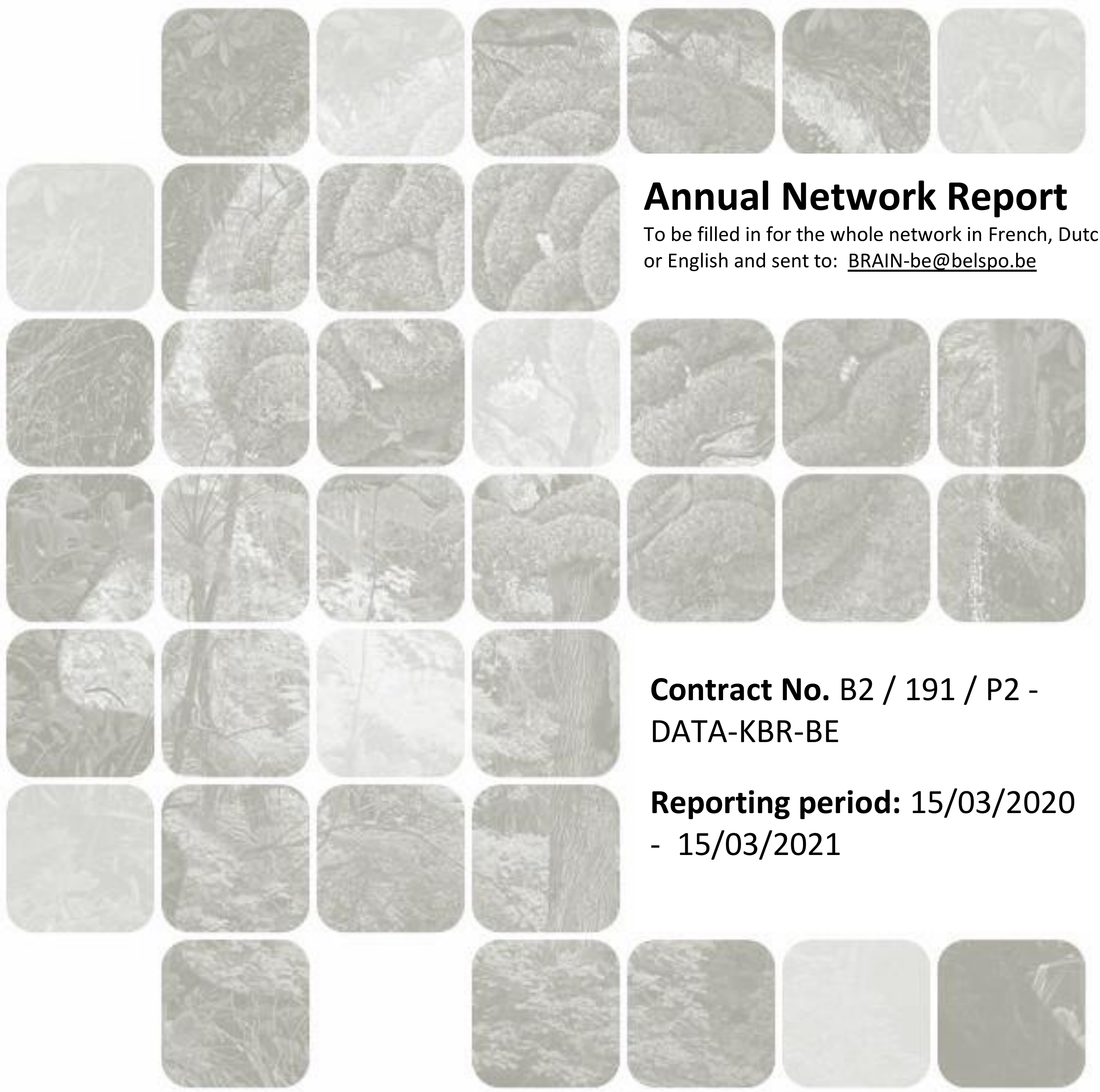
BRAIN-be 2.0

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: 15/03/2020 - 15/03/2021

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

AUTHORS OF THIS REPORT

1. Sally Chambers: KBR, Royal Library of Belgium and Ghent Centre for Digital Humanities, Ghent University

2. Frédéric Lemmers: KBR, Royal Library of Belgium

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

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

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

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

7. Wout Dillen: Antwerp Centre for Digital Humanities and Literary Criticism (ACDC), University of Antwerp

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

9. Antoine Jacquet: KBR, Royal Library of Belgium and Université libre de Bruxelles, Sciences de l'information et de la communication Department

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

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

PROJECT WEBSITE, SOCIAL NETWORKS …

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

TABLE OF CONTENTS

1. EXECUTIVE SUMMARY OF THIS REPORT

DATA-KBR-BE: facilitating data-level access to KBR's Collections for Open Science is a 24 month project (2020-2022) 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. 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 three interdisciplinary research scenarios using thematic datasets extracted from KBR's digitised historical newspaper collection, BelgicaPress.

This report provides an update on the achieved work, intermediary results and preliminary conclusions and recommendations for the first reporting period of the project (15.3.2020 - 15.3.2021). It also outlines the future prospects and planning for the next reporting period. The collaboration with the DATA-KBR-BE Scientific Advisory Board Follow-up Committee, who provide scientific and guidance for the project is also described. Furthermore, valorisation activities, challenges including potential solutions, as well as any modifications to the project planning, e.g. personnel changes, since the initial report are also included.

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 this first reporting period (15.3.2020 - 15.3.2021) per work package are outlined below:

WP1: Co-designing Interdisciplinary Research Scenarios - led by UGent (M1-M6)
The aim of this work package is for the researchers in Ghent and Antwerp to work closely with the KBR's collection, digitisation and ICT experts to co-design *two interdisciplinary research scenarios* that can be used as *a basis for extracting relevant thematic datasets* in WP2. Building on the descriptions of the research scenarios described in the original proposal, detailed descriptions of the research scenarios were prepared, see: 1) Collective Action Belgium, led by GhentCDH and 2) Feuilleton in Belgium, led by ACDC.

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

From the outset, the data science team at IDLab (UGent) started to experiment with an *initial test dataset* from the KBR's collection of digitised historical newspapers, BelgicaPress. Due to lockdown restrictions as a result of the Covid-19 pandemic, remote access to the KBR's digitised collections meant that creative solutions needed to be found to prepare the initial test dataset. Luckily, the KBR had already extracted a dataset to a hard disk for the GhentCDH team for previous research on the newspaper collections. This existing dataset consisted of one French language (Le Peuple) and one Dutch Language (Vooruit) newspaper from BelgicaPress for the period 1886-1938. This enabled colleagues from IDLab to undertake their initial experiments despite the lockdown. (For further details about the *DATA-KBR-BE Datasets, see WP2*).

The first experiments that were undertaken with initial test dataset related to the *automatic detection and extraction* of the different 'parts' of a newspaper, e.g. images, captions, titles, text blocks, columns etc., using the XML-ALTO files, in combination with the pdfs and images to detect the different article sections, text blocks, columns etc. A number of activities were undertaken by the IDLab team:

- *Detecting and extracting images (e.g. photographs, drawings etc.) and their related caption* along with metadata related to the extracted images, (e.g. which newspaper, which edition, on which page etc. the image appeared). In future phases of the project, it would be interesting to see if the International Image Interoperability Framework (IIIF) could be useful to create links from the dataset to the main collection, BelgicaPress. GhentCDH's IIIF annotation and crowdsourcing platform, MADOC, might be useful here.
- *Author signature detection*: detection and recognition of the 'journalist signatures' (i.e. the authors initials at the end of a newspaper article). The, initially controversial, practice of adding initials to newspaper articles came relatively late in Belgium (ca. from 1890s onwards). Links could be provided to the database of Belgian Journalists being created within CAMille.
- Experimenting with *image similarity algorithm***s** to identify recurring sections of the newspaper, e.g. images and captions or literary supplements (Feuilletons). Once training models have been built, they can be used for identifying other Feuilletons within BelgicaPress.

