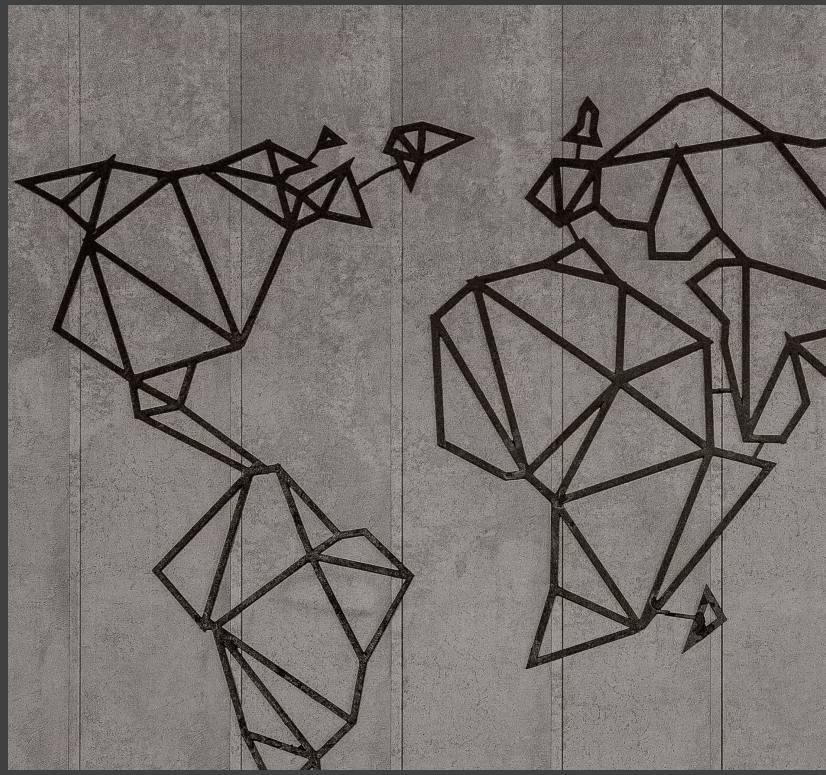
The Alan Turing Institute

Beyond Digital Humanities:
Weaving Humanities, Research Software Engineering and AI

David Beavan
T: @DavidBeavan | M: @davidbeavan@fosstodon.org
University of Georgia (UGA) 2023-03-15

Outline

- Introduction
- Living with Machines
- Highlights
- Challenges
- Future of Digital Research Infrastructure



The Alan Turing Institute

W: turing.ac.uk | T: @turinginst

The Alan Turing Institute

- Established 2015
- National Institute for Data Science and AI
- Over 400 researchers plus many other roles
- All disciplines

