CLARIN and Libraries

9 May 2022 - 10 May 2022

CLARIN

DATA-KBR-BE: Data-level access to digitised collections for digital humanities research

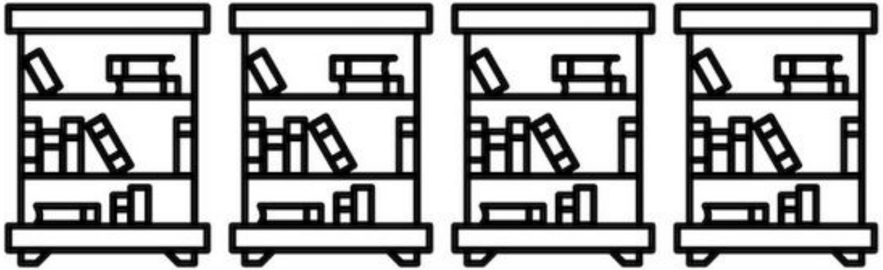
**Sally Chambers, KBR, Royal Library of Belgium and
Ghent Centre for Digital Humanities 🐦 @schambers3**

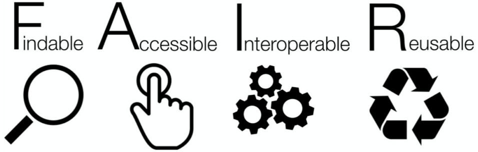
KBR Where time is treasured

GHENT UNIVERSITY

Belgian Science Policy Office
belspo

Collections as Data: “Always Already Computational” - Final Report

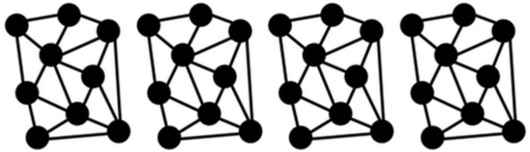




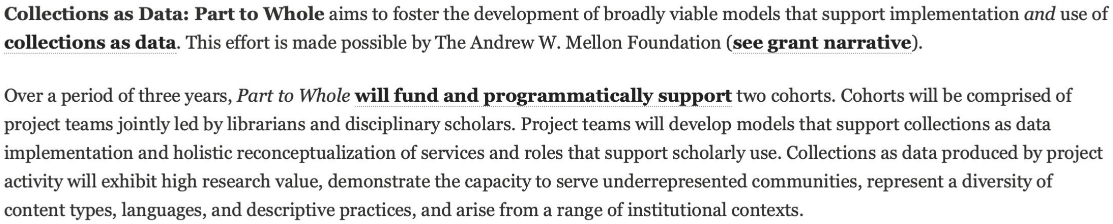
Padilla, Thomas, Allen, Laurie, Frost, Hannah, Potvin, Sarah, Russey Roke, Elizabeth, & Varner, Stewart. (2019, May 22). Final Report --- Always Already Computational: Collections as Data (Version 1). Zenodo. http://doi.org/10.5281/zenodo.3152935

Always Already Computational: Collections as Data final report and project deliverables: https://osf.io/mx6uk/wiki/home/

Collections as Data: Part to Whole



https://collectionsasdata.github.io/part2whole/

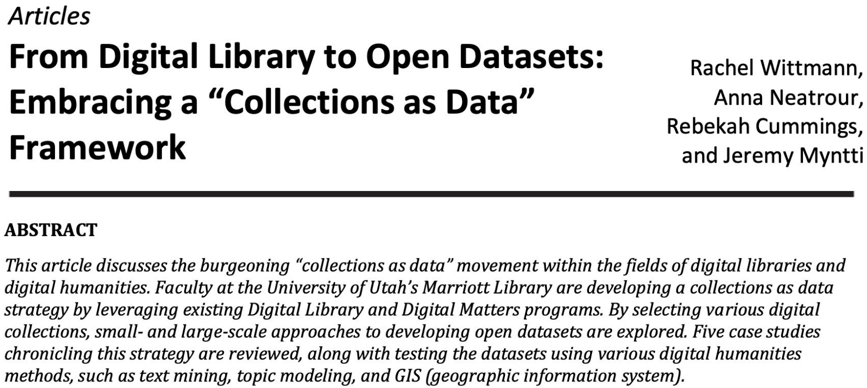


Collections as Data: Part to Whole aims to foster the development of broadly viable models that support implementation *and* use of **collections as data**. This effort is made possible by The Andrew W. Mellon Foundation (**see grant narrative**).

Over a period of three years, *Part to Whole* **will fund and programmatically support** two cohorts. Cohorts will be comprised of project teams jointly led by librarians and disciplinary scholars. Project teams will develop models that support collections as data implementation and holistic reconceptualization of services and roles that support scholarly use. Collections as data produced by project activity will exhibit high research value, demonstrate the capacity to serve underrepresented communities, represent a diversity of content types, languages, and descriptive practices, and arise from a range of institutional contexts.

COLLECTIONS AS DATA

Wittmann, R., Neatrour, A., Cummings, R., & Myntti, J. (2019). From Digital Library to Open Datasets. *Information Technology and Libraries*, *38* (4), 49-61. https://doi.org/10.6017/ital.v38i4.11101



Articles

From Digital Library to Open Datasets: Embracing a "Collections as Data" Framework

Rachel Wittmann,
Anna Neatrour,
Rebekah Cummings,
and Jeremy Myntti

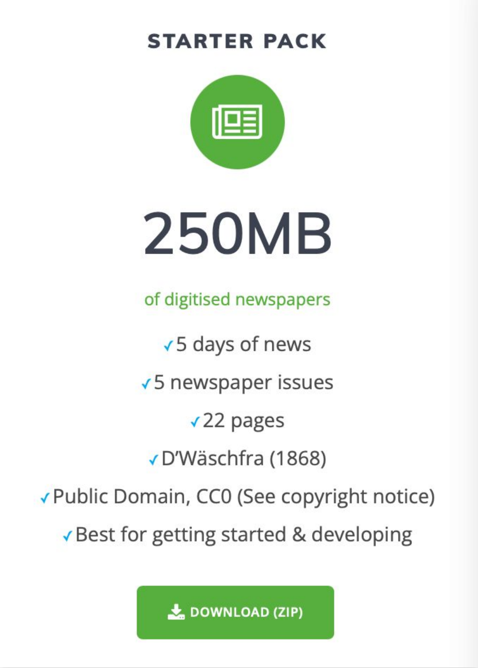
ABSTRACT

This article discusses the burgeoning "collections as data" movement within the fields of digital libraries and digital humanities. Faculty at the University of Utah's Marriott Library are developing a collections as data strategy by leveraging existing Digital Library and Digital Matters programs. By selecting various digital collections, small- and large-scale approaches to developing open datasets are explored. Five case studies chronicling this strategy are reviewed, along with testing the datasets using various digital humanities methods, such as text mining, topic modeling, and GIS (geographic information system).

Newspapers as Data

Bibliothèque nationale du Luxembourg
Open Data

https://data.bnl.lu/data/historical-newspapers/

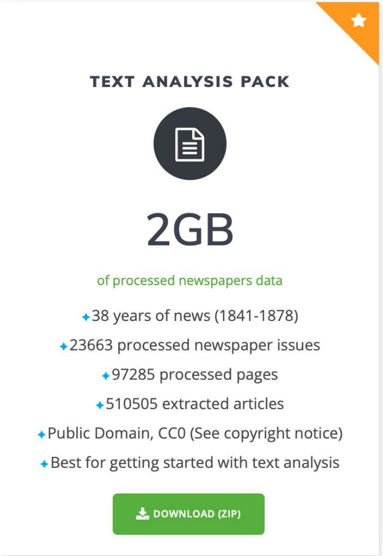


STARTER PACK

250MB

of digitised newspapers

✓5 days of news
✓5 newspaper issues
✓22 pages
✓D'Wäschfra (1868)
✓Public Domain, CC0 (See copyright notice)
✓Best for getting started & developing



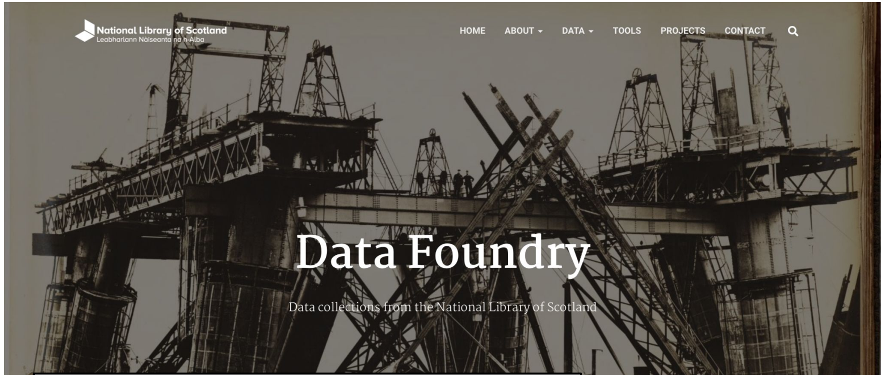
TEXT ANALYSIS PACK

2GB

of processed newspapers data

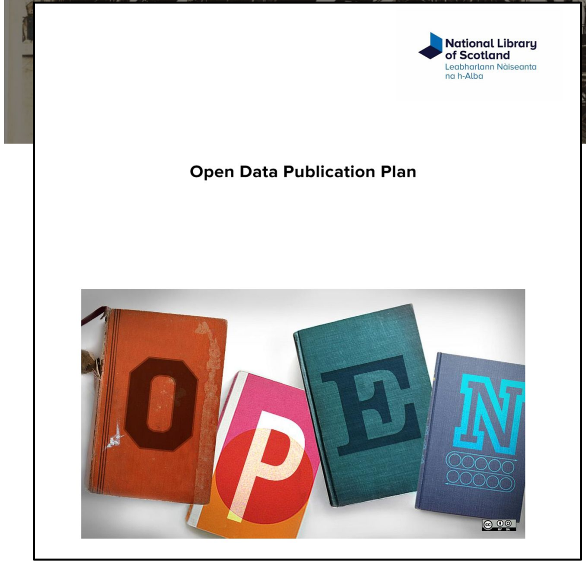
✦38 years of news (1841-1878)
✦23663 processed newspaper issues
✦97285 processed pages
✦510505 extracted articles
✦Public Domain, CC0 (See copyright notice)
✦Best for getting started with text analysis

National Library of Scotland: Data Foundry



Data Foundry

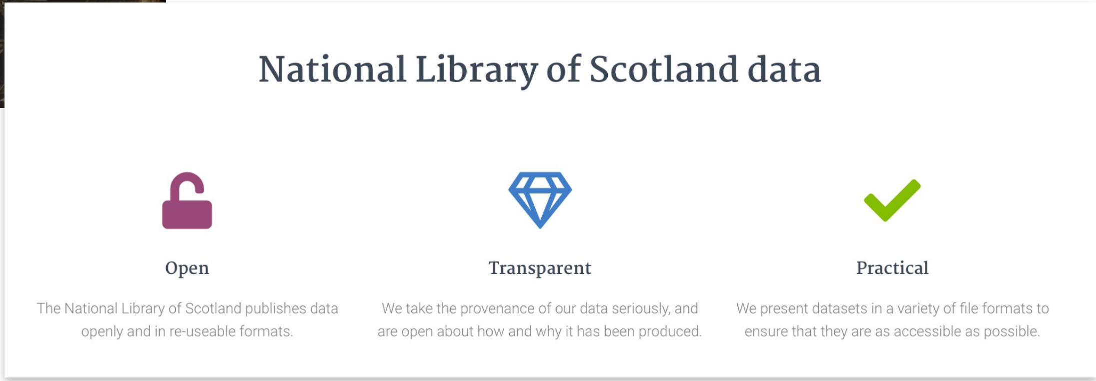
Data collections from the National Library of Scotland



Open Data Publication Plan

Ames, S. (2021) "Transparency, Provenance and Collections as Data: The National Library of Scotland's Data Foundry." LIBER Quarterly 31 (1): 1–13. https://doi.org/10.18352/lq.10371.

https://data.nls.uk



National Library of Scotland data

Open

The National Library of Scotland publishes data openly and in re-useable formats.

Transparent

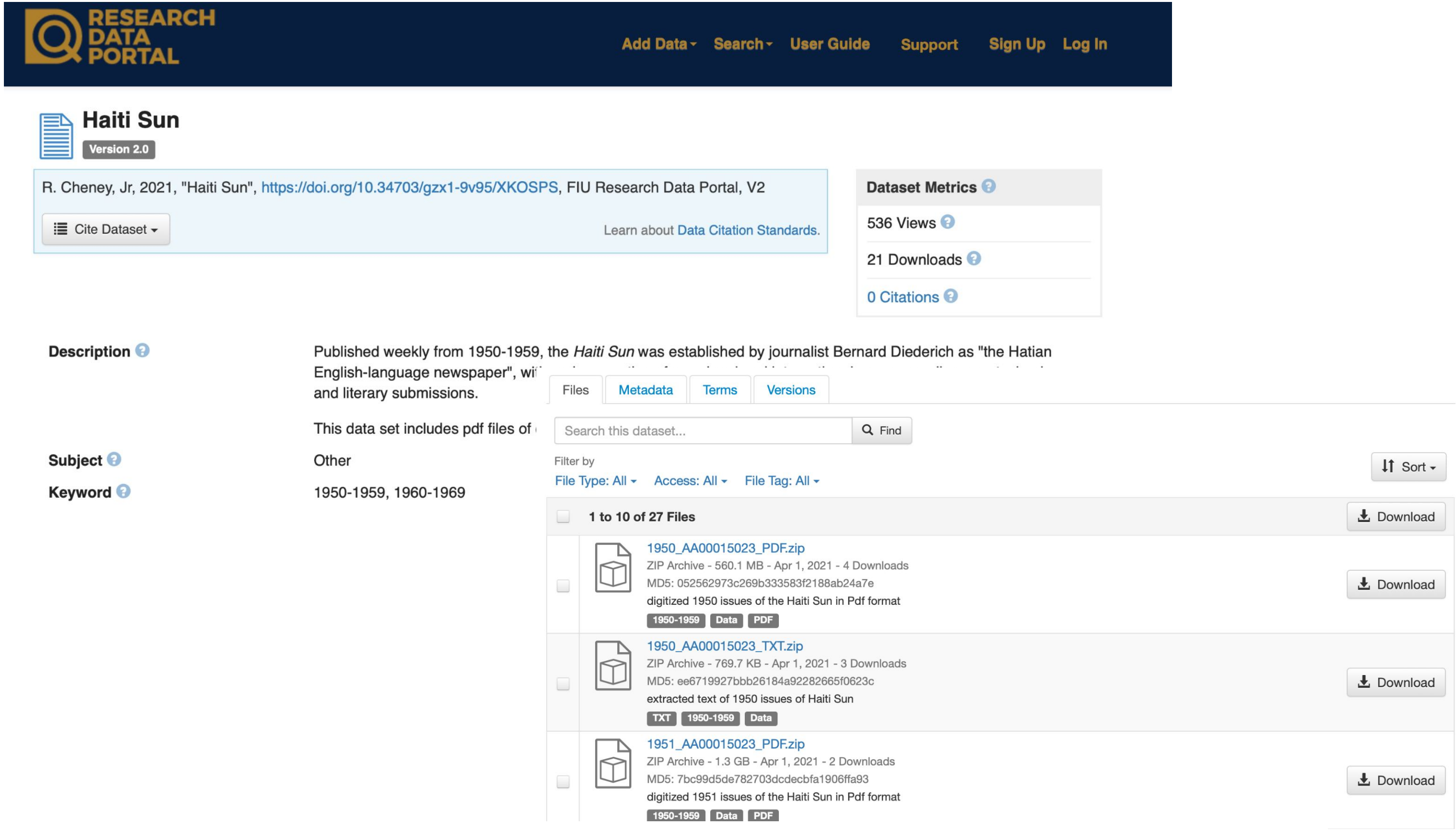
We take the provenance of our data seriously, and are open about how and why it has been produced.

Practical

We present datasets in a variety of file formats to ensure that they are as accessible as possible.

https://data.nls.uk/download/national-library-of-scotland-open-data-publication-plan.pdf

Digital Library of the Caribbean as Data



Haiti Sun

Version 2.0

R. Cheney, Jr, 2021, "Haiti Sun", https://doi.org/10.34703/gzx1-9v95/XKOSPS, FIU Research Data Portal, V2

Learn about Data Citation Standards.

Dataset Metrics

536 Views

21 Downloads

0 Citations

Description Published weekly from 1950-1959, the *Haiti Sun* was established by journalist Bernard Diederich as "the Hatian English-language newspaper", wi

and literary submissions.

