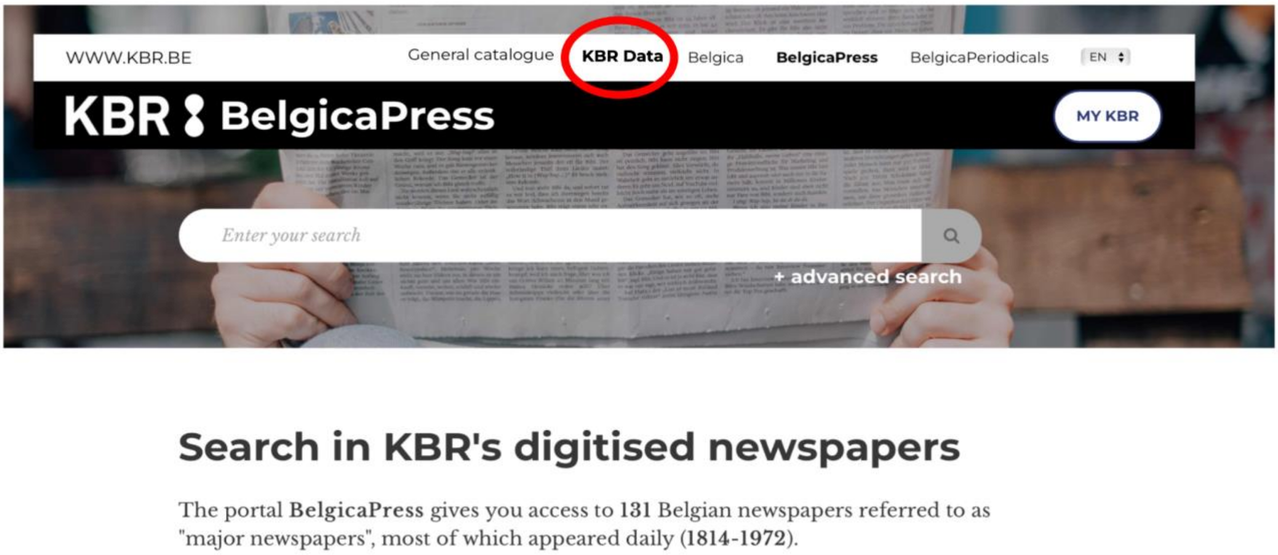
Currently, such requests for data extractions are handled on a case-by-case basis within the framework of the DATA-KBR-BE project. One of the future recommendations of the DATA-KBR-BE project could be for KBR to develop a **data extraction service for researchers**. Such a service would need to: a) identify relevant members of staff who could undertake the data extraction; b) explore how the data could be transferred to the researcher, perhaps Belnet's Filesender Service) could be an option; c) the copyright status of the documents requested and the status of the person requesting the data (e.g. are they using the data for research and education purposes) would need to be taken into account and finally d) a specific webpage, perhaps similar to KBR's Digitisation on Demand service would need to be set up for researchers to submit their requests.

In the future such a form could be published on the DATA-KBR-BE Platform once it is available. Regarding long-term service provision, it could be interesting to explore making the 'Excel search' functionality available for researchers to undertake their own data requests. However, the provision of such a service would require a more detailed investigation, for example, regarding technical security (e.g. prevention of malicious attacks on the service) and performance (e.g. size of the files that could be requested, how many requests could be handled by the servers etc.). From the point of view of the researcher, perhaps an 'excel-like' method of data extraction could be more user-friendly than traditional Application Programming Interfaces (APIs) that may require higher levels of digital literacy.

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.

The focus of the activities in this phase of the project has been on the design of the DATA-KBR-BE platform. Regular meetings (every 4 - 6 weeks) between DATA-KBR-BE Project Coordinator Sally Chambers and Thuy-An Pham from KBR's ICT department took place to iteratively develop the requirements for the DATA-KBR-BE Platform.

Firstly, the DATA-KBR-BE Platform is intended to be a sustainable platform to provide access to KBR's 'Collections as Data'. In this initial phase, the focus is on **curated datasets.** Rather than a standalone project website, it is essential that the DATA-KBR-BE Platform is **fully integrated into the KBR website**, alongside KBR's core digital services such as Belgica, BelgicaPress, BelgicaPeriodicals and the General Catalogue. This integration would be undertaken in close collaboration with KBR's Communications Team. An initial mock-up of how the DATA-KBR-BE platform could be integrated into KBR's website is provided in **Fig. 4** below.



Search in KBR's digitised newspapers

The portal **BelgicaPress** gives you access to 131 Belgian newspapers referred to as "major newspapers", most of which appeared daily (1814-1972).

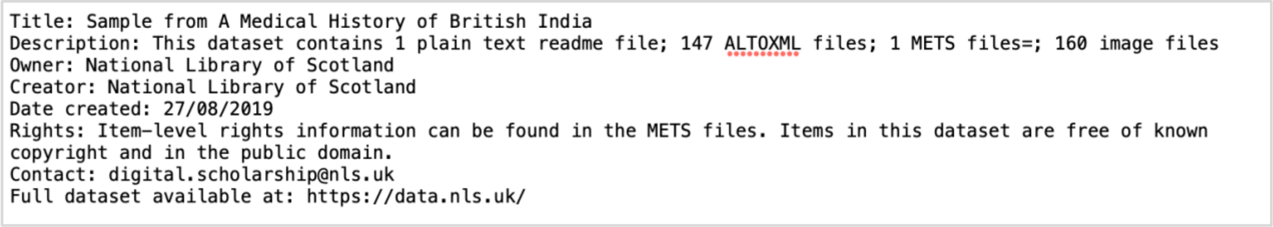
Fig. 4. A mock-up of how the DATA-KBR-BE Platform could be integrated into KBR's website

As the DATA-KBR-BE Platform is intended to be a long-term initiative, beyond the end of the DATA-KBR-BE project itself, a number of phases can already be anticipated. For example, **Phase 1**, which will be developed during the DATA-KBR-BE project, is anticipated as a simple, user-friendly website, where **curated datasets** based on KBR's digitised and born-digital collections will be published. This phase is inspired by existing data platforms such as the National Library of Luxembourg's Open Data Platform and the National Library of Scotland's Data Foundry. In **Phase 2**, which could be funded by a follow-up project, a **searchable repository** of curated datasets could be anticipated. For this phase, sources of inspiration include the British Library's Research Repository, where the British Library's Collection Datasets have been published or the Digital Library of the Caribbean (DLoc)'s DLoc as Data repository. For **Phase 3**, a complementary project to DATA-KBR-BE could focus on providing computational access to KBR's digitised and born-digital collections via Application Programming Interfaces (APIs) or 'data on demand' services. As reported in the previous DATA-KBR-BE Annual Report, 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 funding in December 2021. The inspiration for this phase is, for example, the National Library of the Netherlands'

Data Services. Following the establishment of these phases, it was therefore decided to focus on the **technical and functional requirements** for *Phase 1 of the DATA-KBR-BE platform*; a simple, user-friendly website for publishing **curated datasets for download.**

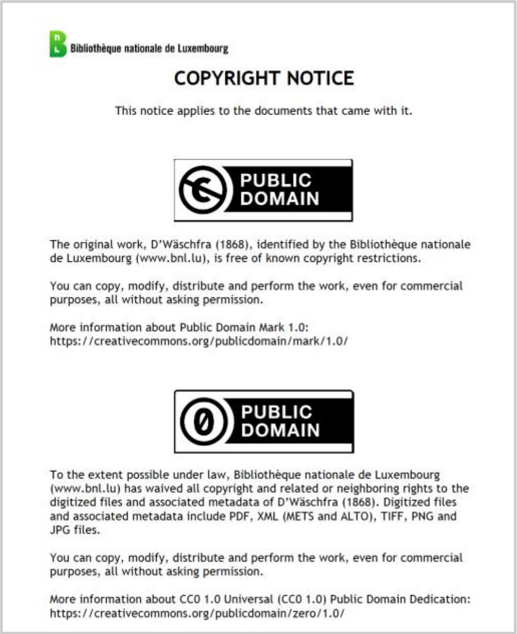
The focus then turned to the **requirements for the datasets** themselves. These requirements included a range of topics such as: a) *documenting the datasets*, e.g. via a README file, b) *data citations* and related *Persistent Identifiers* (PIDs), c) *rights, licensing* and *terms of use* and d) *versioning of datasets*. Each of these topics is explored in further detail below.

Firstly, to document or describe the datasets, several libraries provide a text file accompanying the dataset, such as: a **README file** containing information about the dataset. Examples include the README file from the National Library of Scotland's Data Foundry (**see: Fig. 5** below) or the **Copyright Notice** provided by the National Library of Luxembourg accompanying the datasets published on their Open Data Platform (see: **Fig. 5** below).



```
Title: Sample from A Medical History of British India
Description: This dataset contains 1 plain text readme file; 147 ALTOXML files; 1 METS files=; 160 image files
Owner: National Library of Scotland
Creator: National Library of Scotland
Date created: 27/08/2019
Rights: Item-level rights information can be found in the METS files. Items in this dataset are free of known
copyright and in the public domain.
Contact: digital.scholarship@nls.uk
Full dataset available at: https://data.nls.uk/
```

Fig. 5. An example README file from the National Library of Scotland's Data Foundry



Bibliothèque nationale de Luxembourg

COPYRIGHT NOTICE

This notice applies to the documents that came with it.

PUBLIC DOMAIN

The original work, D'Wäschfra (1868), identified by the Bibliothèque nationale de Luxembourg (www.bnl.lu), is free of known copyright restrictions.

You can copy, modify, distribute and perform the work, even for commercial purposes, all without asking permission.

More information about Public Domain Mark 1.0:
https://creativecommons.org/publicdomain/mark/1.0/

To the extent possible under law, Bibliothèque nationale de Luxembourg (www.bnl.lu) has waived all copyright and related or neighboring rights to the digitized files and associated metadata of D'Wäschfra (1868). Digitized files and associated metadata include PDF, XML (METS and ALTO), TIFF, PNG and JPG files.

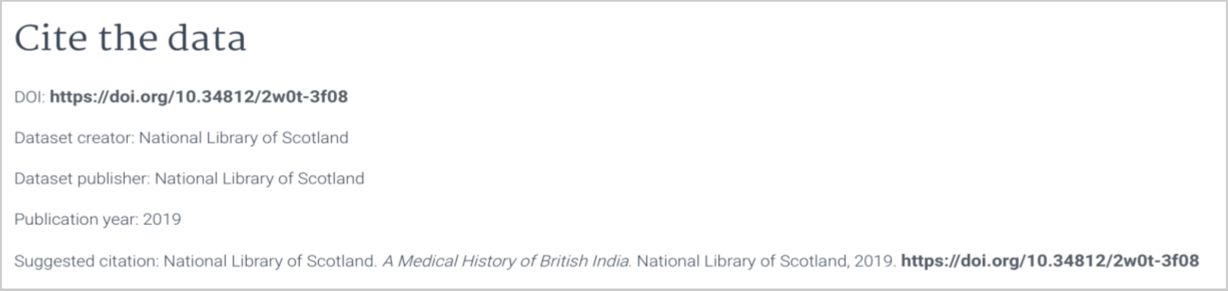
You can copy, modify, distribute and perform the work, even for commercial purposes, all without asking permission.

More information about CC0 1.0 Universal (CC0 1.0) Public Domain Dedication:
https://creativecommons.org/publicdomain/zero/1.0/

Fig. 6. An example Copyright Notice from the National Library of Luxembourg's Open Data Platform

From previous discussions within the DATA-KBR-BE team, requirements for the README file include: that the metadata contained in the file could ideally be automatically generated from KBR's Library Management System Syracuse; the information to **rights, licensing and terms of use** could be automatically provided from **KBR's Rights Management System** which is currently in development. At this stage, the DATA-KBR-BE project focuses on datasets including newspapers that have been published *before 1918* as these resources are freely accessible to all (see: KBR's Copyright page). However, in future phases of the project it may be interesting to investigate how access to *in-copyright datasets* could be provided.

A further area to explore was how to provide a **data citation** for the curated datasets, including a **Persistent Identifier** (PID) for each dataset. Data citations are important for the datasets so that they can be cited in publications where the data has been used. This enables the researcher to provide a clear reference to the dataset, as well as providing a metric for the library to measure the use of the dataset. An example data citation is from National Library of Scotland's Data Foundry, (see **Fig. 7**), below.



Cite the data

DOI: **https://doi.org/10.34812/2w0t-3f08**

Dataset creator: National Library of Scotland

Dataset publisher: National Library of Scotland

Publication year: 2019

Suggested citation: National Library of Scotland. *A Medical History of British India*. National Library of Scotland, 2019. **https://doi.org/10.34812/2w0t-3f08**

Fig. 7. An example Data Citation from the National Library of Scotland's Data Foundry, see: https://doi.org/10.34812/2w0t-3f08

Another challenging area is the assignment of a Persistent Identifier (PID) to the dataset. Currently, KBR uses an internal PID system. For example, each bibliographic record in KBR's catalogue has a PID or 'permalink', e.g. for the Belgian Dutch language newspaper "Vooruit", the permalink to the catalogue record is: https://opac.kbr.be/LIBRARY/doc/SYRACUSE/15981364. Additionally, there is a PID for each digitised edition of the newspaper, for example, the PID for Vooruit from 10th April 1913 is: https://uurl.kbr.be/1545224. At this stage, the creation of 'Collections as Data' datasets is a new activity for KBR. This would require the KBR's PID system to be adapted to be able to assign PIDs at the level of the dataset or for the PID for the datasets to be assigned by another service provider, such as the DOI assigned by the Belgian federal *Data Archive for Social Sciences and the Digital Humanities*, *SODHA*. Potentially, it could be interesting to also include the individual PIDs for various documents that are included in the dataset. **For example, a dataset of the digitised editions of Vooruit from 10th - 15th April 1913 could include:**

- **Permalink of the bibliographic record of the Vooruit**: https://opac.kbr.be/LIBRARY/doc/SYRACUSE/15981364
- **PIDs of each of editions of the Vooruit included in the dataset, e.g.**
    - 10th April 1913: https://uurl.kbr.be/1545224
    - 11th April 1913: https://uurl.kbr.be/1545225
    - 12th April 1913: https://uurl.kbr.be/1545226
    - …
- **Permalink of the dataset itself**: to be created.

A more advanced topic related to datasets is the need for **versioning**. For example, if a researcher would like to add new resources into their dataset, e.g. they would like to compare the newspaper "Vooruit" with "Le Peuple". For example, the Open Science Repository, Zenodo offers functionality for versioning datasets, see: https://help.zenodo.org/#versioning. The Zenodo example could provide inspiration for the versioning of DATA-KBR-BE datasets in the future.

An initial list of "Collections as Data" datasets to be published on the DATA-KBR-BE Platform

As reported previously, 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*) became clearer. Initially, it was thought that it would be the 'cultural heritage datasets' that would be published on the data.kbr.be platform, whereas the 'humanities research data' will be published in a Trusted Digital Repository such as SODHA. However, with the further development of the DATA-KBR-BE platform the line between these two types of datasets is becoming increasingly blurred.

