Evolutionary Innovations: Collections as Data in the AI era

Dr Mia Ridge
Digital Curator, British Library
@mia@hcommons.social
@BL_DigiSchol@techhub.social

BRITISH LIBRARY

Digital Research at the British Library

Encouraging access, improving usability of our digital collections by experimenting with digital technologies and collaborating on a wide range of digital projects

- Exploring automated text transcription, data science, crowdsourcing, emerging formats, corpus linguistics
- Supporting IIIF and the Universal Viewer
- Understanding 'reader' needs
- Increasing literacy across the Library – Digital Scholarship Training Programme – and wider sector

The challenge of scale at the British Library

Our mission: 'For research, inspiration and enjoyment'

The British Library is the national library of the UK with **170 - 200 million items**, including:

16 million books; 8 million stamps; 350,000 manuscript volumes; 60 million patents; 4 million maps; 1.6 million music scores; 60 million newspapers; pamphlets, magazines; television and radio recordings; sounds; billions of webpages; terabytes of e-books, e-journals.

Over 3 million physical and born-digital new items are added every year.

What is 'collections as data'?

(You might already be doing it)

Collections as Data

A movement to share openly reusable, computationally accessible data – metadata, images, text, etc – from digitised or born-digital collections

Interior of Townsville library, ca. 1948; State Library of Queensland on Flickr Commons

'There aren't (yet) any maps or guidelines for working on collections as data. Choices are always embedded in specific social and technical contexts.'

Dog seated at table and waiting to eat birthday cake with owner Miss Gault; State Library of Queensland on Flickr Commons

Why share your Collections as Data?

- To support research, pedagogical, and creative uses – for staff, readers, the public
- To support internal work on digital, AI / machine learning (ML) methods

Gloria Huish at her desk at the Public Library of Queensland, Brisbane, ca 1952, State Library of Queensland on Flickr Commons

What is 'AI'? A buzzword for 'machine learning' (ML)

Statistical models of words, images, AV. These models are used to predict classifications or generate new images, text, audio, video

'Training data' includes the best and the worst of online content

It's fancy predictive text - doesn't 'know' anything about the world

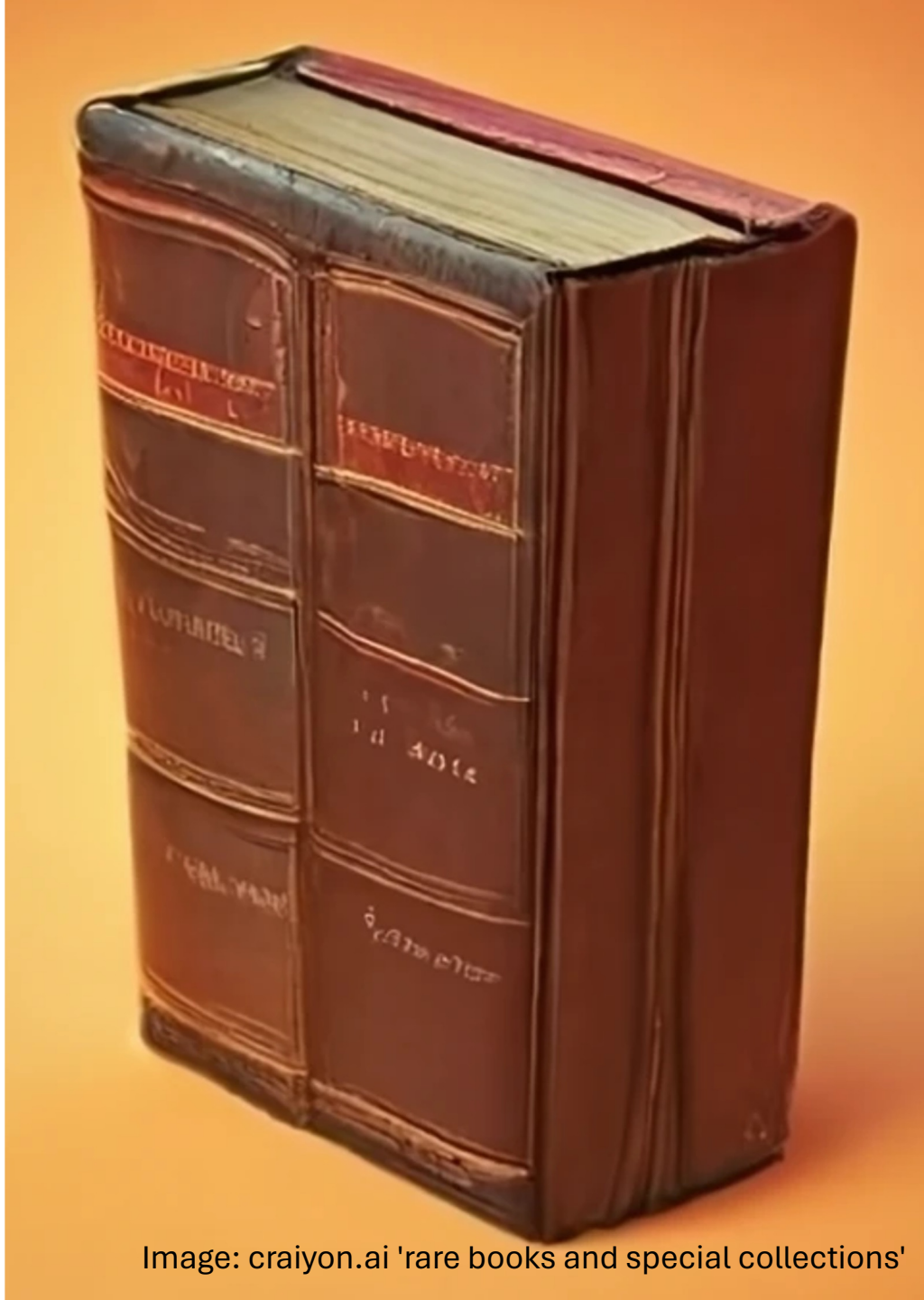


Image: craiyon.ai 'rare books and special collections'