This data set includes pdf files of

Subject Other

Keyword 1950-1959, 1960-1969

□ **1 to 10 of 27 Files** Download

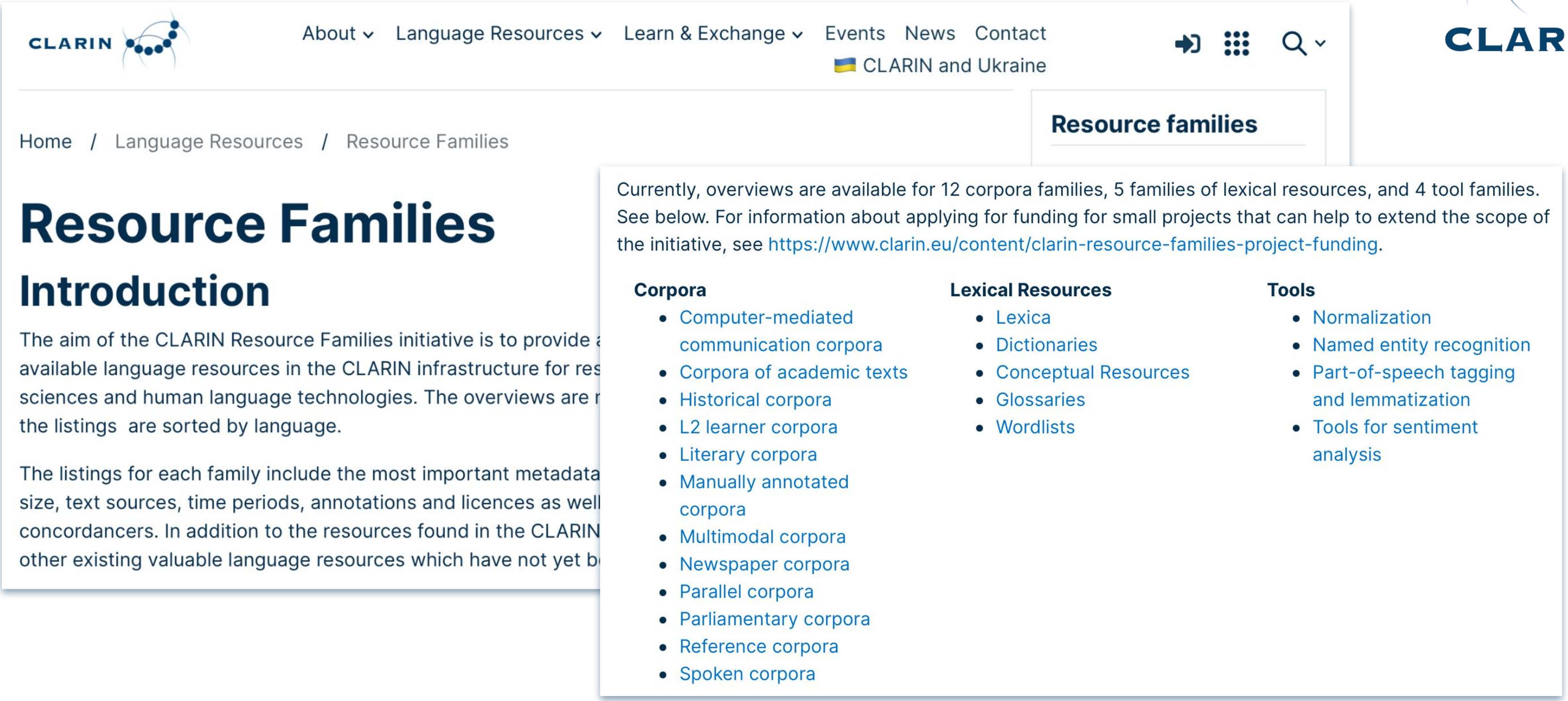
□ 1950_AA00015023_PDF.zip
ZIP Archive - 560.1 MB - Apr 1, 2021 - 4 Downloads
MD5: 052562973c269b333583f2188ab24a7e
digitized 1950 issues of the Haiti Sun in Pdf format
1950-1959 Data PDF
Download

□ 1950_AA00015023_TXT.zip
ZIP Archive - 769.7 KB - Apr 1, 2021 - 3 Downloads
MD5: ee6719927bbb26184a92282665f0623c
extracted text of 1950 issues of Haiti Sun
TXT 1950-1959 Data
Download

□ 1951_AA00015023_PDF.zip
ZIP Archive - 1.3 GB - Apr 1, 2021 - 2 Downloads
MD5: 7bc99d5de782703dcdecbfa1906ffa93
digitized 1951 issues of the Haiti Sun in Pdf format
1950-1959 Data PDF
Download

https://dataverse.fiu.edu/dataverse/dloc-as-data

CLARIN Resource Families



Resource Families

Introduction

The aim of the CLARIN Resource Families initiative is to provide a
available language resources in the CLARIN infrastructure for res
sciences and human language technologies. The overviews are r
the listings are sorted by language.

The listings for each family include the most important metadata
size, text sources, time periods, annotations and licences as well
concordancers. In addition to the resources found in the CLARIN
other existing valuable language resources which have not yet b

Currently, overviews are available for 12 corpora families, 5 families of lexical resources, and 4 tool families. See below. For information about applying for funding for small projects that can help to extend the scope of the initiative, see https://www.clarin.eu/content/clarin-resource-families-project-funding.

Corpora

- Computer-mediated communication corpora
- Corpora of academic texts
- Historical corpora
- L2 learner corpora
- Literary corpora
- Manually annotated corpora
- Multimodal corpora
- Newspaper corpora
- Parallel corpora
- Parliamentary corpora
- Reference corpora
- Spoken corpora

Lexical Resources

- Lexica
- Dictionaries
- Conceptual Resources
- Glossaries
- Wordlists

Tools

- Normalization
- Named entity recognition
- Part-of-speech tagging and lemmatization
- Tools for sentiment analysis

https://www.clarin.eu/resource-families

DATA-KBR-BE

Facilitating data-level access to KBR's digitised and born-digital collections for digital humanities research

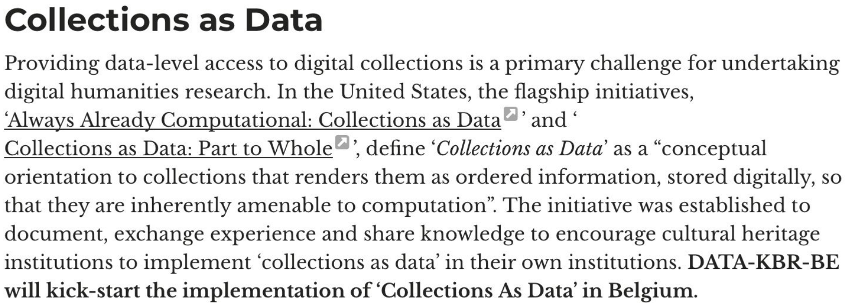
COLLECTIONS AS DATA

https://www.kbr.be/en/projects/data-kbr-be/

Belgian Science Policy Office
belspo

KBR Where time
is treasured

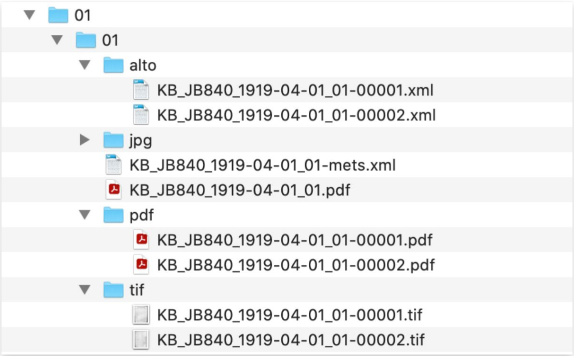
Digitised Historical Newspapers as Data



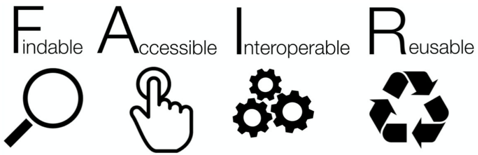
Collections as Data

Providing data-level access to digital collections is a primary challenge for undertaking digital humanities research. In the United States, the flagship initiatives, 'Always Already Computational: Collections as Data' and 'Collections as Data: Part to Whole', define '*Collections as Data*' as a "conceptual orientation to collections that renders them as ordered information, stored digitally, so that they are inherently amenable to computation". The initiative was established to document, exchange experience and share knowledge to encourage cultural heritage institutions to implement 'collections as data' in their own institutions. **DATA-KBR-BE will kick-start the implementation of 'Collections As Data' in Belgium.**

https://collectionsasdata.github.io



```
▼ 01
    ▼ 01
        ▼ alto
                KB_JB840_1919-04-01_01-00001.xml
                KB_JB840_1919-04-01_01-00002.xml
        ▶ jpg
            KB_JB840_1919-04-01_01-mets.xml
            KB_JB840_1919-04-01_01.pdf
        ▼ pdf
                KB_JB840_1919-04-01_01-00001.pdf
                KB_JB840_1919-04-01_01-00002.pdf
        ▼ tif
                KB_JB840_1919-04-01_01-00001.tif
                KB_JB840_1919-04-01_01-00002.tif
```



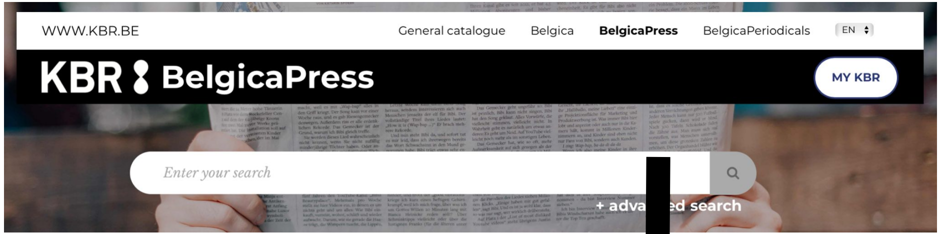
Providing data-level access to the underlying files of digitised and born-digital cultural heritage resources to facilitate data analysis by means of tools and methods developed in the field of digital humanities

https://www.kbr.be/en/projects/data-kbr-be/

GHENT UNIVERSITY

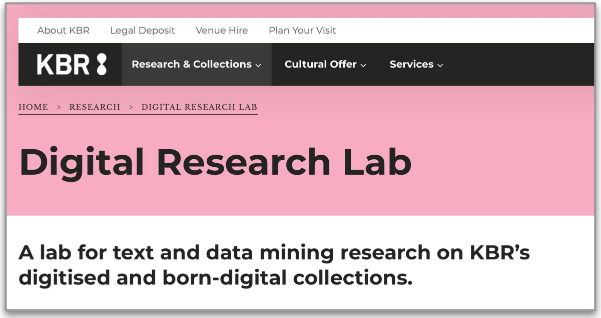
KBR

Digitised Historical Newspapers as Data



WWW.KBR.BE General catalogue Belgica **BelgicaPress** BelgicaPeriodicals EN

KBR BelgicaPress MY KBR



About KBR Legal Deposit Venue Hire Plan Your Visit

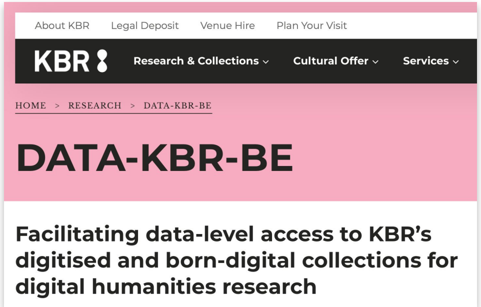
KBR Research & Collections Cultural Offer Services

HOME > RESEARCH > DIGITAL RESEARCH LAB

Digital Research Lab

A lab for text and data mining research on KBR's digitised and born-digital collections.

https://www.kbr.be/en/projects/digital-research-lab/



About KBR Legal Deposit Venue Hire Plan Your Visit

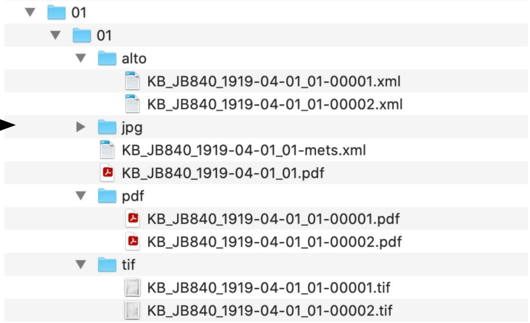
KBR Research & Collections Cultural Offer Services

HOME > RESEARCH > DATA-KBR-BE

DATA-KBR-BE

Facilitating data-level access to KBR's digitised and born-digital collections for digital humanities research

https://www.kbr.be/en/projects/data-kbr-be/



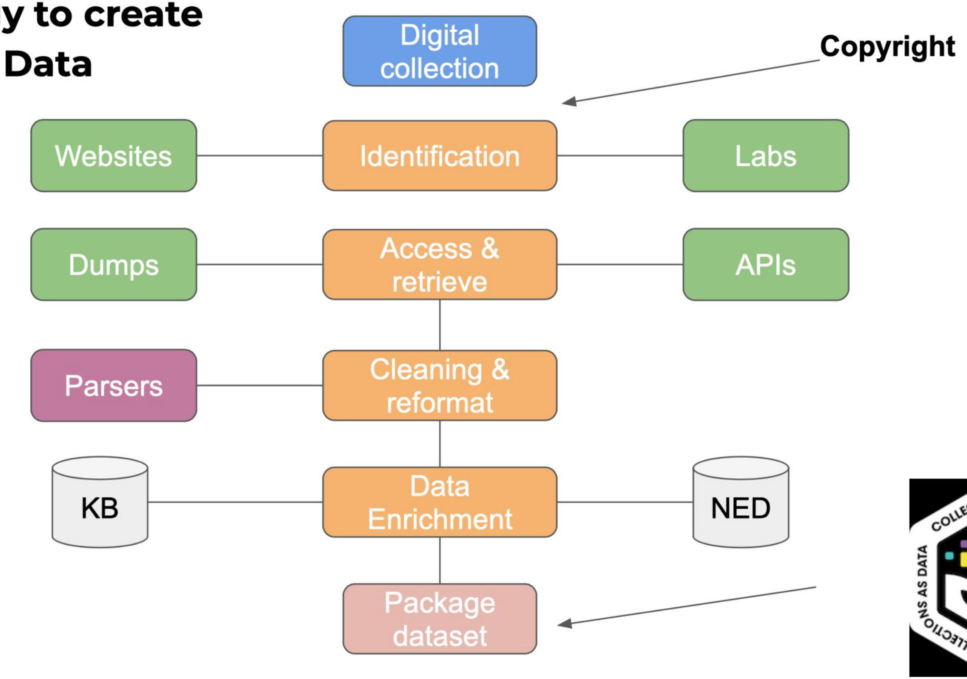
```
▼ 01
  ▼ 01
    ▼ alto
        KB_JB840_1919-04-01_01-00001.xml
        KB_JB840_1919-04-01_01-00002.xml
    ▶ jpg
      KB_JB840_1919-04-01_01-mets.xml
      KB_JB840_1919-04-01_01.pdf
    ▼ pdf
        KB_JB840_1919-04-01_01-00001.pdf
        KB_JB840_1919-04-01_01-00002.pdf
    ▼ tif
        KB_JB840_1919-04-01_01-00001.tif
        KB_JB840_1919-04-01_01-00002.tif
```

