Newspapers in Living with Machines

Congrès Médias 19 - Numapresse : Presses anciennes et modernes à l'ère du numérique La BnF, 3 juin. 2022

Mia Ridge, British Library @mia_out For the *Living with Machines* team @LivingWMachines

Thanks to Dr Hannah Williams (@DrHanWill) for translation assistance

The Alan Turing Institute













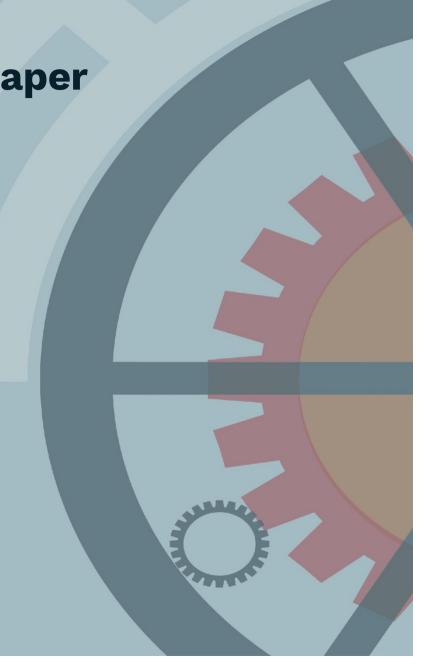
About the British Library's newspaper collection

Over **34,000 titles** from the UK and overseas, **60 million individual issues**

Newspapers from 1603; collection more comprehensive from 1820s

Since 1869 - British and Irish newspaper publishers required to send one copy of each issue to the Library for 'legal deposit'

Currently collecting c1200 UK and Irish newspapers, c100 from overseas daily or weekly



À propos de la collection de journaux au British Library

Plus de **34,000 titres** du Royaume-Uni et d'outre-mer; **60 millions de numéros** individuels

Les journaux datent de 1603; la collection est plus complète à partir des années 1820

Depuis 1869 - éditeurs de journaux britanniques et irlandais sont tenus d'envoyer un exemplaire de chaque numéro au BL pour le "legal deposit"

Actuellement - le BL collectionne environ 1200 journaux britannique et irlandais, et environ 100 quotidiens ou hebdomadaires étrangers

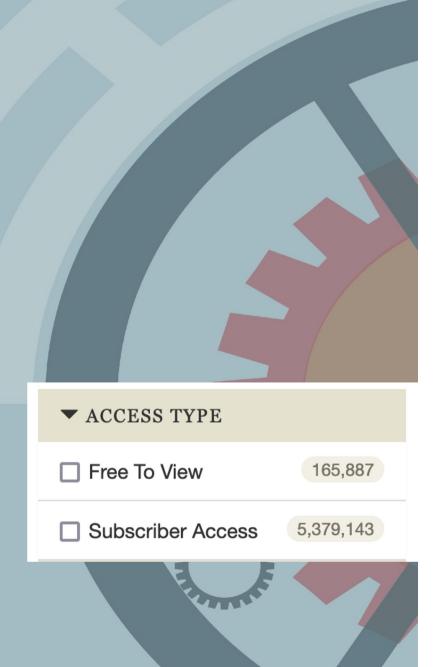
The British Newspaper Archive

Digitisation partnership between the BL and FindMyPast

Launched in 2011; renewed in 2021. 50 million+ pages.

