National Library of Scotland
Leabharlann Nàiseanta na h-Alba

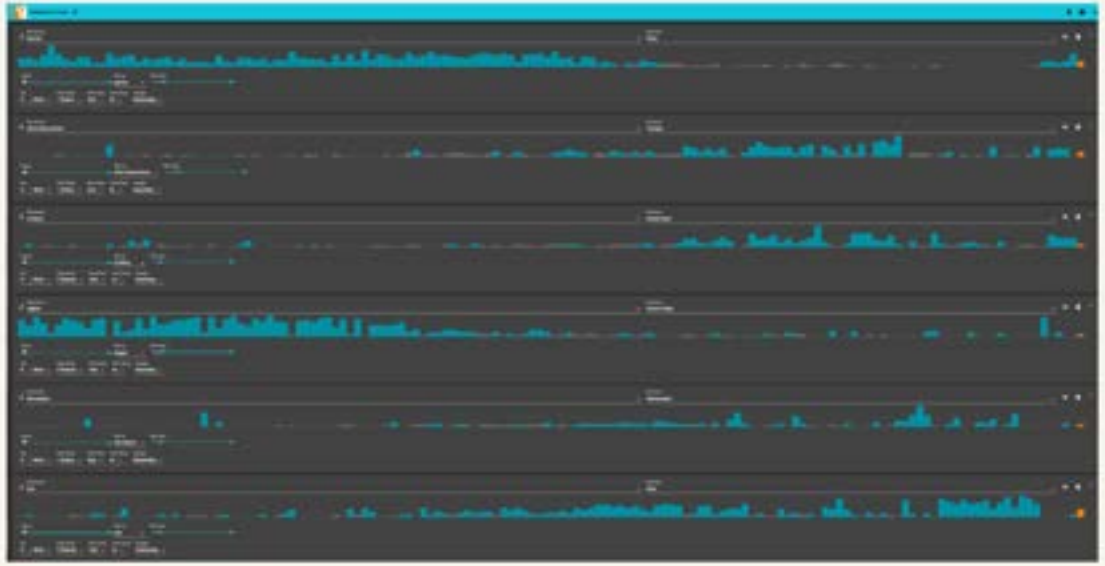
AI-READY COLLECTIONS AT THE NATIONAL LIBRARY OF SCOTLAND

INES BYRNE

DIGITAL TRANSITION MANAGER

NATIONAL LIBRARY OF SCOTLAND

JISC *MAKING YOUR COLLECTIONS AI-READY* WEBINAR 22 NOVEMBER 2023



A Song of Scottish Publishing, 1671-1893

Posted on 15 Dec 2019 by Shawn

The Scottish National Library has made available a collection of chapbooks printed in Scotland, from 1671 – 1893, on their website here. That's nearly 11 million words' worth of material. The booklets cover an enormous variety of subjects. So, what do you do with it?

National Library of Scotland
Leabharlann Nàiseanta na h-Alba

Shawn Graham, Associate Professor of Digital Humanities at Carleton University, Canada

A Song of Scottish Publishing: https://soundcloud.com/shawn-graham-60451318/chapbook-music

https://electricarchaeology.ca/2019/12/15/a-song-of-scottish-publishing-1671-1893/

National Library of Scotland
Leabharlann Nàiseanta na h-Alba

Digital Scholarship Service Levels

Level 1: self-service (making datasets available)

Level 2: plus short consultation (aligns with 121time we give readers in the reading room)

Level 3: funded service (Library as collaborator / partner in funded projects)

National Library of Scotland
Leabharlann Nàiseanta na h-Alba

Data 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.

