



ONB Labs
Symposium 2024

Towards sustainable workflows for newspapers as datasets: *an infrastructural perspective*

Sally Chambers

DARIAH-EU and The British Library

 **DARIAH-EU**
Digital Research Infrastructure
for the Arts and Humanities

Wiener Musik-Zeitung.

Herausgeber und Redacteur August Schmidt.

Nr. 3.

Donnerstag, den 7. Jänner

1841.

Eine Soirée.

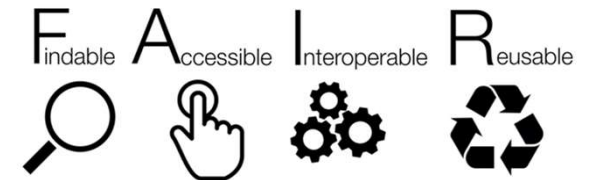
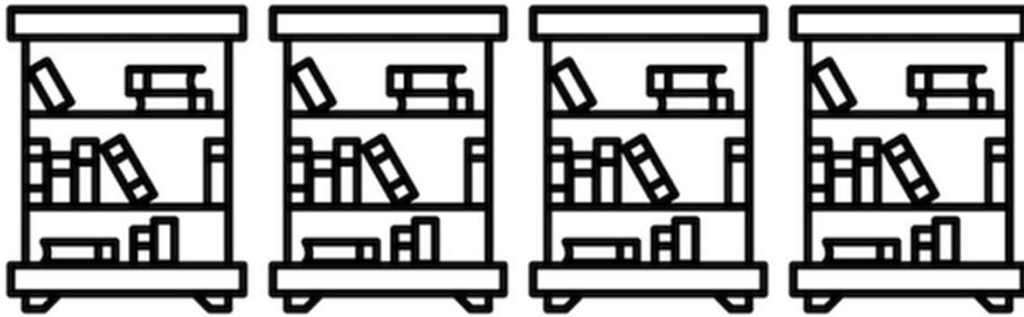
(Schluss)

Am 7. Jänner war eine wundervolle Sommernacht. Die Sterne blühten freundlich herab, und jeder Strahl schien mir zu sagen: Auch wenn ihr Menschen und nicht lebt, leuchtet die Nacht.

BRITISH
LIBRARY

Collections as Data: “Always Already Computational” and

“Part to Whole”



Padilla, Thomas, Allen, Laurie, Frost, Hannah, Potvin, Sarah, Russey Roke, Elizabeth, & Varner, Stewart. (2019). Final Report --- Always Already Computational: Collections as Data.

<http://doi.org/10.5281/zenodo.3152935>

Always Already Computational: Collections as Data final report and project deliverables:

<https://osf.io/mx6uk/wiki/home/>

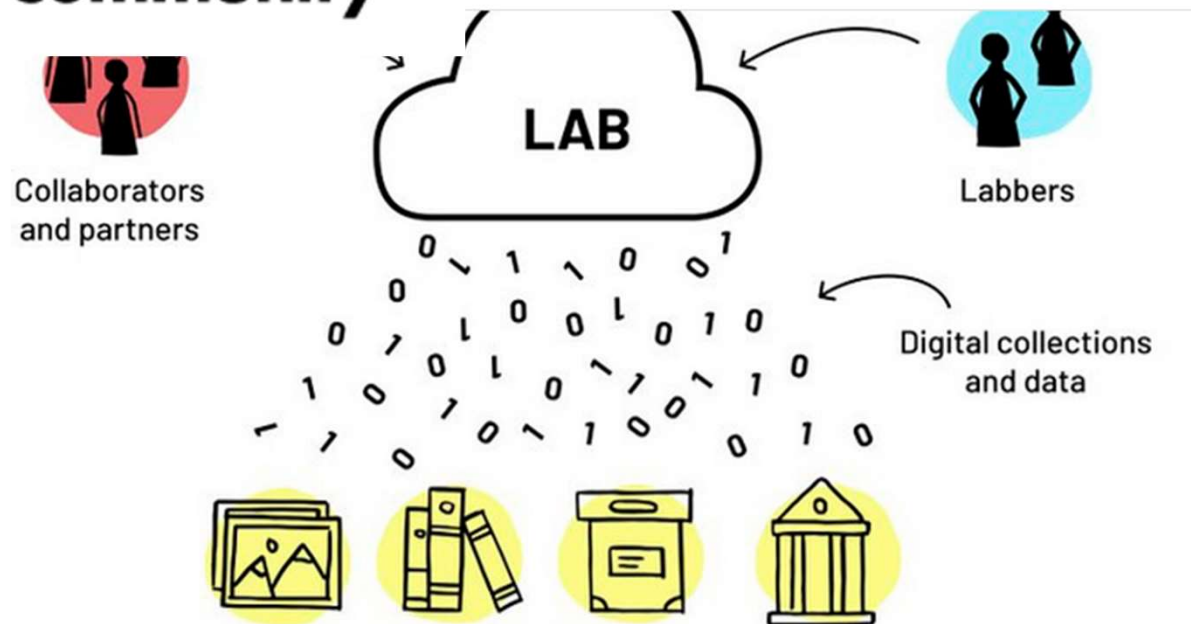
Padilla, T., Scates Kettler, H., Varner, S., & Shorish, Y. (2023). Vancouver Statement on Collections as Data. Zenodo.

