A brief overview: AI at the British Library

From utility tools to new research methods
May 2024

Dr Mia Ridge, Digital Curator, British Library
@mia@hcommons.social @miaout.bsky.social
https://blogs.bl.uk/digital-scholarship/
This presentation DOI: 10.5281/zenodo.11390935

In this presentation

- The British Library's Digital Research team and the *Living with Machines* project

- From utility tools to new research methods: AI (and big data, deep learning, machine learning, data science) examples from the British Library and *Living with Machines*

Find out more:

- Blog posts https://blogs.bl.uk/digital-scholarship/
- British Library's Research Repository https://bl.iro.bl.uk/
- *Living with Machines* https://livingwithmachines.ac.uk/
- *LwM* code https://github.com/Living-with-machines

The British Library's Digital Research team

Enabling the use of the Library's digital collections for research, inspiration, creativity, and enjoyment

- Collaborating to explore new and emerging methods
- Understanding 'reader' needs to improve access and usability
- Supporting IIIF and the Universal Viewer
- **Increasing digital / AI literacy across the Library** – Digital Scholarship Training Programme – **and wider sector**

Living with Machines (2018-23)

https://livingwithmachines.ac.uk

Our Partners

The Alan Turing Institute

British Library

University of Cambridge

University of East Anglia

University of Exeter

Queen Mary University of London

Our Funders

Arts & Humanities Research Council

UK Research and Innovation

How did machines change 19th century lives? A 'data-driven history project, and a historically-informed data science project'

Principal and Co-Investigators

Ruth Ahnert
(QMUL)

David Beavan
(Turing)

Emma Griffin
(UEA)

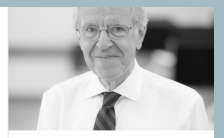
Timothy Hobson
(Turing)

Jon Lawrence
(Exeter)

Maja Maricevic
(British Library)

Barbara McGillivray
(Turing / King's College

Mia Ridge
(British Library)



Alan Wilson
(Turing)

Project Team and Alumni

Claire Austin
Rights Assurance

Kaspar Beelen
Digital Humanities Senior Research Associate

Mariona Coll Ardanuy
Computational Linguistics Senior Research Associate

Léllé Demertzi
Programme Coordinator

Luke Hare
Research Data Scientist

Sherman Lo
Research Data Scientist

Katie McDonough
Senior Research Associate

Federico Nanni
Senior Research Data Scientist

Nilo Pedrazzini
Corpus-Based Digital Humanities Research Assistant

André Piza
Research Project Manager

Griffith Rees
Research Data Scientist

Kalle Westerling
Research Software Engineer

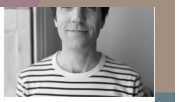
Filipe Bento
Research Associate

Giovanni Colavizza
(Turing)

Joel Dearden (1979-2020)
Research Software Engineer

Michael Jackson
(EPCC), Software Architect

Joshua Rhodes
Senior Research Associate



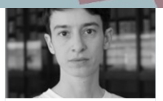
Yann Ryan
British Library Curator, Digital Newspapers

Adam Farquhar
(British Library)

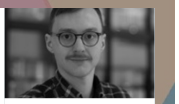
Rosa Filgueira
(EPCC), Data Architect

Sarah Gibson
Research Software Engineer

Guy Solomon
Research Associate



Giorgia Tolfo
Data And Content Manager

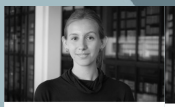


Daniel Van Strien
Digital Curator

Lucy Havens
Research Software Engineer Intern

James Hetherington
(Turing)

Kasra Hosseini
Research Data Scientist



Olivia Vane
Digital Humanities Research Software Engineer

Daniel Wilson
History Senior Research Associate

Rosie Wood
Research Data Scientist

AI tools and research with collections in *LwM*

Natural Language Processing (NLP) and LLM methods:

- Toponym resolution (T-Res) – finding, disambiguating place names in text
- **Algorithms to 'link' individuals across census years**
- Linguistic methods to find machines assigned human-like agency; semantic shifts as words change over time
- **Understanding 'bias' in digitised newspapers corpora with paradata from press directories**
- Deep learning for searching poor OCR (DeezyMatch)
- 'Figurative search' with LLMs ('Living Machine')

'Reading' maps with computer vision models

Helper tools and utilities with AI / machine learning with British Library collections