GHENT UNIVERSITY

KBR

Collections as Data Methodology

A methodology to create Collections as Data



Digital collection

Copyright

Websites

Identification

Labs

Dumps

Access & retrieve

APIs

Parsers

Cleaning & reformat

KB

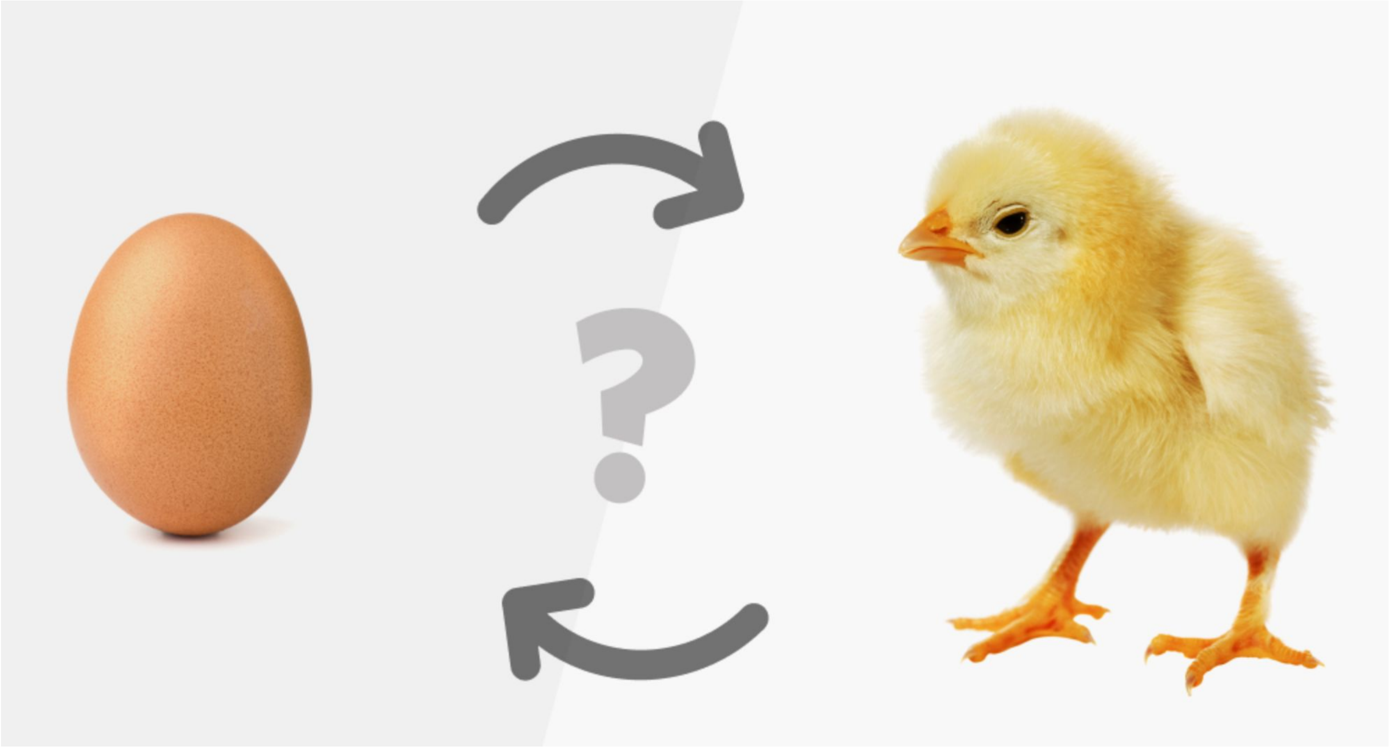
Data Enrichment

NED

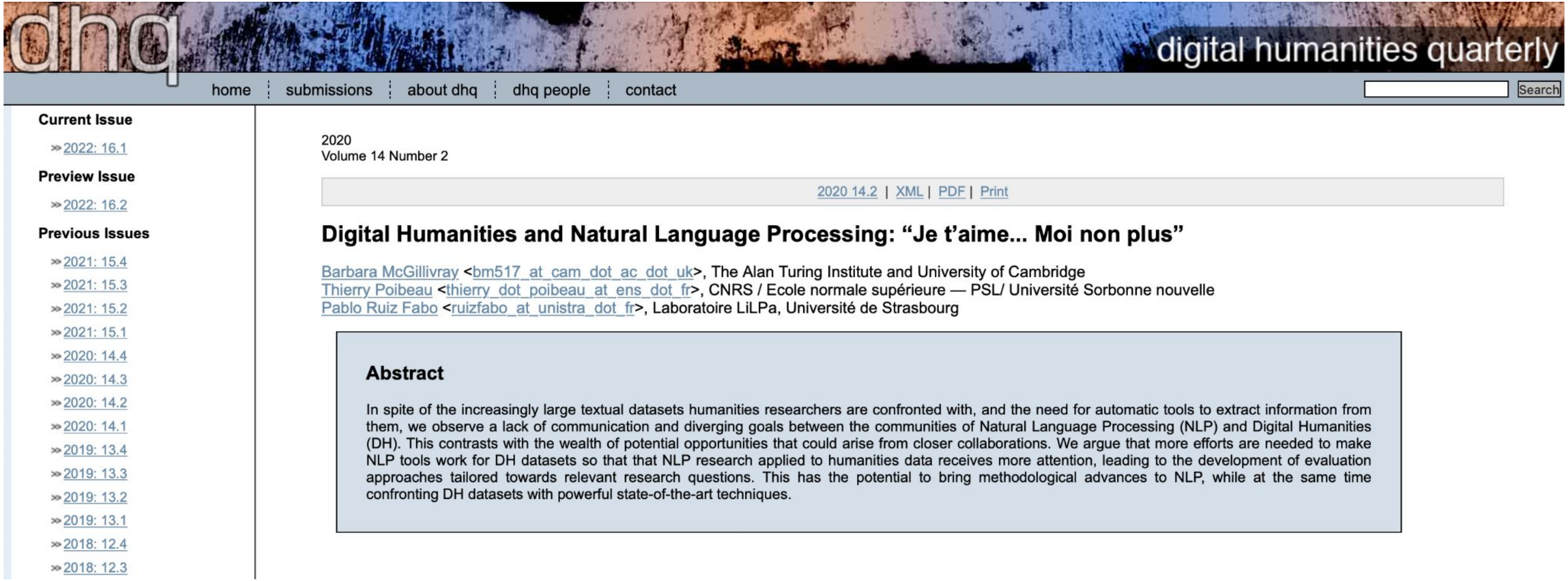
Package dataset

Candela, G., Sáez, M. D., Escobar Esteban, Mp., & Marco-Such, M. (2020). *Reusing digital collections from GLAM institutions.* Journal of Information Science: https://doi.org/10.1177/0165551520950246 & http://rua.ua.es/dspace/handle/10045/109460

Collections or Corpora?



What is a corpus … ?



2020
Volume 14 Number 2

Digital Humanities and Natural Language Processing: “Je t’aime... Moi non plus”

Barbara McGillivray <bm517_at_cam_dot_ac_dot_uk>, The Alan Turing Institute and University of Cambridge
Thierry Poibeau <thierry_dot_poibeau_at_ens_dot_fr>, CNRS / Ecole normale supérieure — PSL/ Université Sorbonne nouvelle
Pablo Ruiz Fabo <ruizfabo_at_unistra_dot_fr>, Laboratoire LiLPa, Université de Strasbourg

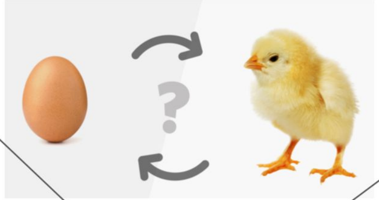
Abstract

In spite of the increasingly large textual datasets humanities researchers are confronted with, and the need for automatic tools to extract information from them, we observe a lack of communication and diverging goals between the communities of Natural Language Processing (NLP) and Digital Humanities (DH). This contrasts with the wealth of potential opportunities that could arise from closer collaborations. We argue that more efforts are needed to make NLP tools work for DH datasets so that that NLP research applied to humanities data receives more attention, leading to the development of evaluation approaches tailored towards relevant research questions. This has the potential to bring methodological advances to NLP, while at the same time confronting DH datasets with powerful state-of-the-art techniques.

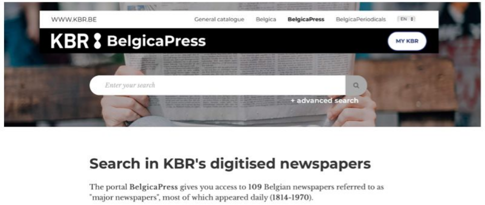
McGillivray, B., Poibeau, T., & Ruiz Fabo, P. (2020). Digital Humanities and Natural Language Processing:“Je t’aime... Moi non plus”. DHQ, Digital Humanities Quarterly. 2020, 14.2
http://www.digitalhumanities.org/dhq/vol/14/2/000454/000454.html

Collections as Data @ KBR

Research Scenarios

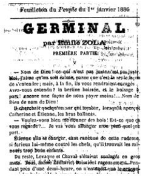


Digital collection



Search in KBR's digitised newspapers

The portal **BelgicaPress** gives you access to **109** Belgian newspapers referred to as "major newspapers", most of which appeared daily (**1814-1970**).



Feuilleton du *Peuple* du 1er janvier 1886

GERMINAL

par Emile ZOLA

PREMIÈRE PARTIE

V

— Nom de Dieu ! ce qui n'est pas juste n'est pas juste. Moi, j'aime qu'on soit calme, parce que c'est la seule façon de s'entendre ; mais, à la fin, ils vous rendraient enragés. Avez-vous entendu ? la berline boisée, et le boisage à part ! encore une façon de nous payer moins !... Nom de Dieu de nom de Dieu !

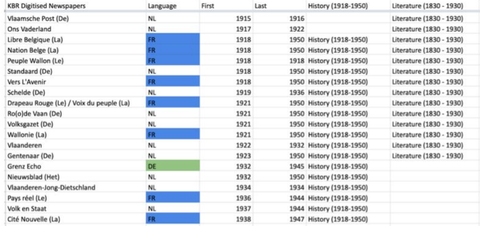
Il cherchait quelqu'un sur qui tomber, lorsqu'il aperçut Catherine et Étienne, les bras ballants.

— Voulez-vous bien me donner des bois ! Est-ce que ça vous regarde ?... Je vas vous allonger mon pied quelque part.

Étienne alla se charger, sans rancune de cette rudesse, si furieux lui-même contre les chefs, qu'il trouvait les mineurs trop bons enfants.

Du reste, Levaque et Chaval s'étaient soulagés en gros mots. Tous, même Zacharie, boisaient rageusement. Pendant près d'une demi-heure, on n'entendit que le craque-

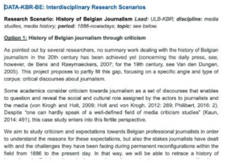
Copyright?



KBR Digitised Newspapers	Language	First	Last	History (1918-1950)	Literature (1830 - 1930)
Vlaamsche Post (De)	NL	1915	1916		Literature (1830 - 1930)
Ons Vaderland	NL	1917	1922		Literature (1830 - 1930)
Libre Belgique (La)	FR	1918	1950	History (1918-1950)	Literature (1830 - 1930)
Nation Belge (La)	FR	1918	1950	History (1918-1950)	Literature (1830 - 1930)
Peuple Wallon (Le)	FR	1918	1918	History (1918-1950)	Literature (1830 - 1930)
Standaard (De)	NL	1918	1950	History (1918-1950)	Literature (1830 - 1930)
Vers L'Avenir	FR	1918	1950	History (1918-1950)	Literature (1830 - 1930)
Schelde (De)	NL	1919	1936	History (1918-1950)	Literature (1830 - 1930)
Drapeau Rouge (Le) / Voix du peuple (La)	FR	1921	1950	History (1918-1950)	Literature (1830 - 1930)
Ro(o)de Vaan (De)	NL	1921	1950	History (1918-1950)	Literature (1830 - 1930)
Volksgazet (De)	NL	1921	1950	History (1918-1950)	Literature (1830 - 1930)
Wallonie (La)	FR	1921	1950	History (1918-1950)	Literature (1830 - 1930)
Vlaanderen	NL	1922	1932	History (1918-1950)	Literature (1830 - 1930)
Gentenaar (De)	NL	1923	1950	History (1918-1950)	Literature (1830 - 1930)
Grenz Echo	DE	1932	1945	History (1918-1950)	
Nieuwsblad (Het)	NL	1932	1950	History (1918-1950)	
Vlaanderen-Jong-Dietschland	NL	1934	1934	History (1918-1950)	
Pays réel (Le)	FR	1936	1944	History (1918-1950)	
Volk en Staat	NL	1937	1944	History (1918-1950)	
Cité Nouvelle (La)	FR	1938	1947	History (1918-1950)	

Identification

Formats?



DATA-KBR-BE: Interdisciplinary Research Scenarios

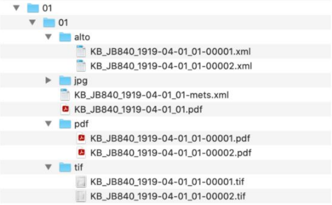
Research Scenario: History of Belgian Journalism ***Lead:*** *ULB-KBR;* ***discipline:*** *media studies, media history;* ***period:*** *1886-nowadays;* ***topic:*** *see below.*

Option 1: History of Belgian journalism through criticism

As pointed out by several researchers, no summary work dealing with the history of Belgian journalism in the 20th century has been achieved yet (concerning the daily press, see, however, de Bens and Raeymaeckers, 2007; for the 19th century, see Van den Dungen, 2005). This project proposes to partly fill this gap, focusing on a specific angle and type of corpus: critical discourses about journalism.

Some academics consider criticism towards journalism as a set of discourses that enables to question and reveal the social and cultural role assigned by the actors to journalists and the media (von Krogh and Holt, 2009; Holt and von Krogh, 2012: 289; Philibert, 2016: 2). Despite "one can hardly speak of a well-defined field of media criticism studies" (Kaun, 2014: 491), this case study enters into this fertile perspective.

We aim to study criticism and expectations towards Belgian professional journalists in order to understand the reasons for these expectations, but also the stakes journalists have dealt with and the challenges they have been facing during permanent reconfigurations within the field from 1886 to the present day. In that way, we will be able to retrace a history of

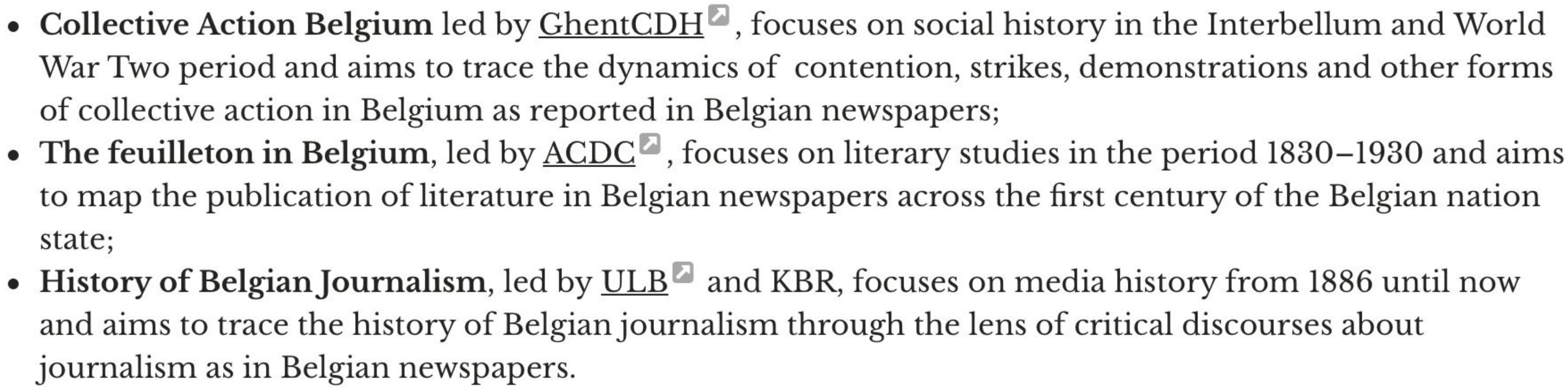


```
▼ 01
  ▼ 01
    ▼ alto
        KB_JB840_1919-04-01_01-00001.xml
        KB_JB840_1919-04-01_01-00002.xml
    ▶ jpg
      KB_JB840_1919-04-01_01-mets.xml
      KB_JB840_1919-04-01_01.pdf
    ▼ pdf
        KB_JB840_1919-04-01_01-00001.pdf
        KB_JB840_1919-04-01_01-00002.pdf
    ▼ tif
        KB_JB840_1919-04-01_01-00001.tif
        KB_JB840_1919-04-01_01-00002.tif
```

Access & retrieve

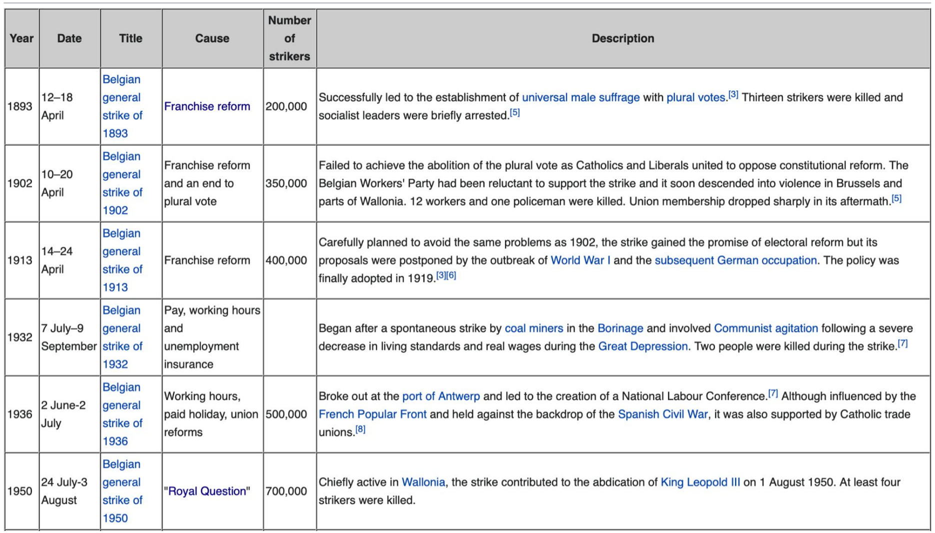
Candela, G., Sáez, M. D., Escobar Esteban, Mp., & Marco-Such, M. (2020). *Reusing digital collections from GLAM institutions*. Journal of Information Science: https://doi.org/10.1177/0165551520950246 & http://rua.ua.es/dspace/handle/10045/109460

Interdisciplinary Research Scenarios



- **Collective Action Belgium** led by GhentCDH, focuses on social history in the Interbellum and World War Two period and aims to trace the dynamics of contention, strikes, demonstrations and other forms of collective action in Belgium as reported in Belgian newspapers;
- **The feuilleton in Belgium**, led by ACDC, focuses on literary studies in the period 1830–1930 and aims to map the publication of literature in Belgian newspapers across the first century of the Belgian nation state;
- **History of Belgian Journalism**, led by ULB and KBR, focuses on media history from 1886 until now and aims to trace the history of Belgian journalism through the lens of critical discourses about journalism as in Belgian newspapers.

