

# Repurpose, remodel and recast: the National Library of Scotland's Data Foundry

**Dr Sarah Ames**

Digital Scholarship Librarian

National Library of Scotland

@semames1 | #NLSdata | sarah.ames@nls.uk

# Today's talk

- Data Foundry and Digital Scholarship at NLS
- Engagement and collaborative activity and outcomes
- The way forward
- Challenges

# The Scottish context

---

## 2015 Open Data Strategy, Scottish Government

### 2018 City Region Deal

- £1.3bn/15 years (government, universities, private & third sector)
- £60 million for data driven innovation
- Skills/Entrepreneurship/Growth

### University of Edinburgh

- Bayes Institute
- Edinburgh Futures Institute
- Creative Informatics
- Centre for Data, Culture and Society

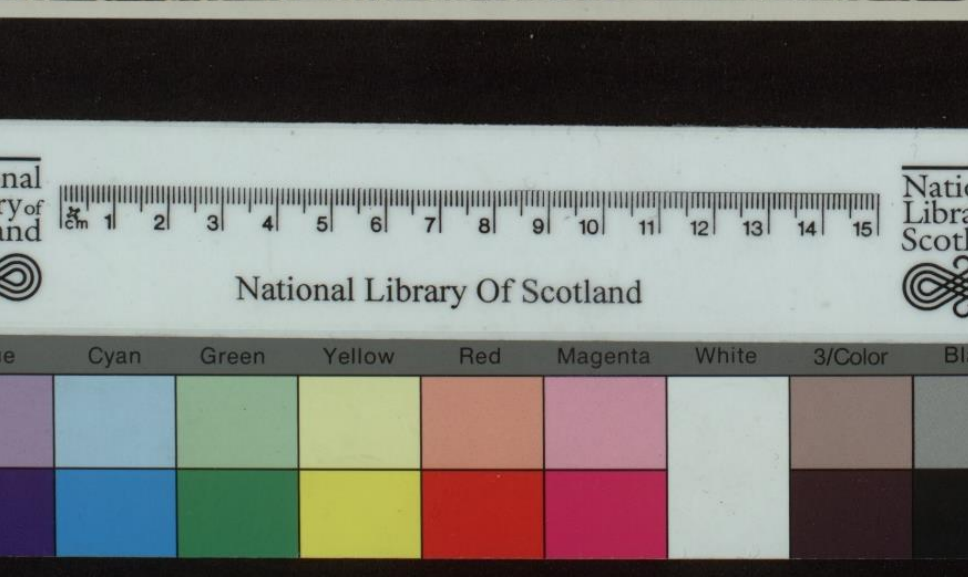
### University of Glasgow

- Information Studies/Digital Humanities

### Universities of Aberdeen, St Andrews, Stirling, Dundee

- Big data/data science

### CodeBase/CodeClan





# The National Library of Scotland

---

Over 31 million items in collection (excluding web archive, and owned-in-perpetuity digital collections)

'One Third Digital' by 2025

In-house mass digitisation programme

Strategy 2020-2025, 'Reaching People': strong focus on engaging with new audiences through the collections

# Digital Scholarship Service



**ENCOURAGE, ENABLE &  
SUPPORT USE OF  
COMPUTATIONAL  
RESEARCH METHODS  
WITH THE COLLECTIONS**



**ENSURE THAT THE  
COLLECTIONS ARE USED  
TO THEIR FULL POTENTIAL**



**ESTABLISH A LIBRARY  
CULTURE WHICH  
UNDERSTANDS DIGITAL  
SCHOLARSHIP**



**PRACTISE AND PROMOTE  
TRANSPARENCY IN OUR  
DATA CREATION  
PROCESSES**



**ANTICIPATE THE FUTURE  
OF RESEARCH**

# Data Foundry



# No-nonsense data

# Identifying user needs

Pro	Intermediate	Beginner
All the tech skills!	Limited tech skills	No tech skills
Will find a way to get the data no matter how presented	Understands value of different formats and approaches for research questions: theoretical rather than practical understanding	Wants to use online tools to explore datasets
But – has expectations of standards (where they exist) & consistency	Wants to get hold of the data easily to check what's there	Just wants the text
	Likely to employ an RA to do the work	

Plus broader audience includes other libraries (standards, presentation of data etc)

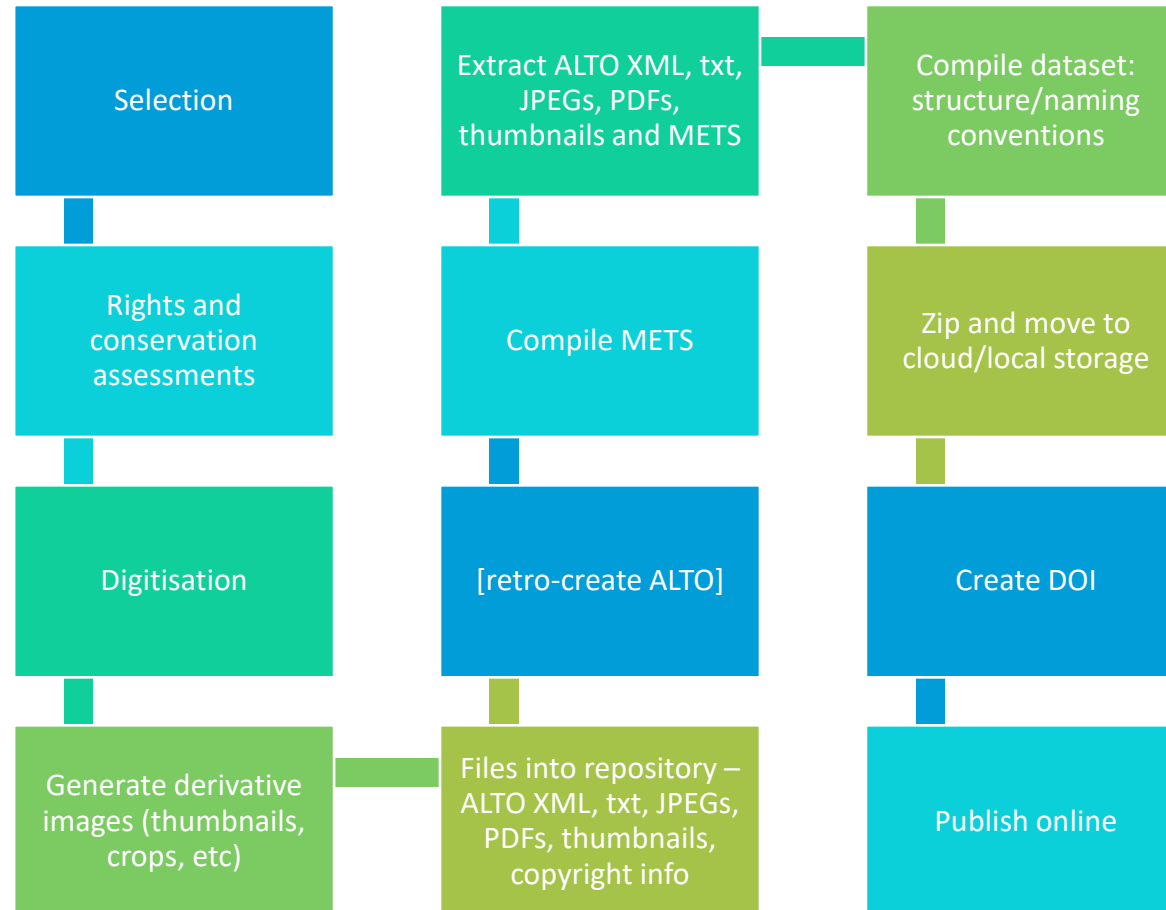


# Identifying user needs

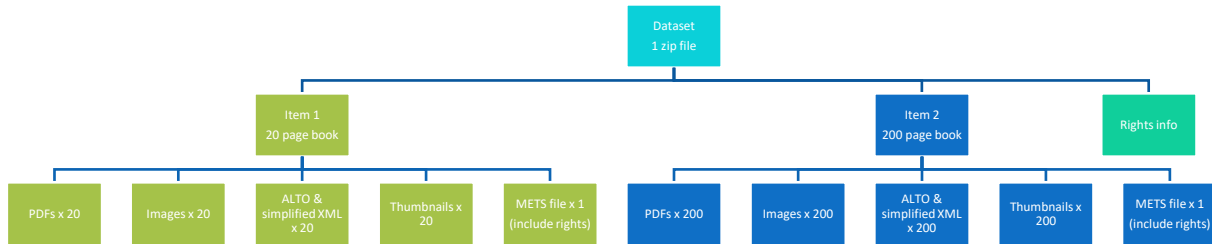
Pro	Intermediate	Beginner
All the tech skills!	Limited tech skills	No tech skills
Will find a way to get the data no matter how presented	Understands value of different formats and approaches for research questions: theoretical rather than practical understanding	Wants to use online tools to explore datasets
But – has expectations of standards (where they exist) & consistency	Wants to get hold of the data easily to check what's there	Just wants the text
	Likely to employ an RA to do the work	