National Library of Scotland
Leabharlann Nàiseanta na h-Alba

Making data available

TEXT

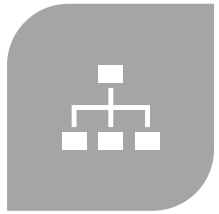
IMAGE

METADATA

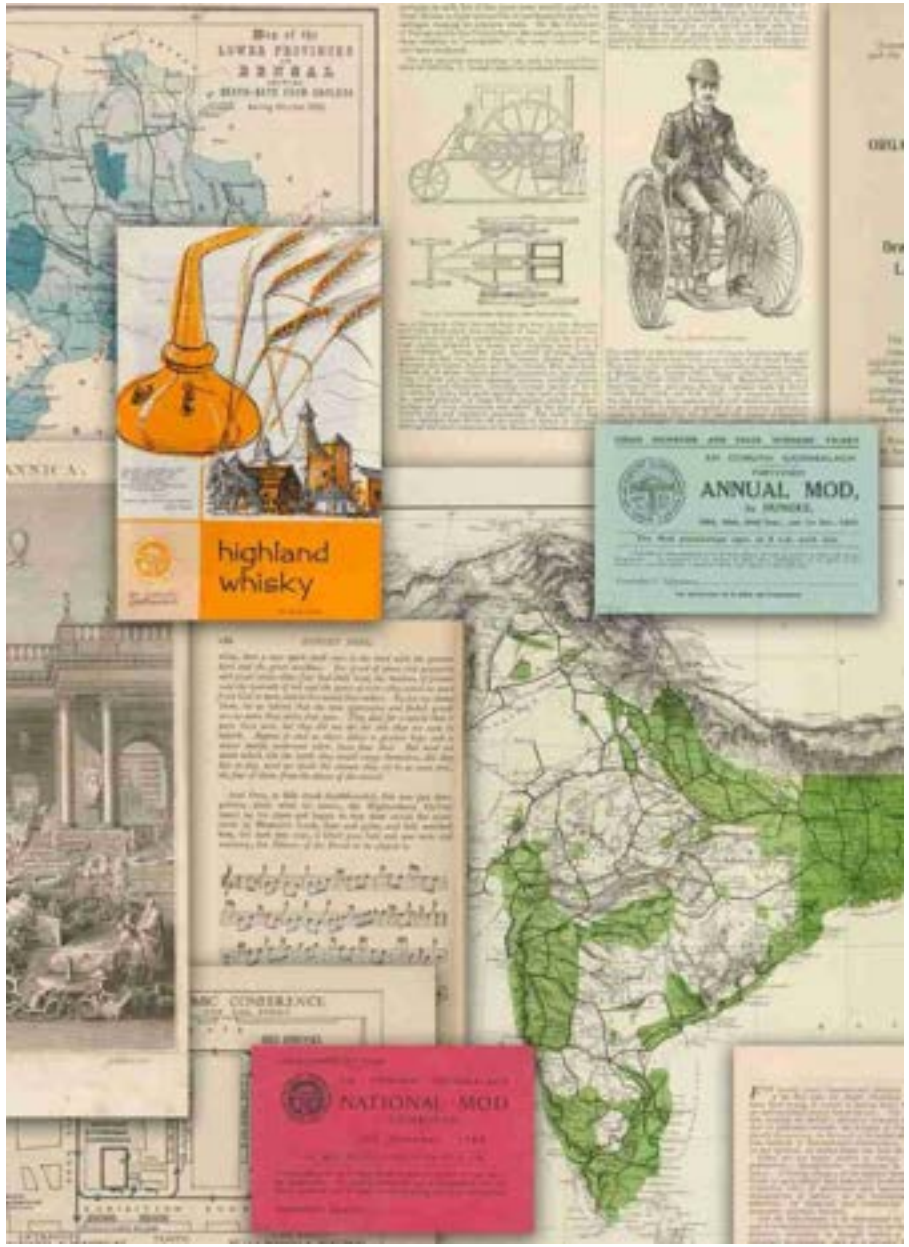
AUDIOVISUAL

MAPS

WEB ARCHIVE



CORPORATE



National Library of Scotland
Leabharlann Nàiseanta na h-Alba

PREREQUISITES

dozens of digitised collections, mainly OCRed texts