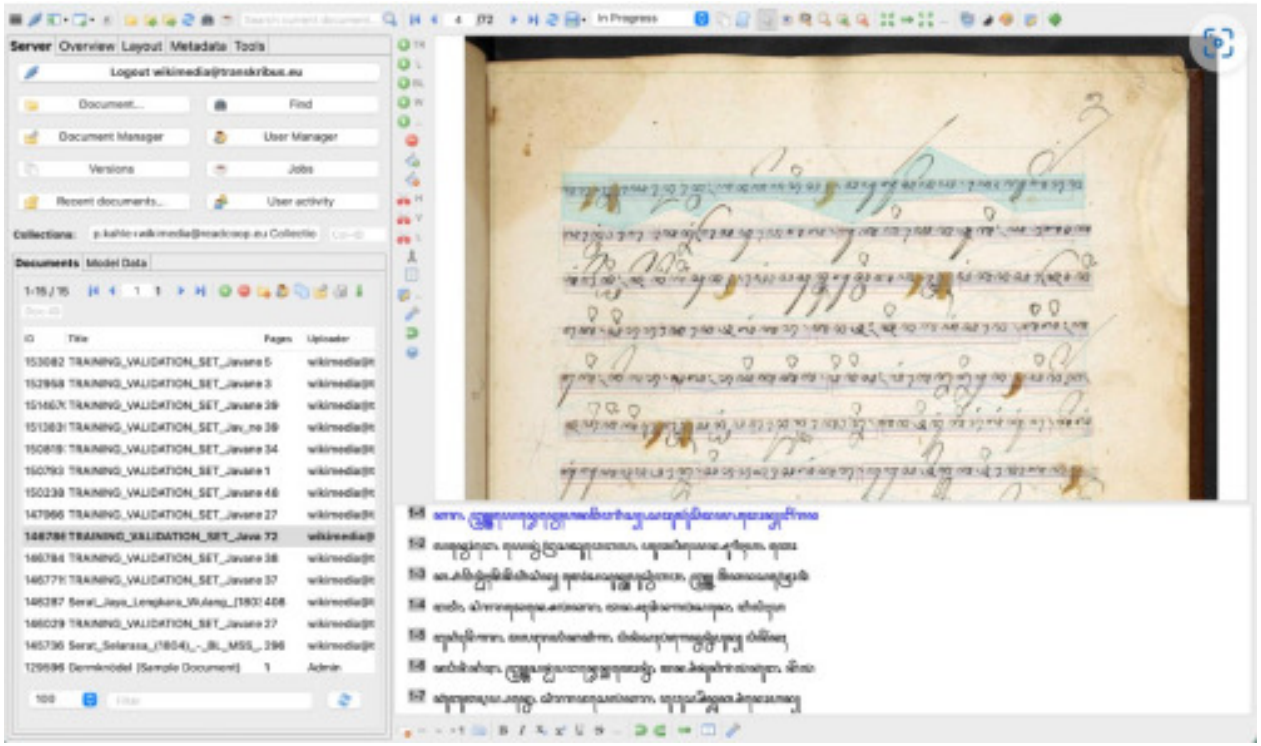
Transkribus: ML for text transcription

Handwritten or printed text

Can be trained to recognise most languages / hands

Constantly improved text and layout recognition – improve performance by training your own model