From collections to corpora



Year	Date	Title	Cause	Number of strikers	Description
1893	12–18 April	Belgian general strike of 1893	Franchise reform	200,000	Successfully led to the establishment of universal male suffrage with plural votes.[3] Thirteen strikers were killed and socialist leaders were briefly arrested.[5]
1902	10–20 April	Belgian general strike of 1902	Franchise reform and an end to plural vote	350,000	Failed to achieve the abolition of the plural vote as Catholics and Liberals united to oppose constitutional reform. The Belgian Workers' Party had been reluctant to support the strike and it soon descended into violence in Brussels and parts of Wallonia. 12 workers and one policeman were killed. Union membership dropped sharply in its aftermath.[5]
1913	14–24 April	Belgian general strike of 1913	Franchise reform	400,000	Carefully planned to avoid the same problems as 1902, the strike gained the promise of electoral reform but its proposals were postponed by the outbreak of World War I and the subsequent German occupation. The policy was finally adopted in 1919.[3][6]
1932	7 July–9 September	Belgian general strike of 1932	Pay, working hours and unemployment insurance		Began after a spontaneous strike by coal miners in the Borinage and involved Communist agitation following a severe decrease in living standards and real wages during the Great Depression. Two people were killed during the strike.[7]
1936	2 June-2 July	Belgian general strike of 1936	Working hours, paid holiday, union reforms	500,000	Broke out at the port of Antwerp and led to the creation of a National Labour Conference.[7] Although influenced by the French Popular Front and held against the backdrop of the Spanish Civil War, it was also supported by Catholic trade unions.[8]
1950	24 July-3 August	Belgian general strike of 1950	"Royal Question"	700,000	Chiefly active in Wallonia, the strike contributed to the abdication of King Leopold III on 1 August 1950. At least four strikers were killed.

https://en.wikipedia.org/wiki/General_strikes_in_Belgium

GHENT UNIVERSITY

KBR

Which newspapers are digitised?

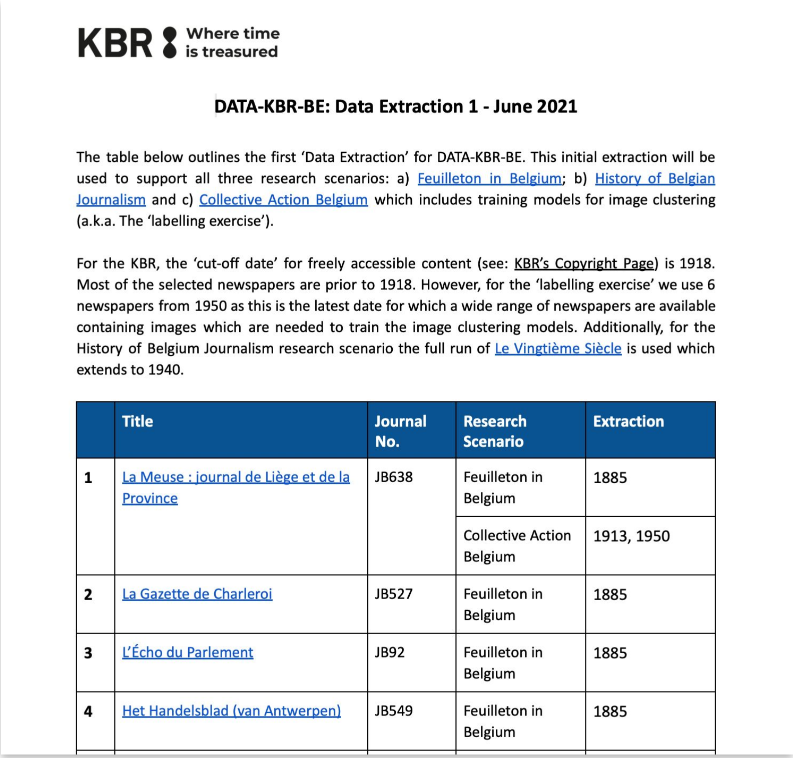
	Title	Journal No.	Digitally availability	1886	1893	1902	1913	1932	1936	1950	Extraction
1	Vooruit: socialistisch dagblad	JB 809	1884-1889; 1901-1902; 1911-1950	Yes	No	Yes	Yes	Yes	Yes	Yes	1913, 1950
2	Het Volk : antisocialistisch dagblad	JB 785	1911-1916; 18 apr. 1918-31 maart 1927; 1931-1950.	No	No	No	Yes	Yes	Yes	Yes	1913, 1950
3	Vaderland	JB 310	16 maart 1910-30 dec. 1913.	No	No	No	Yes	No	No	No	1913, 1950
4	Le Peuple : organe quotidien de la démocratie socialiste	JB837	1885-1907; 1911-1914; 1918-1940; 1944-1950.	Yes	Yes	Yes	Yes	Yes	Yes	Yes	1913, 1950
5	Le Vingtième Siècle	JB729	6 juin 1895-13 mai 1940	Yes	Yes	Yes	Yes	Yes	Yes	No	1913, 1950
6	La Meuse : journal de Liège et de la Province	JB638	1856; 1858-1882; 1884-5 août 1914; 28 nov. 1918-11 mai 1940; 9 sept. 1944-1950.	No	Yes	Yes	Yes	Yes	Yes	Yes	1913, 1950

Importance of historical context

> “The newspaper was founded in Ghent in 1884 with the support of the socialist cooperative Vooruit. When the Parti ouvrier belge - Belgische Werkliedenpartij (“Belgian Workers Party”, POB-BWP, 1885) was founded, it was recognised as its official organ for the Flemish part of the country. The Ghent socialist leader Edward Anseele was editor-in-chief, but he also worked as a typographer. *Vooruit* was published under German censorship during the two world wars. Having reached its peak in the 1950s, it began a slow decline. It was succeeded by the daily *De Morgen* in 1978.”

Vooruit: socialistisch dagblad

DATA-KBR-BE Data Extraction 1



KBR Where time is treasured

DATA-KBR-BE: Data Extraction 1 - June 2021

The table below outlines the first ‘Data Extraction’ for DATA-KBR-BE. This initial extraction will be used to support all three research scenarios: a) Feuilleton in Belgium; b) History of Belgian Journalism and c) Collective Action Belgium which includes training models for image clustering (a.k.a. The ‘labelling exercise’).

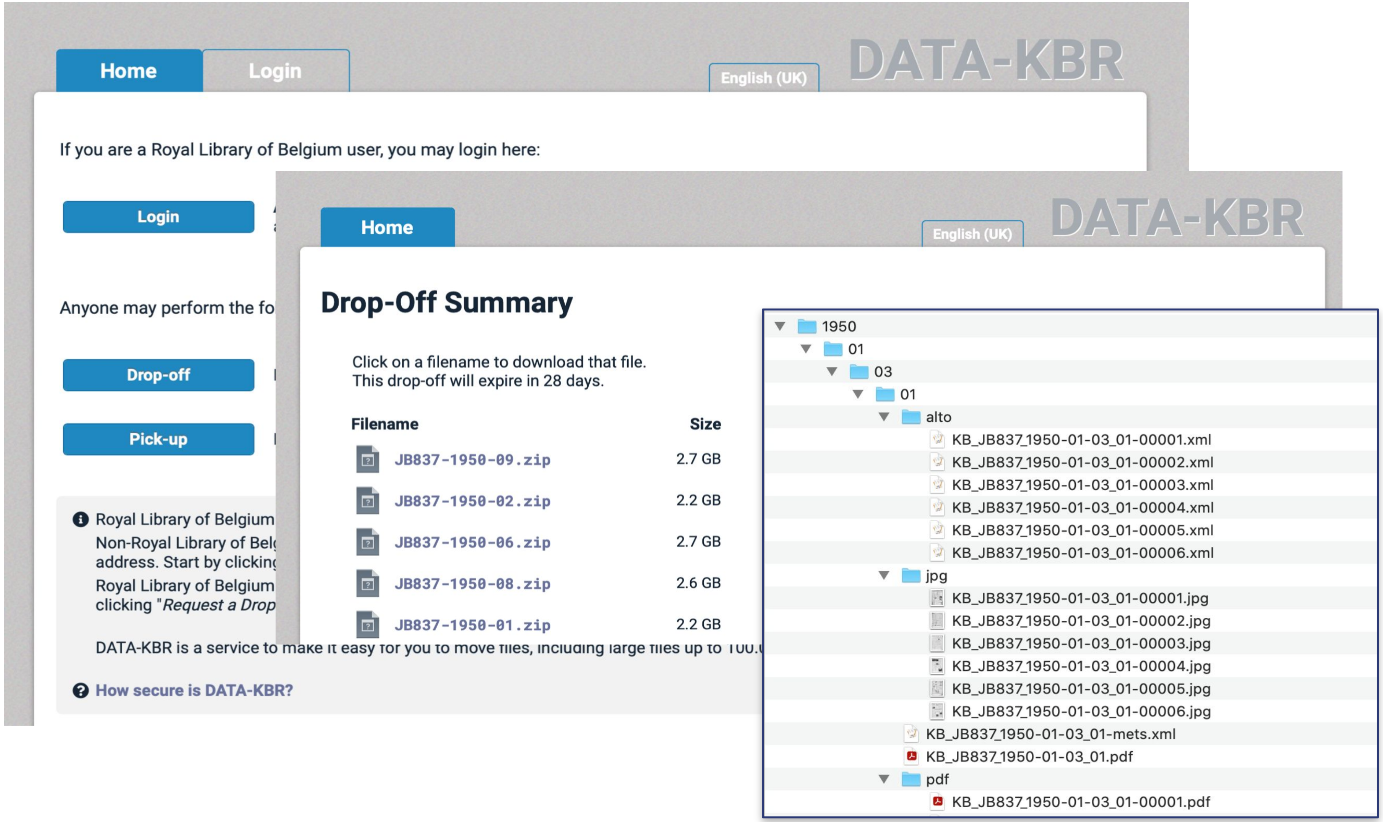
For the KBR, the ‘cut-off date’ for freely accessible content (see: KBR’s Copyright Page) is 1918. Most of the selected newspapers are prior to 1918. However, for the ‘labelling exercise’ we use 6 newspapers from 1950 as this is the latest date for which a wide range of newspapers are available containing images which are needed to train the image clustering models. Additionally, for the History of Belgium Journalism research scenario the full run of Le Vingtième Siècle is used which extends to 1940.

	Title	Journal No.	Research Scenario	Extraction
1	La Meuse : journal de Liège et de la Province	JB638	Feuilleton in Belgium	1885
			Collective Action Belgium	1913, 1950
2	La Gazette de Charleroi	JB527	Feuilleton in Belgium	1885
3	L’Écho du Parlement	JB92	Feuilleton in Belgium	1885
4	Het Handelsblad (van Antwerpen)	JB549	Feuilleton in Belgium	1885

- 3 Research Scenarios
- 12 Newspaper Titles
- 3 years: 1885, 1913, 1950
- 18 ‘drop offs’

(list of titles for extraction)

Methods for Data Sharing: KBR Send Data



If you are a Royal Library of Belgium user, you may login here:

Anyone may perform the fo

Royal Library of Belgium
Non-Royal Library of Belg
address. Start by clicking
Royal Library of Belgium
clicking "*Request a Drop*

DATA-KBR is a service to make it easy for you to move files, including large files up to 100.

How secure is DATA-KBR?

Drop-Off Summary

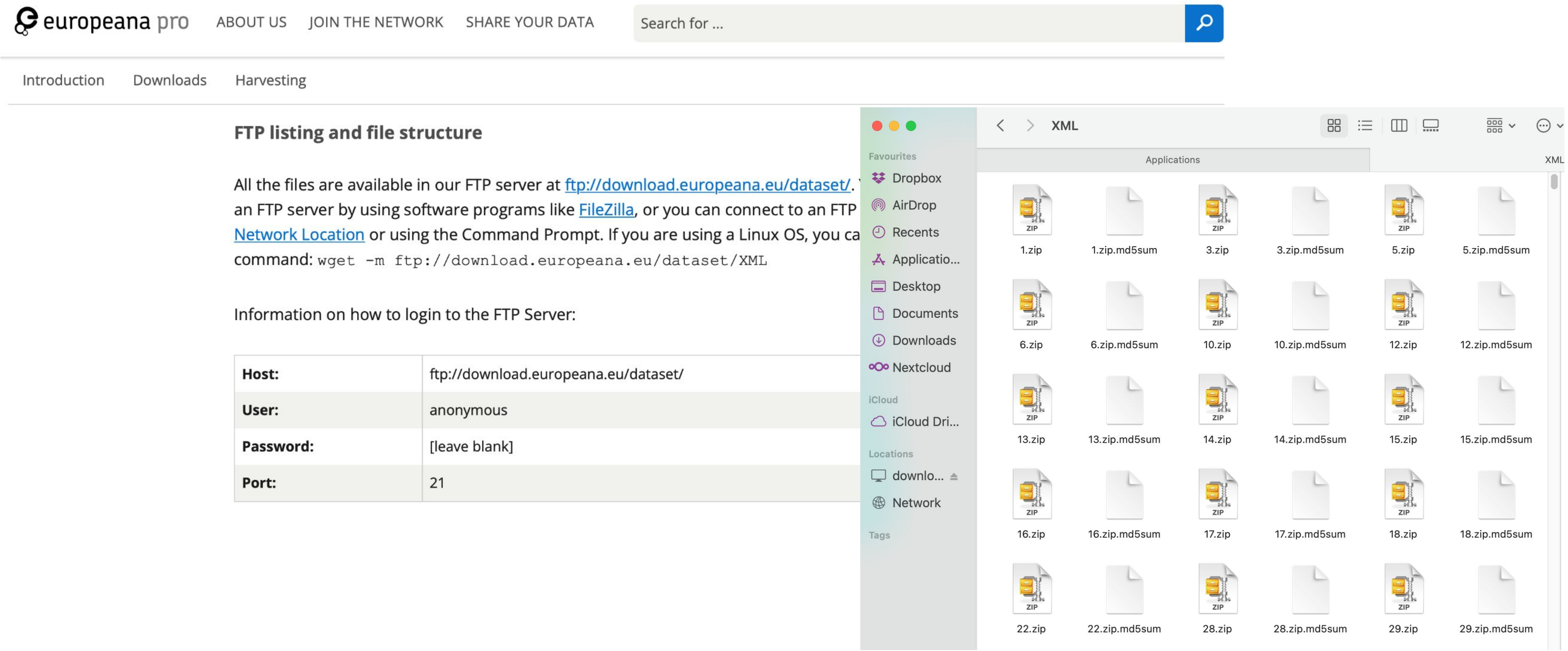
Click on a filename to download that file.
This drop-off will expire in 28 days.

Filename	Size
JB837-1950-09.zip	2.7 GB
JB837-1950-02.zip	2.2 GB
JB837-1950-06.zip	2.7 GB
JB837-1950-08.zip	2.6 GB
JB837-1950-01.zip	2.2 GB

```
▼ 1950
  ▼ 01
    ▼ 03
      ▼ 01
        ▼ alto
            KB_JB837_1950-01-03_01-00001.xml
            KB_JB837_1950-01-03_01-00002.xml
            KB_JB837_1950-01-03_01-00003.xml
            KB_JB837_1950-01-03_01-00004.xml
            KB_JB837_1950-01-03_01-00005.xml
            KB_JB837_1950-01-03_01-00006.xml
        ▼ jpg
            KB_JB837_1950-01-03_01-00001.jpg
            KB_JB837_1950-01-03_01-00002.jpg
            KB_JB837_1950-01-03_01-00003.jpg
            KB_JB837_1950-01-03_01-00004.jpg
            KB_JB837_1950-01-03_01-00005.jpg
            KB_JB837_1950-01-03_01-00006.jpg
          KB_JB837_1950-01-03_01-mets.xml
          KB_JB837_1950-01-03_01.pdf
        ▼ pdf
            KB_JB837_1950-01-03_01-00001.pdf
```

https://datasend.kbr.be

Methods for Data Sharing



FTP listing and file structure

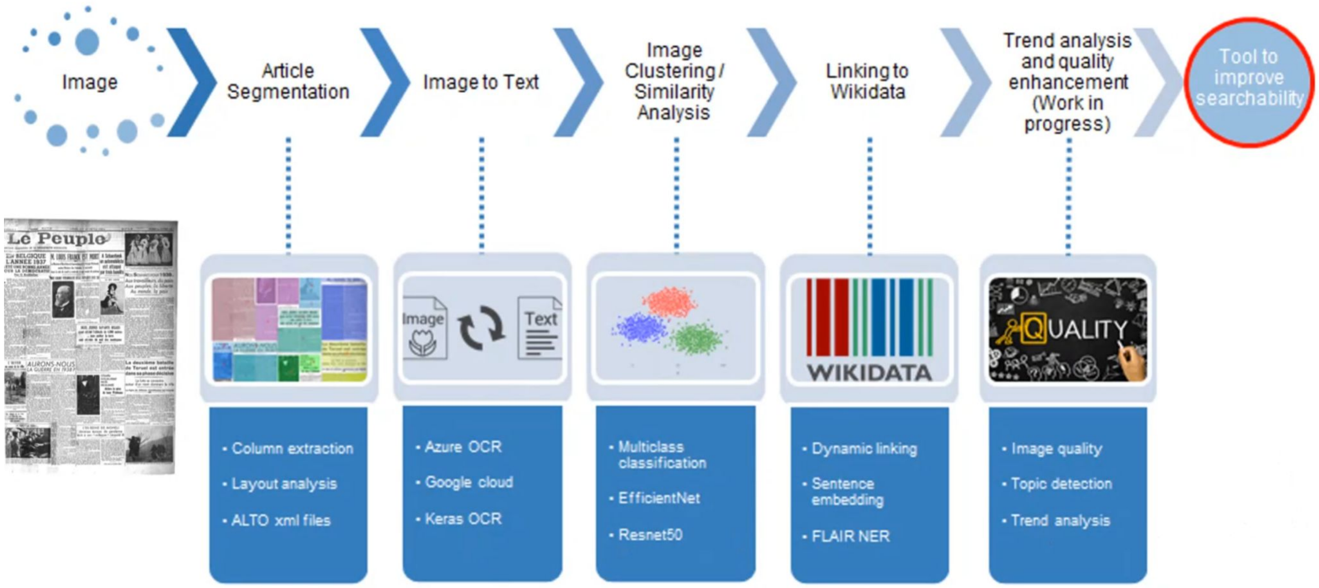
All the files are available in our FTP server at ftp://download.europeana.eu/dataset/. \
an FTP server by using software programs like FileZilla, or you can connect to an FTP
Network Location or using the Command Prompt. If you are using a Linux OS, you ca
command: `wget -m ftp://download.europeana.eu/dataset/XML`