tens of thousands of digitised and geo-referenced maps

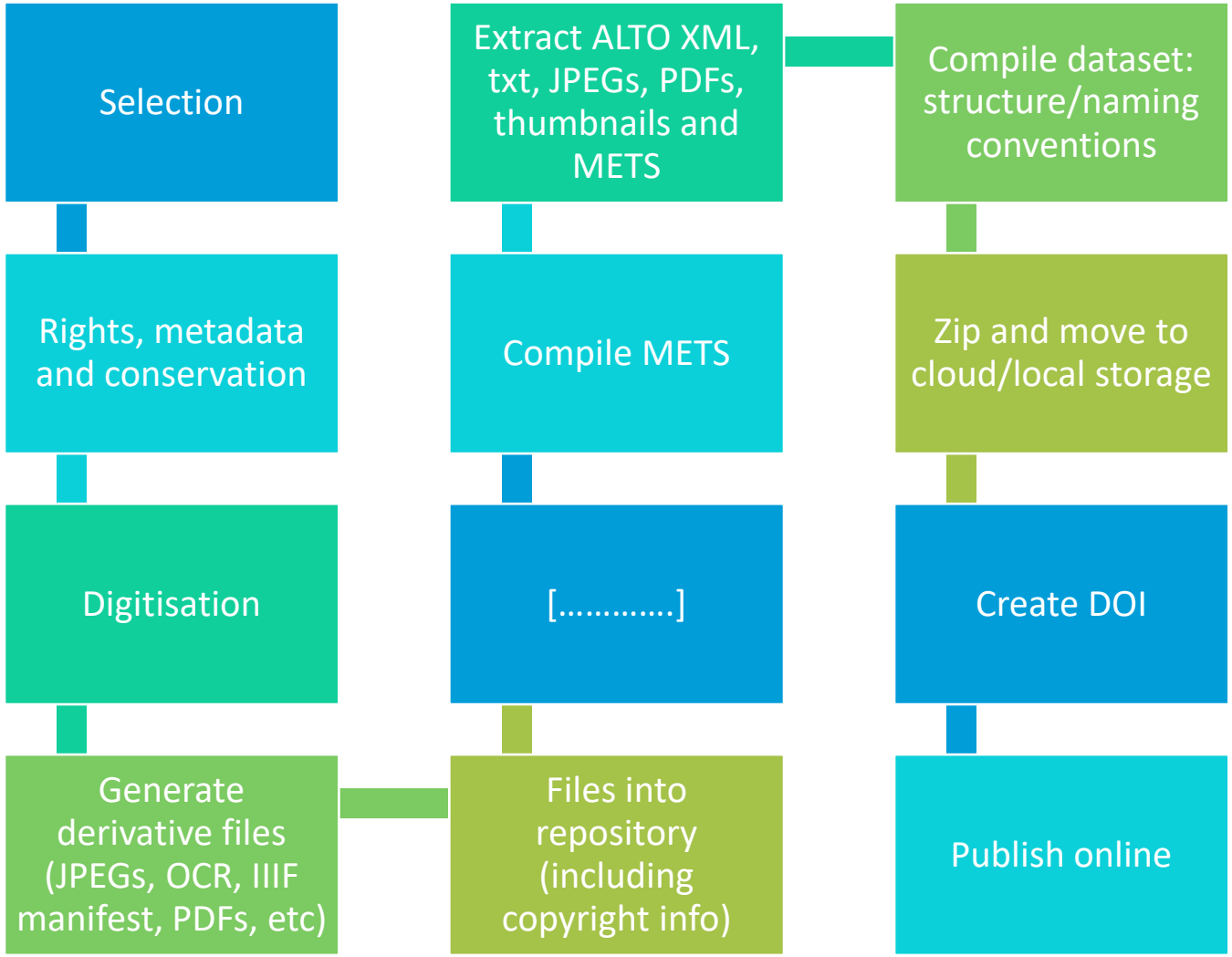
an ongoing digitisation workflow

a rights assessment process

a DAMS

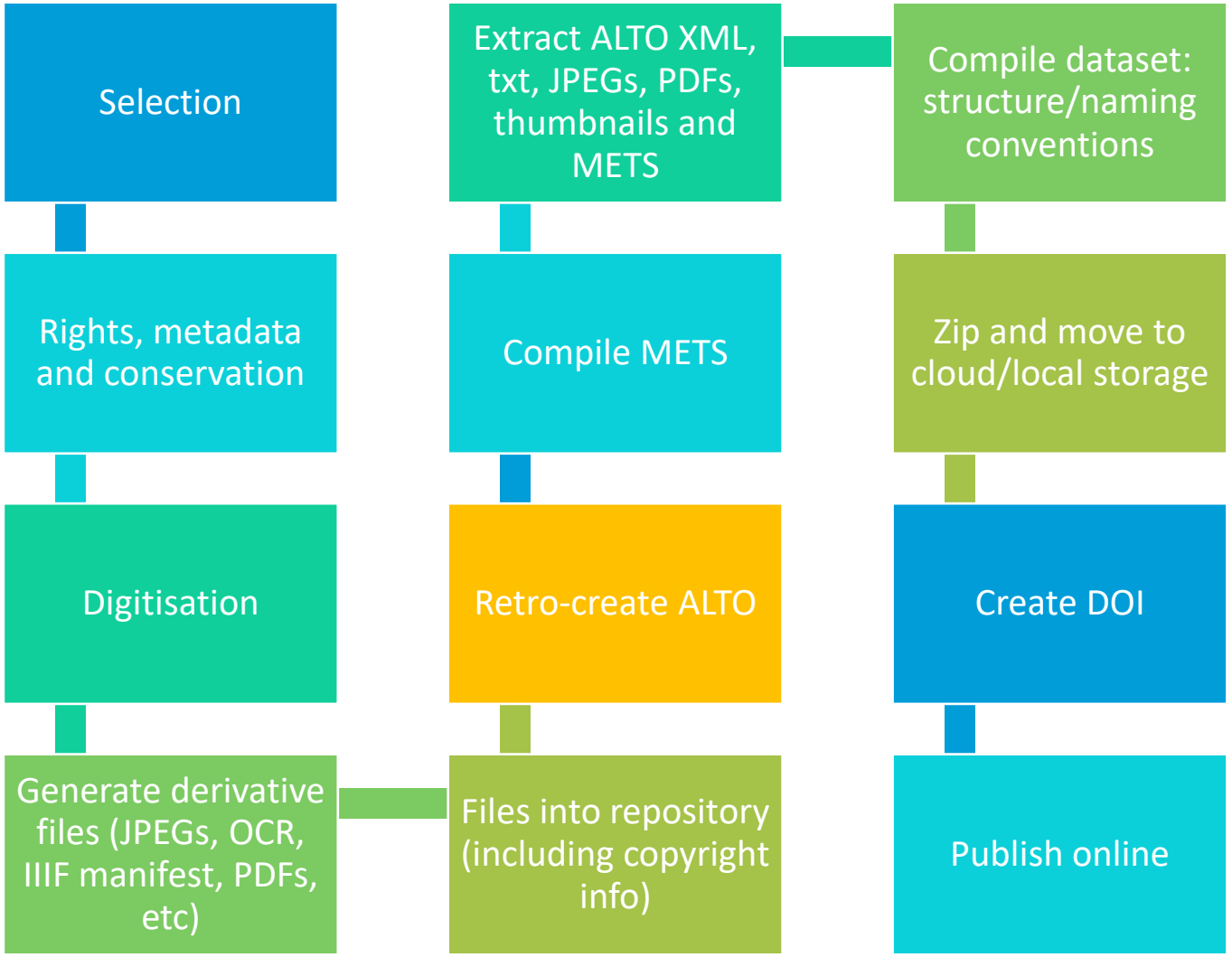
National Library of Scotland
Leabharlann Nàiseanta na h-Alba

WORKFLOW: digitisation to data



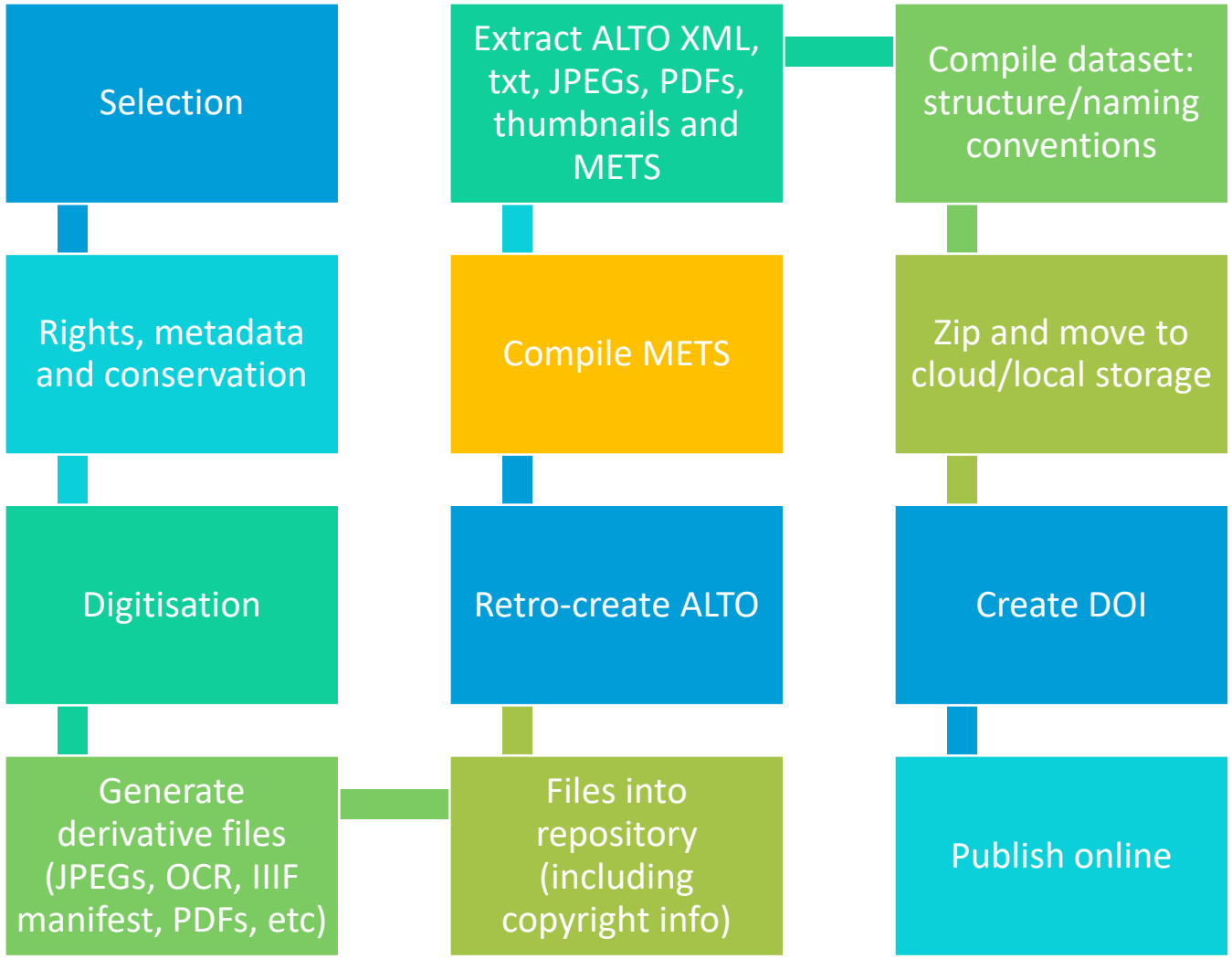
National Library of Scotland
Leabharlann Nàiseanta na h-Alba