<https://doi.org/10.5281/zenodo.8342171>

“to support responsible development and computational use of collections as data”

<https://collectionsasdata.github.io/>

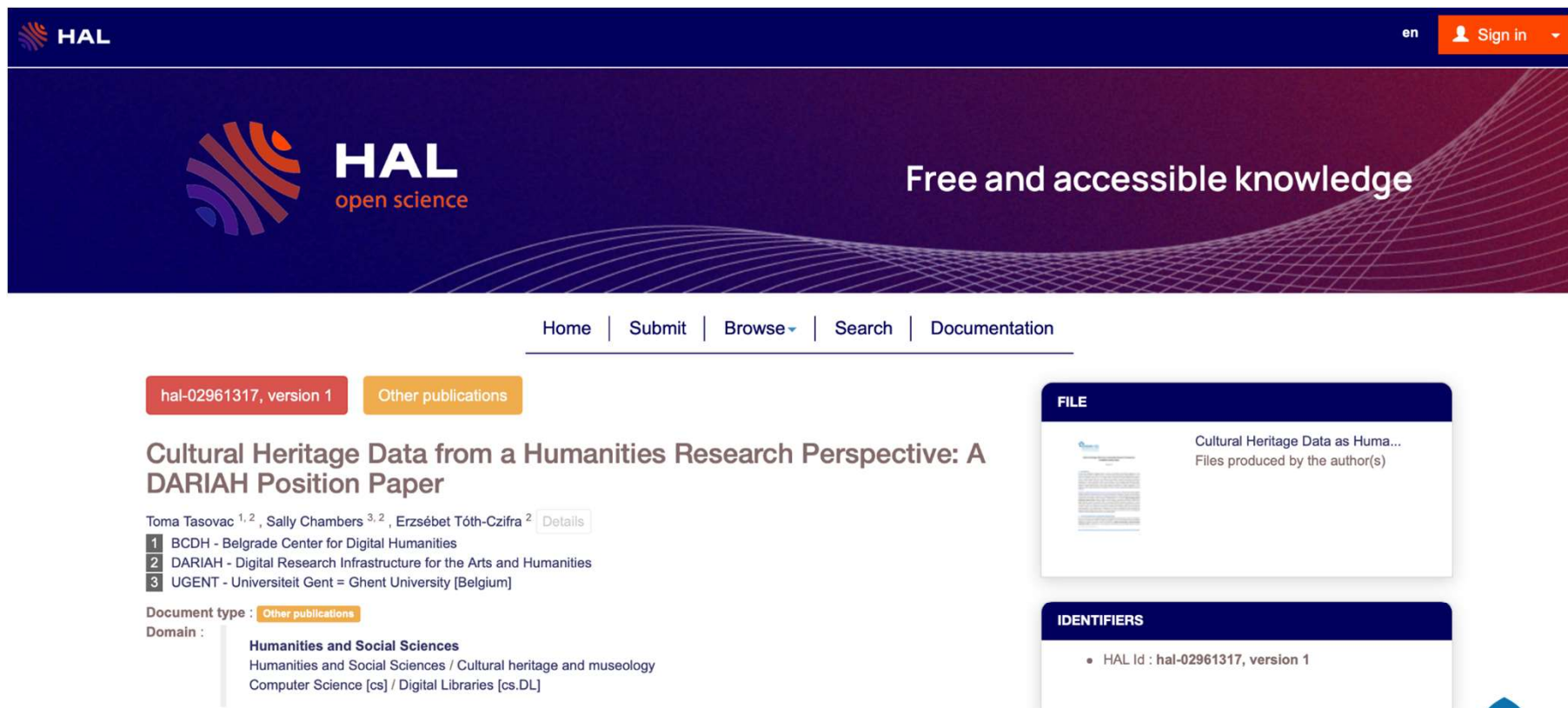
International GLAM Labs Community



Mahey, M., Al-Abdulla, A., Ames, S., Bray, P., Candela, G., Chambers, S., Derven, C., Dobрева-McPherson, M., Gasser, K., Karner, S., Kokegei, K., Laursen, D., Potter, A., Straube, A., Wagner, S-C. and Wilms, L., with forewords by: Al-Emadi, T. A., Broady-Preston, J., Landry, P. and Papaioannou, G. (2019) *Open a GLAM Lab. Digital Cultural Heritage Innovation Labs*, Book Sprint, Doha, Qatar, 23-27 September, 2019.

<https://glamlabs.io/books/>

Cultural Heritage Data as Humanities Research Data



The screenshot shows the HAL open science website interface. At the top, there is a dark blue header with the HAL logo on the left and a 'Sign in' button on the right. Below the header is a large banner with the HAL logo and the text 'Free and accessible knowledge'. A navigation bar contains links for Home, Submit, Browse, Search, and Documentation. The main content area displays the publication 'Cultural Heritage Data from a Humanities Research Perspective: A DARIAH Position Paper' by Toma Tasovac, Sally Chambers, and Erzsébet Tóth-Czifra. It includes a 'FILE' section with a thumbnail of the paper and an 'IDENTIFIERS' section showing the HAL ID: hal-02961317, version 1. The domain is listed as Humanities and Social Sciences, and the document type is 'Other publications'.