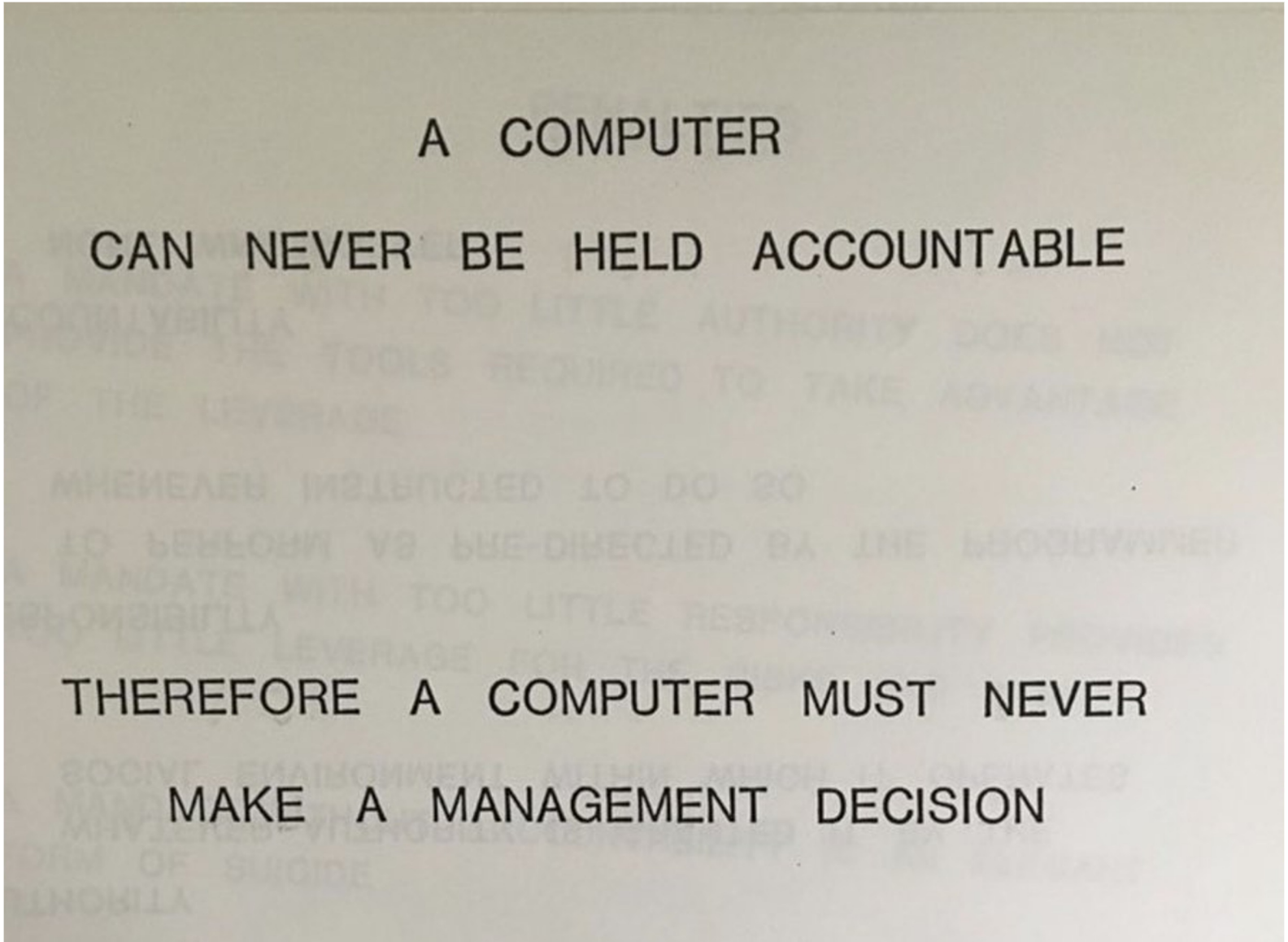
Whose content is 'generative' AI regurgitating?

Are we ok with that?

Whose lives and stories aren't represented?

https://www.vintag.es/2015/09/amazing-vintage-wedding-photos-in.html https://www.flickr.com/photos/147326970@N05/32446909980
https://www.flickr.com/photos/47125576@N00/24108399394 https://www.flickr.com/photos/bokelicious/13981327606/

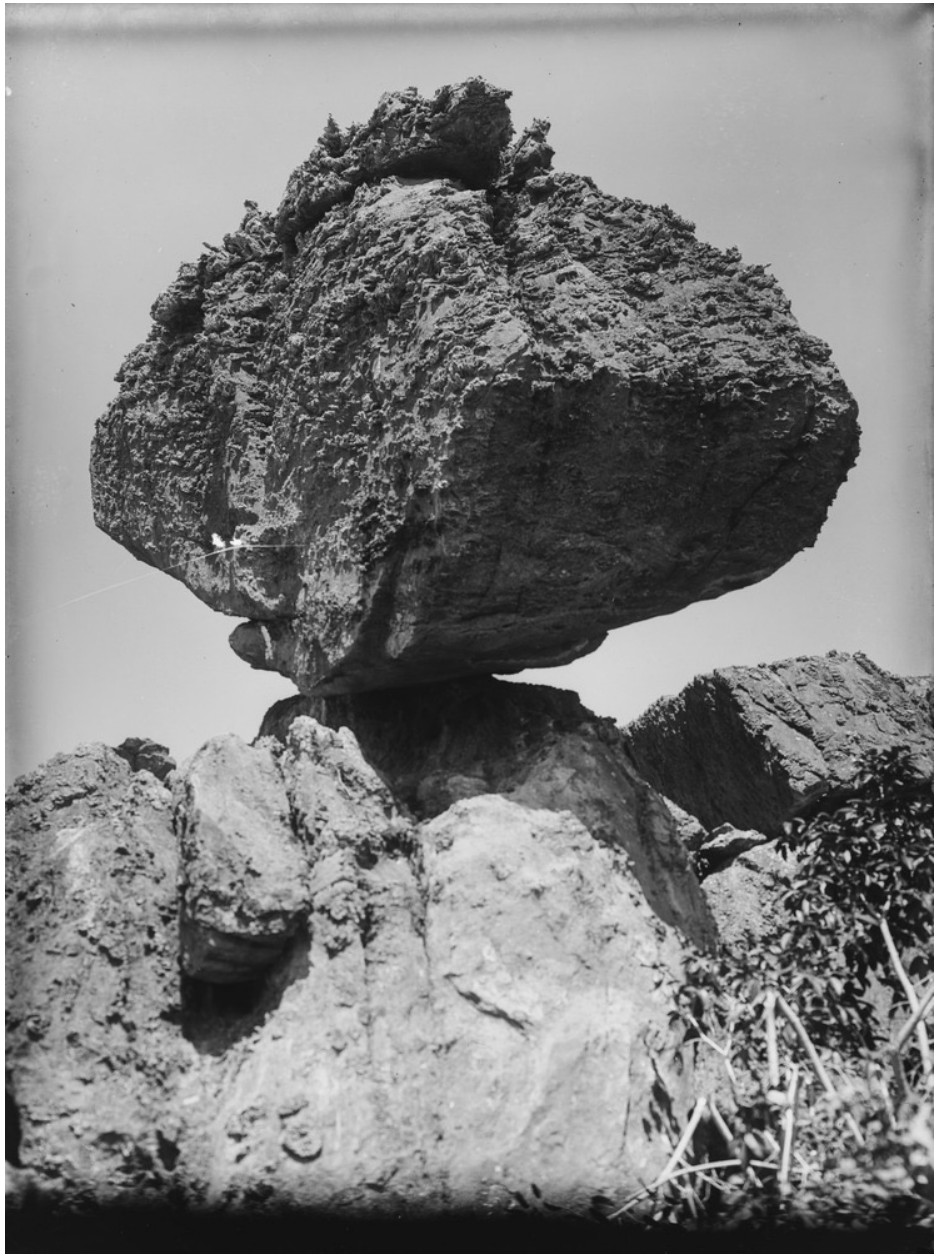
Consequential decisions are still our responsibility



How do we hold companies accountable for infrastructure we can't see?