CHANGING PROCESS: digitisation to data



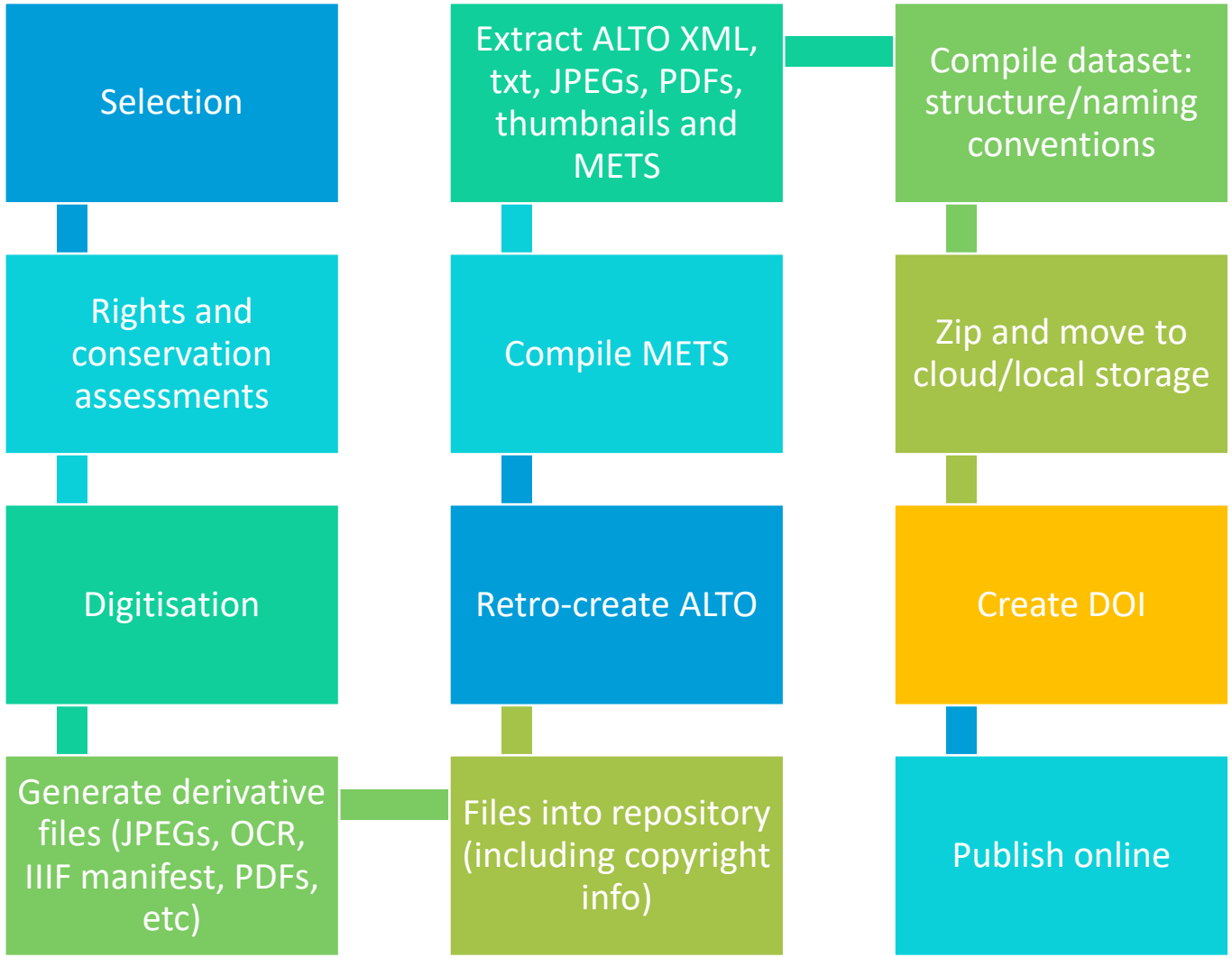
National Library of Scotland
Leabharlann Nàiseanta na h-Alba