In the original project proposal, the DATA-KBR-BE project team committed to co-curating and publishing at **least three Open Science datasets** on the data.kbr.be platform. It is anticipated that these three datasets are likely to be related to the two original interdisciplinary research scenarios: 1) Collective Action Belgium, led by GhentCDH and 2) Feuilleton in Belgium, led by ACDC as well as a dataset related to the additional research scenario, the History of Belgian Journalism. However, when discussing the data.kbr.be platform with KBR colleagues, the potential list of "Collections as Data" datasets already started to increase.

Within WP3 of DATA-KBR-BE, the deliverable: *D3.3 Digital Asset Registry: inventory of KBR's Digital Collections* is intended to create an inventory of KBR's collections which could be published as 'Collections as Data'. Here, the National Library of Scotland's Open Data Publication Plan, could provide inspiration. At this stage it is not clear if it will be possible to publish more than three datasets, however, *D3.3* is intended to facilitate the dataset publication process both during and after the end of the project. KBR's Digital Asset Registry should be a living document which is updated regularly, e.g. every 6-12 months.

An initial overview of the types of datasets that could be published on the DATA-KBR-BE Platform is provided below:

1. **Thematic datasets**

Datasets related to the interdisciplinary research scenarios within DATA-KBR-BE: a) Collective Action Belgium, b) Feuilleton in Belgium and History of Belgian Journalism. Datasets related to research projects, e.g. funded by BELPSO or other funding providers: a) Artpresse: an intermedial study of Belgian art as a network structured as seen through the lens of the mass media magazines in the interbellum years, b) Google Books @ KBR: digitisation of Belgium's rich cultural and historical heritage, c) IMPRESS: beyond the philosophical conflict: religion and liberalism in the Belgian medical press from 1840 to 1914 and d) Photolit: Belgian photo novel: local reuse of a European cultural practice.

2. **Collections as Data datasets**

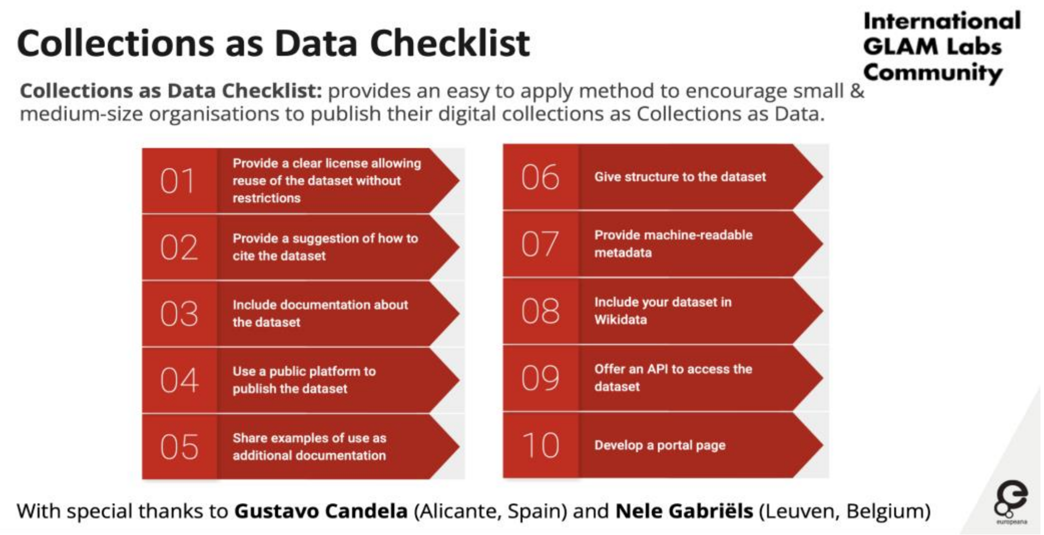
Collections as Data datasets can include any data that is related to KBR's Collections. For example, as part of KBR's emerging Digital Data Strategy, nine "**data types**", related to KBR's collections have been identified: *Bibliographic data; Authority data; Heritage data; Digital data; Collections as data; Born-digital data; Electronic resources; Research data* and *Lab data (code).* It could be interesting to consider publishing one dataset for each data type on the DATA-KBR-BE platform. Example KBR's Collections as Data datasets include:

- **Digital Data:** sample datasets related to KBR's Collections that have been digitised and made available as part of KBR's digital library, Belgica: such as: Manuscripts and Rare Books; Maps and Plans or Music
- **Bibliographic / Authority Data**: e.g. a dataset including **Belgian authors from the 19th Century**. Additionally, links to APIs (e.g. Z39.50 or OAI-PMH) providing access to KBR bibliographic data could also be provided.
- **Born-digital Data:** e.g. a dataset related to **25 years of the KBR website** (see: https://web.archive.org/web/19980129080139/http://www.kbr.be/)

Inspired by the National Library of Scotland's Data Foundry, see for example, A Medical History of British India, each dataset published on the data.kbr.be platform could have a specific webpage providing further information about the dataset in a user-friendly and visually-attractive way. A '**dataset template**' could be set-up for the DATA-KBR-BE platform so that adding new datasets would not be overly time-consuming.

A framework for the development of the DATA-KBR-BE platform

In Summer 2022, the International Galleries, Libraries, Archives and Museums (GLAM) Labs Community started to explore the challenges that cultural heritage institutions are experiencing when attempting to publish their collections as data. This activity led to the emergence of the "**Checklist to publish Collections as Data in GLAM Institutions**", which is intended as an easy to apply method to encourage small and medium sized organisations to publish their digital collections as *Collections as Data*. The checklist (see: **Fig. 8 below**) includes 10 items of key issues that cultural heritage institutions need to consider when publishing their collections as data.



Collections as Data Checklist

International GLAM Labs Community

Collections as Data Checklist: provides an easy to apply method to encourage small & medium-size organisations to publish their digital collections as Collections as Data.

01 Provide a clear license allowing reuse of the dataset without restrictions

02 Provide a suggestion of how to cite the dataset

03 Include documentation about the dataset

04 Use a public platform to publish the dataset

05 Share examples of use as additional documentation

06 Give structure to the dataset

07 Provide machine-readable metadata

08 Include your dataset in Wikidata

09 Offer an API to access the dataset

10 Develop a portal page

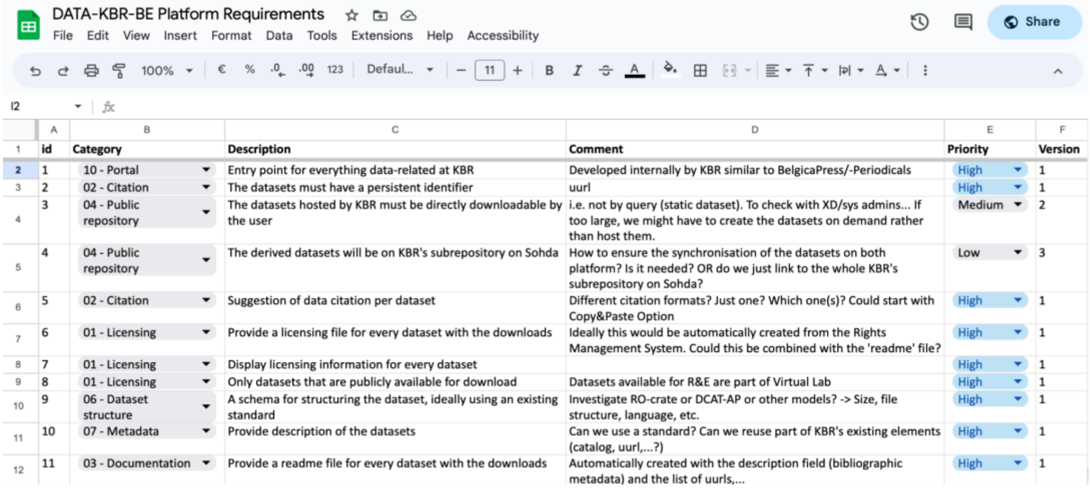
With special thanks to **Gustavo Candela** (Alicante, Spain) and **Nele Gabriëls** (Leuven, Belgium)

Fig. 8 - International GLAM Labs Community's Collections as Data Checklist

The emergence of this checklist arrived at an ideal moment for the DATA-KBR-BE project team as it provided a very useful framework to help structure the development of the functional and technical requirements for the data.kbr.be platform. To raise awareness of the checklist within the wider cultural heritage community, a webinar was organised in October 2022, co-organised by the DATA-KBR-BE project team. The aim of the webinar was to invite the cultural heritage community to provide feedback on the checklist from the viewpoint of its application within their own institution.

To prepare for the webinar, an initial analysis of the checklist was undertaken to assess which checklist items were most relevant for the DATA-KBR-BE project. Initially, *item 10: develop a portal page* and *item 6: give structure to the dataset* were identified as the most relevant items for the project team, as our aim was to develop the DATA-KBR-BE platform and we would also need to understand how to structure the datasets that will be published there. However, it soon became relevant that many, if not all, the checklist items would support the development of the DATA-KBR-BE platform. For example, *item 2: provide a suggestion of how to cite your dataset* and *item 1: provide a licence allowing reuse of the dataset* were quickly seen as essential.

To use the checklist more systematically, a collaborative spreadsheet was designed to capture each of the functional and technical requirements for the DATA-KBR-BE Platform, as shown in **Fig. 9**, below. Column B, is used for categorising each of the requirements based on the checklist list, e.g. *Requirement 1: Entry point for everything data-related at KBR*, has been categorised in relation to checklist *item 10: develop a portal page*. This approach enabled us to: a) group the requirements by category, b) to ensure that our requirements analysis was as exhaustive as possible by considering each of the checklist items and c) to provide feedback to the International GLAM Labs Community to further improve the checklist.



DATA-KBR-BE Platform Requirements

	A	B	C	D	E	F
1	id	Category	Description	Comment	Priority	Version
2	1	10 - Portal	Entry point for everything data-related at KBR	Developed internally by KBR similar to BelgicaPress/-Periodicals	High	1
3	2	02 - Citation	The datasets must have a persistent identifier	uurl	High	1
4	3	04 - Public repository	The datasets hosted by KBR must be directly downloadable by the user	i.e. not by query (static dataset). To check with XD/sys admins... If too large, we might have to create the datasets on demand rather than host them.	Medium	2
5	4	04 - Public repository	The derived datasets will be on KBR's subrepository on Sohda	How to ensure the synchronisation of the datasets on both platform? Is it needed? OR do we just link to the whole KBR's subrepository on Sohda?	Low	3
6	5	02 - Citation	Suggestion of data citation per dataset	Different citation formats? Just one? Which one(s)? Could start with Copy&Paste Option	High	1
7	6	01 - Licensing	Provide a licensing file for every dataset with the downloads	Ideally this would be automatically created from the Rights Management System. Could this be combined with the 'readme' file?	High	1
8	7	01 - Licensing	Display licensing information for every dataset		High	1
9	8	01 - Licensing	Only datasets that are publicly available for download	Datasets available for R&E are part of Virtual Lab	High	1
10	9	06 - Dataset structure	A schema for structuring the dataset, ideally using an existing standard	Investigate RO-crate or DCAT-AP or other models? -> Size, file structure, language, etc.	High	1
11	10	07 - Metadata	Provide description of the datasets	Can we use a standard? Can we reuse part of KBR's existing elements (catalog, uurl,...?)	High	1
12	11	03 - Documentation	Provide a readme file for every dataset with the downloads	Automatically created with the description field (bibliographic metadata) and the list of uurls,...	High	1

Fig. 9. A collaborative spreadsheet for capturing the technical and functional requirements for the DATA-KBR-BE Platform

In the coming months, the DATA-KBR-BE project team will continue to use the checklist to structure the functional and technical requirements for the data.kbr.be platform.

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. In the third phase of the project, the majority of the work has been focused on the *Design of the data.kbr.be Platform*, however, further advances have been made on the analysis of the data which was extracted for the *interdisciplinary research scenarios* as part of Data Extraction 1. Additionally, some challenges related to the quality of the text which has been automatically extracted using Optical Character Recognition (OCR) have been raised which require further investigation. Currently, it is not anticipated that further large data extractions will be required for the *interdisciplinary research scenarios*, but more likely smaller, more specific, **customised data extractions** as the analysis of the data progresses.

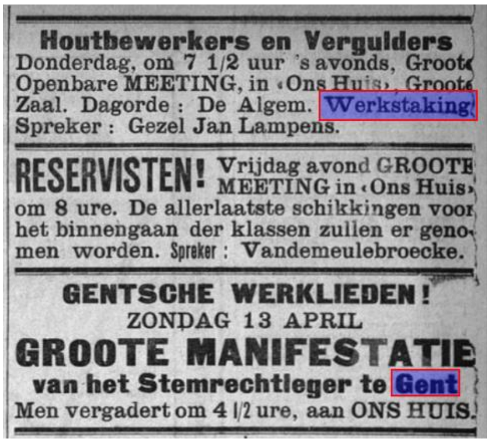
Exploitation of Research Results: Collective Action Belgium

Following the initial analysis of the data extracted in Data Extraction 1, the focus of this reporting period was to further analyse the extracted data to prepare it for scientific exploitation.

For **Collective Action Belgium**, the first step was to build a dataset of all the articles related to "collective action" e.g. strikes, demonstrations and protests. The research team is particularly interested in the strikes that took place in Ghent. The identification 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. For this phase of analysis, it was agreed to focus on the test dataset, the newspaper, the "Vooruit" from April 1913, which was available to the research team on the GhentCDH NextCloud environment.

To facilitate the searching of the test dataset to find the relevant articles related to collective action, Dilawar Ali created a *Jupyter Notebook*. However, to initially explore the dataset, it was decided to undertake the keyword searches using the **advanced search interface of the BelgicaPress.** From the initial list of keywords created by the research team, it was agreed to select three "search phases" for the initial exploration. Firstly, a single keyword "werkstaking". Secondly two keywords "werkstaking" and "gent". Thirdly, "ons huis" or in full "Ons Huis, Socialistische Werkersvereenigingen", a building in Ghent where activities of the Socialist Workers Associations were organised.

Using the *advanced search interface of the BelgicaPress*, for the first single keyword search for "werkstaking", in the *Vooruit* newspaper from April 1913, a total of 140 pages were identified which contained the keyword "werkstaking". The most occurrences appeared in the editions of *10th April, 19th April* and *20th April 1913* with 7 occurrences each. A challenge of the BelgicaPress search interface is that the keywords are provided at page level, not at article level. As a result of the extraction of the data, it is also possible to search directly in the pdfs for the various editions of the newspaper. For example, the Vooruit of 10th April 1913 contains 26 occurrences of "werkstaking", as opposed to 7.



Houtbewerkers en Vergulders
Donderdag, om 7 1/2 uur 's avonds, Groote Openbare MEETING, in ‹Ons Huis›, Groote Zaal. Dagorde: De Algem. Werkstaking. Spreker: Gezel Jan Lampens.