Initially BNA aimed for wide geographic coverage; now topical coverage

From 2021, 1 million free-to-view pages; more to come



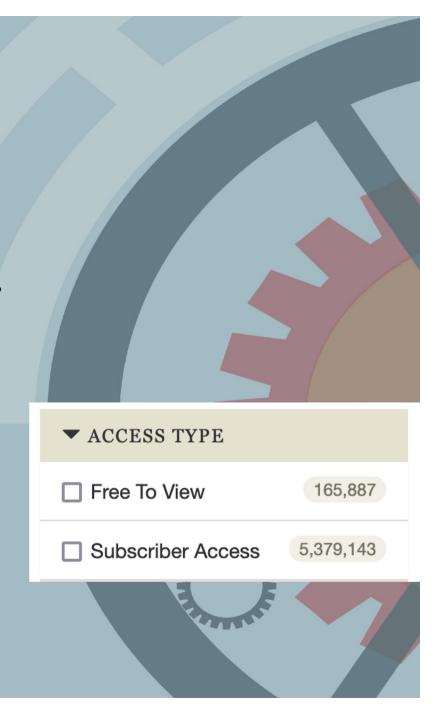
Le British Newspaper Archive

Collaboration entre BL et FindMyPast pour la numérisation.

Lancé en 2011; renouvelé en 2021. Plus de 50 millions de pages.

Objectif initial: couverture géographique large. Maintenant: couverture thématique.

À partir de 2021, 1 million de pages en libre accès; d'autres à venir

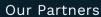


What can you do with 50 million pages?

The Living with Machines project
How did the wider introduction of
machines change ordinary lives in 19th
century Britain?

LwM looks at mechanisation from 1780 - 1920 via digitised newspapers, books, maps and other historical sources.

A 5 year collaboration between curators, data scientists, linguists, historians and other specialists - and the public



















UK Research and Innovation

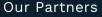
Que peut-on faire avec 50 millions de pages?

Le projet *Living with Machines*

Comment l'introduction des machines a-telle changé la vie quotidienne dans la Grande Bretagne du XIXe siècle ?

LwM porte sur la mécanisation de 1780 à 1920, à partir des journaux numérisés, livres, plans, et autres sources historiques.

Une collaboration (5 ans) entre conservateurs, data scientists, linguistes, historiens, and autres spécialistes - et le public.



















UK Research

Principal and Co-Investigators



Ruth Ahnert

(QMUL)

Read more



David Beavan

(Turing)

Read more



Emma Griffin

(UEA)

Read more



Timothy Hobson

(Turing), Research Software Engineer

Read more



Jon Lawrence

(Exeter)

Read more



Maja Maricevic

(British Library)

Read more



Barbara McGillivray

(Turing / Cambridge)



Mia Ridge

(British Library)



Alan Wilson

(Turing)

Project team



Claire Austin

Rights Assurance

Read more



Kaspar Beelen

Digital Humanities Research Associate

Read more



Mariona Coll Ardanuy

Computational Linguistics Research Associate

Read more



Katie McDonough

History Research Associate

Read more



Federico Nanni

(Turing), Research Data Scientist

Read more



Nilo Pedrazzini

Corpus-Based Digital Humanities Research Assistant

Read more



André Piza

Research Project Manager

Read more



Joshua Rhodes

Research Associate

Read more



Giorgia Tolfo

Data And Content Manager

Read more



Daniel Van Strien



Kalle Westerling



Daniel Wilson

Anciens collaborateurs

- GiovanniColavizza
- Joel Dearden
- AdamFarquhar
- Rosa Filgueira
- Christina Last
- Sarah Gibson
- JamesHetherington
- KasraHosseini
- Michael Jackson
- Yann Ryan
- Olivia Vane

Working with digitised newspapers

From FindMyPast we get:

- Images of digitised pages
- Metadata (various formats)
- Automatically transcribed text (OCR) in various formats

Internal infrastructure for analysis:

- Azure storage and compute
- Metadata database
- 'Defoe' system for full text search
- ALTO -> plain text processing

Different researchers need different formats

- Crowdsourcing images, text, metadata
- Text analysis OCR, plain text, metadata



Travailler avec les journaux numérisés

De FindMyPast, on obtient:

- Images de pages numérisées
- Metadata (divers formats)
- Texte transcrit automatiquement (OCR) idans différents formats

Infrastructure interne pour l'analyse :

- Azure storage and compute
- Metadata database
- 'Defoe' system pour la recherche en texte intégral
- ALTO -> plain text processing