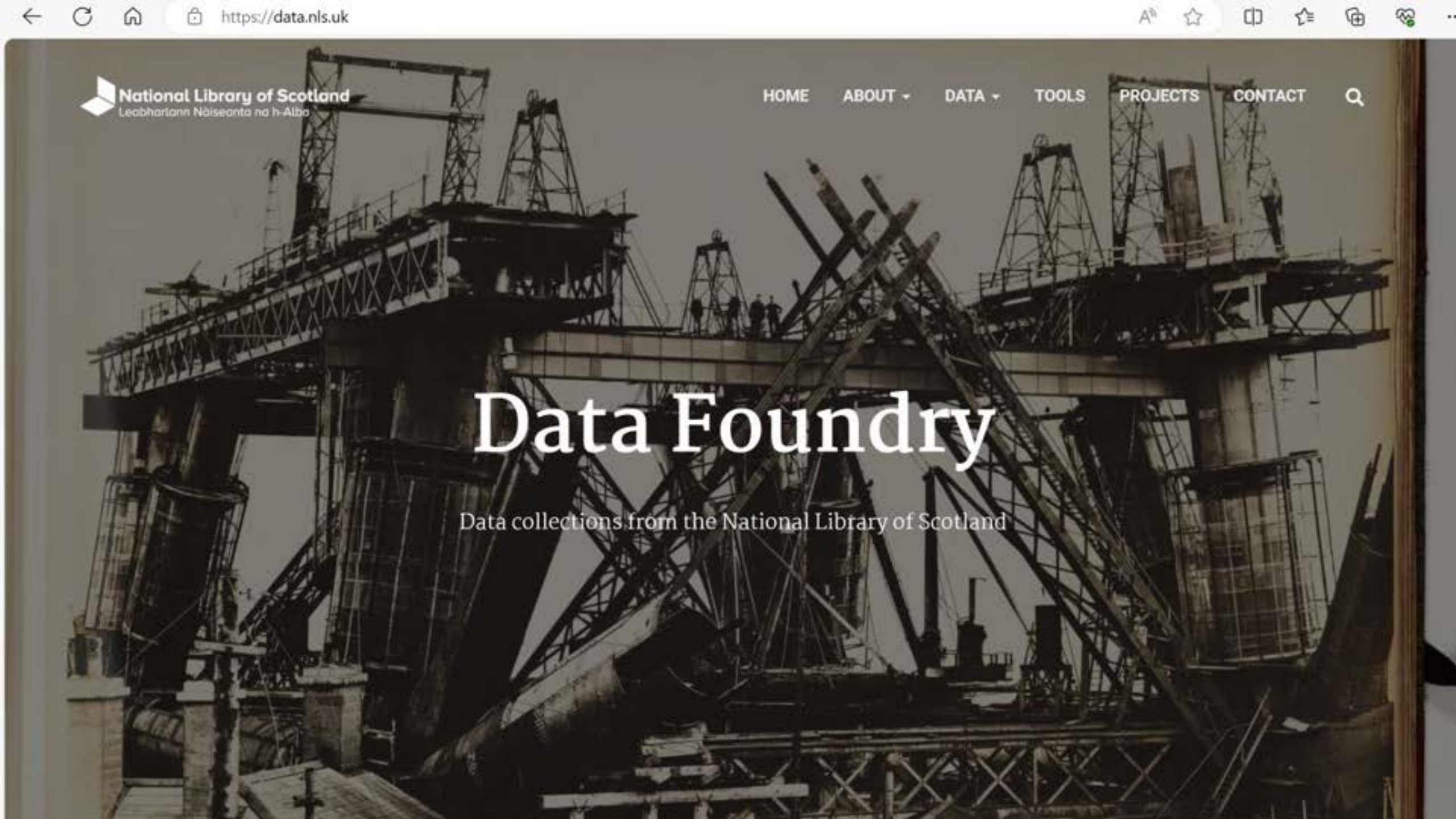
CHANGING PROCESS: digitisation to data



National Library of Scotland
Leabharlann Nàiseanta na h-Alba

CHANGING PROCESS: digitisation to data




https://data.nls.uk
National Library of Scotland
Leabharlann Nàiseanta na h-Alba
HOME
ABOUT
DATA
TOOLS
PROJECTS
CONTACT
Data Foundry
Data collections from the National Library of Scotland

Encyclopaedia Britannica

Original OCR: no clean-up

155,388 ALTO XML files at page level

155,388 image files

METS metadata files at item level

19,257,785 lines and 166,729,009 words

Covers years 1768-1860

The first eight editions of Encyclopaedia Britannica, issued from 1768-1860, comprise a total of 143 volumes. The Britannica was first issued in Edinburgh in 100 weekly parts (forming 3 volumes) from 1768 to 1771 and illustrated with 160 copperplate engravings. The enterprise was undertaken by the partnership of printer Colin Macfarquhar (1744-1793) and engraver Andrew Bell (1726-1809) who paid


Encyclopædia Britannica;

National Library of Scotland
Leabharlann Nàiseanta na h-Alba

Rights information

Encyclopaedia Britannica: up to 1853

Items in this collection up to 1853 are free of known copyright restrictions. For details visit the Library's **copyright page**.

Encyclopaedia Britannica: 1854–1860

Items in this collection between 1854 and 1860 are likely to be free of known copyright restrictions. For details visit our **copyright page**.

National Library of Scotland
Leabharlann Nàiseanta na h-Alba

Download the data

Trial the data

Download a sample of the dataset for initial evaluation.

File contents: 1 plain text readme file; 832 ALTO XML files; 1 METS file; 832 image files.

File size: 132 MB compressed (220 MB uncompressed)

Download sample dataset

All the data

File contents: 1 plain text readme file; 1 CSV inventory file; 155,388 ALTO XML files; 195 METS files; 155,388 image files.

File size: 23.5 GB compressed (44.0 GB uncompressed)

Caution: large dataset

Download full dataset

Just the text

File contents: 1 plain text readme file; 1 CSV inventory file; 195 plain text files.

File size: 336 MB compressed (946.88 MB uncompressed)