Plus broader audience includes other libraries (standards, presentation of data etc)

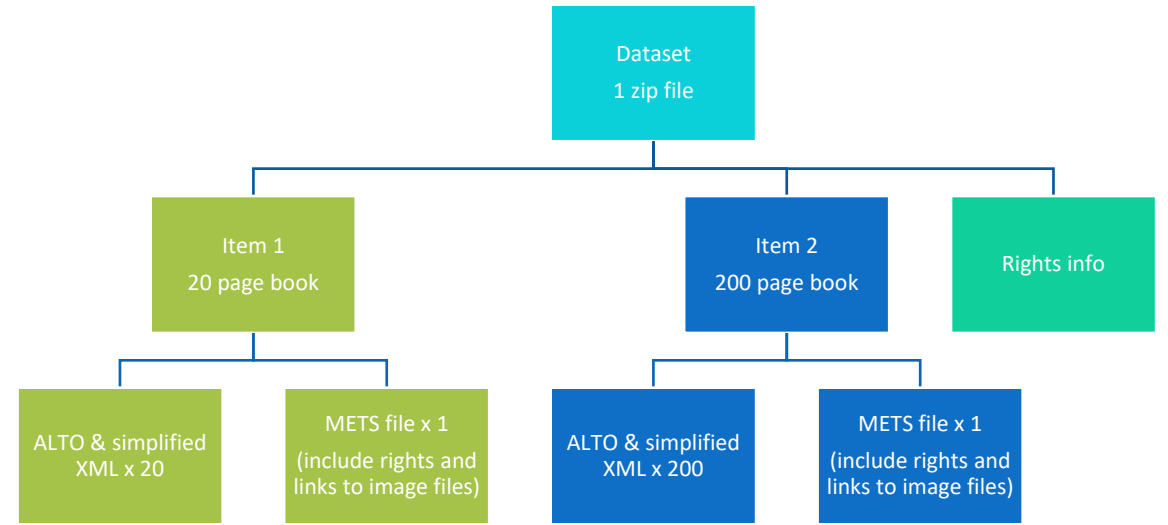
# Changing processes: digitisation to data



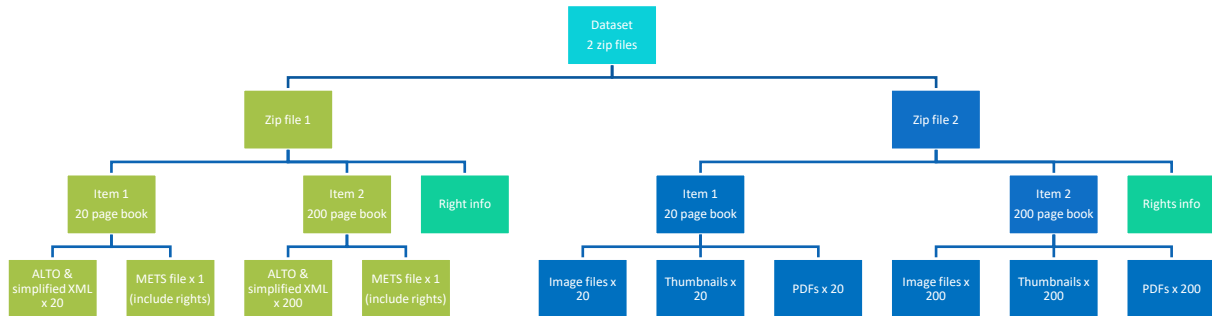
A



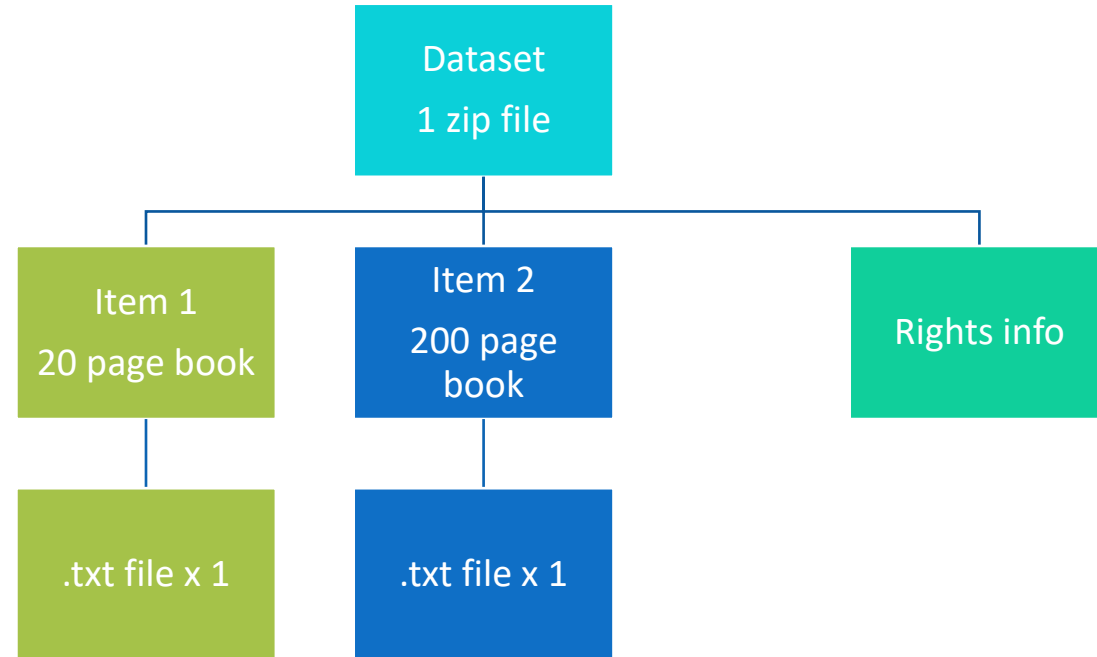
B



C



D



# Making decisions

- Standards:
  - METS/ALTO and Plain text
  - MARC/Dublin Core
  - Tiered downloads
- Image sizes/quality to include
- Storage (local/cloud)
- What metadata to include and what is available
- How to be transparent: gathering and presenting dataset context
- Now moving on to more metadata collections, web archive data, spatial data

# Data Foundry

Data collections from the National Library of Scotland

<https://data.nls.uk/>

# Our principles

## National Library of Scotland data



### Open

The National Library of Scotland publishes data openly and in re-useable formats.



### Transparent

We take the provenance of our data seriously, and are open about how and why it has been produced.



### Practical

We present datasets in a variety of file formats to ensure that they are as accessible as possible.