For example, the research undertaken by Dilawar Al (IDLab), under the supervision of Prof. Steven Verstockt, initially focussed on *photographs and caption detection* (see Figure 1 below):

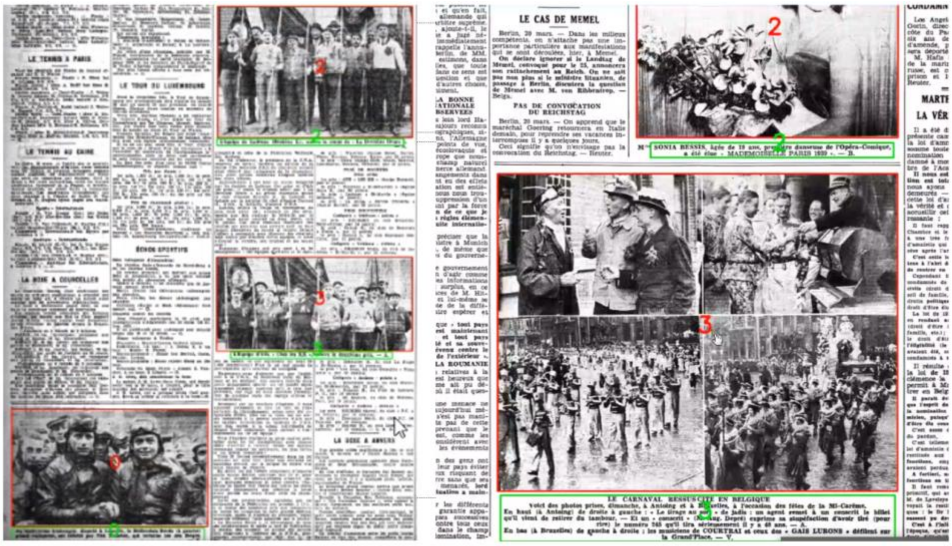


Figure 1. Initial results to automatically detect photographs and captions in digitised historical newspapers undertaken by IDLab.

From these initial experiments, it could be observed that *poor quality Optical Character Recognition (OCR)* could cause issues, such as in identifying the bounding boxes, spacing and overlapping boxes (e.g. picture captions being merged with article text). Some inconsistency in the XML-ALTO files were also found (e.g. between the 'element types', even within the same newspaper title). It was suggested that the *re-OCRing of some selections of the sample texts* (e.g. re-OCRing of the captions) with State-of-the-Art OCR software may improve the results. The idea would not be to re-OCR all the whole sample, but to demonstrate how State-of-the-Art newspaper techniques could improve performance. These initial tests have created the first building blocks for a *pipeline for generic image extraction*, which are also a first step towards extracting the literary supplements for the Feuilleton in Belgium research scenario.

In Summer 2020, the IDLab team requested a larger dataset to further improve their algorithms. To create this dataset it was important to get a better idea of which newspapers include illustrations. To do this, the DATA-KBR-BE team decided to analyse the editions of the newspaper from 1 January, 1 June and 1 December for each of the years available in the initial test dataset. As a result of this manual visual analysis, it could be seen that the first illustrations (i.e. drawings rather than photographs) started to be published from around 1901/1902 onwards. From 1924 onwards the first photographs started appearing. Le Peuple from 1938 was the latest edition of the newspaper in the initial test dataset and it seemed to contain the most illustrations. It was therefore chosen for the larger dataset. Due to the lockdown restrictions, this transfer was undertaken on a home internet connection (4 transfers, each of one per quarter year using Belnet FileSender total of 34 GB of the TIFF files of Le Peuple from 1938). Following the sharing of a large test dataset, Dilawar Ali built his training model based on 2500 images from one month (January 1938) of the newspaper, using the YOLO (You Only Look Once) image detection framework. These initial results were presented by Dilawar Ali at the What's Past is Prologue: the NewsEye International Conference in March 2021 and showcased in the NewspAIper demonstrator (see Figure 2 below).

The NewspAIper demonstrator, which uses Le Peuple from 1938, demonstrates the potential of: a) article segmentation, b) linking text recognition with open data, c) finding similar images across the collection and d) initial ideas for an interactive filter which could be implemented within BelgicaPress. This is also a first step towards the development of an automated pipeline for the enrichment of digitised historical newspaper collections.

NewspAIper

Metadata extraction of historical newspaper archives.

Figure 2. NewspAIper Demonstrator developed by the ID Lab team in the context of DATA-KBR-BE

WP2: Preparation of Datasets - led by KBR (M7-M24)

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.

Within this first phase of the project, it was agreed to organise a *DATA-KBR-BE Brainstorming Workshop,* which was held online on 27th November 2020*.* The aim of the workshop was to bring together the DATA-KBR-BE team to brainstorm ideas to help us to develop a sustainable workflow to facilitate research use of KBR's collections. The programme of the workshop, which lasted 2.5 hours, was divided into four sections: *1) Scoping Exercise: What did we promise?, 2) Review of Existing Library Data Platforms, 3) Preparation of KBR Datasets* and *4) Sustainable Data Extraction Workflow.* The Workshop Presentation slides were used to facilitate the discussion. A workshop report was drafted following the meeting.

Key considerations: The workshop resulted in a very fruitful discussion which helped to identify issues that the DATA-KBR-BE project needs to consider, for example:

- Should DATA-KBR-BE provide access to *in-copyright material* or only material that is already in the *public domain*? Are there examples of existing library data platforms that have provided access to in-copyright materials?
- How do existing library data platforms *licence* and provide *data citation* for their datasets?
- It is crucial to clearly describe which *file formats* (e.g. jpegs, tiffs, pdfs, .txt files etc.) are needed for the research scenarios.
- Further information about the *OCR quality* of the various newspapers in BelgicaPress is needed. The quality of the OCR may be one of the criteria to take into account when making a choice as to which newspapers to use? Perhaps we could run some statistical texts on the OCR quality prior to selection?
- Would providing data-level access to all Belgica Press be feasible by the end of the DATA-KBR-BE project? How could DATA-KBR-BE *provide an extended service* for users of BelgicaPress? For example, would *fully dynamic data extraction* be an option? e.g. a researcher searches in Belgica Press and then is able to do an export of all the data they choose?
- *Would a 'Dataset on Demand' like service be possible at the KBR within the DATA-KBR-BE project?* The Datasets on Demand from Royal Library of Denmark could be an inspiration here? Perhaps KBR's Digit on Demand service, which was scaled up as a result of the Covid-19 pandemic, provides a good starting point for moving towards 'Datasets on Demand'.

Finally, the approach for extracting the thematic datasets to support the research scenarios was discussed. Three potential scenarios were proposed: i) *Scenario 1:* one dataset per research scenario, if so, which one to start with?, ii) *Scenario 2:* Iterative: shared dataset with 3 iterations and iii) *Scenario 3:* 1st version = test dataset and decide later 2nd and 3rd datasets. The DATA-KBR-BE team decided to move forward with *Scenario 3*: first a shared test dataset and on the basis of the experience gained with this initial dataset to decide later how to proceed with the 2nd and 3rd datasets.

Extraction of thematic datasets to support research scenarios

Using the approach (scenario 3) identified as a result of the *Brainstorming workshop* in November 2020, the DATA-KBR-BE team set about to design the first (DATA-KBR-BE Dataset 1) of the three thematic datasets to support the interdisciplinary research scenarios. To do this, two workshops were organised in March and April 2021.

DATA-KBR-BE Dataset 1 Workshop, 30th March 2021: The aim of this workshop was to: a) design a 'test' dataset that can be used to support all three research scenarios i) Feuilleton in Belgium; ii) History of Belgian Journalism and iii) Collective Action Belgium) and can act as a test dataset for experimentation; b) agree the criteria for developing DATA-KBR-BE datasets, c) agree a 'first draft' of the DATA-KBR-BE Dataset 1 and d) agree concrete action and next steps (see: Workshop Agenda).

The first part of the workshop focussed on designing criteria for developing the DATA-KBR-BE datasets. Together the group brainstormed a number of criteria: *a) Research criteria*: research scenario, research questions, research tools and methods and outputs, *b) Collection criteria*: which collection, legal and ethical considerations, which newspaper titles, *c) Digitisation criteria***:** OCR quality, *d) Extraction criteria:* newspaper titles (including time period, issues and pages), whether images are needed, languages, file formats, file size, *e) Data Sharing:* e.g. KBR Send File, Belnet FileSender, hard disk etc. Other more general criteria were also considered such as: whether to *create a Data Management plan per research scenario***,** how to *cite the dataset* etc. The interconnection between the different criteria was also highlighted.

Following agreeing the overall set of criteria for the dataset, the group worked towards preparing a 'first draft' of DATA-KBR-BE Dataset 1. Research outputs prompted a lot of discussion, with the general consensus being that if feasible, *one peer-reviewed article per research scenario*. It was noted that for the computer and data science colleagues that such articles should be published in high-ranking journals recognised by the Computer

Science community. In order to finalise the draft of dataset 1, it was decided that further details, including where possible examples, were needed. It was agreed to organise a second follow-up workshop to discuss the research scenarios in more depth.

DATA-KBR-BE Dataset 1 - Follow-up Workshop, 22nd April 2021: During the follow-up workshop related to Dataset 1 (see: Workshop Agenda), further details were provided on each of the research scenarios. The coordinators of each of the research scenarios outlined their research scenarios in more detail, providing examples where possible.