Information on how to login to the FTP Server:

Host:	ftp://download.europeana.eu/dataset/
User:	anonymous
Password:	[leave blank]
Port:	21

https://pro.europeana.eu/page/harvesting-and-downloads#downloads

Interdisciplinary collaboration with data scientists



Image

Article Segmentation
- Column extraction
- Layout analysis
- ALTO xml files

Image to Text
- Azure OCR
- Google cloud
- Keras OCR

Image Clustering / Similarity Analysis
- Multiclass classification
- EfficientNet
- Resnet50

Linking to Wikidata
- Dynamic linking
- Sentence embedding
- FLAIR NER

Trend analysis and quality enhancement (Work in progress)
- Image quality
- Topic detection
- Trend analysis

Tool to improve searchability

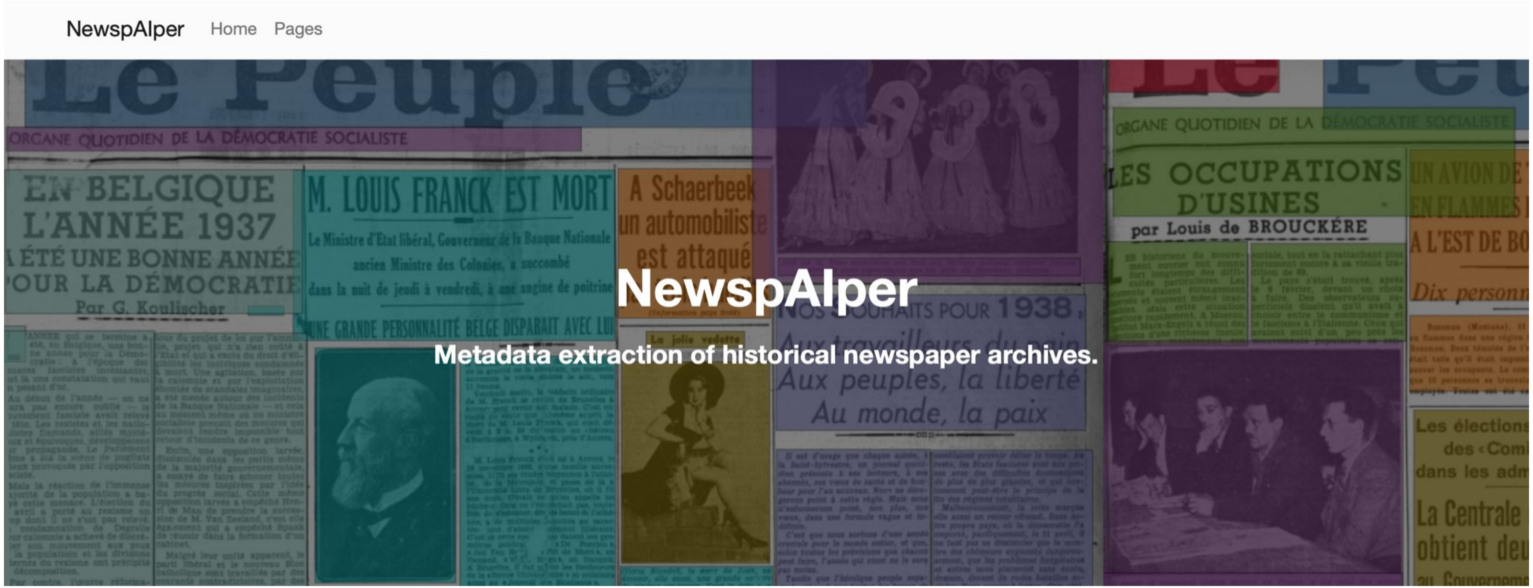
Dilawar Ali, Kenzo Milleville, Alec Van de Broeck & Steven Verstockt, IDLab, UGent

GHENT UNIVERSITY

KBR

NewspAIper Demonstrator

Dilawar Ali, Kenzo Milleville, Alec Van de Broeck & Steven Verstockt, IDLab, UGent



NewspAIper

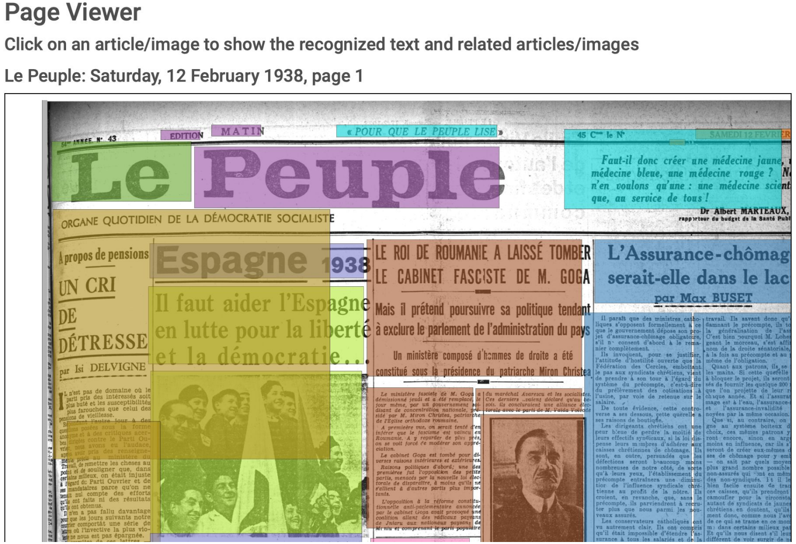
Metadata extraction of historical newspaper archives.

GHENT UNIVERSITY

KBR

https://tw06v072.ugent.be/kbr/

NewspAIper Demonstrator: Document Layout Analysis & Article Separation

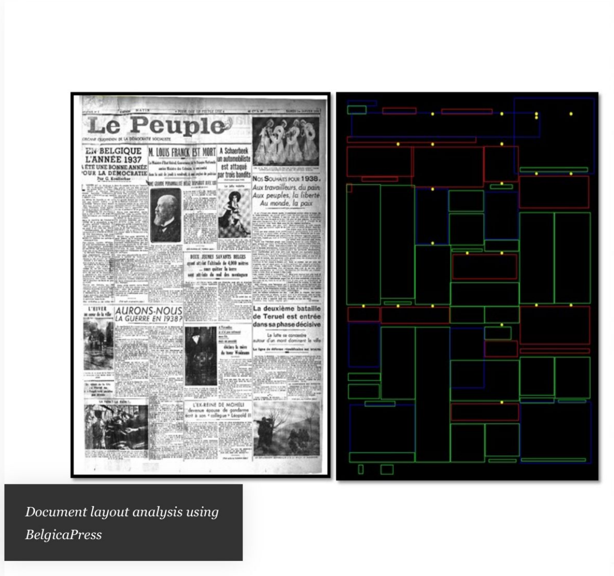


Page Viewer

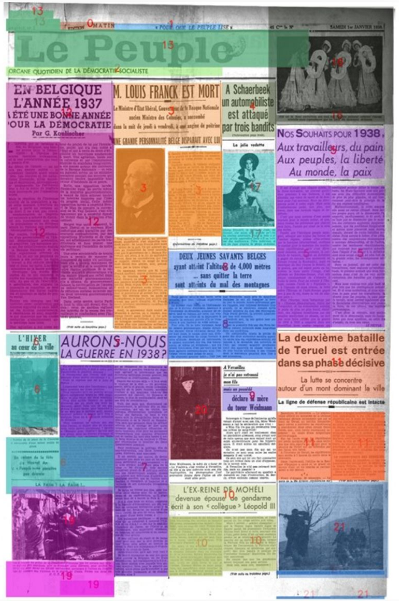
Click on an article/image to show the recognized text and related articles/images

Le Peuple: Saturday, 12 February 1938, page 1

Le Peuple
ORGANE QUOTIDIEN DE LA DÉMOCRATIE SOCIALISTE
EDITION MATIN
« POUR QUE LE PEUPLE LISE »
45 Cmes le N°
SAMEDI 12 FEVRIER
Faut-il donc créer une médecine jaune, médecine bleue, une médecine rouge ? Nous n'en voulons qu'une : une médecine scientifique, au service de tous !
Dr Albert MARTEAUX, rapporteur du budget de la Santé Publique
A propos de pensions
UN CRI DE DÉTRESSE
par Isi DELVIGNE
Espagne 1938
Il faut aider l'Espagne en lutte pour la liberté et la démocratie...
LE ROI DE ROUMANIE A LAISSÉ TOMBER LE CABINET FASCISTE DE M. GOGA
Mais il prétend poursuivre sa politique tendant à exclure le parlement de l'administration du pays
Un ministère composé d'hommes de droite a été constitué sous la présidence du patriarche Miron Christea
L'Assurance-chômage serait-elle dans le lac
par Max BUSET



Document layout analysis using BelgicaPress



Le Peuple
ORGANE QUOTIDIEN DE LA DÉMOCRATIE SOCIALISTE
EN BELGIQUE L'ANNÉE 1937 A ÉTÉ UNE BONNE ANNÉE POUR LA DÉMOCRATIE
M. LOUIS FRANCK EST MORT
A Schaerbeek un automobiliste est attaqué par trois bandits
NOS SOUHAITS POUR 1938
Aux travailleurs, du pain
Aux peuples, la liberté
Au monde, la paix
DEUX JEUNES SAVANTS BELGES
AURONS-NOUS LA GUERRE EN 1938 ?
La deuxième bataille de Teruel est entrée dans sa phase décisive
L'EX-REINE DE MOHELI

https://tw06v072.ugent.be/kbr/

GHENT UNIVERSITY

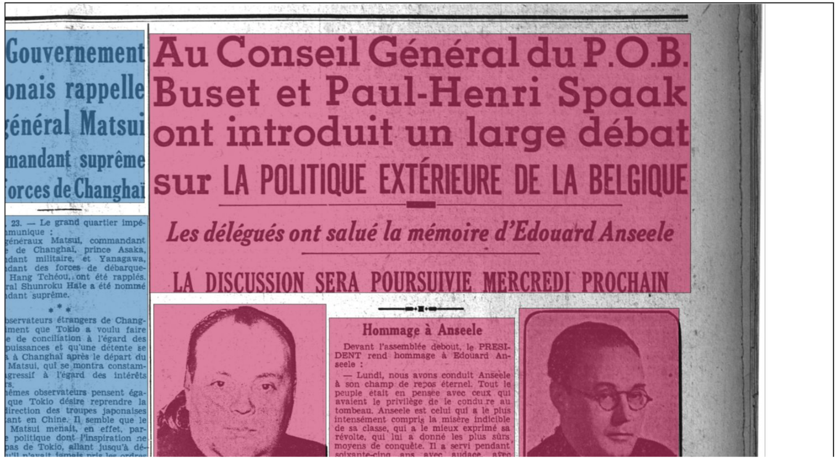
KBR

NewspAIper Demonstrator: Named Entity Recognition & Linking

Page Viewer

Click on an article/image to show the recognized text and related articles/images

Le Peuple: Thursday, 24 February 1938, page 1



Gouvernement
onais rappelle
énéral Matsui
mandant suprême
forces de Changhaï

23. — Le grand quartier impé-
mmunique :
généraux Matsui, commandant
e de Changhaï, prince Asaka,
ndant militaire, et Yanagawa,
ndant des forces de débarque-
Hang Tchéou, ont été rappelés.
ral Shunroku Hate a été nommé
ndant suprême.

* * *

bservateurs étrangers de Chang-
iment que Tokio a voulu faire
e de conciliation à l'égard des
puissances et qu'une détente se
à Changhaï après le départ du
Matsui, qui se montra constam-
gressif à l'égard des intérêts
rs.
mêmes observateurs pensent éga-
que Tokio désire reprendre la
direction des troupes japonaises
ant en Chine. Il semble que le
Matsui menait, en effet, par-
e politique dont l'inspiration ne
pas de Tokio, allant jusqu'à dé-

Au Conseil Général du P.O.B. Buset et Paul-Henri Spaak ont introduit un large débat sur LA POLITIQUE EXTÉRIEURE DE LA BELGIQUE

Les délégués ont salué la mémoire d'Edouard Anseele

LA DISCUSSION SERA POURSUIVIE MERCREDI PROCHAIN

Hommage à Anseele

Devant l'assemblée debout, le PRESIDENT rend hommage à Edouard Anseele :

— Lundi, nous avons conduit Anseele à son champ de repos éternel. Tout le peuple était en pensée avec ceux qui avaient le privilège de le conduire au tombeau. Anseele est celui qui a le plus intensément compris la misère indicible de sa classe, qui a le mieux exprimé sa révolte, qui lui a donné les plus sûrs moyens de conquête. Il a servi pendant soixante-cinq ans avec audace, avec

Most similar articles

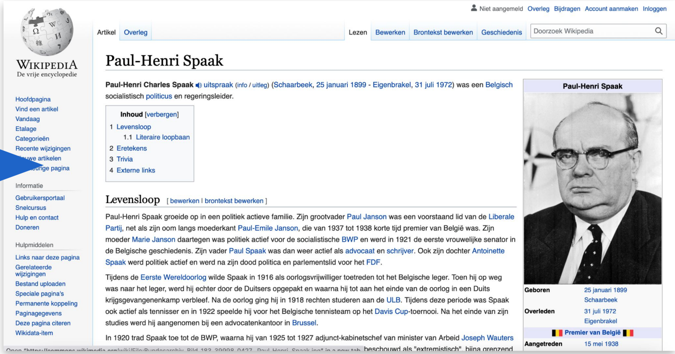
Samedi matin, devant de nombreux délégués réunis à la Maison du Peuple de 'Bruxelles le camarade D.

TROISIÈME JOURNÉE DE DISCUSSION sur la Politique étrangère au Conseil Général du P.O.B. CORNEILLE ME

DEVANT LE PARLEMENT Dans un discours vivement acclamé Paul-Henri Spaak souligne les mérites de la po

Article text:

Au Conseil Général du P.O.B. Buset Paul-Henri Spaak ont introduit un large débat sur LA POLITIQUE EXTÉRIEURE DE LA BEL GIQUE Les délégués ont salué la mémoire d'Edouard Anseele LA DISCUSSION SERA POURSUIVIE MERCREDI PROCHAIN



Paul-Henri Spaak

Paul-Henri Charles Spaak uitspraak (info / uitleg) (Schaarbeek, 25 januari 1899 - Eigenbrakel, 31 juli 1972) was een Belgisch socialistisch politicus en regeringsleider.

Inhoud [verbergen]

1 Levensloop
  1.1 Literaire loopbaan
2 Eretekens
3 Trivia
4 Externe links

Levensloop [bewerken | brontekst bewerken]

Paul-Henri Spaak groeide op in een politiek actieve familie. Zijn grootvader Paul Janson was een voorstaand lid van de Liberale Partij, net als zijn oom langs moederkant Paul-Emile Janson, die van 1937 tot 1938 korte tijd premier van België was. Zijn moeder Marie Janson daartegen was politiek actief voor de socialistische BWP en werd in 1921 de eerste vrouwelijke senator in de Belgische geschiedenis. Zijn vader Paul Spaak was dan weer actief als advocaat en schrijver. Ook zijn dochter Antoinette Spaak werd politiek actief en werd na zijn dood politica en parlementslid voor het FDF.

Tijdens de Eerste Wereldoorlog wilde Spaak in 1916 als oorlogsvrijwilliger toetreden tot het Belgische leger. Toen hij op weg was naar het leger, werd hij echter door de Duitsers opgepakt en waarna hij tot aan het einde van de oorlog in een Duits krijgsgevangenenkamp verbleef. Na de oorlog ging hij in 1918 rechten studeren aan de ULB. Tijdens deze periode was Spaak ook actief als tennisser en in 1922 speelde hij voor het Belgische tennisteam op het Davis Cup-toernooi. Na het einde van zijn studies werd hij aangenomen bij een advocatenkantoor in Brussel.