# A whole-Library effort

Developers

Curators

Metadata

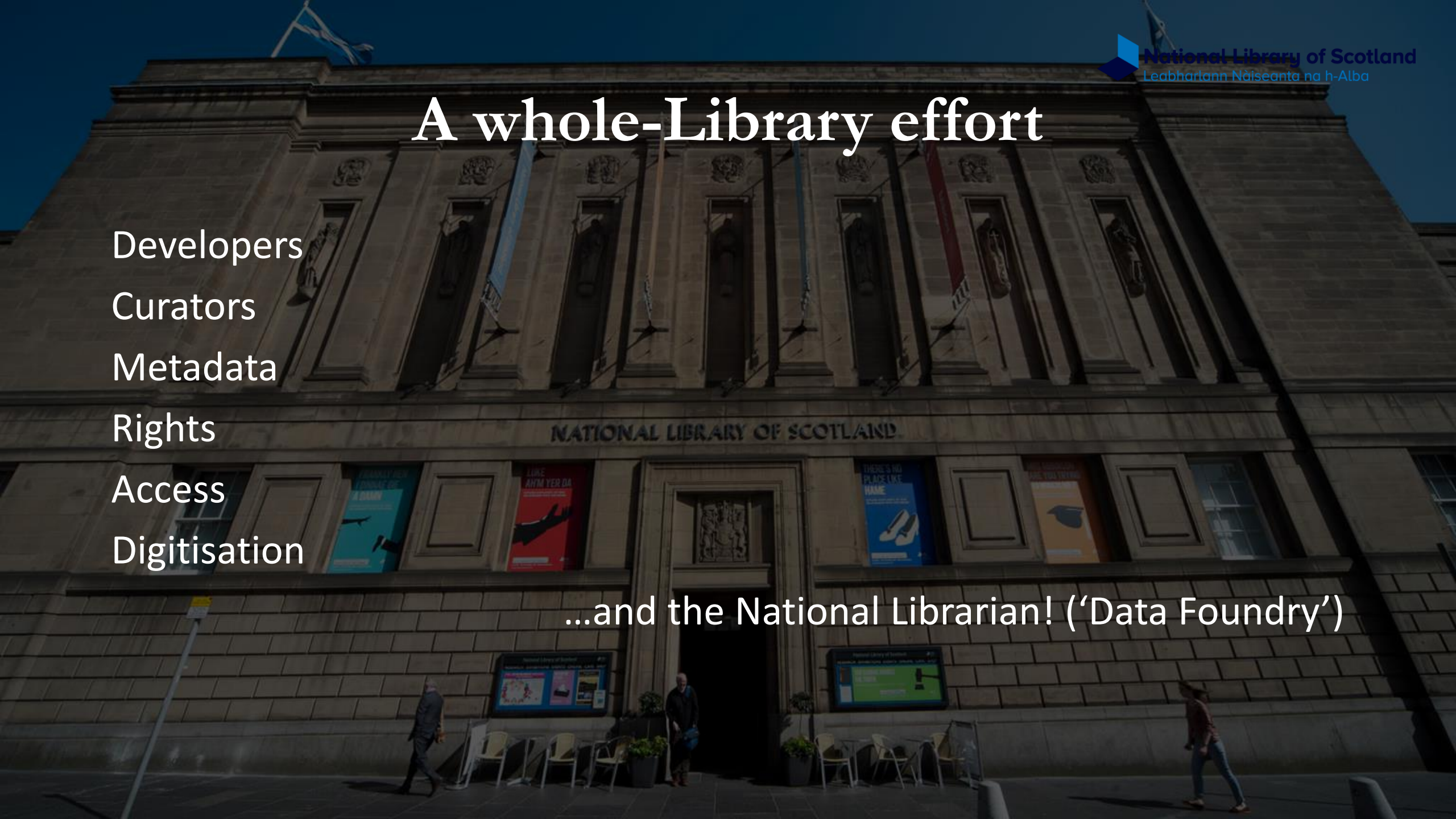
Rights

Access

Digitisation

NATIONAL LIBRARY OF SCOTLAND

...and the National Librarian! (“Data Foundry”)