AI in Collections as Data:

Benefits and limitations



Harriett Brims Collection; State Library of Queensland on Flickr Commons

AI tools can enrich GLAM records

- Transcribe text from images and audio
- Create structured data from unstructured text
- Segment page regions by type
- Translate into other languages, audiences
- Detect objects in images; generate keywords, labels, descriptions
- Detect and link entities - people, places, dates, concepts
- Cluster similar images, texts / improve search by expanding input keywords, images

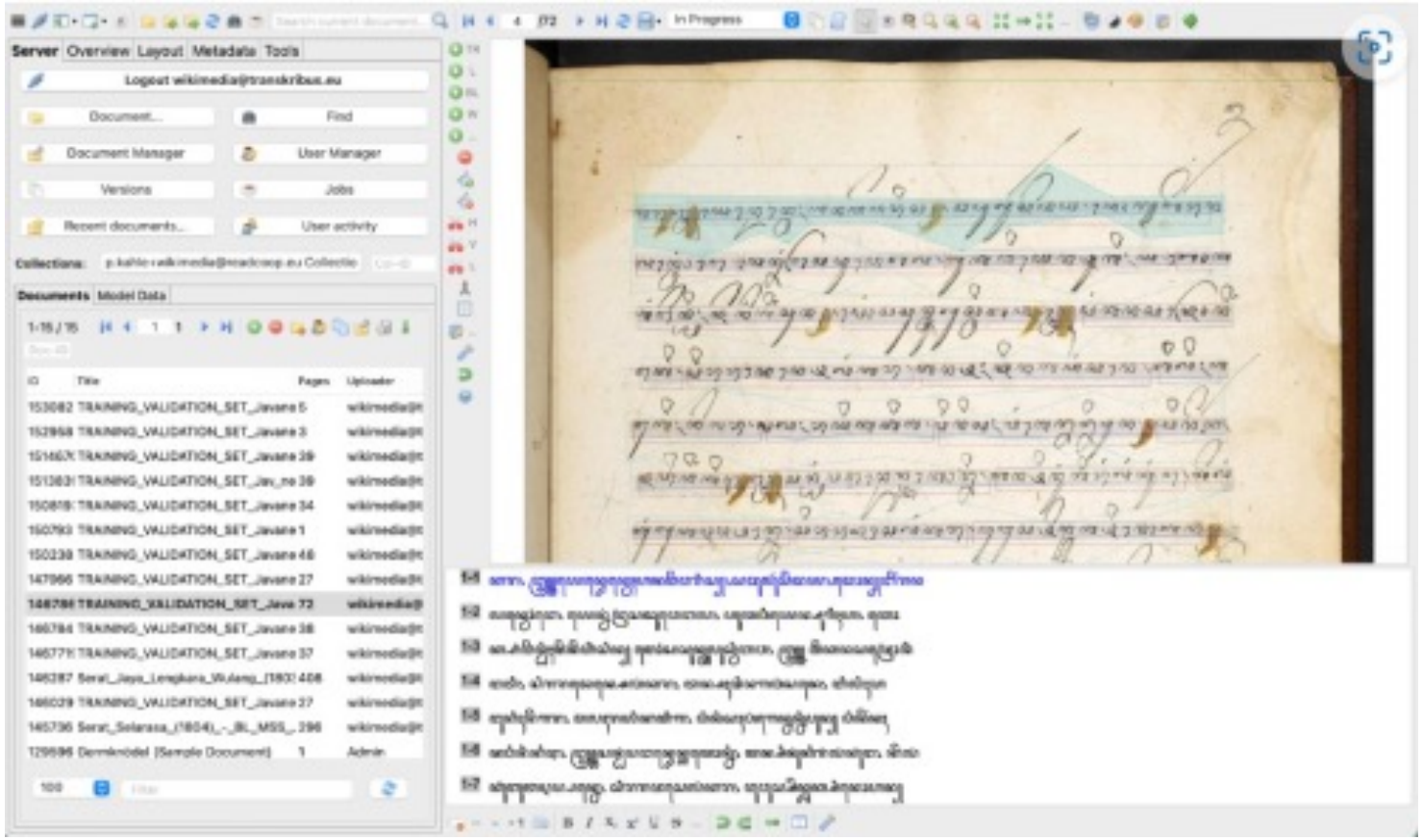
Transkribus: ML for text transcription

Handwritten or printed text

Can be trained to recognise most languages

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