Turing Strategic and Project Partners

University partners

– 13 Founding universities

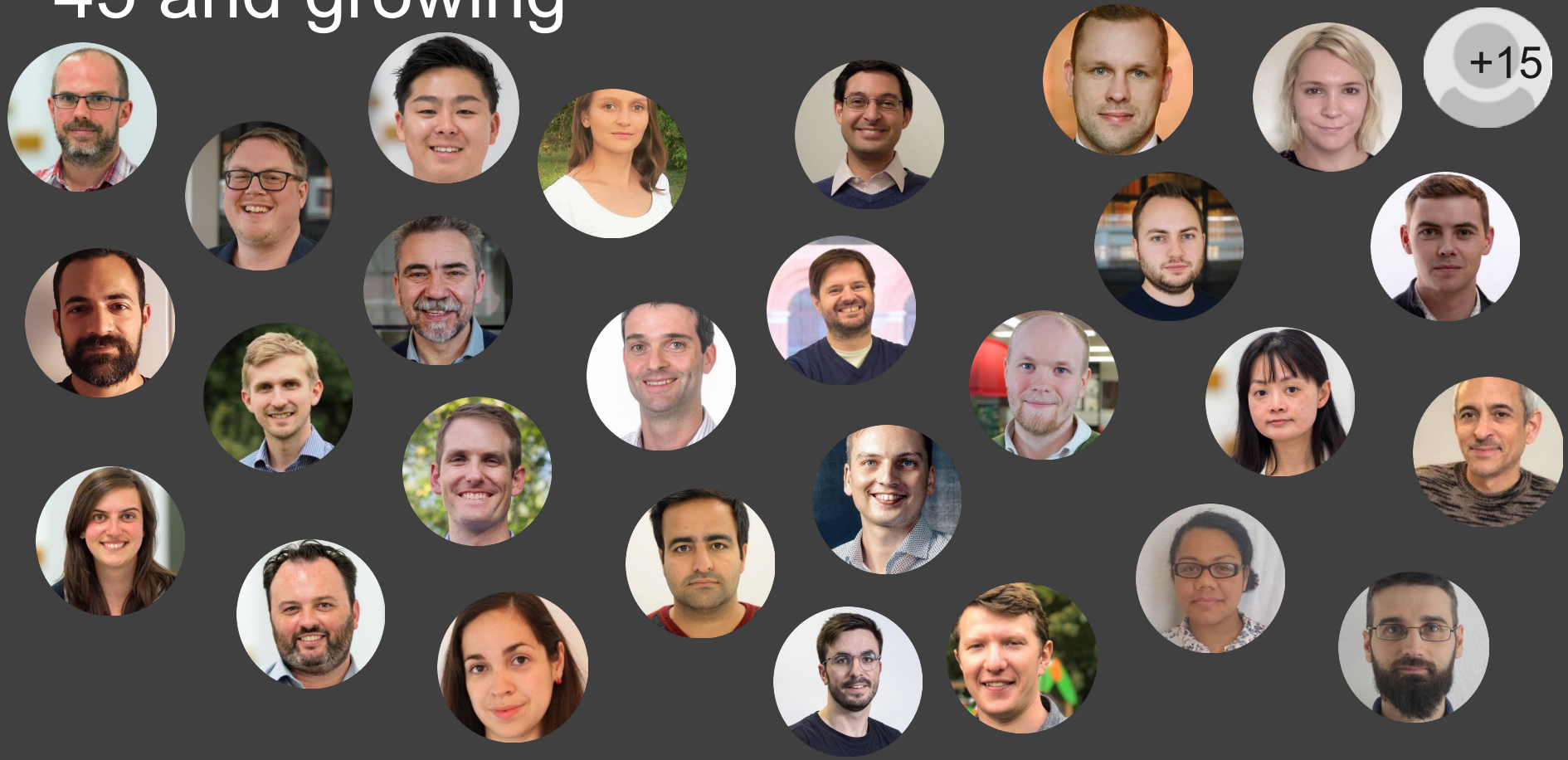
– 25 Additional Network Development Awards

– From 2023 truly national reach

Research Engineering Group (REG)