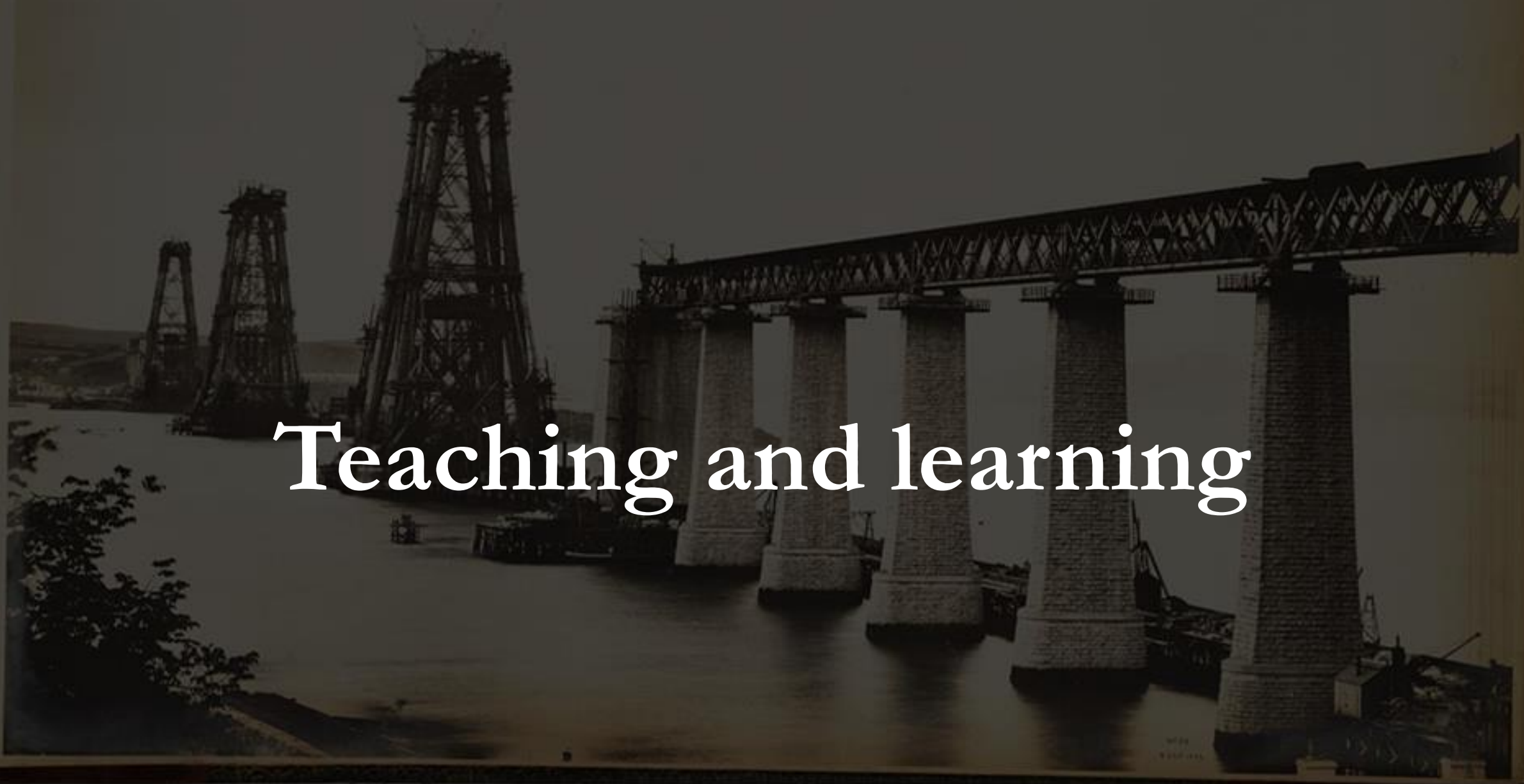




# Engagement activity and outcomes



# Teaching and learning



1870 - 1900  
*Birth of poster  
& Art Nouveau*

*for roughs*  
1900 - 1914  
The New Century  
& Early Modernism

1914 - 1919  
PROPAGANDA

1919 - 1950  
MODERNISM  
& ART DECO

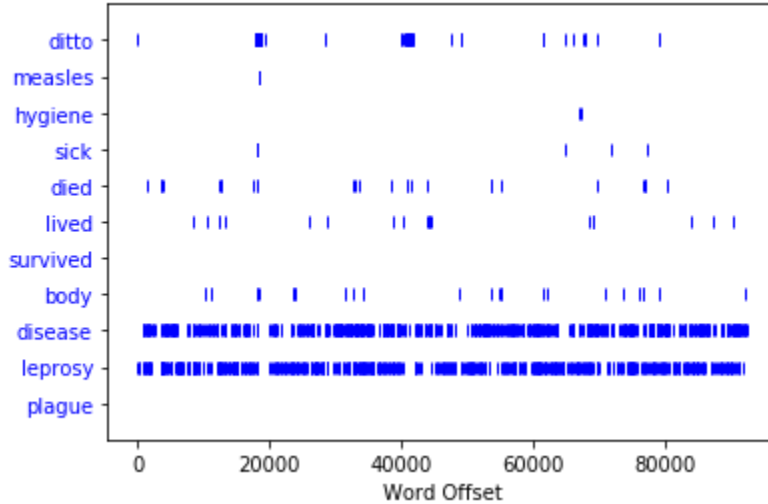
1950 - 1972

1972 - 1990  
The 1980s  
Post Modernism

GLOBAL  
POSTER  
COLLECTION

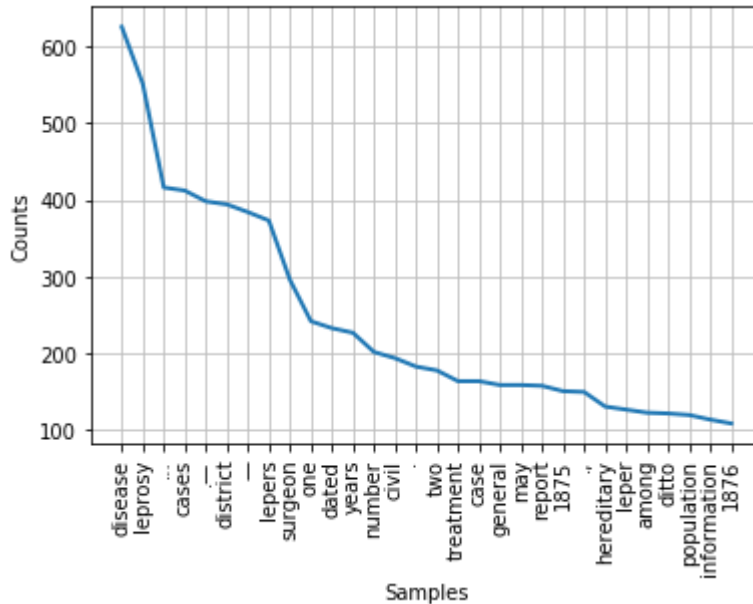


Lexical Dispersion Plot



# TDM Carpentry

- A Medical History of British India dataset
- Dr Bea Alex (University of Edinburgh)



## 2.2 Downloading and Processing Data

### For a Single Document

Instead of providing a string of text we can download one from an corpus. Here we will, load a file from your local machine, tokenise and lowercase it.

Firstly we download a data set (and make a note of where it is saved). We will use the Medical History of British India collection provided by the National Library of Scotland as an example:

<https://data.nls.uk/data/digitised-collections/a-medical-history-of-british-india/>

We can use the `open()` method to open a file in this collection. You need to specify the path to a file in the downloaded collection (this will be different to the one below depending on where you saved it) and the mode of opening it ('r' for read).

The `read()` method is used to read the file. It is then stored as a string variable called `india_raw`.

We then tokenize it as above and normalize it into lowercase. We can check it has worked by printing out a slice of the list `lower_india_tokens`.

```
[19]: 1 file = open('../Downloads/nls-text-indiaPapers/74457530.txt', 'r')
      2 india_raw = file.read()
      3 india_tokens = word_tokenize(india_raw)
      4 lower_india_tokens = [word.lower() for word in india_tokens]
      5 lower_india_tokens[0:10]
```

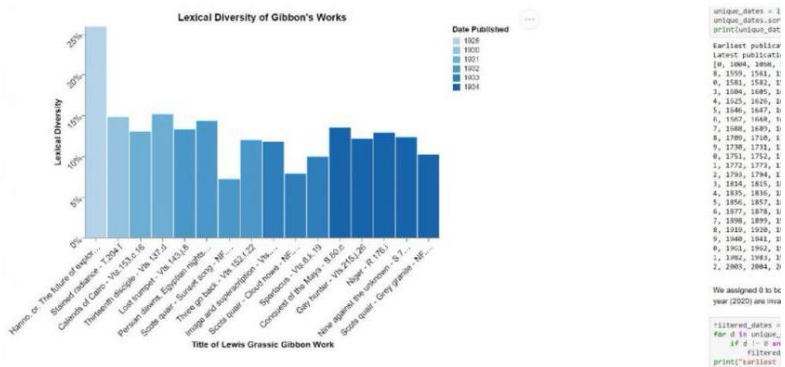
```
: [19]: ['no', '.', '1111', '(', 'sanitary', ')', ',', 'dated', 'ootacamund', ',', '']
```

# Jupyter Notebooks

# Jupyter Notebooks

Initial, exploratory analysis of some of the Library's datasets. No prior programming access and use these Notebooks.

- Series of Jupyter notebooks exploring some of the text and metadata datasets
- For students, learners, non-coders
- By @lucy\_havens



Exploring Lewis Grassic Gibbon First Editions

```
unique_dates = 1
unique_dates.sum()
print(unique_dates)

earliest_publication
latest_publication
[6, 1804, 1804, 1, 8, 1559, 1561, 1, 6, 1581, 1582, 1, 7, 1604, 1605, 1, 4, 1625, 1625, 1, 5, 1640, 1647, 1, 6, 1657, 1668, 1, 7, 1668, 1689, 1, 8, 1700, 1719, 1, 9, 1750, 1751, 1, 8, 1751, 1752, 1, 1, 1752, 1779, 1, 2, 1793, 1794, 1, 3, 1834, 1835, 1, 4, 1835, 1836, 1, 5, 1856, 1857, 1, 6, 1857, 1876, 1, 7, 1898, 1899, 1, 8, 1920, 1920, 1, 9, 1940, 1943, 1, 8, 1961, 1962, 1, 1, 1982, 1983, 1, 2, 2003, 2004, 2]

We assigned 0 to be year (2020) are meta

**filtered_dates**
for of the unique,
if of 0:
    filtered
print(filtered)
```