In 1920 trad Spaak toe tot de BWP, waarna hij van 1925 tot 1927 adjunct-kabinetschef van minister van Arbeid Joseph Wauters

Open "https://commons.wikimedia.org/wiki/File:Bundesarchiv_Bild_183-39998-0427,_Paul_Henri_Spaak.jpg" in a new tab beschouwd als "extremistisch", bijna grenzend

Paul-Henri Spaak	
Geboren	25 januari 1899 Schaarbeek
Overleden	31 juli 1972 Eigenbrakel
Premier van België	
Aangetreden	15 mei 1938

https://tw06v072.ugent.be/kbr/

GHENT UNIVERSITY

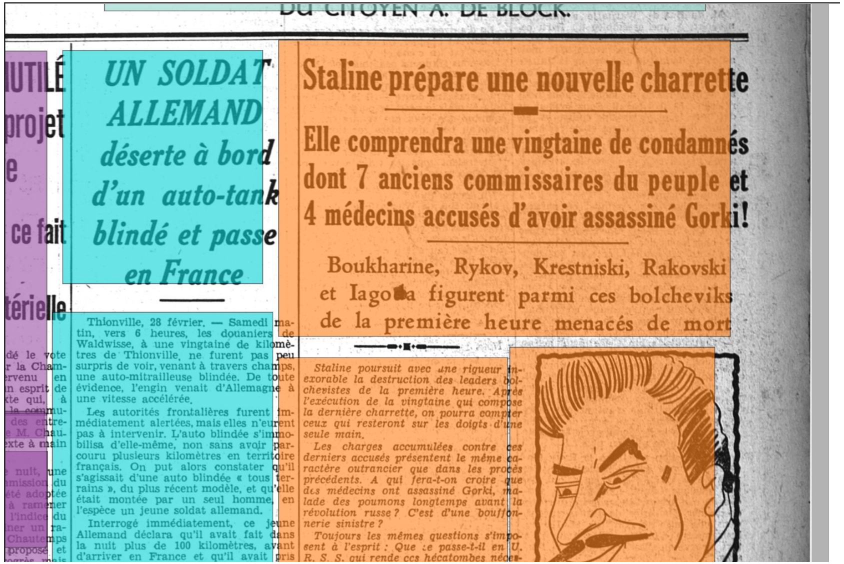
KBR

NewspAIper Demonstrator - Image similarity

Page Viewer

Click on an article/image to show the recognized text and related articles/images

Le Peuple: Tuesday, 1 March 1938, page 1



DU CITOYEN A. DE BLOCK.

UTILÉ
projet
e
ce fait
térielle

dé le vote
r la Cham-
rvenu en
n esprit de
xte qui, à
la commu-
des entre-
e M. Chau-
exte à main

e nuit, une
mission du
été adoptée
à ramener
l'indice du
ner un ra-
Chautemps
proposé et

UN SOLDAT ALLEMAND déserte à bord d'un auto-tank blindé et passe en France

Thionville, 28 février. — Samedi matin, vers 6 heures, les douaniers de Waldwisse, à une vingtaine de kilomètres de Thionville, ne furent pas peu surpris de voir, venant à travers champs, une auto-mitrailleuse blindée. De toute évidence, l'engin venait d'Allemagne à une vitesse accélérée.

Les autorités frontalières furent immédiatement alertées, mais elles n'eurent pas à intervenir. L'auto blindée s'immobilisa d'elle-même, non sans avoir parcouru plusieurs kilomètres en territoire français. On put alors constater qu'il s'agissait d'une auto blindée « tous terrains », du plus récent modèle, et qu'elle était montée par un seul homme, en l'espèce un jeune soldat allemand.

Interrogé immédiatement, ce jeune Allemand déclara qu'il avait fait dans la nuit plus de 100 kilomètres, avant d'arriver en France et qu'il avait pris

Staline prépare une nouvelle charrette

Elle comprendra une vingtaine de condamnés dont 7 anciens commissaires du peuple et 4 médecins accusés d'avoir assassiné Gorki!

Boukharine, Rykov, Krestniski, Rakovski et Iagoda figurent parmi ces bolcheviks de la première heure menacés de mort

Staline poursuit avec une rigueur inexorable la destruction des leaders bolchevistes de la première heure. Après l'exécution de la vingtaine qui compose la dernière charrette, on pourra compter ceux qui resteront sur les doigts d'une seule main.

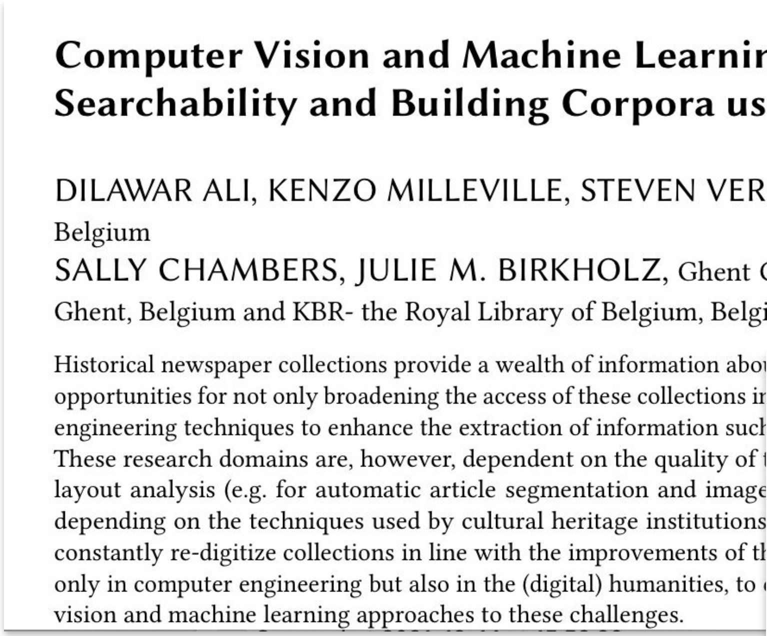
Les charges accumulées contre ces derniers accusés présentent le même caractère outrancier que dans les procès précédents. A qui fera-t-on croire que des médecins ont assassiné Gorki, malade des poumons longtemps avant la révolution russe? C'est d'une bouffonnerie sinistre?

Toujours les mêmes questions s'imposent à l'esprit : Que se passe-t-il en U. R. S. S. qui rende ces hécatombes néces-

Most similar images:

Most similar images:

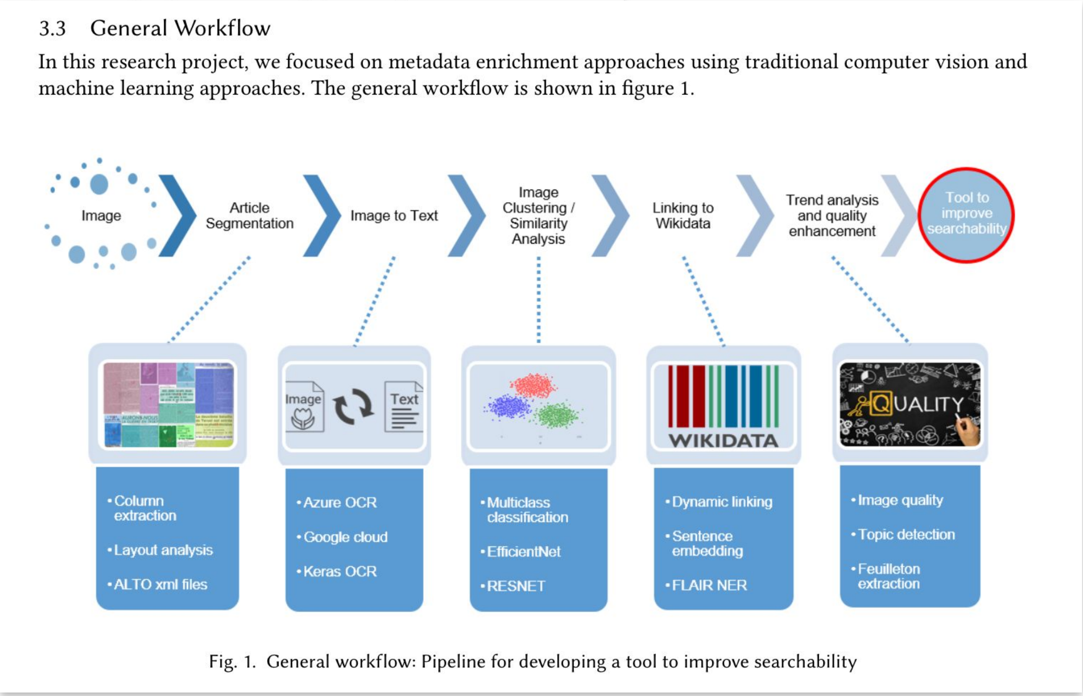
https://tw06v072.ugent.be/kbr/



Computer Vision and Machine Learning Approaches to Improve Searchability and Building Corpora using Historical Newspapers

DILAWAR ALI, KENZO MILLEVILLE, STEVEN VERSTOCKT, IDLab, Ghent University-imec, Ghent, Belgium
SALLY CHAMBERS, JULIE M. BIRKHOLZ, Ghent Centre for Digital Humanities, Ghent University, Ghent, Belgium and KBR- the Royal Library of Belgium, Belgium

Historical newspaper collections provide a wealth of information abou
opportunities for not only broadening the access of these collections ir
engineering techniques to enhance the extraction of information such
These research domains are, however, dependent on the quality of t
layout analysis (e.g. for automatic article segmentation and image
depending on the techniques used by cultural heritage institutions
constantly re-digitize collections in line with the improvements of th
only in computer engineering but also in the (digital) humanities, to c
vision and machine learning approaches to these challenges.

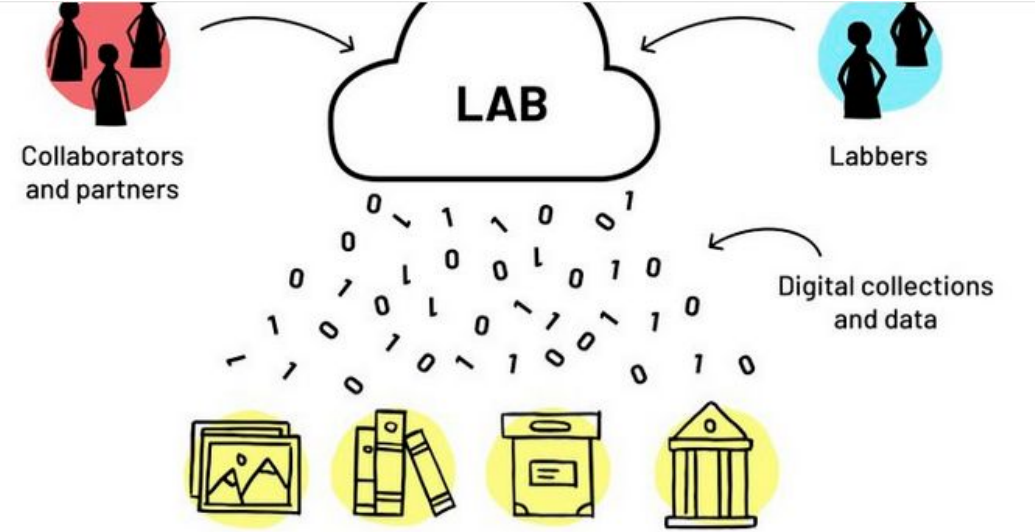


3.3 General Workflow

In this research project, we focused on metadata enrichment approaches using traditional computer vision and machine learning approaches. The general workflow is shown in figure 1.

Fig. 1. General workflow: Pipeline for developing a tool to improve searchability

International GLAM Labs Community



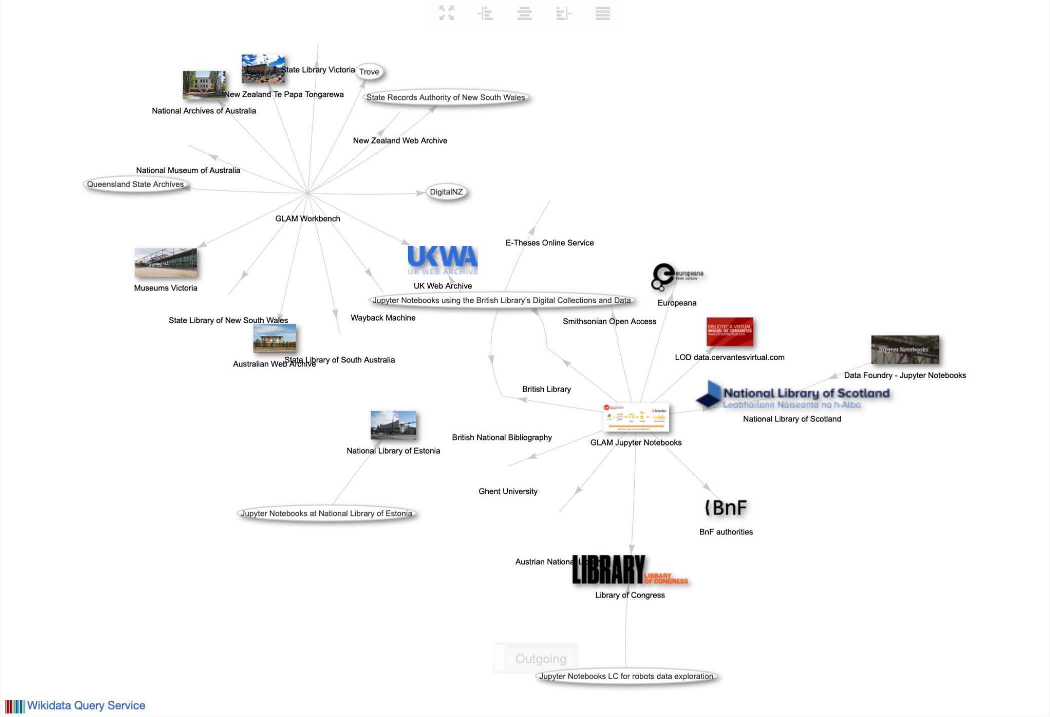
OPEN A GLAM LAB

https://glamlabs.io/books/

Chapter: Rethinking Collections as Data: https://glamlabs.pubpub.org/pub/urhm68yr/release/1

Computational access to digital collections

International
GLAM Labs
Community

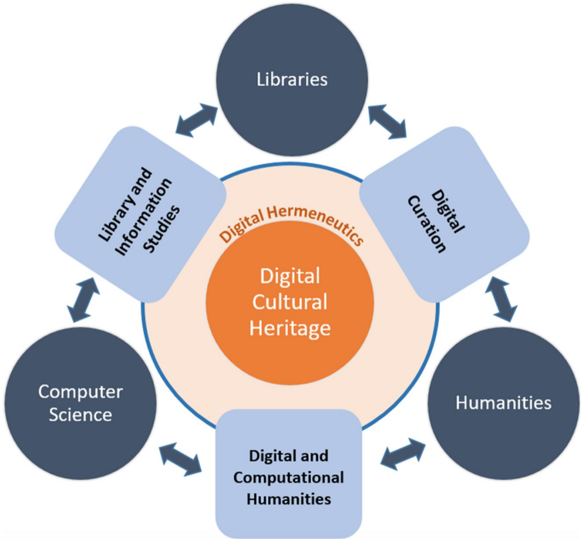


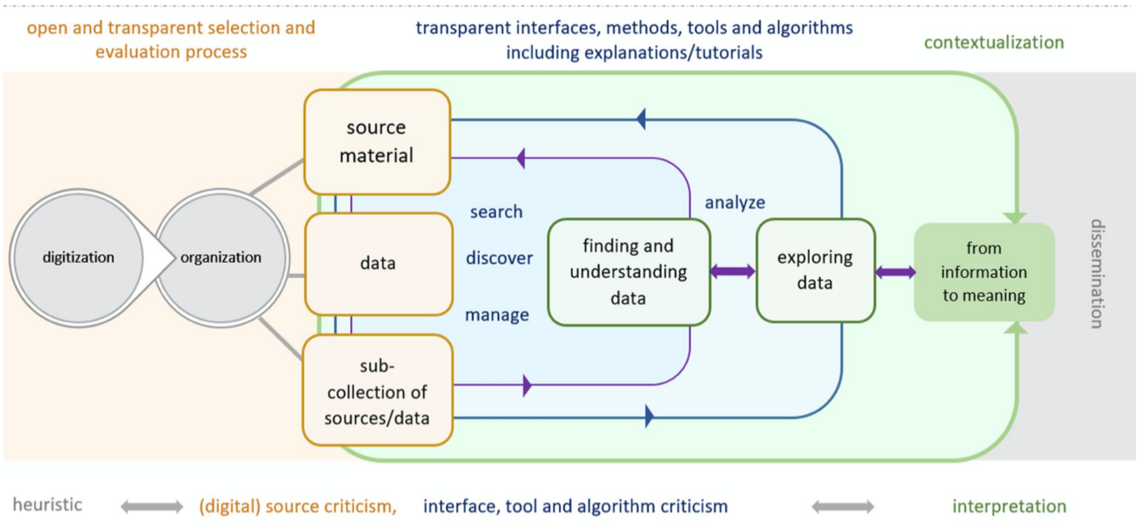
State Library Victoria
Trove
New Zealand Te Papa Tongarewa
National Archives of Australia
State Records Authority of New South Wales
New Zealand Web Archive
National Museum of Australia
Queensland State Archives
DigitalNZ
GLAM Workbench
E-Theses Online Service
UKWA
UK WEB ARCHIVE
UK Web Archive
europeana
Europeana
Museums Victoria
Jupyter Notebooks using the British Library's Digital Collections and Data
State Library of New South Wales
Wayback Machine
Smithsonian Open Access
LOD data.cervantesvirtual.com
Australian Web Archive
State Library of South Australia
Data Foundry - Jupyter Notebooks
National Library of Scotland
Leabharlann Nàiseanta na h-Alba
British Library
National Library of Scotland
National Library of Estonia
British National Bibliography
GLAM Jupyter Notebooks
Ghent University
Jupyter Notebooks at National Library of Estonia
BnF
BnF authorities
Austrian Nationa
LIBRARY
LIBRARY OF CONGRESS
Library of Congress
Outgoing
Jupyter Notebooks LC for robots data exploration
Wikidata Query Service