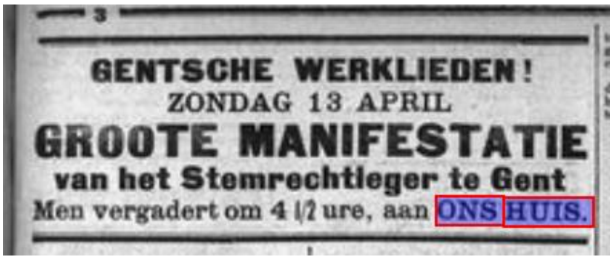
RESERVISTEN! Vrijdag avond GROOTE MEETING in ‹Ons Huis› om 8 ure. De allerlaatste schikkingen voor het binnengaan der klassen zullen er genomen worden. Spreker: Vandemeulebroecke.

GENTSCHE WERKLIEDEN!
ZONDAG 13 APRIL
GROOTE MANIFESTATIE
van het Stemrechtleger te Gent
Men vergadert om 4 1/2 ure, aan ONS HUIS.

For the second keyword search, including two keywords "werkstaking" (26 occurrences) and "gent" (34 occurrences), it was possible to search for both keywords simultaneously using the BelgicaPress advanced search interface. This provides results if the two keywords appear on the same page (see **Fig. 10** to the left). However, it is not currently possible to undertake a **proximity search**, which shows when "werkstaking" and "gent" are in the same article or located within 5 words of each other, which would be more useful for the researcher.

Fig 10. **"Werkstaking" and "Gent" in the Vooruit from 11th April 1913:** https://uurl.kbr.be/1545225

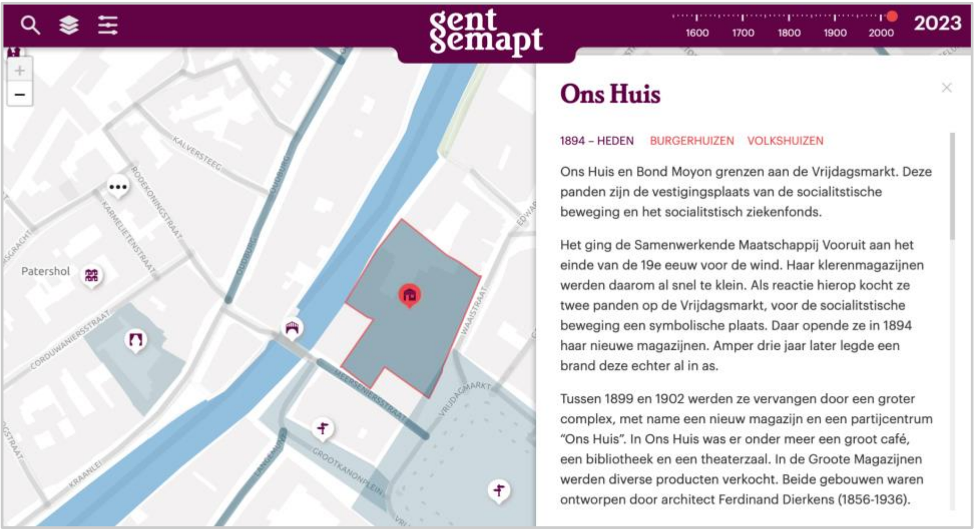
Finally, when searching for a *specific location* and in this case "ons huis", using the "*exact phrase*" search in April 1913 via the *BelgicaPress advanced search interface*, results were found on 67 pages, for example in the edition of 10th April 1913 (see **Fig 11.** to the right).
Fig 11 "ons huis" in the Vooruit from 11th April 1913: https://uurl.kbr.be/1545224



3

GENTSCHE WERKLIEDEN!
ZONDAG 13 APRIL
GROOTE MANIFESTATIE
van het Stemrechtleger te Gent
Men vergadert om 4 1/2 ure, aan ONS HUIS.

The **historical geo-localisation of the collective actions** that took place as reported in the digitised historical newspapers is of particular interest to the GhentCDH team. For example, it would be interesting to geo-locate the locations of the various strikes and demonstrations on a historical map. Such geo-location could be undertaken in collaboration with GhentCDH's GentGemapt project. For example, **Fig. 12** below shows "Ons Huis" geo-located and contextualised on the GentGemapt platform.



Ons Huis

1894 – HEDEN BURGERHUIZEN VOLKSHUIZEN

Ons Huis en Bond Moyon grenzen aan de Vrijdagsmarkt. Deze panden zijn de vestigingsplaats van de socialitstische beweging en het socialitstisch ziekenfonds.

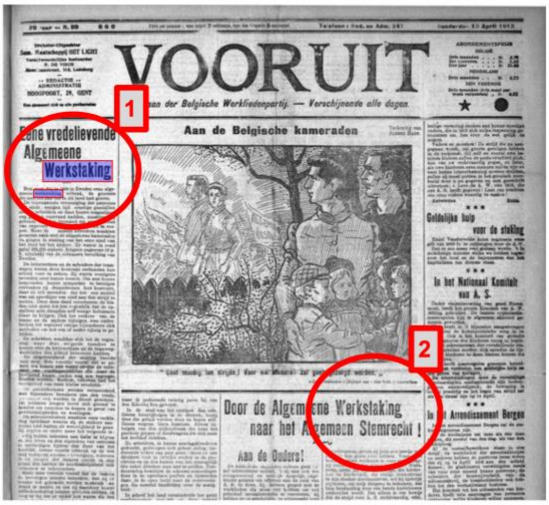
Het ging de Samenwerkende Maatschappij Vooruit aan het einde van de 19e eeuw voor de wind. Haar klerenmagazijnen werden daarom al snel te klein. Als reactie hierop kocht ze twee panden op de Vrijdagsmarkt, voor de socialitstische beweging een symbolische plaats. Daar opende ze in 1894 haar nieuwe magazijnen. Amper drie jaar later legde een brand deze echter al in as.

Tussen 1899 en 1902 werden ze vervangen door een groter complex, met name een nieuw magazijn en een partijcentrum "Ons Huis". In Ons Huis was er onder meer een groot café, een bibliotheek en een theaterzaal. In de Groote Magazijnen werden diverse producten verkocht. Beide gebouwen waren ontworpen door architect Ferdinand Dierkens (1856-1936).

Fig. 12 "Ons Huis", geo-located on the GentGemapt platform: https://kaart.gentgemapt.be/plaats/3347

Currently there is no article-level indexing of the digitised historical newspapers in BelgicaPress. From this initial phase of investigation, even the *indexing of the titles of the articles* could be an additional enrichment to the BelgicaPress search functionality, as well as the inclusion of *proximity searching*.

Additionally, the researchers expressed some concerns regarding the *quality of the Optical Character Recognition (OCR)* of the dataset from April 1913, which would lead to non-identification of relevant articles for the thematic dataset.


VOORUIT
1
2
Eene vredelievende Algemeene Werkstaking
Aan de Belgische kameraden
Door de Algemeene Werkstaking naar het Algemeen Stemrecht!
Aan de Ouders!

For example, it was possible to see visually that "werkstaking" was not always identified in the OCR, see for example, the Vooruit from 10th April 1913 (see **Fig. 13**) to the left. Here you can see at point **[1]** that "werkstaking" is correctly identified, but at point **[2]** that "werkstaking" has not been recognised by the OCR. In this way, when building the corpus, this particular article would not have been included in the dataset.

Fig. 13 - Vooruit from 10th April 1913
https://uurl.kbr.be/1545224

The possibility of reOCRing the test dataset will be explored in the final phase of the project. However, for this exercise to be useful, it would be important to identify additional examples of where poor quality OCR is preventing the research team from undertaking their research, as well investigating a method to measure the quality of the original OCR, which can be used as a benchmark for measuring how much the OCR has improved following the reOCRing process. If the OCR has improved sufficiently, it would be interesting to explore how this improved OCR could be ingested back into BelgicaPress for the benefit of all users.

The next step is to start investigating which data could be included in the **thematic dataset** to be published on the DATA-KBR-BE platform. For example, the extracted data for the *Vooruit* from 1913, which includes the pdfs of the various editions; the pdfs, tiffs and JPEGs of each page, as well as the related XML-ALTO and METS-ALTO files. All of these files for April 1913 total 3.5 GB of data.

Collective Action Belgium internship: Léon Castelein, Masters Student in History, Ghent University

In the context of the *Collective Action Belgium* research scenario, Léon Castelein, a Masters' student in History from Ghent University, is undertaking an internship (14.11.2022 - 1.9.2023) at KBR as part of the DATA-KBR-BE project. For his Masters' thesis, Léon is undertaking a comparison between the general strikes related to the introduction of universal suffrage which took place in Belgium in 1902 and 1913. For this research, Léon is primarily using the Dutch language Belgian daily newspapers the "Vooruit: Socialistisch dagblad" and "Het Volk: Antisocialistisch dagblad".

The main goal of Léon's internship is to gain a better understanding of how a national library such as KBR digitises their historical newspaper collection and in particular how the digitised files could be sustainably extracted and shared with researchers. During his internship, Léon will gain practical insight into: a) KBR's digitisation workflows, b) which tools could be used to extract the necessary data, c) the legal framework for this, d) how data can be shared securely and in a user-friendly way and e) how this fits in the context of developing a data management plan (DMP) and good practices on research data management.

As an important first step, Léon needed to create a dataset for his research. Within the context of the DATA-KBR-BE project both *Vooruit* and *Het Volk* for 1913 had already been extracted as part of Data Extraction 1 and transferred to the GhentCDH Next Cloud environment. Léon could use these datasets for his research. However, for 1902, the data extraction was more complex.

For the *Vooruit* for 1902, the 10th - 19th April is digitally available via BelgicaPress, however, the edition for 20 April 1902 is not available. To create his dataset, Léon was able to use KBR's data extraction tool to extract the Vooruit for 1902, in collaboration with KBR staff. However, the missing edition from 20th April 1902 has also been digitised by the AMSAB Institute of Social History and has been made available online via the Ghent University Library for Ghent University students and staff. The 20th April 1902 was downloaded from Ghent University Library and made available via the GhentCDH Next Cloud environment. In the next phase of the project, it will be investigated as to whether it would be possible to add the missing edition to BelgicaPress.

For *Het Volk* for 10th - 20th April 1902, the situation is even more complex, as KBR's holdings are more sparse for this newspaper title and do not include 1902. For this Léon needed to search further afield and eventually found Het Volk: katholiek social dagblad for 1902 available on *microfilm* (but not digitised) at KADOC Documentation and Research Centre on Religion Culture and Society.

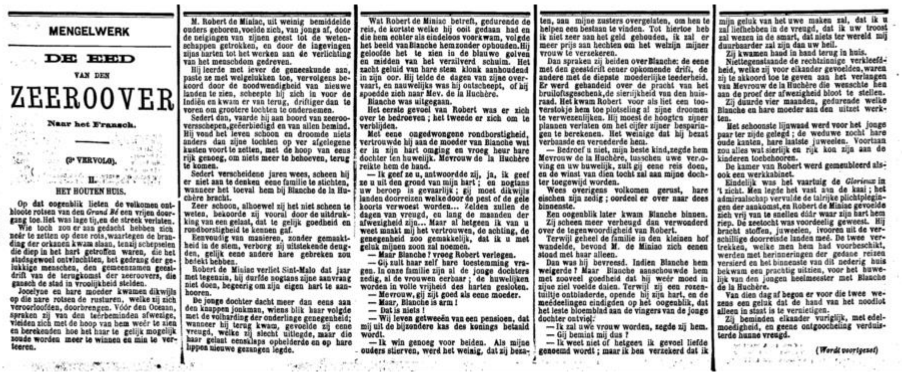
A second important step was for Léon to create a **data management plan (DMP)** for his dataset. To do this, he used the DMP Online, provided by Belnet. KBR staff gave feedback on his draft DMP, which will be further iterated during Léon's internship.

Léon's internship will continue during the final reporting period of DATA-KBR-BE and will culminate in a presentation for KBR Staff on his research, which is scheduled to take place in June 2023. The possibility of publishing a dataset on the DATA-KBR-BE platform related to Léon's research will also be explored.

Exploitation of Research Results: Feuilleton in Belgium

As reported in the previous period, *'Het Handelsblad'* from January 1885 was used as the test dataset to undertake the initial analysis of the data extracted as part of Data Extraction 1. For the *Feuilleton in Belgium* research scenario, not only is it necessary to undertake *article segmentation* of the test dataset, but also to *automatically identify and extract the feuilletons* as a specific type of article. Using Dilawar Ali's *article segmentation pipeline*, it was possible to correctly identify 16 out of the 18 feuilletons from the editions of *'Het Handelsblad'* from January 1885. Another aspect that was analysed was the metadata which was provided related to each extracted feuilleton. In addition to a separate text file (.txt) containing the text of each article, a JSON file containing the related metadata was also provided. This metadata included fields such as the *date* of the newspaper edition (e.g. 3rd January 1885), the "articleID" (e.g. the *identifier of the article* which is composed of a number of text blocks) and *type* of data extracted (e.g. Feuilleton).

Building on the analysis of this initial test dataset, the next step was to **develop a data model** to encode the individual issues of a particular literary work published in the feuilleton using the Text Encoding Initiative (TEI), in order to create a **Digital Scholarly Edition**. This involved a number of steps. In addition to the metadata provided as part of the article segmentation process, a number of **additional metadata fields** would be needed. Firstly, metadata related to the newspaper itself. For example, the *name of the newspaper*, *date* of the newspaper edition, the *page numbers* in the issue where the feuilleton is published and a *permalink* to the specific issue, e.g. *Het Handelsblad, 7 January 1885, page 1, https://uurl.kbr.be/1101467*, as depicted in **Fig. 14**, below.



MENGELWERK

DE EED

VAN DEN

ZEEROOVER

Naar het Fransch.

(3e VERVOLG).

II.

HET HOUTEN HUIS.

Op dat oogenblik lieten de volkomen ontbloote rotsen van den *Grand Bé* een vrijen doorgang toe. Het was lage tij, en de streek verlaten.

Wie toch zou er aan gedacht hebben zich neêr te zetten op deze rots, waartegen de branding der orkanen kwam slaan, tenzij schepselen die diep in het hart getroffen waren, die het stadsgewoel ontvluchtten, het gedrang der gelukkige menschen, den gemeenzamen geestdrift van de terugkomst der zeeroovers, die gansch de stad in vroolijkheid stelden.

Jocelyne en hare moeder kwamen dikwijls op die nare rotsen de rusturen, welke zij zich veroorloofden, doorbrengen. Vóór den Oceaan, spraken zij van den teêrbeminden afwezige, vleiden zich met de hoop van hem weêr te zien en berekenden hoe het haar te gelijk mogelijk zoude worden meer te winnen en min te verteeren.

M. Robert de Miniac, uit weinig bemiddelde ouders geboren, voelde zich, van jongs af, door de neigingen van zijnen geest tot de wetenschappen getrokken, en door de ingevingen zijns harten tot het werken aan de verlichting van het menschdom gedreven.