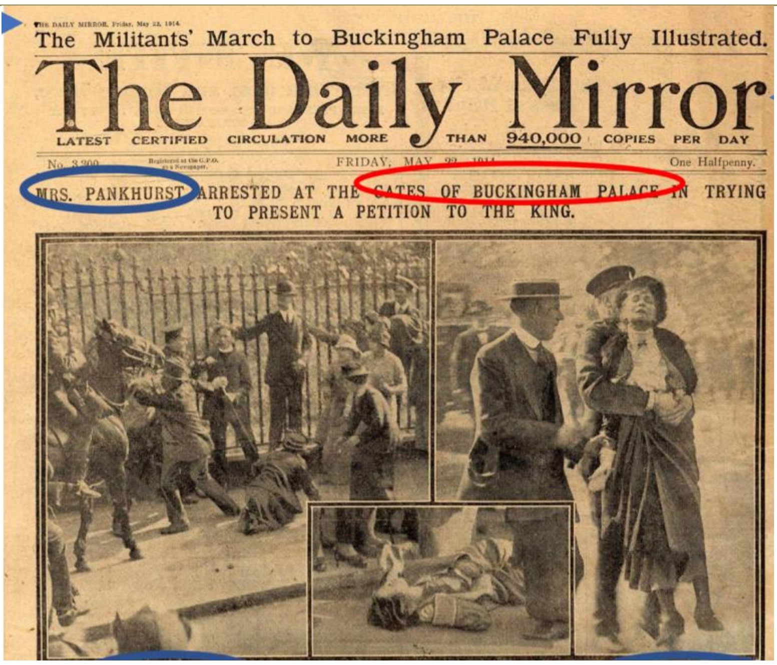
1. **Feuilleton in Belgium: (see Presentation)**

Emerging in France and starting out as non-news articles and non-political information, Feuilleton (or Mengelwerk in Dutch) became focussed on fiction, both new and republished, becoming the 'home of the serialised novel'. These literary supplements received some criticism from the Church, as they had commercial aspirations and were partially intended to keep people interested in buying the newspapers. They became ubiquitous forms of fiction, featuring prominent authors, particularly in Belgium's formative years. As the Feuilletons' have a *combination* of distinct visual features, intended to enhance their commercial advantage, they regularly appeared in the bottom third of the page, often extended the full breadth of the page and appeared on the front or the second page. They had distinctive titles and ended with words such as 'to be continued' (Wordt voortgezet). As they were serialised literary works the number of the series was often mentioned (e.g. 5e vervolg).

It is proposed to use all of the KBR's digitised issues of six newspapers from 1885. 1885 was selected as this was the first year where there were regularly at least three Dutch language (Het Handelsblad, Vooruit and De Koophandel) and three French language (Gazette de Charleroi, La Meuse and L'Echo du Parlement) newspapers. By 1885 the Feuilleton 'fashion' had become fairly standard. Furthermore 1885 was within the first century of the Belgian nation. Using the corpus defined above, the next step would be to train a model to *identify and extract the Feuilletons into a corpus.*

2. **Collective Action Belgium (see: Presentation)**

A core aspect of the Collective Action Belgium research scenario focuses on the modelling of historical events, such as strikes, demonstrations, protests and other forms of collective action as reported in historical newspapers. A first step towards this is to build a corpus based on relevant keywords in the titles and full-text of the newspaper articles from a certain period. For example, 1st May - Labour Day - is known to be a popular date for various forms of collective action. Extracting editions of newspapers within, for example, a two week period before and after 1st May could be one way to build a relevant corpus. For *Historical Event Modelling*, it is important to be able to capture key metadata of interest, such as: ***who** was there? **What** happened**?** **When** was it? **Where** was it?* For example: Mrs Pankhurst (**Who**), was arrested (**What**) at the gates of Buckingham Palace (**Where**) as well as metadata about where this information was reported (e.g. publication, date, page number), i.e. on the front page (p1) of the Daily Mirror of Friday 22 May 1914, see Figure 3. below.



THE DAILY MIRROR, Friday, May 22, 1914.

The Militants' March to Buckingham Palace Fully Illustrated.

The Daily Mirror

LATEST CERTIFIED CIRCULATION MORE THAN 940,000 COPIES PER DAY

No. 3,200 Registered at the G.P.O. as a Newspaper. FRIDAY, MAY 22, 1914 One Halfpenny.

MRS. PANKHURST ARRESTED AT THE GATES OF BUCKINGHAM PALACE IN TRYING TO PRESENT A PETITION TO THE KING.

Figure 3. Examples of metadata (e.g. publication, date, page number, who, where) relevant for historical event modelling

Other techniques to explore this research scenario could include, for example, plotting a histogram based on the frequency of words, over time including locations. KBR's Digital Research Lab will play an active role in this research scenario, in terms of translating research needs into data needs and workflows. Alternatively, simple modelling of the event could include plotting the event as a single point on a map.

In order to build the corpus, the years of the General Strikes in Belgium (1886, 1893, 1902, 1913, 1932 and 1936) were used as a first criterion in selecting which newspaper titles to use. As a second step, we looked at which newspapers in Belgica Press were published in Ghent, as this is the location of focus for this research scenario. There were five daily newspapers: De Gentenaar : katholiek dagblad; Vooruit: socialistisch dagblad; Vaderland; De Vlaamsche post : algemeen dagblad voor Vlaanderen and Het Volk : antisocialistisch dagblad Additionally, there was one weekly newspaper: Vlaanderen - Jong Dietschland. We then compared the data of the General Strikes with the years of the five daily newspapers which had been digitised. As a result, for the first extraction, 1913 (which was also the 2nd year of strikes on electoral reform and also the year of the World Fair in Ghent) was the year with the most digital coverage of the newspapers. To ensure the comparability of the approach with the Feuilleton in Belgium research scenario, we decided to select three Dutch language newspapers: 1) Vooruit: socialistisch dagblad, 2) Het Volk : antisocialistisch dagblad and 3) Vaderland and three comparative French language newspapers 1) Le Peuple : organe quotidien de la démocratie socialiste, 2) Le Vingtième Siècle and 3) La Meuse : journal de Liège et de la Province. In later data extraction, these initial 6 newspapers could be complemented with data from other strike years.

Following the first data extraction, a first step would include a full-text search to find relevant keywords in articles and titles. Once the relevant articles (from the relevant titles over the relevant period) have been identified, Named Entity Recognition (NER) to find the people or key actors in the historical events, as well as Geographic Entity Recognition (GER), where these events happened. Additional use cases within the research scenario could include developing object recognition models for grayscale images, to classify them and make them searchable. This research scenario will connect with Vincent Ducatteeuw's PhD research to create an authoritative list of place names in Ghent (including street names), as a first step towards creating an urban gazetteer for Ghent. The data extracted from the relevant digitised newspapers could be enriched and linked to external references, such as in Wikidata. This research scenario will also connect with GhentCDH's Gent Gemapt project. An important first step would be to understand in what format the data that is extracted from the historical newspapers is needed in order to create the urban gazetteer, e.g. .csv format.

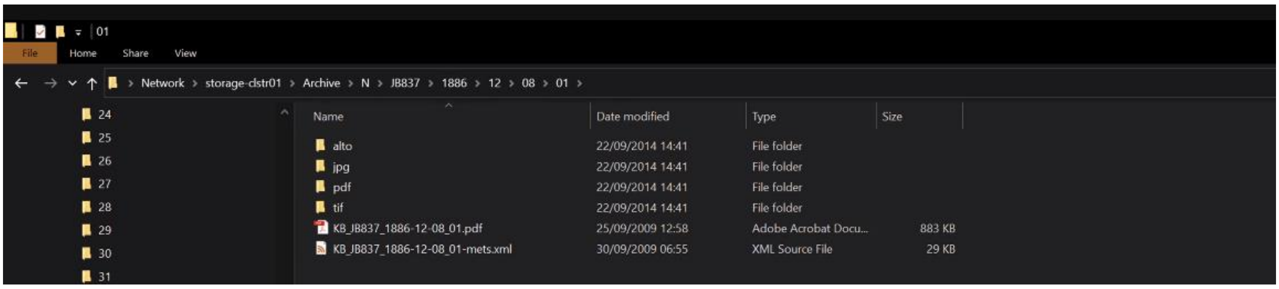
3. History of Belgian Journalism

The 'History of Belgian Journalism' research scenario is part of the broader programme of research activities within CAMille, the Centre for Archives on the Media and Information, ULB-KBR. For this initial case study within DATA-KBR-BE, Antoine provided further details about his text analysis case study using the French-language Roman Catholic newspaper Le Vingtième Siècle (JB729), which focuses on extracting Named Entities using *R Studio* for the period 1895 - 1940. To date, the CAMille research team has a list of 45 journalists with links to *Le Vingtième Siècle*. The goal of this use case is to identify 'new names' in addition to this initial list. The CAMille team are interested in how the newspaper referred to the journalists within the newspaper itself, for example, "our excellent journalist" and whether this changed over time, for example after their retirement. The aim is to (re)trace the careers of Belgian journalists (e.g. Léon Degrelle) through the newspapers that they wrote for. Other use cases include the detection of "journalistic signatures", i.e. initials at the end of articles, including in which years the journalist was active. The obituaries of journalists could also be of interest. For this initial research scenario, Antoine is using the text files for the whole run of the newspaper (1895 - 1940), which amounts to ca. 3GB. For close-reading of an article, he refers to the pdf (which he also has in his corpus). A further research question is to investigate the meta-discourse about journalism in Belgium. The quality of the OCR is important for Antoine's research scenario. So far, the list of 45 journalists led to 14,000 references in newspaper articles. Article segmentation would be useful for Antoine's research. For this, the IDLab team would need a sample list of 100-200 articles, ideally spread over time.