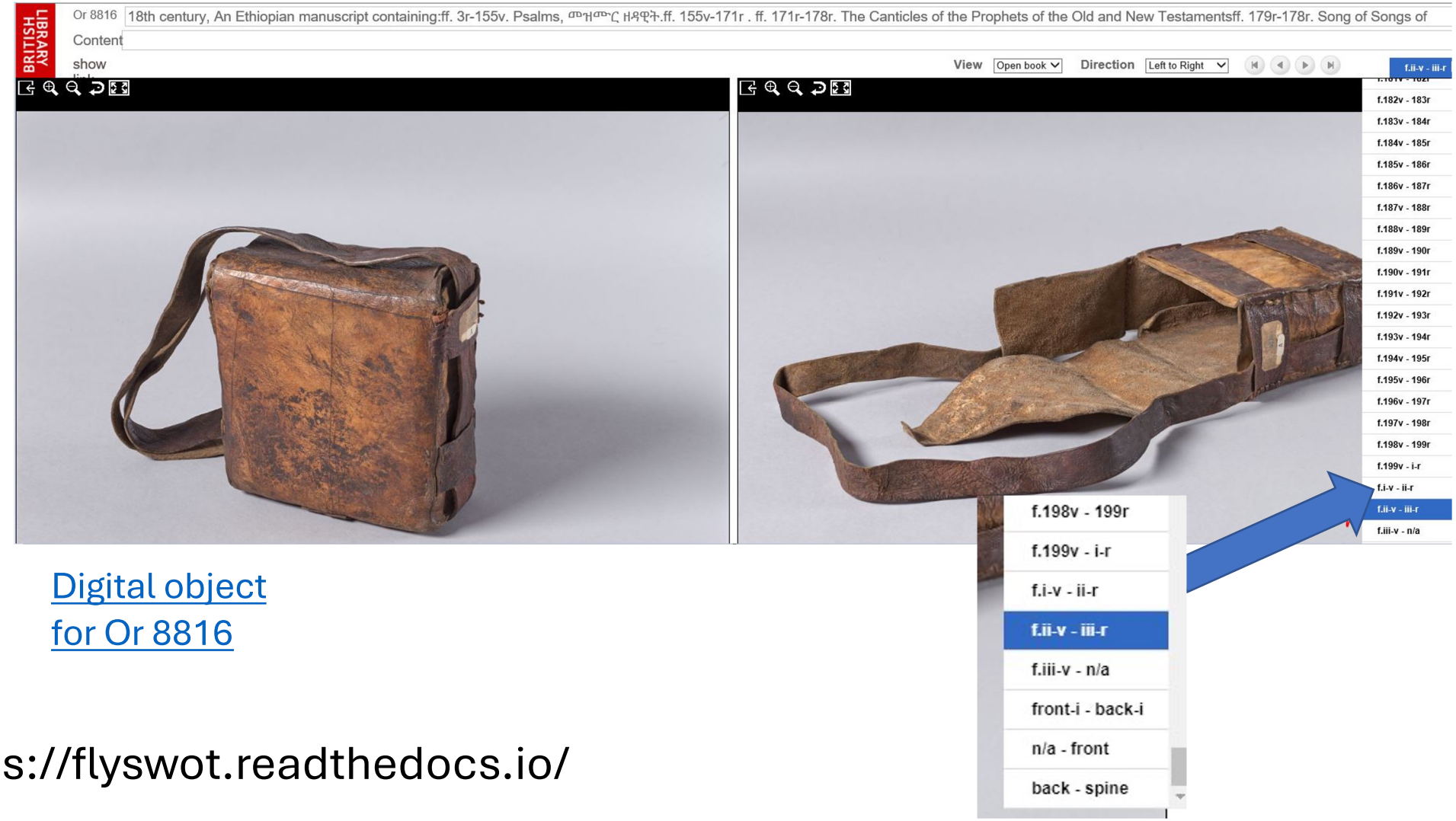
Current work: Javanese manuscripts transcribed on Wikisource are being used to create a HTR model



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

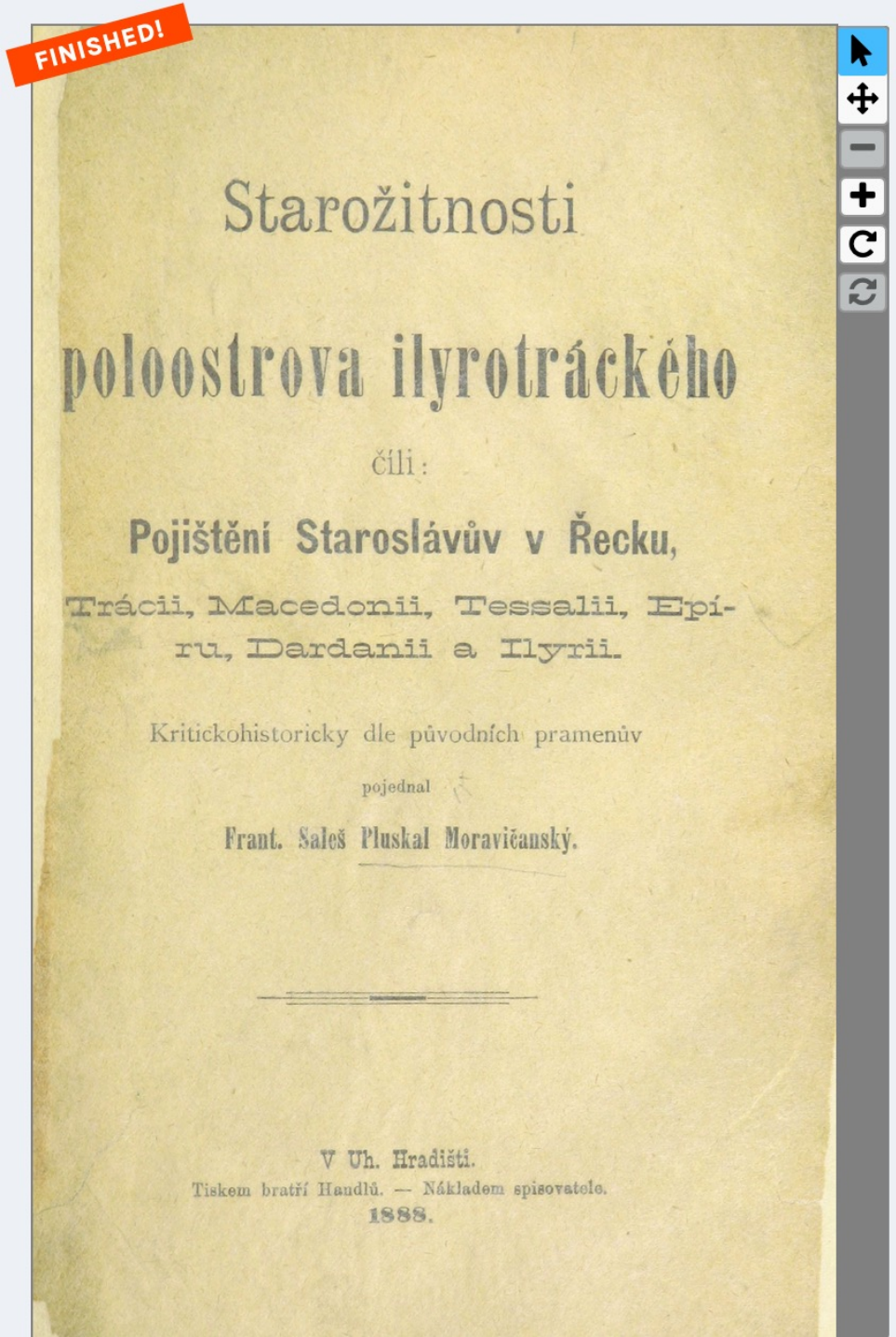


[Digital object for Or 8816]

https://flyswot.readthedocs.io/

More AI/ML at the British Library

- Hands-on 'Hack & Yacks' to explore specific tools
- UK Web Archive - word vectors to track changing meanings for words over time
- Language identification for print collections - used probabilistic methods to predict the language of a book; niche-sourced review on Zooniverse
- Google Cloud Vision API used to extract text from hand-drawn maps
- Personal experiments with ChatGPT to write simple code for e.g. data processing