Download text data

National Library of Scotland
Leabharlann Nàiseanta na h-Alba

Cite the data

DOI: **https://doi.org/10.34812/cg4r-dn40**

Dataset creator: National Library of Scotland

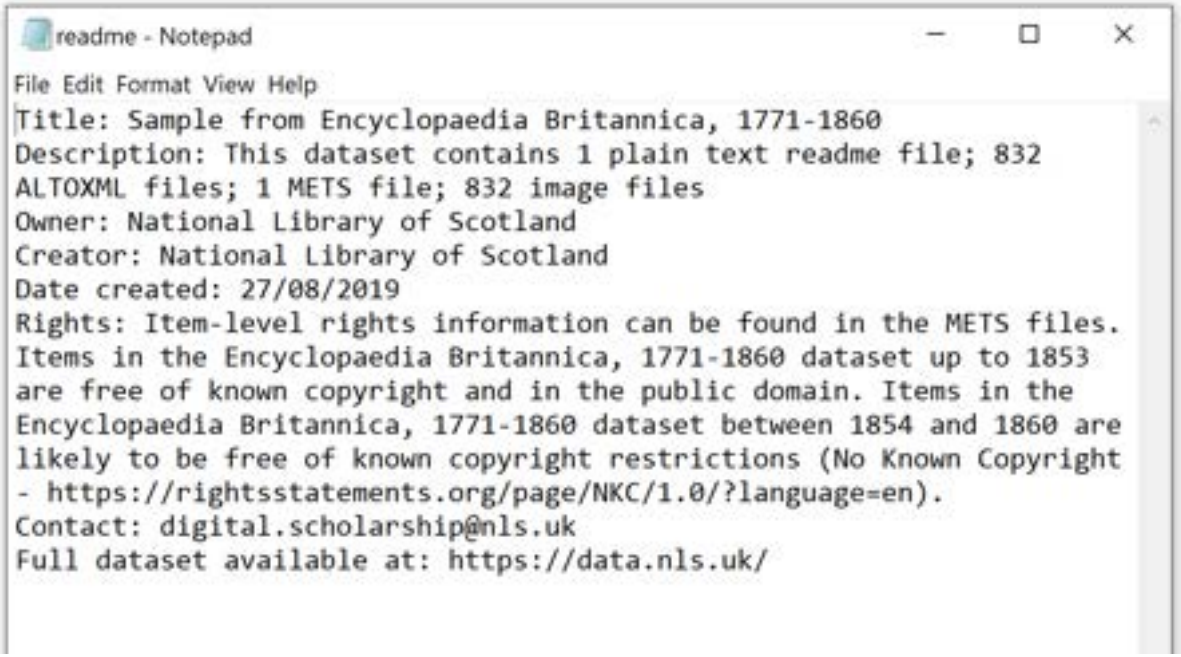
Dataset publisher: National Library of Scotland

Publication year: 2019

Suggested citation: National Library of Scotland. *Encyclopaedia Britannica*. National Library of Scotland, 2020. **https://doi.org/10.34812/cg4r-dn40**

DIGITISED COLLECTIONS DATASET STRUCTURE

Name	Size	Packe...	Modified	Creat...	Acces...	Attr...	Encry...	Com...	CRC	Method	Chara...	Host OS	Version	Volu...	Offset	Folders	Files
alto	105 8...	15 37...	2019-08-27 15:01			D	-		47C13214	Store		FAT	20	0	127 082	0	832
image	123 6...	123 3...	2019-08-27 15:02			D	-		E1385B9B	Store		FAT	20	0	15 579 322	0	832
144133901-mets.xml	1 943...	126 9...	2019-08-26 20:06			A	-		CFA7CFC9	Deflate		FAT	20	0	72		



National Library of Scotland
Leabharlann Nàiseanta na h-Alba

DIGITISED COLLECTIONS DATASET STRUCTURE

Name	Size	Packe...	Modified	Creat...	Acces...	Attr...	Encry...	Com...	CRC	Method	Chara...	Host OS	Version	Volu...	Offset	Folders	Files
188082722.34.xml	714	414	2019-07-18 13:30			A	-		F7DD9475	Deflate		FAT	20	0	127 159		
188082735.34.xml	8 673	1 738	2019-07-18 13:30			A	-		0590254C	Deflate		FAT	20	0	127 666		
188082748.34.xml	9 350	1 909	2019-07-18 13:30			A	-		7EFF7FE6	Deflate		FAT	20	0	129 497		
188082761.34.xml	714	412	2019-07-18 13:30			A	-		AD836F5C	Deflate		FAT	20	0	131 499		
188082774.34.xml	714	413	2019-07-18 13:30			A	-		7EDEF029	Deflate		FAT	20	0	132 004		
188082787.34.xml	714	413	2019-07-18 13:30			A	-		B636CB28	Deflate		FAT	20	0	132 510		
188082800.34.xml	1 808	723	2019-07-18 13:30			A	-		9FAD0CFA	Deflate		FAT	20	0	133 016		
188082813.34.xml	3 767	988	2019-07-18 13:30			A	-		4ED2689A	Deflate		FAT	20	0	133 832		
188082826.34.xml	16 583	3 081	2019-07-18 13:30			A	-		6403F679	Deflate		FAT	20	0	134 913		
188082839.34.xml	1 374	613	2019-07-18 13:30			A	-		710028C3	Deflate		FAT	20	0	138 087		
188082852.34.xml	56 904	8 991	2019-07-18 13:30			A	-		D4C7556D	Deflate		FAT	20	0	138 793		
188082865.34.xml	63 177	10 006	2019-07-18 13:30			A	-		A816DE10	Deflate		FAT	20	0	147 877		
188082878.34.xml	41 729	7 321	2019-07-18 13:30			A	-		4F6C8502	Deflate		FAT	20	0	157 976		
188082891.34.xml	37 631	6 759	2019-07-18 13:30			A	-		8BB3AA68	Deflate		FAT	20	0	165 390		
188082904.34.xml	92 756	14 058	2019-07-18 13:30			A	-		740D8485	Deflate		FAT	20	0	172 242		
188082917.34.xml	148 6...	22 234	2019-07-18 13:30			A	-		C5742529	Deflate		FAT	20	0	186 393		
188082930.34.xml	141 6...	21 601	2019-07-18 13:30			A	-		6D2C6CFC	Deflate		FAT	20	0	208 720		
188082943.34.xml	160 2...	23 801	2019-07-18 13:30			A	-		F6DA8BD7	Deflate		FAT	20	0	230 414		
188082956.34.xml	144 8...	22 266	2019-07-18 13:30			A	-		FA21FBFB	Deflate		FAT	20	0	254 308		
188082969.34.xml	165 7...	24 108	2019-07-18 13:30			A	-		A3A91A1E	Deflate		FAT	20	0	276 667		
188082982.34.xml	143 3...	21 762	2019-07-18 13:30			A	-		5B317035	Deflate		FAT	20	0	300 868		
188082995.34.xml	145 1...	21 992	2019-07-18 13:30			A	-		40A6CC45	Deflate		FAT	20	0	322 723		
188083008.34.xml	161 8...	23 886	2019-07-18 13:30			A	-		2D40AF0A	Deflate		FAT	20	0	344 808		
188083021.34.xml	154 6...	22 791	2019-07-18 13:30			A	-		2E800ED4	Deflate		FAT	20	0	368 787		