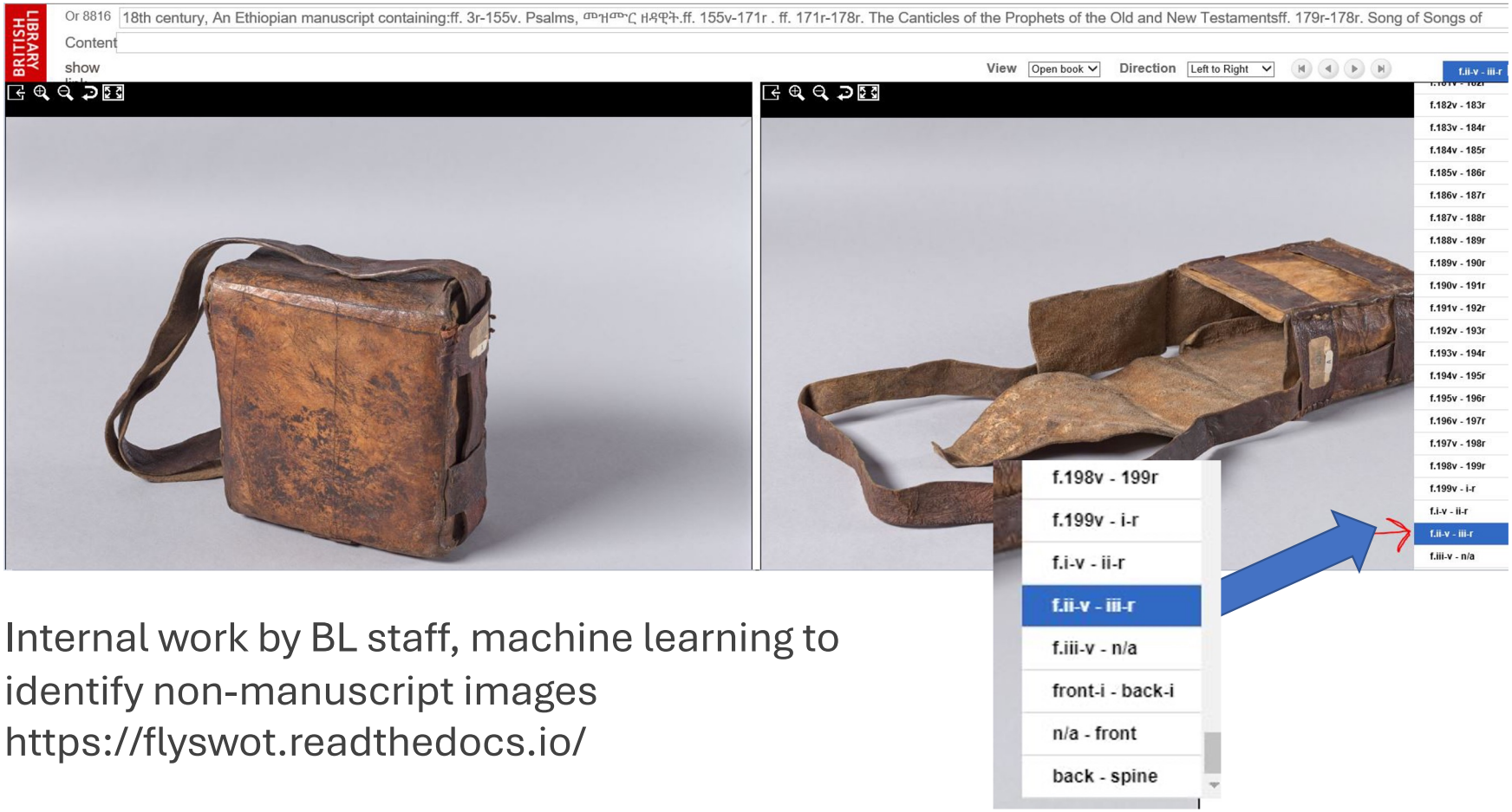
Example: Javanese manuscripts transcribed on Wikisource are being used to create a HTR model



Dr. Adi Keinan-Schoonbaert, Digital Curator

https://blogs.bl.uk/digital-scholarship/2023/08/the-british-library-loves-manuscripts-on-wikisource.html

Flyswot: Detecting ‘fake flysheets’ with ML



Internal work by BL staff, machine learning to identify non-manuscript images
https://flyswot.readthedocs.io/

Languid: Language Identification Project

BRITISH LIBRARY

	afr	alb	amh	ara	arm	aze	baq	bel	ben	bos	bul	cat	ceb	chi	cos	cze	dan	dut	eng	epo	est	fin	fre	fry	geo	ger	gla	gle	glg	gre	guj
afr	85%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
alb	0%	93%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
amh	0%	0%	43%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
ara	0%	0%	0%	61%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
arm	0%	0%	0%	0%	81%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
aze	0%	0%	0%	0%	0%	27%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
baq	0%	0%	0%	0%	0%	0%	52%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
bel	0%	0%	0%	0%	0%	0%	0%	92%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
ben	0%	0%	0%	0%	0%	0%	0%	0%	46%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
bos	0%	0%	0%	0%	0%	0%	0%	0%	0%	57%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
bul	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	92%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
cat	0%	0%	0%	1%	0%	0%	0%	0%	0%	0%	0%	79%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
ceb	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
chi	0%	0%	0%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	69%	0%	0%	0%	0%	1%	1%	0%	3%	0%	0%	0%	0%	0%	0%	0%	0%	0%
cos	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
cze	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	0%	0%	0%	0%	91%	0%	0%	1%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
dan	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	90%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
dut	9%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	2%	0%	2%	0%	0%	0%	92%	0%	0%	0%	0%	0%	22%	0%	0%	0%	0%	0%	0%	0%
eng	1%	1%	33%	16%	6%	9%	0%	0%	49%	0%	1%	4%	0%	25%	0%	3%	2%	2%	83%	8%	5%	4%	2%	0%	12%	4%	15%	28%	0%	4%	52%
epo	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	78%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
est	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	76%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
fin	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	79%	0%	0%	0%	0%	0%	0%	0%	0%	0%
fre	0%	0%	0%	1%	4%	0%	16%	1%	0%	0%	1%	3%	0%	0%	0%	0%	0%	1%	1%	1%	1%	1%	96%	4%	0%	0%	0%	0%	0%	2%	0%
fry	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	67%	0%	0%	0%	0%	0%	0%	0%
geo	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	65%	0%	0%	0%	0%	0%	0%
ger	2%	1%	14%	7%	3%	0%	0%	0%	0%	14%	1%	0%	0%	0%	0%	2%	3%	1%	0%	1%	6%	4%	0%	4%	0%	93%	0%	0%	0%	2%	0%
gla	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	51%	0%	0%	0%	0%
gle	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	17%	72%	0%	0%	0%
glg	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	33%	0%	0%
gre	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	89%	0%
guj	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	44%