Living with Machines

2018 - 23 project between The Alan Turing Institute and British Library with partner universities; funded by AHRC and UK Research and Innovation (UKRI)

Collaboration between data science, history, digital humanities to study the impact of mechanisation via new 'machines'

A data-intensive project working with digitised sources at scale

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

Living with Machines: a research library doing AI / data science / digital humanities

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)
Read more

Mia Ridge
(British Library)
Read more

Alan Wilson
(Turing)
Read more

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

Kasra Hosseini
(Turing), Research Data Scientist
Read more

Katie McDonough
History Research Associate
Read more

Federico Nanni
(Turing), Research Data Scientist
Read more

André Piza
Research Project Manager
Read more

Giorgia Tolfo
Data And Content Manager
Read more

Daniel Van Strien
Digital Curator
Read more

Olivia Vane
Digital Humanities Research Software Engineer
Read more

Daniel Wilson
History Research Associate
Read more

Past Collaborators

Giovanni Colavizza
(Turing)
Read more

Joel Dearden
(Turing), Research Software Engineer
Read more

Adam Farquhar
(British Library)
Read more

Rosa Filgueira
(EPCC), Data Architec
Read more

Sarah Gibson
(Turing), Research Software Engineer
Read more

James Hetherington
(Turing)
Read more

Michael Jackson
(EPCC), Software Architect
Read more

Yann Ryan
British Library Curator, Digital Newspapers
Read more

AI methods for research with collections in LwM

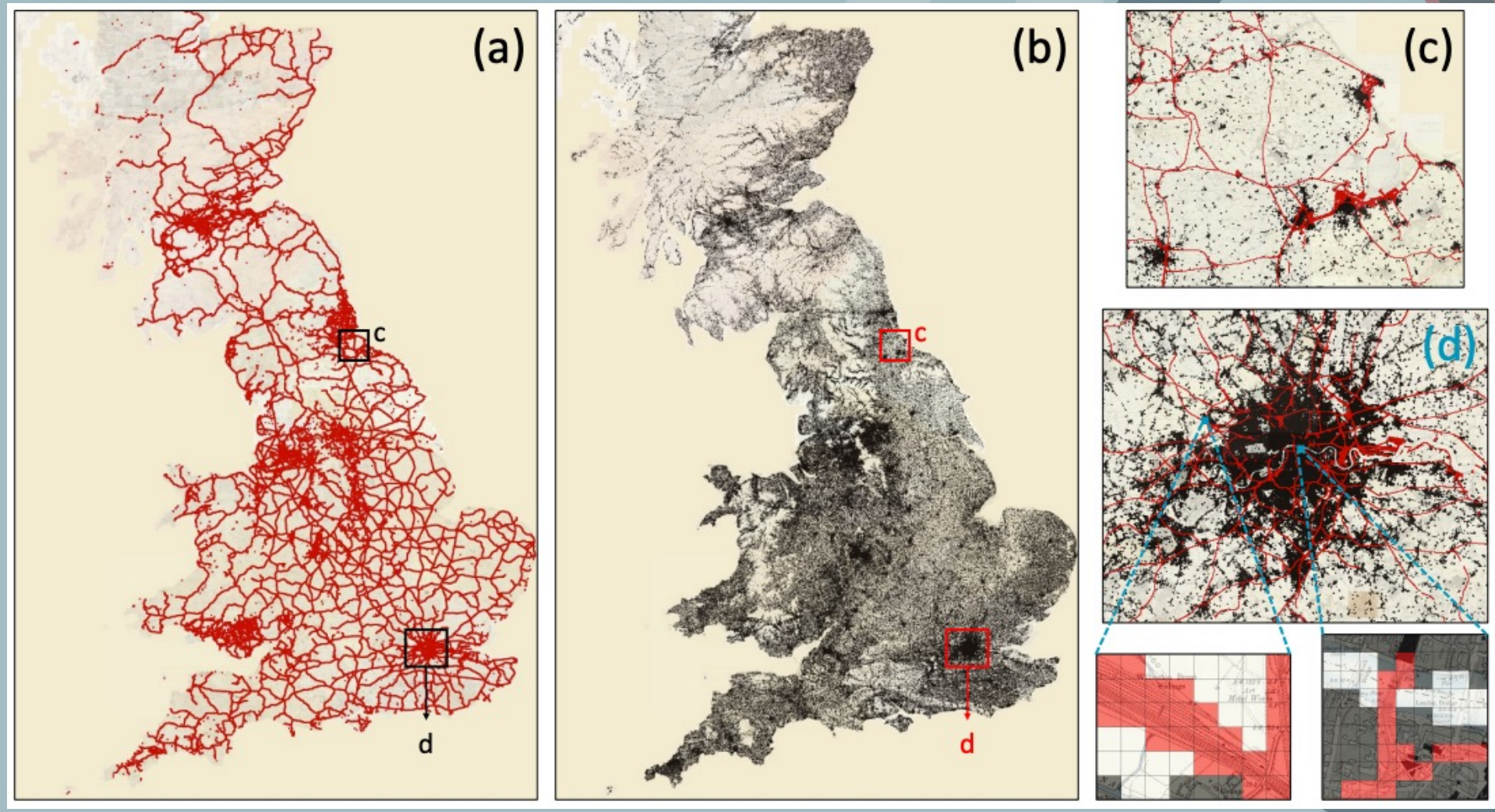
Analysing text with Natural Language Processing (NLP), Large Language Models (LLM):

- Algorithms to 'link' individuals across census years
- Linguistic research to find machines assigned human-like agency; semantic shifts as words change over time
- Contextualising data (newspapers) with historical paradata (Mitchells press directories)
- Trained an LLM (BERT) on digitised BL 19thC books ('BLERT') to explore the uses of a language model for historical research – moving beyond keyword search

https://livingwithmachines.ac.uk/achievements

MapReader / Railspace: computer vision + ML

- Team annotated 62020 patches of maps with yes / no railway
- Trained an ML model with 60% of the patches
- Able to scale up to predict rail across GB