National Library of Scotland
Leabharlann Nàiseanta na h-Alba

DIGITISED COLLECTIONS DATASET STRUCTURE

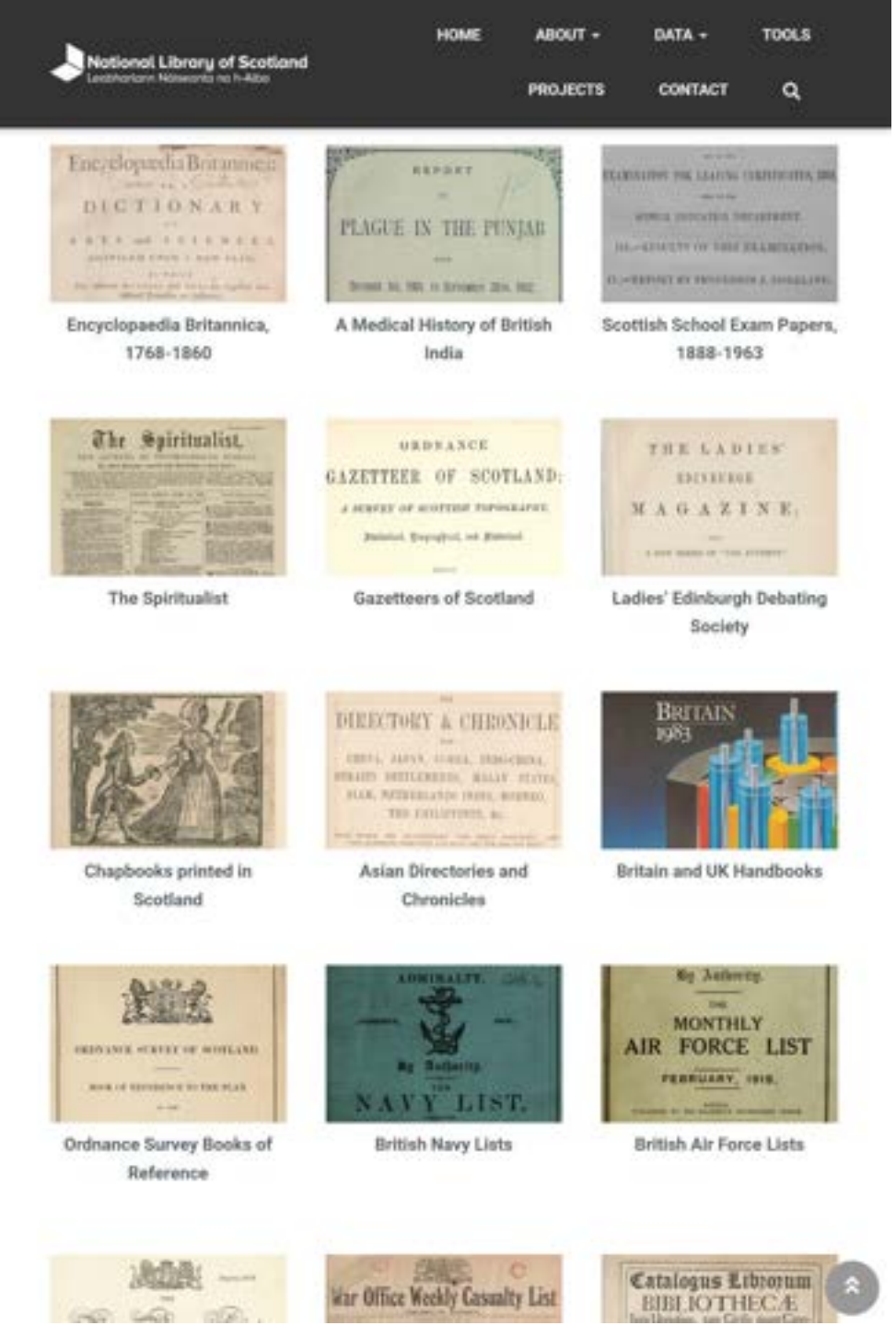
This XML file does not appear to have any style information associated with it. The document tree is shown below.


```
▼<alto xmlns="http://www.loc.gov/standards/alto/v3/alto.xsd">
  ▼<Description>
    <MeasurementUnit>pixel</MeasurementUnit>
    ▼<sourceImageInformation>
      <fileName>./data/pdfs/c_188936619/i_144133901/188082693.23.pdf</fileName>
    </sourceImageInformation>
    ▼<OCRProcessing ID="IdOcr">
      ▼<ocrProcessingStep>
        <processingDateTime>Fri Jul 12 10:36:11 2019 </processingDateTime>
        ▼<processingSoftware>
          <softwareCreator>CONTRIBUTORS</softwareCreator>
          <softwareName>pdfalto</softwareName>
          <softwareVersion>0.1</softwareVersion>
        </processingSoftware>
      </ocrProcessingStep>
    </OCRProcessing>
  </Description>
  ▼<Styles>
    <TextStyle ID="font0" FONTFAMILY="times" FONTSIZE="8.000" FONTTYPE="serif" FONTWIDTH="proportional" FONTCOLOR="#WWWWWW" FONTSTYLE="bold"/>
    <TextStyle ID="font1" FONTFAMILY="times" FONTSIZE="5.000" FONTTYPE="serif" FONTWIDTH="proportional" FONTCOLOR="#WWWWWW" FONTSTYLE=""/>
    <TextStyle ID="font2" FONTFAMILY="times" FONTSIZE="5.000" FONTTYPE="serif" FONTWIDTH="proportional" FONTCOLOR="#WWWWWW" FONTSTYLE="italics"/>
    <TextStyle ID="font3" FONTFAMILY="times" FONTSIZE="4.000" FONTTYPE="serif" FONTWIDTH="proportional" FONTCOLOR="#WWWWWW" FONTSTYLE="bold"/>
    <TextStyle ID="font4" FONTFAMILY="times" FONTSIZE="6.000" FONTTYPE="serif" FONTWIDTH="proportional" FONTCOLOR="#WWWWWW" FONTSTYLE="bold"/>
    <TextStyle ID="font5" FONTFAMILY="times" FONTSIZE="7.000" FONTTYPE="serif" FONTWIDTH="proportional" FONTCOLOR="#WWWWWW" FONTSTYLE="bold"/>
    <TextStyle ID="font6" FONTFAMILY="times" FONTSIZE="3.200" FONTTYPE="serif" FONTWIDTH="proportional" FONTCOLOR="#WWWWWW" FONTSTYLE="bold"/>
    <TextStyle ID="font7" FONTFAMILY="times" FONTSIZE="3.200" FONTTYPE="serif" FONTWIDTH="proportional" FONTCOLOR="#WWWWWW" FONTSTYLE=""/>
    <TextStyle ID="font8" FONTFAMILY="times" FONTSIZE="4.000" FONTTYPE="serif" FONTWIDTH="proportional" FONTCOLOR="#WWWWWW" FONTSTYLE=""/>
    <TextStyle ID="font9" FONTFAMILY="times" FONTSIZE="5.000" FONTTYPE="serif" FONTWIDTH="proportional" FONTCOLOR="#WWWWWW" FONTSTYLE="bold"/>
    <TextStyle ID="font10" FONTFAMILY="times" FONTSIZE="4.800" FONTTYPE="serif" FONTWIDTH="proportional" FONTCOLOR="#WWWWWW" FONTSTYLE="bold"/>
  </Styles>
  ▼<Layout>
    ▼<Page ID="Page24" PHYSICAL_IMG_NR="24" WIDTH="2456" HEIGHT="3337">
      ▼<PrintSpace>
        ▼<TextLine WIDTH="235" HEIGHT="60" ID="p24_t1" HPOS="581" VPOS="199">
          <String ID="p24_w1" CONTENT="A" HPOS="581" VPOS="199" WIDTH="42" HEIGHT="60" STYLEREFS="font0"/>
          <SP WIDTH="60" VPOS="199" HPOS="623"/>
          <String ID="p24_w2" CONTENT="C" HPOS="683" VPOS="199" WIDTH="34" HEIGHT="60" STYLEREFS="font0"/>
          <SP WIDTH="58" VPOS="199" HPOS="717"/>
          <String ID="p24_w3" CONTENT="A" HPOS="775" VPOS="199" WIDTH="41" HEIGHT="60" STYLEREFS="font0"/>
```