Des chercheurs différents ont besoin de formats différents

- Crowdsourcing images, texte, metadata
- Analyse de text OCR, plain text, metadata

Challenges in working with digitised newspapers

Scale

- Millions of pages: hard to process, store
- 60 billion+ word tokens

Copyright

Unknown status after 'safe date' c1882

Licensing

 Project access provided by FindMyPast; not widely available

Security requirements

Access to full dataset is complex

Les challenges pour travailler avec les journaux numérisés

Échelle

- Des millions de pages: difficiles à traiter et à stocker
- Plus de 60 milliards de word tokens

Copyright

 Statut inconnu après la 'safe date' environ 1882

Licences

 Accès au projet fourni par FindMyPast; non disponible à grande échelle

Exigences de sécurité

Accès au dataset est complexe

'Environmental scan'

Kaspar Beelen

Situating the digitised press in the newspapers 'landscape'

 How representative is the digital 'sample' of the newspaper population? Where are imbalances or gaps?

Automatic metadata generation

 Extract metadata about newspapers from contextual resources (press directories)

CHATH

Is a parliamentary borough, a Medway. It is celebrated military depots, marine bar ment. The dock yard has been

THATHAM NEWS.

CIRCULATES in Chatham, Gillingham, and through the

ADVOCATES local interest district, and several column news, and miscellanies.

PROPRIETORS—Joseph (Advt., p. 164.)

Scan de l'environnement

Kaspar Beelen

Situer la presse numérisée dans le contexte des journaux

 Quelle est la représentivité de l'échantillon numérique de journaux en général ? Où se trouve les déséquilibres ou les lacunes ?

Génération automatique de metadata

 Extraire le metadata sur les journaux à partir de ressources contextuelles (répertoires de presse)

CHATH

Is a parliamentary borough, as Medway. It is celebrated military depots, marine bar ment. Thedockyard has been

THATHAM NEWS.

CIRCULATES in Chatham, Gillingham, and through the

Advocates local interest district, and several column news, and miscellanies.

PROPRIETORS—Joseph (Advt., p. 164.)

Newspapers and public participation

Proven public interest in digitised historical newspaper articles

Build on BL crowdsourcing as productive, public engagement

Designed crowdsourced tasks to classify and annotate articles

- Build datasets for research and analysis; exhibition
- See project themes and explore practical applications of AI / machine learning

Mia Ridge, Barbara McGillivray, Giorgia Tolfo, Kalle Westerling, Nilo Pedrazzini with contributions from Olivia Vane, Christina Last, Federico Nanni, Sarah Gibson, Kasra Hosseini

there have been four teen casualties, arising from year persons accidentally getting into the water. Among them were three perambulator cases, in which boys and girls, having been intrusted with these vehicles containing little children securely strapped in, have run races on the banks of the lake, and in the final grand push to be first, the machine has been capsized in the water, turning bottom upwards, but in each case the children were providentially rescued. Two of the total of 14 were drowned through the swamping of a sailing boat in a sudden squall. Not a single death from suicide has occurred during the year—a very

Journaux et participation du public

Intérêt avéré du public pour les articles de journaux historiques numérisés.

S'appuyer sur le crowdsourcing au BL pour l'engagement public productif.

Nouvelles activités de crowdsourcing pour classer et annoter les articles.

- Construire datasets pour la recherche et l'analyse; exposition
- Découvre les thèmes des projets et les applications pratiques

Mia Ridge, Barbara McGillivray, Giorgia Tolfo, Kalle Westerling, Nilo Pedrazzini avec les contributions de Olivia Vane, Christina Last, Federico Nanni, Sarah Gibson, Kasra Hosseini

there have been four teen casualties, arising from year persons accidentally getting into the water. Among them were three perambulator cases, in which boys and girls, having been intrusted with these vehicles containing little children securely strapped in, have run races on the banks of the lake, and in the final grand push to be first, the machine has been capsized in the water, turning bottom upwards, but in each case the children were providentially rescued. Two of the total of 14 were drowned through the swamping of a sailing boat in a sudden squall. Not a single death from suicide has occurred during the year—a very

New question from 'accidents' task: what was a machine in the 19th century?