HAL open science

Free and accessible knowledge

Home | Submit | Browse | Search | Documentation

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-Czifra² [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]

FILE

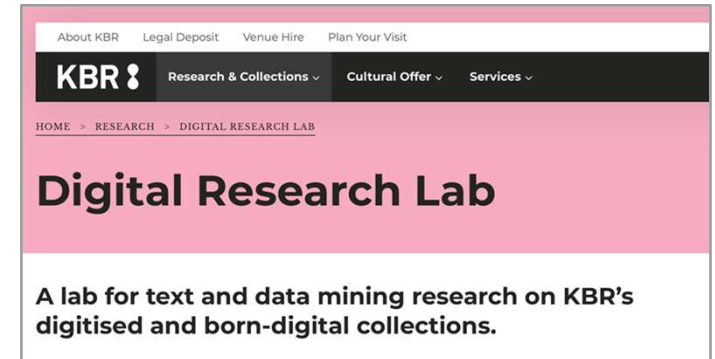
Cultural Heritage Data as Huma...
Files produced by the author(s)

IDENTIFIERS

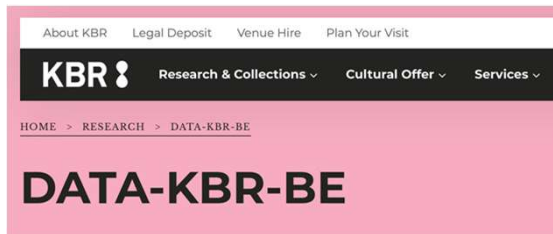
• HAL Id : hal-02961317, version 1

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>

Collections as Data & Labs @ KBR



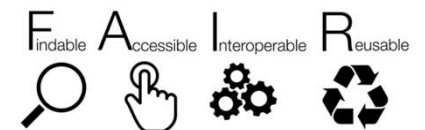
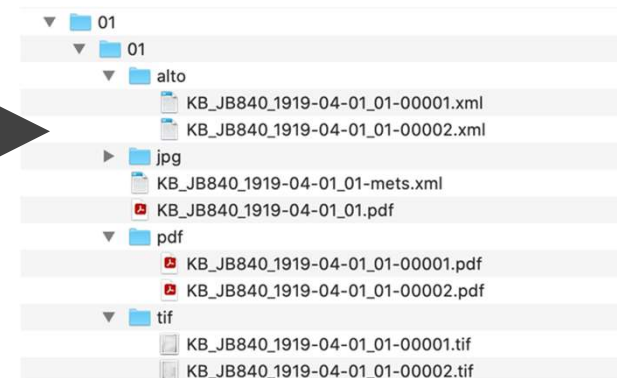
<https://www.kbr.be/en/projects/digital-research-lab/>



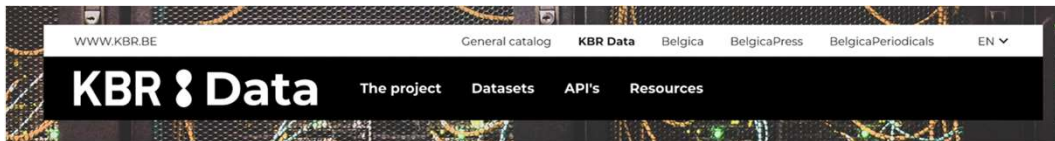
Facilitating data-level access to KBR's digitised and born-digital collections for digital humanities research

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

KBR Where time is treasured



Collections as Data & Humanities Research Data



Home > Datasets > Collective action in Ghent 1913

Collective actions Belgium

digitisation newspapers maps & plans history bla bla bla .jpeg .tiff .xml

Description

Events can broadly be defined as "things that happen". This challenge becomes increasingly complex when modelling events of the past. Such research aims to detect, identify, trace and model events of collective action (e.g. unionisation, strikes and demonstrations). For example, techniques such as Named Entity Recognition (NER) and Geographic Entity Recognition (GER) can be used to identify key actors and places in the BelgicaPress corpus. These entities can then be inter-linked with reports of strikes or demonstrations, mined from the historical newspapers.

The datasets "Collective actions during the general strikes of 1902 and 1913 in Ghent" were created to support the research for the master's thesis "Die kalme strijd van vree en rust. Een Ruimtelijke vergelijking tussen de Algemene werkstakingen van 1902 en 1913 in Gent" by Léon Castelein, Ghent University in 2023. The data exists of collective actions, their locations and additional information gathered from the newspapers "De Vooruit" and "Het Volk" from 14 April 1902 to 20 April 1902 and 14 April 1913 to 23 April 1913.

Downloads

<https://doi.org/10.34934/DVN/PQ83W7>

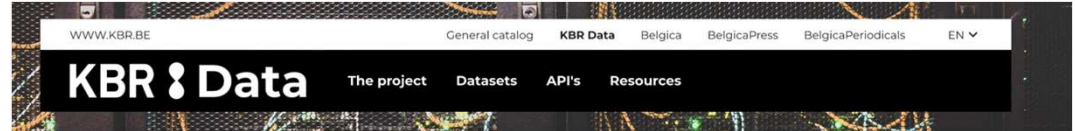
[Sample \(26Mb\)](#)

Licensing and citation

Bla bla bla (or only on SODHA so that there is no issue with synchronisation?)

Publications

Publication 1
Publication 2
Publication 3



Home > Datasets > Collective action in Ghent 1913

Collective actions during the general strikes of 1902 and 1913 in Ghent

digitisation newspapers maps & plans history bla bla bla .jpeg .tiff .xml

Description

Events can broadly be defined as "things that happen". This challenge becomes increasingly complex when modelling events of the past. Such research aims to detect, identify, trace and model events of collective action (e.g. unionisation, strikes and demonstrations). For example, techniques such as Named Entity Recognition (NER) and Geographic Entity Recognition (GER) can be used to identify key actors and places in the BelgicaPress corpus. These entities can then be inter-linked with reports of strikes or demonstrations, mined from the historical newspapers.