A 'confusion matrix' of languages commonly confused with each other e.g. Afrikaans and Dutch. Figure 8 from Automated Language Identification of Bibliographic Resources (2019).

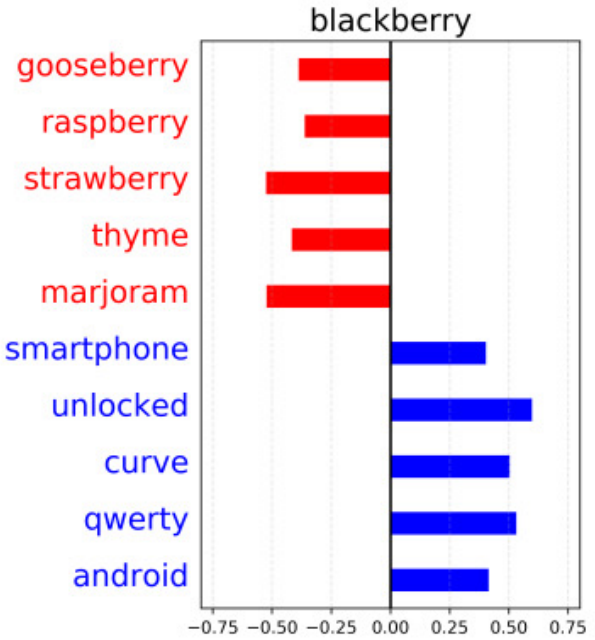
Internal work by Victoria Morris using machine learning techniques to assign language codes to MARC catalogue records.

Language codes were assigned to 1.15 million records with 99.7% confidence. 471 languages identified, 141 of which were not previously represented.

More AI/ML at the British Library

- eScriptorium for handwritten text recognition (HTR) of historical Chinese manuscripts
- Training including hands-on 'Hack & Yacks' to explore specific tools
- Text extracted from historical hand-drawn maps with Google Cloud Vision API
- Personal experiments by staff using ChatGPT to write simple code for e.g. data processing

- Research with the UK Web Archive - word vectors to track changing meanings for words over time



Making bad OCR text searchable

Some of our items were digitised decades ago. Poor automatic transcription (OCR) hinders small and large-scale research

DeezyMatch improves search - a Living with Machines tool shared for wider use

A Flexible Deep Neural Network Approach to Fuzzy String Matching

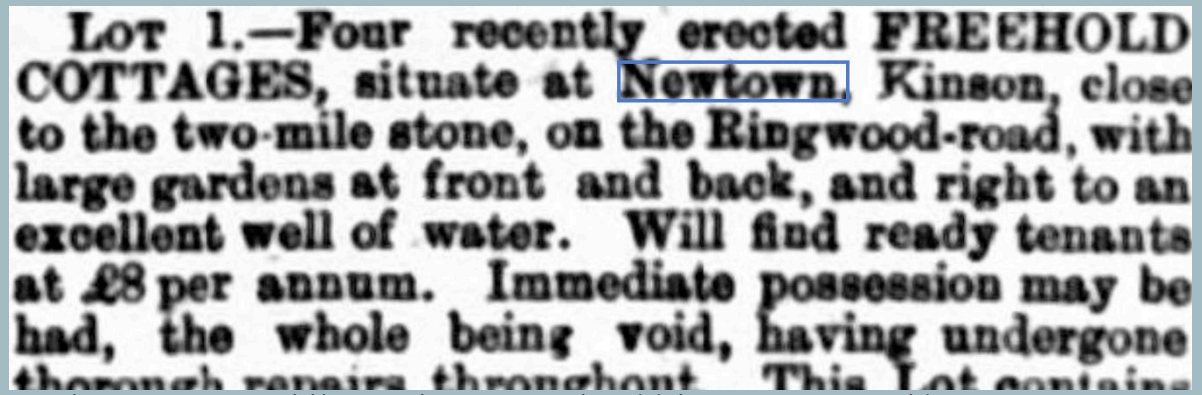
DeezyMatch can be used in the following tasks:

- Fuzzy string matching
- Candidate ranking/selection
- Query expansion
- Toponym matching

Or as a component in tasks requiring fuzzy string matching and candidate ranking, such as:

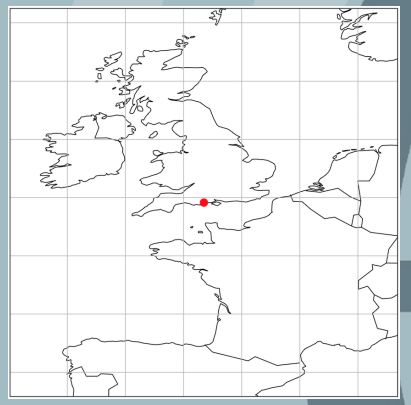
- Record linkage
- Entity linking

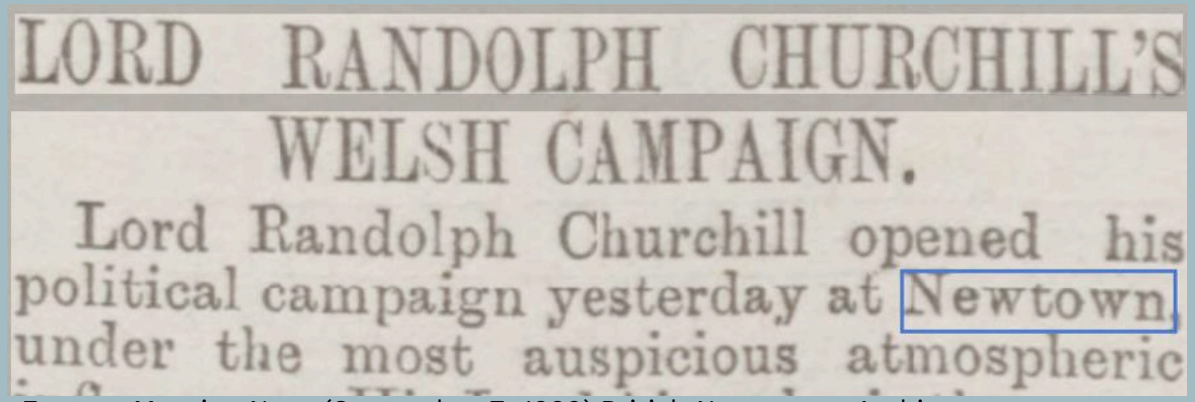
Finding, disambiguating and locating place names in texts (toponym resolution)



LOT 1.—Four recently erected FREEHOLD COTTAGES, situate at Newtown, Kinson, close to the two-mile stone, on the Ringwood-road, with large gardens at front and back, and right to an excellent well of water. Will find ready tenants at £8 per annum. Immediate possession may be had, the whole being void, having undergone thorough repairs throughout. This Lot contains

Poole & Dorset Herald(November 23, 1882), British Newspaper Archives

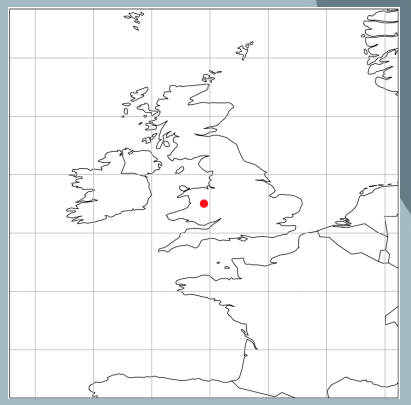




LORD RANDOLPH CHURCHILL'S WELSH CAMPAIGN.

Lord Randolph Churchill opened his political campaign yesterday at Newtown, under the most auspicious atmospheric