Expl

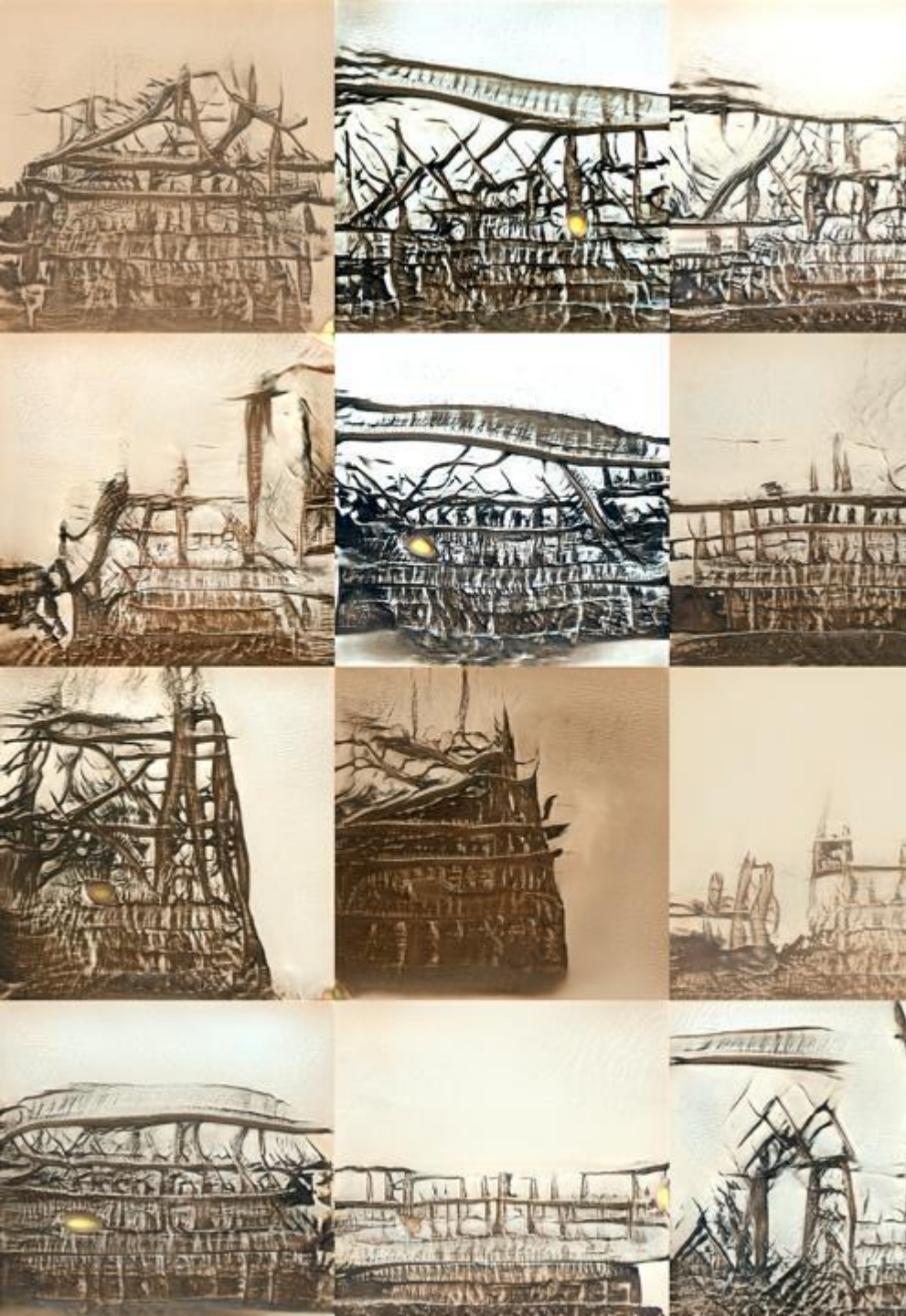
# Creative partnerships

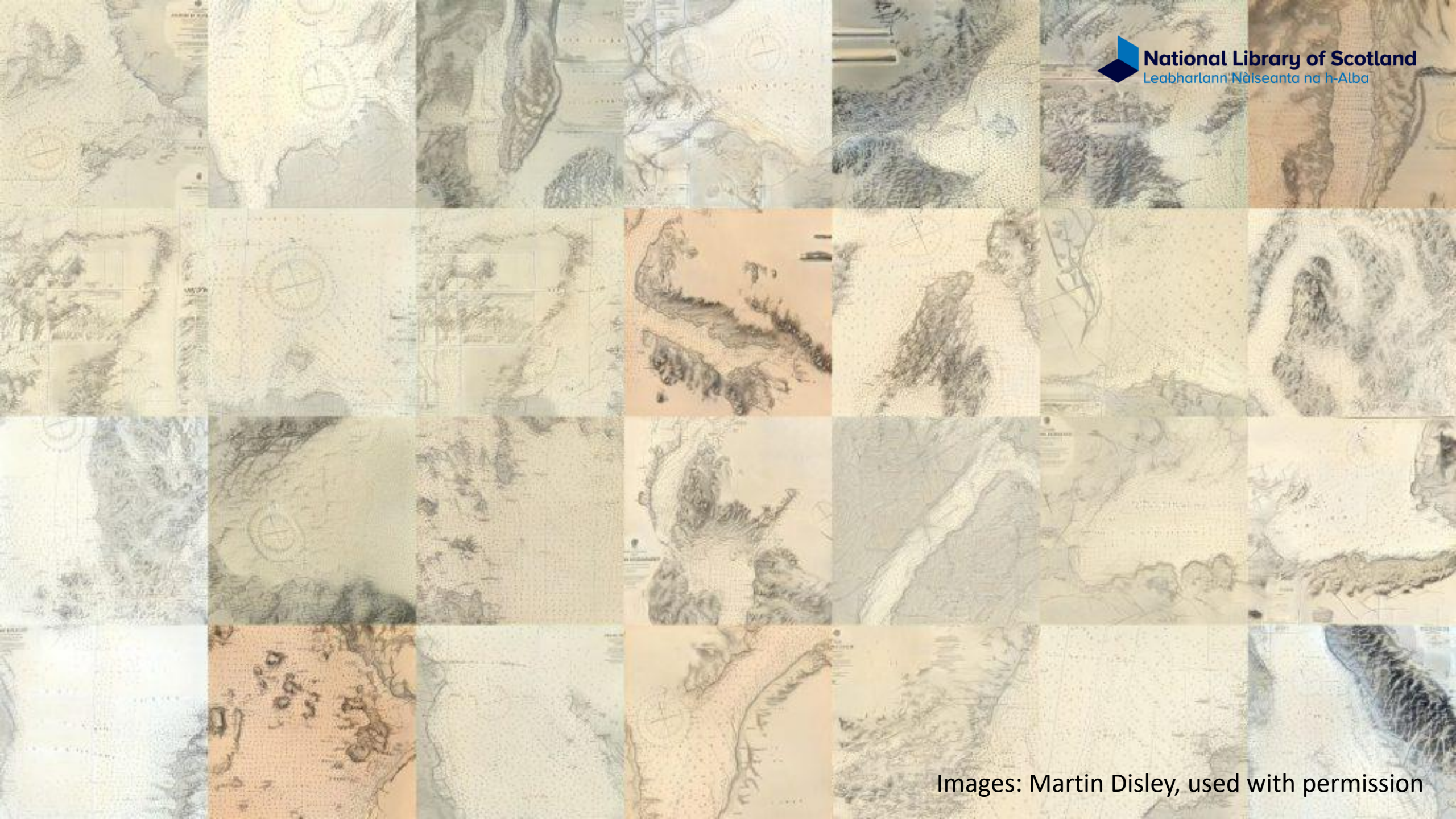


# the dataset is not the map is not the territory

---

- Martin Disley @martin\_disley
- Funded by Creative Informatics (University of Edinburgh)
- Using GAN techniques with the collections
- <https://data.nls.uk/projects/artist-in-residence/>





# Is it true? The post-truth archive factory

---

- Marion Carré @Ma\_rionC
- Collaboration with Goethe-Institut Glasgow, Alliance Française Glasgow and Institut Français d’Ecosse
- Challenge: how does AI open up new ways of interacting with library and archival collections and what are the challenges and dangers of using this technology in archival research?
- Will you archive ‘fact’ or fake news?
- Take part: <https://data.nls.uk/projects/artist-in-residence-marion-carre/>



...

How do we construct an archive? How do we construct truth? Using Artificial Intelligence to create a fake archive, our Artist in Residence @Ma\_rionC asks these questions using #NLSdata in her new work, Selective Memories.



2:55 PM · Nov 11, 2021 · Twitter Web App





# Research collaborations



T H E

# *Farmer's Son*

O R, T H E

## Unfortunate Lovers.

TO WHICH ARE ADDED,

THE CAPTAIN OF LOVE.

MY MOTHER DID SO BEFORE ME.

JUST THE THING.

THE WILD ROVER.



## Annual fellowship 20-21

---

- Dr Giles Bergel (University of Oxford)
- VGG computer vision tools
- Chapbooks dataset
- Exploring what can be learnt from their illustrations about chapbooks' origins; about relationships between chapbook printers, publishers and distributors; and about the type and range of imagery available to their readers.

<https://data.nls.uk/projects/the-national-librarians-research-fellowship-in-digital-scholarship>

# Annual fellowship 21-22

---

- Dr Rosa Filgueira (St Andrews University)
  - AI toolkit for the collections: bringing AI tools to those who can't code
  - Initially with the Encyclopaedia Britannica dataset

<https://data.nls.uk/projects/the-national-librarians-research-fellowship-in-digital-scholarship-2021-22/>

# Annual fellowship 22-23

- Dr Gustavo Candela (University of Alicante)
  - Exploring the Semantic Web and Wikidata with Library metadata collections

<https://data.nls.uk/projects/the-national-librarians-research-fellowship-in-digital-scholarship-2022-23/>



# PhD studentships



## **Joe Nockels**

'Adopting Transkribus in the National Library of Scotland: Understanding how handwritten text recognition will change management and use of digitised manuscripts'.

Supervisors: Professor Melissa Terras (University of Edinburgh), Dr Paul Gooding (University of Glasgow), Dr Sarah Ames and Stephen Rigden (National Library of Scotland).