DATA-KBR-BE Data Extraction Workflow meeting, 7th May 2021 (Thuy-An Pham, ICT Department and Sally Chambers, DATA-KBR-BE Project Coordinator). The aim of this meeting was to take the first steps in understanding how 'DATA-KBR-BE Dataset 1' would practically be *extracted* from the KBR's ICT Infrastructure and *shared* with the DATA-KBR-BE project researchers. This first meeting took the format of a semi-structured interview / informal conversation to understand 'behind the scenes' of the KBR's ICT infrastructure.

Close inter-departmental collaboration within KBR is essential for the long-term success of DATA-KBR-BE. A key activity within the project is to bring together experts from across the different departments across the KBR including digitisation, ICT, collection managers and metadata experts. A large-scale project within the KBR is currently underway to optimise the technical infrastructure that supports the KBR's digitisation workflow, digital asset management and long-term archiving. This process provides the DATA-KBR-BE team with an ideal opportunity to directly contribute the requirements for the design of a sustainable data extraction workflow within KBR's data infrastructural activities. To this end, the project can embed its research results into KBR's day-to-day operations from the outset.

As the DATA-KBR-BE project focuses on the KBR's Digitised Newspaper Collection, BelgicaPress in the first instance, the first step was to understand how the digital files related to BelgicaPress are stored. Within the KBR's existing Data Infrastructure, all the digitised newspapers are organised in a single archive "N", organised by Newspaper Title / Shelf Number (e.g. JB 837 for *Le Peuple*), then by year, month, day with separate folders for the various related files, e.g. XML-ALTO, JPG, PDF, TIFFs etc., see Figure 3 below.



Network › storage-dstr01 › Archive › N › JB837 › 1886 › 12 › 08 › 01 ›

Name	Date modified	Type	Size
alto	22/09/2014 14:41	File folder	
jpg	22/09/2014 14:41	File folder	
pdf	22/09/2014 14:41	File folder	
tif	22/09/2014 14:41	File folder	
KB_JB837_1886-12-08_01.pdf	25/09/2009 12:58	Adobe Acrobat Docu...	883 KB
KB_JB837_1886-12-08_01-mets.xml	30/09/2009 06:55	XML Source File	29 KB

Figure 3. A screenshot of the file structure for KBR's digitised newspaper archive (before 'Stock')

This file structure makes it easy to transfer (even if time-consuming) to transfer large batches of data, e.g. Le Peuple for 1886. However, it is less conducive for extracting particular subsets of the data, e.g. edition of the

1st May for each year of publication of the newspaper. Once the new 'stock' system will be in place it will be easier to extract specific documents (for the KBR's newspaper collection one document = one day of a particular newspaper. It will be possible for users to select up to 10 documents (i.e. 10 days of news) and download the related files. The 'stock' system creates a copy of files and adds them to 'queue of downloads' for sending to the user for example, using a secure File Transfer Service such as KBR Send File or Belnet's FileSender. This process could be automated so that the 'dataset' is emailed automatically to the user without need for human interaction.

In the first version of 'Stock' scheduled for release in Summer 2021, this functionality will initially be for KBR staff only. However, in later versions of the system, this '*corpus building functionality*' could be potentially extended to external KBR users, for example, via the BelgicaPress interface. This functionality could be the first step towards a 'dataset on demand' service, inspired by the KBR's 'digitisation on demand' service which has been particularly successful especially when due to the global Covid-19 pandemic, the KBR could not be open for readers. Furthermore, digitised historical newspapers are 'heavy' in terms of file sizes. For example, for the full run (1856 - 1950) of La Meuse, is estimated at ca. 183 GB (uncompressed), 168 GB (compressed). Due to file transfer speeds, it may still be necessary to share large datasets with researchers using more conventional methods such as copying the files to a hard-disk. However, an online secure environment of 'data sandbox' could be an interesting option to explore.

DATA-KBR-BE Dataset 1: As the migration of the data to the new 'Stock' system is planned for Summer 2021, and the downloading of larger datasets may be more difficult after the migration, then perhaps we should transfer as much as possible prior to the migration. Regarding practicalities, it could be interesting to explore whether the files in the DATA-KBR-BE Dataset 1 could be transferred to a 'DATA-KBR-BE' folder on the KBR's file server. We could then create 'subsets' of the data as needed for the various research scenarios, which could be sent to the researchers involved via secure file transfer or via hard-disk. The GhentCDH hardisk has, for example, storage capacity of 2TB (with 623 GB) remaining. However, we could buy a new hard disk(s) if needed for the various research scenarios.

Regarding the more 'sustainable' solution for data sharing, this could be anticipated in a number of phases, which could be aligned with the planning of KBR's ICT department. For example, remote access to the 'DATA-KBR-BE folder on the KBR's file server for DATA-KBR-BE affiliated researchers (e.g. at Ghent or Antwerp University). Remote access to the KBR's file servers is currently only permitted via a VPN from a KBR Laptop. Within DARIAH-BE (via CLARIAH-Flanders), a pilot project is currently underway to provide secure access to humanities research datasets via the Flemish SuperComputer. This would also facilitate computational analysis of the datasets (either using High Performance Computing Capacity or via lighter-weight solutions such as via Jupyter Notebooks). The Royal Danish Library's Cultural Heritage Cluster is a source of inspiration for Belgium.

WP3: Data access via data.kbr.be - led by KBR (M4-M24, October 2020 - June 2022)
The goal of WP3 is to: a) design, b) implement and c) test the usability of the new data.kbr.be data platform. The implementation of the *data.kbr.be platform* will include the Open Humanities datasets prepared in WP2, and the development of a *Digital Asset Registry* to inventorise KBR's digital (digitised and born-digital) collections as well as collections that are currently in KBR's digitisation pipeline.
In the *DATA-KBR-BE Brainstorm Workshop* which took place in November 2020 (see further details on above), a review of the functionality of existing library data platforms was undertaken. Prior to the workshop, a selection of data platforms were provided to participants:

- National Library of Scotland's Data Foundry
- Austrian National Library's ONB Labs Datasets
- British Library's Digital Collections and Data
- National Library of the Netherland's Data Services and APIs
- Danish Cultural Heritage Cluster (large scale data access)
- British National Bibliography (BNB) Linked Data Platform (Linked Open Data/SPARQL endpoint)

However, during the workshop itself, it was decided to focus on one particular platform, the National Library of Scotland's Data Foundry which offered many of the anticipated functionalities needed for data.kbr.be, as well as to focus the conversation. There are three core underlying values to the Data Foundry: 1) *Open*: the National Library of Scotland publishes data openly and in re-usable formats, 2) *Transparent*: the provenance of the data is taken seriously, and the library is open about *how* and *why* it has been produced and 3) *Practical***:** the datasets are presented in a variety of file formats to ensure that they are accessible as possible. At the time of the workshop (November 2020) the Data Foundry provided access to four core data collections*: digitised collections, metadata collections, map and spatial data and organisational data***.** As DATA-KBR-BE will initially focus on the Belgica Press, the KBR Digitised Newspaper Collections, the Data Foundry's digitised collections data collection was explored in more detail. More detailed information about the Data Foundry and its role in digital scholarship at the National Library of Scotland can be found in this article: Ames, S., & Lewis, S. (2020). *Disrupting the library: Digital scholarship and Big Data at the National Library of Scotland*. Big Data & Society. https://doi.org/10.1177/2053951720970576. Additionally, the *National Library of Scotland's Open Data Publication Plan* was also particularly of interest to the DATA-KBR-BE project team.

WP4: Scientific exploitation and valorisation - led by UAntwerpen (M10-M24, April 2021 - June 2022)

The aim of WP4 is to *carry out the interdisciplinary research scenarios* co-designed in WP1 using the *thematic datasets* extracted in WP2 and published on *data.kbr.be* in WP3. Originally in the project proposal it was anticipated that these three steps: a) research scenario design, b) dataset preparation and c) scientific exploitation would occur sequentially. However, after the initial start-up phase of the project, this process was more iterative rather than linear or sequential. The DATA-KBR-BE project team therefore agreed that the three steps should continuously run in parallel for each of the research scenarios throughout the project. As agreed in the 'DATA-KBR-BE Dataset 1 workshop', our aim is to prepare at least one interdisciplinary, peer-reviewed journal article as the DATA-KBR-BE team, potentially for DSH: Digital Scholarship in the Humanities. If possible, we would like to additionally prepare one peer-reviewed journal article per research scenario.