Machine learning for bad OCR

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

DeezyMatch improves search

A Flexible Deep Neural Network Approach to Fuzzy String Matching

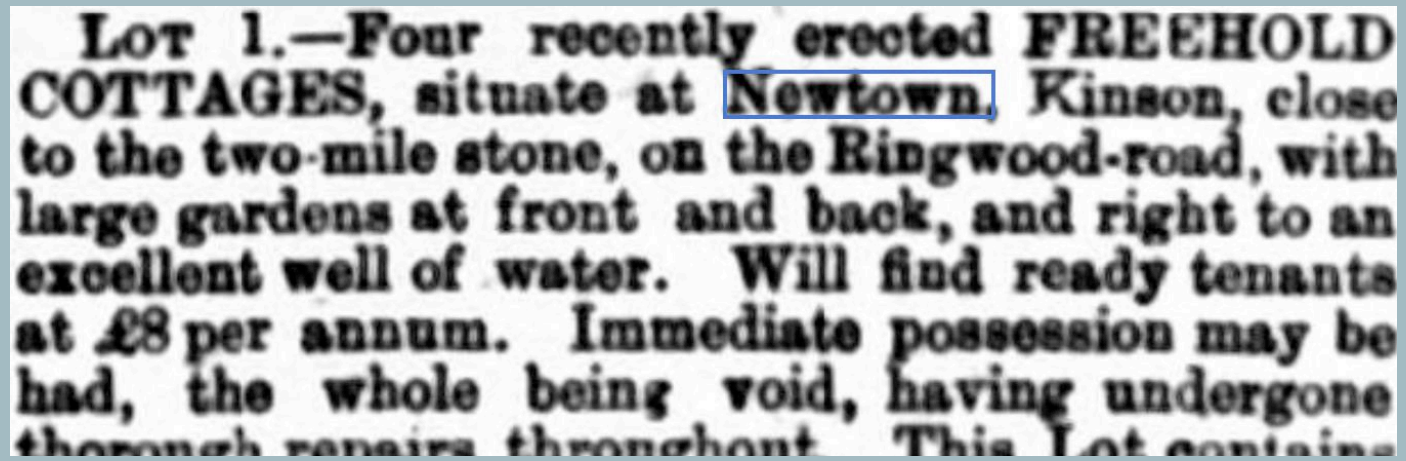
DeezyMatch can be used in the following tasks:

- 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:

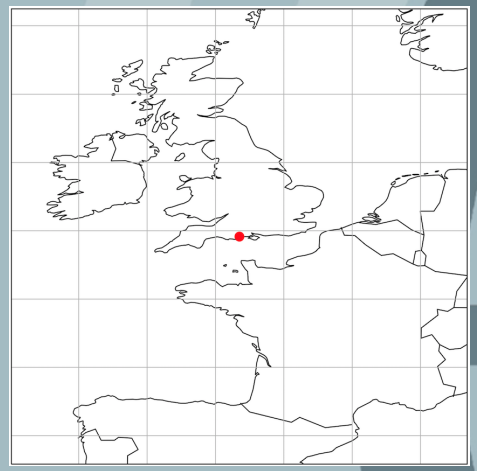
- Record linkage
- Entity linking

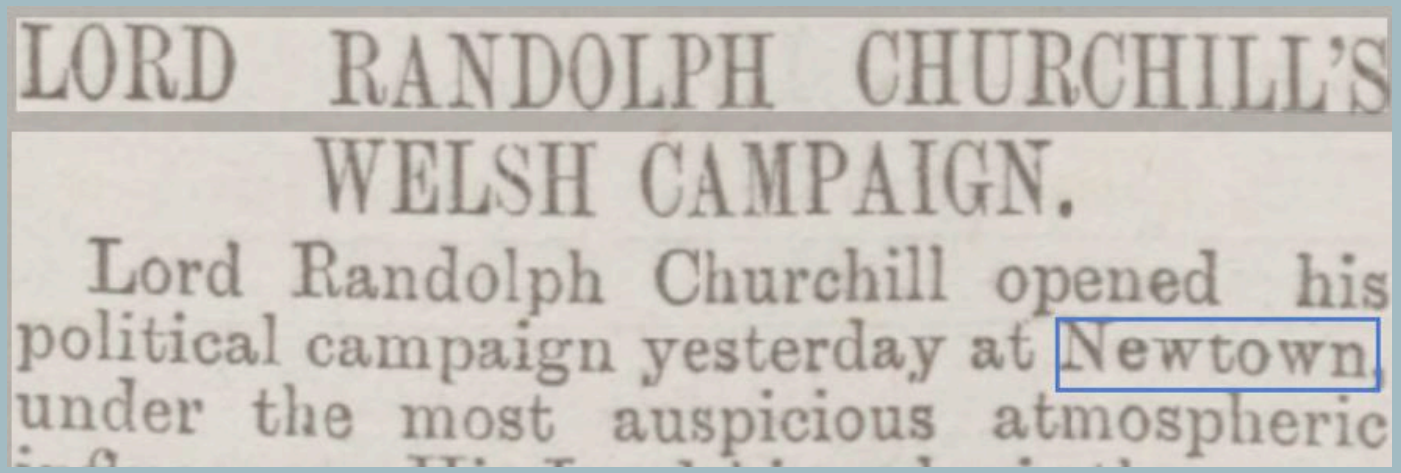
Finding, disambiguating and locating place names in texts (toponym resolution)



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

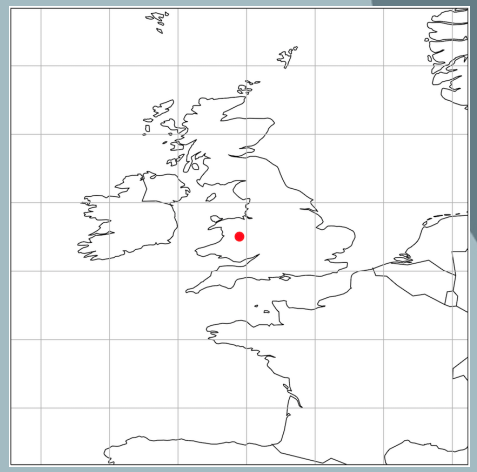




LORD RANDOLPH CHURCHILL'S WELSH CAMPAIGN.