W: turing.ac.uk/research-engineering | T: @turinghut23

45 and growing



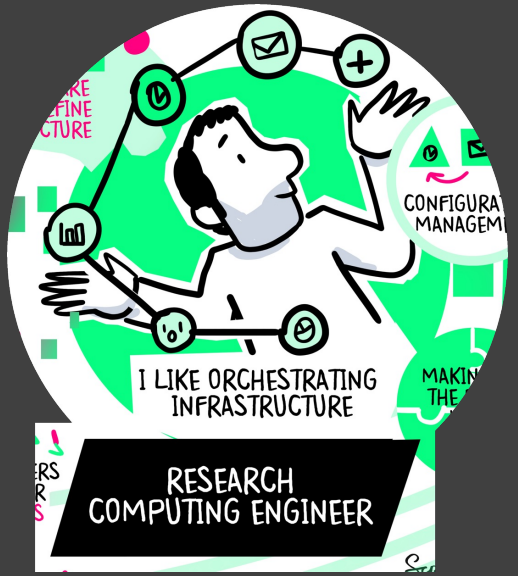


Image credits: Scriberia, created for the Research Engineering Group

Research Software Engineer: The Craftsperson and the Scholar

– James Hetherington

Living with Machines

W: livingwithmachines.ac.uk
T: @livingwmachines

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

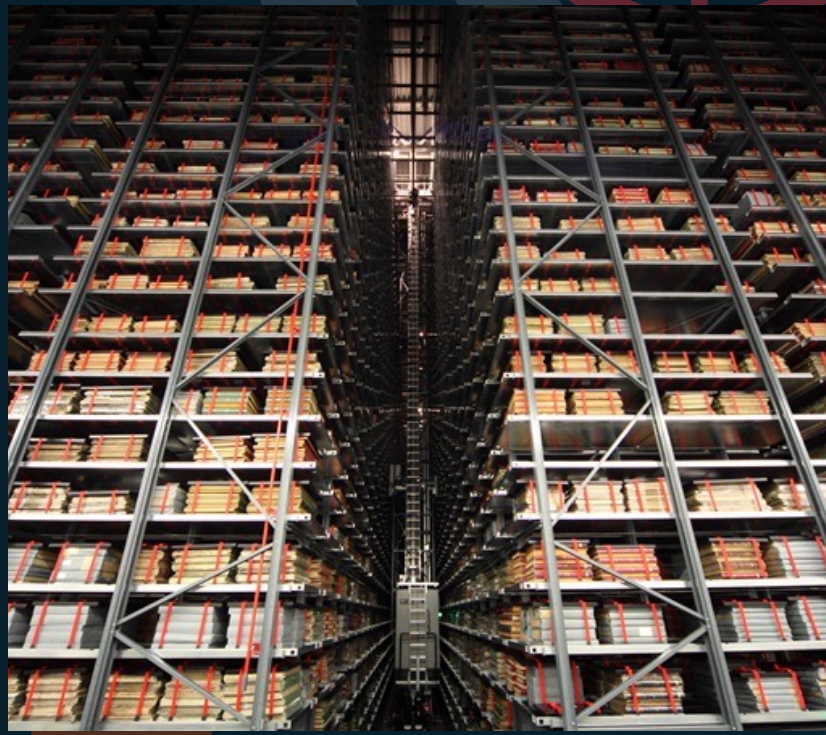
UK Research and Innovation

Living with Machines

- British Industrial Revolution (1780-1920)
- New historical perspectives
- Use digital methods
- New tools and code
- New computational techniques
- Enrich national data holdings
- Advance public awareness

Living with Machines

- Going beyond Digital Humanities
- 5 year project
- 2018-2023
- Arts & Humanities Research Council funded
- £9+ million
- Six partner institutions



Ruth Ahnert

(QMUL)

Read more

David Beavan

(Turing)

Read more

Emma Griffin

(UEA)

Read more

Timothy Hobson

(Turing)

Read more

Jon Lawrence

(Exeter)

Read more

Maja Maricevic

(British Library)

Read more

Barbara McGillivray

(Turing / King's College London)

Read more

Mia Ridge

(British Library)

Read more

Alan Wilson

(Turing)

Read more

STOP PRESS
PLUS
NEW STARTER
JOE PALMER

Claire Austin

Rights Assurance

Read more

Kaspar Beelen

Digital Humanities Senior Research Associate

Read more

Mariona Coll Ardanuy

Computational Linguistics Senior Research Associate

Read more

Léllé Demertzi

Programme Coordinator

Read more

Luke Hare

Research Data Scientist

Read more

Sherman Lo

Research Data Scientist

Read more

Katie McDonough

Senior Research Associate

Read more

Federico Nanni

Senior Research Data Scientist

Read more

Nilo Pedrazzini

Corpus-Based Digital Humanities Research Assistant

Read more

André Piza

Research Project Manager

Read more

Griffith Rees

Research Data Scientist

Read more

Kalle Westerling

Research Software Engineer

Read more

Daniel Wilson

History Senior Research Associate

Read more

Rosie Wood

Research Data Scientist

Read more

Filipe Bento

Research Associate

Read more

Giovanni Colavizza

(Turing)

Read more

Joel Dearden (1979-2020)

Research Software Engineer

Read more

Adam Farquhar

(British Library)

Read more

Rosa Filgueira

(EPCC), Data Architect

Read more

Sarah Gibson

Research Software Engineer

Read more

Lucy Havens

Research Software Engineer Intern

Read more

James Hetherington

(Turing)