Further activities in this work package included initial brainstorming regarding the DATA-KBR-BE Hackathon (e.g. potential timing, Save the Date announcement, Audience, Goals etc) which is scheduled to take place towards the end of the project. This work will be following the extraction of DATA-KBR-BE dataset 1. Finally, in October 2020, an initial meeting took place between KBR staff (including members of the DATA-KBR-BE project team) and colleagues from the State Archives of Belgium, about the KBR's requirements for publishing research data in the SODHA, Social Sciences and Digital Humanities Archives, data repository. This meeting instigated a wider discussion where questions such as: how is research data distinct from all the other data types held by the KBR? should it be valorised differently? how does it connect to the current and future KBR Action Plan etc. Consequently there emerged a need to map the data we have at KBR and develop a digital strategy for managing this data.

Starting in January 2021 a number of interviews were conducted with those present at the original meeting in December. In these semi-structured interviews (where everyone received the same set of questions, which could also be expanded on depending on the conversation) or during the gathering of data stories, we spoke about what is data based on each person's expertise, how it is used and reused, what is needed, what is missing, etc. The results of these interviews were summarised to that group by Julie and Sally in March 2021, with a proposal for next steps for further developing a first proposal of this strategy were presented to the KBR's Management Team in May 2021. It is anticipated that the DATA-KBR-BE project will have a pivotal role in the further development of KBR's Digital Data Strategy.

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

The aim of WP5 is to ensure the timely management and monitoring of the project, including liaison with BELSPO, organisation of Follow-up Committee meetings and communication activities. The DATA-KBR-BE project team has met on a regular basis since February 2020. These meetings take place, usually about once a month, depending on the needs of the project and during this reporting period have more or less exclusively been online. Informal notes and action points from the meetings are written in the DATA-KBR-BE Rolling Notes document. This regular meeting schedule seems to work well. As and when needed, meetings about specific subjects are organised. Further details related to the Follow-up Committee can be found in Section 6 and to communication activities in Section 7. During this first reporting period DATA-KBR-BE has encountered some challenges - primarily as a result of the global Covid-19 pandemic - however, nothing which has significantly affected the progress of the project. The challenges that DATA-KBR-BE has encountered and the solutions that the project team found are detailed further in Section 8.

3. INTERMEDIARY RESULTS

WP1: Co-designing Interdisciplinary Research Scenarios
D1.1 Report describing Co-Designed Interdisciplinary Research Scenarios (M6, December 2020)
There is one deliverable foreseen in WP1: *D1.1 Report describing Co-Designed Interdisciplinary Research Scenarios*. By December 2020, detailed descriptions of the two interdisciplinary research scenarios outlined in the DATA-KBR-BE project proposal had been prepared: 1) Collective Action Belgium, led by GhentCDH and 2) Feuilleton in Belgium, led by ACDC. Furthermore an additional research scenario was added, led by Camille on the History of Belgian Journalism. As detailed in section 2 above, during the initial start-up phase of the project, it quickly became clear that the three steps originally anticipated in the project proposal: a) research scenario design (WP1), b) dataset preparation (WP2) and c) scientific exploitation (WP4) would take place iteratively, rather than sequentially. The DATA-KBR-BE project team therefore agreed that the three steps should continuously run in parallel for each of the research scenarios throughout the project. In Spring 2021, detailed presentations of the two core research scenarios were prepared: 1) Collective Action Belgium) and 2) Feuilleton in Belgium for the Data Extraction Workshops (see WP2). The work that is undertaken as part of the research scenarios will continue to be documented throughout the project.

WP2: Preparation of Datasets - led by KBR (M7-M24, January 2021 - June 2022)
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 (M9, M12, M16)
The preparation for this deliverable was initiated during the *DATA-KBR-BE Brainstorming Workshop* which was held online on 27th November 2020. The aim of the workshop was to bring together the DATA-KBR-BE team to brainstorm ideas to help us to develop a sustainable workflow to facilitate research use of KBR's collections. This workshop led to two further workshops to further design the initial dataset for extraction.

The first *DATA-KBR-BE Dataset 1 Workshop* (see: Workshop Agenda) took place on 30th March 2021. The aim of this workshop was to: a) design a 'test' dataset that can be used to support all three research scenarios and can act as a test dataset for experimentation; b) agree the criteria for developing DATA-KBR-BE datasets, c) agree a 'first draft' of the DATA-KBR-BE Dataset 1 and d) agree concrete action and next steps. This first workshop was followed up with a second workshop *DATA-KBR-BE Dataset 1 - Follow-up Workshop* which took place on 22nd April 2021 (see: Workshop Agenda), where further details were provided on each of the research scenarios, providing examples where possible. These two workshops led to the finalisation of the DATA-KBR-BE Data Extraction 1 in June 2021, which is scheduled to be validated by the DATA-KBR-BE project team during their July 2021 meeting.

D2.2 Sustainable data extraction workflow design (M24).
Focussing more on the technical and practical details of the data extraction, a *DATA-KBR-BE Data Extraction Workflow meeting* took place on 7th May 2021. During this meeting Thuy-An Pham, KBR's ICT Department and Sally Chambers, DATA-KBR-BE Project Coordinator met to take the first steps in understanding how DATA-KBR-BE Data Extraction 1 would practically be *extracted* from the KBR's ICT Infrastructure and *shared* with the DATA-KBR-BE project researchers. The transfer of this initial dataset can then take place in Summer 2021. Documenting this process will be the first important step towards *D2.2 Sustainable data extraction workflow design (M24)*.

WP3: Data access via data.kbr.be - led by KBR (M4-M24, October 2020 - June 2022)
There are three deliverables foreseen in WP3: *D3.1 Design of the KBR Open Data Platform: data.kbr.be (M9, March 2021)*; *D3.2 Implementation of data.kbr.be including dataset publication (M24; alpha version M18)* and *D3.3 Digital Asset Registry: inventory of KBR's Digital Collections (M24)*. In this reporting period, the focus of our activities has primarily been on D3.1.

D3.1 Design of the KBR Open Data Platform: data.kbr.be (M9, March 2021)
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. Work towards this was primarily undertaken in the context of the **DATA-KBR-BE Brainstorm Workshop** which took place in November 2020 (see further details on above), a review of the functionality of existing library data platforms was undertaken including: National Library of Scotland's Data Foundry; Austrian National Library's ONB Labs Datasets; British Library's Digital Collections and Data; National Library of the Netherland's Data Services and APIs; Danish Cultural Heritage Cluster and British National Bibliography (BNB) Linked Data Platform. During this reporting period, more time needed to be dedicated to the dataset extraction. This means that the design of the KBR Open Data Platform is a little behind the initial schedule anticipated in the project proposal. However, it is not anticipated that this is a significant risk for the ultimate delivery of the platform.

WP4: Scientific exploitation and valorisation - led by UAntwerpen (M10-M24, April 2021 - June 2022)
There are two deliverables foreseen in *WP4: D4.1 Publication of Open Datasets in a Trusted Digital Repository (M24)* and *D4.2 Report of the High Profile Hackathon (M24)*.

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. As both of the WP4 deliverables are due at the end of the project, only initial steps have been taken towards their realisation at this stage. Regarding D4.1 in October 2020, an initial meeting took place between KBR staff (including members of the DATA-KBR-BE project team) and colleagues from the State Archives of Belgium, about the KBR's requirements for publishing research data in the SODHA, Social Sciences and Digital Humanities Archives, data repository. SODHA is anticipated to be the Trust Digital Repository where the open datasets produced within DATA-KBR-BE will be published. This initial meeting led to the formation of an inter-departmental task force related to the development of a Digital Data Strategy for KBR. With regards to D4.2 initial brainstorming regarding the DATA-KBR-BE Hackathon (e.g. potential timing, Save the Date announcement, Audience, Goals etc) which is scheduled to take place towards the end of the project. This work will be continued following the extraction of DATA-KBR-BE dataset 1 in Summer 2021.

Finally, in October 2020, an initial meeting took place between KBR staff (including members of the DATA-KBR-BE project team) and colleagues from the State Archives of Belgium, about the KBR's requirements for publishing research data in the SODHA, Social Sciences and Digital Humanities Archives, data repository. This meeting instigated a wider discussion where questions such as: how is research data distinct from all the other data types held by the KBR? should it be valorised differently? how does it connect to the current and future KBR Action Plan etc. Consequently there emerged a need to map the data we have at KBR and develop a digital strategy for managing this data.