The datasets "Collective actions during the general strikes of 1902 and 1913 in Ghent" were created to support the research for the master's thesis "Die kalme strijd van vree en rust. Een Ruimtelijke vergelijking tussen de Algemene werkstakingen van 1902 en 1913 in Gent" by Léon Castelein, Ghent University in 2023. The data exists of collective actions, their locations and additional information gathered from the newspapers "De Vooruit" and "Het Volk" from 14 April 1902 to 20 April 1902 and 14 April 1913 to 23 April 1913.

Downloads

<https://doi.org/10.34934/DVN/PQ83W7>

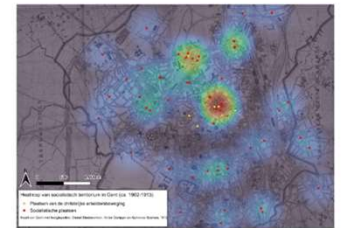
[Sample \(26Mb\)](#)

Licensing and citation

Bla bla bla (or only on SODHA so that there is no issue with synchronisation?)

Publications

Publication 1
Publication 2
Publication 3



KBR

DATA-KBR-BE: <https://www.kbr.be/en/projects/data-kbr-be/>



Collective actions during the general strikes of 1902 and 1913 in Ghent

digitisationnewspapersmaps & planshistorybla bla bla.jpeg.tif.xml

Description

Events can broadly be defined as “things that happen”. This challenge becomes increasingly complex when modelling events of the past. Such research aims to detect, identify, trace and model events of collective action (e.g. unionisation, strikes and demonstrations). For example, techniques such as Named Entity Recognition (NER) and Geographic Entity Recognition (GER) can be used to identify key actors and places in the BelgicaPress corpus. These entities can then be inter-linked with reports of strikes or demonstrations, mined from the historical newspapers.

The datasets “Collective actions during the general strikes of 1902 and 1913 in Ghent” support the research for the master's thesis “Die kalme strijd van vree en rust.” Een R tussen de Algemene werkstakingen van 1902 en 1913 in Gent” by Léon Castelein, Ghe The data exists of collective actions, their locations and additional information gathe newspapers “De Vooruit” and “Het Volk” from 14 April 1902 to 20 April 1902 and 14 A 1913.

Downloads

<https://doi.org/10.34934/DVN/PQQ3W7>
[Sample \(26Mb\)](#)

Licensing and citation

Bla bla bla (or only on SODHA so that there is no issue with synchronisation?)

Publications

- Publication 1
- Publication 2
- Publication 3



SODHAAdd DataSearchSODHA GuideSign UpLog In

SodhaSocial Sciences and Digital Humanities Archive – SODHAHosted by the State Archives of Belgium

Social Sciences and Digital Humanities Archive – SODHA >

Collectieve acties tijdens de algemene werkstakingen van 1902 en 1913 in Gent

Version 2.0

Castelein, Léon, 2023, “Collectieve acties tijdens de algemene werkstakingen van 1902 en 1913 in Gent”, <https://doi.org/10.34934/DVN/PQQ3W7>, Social Sciences and Digital Humanities Archive – SODHA, V2, UNF:6:F+3NuyzbAWbmFB026pwhzQ== [fileUNF]

Cite Dataset

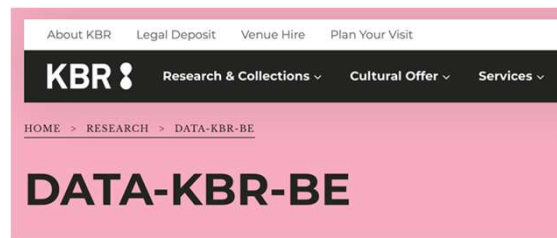
Learn about Data Citation Standards.

Description

Datasets opgesteld ter ondersteuning van mijn masterproof. Hierin maak ik een ruimtelijke vergelijking tussen de algemene werkstakingen van 1902 en 1913 in Gent. De data is afkomstig uit de Vooruit en Het Volk tussen 10 en 22 april 1902 en tussen 14 en 24 april 1913. Drie datasets zijn opgesteld: de verzameling van collectieve acties in de kranten van 1902, de verzameling van collectieve acties in de kranten van 1913 en de verzameling van plaatsen waar acties plaatsvonden, aangevuld met belangrijke locaties in de handen van Gentse socialistische en christelijke arbeidersbewegingen (aangeduid als socialistisch of antisocialistisch territorium). De datasets met collectieve acties bevatten de velden activetyp, deelnemende groep, ideologie, plaats, startuur, duur, startdatum, einddatum, toelichting, opmerkingen en bronnen. De dataset met plaatsen bevat de plaatsnaam, alternatieve namen, tot welk territorium het eventueel behoort, het locatietype, start- en einddatum, de coördinaten in GeoJSON formaat, opmerkingen en bronnen. Meer toelichting over de gegevenstypes is in de bijhorende documentatie te vinden. (2023-08-11)