Eastern Morning News(September 7, 1889),British Newspaper Archives



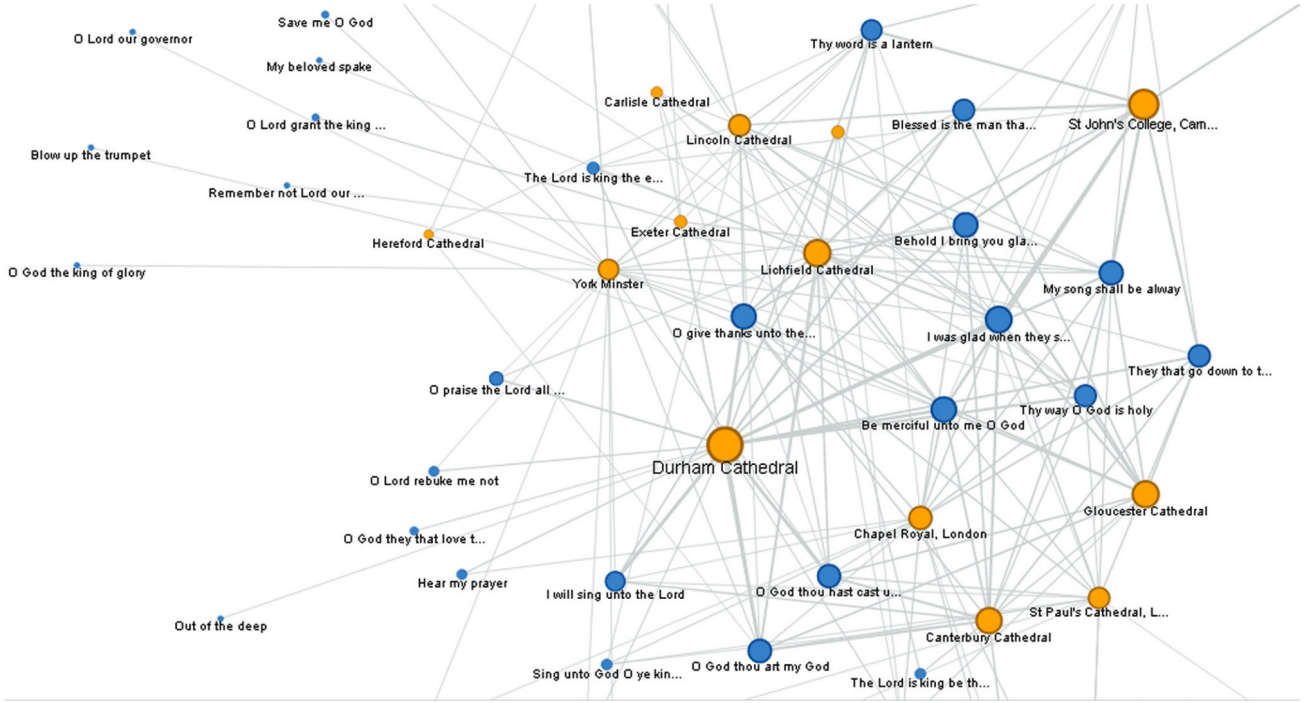
Research and creative AI / machine learning with British Library collections

A Big Data History of Music

AHRC-funded collaboration with Royal Holloway 2014 – 15

Made BL catalogue records for printed and manuscript music available as open data

Case studies and visualisations 'show the potential of quantitative approaches to complement the close or contextual readings customarily used by historical musicologists'



8 Manuscripts of Purcell's anthems in cathedral and chapel libraries. Data from RISM A/II

Flickr image similarity

'16 Very Sad Girls'
Artworks and Findings using Flickr Commons (2014-2016)

Work by Mario Klingemann (Quasimondo) using semi-automatic image classification, vector space clustering, machine learning