EXCITING SCENE.—A few days since, as a bricklayer + and his labourer were at work on the chimney of a new foundry in course of erection at Heywood, near c Manchester, the triangle on the top of the scaffolding gave way and fell over the side. A rope connected with the machine dragged the labourer off by the leg, but as the other end was fast to a support, he hung in mid air, head downwards, a few yards from the top, and about 50 feet from the ground. His leg was broken by the jerk. As no assistance could be rendered from below, his comrade contrived to hook him by the braces, but as he was drawing him up they broke, and he again fell the same distance. After about half an hour's suspension, he was rescued by his comrade throwing down the brickwork until he reached him. He was removed to Manchester Infirmary.

PAIRING and FITTING in a machine room. Also good STOOK

TASK

TUTORIAL

Classify the machine

Look for the outlined 'machine' phrase. Does the image contain readable and relevant text for you to classify a physical or metaphorical machine? Choose the best option below.

No - the image is too hard to read

No - the image doesn't contain enough information to classify the type or purpose of the 'machine', or it is not about a physical or metaphorical machine

No - the highlighted text is a reference to a place such as a workshop or office

Yes - the image includes enough information to describe or classify the machine

NEED SOME HELP WITH THIS TASK?

Done

*

FIELD GUIDE

Nouvelle question de l'activité 'accidents' : qu'est-ce qu'une machine au XIXe siècle?

EXCITING SCENE.—A few days since, as a bricklayer + and his labourer were at work on the chimney of a new foundry in course of erection at Heywood, near c Manchester, the triangle on the top of the scaffolding gave way and fell over the side. A rope connected with the machine dragged the labourer off by the leg, but as the other end was fast to a support, he hung in mid air, head downwards, a few yards from the top, and about 50 feet from the ground. His leg was broken by the jerk. As no assistance could be rendered from below, his comrade contrived to hook him by the braces, but as he was drawing him up they broke, and he again fell the same distance. After about half an hour's suspension, he was rescued by his comrade throwing down the brickwork until he reached him. He was removed to Manchester Infirmary.

PAIRING and FITTING in a machine room. Also good STOOK

TASK

TUTORIAL

Classify the machine

Look for the outlined 'machine' phrase. Does the image contain readable and relevant text for you to classify a physical or metaphorical machine? Choose the best option below.

No - the image is too hard to read

No - the image doesn't contain enough information to classify the type or purpose of the 'machine', or it is not about a physical or metaphorical machine

No - the highlighted text is a reference to a place such as a workshop or office

Yes - the image includes enough information to describe or classify the machine

NEED SOME HELP WITH THIS TASK?

Done

*

FIELD GUIDE

OED machine entry analysis: 'Classify it!'

machine, n.

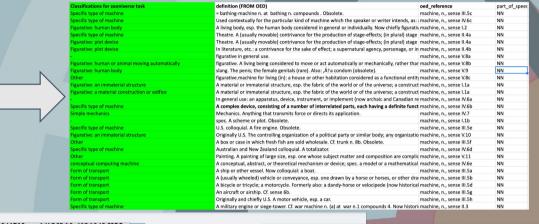
Pronunciation: Brit. // məˈʃiːn/, U.S. // məˈʃin/

Forms: 1500s machyne, 1500s-1700s machin, 1500s-machine, 1600s macheen.

Frequency (in current use): •••••

Origin: A borrowing from French. Etymon: French machine.

Etymology: < Middle French, French machine < classical Latin $m\bar{a}china$ (compare MACIG ancient Greek (Attic) $μηχαν\dot{η}$: see MECHANIC adj. and n.), probably related to $μ\dot{η}χος$ 'means, European base of MAY v.' Compare Spanish $m\acute{a}quina$ (1444), Italian macchina (15th cent.); are all 17th-cent. borrowings from French.