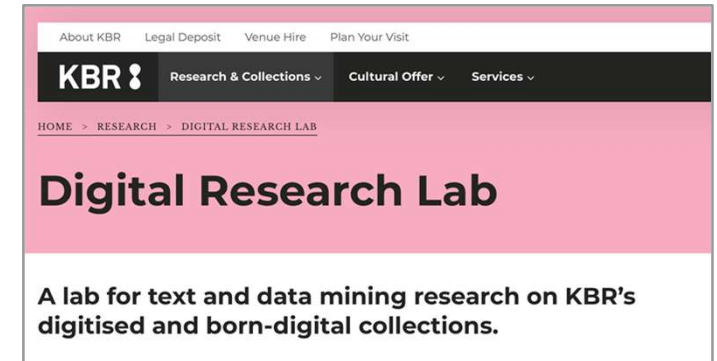
1 to 4 of 4 Files		Download
	Collectieve acties in Gent tijdens de algemene staking van 1902 uit de Vooruit en Het Volk van 10 tot 22 april 1902.tab Tabular Data - 50.7 KB Published Aug 30, 2023 1 Download	
	13 Variables, 180 Observations UNF:6:qWUp...LFRev... CSV met de collectieve acties in Gent tijdens de algemene staking van 1902 uit de Vooruit en Het Volk van 10 tot 22 april 1902. HistoryGhentCollective actionData	
	Collectieve acties in Gent tijdens de algemene staking van 1913 uit de Vooruit en Het Volk van 14 tot 24 april 1913.tab Tabular Data - 88.3 KB Published Aug 30, 2023 1 Download	
	13 Variables, 354 Observations UNF:6:u7bT...LFRev... CSV met de collectieve acties in Gent tijdens de algemene staking van 1913 uit de Vooruit en Het Volk van 14 tot 24 april 1913. HistoryGhentCollective actionData	
	Documentatie van de datasets collectieve acties tijdens de algemene stakingen van 1902 en 1913 in Gent.pdf Adobe PDF - 352.2 KB Published Aug 30, 2023 1 Download MD5: 122...Fax Documentatie bij de datasets HistoryGhentCollective actionDocumentation	
	Plaatsen van collectieve acties, socialistisch en katholiek territorium tijdens de algemene werkstakingen van 1902 en 1913.tab Tabular Data - 121.3 KB Published Aug 30, 2023 1 Download 12 Variables, 248 Observations UNF:6:AvVM...LFRev... CSV met alle plaatsen en routes waar de acties plaatsvonden. HistoryGhentCollective actionData	

Digitised Historical Newspapers as Data

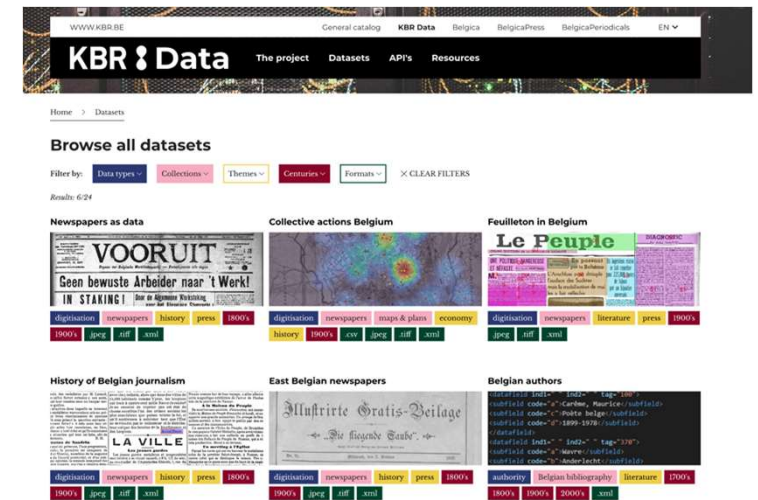


Facilitating data-level access to KBR's digitised and born-digital collections for digital humanities research

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



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



The British Library Research Repository

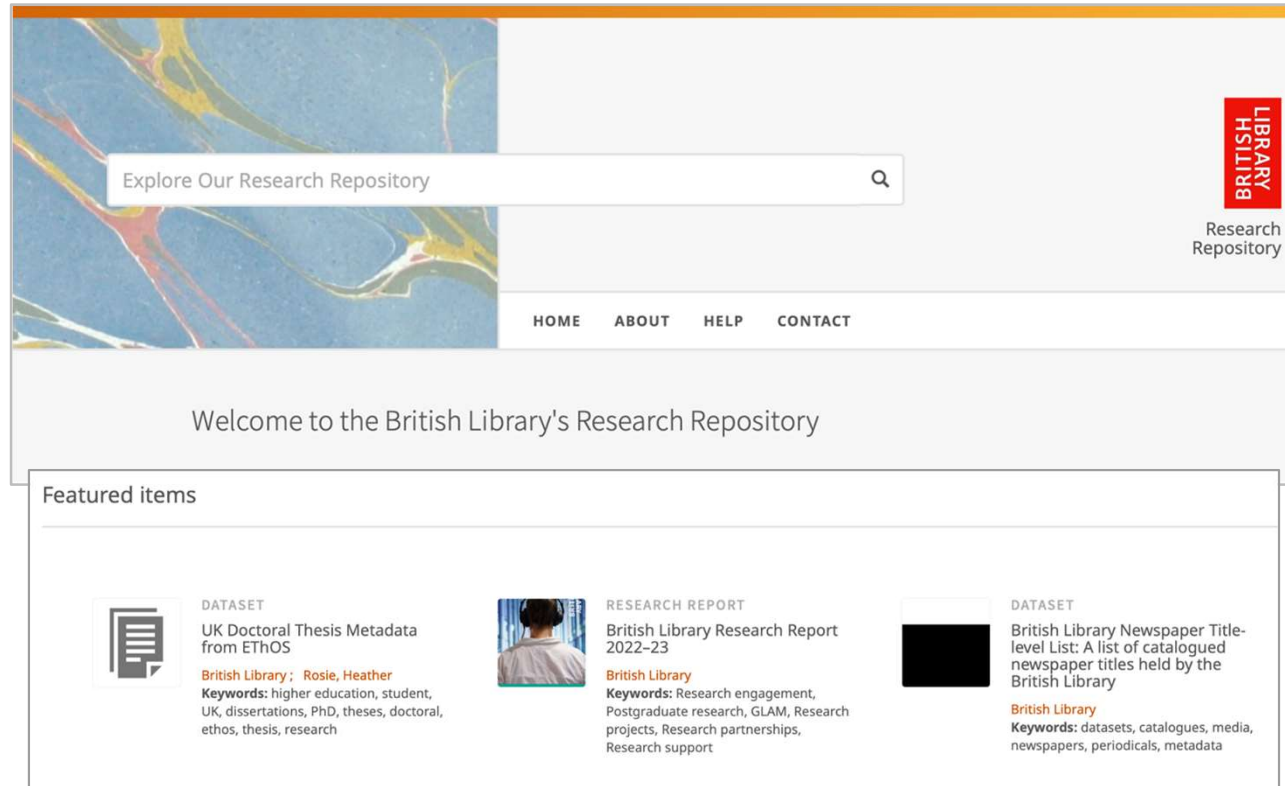
The British Library Research Repository is an **open access repository** for the research produced by staff and research associates of the British Library.

