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.
- Practice and promote transparency in our data creation processes.
- Anticipate the future of research.
No-nonsense data
## Identifying user needs

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

Plus broader audience includes other libraries (standards, presentation of data etc)
# Identifying user needs

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

Plus broader audience includes other libraries (standards, presentation of data etc)
Changing processes: digitisation to data

Selection
- Rights and conservation assessments
- Digitisation
- Generate derivative images (thumbnails, crops, etc)

Extract ALTO XML, txt, JPEGs, PDFs, thumbnails and METS
- [retro-create ALTO]
- Compile METS
- Files into repository – ALTO XML, txt, JPEGs, PDFs, thumbnails, copyright info

Compile dataset: structure/naming conventions
- Zip and move to cloud/local storage
- Create DOI
- Publish online

Files into repository – ALTO XML, txt, JPEGs, PDFs, thumbnails, copyright info

Generate derivative images (thumbnails, crops, etc)
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

...and the National Librarian! (‘Data Foundry’)

...and the National Librarian! (‘Data Foundry’)
Engagement activity and outcomes
Teaching and learning
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 a 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 tokenise 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`.

```python
[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

• Series of Jupyter notebooks exploring some of the text and metadata datasets
• For students, learners, non-coders
• By @lucy_havens
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/
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.

Is it true? The post-truth archive factory

- Marion Carré @Ma_rionC
- Collaboration with Goethe-Institut Glasgow, Alliance Française Glasgow and Institut Français d’Écosse
- 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/
‘Is it true? The Post-truth Archive Factory’

Marion Carré

7th December 2021 — 6th January 2022

Images: Marion Carré, used with permission
Research collaborations
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.

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

Annual fellowship 22-23

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

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
Projects

Examples of collaborative and individual projects over the past 6 months

https://data.nls.uk/projects/
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’
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
(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