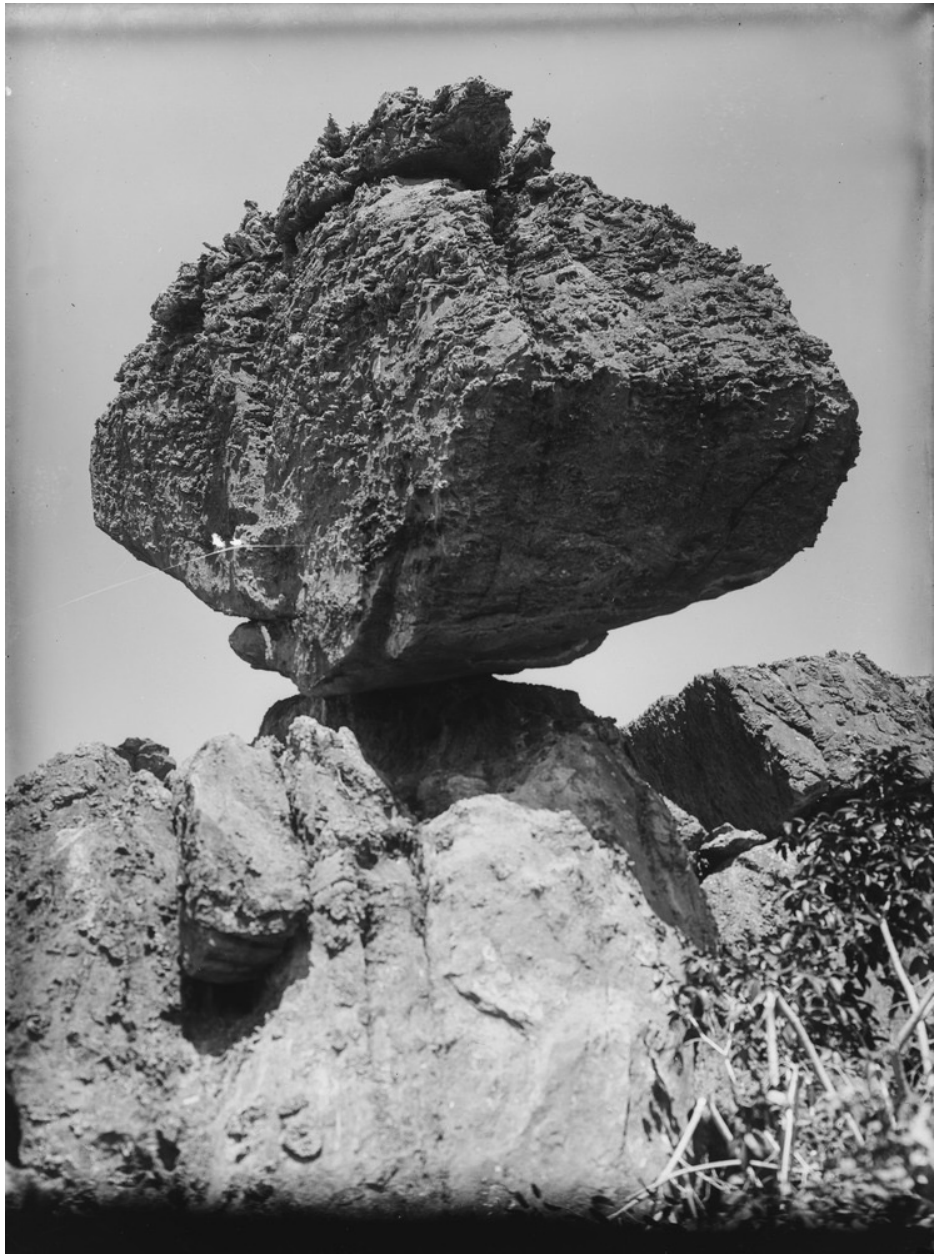
Lord Randolph Churchill opened his political campaign yesterday at Newtown, under the most auspicious atmospheric

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



AI in Collections as Data:

Benefits and **limitations**



Harriett Brims Collection; State Library of Queensland on Flickr Commons

Reflections from *Living with Machines*

It's hard to convey scale and complexity of GLAM collections, the impact of copyright, previous collecting and digitisation decisions

Who's responsible for preparing data for analysis? (How clean is 'clean'?) Who designs and hosts infrastructure to analyse it?

'Infrastructure' means different things in libraries and academia

Publishing reusable datasets and code increases value and impact

Talks, hands-on training and time to experiment increase literacy

Lessons learnt – planning AI in GLAMs

Allow time to learn enough to define goals; plan timelines and skills from there

Talk to as many people as possible to understand challenges and impact on different teams

Think about workflows for data, managing rights, documentation

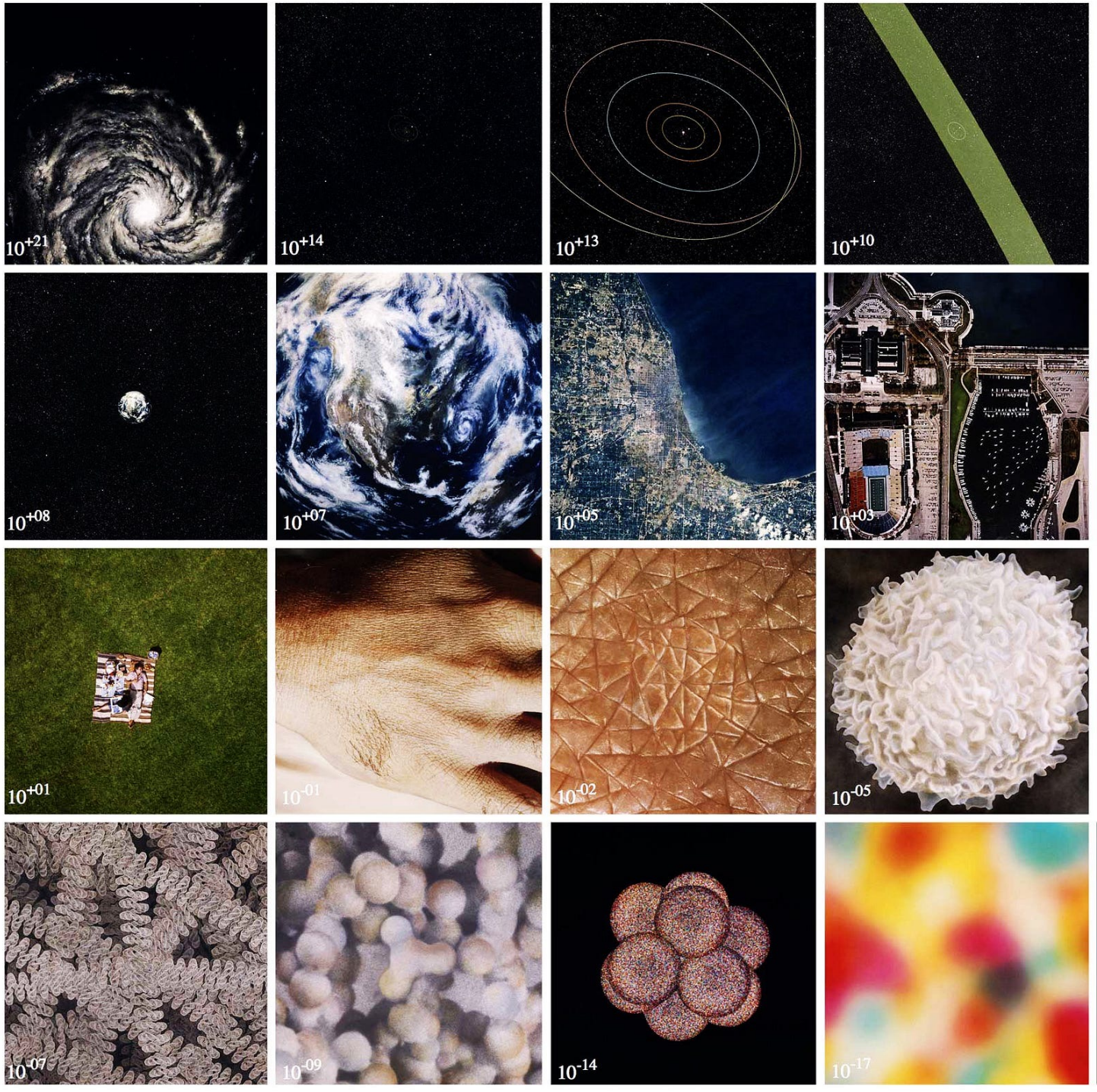
The only constant is change. Data science / machine learning / AI will dramatically each ~~year~~ quarter. Hold infrastructure lightly