Funder: Scottish Graduate School for Arts and Humanities, AHRC Collaborative Doctoral Award

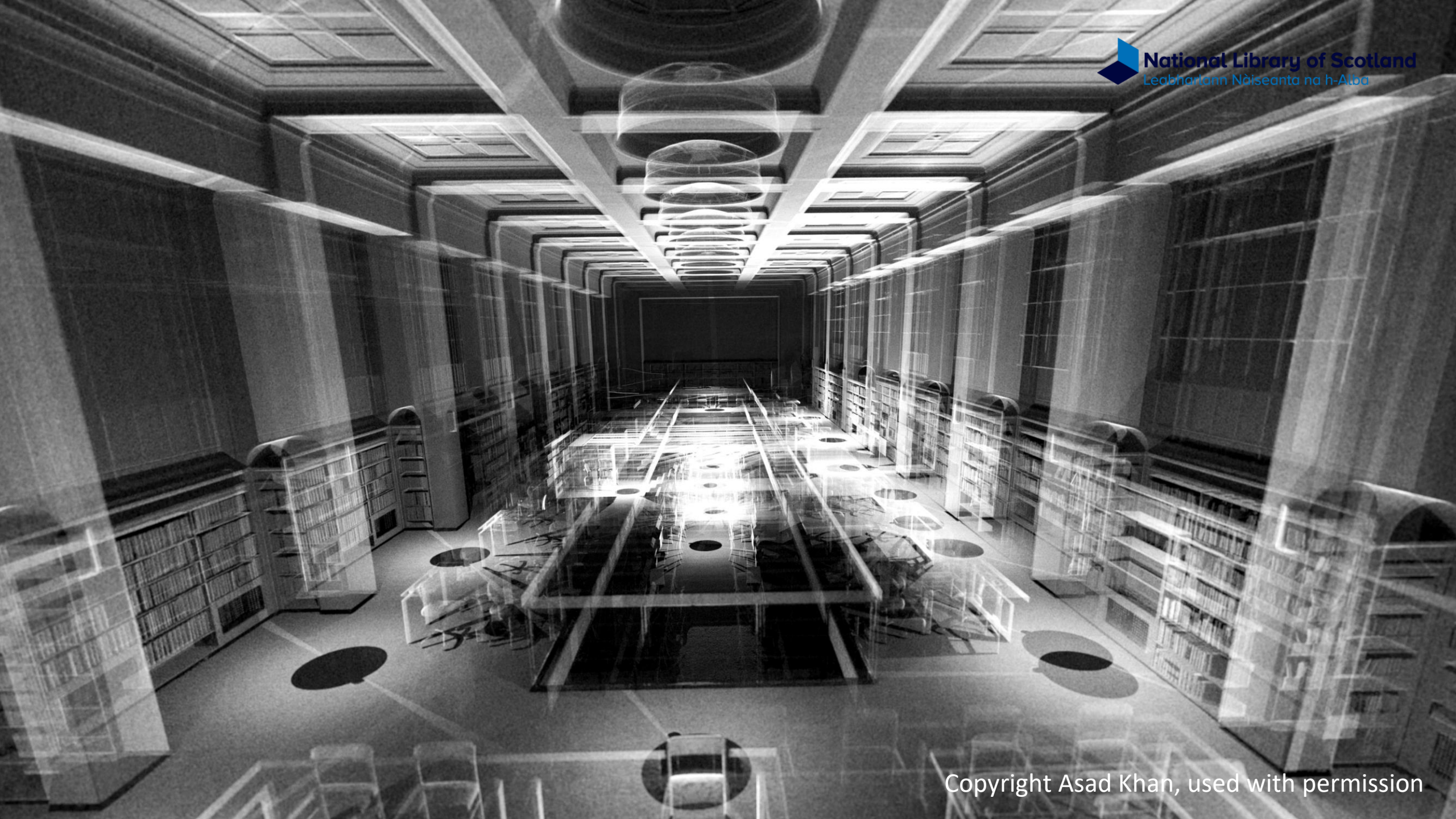


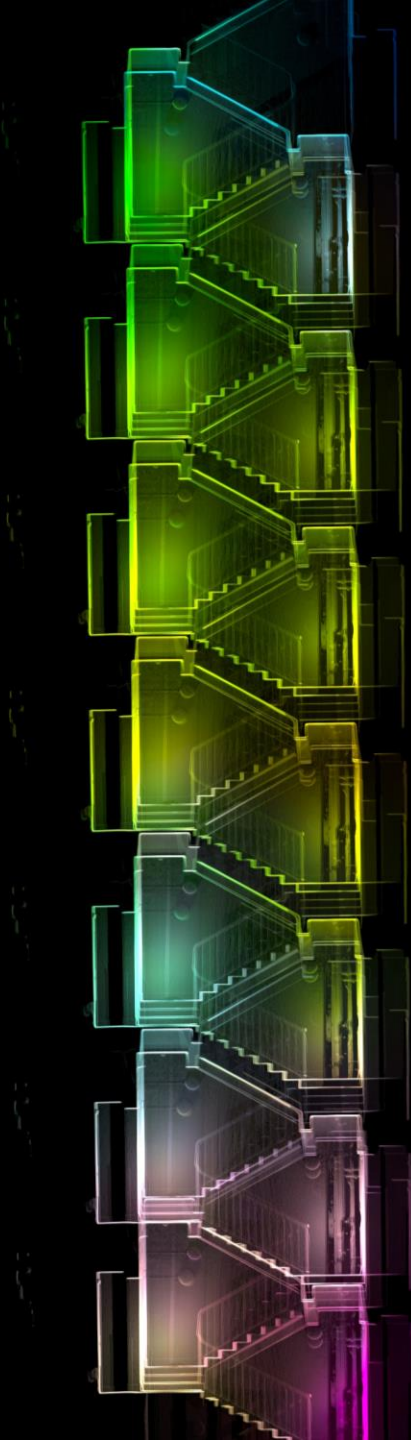
## **Ash Charlton**

'Slavery and Race in the Encyclopaedia Britannica (1768-1860): A Text Mining Approach'.

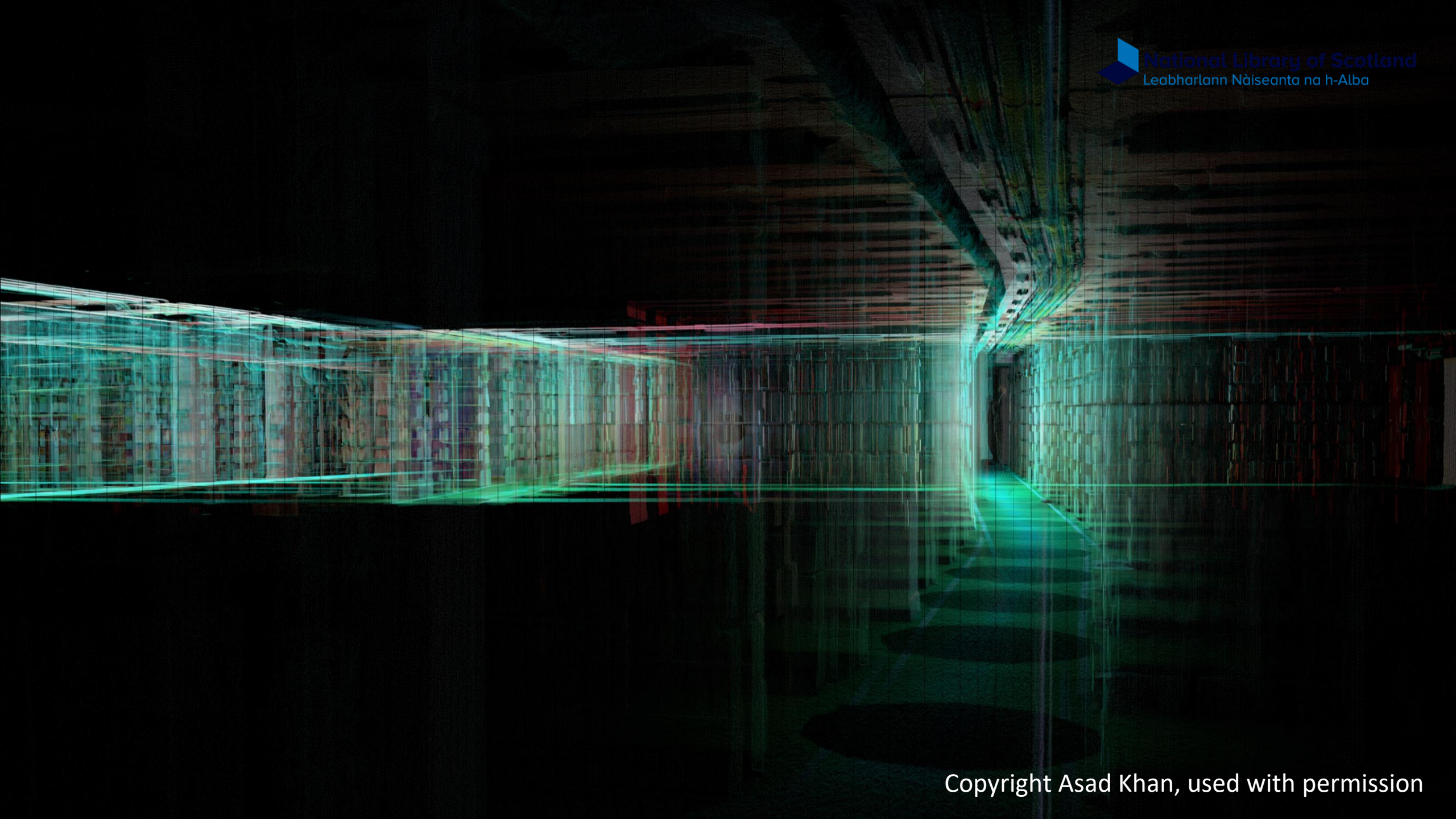
Supervisors: Professor Melissa Terras and Professor Diana Paton (University of Edinburgh), Dr Sarah Ames and Robert Betteridge (National Library of Scotland)