Content includes: journal articles, conference papers, books and book chapters, reports, **datasets**, images, exhibition texts and blog posts.

Resource Type

Dataset

270



<https://bl.iro.bl.uk/>

The British Library Research Repository

British Library News Datasets

The Library holds a newspaper collection of over 34,000 titles dating from 1619 to the present day, and growing collections of radio, television and web news. The datasets here relate to these collections, either in digital form or hardcopy, and we hope will be useful to explore further or to analyse the collection through its metadata. More datasets will be added over time. Contact us at newspapers@bl.uk or on Twitter @BL_newsroom if you have questions, or to let us know how you used our data.

More information on our news collections can be found here: <https://www.bl.uk/subjects/news-media>

Digitised newspapers from the British Library are available on <https://www.britishnewspaperarchive.co.uk>, which is free to use on the British Library premises.

<https://bl.iro.bl.uk/collections/353c908d-b495-4413-b047-87236d2573e3>

Newspapers

User Collection



Collection Details

Total items 57

Resource type Collection Dataset

<https://bl.iro.bl.uk/collections/9a6a4cdd-2bfe-47bb-8c14-c0a5d100501f>

DATASET



British Library Newspaper Title-level List:
A list of catalogued newspaper titles held
by the British Library

<https://bl.iro.bl.uk/>



DATASET

British and Irish Newspapers



DATASET

The Newspaper Press Directory (1881-
1920)

BRITISH LIBRARY

The British Library Research Repository



DATASET

The Newspaper Press Directory (1846-1920) - enriched and structured version

Mitchell's Newspaper Press Directories contained an almost complete list of newspapers published in England, Wales, Scotland and Ireland. It was published regularly from 1846 onwards and provided a detailed description of the newspaper landscape over time. This version contains a structured, tabular representation of the directories (as CSV or Excel...

2023

C. Mitchell and Co. ; British Library
press directories and newspapers



DATASET

Language of Mechanisation: annotated historical newspaper articles

Datasets created through crowdsourcing tasks created on the Zooniverse crowdsourcing platform by the Living with Machines 'language of mechanisation' project team. Building on earlier work classifying machines by function, we asked volunteers on Zooniverse 'how did the word x change over time and place?' and presented them with options for...

2023

British Library ; Ridge, Mia ; Pedrazzini, Nilo ; McGillivray, Barbara
crowdsourcing, 19th century British English, annotation, historical newspapers, mechanisation, data visualisation, historical semantics, and transport history



Heritage Made Digital
Newspapers
HeritageMadeDigitalNewspapers

Follow

Overview Repositories 7 Projects Packages Stars 6

Popular repositories

[text-mine-title-list](#)

Public

What words in newspaper titles were most unique to each century? To each country?

[mapping-titles](#)

Public

Using R to count and map newspaper titles, by city

[newspaper_analyser](#)

Public

[basic-stats](#)

Public

A demonstration of how to sort, filter and count the dataset using R to answer some basic statistical questions

[title_constellation](#)

Public

[COVID-19](#)