Enriched metadata is a challenge for traditional catalogue systems

Digital methods address lines and regions on a page – or entire datasets

Catalogues tend to know about the 'deliverable unit'

Moving between scales is hard



Another challenge: 'meaningful human control at the right stage'

'When you use generative AI [make sure] there are processes for **quality assurance controls** which include an **appropriately trained and qualified person to review your generative AI tool's outputs** and validation of all **decision making** that generative AI outputs have fed into.

...

You must have fully tested the product before deployment, and have robust assurance and regular checks of the live tool in place. Since it is not possible to build models that never produce unwanted or fictitious outputs (i.e. hallucinations), incorporating end-user feedback is vital. Put mechanisms into place that allow **end-users to report content and trigger a human review process**.'

- Generative AI Framework for HMG, UK

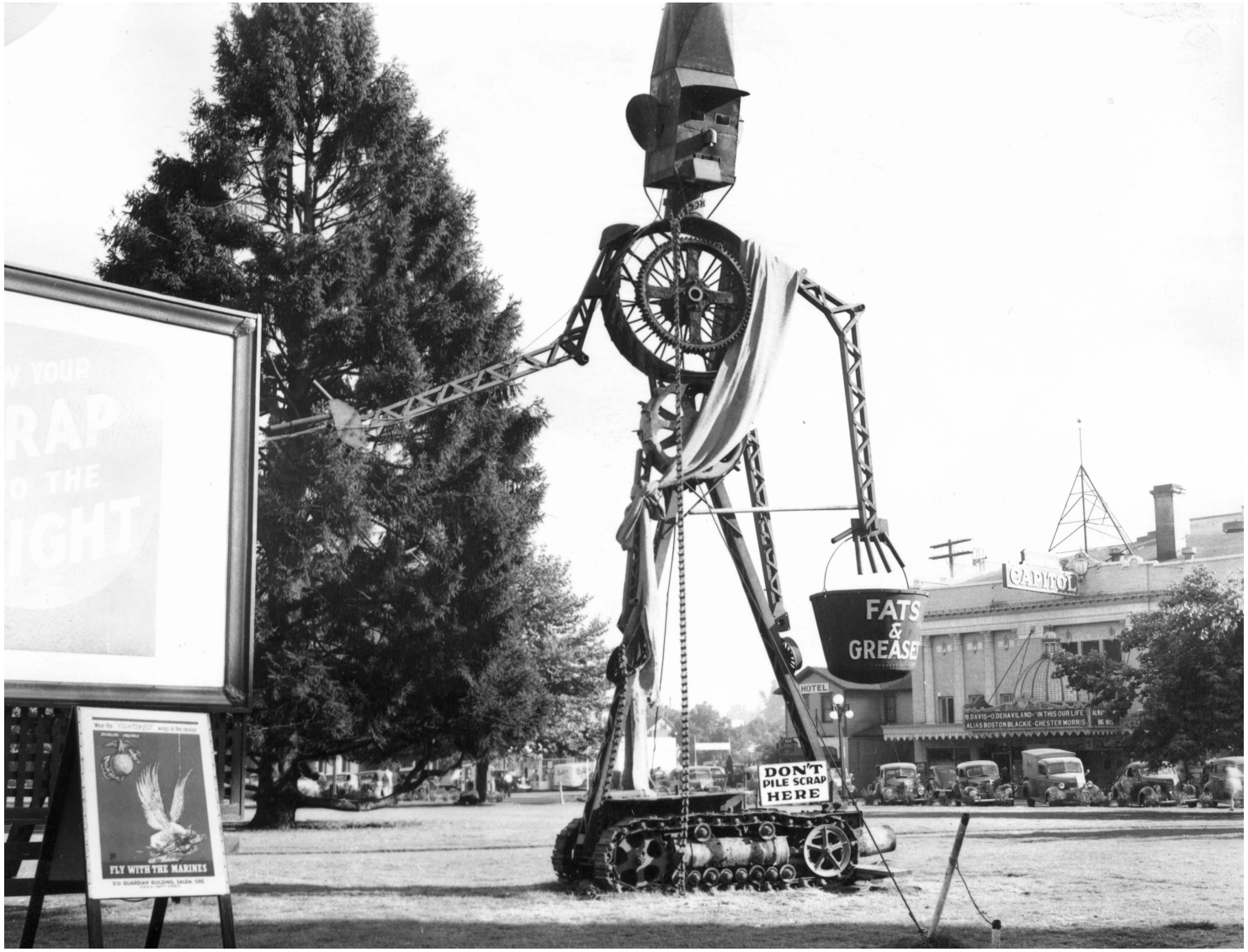
AI isn't great at 'rare' or 'special'

'AI tends to sand away the unusual. It's trained to answer with the most likely answer to your question, which is not necessarily the most correct [or interesting] answer.'

Janelle Shane, https://www.aiweirdness.com/ai-vs-a-giraffe-with-no-spots/

Image: Brights Zoo

AI can get you there - but it can't make meaning



Scrappo, mechanical scrap metal creation, Salem, Oregon, 1942; OSU Special Collections & Archives on Flickr Commons

Thank you!
Questions?

BRITISH LIBRARY

Generative AI Framework for HMG

'ten common principles to guide the safe, responsible and effective use of generative AI in government organisations' published January 18:

- Principle 1: You know what generative AI is and what its limitations are
- Principle 2: You use generative AI lawfully, ethically and responsibly
- Principle 3: You know how to keep generative AI tools secure
- Principle 4: You have meaningful human control at the right stage
- Principle 5: You understand how to manage the full generative AI lifecycle
- Principle 6: You use the right tool for the job
- Principle 7: You are open and collaborative
- Principle 8: You work with commercial colleagues from the start
- Principle 9: You have the skills and expertise needed to build and use generative AI
- Principle 10: You use these principles alongside your organisation’s policies and have the right assurance in place