Read more

Kasra Hosseini

Research Data Scientist

Read more

Michael Jackson

(EPCC), Software Architect

Read more

Joshua Rhodes

Senior Research Associate

Read more

Yann Ryan

British Library Curator, Digital Newspapers

Read more

Guy Solomon

Research Associate

Read more

Giorgia Tolfo

Data And Content Manager

Read more

Daniel Van Strien

Digital Curator

Read more

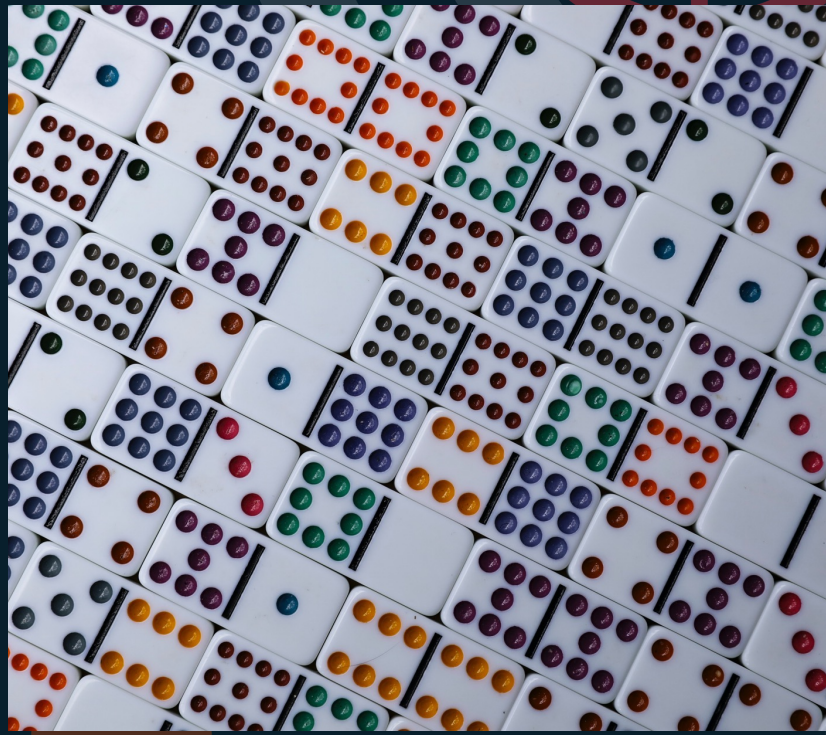
Olivia Vane

Digital Humanities Research Software Engineer

Read more

In numbers

- 109 billion words of historic newspapers
- 129 thousand map sheets
- 180 million household census entries
- plus trade directories, newspaper directories, C19 novels corpus



Language

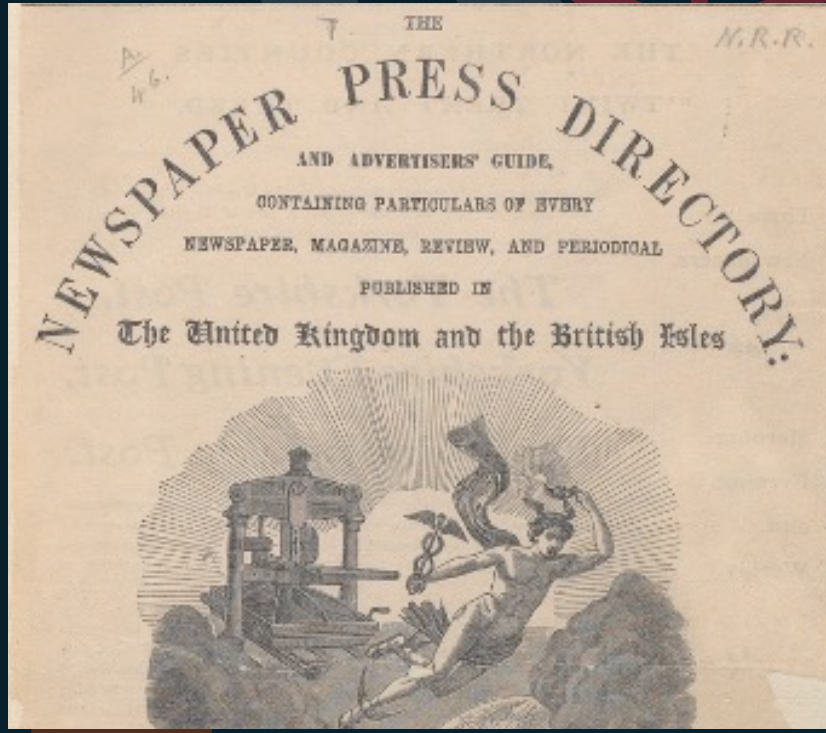
Environmental Scan

Voices in print

- Digitisation selection
- Preservation selection
- Optical character recognition
- Political bias
- Geographical spread
- Price point