It is "remarkably easy to work."—Vide Gardeners' Magazine.
"The quickest, most simple, and most efficient Mower ever used."
Vide Gardeners' Chronicle. And "far superior to any of ours."
Vide The Field. Owing to the unprecedented success this Mower met with last year, increased facilities for manufacture have been required, also many important improvements in the machine have been made, and the prices for 1871 greatly reduced. Send for Catalogue. Williams and Co., Manufacturers and Patentees.
1. City Road, John G. Rollins, Old Swan Wharf, London, E.C., Sole Wholesale Agent for England.

MARAVILLA COCOA. -- No breakfast table is complete without this delicious becerage. - The Globe says: "Taylor

TASK

TUTORIAL

Select the classification that best matches the highlighted 'machine'. If there are several relevant options, choose the most specific classification.

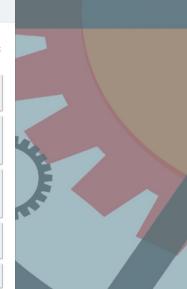
War and military - machines specialised for use in war

Figurative - descriptive, poetical or metaphorical uses; e.g. comparing machines and something else, or vice versa

Transport - includes ships, planes and wheeled vehicles whose primary purpose is transport

Mechanical apparatus - a complex device that performs a certain kind of work

Other



'Classify it!': Analyse du terme 'machine' dans les dictionnaires

machine, n.

Pronunciation: Brit. // məˈʃiːn/, U.S. // məˈʃin/

Forms: 1500s machyne, 1500s–1700s machin, 1500s– machine, 1600s macheen.

Frequency (in current use): •••••

Origin: A borrowing from French. Etymon: French machine.

Etymology: < Middle French, French machine < classical Latin $m\bar{a}china$ (compare MACIG ancient Greek (Attic) $μηχαν\dot{η}$: see MECHANIC $ad\dot{j}$. and n.), probably related to $μ\dot{η}χος$ 'means, European base of MAY v.' Compare Spanish $m\acute{a}quina$ (1444), Italian macchina (15th cent.); are all 17th-cent. borrowings from French.



It is "remarkably easy to work."—Vide Gardeners' Magazine.
"The quickest, most simple, and most efficient Mower ever used."
Vide Gardeners' Chronicle. And "far superior to any of ours."
Vide The Field. Owing to the unprecedented success this Mower met with last year, increased facilities for manufacture have been required, also many important improvements in the machine have been made, and the prices for 1871 greatly reduced. Send for Catalogue. Williams and Co., Manufacturers and Patentees.
1. City Road, John G. Rollins, Old Swan Wharf, London, E.C., Sole Wholesale Agent for England.

MARAVILLA COCOA. -- No breakfast table is complete without this delicious becerage. -- The Globe says: "Taylor

TASK

TUTORIAL

Select the classification that best matches the highlighted 'machine'. If there are several relevant options, choose the most specific classification.

War and military - machines specialised for use in war

Figurative - descriptive, poetical or metaphorical uses; e.g. comparing machines and something else, or vice versa

Transport - includes ships, planes and wheeled vehicles whose primary purpose is transport

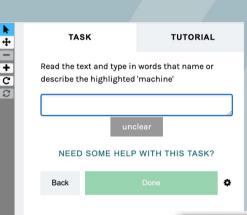
Mechanical apparatus - a complex device that performs a certain kind of work