WP5: Project Management and Communication - led by KBR (M1-M24, July 2020 - June 2022)
There are two deliverables foreseen in WP5*: D5.1 Annual Report (M12, 15th March 2021)* and *D5.2 Final Report (M12)*. ***D5.1 Annual Report*** (the report that you are reading now) was successfully delivered in BELSPO in Summer 2021. This was a little later than originally anticipated in the project, primarily due to the fact that the DATA-KBR-BE project coordinator could only be appointed in October 2020 in a part-time position (50% FTE). However, a draft report had been delivered earlier to the BELSPO project officer in anticipation of the full report.

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 first reporting period, the project has already undertaken some core activities (see Section 2) and achieved some significant results (see Section 3). On the basis of these activities and results, the following preliminary conclusions and related recommendations can be drawn:

Stimulating and facilitating academic research using KBR's collections

KBR, the Royal Library of Belgium, is the national research library. KBR collects all Belgian publications and preserves, manages and studies more than 8 million documents of Belgium's rich cultural and historical heritage. KBR provides access to all information in its collections, facilitates research and offers a broad cultural experience. KBR's mission is to support three core target audiences: a) *Culture lovers* via the KBR Museum), b) *General Public* and the 3) *Research Community* with a focus on (digital) Humanities researchers in particular. Through DATA-KBR-BE, KBR has been able to take its first concrete steps towards extending KBR's services for researchers.

Increased understanding of the different needs and research cultures in interdisciplinary teams

DATA-KBR-BE is positioned at the intersection between digital cultural heritage, digital humanities and data science. This entails building an interdisciplinary team of researchers with *different research cultures*, needs and ways of working. Below are a list of preliminary conclusions regarding each of the disciplinary communities within DATA-KBR-BE:

(Digital) Humanities researchers need data-level access to KBR's digitised and born-digital collections for their research. However, creating datasets (or corpora), based on KBR's collections, is a more challenging task than originally anticipated. Corpus development is an ongoing process, which is grounded in the historical context of the period being studied and iteratively develops through exploration. It is often quite 'organic', rather than methodical and is seldom documented. *Questions arise such as:* Which newspaper titles to choose? In which languages? For which years? What is the historical context of these newspapers? Are these titles and years digitally available? What is the OCR quality of those titles?

Data Scientists are fast-paced. They want easy access to the data quickly, so they can move forward with developing and training models. As a result there is some level of frustration at the speed at which data can be delivered. This also sometimes leads to the data scientists working in isolation. They are keen to understand the details of what the humanities researchers need to answer their research questions, ideally illustrated by specific examples. However, due to the iterative research practices of humanities researchers, expressing this information in clear requirements is not always easy. Publishing the findings of their research in high-impact journals in computer science is a key motivating factor. It is important that the research scenarios are intellectually stimulating for the computer and data scientists and that they are not 'merely' seen as supporting the humanities researchers with their technical knowledge and expertise.

***Cultural Heritage professionals**.* There are a range of experts working at the KBR: digitisation experts, heritage collections experts, ICT experts and digital curators. These experts often work in different departments within the library. Below is an overview of the key observations about the different cultural heritage experts: a) *heritage collections expert*s have deep, but sometimes tacit knowledge about the collections, e.g. historical context of the newspapers. Could this tacit knowledge be made explicit?, b) *digitisation experts* often have a lot of knowledge about the digitisation process, including detailed technical information about the digitisation of particular materials in the library. However, this information is often in internal systems and not necessarily available for researchers wishing to use the digital collections, c) *library ICT experts* are heavily in demand. At the same time they need large-scale, production-ready solutions that can be rolled out as services for the library collection as a whole (several millions items). The 'ad hoc' requests of researchers are often challenging to respond to given the capacity of the team and d) *digital curators:* there is an emerging need for a new type of role at KBR here initially named as *digital curators*. These people have a good level of digital literacy and understanding of the digital aspects of the collections, e.g. file formats, the digitisation process, metadata and data literacy etc.. They are also highly skilled communicators in order to act as mediators or brokers between the different experts involved in an interdisciplinary project such as DATA-KBR-BE. Their role is to facilitate access to the KBR's digitised and born-digital collections for digital humanities research.

Corpus building is an ongoing, contextual and iterative process
The interdisciplinary research scenarios provide the validation framework for the whole project. In other words, they ensure that the sustainable data extraction workflows and the data.kbr.be platform support the real needs of digital humanities researchers. Originally in the project proposal it was anticipated that these research scenarios would be undertaken in three steps: a) research scenario design, b) dataset preparation and c) scientific exploitation, which would occur sequentially. However, after the initial start-up phase of the project, this process was more iterative rather than linear or sequential. The DATA-KBR-BE project team therefore recommends that the three steps should continuously run in parallel for each of the research scenarios throughout the project.

Contextual transparency of KBR's collections
KBR's cultural heritage experts possess a wealth of knowledge about KBR's collections. For example, about the historical context of the newspapers in KBR's collections or technical details about the digitisation of these collections (e.g. information about the OCR quality of the digitised collections). This both tacit knowledge, in the case of the historical context of the newspapers or the internal details about the digitisation process of the collections of the kind of information that would also be useful to digital humanities researchers to help them make informed choices when building their research corpora. In the second phase of DATA-KBR-BE, we can explore how this until now buried knowledge can be made transparent for researchers and other users of KBR's collections. Related to **OCR quality** in particular. While it would already be useful to make the technical information about quality of the OCR transparent to the users, the question was also raised as to whether the OCR quality of some parts of KBR's digitised collections was sufficient for the application of some digital methods. This will be investigated further in the second part of the project.

Increased understanding of what 'collections as data' means for KBR
The DATA-KBR-BE project proposal outlines that the datasets resulting from the scientific exploitation of the open science datasets will be fully documented and deposited in an appropriate Trust Digital Repository for (scholarly) reuse. We already anticipated that this work would be undertaken in liaison with the BELSPO-funded BISHOPS (Belgian Infrastructure for Social Sciences and Humanities Open Science), led by the State Archives and the Social Sciences and Digital Humanities Archive (SODHA). During the first reporting phase of DATA-KBR-BE, the importance of distinguishing between 'cultural heritage datasets', 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 is the data that resulting from analysing the 'cultural heritage datasets', with digital humanities research methods, became clear. It is the 'cultural heritage datasets' that will be published on the data.kbr.be platform, whereas the 'humanities research data' are *derived datasets* resulting from their analysis in the interdisciplinary research scenarios (in the context of *WP4 - Scientific exploitation and valorisation*). These 'humanities research datasets' will be published in a Trusted Digital Repository such as SODHA.

Transferability of the results to other types of material
In the first instance, DATA-KBR-BE focuses on the digitised newspapers in BelgicaPress. However, as recommended by the Follow-up Committee, digitised newspapers are one specific example of digital collections. In addition to other digitised materials there are other types of collections, such as born-digital materials (e.g. web-archives and social media archives) as well as licenced eresources. Already within the DATA-KBR-BE project team other examples of collections which could be published as datasets have been suggested, e.g. publication of datasets from previous BELSPO Brain projects (e.g. IMPRESS, Photo-Lit etc.) and publication of at least one born-digital dataset (e.g. from the PROMISE or BE-Social projects).

Broadening of the level and depth of legal expertise
There is some legal expertise in the library. For example, to understand and ensure compliance with the relevant legal frameworks (e.g. European and Belgian copyright law, Privacy legislation (e.g. GDPR), the Open Data Directive, exemptions for Text and Data Mining etc.) However this expertise often is limited to one or two individuals. The Follow-up Committee recommended that DATA-KBR-BE should ensure that they have (access to) a legal expert for the project. Additionally, it may be useful to explore how the level of legal expertise across the project team, including library staff, digital humanities researchers and data scientists could be raised in the second phase of the project.

Improved level of understanding of the KBR's data infrastructure
Starting with the presentation of the existing and emerging KBR Data Infrastructure in June 2020 by Xavier Delor and Thuy-An Pham of KBR's ICT team in June 2020, the level of understanding about the KBR's data infrastructure has gradually improved over the project. This has included increased knowledge of the different components of the existing data infrastructure; the project to optimise the KBR's data infrastructure (internally known as the 'Stock' project) as well as plans for future development of the infrastructures (e.g. implementation of IIIF or the automation of the KBR's digitisation workflow. This increased knowledge, as well as the close collaboration with the KBR's ICT has primarily thanks to the appointment of Thuy-An Pham as the ICT Departmental liaison for the DATA-KBR-BE project.

Data sharing: implementing and documenting the extraction of the initial DATA-KBR-BE dataset.
As highlighted elsewhere in this report, the selection of the thematic datasets to support the research scenarios was more challenging than originally anticipated. However, a second challenge was related to *how best to share the DATA-KBR-BE Dataset* once it had been selected. Currently, if researchers request a data extraction from KBR's collections, this is handled on a case-by-case basis. Depending on the size of the dataset, the files are either transferred by KBR's SendFile service, which is also used for KBR's Digitisation on Demand service, or the files are copied onto hard disk(s) for the user (which during the Covid-19 pandemic is less than ideal). Exactly how to transfer the files to researchers in an easy and efficient manner is one of the first issues that will be explored in the next phase of DATA-KBR-BE.