Newspaper press directories

- Mitchell's Newspaper Press Directory and Advertisers' Guide
- 1846 -
- £ we digitised, via BL, we would like data to be open



CHATHAM.—(*Kent.*)

Is a parliamentary borough, and river port, on the east bank of the Medway. It is celebrated for its extensive dockyard, naval arsenal, military depots, marine barracks, and Royal Engineers' establishment. The dockyard has been vastly extended.—*Papers for the district*—

CHATHAM NEWS. Friday and Saturday, 1*d.*

Independent.—Established July 2, 1859.

CIRCULATES in Chatham, Brompton, Rochester, Strood, Gillingham, and through the north district of Kent.

ADVOCATES local interests; and unites the news of the district, and several columns of carefully-selected general news, and miscellanies.

PROPRIETORS—Joseph Clayton and Henry Foster.

(Advt., p. 164.)

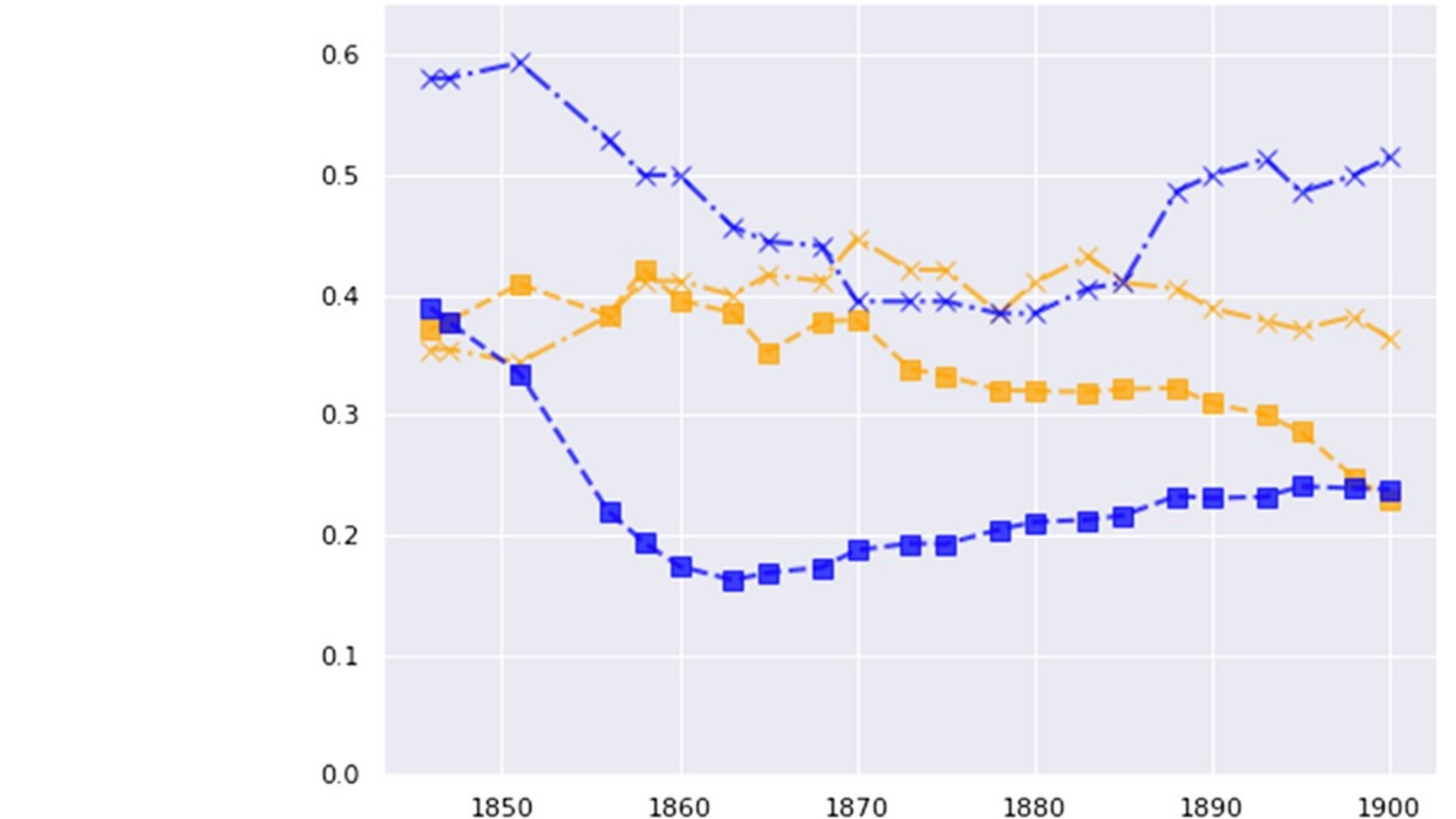
AI methods

- Image segmentation
- OCR (including identification and rescaling of drop caps)
- Text segmentation
- Semantic annotation
- External record linking
- AI deep learning
- Human-in-the-loop training

Structured data

Time

S-TITLE	S-POL	E-PER	year
PORTSMOUTH TIMES AND NAVAL GAZETTE	Liberal	Alfred King	1855
PORTSMOUTH TIMES AND NAVAL GAZETTE	Liberal	Alfred King	1856
PORTSMOUTH TIMES AND NAVAL GAZETTE	Liberal	Alfred King	1857
PORTSMOUTH TIMES	Conservative	Alfred King	1858
BORTSMOUTH TIMES	Conservative	Alfred King	1860
PORTSMOUTH TIMES	Conservative	Richard Holbrook	1861
PORTSMOUTH TIMES	Liberal-Conservative	R . Holbrook	1865