Public

Forked from CSSEGISandData/COVID-19

Novel Coronavirus (COVID-19) Cases, provided by JHU CSSE

<https://github.com/HeritageMadeDigitalNewspapers>

<https://bl.iro.bl.uk/>

Social Sciences and Humanities Open Marketplace



Tools & services

Training materials

Publications

Datasets

Workflows

Browse

Contribute

Report an issue

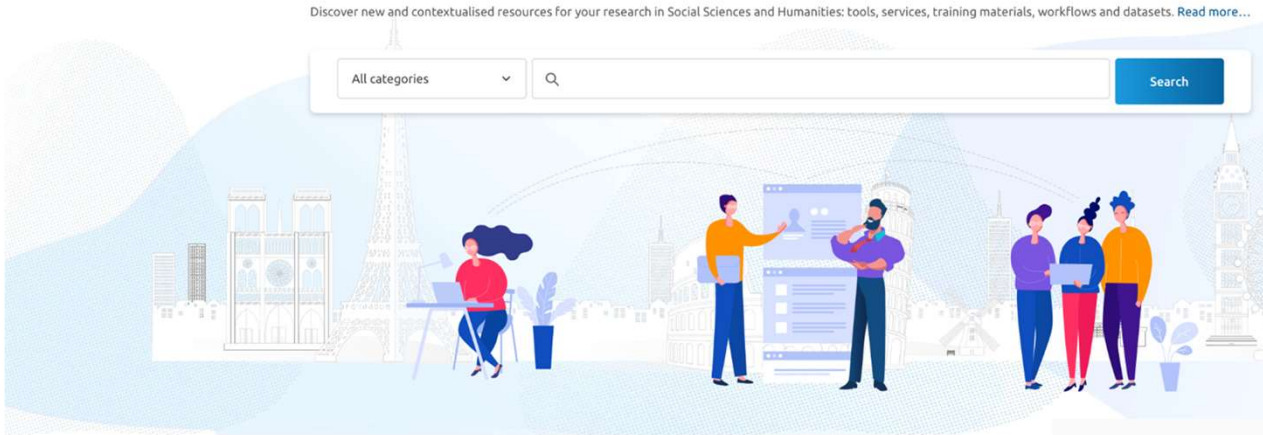
3 Guiding Principles:

- Contextualisation
- Curation
- Community

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...](#)

All categories



Barbot, L., Dolinar, M., Gray, E. J., Grisot, C., Illmayer, K., Kurzmeier, M., & McGillivray, B. (2024). Contextualizing Research Tools & Services Through Workflows in the SSH Open Marketplace. *Journal of Open Humanities Data*, 10 (1), 22. <https://doi.org/10.5334/johd.192>



<https://marketplace.sshopencloud.eu>

Cultural Heritage Data as Humanities Research Data



The screenshot shows the SSH Open Marketplace website. The header includes the SSH Open Marketplace logo, navigation links (Tools & services, Training materials, Publications, Datasets, Workflows, Browse, Contribute, About), and links for 'Report an issue' and 'Sign in'. A search bar is present. The main content area displays a workflow titled 'A workflow to publish Collections as Data: the case of Cultural Heritage data spaces'. The workflow description states: 'Cultural Heritage institutions have been making their digital collections available for the public for several decades. Advances in technology such as Artificial Intelligence and Machine Learning have provided a new context in which digital collections can be analysed using computational methods. Initiatives such as [Collections as data](#) and the [FAIR data principles](#), have emerged to provide best practices and guidelines for publishing [digital collections suitable for computational use](#). These initiatives are complemented with the [CARE principles](#) to strengthen ethical considerations in data governance and reuse. In parallel, experimental [Labs](#) have been implemented in Galleries, Libraries, Archives and Museums (GLAM) in order to reuse the digital collections. [Data spaces](#) have emerged as an innovative way to publish and reuse digital collections. Based on [previous work undertaken in the context of the GLAM Labs Community](#), this workflow provides a set of steps to publish Collections as data. It aims to guide and encourage cultural heritage institutions, step by step, in publishing their collections, so that they are suitable for computational use. It is important

A workflow to publish Collections as Data: the case of Cultural Heritage data spaces

Details

ACCESS

License [Creative Commons Attribution 4.0 International](#)

CATEGORISATION

Activity [Description](#) [Extracting](#) [Analyzing](#) [Discovering](#) [Gathering](#) [Disseminating](#) [Sharing](#)

Keyword [digital collections](#) [Data](#) [computational methods](#) [Collections as data](#)



Candela, G., Chambers, S. and Irollo, A. (2023) A workflow to publish Collections as Data: the case of Cultural Heritage data spaces.
<https://marketplace.sshopencloud.eu/workflow/I3JvP6>

A Collections as Data Workflow - 10 Steps

- 1 | Provide a clear license allowing reuse of the dataset without restrictions [Expand ▼](#)
- 2 | Provide a suggested citation for the dataset so reusers are aware of how to cite it [Expand ▼](#)
- 3 | Include documentation about the dataset [Expand ▼](#)
- 4 | Use a public platform to make available the dataset for the public
- 5 | Share examples of use to demonstrate how the dataset can be reused



- 6 | Think about a structure for the dataset for a better understanding of how to reuse the content [Expand ▼](#)
- 7 | Include machine-readable metadata about the content provided in the dataset [Expand ▼](#)
- 8 | Use an existing collaborative-edition platform to include the information about the dataset [Expand ▼](#)
- 9 | Provide the dataset by means of an existing API [Expand ▼](#)
- 10 | Create a website to present and describe the dataset to encourage its reuse [Expand ▼](#)

<https://marketplace.sshopencloud.eu/workflow/I3JvP6>

Introduction to Cultural Heritage Data

EN

This course provides the essential knowledge and skills to understand and efficiently use Cultural Heritage data. Guided by Prof. Lorena, a persona created for the course, participants explore the significance of CH data, its types, and formats. They learn to identify sources for data acquisition and apply techniques to enhance data quality. The course also covers methods for organizing CH data, introduces key metadata standards, and examines current trends and technologies in the field.



Read more →



Introduction to Europeana APIs

EN

This course provides a comprehensive understanding of Europeana as a digital platform through a walkthrough of the Application Programming Interfaces (APIs) it offers. It provides the knowledge and skills to understand the purpose they serve and the functionality they have, to exploit them by formulating efficient queries for cultural heritage information retrieval. Building on use cases, it delves into the APIs required to achieve research goals, exploring their features and providing familiarisation with supported data formats.



Read more →



Introduction to Collections as Data

EN

The goal of this course is to introduce the Collections as Data principles in the cultural heritage sector to make available a digital collection suitable for computational use. Students will have a fundamental understanding of the complexities of Collections as Data as well as an appreciation of the diversity of the content provided by cultural heritage institutions. This course will be useful for small and medium-sized institutions willing to make available their digital collections suitable for computational use.