Additionally, other related questions have also been raised. For example: Should DATA-KBR-BE provide access to **in-copyright material** or only material that is already in the **public domain**? What agreements need to be signed for the use of these files before they are transferred to the users? Can these agreements be signed online? Is DATA-KBR-BE only about data extraction? For example, if a dataset was enriched with post-processed OCR correction, could the enriched data be ingested back into the KBR's systems?

Towards and beyond the data.kbr.be platform
While providing access to 'collections as data' via a data platform such as the envisaged data.kbr.be platform, this is only one way of providing access to the data. Depending primarily on the research question and the level of digital literacy of the researcher(s) it may mean that different 'levels' of data access are needed. Additionally, researchers may need to work with several of these levels of access depending on their research question.

Below is a first attempt to describe the potential 'levels' of data access that KBR could provide: **1)** ***Online Corpus building:*** The European project NewsEye's demonstrator platform includes a 'create your own datasets' functionality, which could conceptually be interesting for DATA-KBR-BE. While this kind of functionality may not be scalable at the European level, it may be interesting to explore its implementation within BelgicaPress?, ***2) 'Datasets on Demand'*****:** The idea to explore a 'datasets on demand' service was first discussed at the DATA-KBR-BE brainstorm workshop in November 2021. It is inspired by the Royal Library of Denmark, where they have created the relevant legal agreements for providing 'Datasets on Demand' (see: this video). This potential service could extend the KBR's existing SendFile service (e.g. by increasing the size of the files being shared) and also the KBR's Digitisation on Demand service, ***3) Published Datasets*****:** The National Library of Scotland's Data Foundry and National Library of Luxembourg's Open Data Platform, for example, for Historical Newspapers are great examples of where datasets have been published by the library's themselves for download. Services such were the original inspiration for the data.kbr.be platform. It is anticipated that something similar will be set-up for DATA-KBR-BE by the end of the project. However, it is worth noting that these are 'library-driven' datasets, rather than 'research driven' datasets or 'collections' rather than 'corpora'. It has been noted by researchers that they may be more interested in building their own corpora, rather than working with library-selected collections, **4) Jupyter Notebooks:** The GLAM WorkBench developed by Australian historian Tim Sherratt is an interesting example of providing relatively low-barrier computational access to digital cultural heritage collections via Jupyter Notebooks. The GLAM Workbench has also been recently extended with Notebooks for WebArchives. Additionally, Gustavo Candela's Notebooks from the University of Alicante could provide valuable inspiration for DATA-KBR-BE. It may be interesting to experiment with Jupyter notebooks within the DATA-KBR-BE context, for example, for the Hackathon in the second phase of the project. **5) API access:** For more computation approaches to humanities research access to the KBR's datasets via an Application Programming Interface (API) may be an interesting option to explore. Existing examples of API access to national library collections include the National Library of the Netherlands' Data Services and APIs and the National Library of France's APIs. While this option could be explored in DATA-KBR-BE, it would not be feasible to develop API services for the KBR within the context of the DATA-KBR-BE and **6) 'Cloud access':** Finally, it may be interesting to explore the feasibility of developing a secure online environment where 'access' and 'analysis' are brought together. This is something that is already being experimented with in the CLARIAH-Flanders project. This 'cloud access' is also inspired by the Danish Cultural Heritage Cluster.

5. FUTURE PROSPECTS AND PLANNING

WP1: Co-designing Interdisciplinary Research Scenarios
There is one deliverable foreseen in WP1: *D1.1 Report describing Co-Designed Interdisciplinary Research Scenarios*. Work will continue on the three interdisciplinary research scenarios 1) Collective Action Belgium, led by GhentCDH, 2) Feuilleton in Belgium, led by ACDC and 3) History of Belgian Journalism, led by Camille. This work will be an iterative process combining a) research scenario design (WP1), b) dataset preparation (WP2) and c) scientific exploitation (WP4). The focus of the work will initially be the analysis of the data transferred in DATA-KBR-BE Data Extraction 1.

WP2: Preparation of Datasets - led by KBR (M7-M24, January 2021 - June 2022)
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)*. With the sign-off of the DATA-KBR-BE Data Extraction 1 by the DATA-KBR-BE project team during their July 2021 meeting, the next step is to transfer this initial dataset in Summer 2021. This data transfer process will be documented and evaluated. In the original project planning, it was anticipated that the extraction of three thematic datasets would: a) be able to happen sooner (M6: March 2021, M12: June 2021 and M16: September 2021). It is anticipated that more time will be needed for the analysis and exploitation of the datasets prior to the next *DATA-KBR-BE Data Extraction 2.*

Following the successful transfer of DATA-KBR-BE Data Extraction 1 by September 2021, it is anticipated that October - December 2021 will be used to undertake a first phase of analysis of the data (cf. WP4) for each of the research scenarios. An update will be provided to the DATA-KBR-BE Follow-up Committee on this first analysis phase during their meeting in December 2021. As noted earlier, during Summer 2021, a migration of KBR's data to the new 'stock' data infrastructure will take place. This will mean that *DATA-KBR-BE Data Extraction 2* will be a 'post-Stock' data extraction and therefore will have a different workflow. However, the evaluation undertaken for the transfer of DATA-KBR-BE Data Extraction 1 will still provide valuable requirements for *DATA-KBR-BE Data Extraction 2*.

WP3: Data access via data.kbr.be - led by KBR (M4-M24, October 2020 - June 2022)
There are three deliverables foreseen in WP3: *D3.1 Design of the KBR Open Data Platform: data.kbr.be (M9, March 2021)*; *D3.2 Implementation of data.kbr.be including dataset publication (M24; alpha version M18)* and *D3.3 Digital Asset Registry: inventory of KBR's Digital Collections (M24)*. Following the first DATA-KBR-BE Data Extraction in Summer 2021, further work will be undertaken to prepare for the design of the data.kbr.be platform. A next concrete step will be to carry out semi-structured interviews with other National Libraries which host data platforms, such as National Libraries of France (BnF), Luxembourg (BnL), Netherlands (KB) and Scotland (NLS). This work will inform the development of the job profile for the appointment of the KBR data scientist (see also: WP5.)

WP4: Scientific exploitation and valorisation - led by UAntwerpen (M10-M24, April 2021 - June 2022)
There are two deliverables foreseen in *WP4: D4.1 Publication of Open Datasets in a Trusted Digital Repository (M24)* and *D4.2 Report of the High Profile Hackathon (M24)*. Following the successful transfer of DATA-KBR-BE Data Extraction 1 by September 2021, it is anticipated that October - December 2021 will be used to undertake a first phase of analysis of the data (cf. WP4) for each of the research scenarios. An update will be provided to the DATA-KBR-BE Follow-up Committee on this first analysis phase during their meeting in December 2021. Furthermore, it is anticipated that an instance (sub-dataverse) of the SODHA repository for KBR will be set up in Autumn 2021. Once this sub-dataverse is available, an initial test dataset will be deposited.

WP5: Project Management and Communication - led by KBR (M1-M24, July 2020 - June 2022)
There is one remaining deliverable foreseen in WP1: *D5.2 Final Report (M24).* The day-to-day coordination of the DATA-KBR-BE project will continue including regular project team meetings, organisation of Follow-up Committee meetings and communication activities. As mentioned in Section 8, the KBR intends to apply to BELSPO to extend the duration of the DATA-KBR-BE project. It is intended that this process will be initiated in September/October 2021. Alongside this, the KBR Data Scientist will be recruited. It is anticipated that if the duration of the DATA-KBR-BE project is approved, an additional *Annual Report* as well as D5.2 Final Report (M24) will be needed.

6. FOLLOW-UP COMMITTEE

DATA-KBR-BE Follow-up Committee Meeting, Wednesday 16th December 2020
The first DATA-KBR-BE Scientific Advisory Board or 'Follow-up Committee' meeting took place online on *Wednesday 16th December 2020 (14:00 - 16:00).* The meeting was attended by all members of the Follow-up Committee as well as the DATA-KBR-BE project team. The agenda for the meeting can be found here. Following a round of introductions, the project team gave a presentation to introduce the DATA-KBR-BE project. This focussed on four key topics: *a) Collections as Data @ KBR, b) Facilitating Research with KBR Collections, c) Sustainable Data Extraction workflow and d) Designing the data.kbr.be platform.* While there was a lot of material to cover in a short online meeting, the Follow-up Committee fully engaged with the content, which led to a valuable discussion and some initial recommendations and next steps.