Hij leerde met iever de geneeskunde aan, paste ze met welgelukken toe, vervolgens bekoord door de noodwendigheid van nieuwe landen te zien, scheepte hij zich in voor de Indiën en kwam er van terug, driftiger dan te voren om grootere tochten te ondernemen.

Sedert dan, vaarde hij aan boord van zeerooverschepen, geëerbiedigd en van allen bemind. Hij vond het leven schoon en droomde niets anders dan zijne tochten op ver afgelegene kusten voort te zetten, met de hoop van eens rijk genoeg, om niets meer te behoeven, terug te komen.

Sedert verscheidene jaren wees, scheen hij er niet aan te denken eene familie te stichten, wanneer het toeval hem bij Blanche de la Huchère bracht.

Zeer schoon, alhoewel zij het niet scheen te weten, bekoorde zij vooral door de uitdrukking van een gelaat, dat te gelijk goedheid en rondborstigheid te kennen gaf.

Eenvoudig van manieren, zonder gemaaktheid in de stem, verborg zij uitstekende deugden, gelijk eene andere hare gebreken zou bedekt hebben.

Robert de Miniac verliet Sint-Malo dat jaar met tegenzin, hij durfde nogtans zijne aanvraag niet doen, begeerig om zijn eigen hart te aanhooren.

De jonge dochter dacht meer dan eens aan den knappen jonkman, wiens blik haar volgde met de volharding der onderlinge genegenheid; wanneer hij terug kwam, gevoelde zij eene vreugd, welke zij slecht uitlegde, maar die haar gelaat eensklaps opbelderde en op hare lippen nieuwe gezangen legde.

Wat Robert de Miniac betreft, gedurende de reis, de kortste welke hij ooit gedaan had en die hem echter als eindeloos voorkwam, volgde het beeld van Blanche hem zonder ophouden. Hij geloofde het te zien in de blauwe golven en midden van het verzilverd schuim. Het zacht geluid van hare stem klonk aanhoudend in zijn oor. Hij telde de dagen van zijne overvaart, en nauwelijks was hij ontscheept, of hij spoedde zich naar Mev. de la Huchère.

Blanche was uitgegaan.

Het eerste gevoel van Robert was er zich over te bedroeven; het tweede er zich om te verblijden.

Met eene ongedwongene rondborstigheid, vertrouwde hij aan de moeder van Blanche wat er in zijn hart omging en vroeg heur hare dochter ten huwelijk. Mevrouw de la Huchère reikte hem de hand.

— Ik geef ze u, antwoordde zij, ja, ik geef ze u uit den grond van mijn hart; en nogtans uw beroep is gevaarlijk; gij moet dikwijls landen doorreizen welke door de pest of de gele koorts verwoest worden... Zelden zullen de dagen van vreugd, en lang de maanden der afwezigheid zijn... Maar al hetgeen ik van u weet maakt mij het vertrouwen, de achting, de genegenheid zoo gemakkelijk, dat ik u met geluk mijnen zoon zal noemen.

— Maar Blanche? vroeg Robert verlegen.

— Gij zult haar zelf hare toestemming vragen. In onze familie zijn al de jonge dochters zedig, al de vrouwen eerbaar; de huwelijken worden in volle vrijheid des harten gesloten.

— Mevrouw, gij zijt goed als eene moeder.

— Maar, Blanche is arm!

— Dat is niets!

— Wij leven getweeën van een pensioen, dat mij uit de bijzondere kas des konings betaald wordt.

— Ik win genoeg voor beiden. Als mijne ouders stierven, werd het weinig, dat zij bezaten, aan mijne zusters overgelaten, om hen te helpen een bestaan te vinden. Tot hiertoe heb ik niet zeer aan het geld gehouden, ik zal er meer prijs aan hechten om het welzijn mijner vrouw te verzekeren.

Dan spraken zij beiden over Blanche: de eene met den geestdrift eener opkomende drift, de andere met de diepste moederlijke teederheid. Er werd gehandeld over de pracht van het bruiloftsgeschenk, de sierlijkheid van den huisraad. Het kwam Robert voor als liet een tooverstokje hem toe plotseling al zijne droomen te verwezenlijken. Hij moest de hoogten zijner plannen verlaten om het cijfer zijner besparingen te berekenen. Het weinige dat hij bezat verbaasde en verneederde hem.

— Bedroef u niet, mijn beste kind, zegde hem Mevrouw de la Huchère, tusschen uwe verloving en uw huwelijk, zult gij eene reis doen, en de winst van dien tocht zal aan mijne dochter toegewijd worden.

Wees overigens volkomen gerust, hare eischen zijn zedig; oordeel er over naar dees binnenste.

Een oogenblik later kwam Blanche binnen. Zij scheen meer verheugd dan verwonderd over de tegenwoordigheid van Robert.

Terwijl geheel de familie in den kleinen hof wandelde, bevond M. de Miniac zich eenen stond met haar alleen.

Dan was hij bevreesd. Indien Blanche hem weigerde? Maar Blanche aanschouwde hem met zooveel goedheid dat hij weêr moed in zijne ziel voelde dalen. Terwijl zij een rozentuiltje ontbladerde, opende hij zijn hart, en de medeedelingen eindigden op het oogenblik, dat het leste bloemblad aan de vingers van de jonge dochter ontviel.

— Ik zal uwe vrouw worden, zegde zij hem.

— Gij bemint mij dus?

— Ik weet niet of hetgeen ik gevoel liefde genoemd wordt; maar ik ben verzekerd dat ik mijn geluk van het uwe maken zal, dat ik u zal liefhebben in de vreugd, dat ik uw troost zal wezen in de smart, dat niets ter wereld mij duurbaarder zal zijn dan uw heil.

Zij kwamen hand in hand terug in huis.

Niettegenstaande de rechtzinnige verkleefdheid, welke zij voor elkander gevoelden, waren zij te akkoord toe te geven aan het verlangen van Mevrouw de la Huchère die wenschte hen aan de proef der afwezigheid bloot te stellen.

Zij duurde vier maanden, gedurende welke Blanche en hare moeder aan den uitzet werkten.

Het schoonste lijnwaad werd voor het jonge paar ter zijde gelegd; de weduwe zocht hare oude kanten, hare laatste juweelen. Voortaan zou alles wat sierlijk en rijk kon zijn aan de kinderen toebehooren.

De kamer van Robert werd gemeubleerd alsook een werkkabinet.

Eindelijk was het vaartuig de *Glorieux* in 't zicht. Men legde het vast aan de kaai; het admiraalschap vervulde de talrijke plichtplegingen der aankomst, en Robert de Miniac gevoelde zich vrij van te snellen dáár waar zijn hart hem riep. De zeetocht was voordeelig geweest. Hij bracht stoffen, juweelen, ivooren uit de verschillige doorreisde landen meê. De twee vertrekken, welke men hem had voorbeschikt, werden met herinneringen der gedane reizen versierd en het binnenste van dit nederig huis bekwam een prachtig uitzien, voor het huwelijk van den jongen heelmeester met Blanche de la Huchère.

Van dien dag af begon er voor die twee wezens een geluk dat de hand van het noodlot alleen in staat is te vernietigen.

Zij beminden elkander vuriglijk, met edelmoedigheid, en geene ontgoocheling verduisterde hunne vreugd.

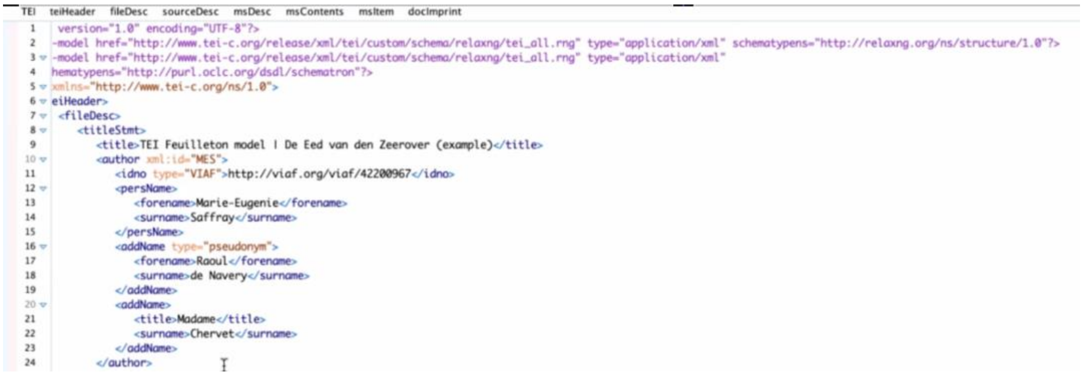
(*Wordt voortgezet*)

Fig. 14 - Het Handelsblad, 7 January 1885, page 1, https://uurl.kbr.be/1101467

Additionally, **metadata related to the specific feuilleton** would also be useful, e.g. *title, author, original language* and *number in sequence*. For example, *title*: De Eed van den Zeeroover, *author*: Raoul de Navery (which was only printed at the end of the *original instalment* published in the edition of 31.1.1884 and not in subsequent instalments), *original language*: French and the *number sequence*: 3rd continuation or "3e Vervolg", which is the 4th instalment of this particular work (the *original instalment* was published in the edition of 31.1.1884, the *1st continuation* in the edition of 3.1.1885, the *2nd continuation* in the edition of 4-5.1.1885 and the *3rd continuation* in the edition of 7.1.1885). If available, it could also be interesting to include in the metadata the different '*sub-titles*' used within the feuilletons, e.g. "II. Het Houten Huis", as depicted in **Fig. 14** above.

Regarding data extraction, it is interesting to note that the first instalment of the "*De Eed van den Zeeroover"* was published on 31.12.1884 and therefore was not originally extracted as part of Data Extraction 1. The additional files for *'Het Handelsblad'* from 31.12.1884 were therefore extracted separately using the KBR's data extraction tool and uploaded to the GhentCDH Next Cloud environment.

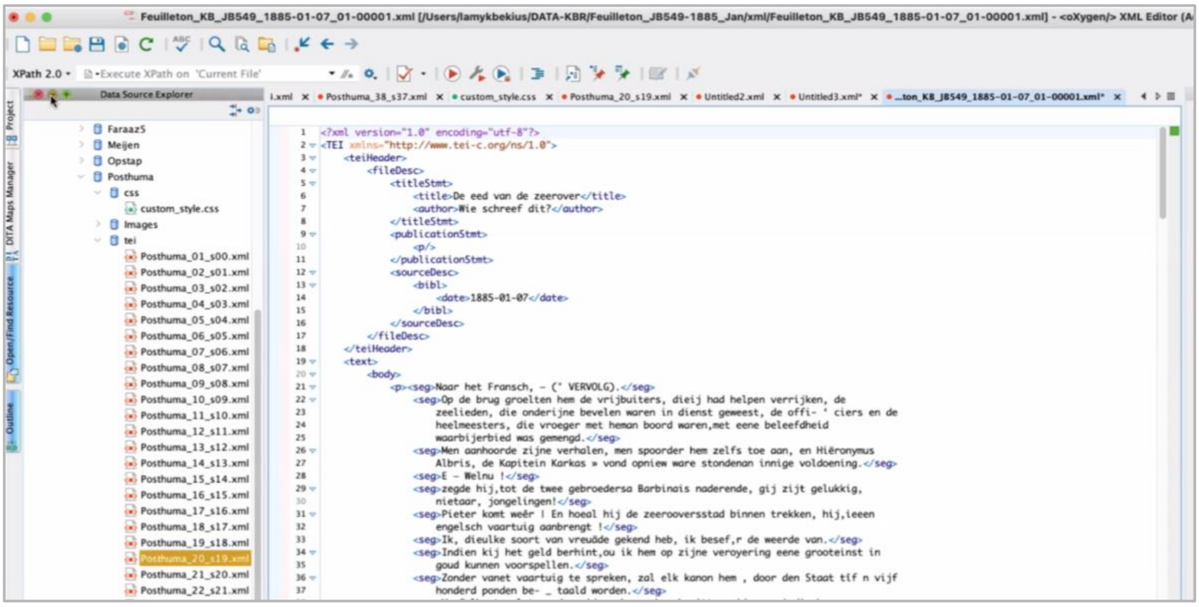
With a clear idea of which metadata would be needed for the digital scholarly edition of each of the feuilletons, a second step was to model this data in TEI-XML. This data modelling was undertaken by Wout Dillen (UAntwerpen / University of Borås, Sweden). For example, details about the *title* and the *author* of the literary work could be encoded into the TEI Header, see **Fig. 15**, below.



```
 version="1.0" encoding="UTF-8"?>
-model href="http://www.tei-c.org/release/xml/tei/custom/schema/relaxng/tei_all.rng" type="application/xml" schematypens="http://relaxng.org/ns/structure/1.0"?>
-model href="http://www.tei-c.org/release/xml/tei/custom/schema/relaxng/tei_all.rng" type="application/xml"
hematypens="http://purl.oclc.org/dsdl/schematron"?>
xmlns="http://www.tei-c.org/ns/1.0">
eiHeader>
 <fileDesc>
    <titleStmt>
       <title>TEI Feuilleton model | De Eed van den Zeerover (example)</title>
       <author xml:id="MES">
          <idno type="VIAF">http://viaf.org/viaf/42200967</idno>
          <persName>
             <forename>Marie-Eugenie</forename>
             <surname>Saffray</surname>
          </persName>
          <addName type="pseudonym">
             <forename>Raoul</forename>
             <surname>de Navery</surname>
          </addName>
          <addName>
             <title>Madame</title>
             <surname>Chervet</surname>
          </addName>
       </author>
```

Fig. 15 - Text Encoding Initiative (TEI) Header including Feuilleton *title* and *author* information