International GLAM Labs Community: Computational Access to Digital Collections

https://glamlabs.io/computational-access-to-digital-collections/

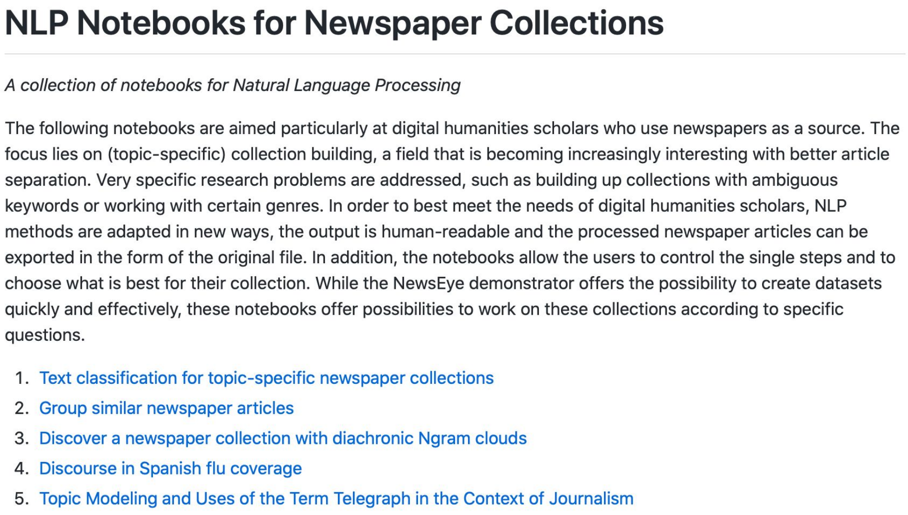
Corpus Building: an interdisciplinary digital hermeneutics workflow


Libraries
Library and Information Studies
Digital Curation
Digital Hermeneutics
Digital Cultural Heritage
Computer Science
Humanities
Digital and Computational Humanities


open and transparent selection and evaluation process
transparent interfaces, methods, tools and algorithms including explanations/tutorials
contextualization
digitization
organization
source material
data
sub-collection of sources/data
search
discover
manage
analyze
finding and understanding data
exploring data
from information to meaning
dissemination
heuristic
(digital) source criticism,
interface, tool and algorithm criticism
interpretation

Oberbichler, S. et al. (2021) *Integrated interdisciplinary workflows for research on historical newspapers: Perspectives from humanities scholars, computer scientists, and librarians*. JASIST, August 2021. https://doi.org/10.1002/asi.24565

Historical Data Analysis using Jupyter Notebooks

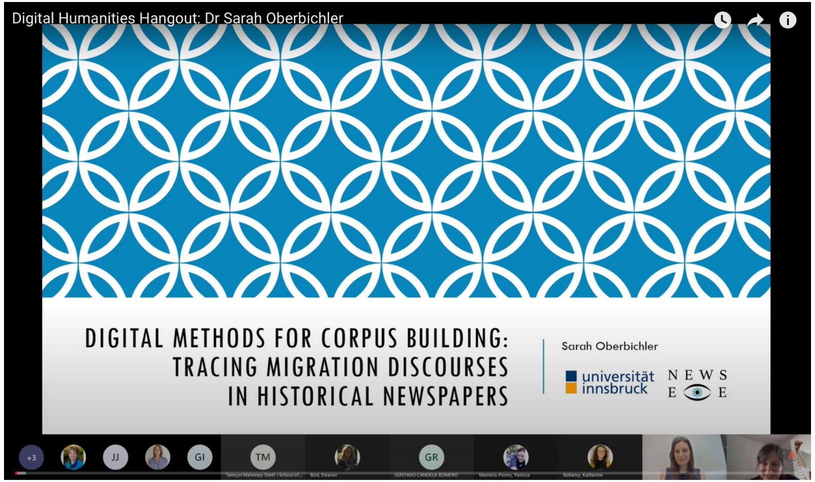


NLP Notebooks for Newspaper Collections

A collection of notebooks for Natural Language Processing

The following notebooks are aimed particularly at digital humanities scholars who use newspapers as a source. The focus lies on (topic-specific) collection building, a field that is becoming increasingly interesting with better article separation. Very specific research problems are addressed, such as building up collections with ambiguous keywords or working with certain genres. In order to best meet the needs of digital humanities scholars, NLP methods are adapted in new ways, the output is human-readable and the processed newspaper articles can be exported in the form of the original file. In addition, the notebooks allow the users to control the single steps and to choose what is best for their collection. While the NewsEye demonstrator offers the possibility to create datasets quickly and effectively, these notebooks offer possibilities to work on these collections according to specific questions.

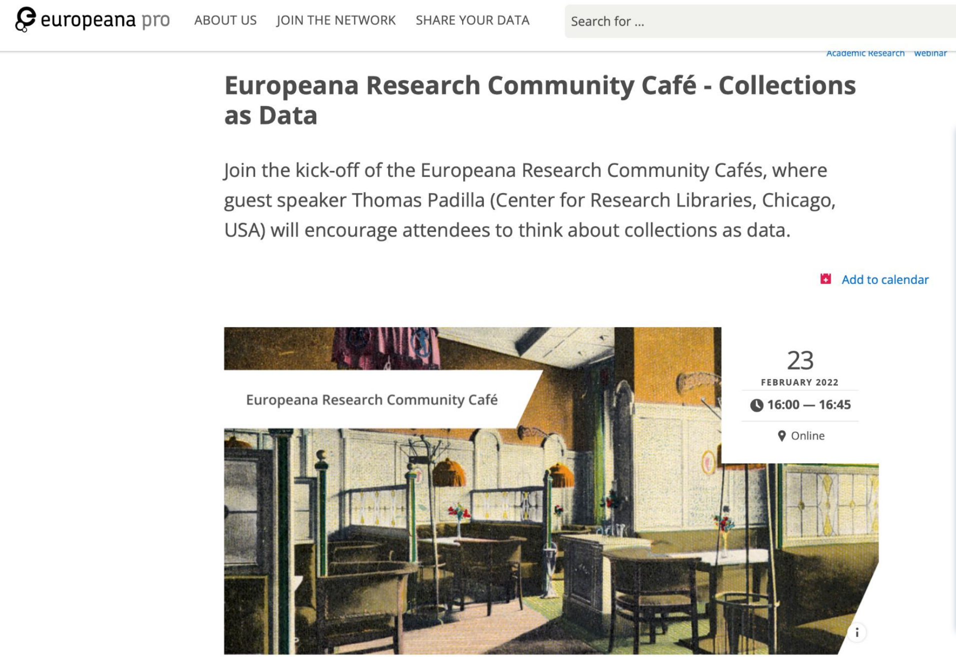
1. Text classification for topic-specific newspaper collections
2. Group similar newspaper articles
3. Discover a newspaper collection with diachronic Ngram clouds
4. Discourse in Spanish flu coverage
5. Topic Modeling and Uses of the Term Telegraph in the Context of Journalism


Digital Humanities Hangout: Dr Sarah Oberbichler
DIGITAL METHODS FOR CORPUS BUILDING:
TRACING MIGRATION DISCOURSES
IN HISTORICAL NEWSPAPERS
Sarah Oberbichler
universität innsbruck
NEWS EYE
+3
JJ
GI
TM
GR

Oberbichler, S. (2020) NLP Notebooks for Newspaper Collections. https://github.com/NewsEye/NLP-Notebooks-Newspaper-Collections

Oberbichler, S. (2021) Digital Methods for Corpus Building: Tracing Migration Discourses in Historical Newspapers, DH Hangout (Lancaster, Ghent, Lisbon), June 2021: https://www.youtube.com/watch?v=5w8X0qXP67M

Europeana Research Community Café - Collections as Data



europeana pro ABOUT US JOIN THE NETWORK SHARE YOUR DATA Search for ...

Academic Research webinar webinars

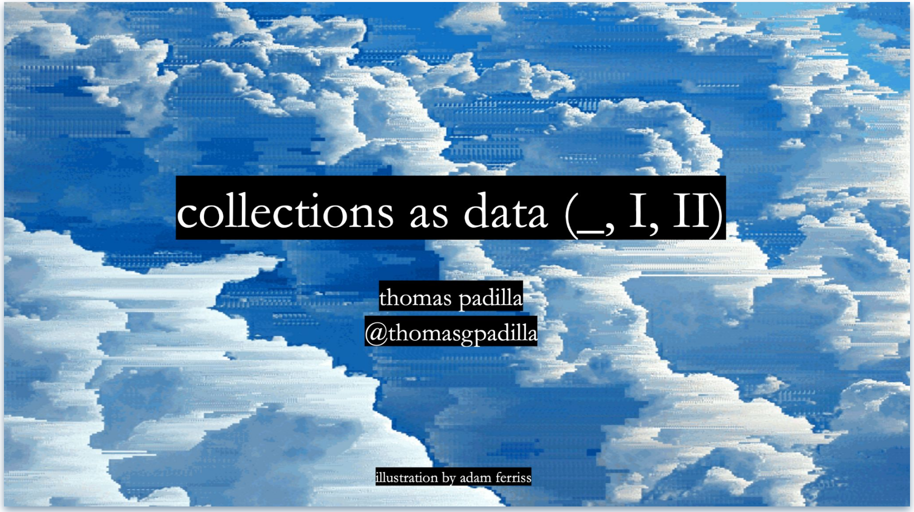
Europeana Research Community Café - Collections as Data

Join the kick-off of the Europeana Research Community Cafés, where guest speaker Thomas Padilla (Center for Research Libraries, Chicago, USA) will encourage attendees to think about collections as data.

Add to calendar

Europeana Research Community Café

23
FEBRUARY 2022
16:00 — 16:45
Online



collections as data (_, I, II)

thomas padilla
@thomasgpadilla

illustration by adam ferriss

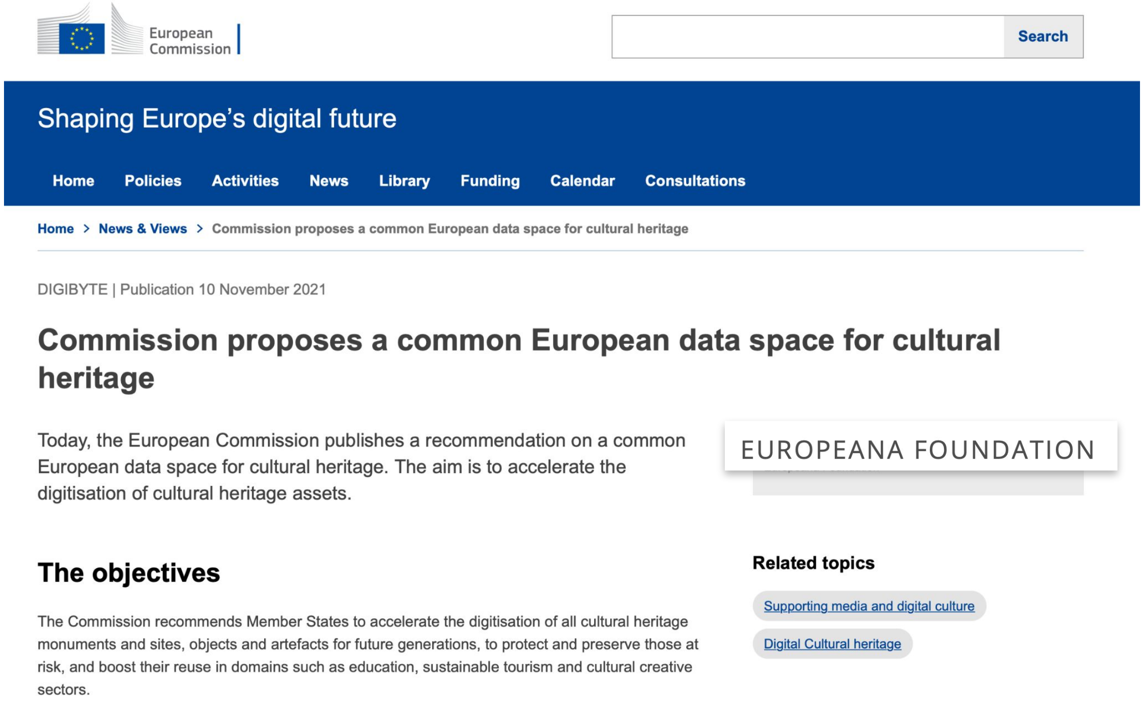
https://pro.europeana.eu/event/europeana-research-community-cafe-collections-as-data

Thomas Padilla’s presentation

https://docs.google.com/presentation/d/1XM71isRXEjKesaWPZjSrZiKytq2XC2QJ-gqKWZF

GHENT UNIVERSITY

A European data space for cultural heritage



Shaping Europe’s digital future

Home > News & Views > Commission proposes a common European data space for cultural heritage

DIGIBYTE | Publication 10 November 2021

Commission proposes a common European data space for cultural heritage

Today, the European Commission publishes a recommendation on a common European data space for cultural heritage. The aim is to accelerate the digitisation of cultural heritage assets.

EUROPEANA FOUNDATION

The objectives

The Commission recommends Member States to accelerate the digitisation of all cultural heritage monuments and sites, objects and artefacts for future generations, to protect and preserve those at risk, and boost their reuse in domains such as education, sustainable tourism and cultural creative sectors.

Related topics

Supporting media and digital culture

Digital Cultural heritage

EUROPEAN COMMISSION

Brussels, 10.11.2021
C(2021) 7953 final

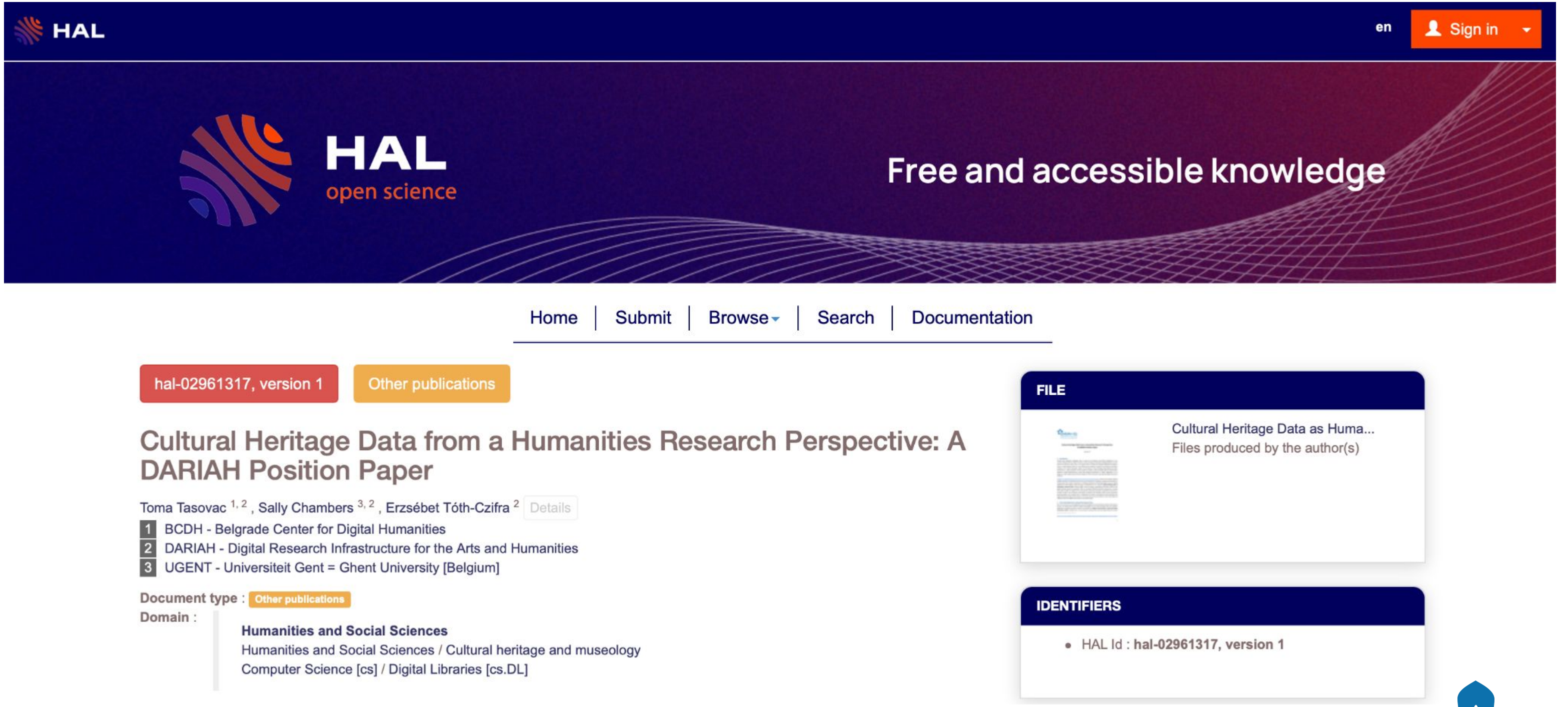
COMMISSION RECOMMENDATION

of 10.11.2021

on a common European data space for cultural heritage

https://digital-strategy.ec.europa.eu/en/news/commission-proposes-common-european-data-space-cultural-heritage