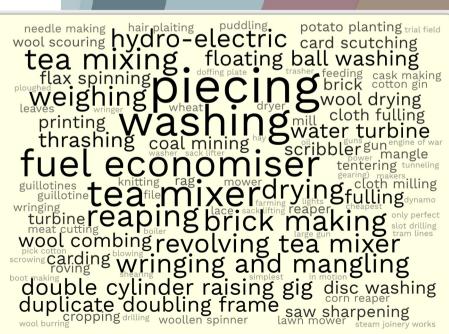
Other



'Describe it!' task

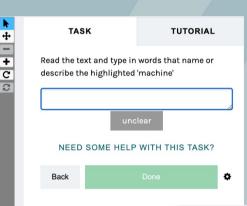
ndles, so that the different grooves can be bught to act on the centre of the hoppers; and, ien necessary to empty them, it is done by aply shifting the rollers from under the hoppers, the same time holding a bag under the delivery be. This machine has three speeds, and can be sown out of gear at the headlands by a moveable teh. The advantages claimed for it are simcity, strength, non-cutting of the seeds, and the gular and continuous delivery. Messrs. M'Kenzie wed a large number of ploughs of almost every aginable description, which all seemed of arreal





'Describe it!' : Activité de transcription

ndles, so that the different grooves can be ought to act on the centre of the hoppers; and, ten necessary to empty them, it is done by aply shifting the rollers from under the hoppers, the same time holding a bag under the delivery be. This machine has three speeds, and can be rown out of gear at the headlands by a moveable tch. The advantages claimed for it are simcity, strength, non-cutting of the seeds, and the cular and continuous delivery. Messrs. M'Kenzie wed a large number of ploughs of almost every aginable description, which all seemed of avoid



needle making hair plaiting puddling. potato planting trial field wool scouring hydro-electric card scutching tea mixing floating ball washing flax spinning plecing ball washing ball washing flax spinning plecing brick cotton gin weighing wheat dryer wool drying wool drying printing washing coal mining water turbine thrashing coal mining scribble water turbine thrashing coal mining scribble guns gun engine of war scribble guns gun engine of w

Case Study: Mapping Victorian Accidents

Kaspar Beelen, Mariona Coll Ardanuy, Federico Nanni

Detect and **disambiguate** place names through **toponym resolution** in articles that mention accidents e.g.:

"On Monday afternoon a dense fog prevailed, and a lamentable **accident** occurred at **{Thorpe: Q7796696}**, near **{Norwich: Q130191}**."

They identified 75,000 place names (toponyms) close to the token 'accident' in five newspapers titles.

Result: visualise 'accidents' over time and space



Étude de cas : Mapping les accidents du XIXe siècle

Kaspar Beelen, Mariona Coll Ardanuy, Federico Nanni

Détecter et **désambiguïser** les noms de lieux par **toponym resolution** dans les articles qui mentionnent des accidents. Par exemple :

"On Monday afternoon a dense fog prevailed, and a lamentable **accident** occurred at **{Thorpe: Q7796696}**, near **{Norwich: Q130191}**."

Ils ont identifié 75,000 noms de lieux (toponyms) proches du token 'accident' dans 5 titres de journaux.

Resultat : visualisation des 'accidents' dans le temps et l'espace.



Future work

Detecting ads (using crowdsourced 'ad or not' dataset, compound nouns)

Identifying copyright-free 'unauthored' works

Publishing derived datasets for wider use Visualisations in exhibition (Leeds; July 2022 - January 2023)

'Convergence' work to unite methods and sources across the project

Integration with BL systems, training



Travail futur

Détection des pubs (avec un crowdsourced dataset - 'pub ou non'; noms composés)

Identification d'œuvres 'sans auteur' et libre de copyright

Publication de derived datasets pour une utilisation plus large

Visualisations dans une exposition (Leeds; July 2022 - January 2023)

Travail de 'convergence' pour unir les méthodes et sources à travers le projet

Intégration avec les systèmes du BL; formation



Thanks for listening!

Find out more:

https://livingwithmachines.ac.uk

Try our crowdsourcing task http://bit.ly/LivingWithMachines

We look forward to your questions: digitalresearch@bl.uk

@LivingWMachines @BL_DigiSchol @mia_out

Merci de votre attention!

Pour en savoir plus:

https://livingwithmachines.ac.uk

Essayez notre activité de crowdsourcing: http://bit.ly/LivingWithMachines

Nous attendons vos questions: digitalresearch@bl.uk

@LivingWMachines @BL_DigiSchol @mia_out