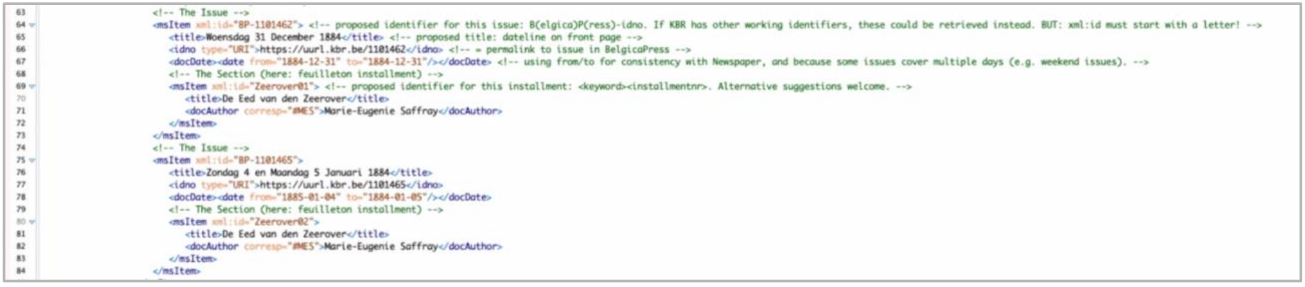
Additionally, the text that had been automatically extracted using the *article segmentation pipeline* can be converted to TEI-XML using a script as shown in **Fig. 16** below. The script and the conversion to TEI-XML was undertaken by Lamyk Bekius (UAntwerpen).



```
<?xml version="1.0" encoding="utf-8"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0">
    <teiHeader>
        <fileDesc>
            <titleStmt>
                <title>De eed van de zeerover</title>
                <author>Wie schreef dit?</author>
            </titleStmt>
            <publicationStmt>
                <p/>
            </publicationStmt>
            <sourceDesc>
                <bibl>
                    <date>1885-01-07</date>
                </bibl>
            </sourceDesc>
        </fileDesc>
    </teiHeader>
    <text>
        <body>
            <p><seg>Naar het Fransch, - (' VERVOLG).</seg>
                <seg>Op de brug groelten hem de vrijbuiters, dieij had helpen verrijken, de
                    zeelieden, die onderijne bevelen waren in dienst geweest, de offi- ' ciers en de
                    heelmeesters, die vroeger met heman boord waren,met eene beleefdheid
                    waarbijerbied was gemengd.</seg>
                <seg>Men aanhoorde zijne verhalen, men spoorder hem zelfs toe aan, en Hiëronymus
                    Albris, de Kapitein Karkas » vond opniew ware stondenan innige voldoening.</seg>
                <seg>E - Welnu !</seg>
                <seg>zegde hij,tot de twee gebroedersa Barbinais naderende, gij zijt gelukkig,
                    nietaar, jongelingen!</seg>
                <seg>Pieter komt weêr | En hoeal hij de zeerooversstad binnen trekken, hij,ieeen
                    engelsch vaartuig aanbrengt !</seg>
                <seg>Ik, dieulke soort van vreuǎde gekend heb, ik besef,r de weerde van.</seg>
                <seg>Indien kij het geld berhint,ou ik hem op zijne veroyering eene grooteinst in
                    goud kunnen voorspellen.</seg>
                <seg>Zonder vanet vaartuig te spreken, zal elk kanon hem , door den Staat tif n vijf
                    honderd ponden be- _ taald worden.</seg>
```

Fig. 16 - Text of "De Eed van den Zeeroover" encoded in TEI-XML

As the goal for this phase of the *Feuilleton in Belgium* is to create a digital scholarly edition of an entire literary work which has been published in instalments within a historical newspaper, it is important to model the interconnections between the different instalments of the feuilleton. This inter-linking can be undertaken in the TEI-Header. **Fig. 17** below displays the 'issue' level metadata, with lines 64-73 providing the details of the first instalment of the "De Eed van den Zeeroover" published on 31.12.1884, as well as the second instalment published in the edition of *Het Handelsblad* for 4-5.1.1885. For example, a persistent identifier could be assigned at the issue level, as a combination of "BP" standing for BelgicaPress combined with the identifier from the UURL (KBR's persistent identifier system) linking to the digital object. For *Het Handelsblad* for 31.12.1884, the proposed identifier would be "BP-1101462". It is important to note that an identifier encoded in the field *xml:id* must start with a letter, hence the additional of "BP" to the UURL of the digital object.

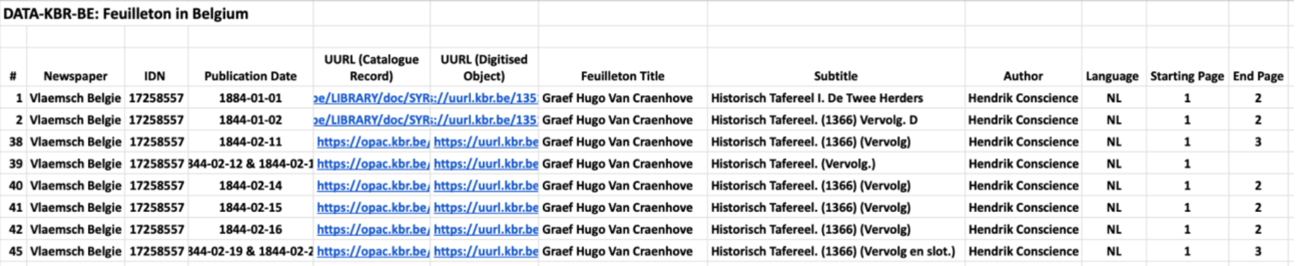


```
<!-- The Issue -->
<msItem xml:id="BP-1101462"> <!-- proposed identifier for this issue: B(elgica)P(ress)-idno. If KBR has other working identifiers, these could be retrieved instead. BUT: xml:id must start with a letter! -->
   <title>Woensdag 31 December 1884</title> <!-- proposed title: dateline on front page -->
   <idno type="URI">https://uurl.kbr.be/1101462</idno> <!-- = permalink to issue in BelgicaPress -->
   <docDate><date from="1884-12-31" to="1884-12-31"/></docDate> <!-- using from/to for consistency with Newspaper, and because some issues cover multiple days (e.g. weekend issues). -->
   <!-- The Section (here: feuilleton installment) -->
   <msItem xml:id="Zeerover01"> <!-- proposed identifier for this installment: <keyword><installmentnr>. Alternative suggestions welcome. -->
      <title>De Eed van den Zeerover</title>
      <docAuthor corresp="#MES">Marie-Eugenie Saffray</docAuthor>
   </msItem>
</msItem>
<!-- The Issue -->
<msItem xml:id="BP-1101465">
   <title>Zondag 4 en Maandag 5 Januari 1884</title>
   <idno type="URI">https://uurl.kbr.be/1101465</idno>
   <docDate><date from="1885-01-04" to="1884-01-05"/></docDate>
   <!-- The Section (here: feuilleton installment) -->
   <msItem xml:id="Zeerover02">
      <title>De Eed van den Zeerover</title>
      <docAuthor corresp="#MES">Marie-Eugenie Saffray</docAuthor>
   </msItem>
</msItem>
```

Fig. 17 Issue level metadata for the 1st and 2nd instalments of "De Eed van den Zeeroover" encoded in TEI-XML.

Following this initial data modelling phase, attention returned to the original aim for the *Feuilleton in Belgium* research scenario; to investigate the prominent literary authors in the first century of the Belgian state. The test dataset of January 1885 is interesting here as the "*De Eed van den Zeeroover*" was written by a French writer, who published under several pseudonyms, including Raoul de Navery (1829-1885). While it is interesting to investigate which international authors were published in Belgian newspapers, this is beyond the scope of the original research question of the *Feuilleton in Belgium.* Wout Dillen therefore proposed to select a different newspaper to continue this work, selecting an alternative newspaper from the same period, the *Vlaemsch Belgie*, a Belgian, Dutch language daily newspaper published in Brussels in 1884, which published *feuilletons* from Belgian writers. It was agreed to take a similar approach as with the original test dataset from *Het Handelsblad.* However, it soon became clear that a different approach would be needed due to the nature of the *Vlaemsch Belgie*.

Firstly, while the *Vlaemsch Belgie*, which was published between 1st January and 22 November 1884, and then continued from 23rd November 1884 as *De vlaemsche Belgen : Dagblad voor staetkunde, koophandel en letterkunde*, had been digitised and was available in BelgicaPress, it had been treated as a single volume, rather than as a newspaper consisting of different editions. This was challenging as Optical Character Recognition (OCR) had therefore not been applied to the newspaper. As a first step, It was therefore agreed to manually identify the *feuilletons* published *Vlaemsch Belgie* for January 1884. This work was undertaken by Sally Chambers (KBR) and Lamyk Bekius (UAntwerpen) on the basis of the digitised edition of the *Vlaemsch Belgie* in BelgicaPress and using a collaborative spreadsheet to collect the relevant metadata, such as *newspaper title, IDN (identifier of the bibliographic record in the KBR catalogue), publication date, UURL (persistent identifier of a) the catalogue record and b) of the digitised object), title of the feuilleton, subtitle, author, language, start and end page of the feuilleton, sequence number, details about the next issue, genre* as well as a specific *notes field* to provide any further information. From this manual analysis of *Vlaemsch Belgie* for January 1884 (see **Fig. 18**, below) it was possible to identify a particular literary work: *Graef Hugo Van Craenhove: Historisch Tafereel* written by Belgian author, Hendrik Conscience, which has been published in 8 instalments from 1st January to 20th February 1884. It was agreed to use this literary work as a case study for developing the *digital scholarly edition*, based on the TEI data model outlined above, for the *Feuilleton in Belgium* research scenario.



DATA-KBR-BE: Feuilleton in Belgium

#	Newspaper	IDN	Publication Date	UURL (Catalogue Record)	UURL (Digitised Object)	Feuilleton Title	Subtitle	Author	Language	Starting Page	End Page
1	Vlaemsch Belgie	17258557	1884-01-01	oe/LIBRARY/doc/SYR	://uurl.kbr.be/135	Graef Hugo Van Craenhove	Historisch Tafereel I. De Twee Herders	Hendrik Conscience	NL	1	2
2	Vlaemsch Belgie	17258557	1844-01-02	oe/LIBRARY/doc/SYR	://uurl.kbr.be/135	Graef Hugo Van Craenhove	Historisch Tafereel. (1366) Vervolg. D	Hendrik Conscience	NL	1	2
38	Vlaemsch Belgie	17258557	1844-02-11	https://opac.kbr.be/	https://uurl.kbr.be	Graef Hugo Van Craenhove	Historisch Tafereel. (1366) (Vervolg)	Hendrik Conscience	NL	1	3
39	Vlaemsch Belgie	17258557	344-02-12 & 1844-02-1	https://opac.kbr.be/	https://uurl.kbr.be	Graef Hugo Van Craenhove	Historisch Tafereel. (Vervolg.)	Hendrik Conscience	NL	1	
40	Vlaemsch Belgie	17258557	1844-02-14	https://opac.kbr.be/	https://uurl.kbr.be	Graef Hugo Van Craenhove	Historisch Tafereel. (1366) (Vervolg)	Hendrik Conscience	NL	1	2
41	Vlaemsch Belgie	17258557	1844-02-15	https://opac.kbr.be/	https://uurl.kbr.be	Graef Hugo Van Craenhove	Historisch Tafereel. (1366) (Vervolg)	Hendrik Conscience	NL	1	2
42	Vlaemsch Belgie	17258557	1844-02-16	https://opac.kbr.be/	https://uurl.kbr.be	Graef Hugo Van Craenhove	Historisch Tafereel. (1366) (Vervolg)	Hendrik Conscience	NL	1	2
45	Vlaemsch Belgie	17258557	344-02-19 & 1844-02-2	https://opac.kbr.be/	https://uurl.kbr.be	Graef Hugo Van Craenhove	Historisch Tafereel. (1366) (Vervolg en slot.)	Hendrik Conscience	NL	1	3

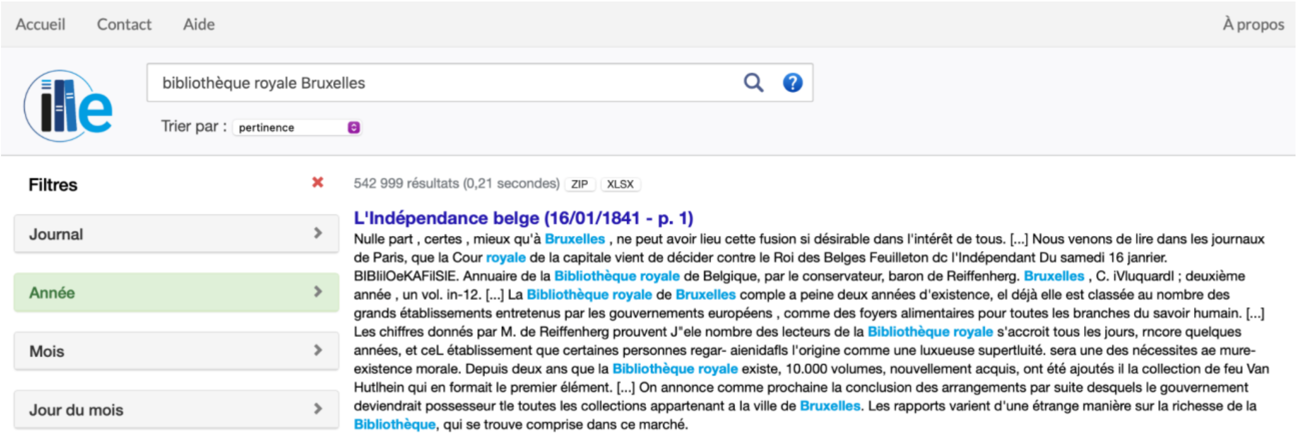
Fig. 18 Metadata related to the literary work *Graef Hugo Van Craenhove: Historisch Tafereel written* by *Hendrik Conscience* as published in the *Vlaemsch Belgie* for January 1884

While the *Vlaemsch Belgie* provides a good case study for the *Feuilleton in Belgium* research scenario, the fact that it has not been OCR'd or indexed within *BelgicaPress* provides additional challenges. However, with the manual identification of the relevant pages that included the *feuilletons* it was possible to ask the KBR's ICT team just for the relevant pages (available in JPEG format) of the newspaper which was uploaded to the GhentCDH NextCloud environment. As a next step towards creating a digital scholarly edition, the DATA-KBR-BE team will investigate if it will be possible to OCR either the *Vlaemsch Belgie* for 1884 or at least the pages containing the literary work: *Graef Hugo Van Craenhove: Historisch Tafereel*.

Exploitation of Research Results: History of Belgian Journalism

As reported in the previous annual report, 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. Following an initial meeting with Brecht in March 2022, 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 (Steven Verstockt and Dilawar Ali) and the DATA-KBR-BE Project Team from KBR (Frédéric Lemmers and Sally Chambers) was organised to explore collaboration possibilities. Access to the DATA-KBR-BE data via the GhentCDH Next Cloud had also been provided to the *History of Belgian Journalism* team.

