Collections as data at the National Library of Scotland: access, engagement, outcomes

Sarah Ames
Digital Scholarship Librarian
National Library of Scotland
@semames1 | #NLSdata | sarah.ames@nls.uk
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
Introducing the Digital Scholarship Service
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
1. Making data available
2. External engagement

Collaboration

Projects

PhDs, residencies, fellowships

Beyond research community
3. Internal engagement

- Awareness
- Training
- Champions
- Culture
Collections as data and the Data Foundry
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 existing 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 existing 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

Compile METS

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

Compile dataset: structure/naming conventions

 Compile METS

[retro-create ALTO]

Zip and move to cloud/local storage

Create DOI

Publish online

Create DOI
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’)
Engagement activity and outcomes
Teaching and learning
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/
‘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

- 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

- Dr Gustavo Candela (University of Alicante)
  - Linked Open Data – Moving Image Archive metadata
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/
Challenges
Change! Workflows, processes, culture...

OCR quality

Copyright

How can we be clear about the provenance of our data?

Resourcing! No developers

What does collections data look like?

Versioning

Scale and scope of digital scholarship
What have we learned so far?

1. It’s worth putting effort and resource into presenting and promoting data collections

2. There is more to digitisation than online image galleries

3. DS has implications across the Library

4. Historic practices affect everything: it’s difficult to break the cycle

5. Our users have exciting, refreshing, innovative approaches to using collections, which we can learn from
Thank you!

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