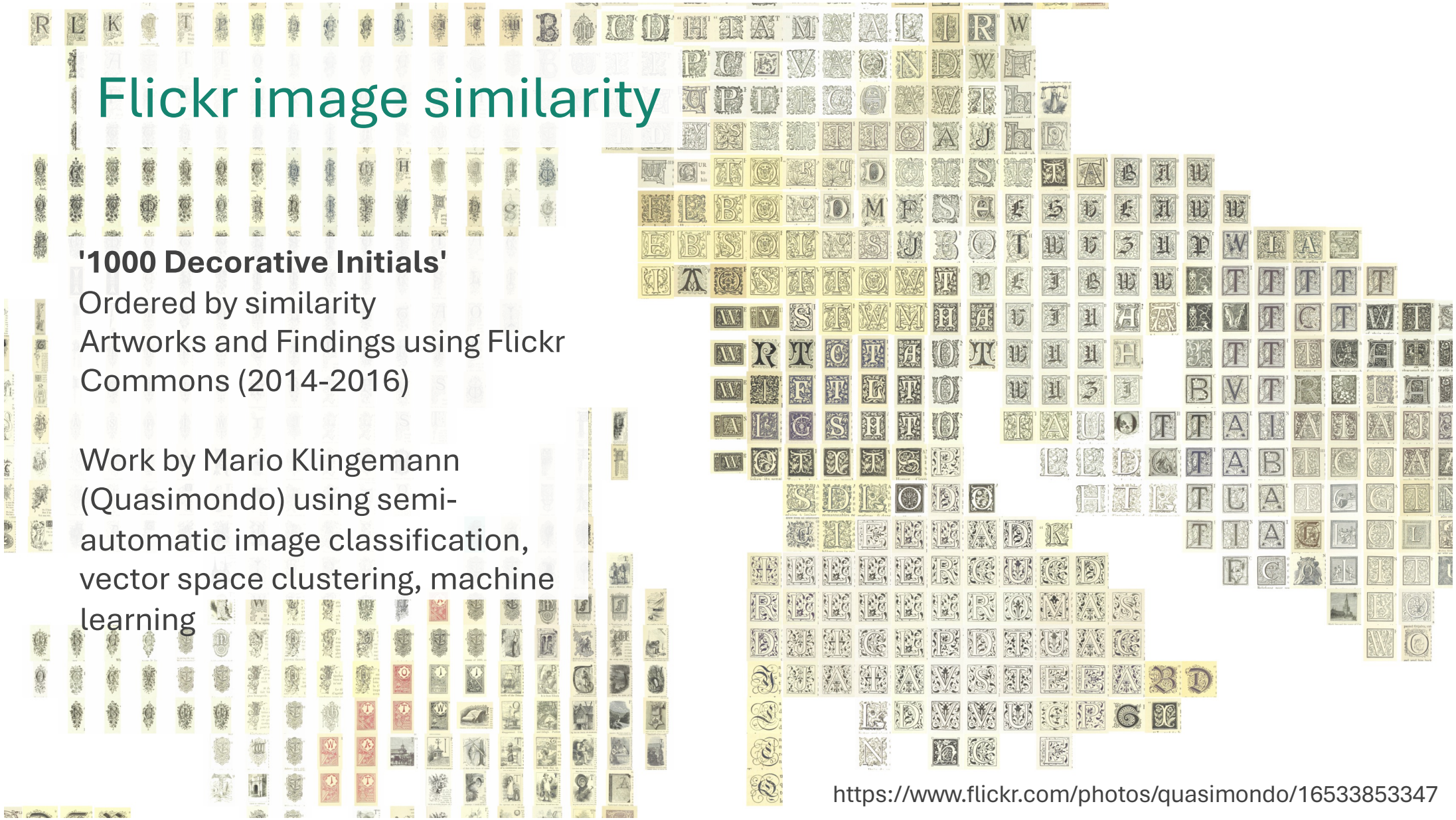
Based on 1 million images released on Flickr Commons

https://www.flickr.com/photos/quasimondo/albums/72157638820730895/

Flickr image similarity

'1000 Decorative Initials'
Ordered by similarity
Artworks and Findings using Flickr Commons (2014-2016)

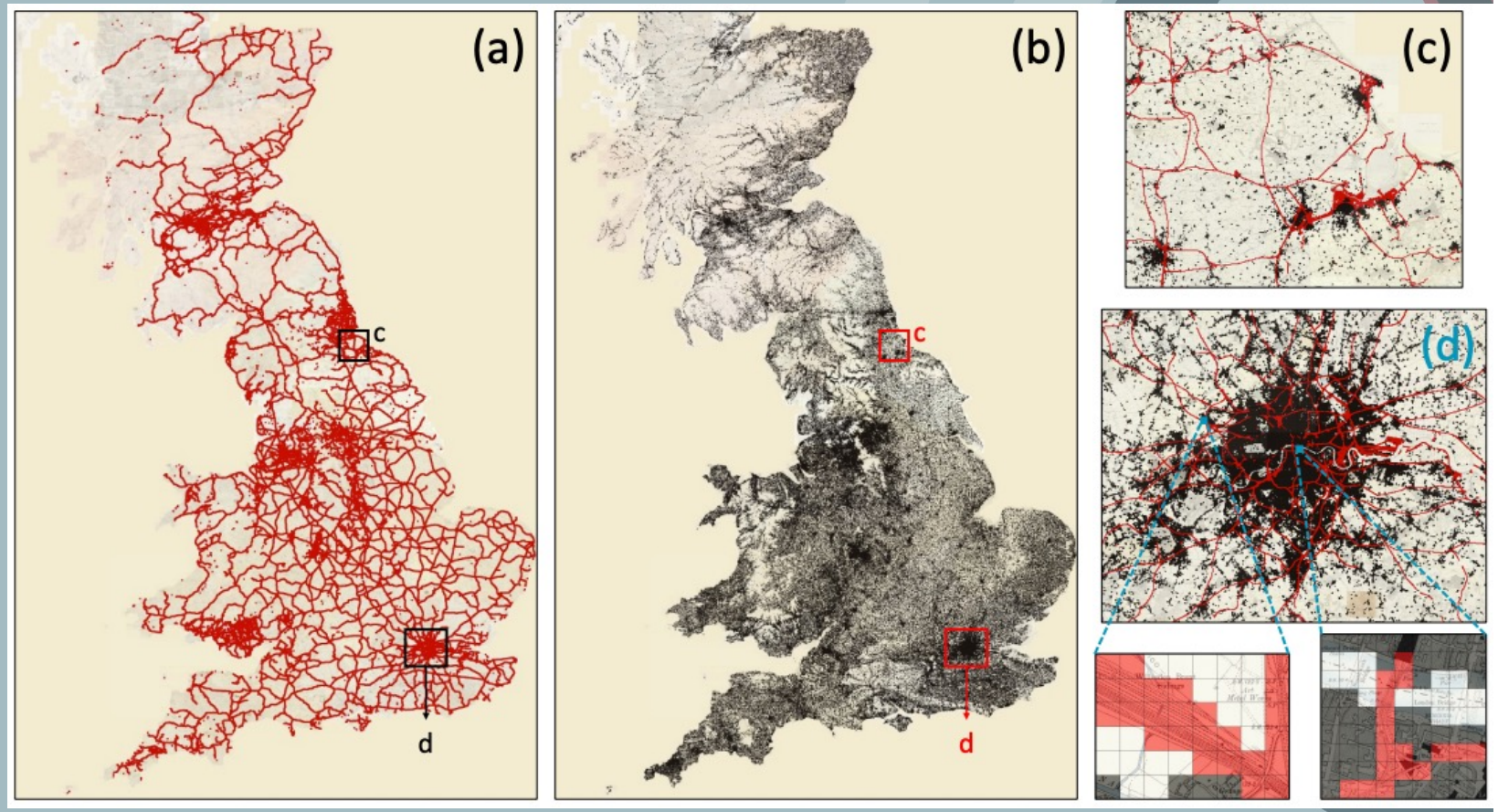
Work by Mario Klingemann (Quasimondo) using semi-automatic image classification, vector space clustering, machine learning