PORTSMOUTH TIMES	Liberal-Conservative	R . Holbrook	1866
PORTSMOUTH TIMES	Liberal-Conservative	NaN	1867
PORTSMOUTH TIMES	Liberal-Conservative	Edward George Holbrook	1868
PORTSMOUTH TIMES	Liberal-Conservative	Edward George Holbrook	1869


0.6
0.5
0.4
0.3
0.2
0.1
0.0
1850
1860
1870
1880
1890
1900

Language of mechanisation

Crowdsourcing

- Diachronic corpus
- Language change over time
- Off the shelf AI methods work on present day English
- Train them on historical data
- Need training data
- Zooniverse

Read the text carefully

Select the option that best matches the use of 'car' in the article. For example, horse-drawn vehicle, railway carriage or motor car.

for Skerton. It is hard lines on Skerton that other parts of the town should have electric cars and that this, one of the most populous and thriving, should be cut off without them. Yet Skerton is not in hopeless case, after all. It has the horse car system, and it is not the Corporation's fault that it cannot be linked up with the electric system this side of the river. What

Place

Toponym disambiguation

Toponym disambiguation

- Resolve OCR issues
- Gather context on place
- Link to gazetteer

Ashton what?

Ashton under Lyne Ashton-under-Lyne Ashtonunder-line ASHTONCNDER-LYNE Ashton-under-lyne Ashtonunder-Lyne ASHTON-UXDER-LYNE Ashton-cnder-Ltne Aditon-under-line Asbtcn-under-Lyne Ashton ASHTON-UNDER-LYNE

Solution

- DEEp fuzZY string MATCHing
- Deezymatch Python library
- Resilient to OCR errors
- Gazetteer with topics seeded from wikipedia