Firstly, further details about the research being undertaken by the CAMille team related to the *History of Belgian Journalism* research scenario were explained in more detail. A first step is the creation of the CAMille platform, where a selection of digitised historical newspapers (currently: *L'Avenir du Luxembourg; La Libre Belgique; L'Indépendance belge; L'Indépendance belge (édité en Angleterre); Journal de Bruxelles; Journal de Charleroi; La Meuse; Le Petit Bleu; Le Vingtième Siècle; Vers l'Avenir; Le Peuple; Le Soir* and *Le Drapeau rouge*) have been indexed as full-text as illustrated in **Fig. 19** below.



bibliothèque royale Bruxelles

Trier par : pertinence

542 999 résultats (0,21 secondes)

L'Indépendance belge (16/01/1841 - p. 1)

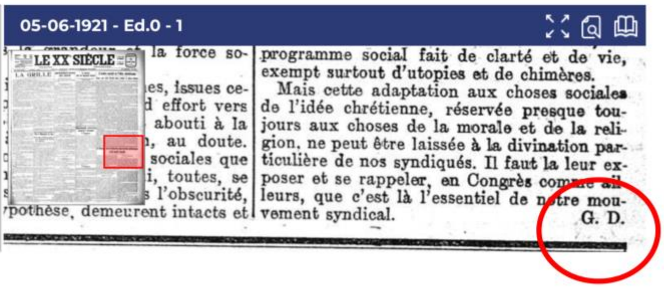
Nulle part , certes , mieux qu'à Bruxelles , ne peut avoir lieu cette fusion si désirable dans l'intérêt de tous. [...] Nous venons de lire dans les journaux de Paris, que la Cour royale de la capitale vient de décider contre le Roi des Belges Feuilleton dc l'Indépendant Du samedi 16 janrier. BIBlilOeKAFilSIE. Annuaire de la Bibliothèque royale de Belgique, par le conservateur, baron de Reiffenherg. Bruxelles , C. iVluquardl ; deuxième année , un vol. in-12. [...] La Bibliothèque royale de Bruxelles comple a peine deux années d'existence, el déjà elle est classée au nombre des grands établissements entretenus par les gouvernements européens , comme des foyers alimentaires pour toutes les branches du savoir humain. [...] Les chiffres donnés par M. de Reiffenherg prouvent J"ele nombre des lecteurs de la Bibliothèque royale s'accroit tous les jours, rncore quelques années, et ceL établissement que certaines personnes regar- aienidafls l'origine comme une luxueuse supertluité. sera une des nécessites ae mure- existence morale. Depuis deux ans que la Bibliothèque royale existe, 10.000 volumes, nouvellement acquis, ont été ajoutés il la collection de feu Van Hutlhein qui en formait le premier élément. [...] On annonce comme prochaine la conclusion des arrangements par suite desquels le gouvernement deviendrait possesseur tle toutes les collections appartenant a la ville de Bruxelles. Les rapports varient d'une étrange manière sur la richesse de la Bibliothèque, qui se trouve comprise dans ce marché.

Fig. 19 - A screen shot from the CAMille platform which includes a selection of digitised historical newspapers which have been indexed as full-text

More specifically, the team is currently working in particular on the entire run of the French language Belgian Roman Catholic newspaper Le Vingtième Siècle (1895 - 1940) with the goal of investigating all of the journalists who worked for the newspaper including an analysis of the changes that have taken place in the newsroom over time, such as changing systems, processes and organisational practices. Furthermore, a number of case studies of particular female journalists will be explored, for example Alice Bron, as well an exploration of the **editorial teams** which were active in Belgian journalism, including who was writing for which newspaper in what period.

Additionally, an existing *Dictionary of Belgian Journalism*, compiled by Pierre Van den Dungen is being converted into a database. 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. For the linking of the authors' signatures identified in newspaper articles to the biographical entries about the Belgian journalists' in the CAMille database, **automatic recognition of signatures** was particularly of interest.

Sébastien De Valeriola has been experimenting with a method to recognise the location of particular strings (e.g. text on the right hand side of the page following a blank space, as depicted in **Fig. 20** to the right). However, he is currently working with digitised newspapers without article segmentation. A collaboration with DATA-KBR-BR regarding article segmentation could be interesting to explore.



05-06-1921 - Ed.0 - 1

la force so-
es, issues ce-
d effort vers
abouti à la
n, au doute.
sociales que
i, toutes, se
s l'obscurité,
pothèse, demeurent intacts et

programme social fait de clarté et de vie, exempt surtout d'utopies et de chimères.
Mais cette adaptation aux choses sociales de l'idée chrétienne, réservée presque toujours aux choses de la morale et de la religion, ne peut être laissée à la divination particulière de nos syndiqués. Il faut la leur exposer et se rappeler, en Congrès comme ailleurs, que c'est là l'essentiel de notre mouvement syndical.
G. D.

Fig. 20 - An example of a journalist's signature at the end of an article

To facilitate this process, the CAMille team has been using a spreadsheet to manually record each of the signatures in *Le Vingtième Siécle* from 1921. Metadata such as *which file* the signature has been identified in, as well as the text of each signature is recorded in a separate field, e.g. *Le Vingtième Siécle* from 1.6.1921: *signature01:* P.M.; *signature 02:* A.M. and *signature 03:* P.D. as well as an *additional notes field*. Furthermore an analysis of the different types of signatures have been undertaken, such as signatures of a *Feuilleton, a fictional character, signature of an author of a letter published in the newspaper, signature of the King, etc.*

Regarding next steps, it was agreed to explore the possibility of extracting initially *Le Vingtième Siècle* from 1921 and at a later stage, *Le Peuple* from 1921, and to run the article segmentation pipeline to investigate whether it would improve the automatic extraction of the author's signature.

DATA-KBR-BE Article Segmentation Pipeline

For Ghent University's Data Scientist, Dilawar Ali (ID Lab, UGent), a core focus of his work was to finalise the writing of his PhD thesis and prepare for the handover of his **article segmentation pipeline** to the DATA-KBR-BE project team. An important step in this process, was the publication of an article, written by Dilawar together with other members of the DATA-KBR-BE team: "*Computer Vision and Machine Learning Approaches for Metadata Enrichment to Improve Searchability of Historical Newspaper Collections*", which was submitted 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). As reported in the previous reporting period, the initial version of the article had been submitted in January 2022 for partial fulfilment of Dilawar's PhD. In September 2022, a minor revision of the article was requested. A revised version of the article was submitted in mid-October 2022. This revised version was accepted for publication in December 2022. The published version of the article in the *Journal of Documentation* can be found here. A preprint version of the article is also available in the Ghent University Academic Bibliography.

Another major achievement was that Dilawar Ali (ID Lab, UGent) successfully defended his PhD at the Faculty of Engineering and Architecture at Ghent University (see **Fig. 21** to the left) with his PhD entitled: "*AI-based methods to enrich, geolocalize, and rephotograph historical newspaper photos*" on 25th January 2023. KBR's digitised historical newspapers in the context of DATA-KBR-BE formed a major case study as part of his PhD, particularly in relation to metadata enrichment methods. Dilawar's **article segmentation pipeline** was a key contribution, both to his PhD and to the DATA-KBR-BE project.

Fig. 21 - Dilawar Ali (ID Lab, UGent) presenting the KBR' Newspaper Collections case study during his PhD defence on 25th January 2023.

Following the successful defence of his PhD, a number of meetings with Dilawar, Sally Chambers and Tan Lu from KBR regarding the handover of the code of the **article segmentation pipeline** to the DATA-KBR-BE project team for further use as part of the project. A handover plan was agreed and will be implemented following Dilawar's return to Pakistan in Spring 2023.

DATA-KBR-BE: Stimulating Further Research Collaborations

Historical Newspapers Research Community and Workshop

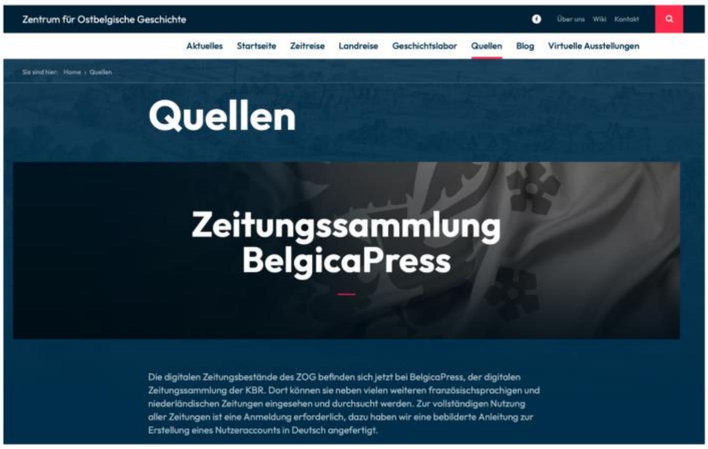
Building further on the History of Belgian Journalism research scenario, FEDtWIN researchers Brecht Deseure and Julie Birkholz took the initiative to establish an *informal research community around historical newspapers in Belgium*. The aim of this community is to exchange approaches, methods and expertise related to research on historical newspapers and to explore future collaborations. In particular, KBR is interested in deepening our understanding of researchers' needs and requirements, with a view to improving our services where possible.

As a first step, a *Newspaper Research Workshop* is being organised and will take place on 20th March 2023 at KBR. The main goal of the workshop is for the researchers to get to know each other and be informed about ongoing research. Presentations, particularly from early career researchers, followed by community discussion are anticipated.

Research Collaboration with the Zentrum für Ostbelgische Geschichte (ZOG)

As reported in the last period, the DATA-KBR-BE project had set up a collaboration with the Zentrum für Ostbelgische Geschichte (ZOG) regarding their collections of **around 50 German-language digitised historical newspapers titles from East Belgium**. The focus of the collaboration, over the medium to long term, is to ingest these digitised newspapers into BelgicaPress. Following a pilot project with three newspapers: *Eupener Zeitung*, *Eupener Nachrichten* and *Grenz-Echo*, it was agreed to continue this collaboration and develop an **Ingestion Plan** to manage the inclusion of all the newspapers.
For *Phase 1 of the Ingestion Plan* the focus would be on the 11 newspapers that has been previously made available via the ZOG's website: *Der Landbote; Die Arbeit; Die Fliegende Taube; Eupener Bürger Zeitung; Eupener Kreisblatt; Eupener Nachrichten; Eupener Zeitung; Korrespondenzblatt; Malmedy-St. Vither Volks-Zeitung; St. Vither Volks-Zeitung and Wochenblatt für den Kreis Malmedy*. This work will be undertaken on a step-by-step basis with a view to have completed Phase 1 of the Ingestion Plan by the end of 2023. To formalise this process a *draft collaboration agreement* has been prepared, which following the approval of the BELSPO and ZOG legal advisers, will be signed by both organisations.

In March 2023 the ZOG launched their new website, including a specific section of the website linking through to their newspapers in BelgicaPress (see **Fig. 22, right**). It will be interesting to monitor how many BelgicaPress users have been redirected from the ZOG website. Additionally, the possibility of translating the BelgicaPress interface into German to support the German language community of users is being explored.



Die digitalen Zeitungsbestände des ZOG befinden sich jetzt bei BelgicaPress, der digitalen Zeitungssammlung der KBR. Dort können sie neben vielen weiteren französischsprachigen und niederländischen Zeitungen eingesehen und durchsucht werden. Zur vollständigen Nutzung aller Zeitungen ist eine Anmeldung erforderlich, dazu haben wir eine bebilderte Anleitung zur Erstellung eines Nutzeraccounts in Deutsch angefertigt.

Fig. 22 - The Zentrum für Ostbelgische Geschichte (ZOG)'s website highlighting their digitised historical newspaper collection in BelgicaPress.

Finally, the possibility of including an **additional research scenario related to the ZOG's East German newspapers** is continuing to be explored. As a first step towards this the researchers from ZOG are members of the *Historical Newspapers Research Community* and will take part in the *Newspaper Research Workshop* which is being organised on 20th March 2023 at KBR. During that workshop, Niklas Stenzel, Research Associate at ZOG and PhD student at the University of Siegen in Germany and C2DH at the University of Luxembourg will give a presentation on: *Discursive Identity Construction in East Belgium: A Linguistic and Discourse Historical Analysis of the Patterns of Language and Communication in East Belgian Mass Media*.

Digital Humanities Masters Internships 2022-2023

During the academic year 2022-2023, a number of masters' students, coordinated by Julie Birkholz and in liaison with the FED-tWIN projects: CAMILLE, LabEL and the Digital Research Lab and the BRAIN projects DATA-KBR-BE and BelgicaWeb, have been welcomed to KBR to undertake internships as part of their studies. In particular, masters' students from the *KU Leuven's Digital Humanities Masters' Programme*, from *Ghent University's Department of History* who are using KBR's collections or doing internships at KBR and from the *Université de Franche-Comté Rare Book and Digital Humanities Masters' Programme*.

In the context of DATA-KBR-BE, and in addition to the internship of Léon Castelein from Ghent University's Department of History who is undertaking his internship directly as part of the DATA-KBR-BE *Collective Action Belgium* research scenario as reported above, the internships on *post-OCR correction of historical newspapers* and *recovering female Belgian journalists using Named Entity Recognition in digitised historical newspapers* were particularly of interest for the DATA-KBR-BE project team. An event was organised in March 2023 to welcome all interns to KBR. This will be followed by a closing event in June 2023 where the students will be able to present the results of their internships at KBR.

WP5: Project Management and Communication - led by KBR

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. During the third reporting period, the DATA-KBR-BE project team focused on specific meetings with relevant team members, e.g. meetings to discuss the development of the DATA-KBR-BE platform; research-scenario specific meetings and meetings to discuss the documentation of article segmentation pipeline and the handover of the code. During this reporting period, a meeting of the DATA-KBR-BE Scientific Advisory Board ('Follow-up Committee') was not held, however the DATA-KBR-BE project coordinator and several members of the DATA-KBR-BE 'Follow-up Committee' took part in a week long seminar at Schloss Dagstuhl – Leibniz Center for Informatics in Germany on *Computational Approaches for Digitised Historical Newspapers* from 17th-22nd July 2022. This provided the opportunity to discuss DATA-KBR-BE's developments. The next DATA-KBR-BE Scientific Advisory Board meeting will take place in Autumn 2023. 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*.

3. INTERMEDIARY RESULTS

This section provides an overview 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 December 2020, in this 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 as well as on the additional research scenario, the History of Belgian Journalism. The focus for this reporting period was on the further analysis of the extracted data including whether further additional data needs to be extracted as the research scenarios evolve. 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

In this reporting period, the work focussed on further analysis of the data extracted as part of Data Extraction 1 including whether further data extractions would be needed as the research scenarios evolve. As the data extracted as part of Data Extraction 1 provided a substantial amount of data for the research teams to analyse as part of their *interdisciplinary research scenarios*, more **customised data extractions** were needed as the research scenarios developed. The details of these additional **customised data extractions** are included in *WP4: Scientific exploitation and valorisation*. Additionally, the researchers reported that the quality of the textual layer of the digitised newspapers which had been created using Optical Character Recognition (OCR) was insufficient for them to undertake the research as they had anticipated. These OCR quality challenges will be investigated further in the final phase of the project.

D2.2 Sustainable data extraction workflow design.

To facilitate the data extraction process, KBR's ICT and Digitisation Departments developed the **KBR Data Extraction Tool.** *KBR's data extraction tool* has two key options: a) a ***simple search*** which is intended for downloading specific files between two publication dates and b) an ***excel search*** which is intended for requesting the extraction of multiple files or datasets from KBR's data infrastructure for download. Following a data extraction request by a KBR member of staff, the extracted files are downloaded from the KBR's file system and uploaded to the **GhentCDH Next Cloud environment** (https://data.ghentcdh.ugent.be) so that they can be accessed by the researcher. During the DATA-KBR-BE project, such requests for data extractions are being handled on a case-by-case basis. However, one of the future recommendations of the DATA-KBR-BE project could be for KBR to develop a **data extraction service for researchers**. The development of such a service could be potentially explored in the context of the recently funded "*KBR Virtual Lab: e-infrastructure for facilitating access and research of KBR's collections as data"* project. DATA-KBR-BE's recommendations for this will be described in the *D2.2 Sustainable data extraction workflow design*, which will be delivered towards the end of the project.