Cultural Heritage Data is Humanities Research Data


HAL
en
Sign in
HAL
open science
Free and accessible knowledge
Home | Submit | Browse | Search | Documentation
FILE
IDENTIFIERS

hal-02961317, version 1 | Other publications

Cultural Heritage Data from a Humanities Research Perspective: A DARIAH Position Paper

Toma Tasovac [1, 2] , Sally Chambers [3, 2] , Erzsébet Tóth-Czifra [2] Details

[1] BCDH - Belgrade Center for Digital Humanities
[2] DARIAH - Digital Research Infrastructure for the Arts and Humanities
[3] UGENT - Universiteit Gent = Ghent University [Belgium]

Document type : Other publications

Domain :

Humanities and Social Sciences
Humanities and Social Sciences / Cultural heritage and museology
Computer Science [cs] / Digital Libraries [cs.DL]

Cultural Heritage Data as Huma...
Files produced by the author(s)

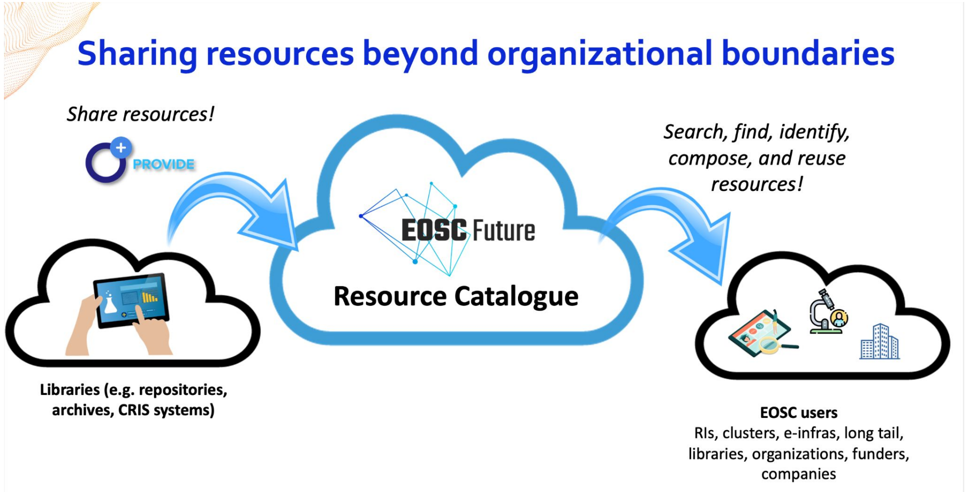
- HAL Id : **hal-02961317, version 1**

Tasovac, T., Chambers, S. and Tóth-Czifra, E. (2020) Cultural Heritage Data from a Humanities Research Perspective: A DARIAH Position Paper. https://hal.archives-ouvertes.fr/hal-02961317

DARIAH-EU
Digital Research Infrastructure
for the Arts and Humanities

European Open Science Cloud (EOSC) and Libraries

EUROPEAN OPEN SCIENCE CLOUD



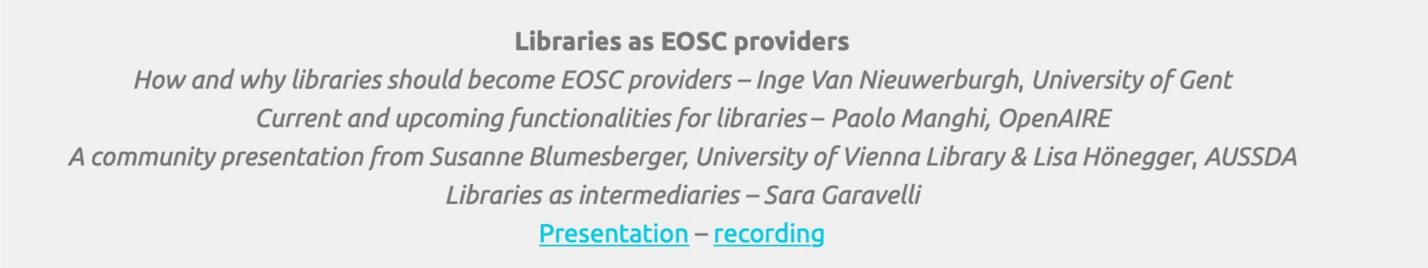
Sharing resources beyond organizational boundaries

Share resources!

Search, find, identify, compose, and reuse resources!

Libraries (e.g. repositories, archives, CRIS systems)

EOSC users
RIs, clusters, e-infras, long tail, libraries, organizations, funders, companies



Libraries as EOSC providers

How and why libraries should become EOSC providers – Inge Van Nieuwerburgh, University of Gent

Current and upcoming functionalities for libraries – Paolo Manghi, OpenAIRE

A community presentation from Susanne Blumesberger, University of Vienna Library & Lisa Hönegger, AUSSDA

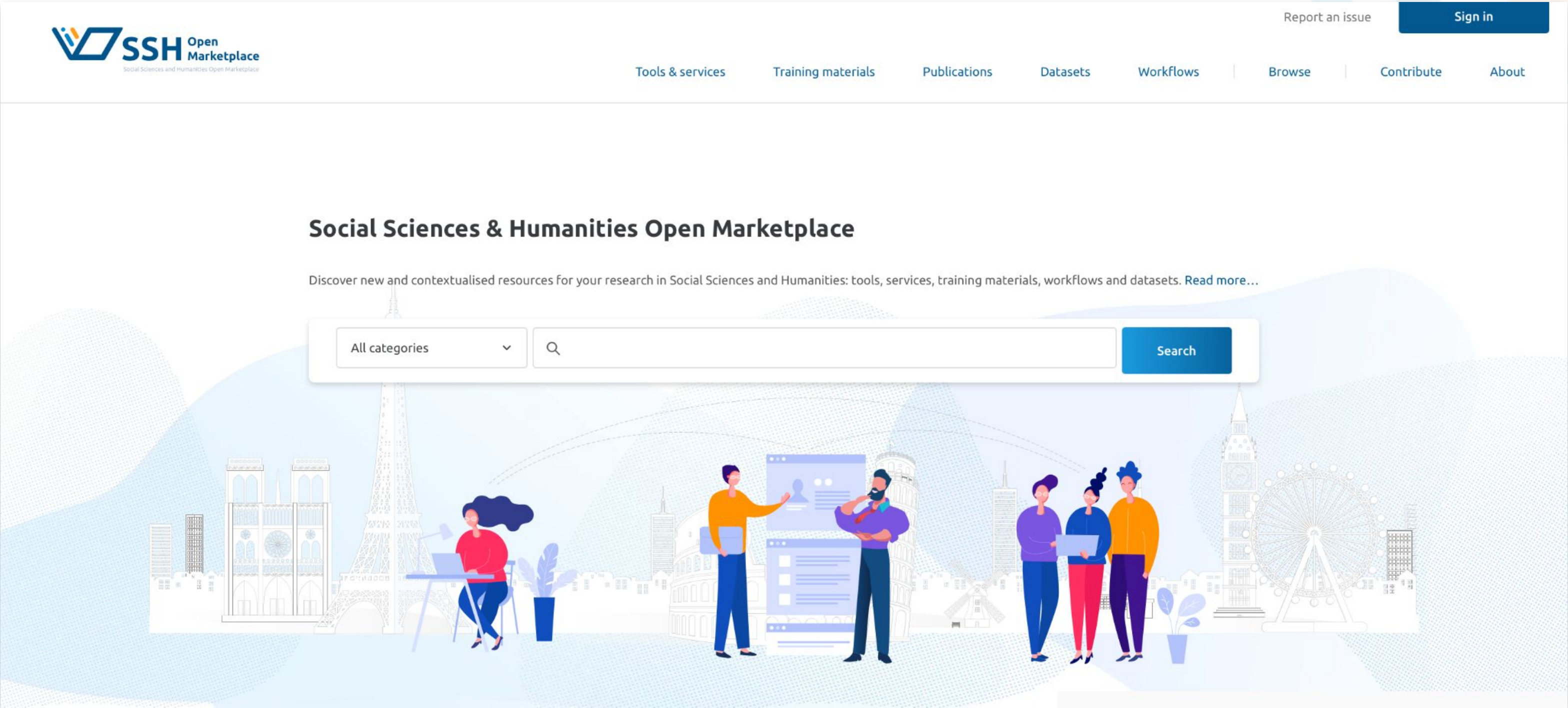
Libraries as intermediaries – Sara Garavelli

Presentation – recording

Presentation: https://eoscfuture.eu/wp-content/uploads/2022/04/ProviderDays_Libraries_Presentation2Inge_clean.pdf

Recording: https://www.youtube.com/watch?v=_vxHJNU78Pw

Social Sciences and Humanities Open Marketplace



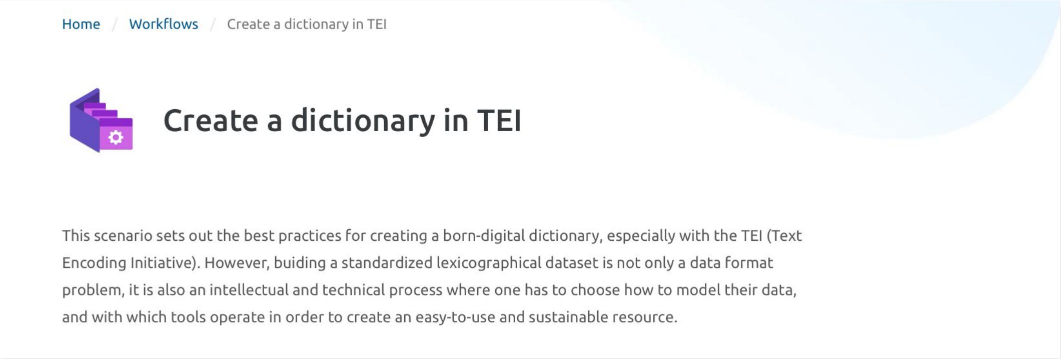
Social Sciences & Humanities Open Marketplace

Discover new and contextualised resources for your research in Social Sciences and Humanities: tools, services, training materials, workflows and datasets. Read more…

https://marketplace.sshopencloud.eu

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC Open Marketplace: from tools to workflows

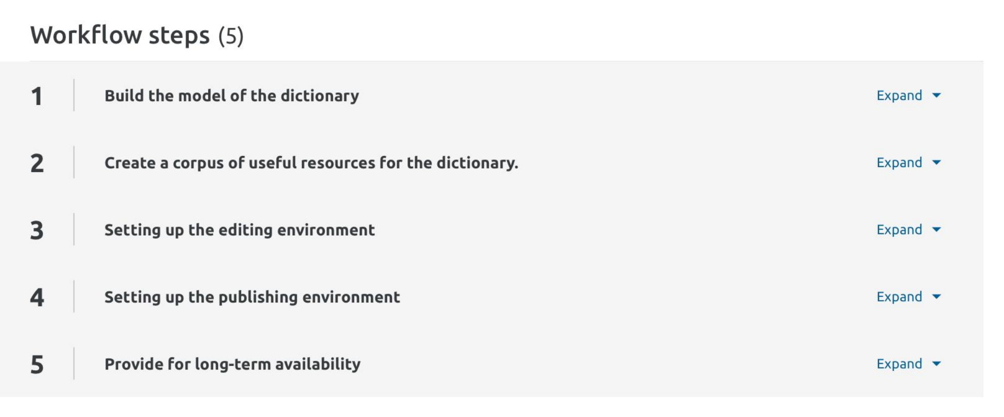


Home / Workflows / Create a dictionary in TEI

Create a dictionary in TEI

This scenario sets out the best practices for creating a born-digital dictionary, especially with the TEI (Text Encoding Initiative). However, buiding a standardized lexicographical dataset is not only a data format problem, it is also an intellectual and technical process where one has to choose how to model their data, and with which tools operate in order to create an easy-to-use and sustainable resource.

Photo by Joshua Hoehne on Unsplash



Workflow steps (5)

1 Build the model of the dictionary — Expand ▾

2 Create a corpus of useful resources for the dictionary. — Expand ▾

3 Setting up the editing environment — Expand ▾

4 Setting up the publishing environment — Expand ▾

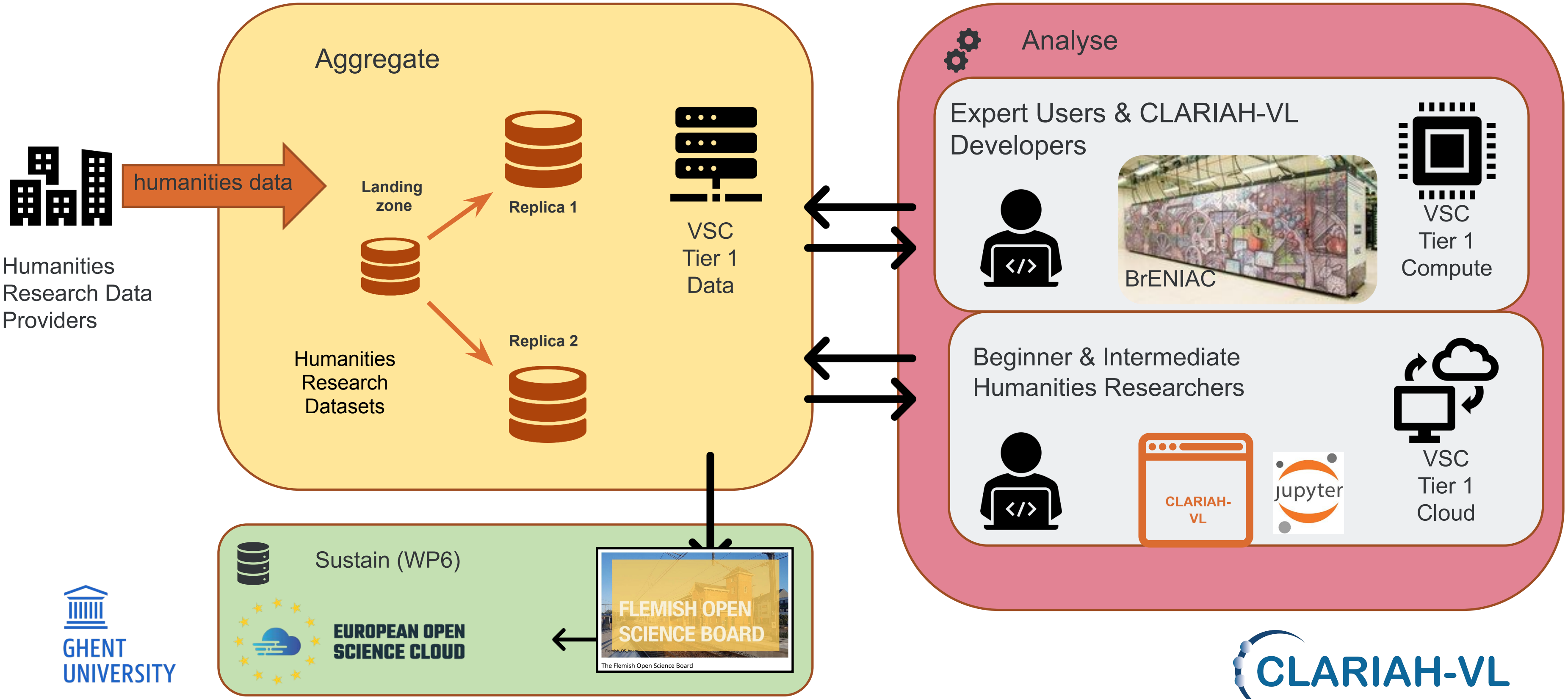
5 Provide for long-term availability — Expand ▾

https://marketplace.sshopencloud.eu/workflow/4qFarh

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC social sciences & humanities open cloud

Data-driven research in the arts and humanities


humanities data
Humanities Research Data Providers
Aggregate
Landing zone
Replica 1
Replica 2
Humanities Research Datasets
VSC Tier 1 Data
Analyse
Expert Users & CLARIAH-VL Developers
BrENIAC
VSC Tier 1 Compute
Beginner & Intermediate Humanities Researchers
CLARIAH-VL
jupyter
VSC Tier 1 Cloud
Sustain (WP6)
EUROPEAN OPEN SCIENCE CLOUD
FLEMISH OPEN SCIENCE BOARD
The Flemish Open Science Board

GHENT UNIVERSITY

CLARIAH-VL
Open Humanities Service Infrastructure

CLARIN and Libraries

9 May 2022 - 10 May 2022

Thanks for listening!

Sally Chambers

Sally.Chambers@kbr.be|Sally.Chambers@UGent.be |

KBR, Royal Library of Belgium and

Ghent Centre for Digital Humanities, Belgium