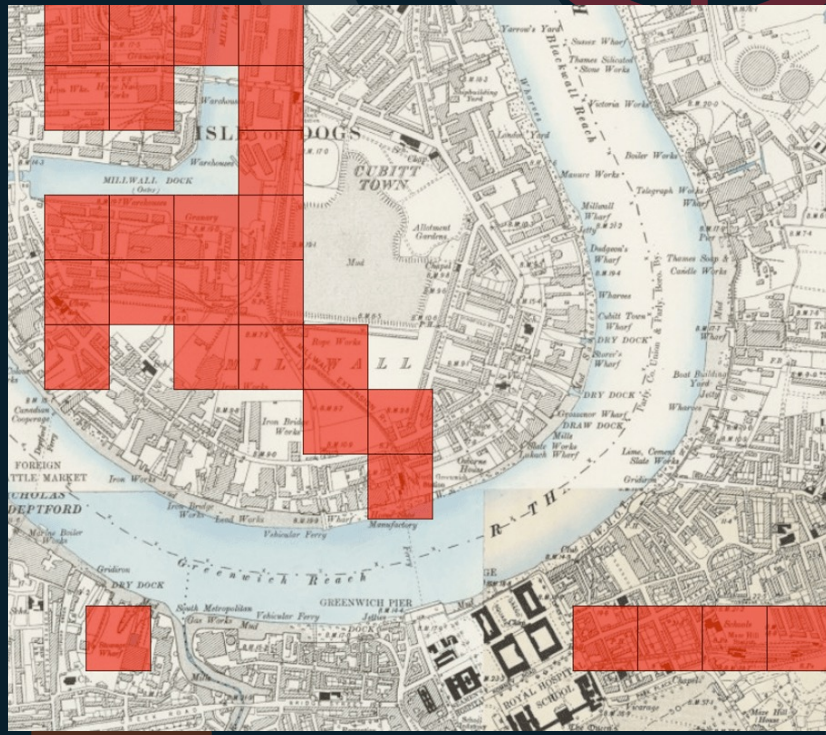
MapReader

No. 6 WARD
SOMERSTOWN
HOUGHTON PLACE
BRIDGEWATER STREET
SIDNEY STREET
CLARENDON STREET
HAMPDEN STREET
PHŒNIX STREET
Coal Depôt
Milk Shed
MIDLAND ROAD
Goods Shed
Goods Shed
BARNBY STREET
GEE STREET
CLARENDON SQUARE
CHARLES STREET
DRUMMOND CRESCENT
LITTLE DRUMMOND STREET
CHAPEL STREET
Urinal
Urinal
Hall
Convent
School
Presb. Church
St. Mary's Church
Railway Clearing Houses
Railway Clearing Houses
Euston Station (Terminus)
Printing Works
Hospital
St. James's
ST. JAMES'S
CARDINGTON STREET
AMPTHILL SQUARE
Signals
Ward Bdy.
Posts
Picture Theatre

1847

MapReader

- AI Computer Vision
- Science to humanities to science
- Annotation and training
- Model inference
- Visualisation



All together now

Challenges and learnings

Explore bias and representativeness

- Question what you work with
- Ethical data science and AI
- Arts & Humanities lead the way
- Correct for bias?
- Or just work with it?
- Map the contours of your data

Be data ready

- Data readiness
- Data wrangling takes time
- Understand intellectual property
- Reproducible research if you can
- Think licences

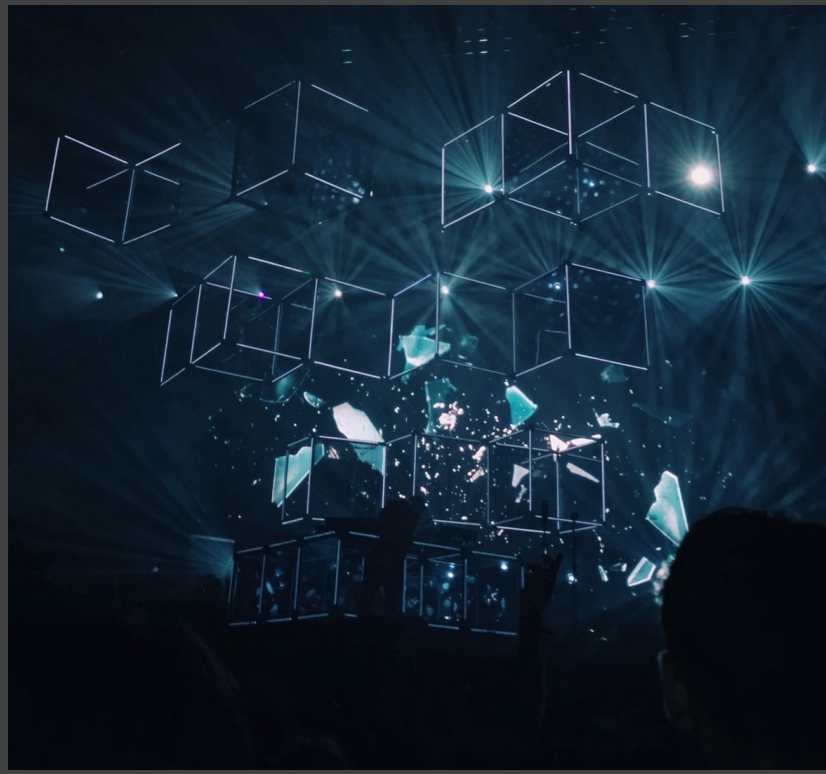
Share what success looks like

- Funders
- Organisations
- Professions
- Career paths
- Values and norms
- Authorship and attribution