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.*

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. The focus of the activities in this phase of the project has been on the design of the DATA-KBR-BE platform and in particular, the **iterative development of the technical and functional requirements**. The key results achieved in this reporting period have included the agreement that the DATA-KBR-BE Platform should be **fully integrated into the KBR website**, alongside KBR's core digital services such as Belgica, BelgicaPress, BelgicaPeriodicals and the General Catalogue. An initial mock-up of how the DATA-KBR-BE platform could be integrated into the KBR website has been provided.

As part of the iterative development of the technical and functional requirements, a number of phases of development for the DATA-KBR-BE platform have been designed. This phased approach to the development of the platform is related to its sustainability. It is intended that the DATA-KBR-BE Platform will be a long-term initiative, beyond the end of the DATA-KBR-BE project itself. For the delivery of the *Phase 1 of the DATA-KBR-BE Platform* at the end of the project, it is recommended that **a simple, user-friendly website, for publishing curated datasets for download based on KBR's digitised and born-digital collections.**

D3.2 Implementation of data.kbr.be including dataset publication

Following the decision to deliver a *simple, user-friendly website for the publication of curated datasets*, the focus then turned to the requirements for the datasets themselves. The requirements investigated in this phase of the project included a range of topics such as: a) *documenting the datasets, e.g. via a README file*, b) *data citations and related Persistent Identifier (PIDs)*, c) *rights, licensing and terms of use* and d) *versioning* *of datasets*. The publication of the International Galleries, Libraries, Archives and Museums (GLAM) Labs Community's "**Checklist to publish Collections as Data in GLAM Institutions**", which is intended as an easy to apply method to encourage small and medium sized organisations to publish their digital collections as Collections as Data has been particularly useful.

D3.3 Digital Asset Registry: inventory of KBR's Digital Collections

The *Digital Asset Registry* is intended to be an inventory of KBR's collections which could be published as 'Collections as Data' on the DATA-KBR-BE platform. From the original project proposal, the DATA-KBR-BE project team committed to co-curating and publishing at **least three Open Science datasets** on the platform. It is anticipated that these three datasets are likely to be related to the two original interdisciplinary research scenarios: 1) Collective Action Belgium, led by GhentCDH and 2) Feuilleton in Belgium, led by ACDC as well as a dataset related to the additional research scenario, the History of Belgian Journalism. At this stage it is not clear if it will be possible to publish more than three datasets, however, D3.3 is intended to facilitate the dataset publication process both during and after the end of the project. KBR's Digital Asset Registry should be a living document which is updated regularly, e.g. every 6-12 months. An **initial overview of the types of datasets that could be published on the DATA-KBR-BE Platform has been prepared.** This includes: a) ***thematic datasets***, e.g. datasets related to the interdisciplinary research scenarios within DATA-KBR-BE or related to research projects, e.g. funded by BELPSO or other funding providers and b) ***Collections as Data datasets***, e.g. datasets can include any data that is related to KBR's Collections. As part of KBR's emerging Digital Data Strategy, nine "**data types**", related to KBR's collections have been identified: *Bibliographic data; Authority data; Heritage data; Digital data; Collections as data; Born-digital data; Electronic resources; Research data* and *Lab data (code).* It could be interesting to consider publishing one dataset for each data type on 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.*

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

The focus of the work for this reporting period has been on the further analysis of the data which was extracted as part of Data Extraction 1. The key results for each *interdisciplinary research scenario* are summarised below.

For **Collective Action Belgium**, a first important step was to build a dataset of all the articles related to "collective action" e.g. strikes, demonstrations and protests, in particular, the strikes that took place in Ghent. This was undertaken in a number of ways. For example by performing **keyword searches on the extracted data.** For this analysis, it was agreed to focus on the test dataset, the newspaper, the "Vooruit" from April 1913. To focus this work, three "search phases" were selected. Firstly, a single keyword "werkstaking". Secondly two keywords "werkstaking" and "gent". Thirdly, "ons huis" or in full "Ons Huis, Socialistische Werkersvereenigingen", a building in Ghent where activities of the Socialist Workers Associations were organised. Even though a *Jupyter Notebook* has been prepared to facilitate this work, it was decided to undertake the keyword searches using the **advanced search interface of the BelgicaPress.** Additional keyword searches on the extracted pdfs could also be undertaken as required.

The historical geo-localisation of the *collective actions* that took place as reported in the digitised historical newspapers is of particular interest to the GhentCDH team. For example, it would be interesting to geo-locate the locations of the various strikes and demonstrations on a historical map. Such geo-location could be undertaken with collaboration with GhentCDH's GentGemapt project.

The key results in this phase included the importance of **article-level indexing of the digitised historical newspapers in BelgicaPress**. Even though it is unlikely that article segmentation could be applied to the whole of BelgicaPress, even the **indexing of the titles of the articles** could be an additional enrichment to the BelgicaPress search functionality. Another useful enhancement would be the inclusion of **proximity searching**, e.g. when "*werkstaking*" and "*gent*" appear in the same article or located within 5 words of each other.

Additionally, the researchers expressed some concerns regarding the **quality of the Optical Character Recognition (OCR)** for "*Vooruit*" from April 1913, which would lead to non-identification of relevant articles for the thematic dataset. The possibility of **reOCRing the test dataset** will be explored in the final phase of the project.

Finally, some initial investigations as to which data could be included in the **thematic dataset for "Collective Action Belgium"** to be published on the DATA-KBR-BE platform were undertaken. Which *file formats* (e.g. pdfs, tiffs, JPEGs, XML-ALTO files) as well as the *volume of the files* (e.g. all files for April 1913 of *Vooruit* totals 3.5 GB of data) are key questions to be addressed. The possibility of publishing a dataset on the DATA-KBR-BE platform related to Léon Castelein's research in the context of "*Collective Action Belgium"* will also be explored.

For the **Feuilleton in Belgium**, the key results in this reporting period have been related to the **development of a data model** to encode the individual instalments of a particular literary work using the **Text Encoding Initiative (TEI)** with the goal of creating a **Digital Scholarly Edition of a particular literary work** as published in the digitised historical newspapers. For this, **metadata at the level of the newspaper** were needed, such as the *name of the newspaper*, *date* of the newspaper edition, the *page numbers* in the issue where the feuilleton is published and a *permalink* to the specific issue, e.g. *Het Handelsblad, 7 January 1885, page 1, https://uurl.kbr.be/1101467*. Additionally, **metadata related to the specific feuilleton** which would also be useful, e.g. *title, author, original language* and *number in sequence* were identified.

Furthermore, as such literary works are published in a number of instalments in a historical newspaper, it was important to **model the interconnections between the different instalments of the feuilleton**. This inter-linking can be undertaken in the TEI-Header. Finally, the text of the feuilleton, which had been automatically extracted using the *article segmentation pipeline* was converted to TEI-XML using a script. To validate the data model, the "De Eed van den Zeeroover" by Raoul de Navery, which was published in a number of instalments in *Het Handelsblad* from December 1884 onwards was encoded using the data model.

Related to the data extraction, following the data modelling phase, it was agreed to **select an alternative newspaper from the same period, the *Vlaemsch Belgie*, a Belgian, Dutch language daily newspaper published in Brussels in 1884**, to create the digital scholarly edition for the *Feuilleton in Belgium.* Following a manual identification of the *feuilletons* published in *Vlaemsch Belgie* in January 1884 a particular literary work was identified. This work: *Graef Hugo Van Craenhove: Historisch Tafereel* written by Belgian author, *Hendrik Conscience*, had been published in 8 instalments from 1st January to 20th February 1884. The research team agreed to use it as a case study for developing the *digital scholarly edition*, based on the TEI data model outlined above, for the *Feuilleton in Belgium* research scenario.

However, this choice of case study is not without its challenges, as the *Vlaemsch Belgie* has not been OCR'd or indexed within *BelgicaPress.* As a next step towards creating a digital scholarly edition, the DATA-KBR-BE team will investigate if it will be possible to OCR either the *Vlaemsch Belgie* for 1884 or at least the pages containing the literary work: *Graef Hugo Van Craenhove: Historisch Tafereel*.

For the **History of Belgian Journalism** key results have included the continued exploration of collaboration between the CAMille, the Centre for Archives on the Media and Information, ULB-KBR team and the DATA-KBR-BE project team. During a follow-up meeting, the research being undertaken by the CAMille team related to the *History of Belgian Journalism* research scenario were explained in more detail. For example: a) the creation of the CAMille platform, where a selection of digitised historical newspapers have been indexed as full-text, b) an analysis of the entire run of the Le Vingtième Siècle (1895 - 1940) to explore changed that had taken place in the newsroom over time, such as a number of case studies of particular journalists, as well an exploration of the **editorial teams** which were active in Belgian journalism, including who was writing for which newspaper in which period. Additionally, the conversion of the existing *Dictionary of Belgian Journalism*, compiled by Pierre Van den Dungen, to a database was explained.

Regarding possible collaborations, the objective of providing links from the **authors signatures** identified in newspaper articles to the **biographical entries** about the Belgian journalists in the database was identified. The CAMille team has been experimenting with digital methods to automatically extract the authors signatures. Regarding next steps, it was agreed to explore the possibility of extracting initially *Le Vingtième Siècle* from 1921 and at a later stage, *Le Peuple* from 1921, and to run the article segmentation pipeline to investigate whether it would improve the automatic extraction of the authors signature.

D4.2 Report of the High Profile Hackathon
At this stage, it is anticipated that the High Profile Hackathon, which would take place at KBR in early 2024, would be an ideal opportunity to present the beta version of the DATA-KBR-BE Platform to the wider community, towards the end of the project. The webinar "Towards implementing Collections as Data in GLAM institutions", which was co-organised by the DATA-KBR-BE project team in October 2022, provided inspiration for the design of the DATA-KBR-BE Hackathon. The possible collaboration with the project team from the common European Data Space for Cultural Heritage and the International GLAM Labs Community is being explored. The possibility of including a legal round table, as suggested by the Follow-up Committee continues to be explored.

WP5: Project Management and Communication
There were originally two deliverables foreseen in WP5*: D5.1 Annual Report 2020-2021* 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. In the final reporting period, *D5.2 Final Report* will be prepared. The DATA-KBR-BE communication activities are detailed in *Section 7*.

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 third 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:

Data preparation: iterative data extraction, data quality and beyond
As we moved towards more in-depth analysis of the extracted data the clearer it became that data extraction is only a first step in the research process. For example, the initial data extraction (Data Extraction 1), one year for 3 Dutch Language and 3 French Language Belgian Newspaper titles were extracted, totaling around 500 GB of data. Following an initial core data extraction, **customised data extractions** seem much more relevant to the research teams, for example to help design a data model, or to develop a corpus. Furthermore, such **data extraction needs to be undertaken on an iterative basis**, possibly throughout the research lifecycle. For example, in an initial data extraction, required data could be missing, e.g. for the *Feuilleton in Belgium* when extracting the relevant feuilletons, the first instalment was at the end of the previous year. Alternatively, additional extractions are needed as a **comparative dataset**, for example for *Collective Action Belgium* a comparative strike year needed to be extracted.

This need for **customised, ongoing and iterative data extractions** points to the importance of robust **data extraction tools**, such as the ***KBR Data Extraction Tool*** developed by KBR's ICT and Digitisation Departments. This tool is simple and user-friendly and does not require a high-level of digital literacy as an Application Programming Interface (API) might. Currently, the *KBR Data Extraction Tool* is only available for selected KBR members of staff. However, in the future, it may be interesting to explore training multiple members of KBR staff (potentially from the Digitisation Department) to undertake '**Datasets on Demand**' data extractions. In the long term, it could be interesting **exploring making such a data extraction tool available for researchers and other users of KBR's collections**. However, the provision of such a service would require a more detailed investigation, for example, regarding *technical security* (e.g. prevention of malicious attacks on the service) and *performance* (e.g. size of the files that could be requested, how many requests could be handled by the servers etc.).

Additionally, it is important to emphasise that a **data sharing platform is essential.** Even though the

DATA-KBR-BE project team sometimes take it for granted now, the use of the **GhentCDH Next Cloud environment** (https://data.ghentcdh.ugent.be) remains crucial for the *interdisciplinary research scenarios.*

Data extraction is also only the first step in the research process. Once the data has been extracted, not only are further data extractions potentially needed, but also **additional processing steps**. For example, on several occasions the research teams noted that the quality of the OCR'd text was not sufficient for their needs or whether article segmentation or Named Entities could be extracted. It could be interesting for KBR to explore the recruitment of a "Research Software Engineer" to work together with researchers to help them prepare their data.

Regarding **OCR Quality**, a better understanding of where poor quality OCR is preventing the research team from undertaking their research is needed. It would be important to investigate a method to measure the quality of the original OCR as well as exploring ways in which the OCR could be improved. While experimenting with Collections as Data has several benefits, extracting the data does mean that the data is also more visible for scrutiny.

Towards a sustainable 'Collections as Data' platform at KBR

The DATA-KBR-BE Platform is intended to be a sustainable platform providing access to KBR's 'Collections as Data'. As the development of the DATA-KBR-BE platform became more concrete, the importance of fully **integrating it into the KBR website**, alongside KBR's core digital services such as Belgica, BelgicaPress, BelgicaPeriodicals and the General Catalogue, became more apparent. Furthemore, as the DATA-KBR-BE Platform is intended to be a long-term initiative, beyond the end of the DATA-KBR-BE project itself, anticipating a number of phases of development was very useful. Yet, at the same time, a focused approach on what is achievable during the lifetime of the DATA-KBR-BE project is paramount. The recommendation to focus on the delivery of **a simple, user-friendly website, for publishing curated datasets for download based on KBR's digitised and born-digital collections** for *Phase 1 of the DATA-KBR-BE Platform* will be key to its success.

However, the delivery of 'Collections as Data' at KBR does not stop at the publication of the platform. The design and publication of the datasets is an essential step in this process. In the original project proposal, the DATA-KBR-BE project team committed to co-curating and publishing at **least three Open Science datasets** on the platform. It is anticipated that these three datasets are likely to be related to the two original interdisciplinary research scenarios: 1) Collective Action Belgium and 2) Feuilleton in Belgium, as well as a dataset related to the additional research scenario, the History of Belgian Journalism. For the DATA-KBR-BE project team, the publication of the International Galleries, Libraries, Archives and Museums (GLAM) Labs Community's "**Checklist to publish Collections as Data in GLAM Institutions**", which is intended as an easy to apply method to encourage small and medium sized organisations to publish their digital collections as Collections as Data, came at an ideal moment.

Finally, there is already much interest to publish a variety of datasets on the emerging DATA-KBR-BE platform. The preparation of an **initial overview of the types of datasets that could be published on the DATA-KBR-BE Platform** which includes: a) *thematic datasets*, e.g. datasets related to the interdisciplinary research scenarios within DATA-KBR-BE or related to research projects, e.g. funded by BELPSO or other funding providers and b) *Collections as Data datasets*, e.g. datasets can include any data that is related to KBR's Collections. As part of KBR's emerging Digital Data Strategy, nine "**data types**", related to KBR's collections have been identified: *Bibliographic data; Authority data; Heritage data; Digital data; Collections as data; Born-digital data; Electronic resources; Research data* and *Lab data (code).* It could be interesting to consider publishing one dataset for each data type on the DATA-KBR-BE platform.