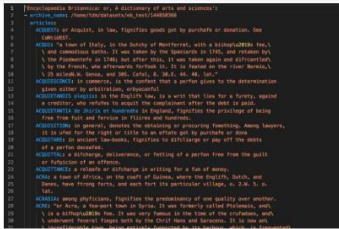
Funder: Scottish Graduate School for Arts and Humanities, AHRC Collaborative Doctoral Award



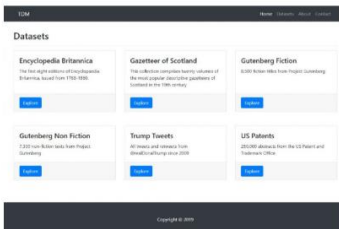




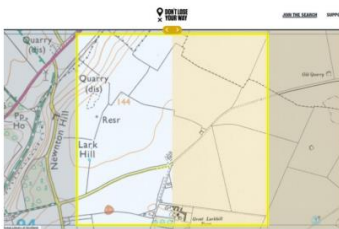




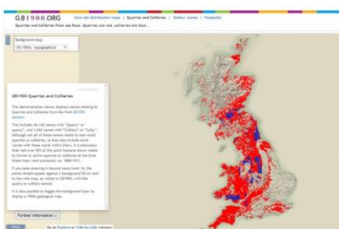
High Performance Computing meets  
Encyclopaedia Britannica



Text and data mining platform



Finding lost footpaths using GB1900



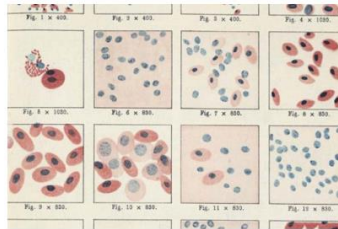
Mapping quarries and collieries



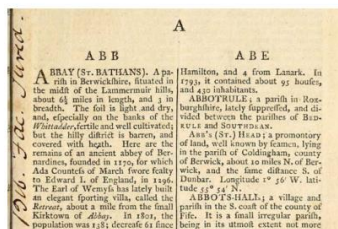
Data visualisation student projects



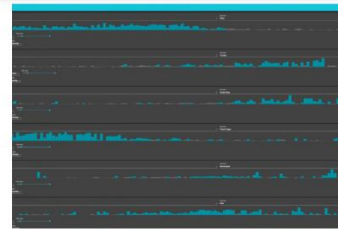
Artist in Residence



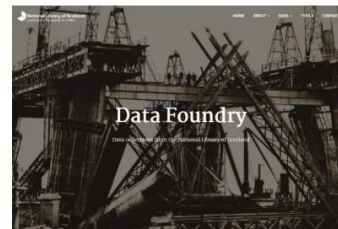
Teaching digital humanities with A  
Medical History of British India



Geoparsing the Gazetteers of Scotland



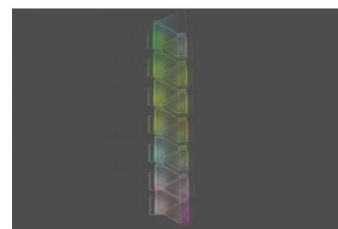
3,000 Scottish chapbooks...as music!



The National Librarian's Research  
Fellowship in Digital Scholarship



AI in Residence



The Library as LiDAR (coming soon)

# Projects

<https://data.nls.uk/projects/>

Examples of collaborative and individual projects over  
the past 6 months



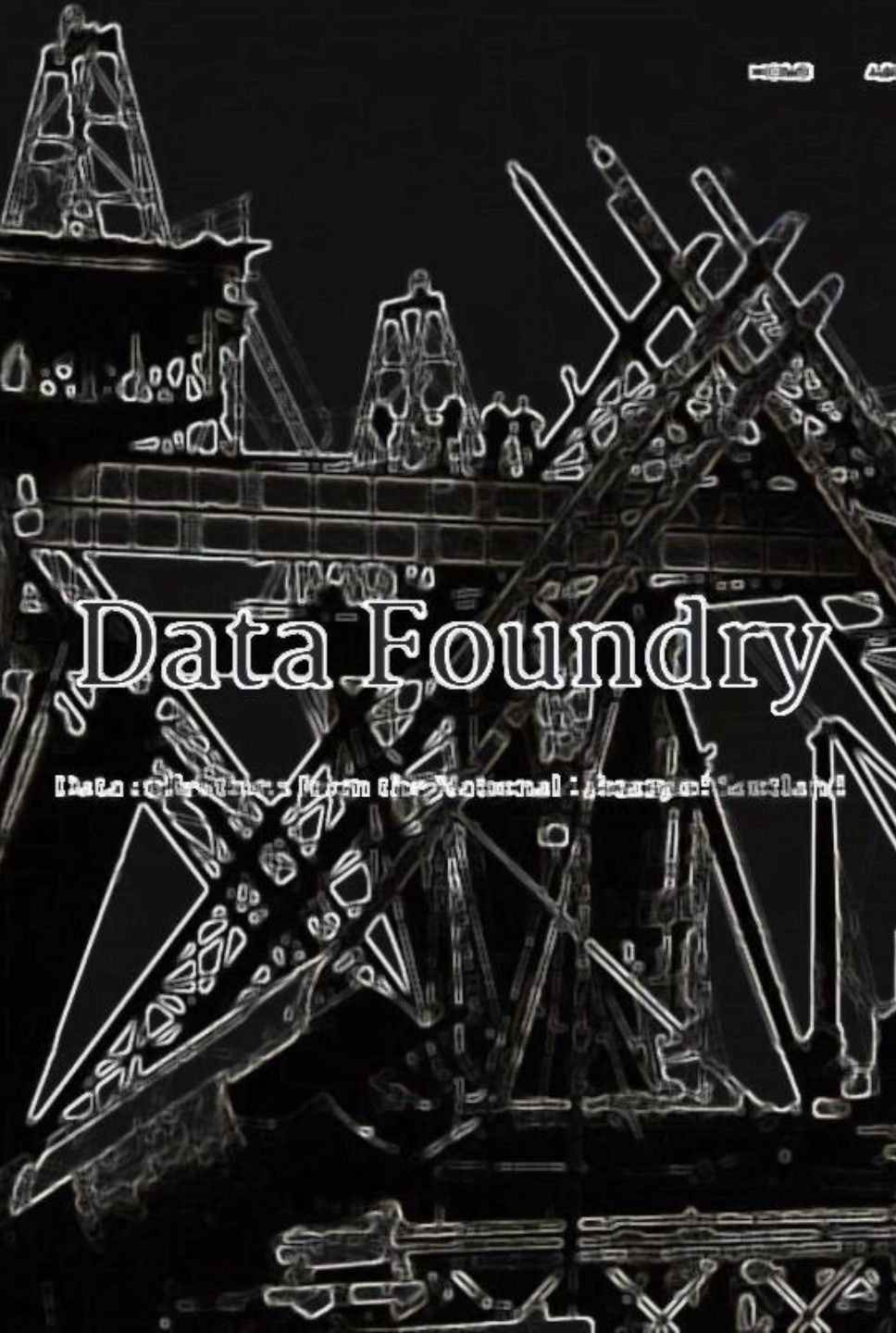
# The way forward

# Beyond the ‘collection’

- Mix ‘n’ match datasets
  - New DAMS will enable staff to create non-collection-driven datasets
  - Eventually extend this to users?
  - User-driven