Digital Research Infrastructure

Digital Research Infrastructure

- Compute
- Storage
- Skills
- Trusted Research Environments



All disciplines

– Physical sciences
– Biological sciences
– Environment
– Economic and Social
– Arts & Humanities

Challenges

- Investment tied to projects
- Software limited in scope
- Skills in short supply
- Collections under used

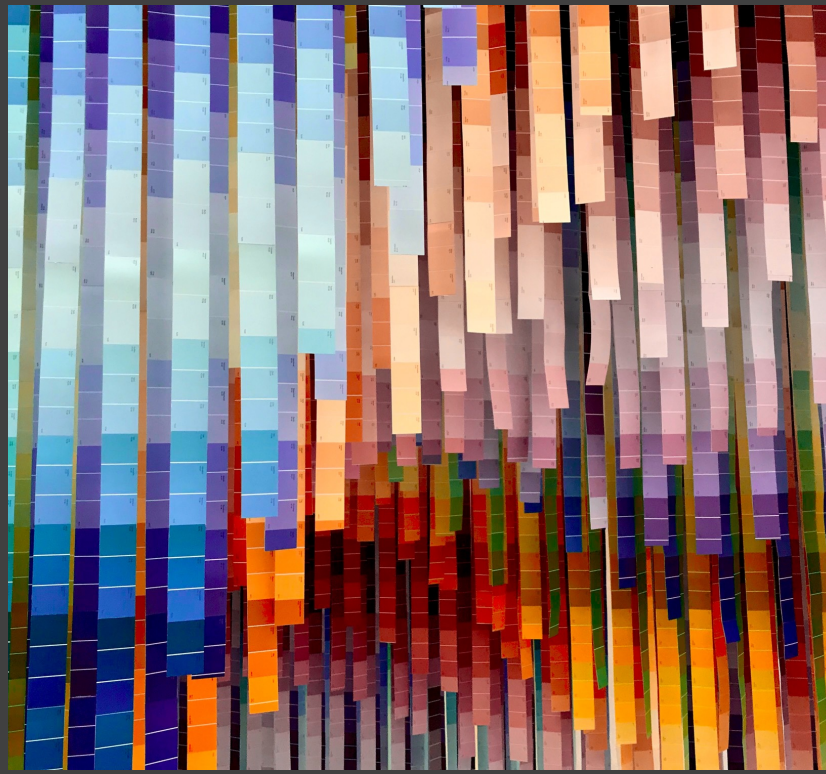
Weaving Humanities, Research Software Engineering and AI

The Alan Turing Institute

Thank you

David Beavan
T: @DavidBeavan | M: @davidbeavan@fosstodon.org
University of Georgia (UGA) 2023-03-15

Image credits



- Photo by Marjan Blan | @marjanblan on Unsplash
- Photo by Giuseppe Ruco on Unsplash
- Photo by Greyson Joralemon on Unsplash
- Photo by fabio on Unsplash
- Photo by Martin Adams on Unsplash
- Photo by Shahadat Rahman on Unsplash
- Photo by Henry & Co. on Unsplash
- Photo by Jené Stephaniuk on Unsplash
- Photo by Daniel K Cheung on Unsplash
- Photo by Mick Haupt on Unsplash
- Photo by Benjamin Davies on Unsplash
- Photo by John Schaidler on Unsplash
- Letter Word Text Taxonomy by Teresa Berndtsson
- Photo by Camylla Battani on Unsplash
- Photo by Adi Goldstein on Unsplash
- Photo by Daniel Fazio on Unsplash
- Photo by Marek Piwnicki on Unsplash
- Maps above reproduced with the permission of the National Library of Scotland

The Alan Turing Institute

Beyond Digital Humanities:
Weaving Humanities, Research Software Engineering and AI

David Beavan
T: @DavidBeavan | M: @davidbeavan@fosstodon.org
University of Georgia (UGA) 2023-03-15