Stimulating and facilitating academic research using KBR's collections

It has been very encouraging to witness the ongoing analysis of the data extracted for DATA-KBR-BE's *interdisciplinary research scenarios.* Not only has Dilawar Ali (IDLab, UGent) successfully defended his PhD using KBR's Digitised Historical Newspaper Collections as a key case study, a number of (Digital) Humanities Masters Students have been undertaking internships at KBR on related topics. While not all of these research scenarios will result in fully-fledged research articles and publications, the process of using the DATA-KBR-BE project to move towards a sustainable data extraction workflow, on the basis of real-life research, has been invaluable. As a result, we understand the need for robust tools to enable iterative data extraction much better and realise that implementing frameworks for measuring, and where needed improving data quality ,will be essential to the success of ***extending KBR's services for researchers.***

5. FUTURE PROSPECTS AND PLANNING

WP1: Co-designing Interdisciplinary Research Scenarios. For the final reporting period, the focus of the work will be on the finalisation of the *interdisciplinary research scenarios*, in particular: a) preparation of the datasets for publication (*WP2: Preparation of Datasets*) and b) preparation specific webpages on the DATA-KBR-BE Platform providing information about the research scenarios (*WP4: Scientific exploitation and valorisation*). The dataset specific pages on the National Library of Scotland's Data Foundry (e.g. the Plague in the Punjab) will be of particular inspiration.

WP2: Preparation of Datasets. During the final reporting period, the DATA-KBR-BE project team will continue to work with researchers undertaking the scenarios to **design and prepare the datasets for each research scenario for publication on the DATA-KBR-BE Platform**. Any further **customised data extractions** will be undertaken as required. Additionally, depending on the time available, the publication of other '**thematic datasets**' as well as '**collections as data datasets**' will also be explored. For this work, the GLAM Labs' Collections as Data Checklist will be particularly useful. As part of the development of the common European data space for Cultural Heritage, this checklist is being transformed into a workflow for the Social Sciences and Humanities Open Marketplace. The DATA-KBR-BE team will use this workflow to guide the development of the datasets for publication. Once the datasets have been prepared for publication, a final description of the **sustainable data extraction pipeline** will be prepared, based on the use of KBR's Data Extraction Tool. At this stage, the **DATA-KBR-BE Data Management Plan** will also be updated. Additionally, **recommendations for the development of a KBR 'dataset on demand' service** will be prepared as well as proposals for **how the DATA-KBR-BE platform could be managed after the end of the project**. This will also consider the work that could be continued as part of the KBR Virtual Lab project.

WP3: Data access via data.kbr.be. With the **technical and functional requirements** well-developed (see: working document), a key focus of the work in the final reporting period is to **implement the DATA-KBR-BE platform**. Due to the challenges in hiring a DATA-KBR-BE Data Scientist (see: *Section 8* for further details) it is anticipated that this work could be undertaken by KBR's ICT Department. A detailed planning for the implementation of the DATA-KBR-BE platform will be prepared in Summer 2023, which will include a number of milestones to help finalise the publication of the platform. For example, an **alpha version of the platform** could be delivered in December 2023. This alpha version could be tested by KBR members of staff. A second **beta version** of the platform could be delivered in February 2024, incorporating the feedback from KBR staff. This beta version could be 'soft launched' during the Hackathon which also could be organised in February 2024 to gather feedback from the wider community. Finally **version 1.0 of the DATA-KBR-BE platform** could be delivered by the end of the project in mid-March 2024.

WP4: Scientific exploitation and valorisation. Complementary to the preparation of the datasets related to the research scenarios in WP2, a specific page on the DATA-KBR-BE Platform will be dedicated to contextualising the datasets resulting from the three interdisciplinary research scenarios: 1) Collective Action Belgium, 2) Feuilleton in Belgium and 3) the History of Belgian Journalism. The specific pages on the National Library of Scotland's Data Foundry (e.g. the Plague in the Punjab) are anticipated to be a particular source of inspiration. These web pages are intended to provide: a) **a contextual description of the datasets resulting from the research scenarios**, b) **provide links to the datasets related to the research scenario** and c) potentially **provide links to articles and other academic publications related to the research scenario.** If there is time available, the possibility of publishing other '**thematic datasets**' and '**collections as data datasets**', along with their related descriptions, will also be explored. These **dataset descriptions** should be visually appealing and written with both researchers and the general public in mind. This work will be undertaken in close collaboration with KBR's Communications Department.

The second focus of this work package in the final phase of the project will be the preparation of the **High**

Profile Hackathon. It is anticipated that the Hackathon, which would take place at KBR in early 2024, would be an ideal opportunity to **present the beta version of the DATA-KBR-BE Platform to the wider community**, towards the end of the project. The Hackathon could potentially be organised in collaboration with the project team from the common European Data Space for Cultural Heritage and the International GLAM Labs Community.

As reported previously, the DATA-KBR-BE team committed to publish at least *one interdisciplinary, peer-reviewed journal article* related to the project. This commitment has already been fulfilled with the publications of the article: "*Computer Vision and Machine Learning Approaches for Metadata Enrichment to Improve Searchability of Historical Newspaper Collections*" in the *Journal of Documentation*. The possibility of publishing an additional article in the final phase of the project will be explored.

WP5: Project Management and Communication. During the final reporting period, the DATA-KBR-BE project team will prepare *D5.2 Final Report*. In addition to reporting on the DATA-KBR-BE achievements throughout the project, it will explore recommendations for **how the DATA-KBR-BE platform could be managed after the end of the project**. It is intended that an online meeting of the Follow-up Committee will be organised in Autumn 2023 to present the **proposed implementation plan** of the DATA-KBR-BE platform for feedback and suggestions. The Scientific Advisory Board will also be invited for a final physical meeting, co-located with the High Profile Hackathon.

6. FOLLOW-UP COMMITTEE

During this reporting period, a meeting of the DATA-KBR-BE Scientific Advisory Board ('Follow-up Committee') was not held, however the DATA-KBR-BE project coordinator and several members of the DATA-KBR-BE 'Follow-up Committee' took part in a week long seminar at Schloss Dagstuhl – Leibniz Center for Informatics in Germany on *Computational Approaches for Digitised Historical Newspapers* from 17th-22nd July 2022. As part of the seminar, the DATA-KBR-BE Project Coordinator Sally Chambers gave an invited talk on *Newspapers as Data: Challenges and Solutions* where key issues related to the DATA-KBR-BE (such as *Collections as Data at KBR and the application to digitised historical newspaper collections; facilitating corpus building for the DATA-KBR-BE interdisciplinary research scenarios* and *methods for data extraction and sharing*) were discussed with key experts from the field. The key outcome of the workshop was an extensive Open Access Workshop Report.

The next DATA-KBR-BE Scientific Advisory Board ('Follow-up Committee') will be organised as an online meeting in Autumn 2023 to present the **proposed implementation plan of the DATA-KBR-BE platform** for feedback and suggestions. The final meeting of the Follow-up Committee will be a physical meeting, co-located with the High Profile Hackathon. Additionally, email updates are provided to the follow-up Committee between meetings.

7. VALORISATION ACTIVITIES

7.1 PUBLICATIONS

The DATA-KBR-BE Publications for this reporting periods are listed below:

Ali, D., Milleville, K., Verstockt, S., Van de Weghe, N., Chambers, S. and Birkholz, J.M. (2023), *Computer vision and machine learning approaches for metadata enrichment to improve searchability of historical newspaper collections*, Journal of Documentation. https://doi.org/10.1108/JD-01-2022-0029. A preprint version of the article is also available in the Ghent University Academic Bibliography.

Ali, D. (2023). AI-based methods to enrich, geolocalize, and rephotograph historical newspaper photos. Ghent University. PhD Thesis. Faculty of Engineering and Architecture, Ghent, Belgium. http://hdl.handle.net/1854/LU-01GRB7ZXBRFWB7VE1CCSWBWVV5

Candela, G., Gabriëls, N., Chambers, S., Pham T-A.; Ames, S., Fitzgerald, N.; Hofmann, K., Harbo, V., Potter, A., Ferriter, M., Manchester, E., Irollo, A., Van Keer, E., Mahey, M., Holowina, O. and Dobreva, M. (2023). *A Checklist to Publish Collections as Data in GLAM Institutions.* https://doi.org/10.48550/arXiv.2304.02603

7.2 PARTICIPATION/ORGANISATION OF SEMINARS (NATIONAL/INTERNATIONAL)

During this reporting period, the DATA-KBR-BE project team (co-)organised or participated in a range of (inter)national seminars and events, which are listed below. Upcoming events are also included.

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.

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.

Verstockt, S. (2022) *DATA-KBR-BE NewspAIper: AI-based metadata enrichment of historical newspapers* and video interview. European TimeMachine project.

Computational Approaches for Digitised Historical Newspapers, Schloss Dagstuhl – Leibniz Center for Informatics, Germany, 17th-22nd July 2022.

"Towards implementing Collections as Data in GLAM institutions", a webinar organised by the International GLAM Labs Community together with the DATA-KBR-BE project team on 25th October 2022.

Sharing and Sustaining Digitisation Knowledge. IMPACT Centre of Competence in Digitisation, 10th Anniversary Workshop and Writing Sprint. University of Alicante, 12-14th December 2022

Upcoming: Collections as Data: State of the Field and Future Directions. Internet Archive Canada, Vancouver, Canada, 25-26 April, 2023.

Upcoming: The DATA-KBR-BE project team submitted a presentation proposal - *Towards the DATA-KBR-BE platform: an iterative approach to publishing 'Collections as Data' at KBR, the Royal Library of Belgium* for the DH Benelux Conference 2023, 31st May - 2nd June 2023, KBR Royal Library of Belgium, Brussels.

Upcoming: Cultural Heritage Data as Humanities Research Data? DARIAH Annual Event 2023, 6th-9th June 2023, Eötvös Loránd University, Faculty of Humanities, Budapest, Hungary.

Upcoming: CENL Dialogue Forum: National Libraries as Data Infrastructures. Presentation at the *Conference of European National Librarians (CENL) Annual General Meeting, Bibliothèque nationale de France (BnF), Paris, 18th-20th June 2023.*

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 for digital humanities researchers. Furthermore, 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**. At the time of writing, both of these strategies are in their final stages of development and are anticipated for publication in Autumn 2023. Both of these activities contribute directly to **KBR's Action Plan for 2022-2024.**

The DATA-KBR-BE team is **continuing to investigate opportunities to continue the work of DATA-KBR-**

BE, both in Belgium, e.g. via other BELSPO funding programmes such as ESFRI-FED or European Funding, e.g. Horizon Europe. As mentioned in the previous reporting period, the ***KBR Virtual Lab: e-infrastructure for facilitating access and research of KBR's collections as data*** has been successfully funded by the Belspo, ESFRI-FED Programme. The project is expected to start in Autumn 2023. The *KBR Virtual Lab* can be seen as a complementary project to DATA-KBR-BE as it focuses on providing computational access to KBR's digitised and born-digital collections via Application Programming Interfaces (APIs) or 'data on demand' services. In the final phase of DATA-KBR-BE, the work plans of the two projects will be aligned to see to what extent the KBR Virtual Lab project can continue to develop and sustain the DATA-KBR-BE Platform.

Regarding European funding, as reported in the last reporting period, 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). Despite good reviews, unfortunately the proposal was not funded on this occasion. A second project proposal entitled: ***ENRICHES: ENRiching and Interlinking Cultural HeritagE Sustainably***, also led by the University of La Rochelle, was submitted in March 2023 to the Horizon Europe Call: *HORIZON-CL2-2023-HERITAGE-01-03: Re-visiting the digitisation of cultural heritage: What, how and why?* This proposal further develops the research undertaken in the DATA-KBR-BE project at European scale. It also links the Collections as Data work at KBR to the emerging large scale European initiatives such as the *Common European Data Space for Cultural Heritage* and the *European Collaboration Cloud for Cultural Heritage*.

Sally Chambers (DATA-KBR-BE Project Coordinator) participated in a meeting organised by BELSPO in relation to the **establishment of a Belgian Federal Open Science Cloud (FedOSC)** in February 2023. As a follow-up from this meeting, she contributed to a nota which was prepared regarding *Strengthening Data Stewardship Expertise at the Belgian Federal Scientific Institutions*. This nota stressed the importance of domain-specific data management expertise to ensure a high-quality Federal Open Science Cloud. The research being undertaken in the context of DATA-KBR-BE could provide a valuable contribution to these discussions.

7.4 OTHER

n/a

8. ENCOUNTERED PROBLEMS AND SOLUTIONS

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

The formal approval by BELSPO of the extension of the DATA-KBR-BE project in February 2022 for a period of 24 months to 15.3.2024 was much welcomed by the project team. This required the scheduling of the project to be adjusted in light of the extension. In principle, the original structure of the timing of tasks and deliverables could be retained. The only change which was required was the duration of the tasks, which needed to be doubled in length. This could be implemented easily without disrupting the flow of the project. Two additional Annual Reports for the periods: *15.03.2021-15.3.2022* and *15.03.2022-15.3.2023* were requested by BELSPO. They were added to the project plan and both delivered successfully, albeit with some minor delays.

As reported in the previous Annual Report, the recruitment of the KBR Data Scientist was intended to be a key activity for this reporting period. However, due to the indexing of the salary costs in line with the rate of inflation following the Covid health crisis, there was a reduction of the available budget for the data scientist (from 66,000€ to 58,000€). Additionally, the KBR Virtual Lab project had experienced one unsuccessful round of recruitment for a Scientific Programmer in Summer 2022. Following consultation with BELSPO, it was agreed to combine the efforts of the DATA-KBR-BE and KBR Virtual Lab projects, to advertise for a Scientific Programme for both projects. This would both strengthen the synergies between the two projects and provide KBR with the opportunity to offer a longer term contract to the successful candidate. A joint call for applicants was launched in March 2023 with a closing date of 26th March 2023.

Finally, the extraction of the datasets for the research scenarios increased the visibility of the quality of the automatically recognised textual layer of the digitised newspapers using Optical Character Recognition (OCR). In some cases, the research teams reported that the quality of the OCR was insufficient for the researchers to undertake the research as they had anticipated. As a result, it was recognised that a better understanding of how the OCR quality is preventing the research teams from undertaking their research is needed. For example, it would be important to investigate a method to measure the quality of the original OCR as well as exploring ways in which the OCR could be improved. The improved OCR layer could then be ingested back into BelgicaPress. These OCR quality issues are not a blocking factor for the DATA-KBR-BE project, but they do represent an issue that requires further investigation in the final period of the project.

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 / C2DH, University of Luxembourg
- Steven Claeyssens, National Library of the Netherlands
- Wout Dillen, University of Borås, Sweden
- 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.