Recommendations & Next Steps from first meeting
Already, the members of the Follow-up committee were able to make a number of valuable suggestions. For example, the importance of including a *copyright expert* in the DATA-KBR-BE project team was stressed. Furthermore a number of *potential collaborations* were proposed, especially with related projects such as the BELSPO-funded ADOCHS, Auditing Digitisation Outputs in the Cultural Heritage Sector project; NumaPresse in France, the OCR-D project in Germany; the Swiss-Luxembourg collaboration IMPRESSO, Media Monitoring of the Past: Mining 200 years of historical newspapers and the European TimeMachine project especially related to spatio-temporal evolutions. These collaborations could focus on the sharing of research methods and

results, such as *sharing of ground truth data*, for example, such as the ground-truth from the National Library of the Netherlands or the IMPACT Centre of Competence on Digitisation. It was advised that DATA-KBR-BE should also consider API-access to the data. One of the follow-up committee members asked about the *transferability of the results to other types of material.*

Regarding next steps, it was proposed that *interviews with the National Library of Luxembourg (BnL) and the National Library of the Netherlands (KB) regarding their data platforms* would provide valuable insights to the DATA-KBR-BE project team. Although it is intended that the DATA-KBR-BE Follow-up Committee will meet once a year, it was agreed that virtual meetings every six months would be valuable.

The next meeting of the DATA-KBR-BE Scientific Advisory Board ('Follow-up Committee') is scheduled for Thursday 10th June 2021 (10:00 - 12:00 CET via Zoom).

7. VALORISATION ACTIVITIES

7.1 PUBLICATIONS

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

Following the DH Benelux 2021: 'The Humanities in a Digital World', 2-4 June 2021, Leiden (Online) a call for paper for a special issue of the DH Benelux Journal is expected. The DATA-KBR-BE project team intends to submit an article based on our presentation at DH Benelux: Collections as Data: interdisciplinary experiments with KBR's digitised historical newspapers: a Belgian case study

We have been liaising with the NewsEye team regarding the possibility of preparing a Special Issue of a journal on Research using Digitised Historical Newspapers, inspired by the NewsEye International conference What's Past is Prologue: the NewsEye International Conference, 16-17 March 2021. The idea would be that the DATA-KBR-BE team would form part of the Editorial Committee of this special issue and submit at least one article about our research. For example, an article based on the presentation given by Dilawar Ali and Steven Verstockt on *Challenges in extraction and classification of news articles from historical newspapers* at the conference.

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

7.2 PARTICIPATION/ORGANISATION OF SEMINARS (NATIONAL/INTERNATIONAL)

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

Chambers, S. (2021) DATA-KBR-BE: experimenting with digital humanities datasets at KBR, Royal Library of Belgium. Presentation at: DH Hangouts, 12 March 2021.

Ali, D., & Verstockt, S. (2021). *Challenges in extraction and classification of news articles from historical newspapers*. Presentation at: What's Past is Prologue: the NewsEye International Conference, 16-17 March 2021.

Lightning talk of DATA-KBR-BE presentation at the 'DH Benelux Edition' of the *Belgian DH Early Career Researchers meeting*, 27th May 2021.

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

Presented DATA-KBR-BE as part of the *"DH related research projects @ KBR"* coordinated by Julie M. Birkholz at the Research Seminar organised by KU Leuven Artes Libraries Presentation, 8 June 2021.

Upcoming: The DATA-KBR-BE project team has been invited to give a presentation as part of the ADOCHS Online Study Day: Image and Data Processing in the Cultural Heritage Sector, which is scheduled to take place online on 14th September 2021.

7.3 SUPPORT TO DECISION MAKING (IF APPLICABLE)

The DATA-KBR-BE project has already started to play a crucial role within the KBR regarding exploring new ways of providing access to its digitised and born-digital collections for the research community, and in particular digital humanities researchers. For example: a) *strengthening the inter-departmental collaboration with KBR's ICT department regarding the optimisation of the KBR's data infrastructure* and b) *contributing to the initiation of a Digital Data strategy at KBR, where 'Collections as Data' has been identified as one of the core data types*. DATA-KBR-BE will therefore directly contribute to the development of *KBR's Action Plan for 2022-2024*. Already, the DATA-KBR-BE team is investigating opportunities to continue the work of DATA-KBR-BE, both in Belgium, e.g. via other Belspo funding programmes such as ESFRI-FED and BRAIN or European Funding, e.g. Horizon Europe. To do this, reaching out to other research groups and projects (e.g. ADOCHS, OCR-D, NumaPress, Impresso, TimeMachine, NewsEye and NieuweTijdingen) and tools e.g. National Library of the Netherlands CHRONIC (Classified Historical Newspaper Images) dataset and related tools and the Newspaper Navigator project at the Library of Congress are just some examples of *potential collaborations.*

7.4 OTHER

n/a

8. ENCOUNTERED PROBLEMS AND SOLUTIONS

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

During this first project period, DATA-KBR-BE has encountered some challenges, but nothing which has significantly affected the progress of the project. The challenges that DATA-KBR-BE has encountered and the solutions that the project team found are detailed below:

Administrative start-up of the project: DATA-KBR-BE was formally approved in December 2019. The administrative start-up of the project began in early 2020, with the first meeting of the DATA-KBR-BE project consortium taking place in February 2020. However, the escalation of the global Covid-19 pandemic from

March 2020 impacted significantly on the administrative start-up of the project, especially with regard to staff recruitment (see below). BELSPO recognised the impact of Covid-19 and issued a statement for project coordinators. Following the recruitment of the DATA-KBR-BE project coordinator (see below), the project was able to pick-up speed from October 2020. The DATA-KBR-BE Initial Report was submitted to BELSPO in December 2020.

Staff recruitment: During the period March 2020 - September 2020, due to the Covid-19 pandemic, staff recruitment was significantly hindered. However, this enabled the DATA-KBR-BE project team to reflect on some alternative solutions for recruiting staff. Following consultation with the DATA-KBR-BE co-supervisors and with BELSPO, a scenario where Sally Chambers (GhentCDH, UGent), who had instigated and led the preparation of the DATA-KBR-BE project proposal, could be appointed on a 50% FTE contract (Scientific Collaborator, Humanities) by KBR to coordinate the project was proposed. This solution was proposed to the Inspector of Finances and subsequently approved. Sally Chambers took up her position as 50% project coordinator from 1 October 2020.

For the recruitment of the *Scientific Collaborator, Data Science for KBR (12 PM)*, a number of possible scenarios were explored, e.g. external consultancy, sub-contracting to UGent. However, as the DATA-KBR-BE project coordinator was only appointed in October 2020 and for 50% instead of 100%, the DATA-KBR-BE project team decided to postpone the recruitment decision until summer 2021. In this way, the DATA-KBR-BE project team would have made progress with the initial phase of the project and in this way would have a clearer idea as to what skills for this position are. Ideally, this position would be a direct recruitment within KBR's ICT department to strengthen the internal capacity of the team. However, such people are challenging to recruit due to the short-term nature of the contracts. The postponement of this recruitment will not adversely affect the project as the KBR's ICT department appointed an internal liaison for the project, Thuy-An Pham. This close collaboration with KBR's ICT department from the outset has been crucial for the development of the sustainable data extraction workflow.

Remote access to KBR's digitised collections: Due to lockdown restrictions as a result of the Covid-19 pandemic, remote access to the KBR's digitised collections, as well as organisational changes needed at KBR to enable remote working, meant that creative solutions needed to be found to prepare the initial test dataset. Luckily, the KBR had already extracted a dataset to a hard disk for the GhentCDH team for previous research on the newspaper collections. This existing dataset consisted of one French language (Le Peuple) and one Dutch Language (Vooruit) newspaper from BelgicaPress for the period 1886-1938). This enabled colleagues from IDLab to undertake their initial experiments despite the lockdown.

9. MODIFICATIONS COMPARED TO THE PREVIOUS REPORT

9.1 PERSONNEL

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

9.2 COMPOSITION OF THE FOLLOW-UP COMMITTEE

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

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

From 1 June 2021, Wout Dillen will leave the University of Antwerp to become Assistant Professor/Senior Lecturer in Library and Information Science at the University of Borås, Sweden. We would like to propose that we invite him to become a member of the DATA-KBR-BE Scientific Advisory Board, so he can continue his work on his research scenario: the Feuilleton in Belgium, as part of his research time in his new position.

10. REMARKS AND SUGGESTIONS

As noted earlier, the DATA-KBR-BE project team has had initial talks with the BELSPO project officer regarding the possibility of extending the duration of the project until the end of 2023. This possible extension, without additional funding, would further strengthen the institutional embedding of the outcomes of DATA-KBR-BE into the day-to-day work of the KBR, and to take full advantage of the implementation of the KBR's 'Stock' infrastructure, the first phase of which is intended to be rolled out in Autumn 2021. This extension would be made possible because the DATA-KBR-BE project coordinator has been appointed in a part-time (50% FTE), rather than a full-time position. However, a draft report had been delivered earlier to the BELSPO project officer in anticipation of the full report. It is intended that this process will be initiated in September/October 2021. The KBR Data Scientist will also be recruited alongside this.