https://www.flickr.com/photos/quasimondo/16533853347

MapReader / Railspace: computer vision + ML

- Annotated 62020 patches of maps with yes / no railway
- Trained ML model with 60% of patches
- Able to scale up to predict rail (or other features) across GB



Living Machines: linguistic methods for a study of atypical animacy

When are inanimate objects - machines - given animate attributes?

Target	Sentence	Animacy	Humanness
engine	In December, the first steam fire **engine** was received, and tried on the shore of Lake Monona, with one thousand feet of hose.	0	0
engine	It was not necessary for Jakie to slow down in order to allow the wild **engine** to come up with him; she was coming up at every revolution of her wheels.	1	1
locomotive	Nearly a generation had been strangely neglected to grow up un-Americanized, and the private adventurer and the **locomotive** were the untechnical missionaries to open a way for the common school.	1	1
machine	The worst of it was, the people were surly; not one would get out of our way until the last minute, and many pretended not to see us coming, though the **machine**, held in by the brake, squeaked a pitiful warning.	1	1
machines	Our servants, like mere **machines**, move on their mercenary track without feeling.	1	0
machinery	We have everywhere water power to any desirable extent, suitable for propelling all kinds of **machinery**	0	0

'The Living Machine: A Computational Approach to the Nineteenth-Century Language of Technology'

- Trained a LLM (BERT) on digitised BL 19thC books ('BLERT') to explore the uses of a language model for historical research
- They devised a method for 'figurative search' to help find figures of speech in nineteenth-century texts that portrayed machines as self-acting, automatic, or alive – or as 'mere machines'

#	Books	Newspapers
1	other (1655)	infernal (748)
2	threshing (1513)	new (589)
3	new (1455)	threshing (506)
4	infernal (1185)	other (472)
5	mere (1065)	mere (319)
6	first (1056)	best (268)
7	whole (785)	whole (226)
8	great (767)	hydraulic (198)
9	electric (657)	old (184)

Lessons learnt / successes: AI at the BL

- After *LwM*, BL **better understands** researchers' needs; copyright, infrastructure, workflows for **AI with collections**
- Outputs, case studies and lessons learnt integrated into existing BL **training** programme and external talks
- Contributed to **new skills**, increased **AI literacy** in staff
- Feeds into current work on **AI strategy, ethics**
- Publishing well-documented, reusable datasets, tools and code increases value and impact; **enables future AI projects**

Thanks for reading!

Questions? digitalresearch@bl.uk

Dr Mia Ridge, Digital Curator, British Library

@mia@hcommons.social @miaout.bsky.social
@BL_DigiSchol@techhub.social @LivingWithMachines@zirk.us

https://blogs.bl.uk/digital-scholarship/