Read more →



<https://campus.dariah.eu/>



CENL Dialogue Forum: *National Libraries as Data Infrastructures*

CENL

To facilitate structural and strategic collaboration between Europe's National Libraries and Research Infrastructures

- Develop 2-3 'Collections as Data' pilots using the [Collections as Data workflow](#)
- Increase the discoverability of 'Collections as Data', e.g. harvest the metadata of cultural heritage [datasets](#) e.g. from existing repositories
- Share these 'collections as data' on research platforms such as the [SSH Open Marketplace](#)

 **DARIAH-EU**
Digital Research Infrastructure
for the Arts and Humanities


CLARIN

Social Sciences and Humanities Open Marketplace

[Report an issue](#)[Sign in](#)[Tools & services](#)[Training materials](#)[Publications](#)[Datasets](#)[Workflows](#)[Browse](#)[Contribute](#)[About](#)

x

[Search](#)[Home](#) / [Search](#)

Search results (140)

Refine your search

[Clear filters](#)

CATEGORIES

- | | | | |
|--------------------------|--|--------------------|-----|
| <input type="checkbox"/> | | Tools & services | 21 |
| <input type="checkbox"/> | | Training materials | 3 |
| <input type="checkbox"/> | | Publications | 1 |
| <input type="checkbox"/> | | Datasets | 115 |



Jupyter notebooks for Europeana newspaper text resource processing with CLARIN NLP tools



Keywords [research notebook](#) [newspapers](#) [NLP](#) [python](#) [language resources](#)

Binder These notebooks have been designed to help getting started with the processing historical text resources (from Europeana Newspapers) with natural language processing (NLP) tools (from CLARIN) using Jupyter notebooks. The easiest way to get started is to click the Launch binder badge above. This will guide you through t...



Europeana Newspapers

Europeana Newspapers has 10 repositories available. Follow their code on GitHub.



Europeana historical newspapers: Austria

Keywords [Newspaper corpora](#)

This corpus contains 147,515 issues of 77 newspapers published in Austria between 1683 and 1930.

<https://marketplace.sshopencloud.eu/search?q=newspapers>

ECCCH

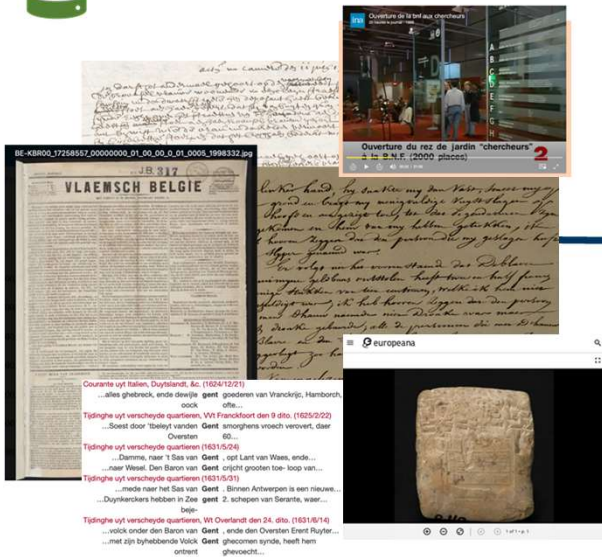
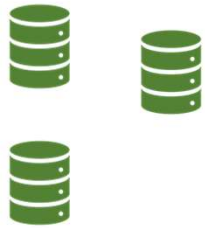
European Collaborative Cloud for Cultural Heritage

**Preserving the Past, Shaping the Future: your Gateway to a
Collaborative and Innovative European Cultural Heritage Community**

<https://www.echoes-eccch.eu/>

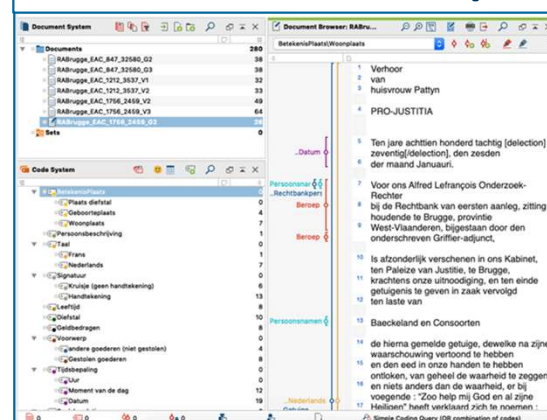


Towards a Virtual Transcription Laboratory



Virtual Transcription Laboratory

Humans-in-the-loop



Automatic Text Recognition
(HTR, OCR)



Digital (Text) Analysis Workflows

- Named Entity Extraction
- Social Network Analysis
- Sentiment & Emotion Mining



Dariah.lab



ONB Labs
Symposium 2024

Thanks for listening!

Sally Chambers | sally.chambers@bl.uk



DARIAH-EU

Digital Research Infrastructure
for the Arts and Humanities

Wiener Musik-Zeitung.

Herausgeber und Redacteur August Schmidt.

Nr. 3.

Donnerstag, den 7. Jänner

1841.

Eine Soirée.

(Schluss)

Am 7. Jänner war eine wundervolle Sommernacht. Die Sterne blühten freundlich herab, und jeder Strahl schien mir zu sagen: Auch wenn ihr Menschen und nicht lebt, leuchtet mir zu.

LIBRARY
HSLIRB