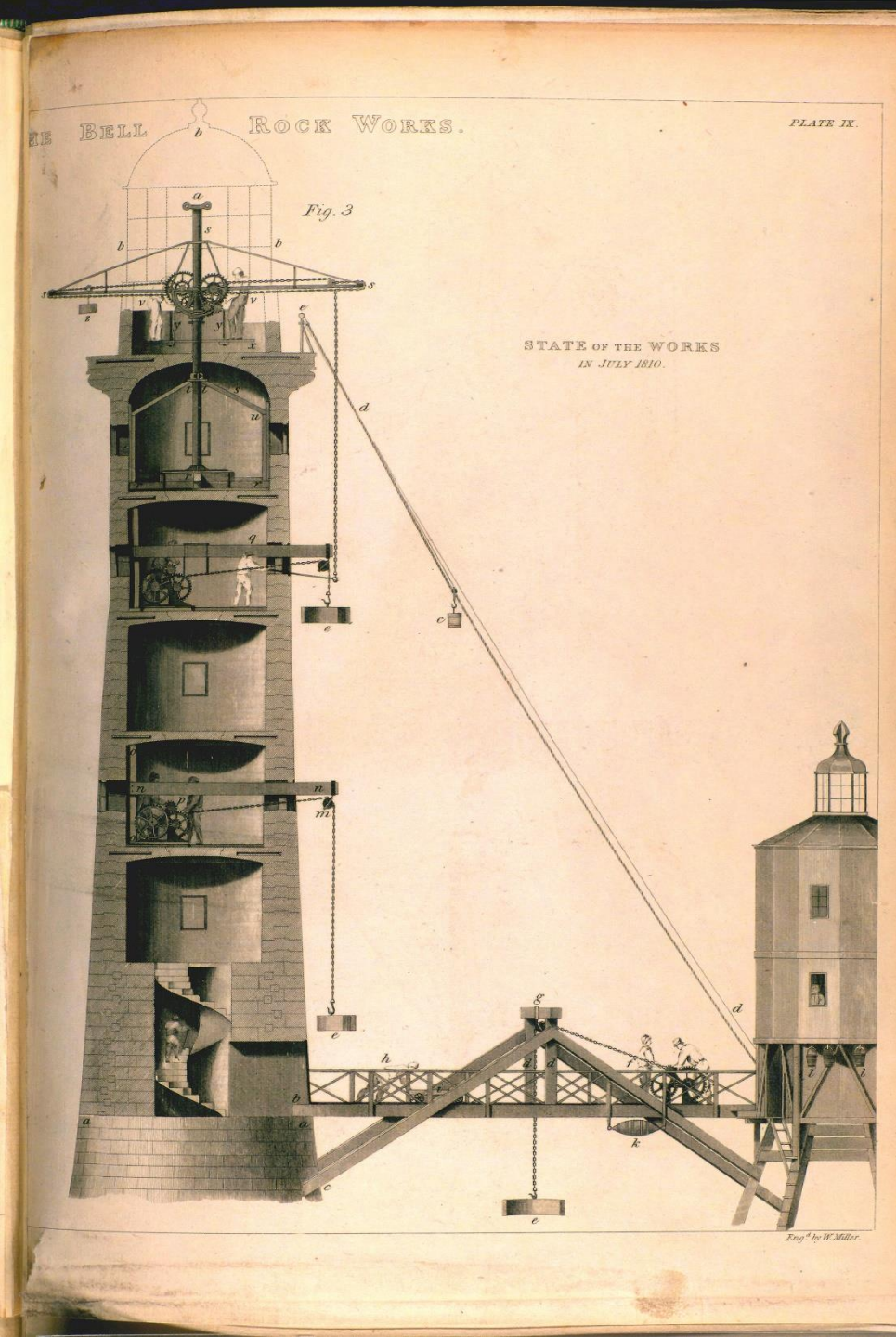
# ‘Dark’ Data Foundry

- Non-open data
- Gradually increase organisational appetite for risk
- Data Safe Haven?



# Improved engineering

- Ingest of modified datasets
- More sophisticated search
- In-house tech support
- Library as data: Library itself as a 'data space'



, signifies goods got by purchase or donation. See

the Dutchy of Montferrat, with a bihop\u2019s see, \t was taken by the Spaniards in 1745, and retaken by \t; but after this, it was taken again and difrcantled \twards forsook it. It is seated on the river Bormio, \t 30S. Cafal, 8. 30.E. 44. 40. lat."

is the consent that a person gives to the determination

on, or by a conful

e English law, is a writ that lies for a surety, against

o acquit the complainant after the debt is paid.

undredth: in England, signifies the privilege of being

e in shires and hundreds.

notes the obtaining or procuring something. Among lawyers,

or title to an estate got by purchase or dona

books, signifies to discharge or pay off the debts

iverance, or setting of a person free from the guilt

e.

discharge in writing for a sum of money.

the coast of Guinea, where the English, Dutch, and

and each sort its particular village, o. 2.W. 5. o.

signifies the predominancy of one quality over another.

town in Syria. It was formerly called Ptolemais, and

It was very famous in the time of the crusadoes, and

is both by the Chris Hans and Saracens. It is now an

ing entirely supported by its harbour, which, is frequented

ions. It is 20 miles S. Tyre, and 37N. Jerusalem, 39.\

, in the Mogul\u2019s dominions, the same with lak,\

f 100,000 rupees; the rupee is of the value of the

s, or 30 fols of Holland; an 100 lacks of rupees make

r 10,000,000 rupees; the pound Sterling is about 8\

ich proportion, a lack of rupees amounts to 12,500\

a measure of land used in several provinces of France,\

y. It is larger or less according to the different\

contains 160 perches. The Acre of woods in France, consists\

erg es; the rood is 40 perches, the perch 24 feet,\

inch 12 lines. Acre, the universal measure of land\

England contains 4 square roods, a rood 40ACR or poles\

tatute. Yet this measure does not prevail in all parts\

th of the pole varies in different counties, and is\

re, the difference running from the 16feet to 28. The\

to 30 square chains, of 22 yards each, that is 4040\

A C R

ACQUEST, or Acquist, in law, signifies

by purchase or donation. See CONQUEST.

ACQUI, a town of Italy, in the Dutchy of

rat, with a bishop's see, and commodious

was taken by the Spaniards in 1745, and r

the Piedmontese in 1746; but after this, it

again and dismantled by the French, who

forsook it. It is seated on the river Bormio

N. W. of Genoa, and 30 S. of Cafal, 8. 30

44. 40. lat.

ACQUIESCENCE, in commerce, is the co

a person gives to the determination given cit

bitration, or by a conful

ACQUIETANDIS *plegiis*, in the English

writ that lies for a surety, against a creditor

ufes to acquit the complainant after the de

ACQUIETANTIA *de spiris et hundredis*, in

signifies the privilege of being free from sui

vice in shires and hundreds.

ACQUISITION, in general, denotes the ob

procuring something. Among lawyers, it i

the right or title to an estate got by purcha

tion.

ACQUITARE, in ancient law-books, signifi

charge or pay off the debts of a person dece

ACQUITTAL, a discharge, deliverance, or

a person free from the guilt or suspicion of a

ACQUITTANCE, a release or discharge i

for a sum of money.

ACRA, a town of Africa, on the coast of Guin

the English, Dutch, and Danes, have str

and each fort its particular village, o. 2.

5. o. lat.

ACRASIA, among physicians, signifies the pr

cy of one quality over another.

ACRE, or ACRA, a sea-port town in Syria

formerly called *Ptolemais*, and is a bishop's

# Analysis layers

- Dr Rosa Filgueira – Fellowship project
- Potential for platforms to be built on top of Data Foundry, given the standardised nature of the data



ccuracy.

that is of a (sharp or pungent taste.

ujl-eaters. It has been much disputed whether the inhabitants

c ever eat locusts. We shall give the substance of\

n this subject, who travelled in Syria and Egypt so\

This ingenious gentleman, who travelled with a view\

ory, informs us, that he asked Franks, and many other\

ng in these countries, whether they had ever heard\

Arabia and Ethiopia, &c. used locusts as food. They

He likewise asked the same question of Armenians,\

ho lived in Arabia, and had travelled in Syria and\

of whom said they heard of such a practice, and others\

en the people eat these insects. He at last obtained\

contains 160 perches.

*The Acre of woods in France*, consists of fe

called *verg es*; the rood is 40 perches, the

feet, the foot 12 inches, the inch 12 lines.

ACRE, the universal measure of land in Brit

acre in England contains 4 square roods, a

# (Inter)national infrastructure?

- The success of Data Foundry depends on community
- Project led by Universities of Glasgow/Sheffield





# Challenges

Change! Workflows, processes, culture...

Resourcing and sustainability

Identifying user needs

(Inter)national funding/partnerships

Ethics





# Thank you!

sarah.ames@nls.uk  
@semames1 | #NLSdata