National Library of Scotland
Leabharlann Nàiseanta na h-Alba

DIGITISED COLLECTIONS DATASET STRUCTURE

```
</TextLine>
▼<TextLine WIDTH="1028" HEIGHT="37" ID="p24_t7" HPOS="209" VPOS="469">
  <String ID="p24_w37" CONTENT="ACADEMY," HPOS="209" VPOS="469" WIDTH="255" HEIGHT="37" STYLEREFS="font1"/>
  <SP WIDTH="25" VPOS="469" HPOS="464"/>
  <String ID="p24_w38" CONTENT="in" HPOS="489" VPOS="469" WIDTH="32" HEIGHT="37" STYLEREFS="font1"/>
  <SP WIDTH="19" VPOS="469" HPOS="521"/>
  <String ID="p24_w39" CONTENT="antiquity," HPOS="540" VPOS="469" WIDTH="174" HEIGHT="37" STYLEREFS="font1"/>
  <SP WIDTH="27" VPOS="469" HPOS="714"/>
  <String ID="p24_w40" CONTENT="a" HPOS="741" VPOS="469" WIDTH="17" HEIGHT="37" STYLEREFS="font1"/>
  <SP WIDTH="24" VPOS="469" HPOS="758"/>
  <String ID="p24_w41" CONTENT="garden" HPOS="782" VPOS="469" WIDTH="120" HEIGHT="37" STYLEREFS="font1"/>
  <SP WIDTH="28" VPOS="469" HPOS="902"/>
  <String ID="p24_w42" CONTENT="or" HPOS="930" VPOS="469" WIDTH="38" HEIGHT="37" STYLEREFS="font1"/>
  <SP WIDTH="21" VPOS="469" HPOS="968"/>
  <String ID="p24_w43" CONTENT="villa," HPOS="989" VPOS="469" WIDTH="89" HEIGHT="37" STYLEREFS="font1"/>
  <SP WIDTH="25" VPOS="469" HPOS="1078"/>
  <String ID="p24_w44" CONTENT="fituated" HPOS="1103" VPOS="469" WIDTH="134" HEIGHT="37" STYLEREFS="font1"/>
</TextLine>
▼<TextLine WIDTH="980" HEIGHT="37" ID="p24_t8" HPOS="256" VPOS="515">
  <String ID="p24_w45" CONTENT="within" HPOS="256" VPOS="515" WIDTH="111" HEIGHT="37" STYLEREFS="font1"/>
  <SP WIDTH="18" VPOS="515" HPOS="367"/>
  <String ID="p24_w46" CONTENT="a" HPOS="385" VPOS="515" WIDTH="18" HEIGHT="37" STYLEREFS="font1"/>
  <SP WIDTH="15" VPOS="515" HPOS="403"/>
  <String ID="p24_w47" CONTENT="mile" HPOS="418" VPOS="515" WIDTH="75" HEIGHT="37" STYLEREFS="font1"/>
  <SP WIDTH="17" VPOS="515" HPOS="493"/>
  <String ID="p24_w48" CONTENT="of" HPOS="510" VPOS="515" WIDTH="40" HEIGHT="37" STYLEREFS="font1"/>
  <SP WIDTH="12" VPOS="515" HPOS="550"/>
  <String ID="p24_w49" CONTENT="Athens," HPOS="562" VPOS="515" WIDTH="139" HEIGHT="37" STYLEREFS="font1"/>
  <SP WIDTH="26" VPOS="515" HPOS="701"/>
  <String ID="p24_w50" CONTENT="where" HPOS="727" VPOS="515" WIDTH="107" HEIGHT="37" STYLEREFS="font1"/>
  <SP WIDTH="17" VPOS="515" HPOS="834"/>
  <String ID="p24_w51" CONTENT="Plato" HPOS="851" VPOS="515" WIDTH="96" HEIGHT="37" STYLEREFS="font1"/>
  <SP WIDTH="17" VPOS="515" HPOS="947"/>
  <String ID="p24_w52" CONTENT="and" HPOS="964" VPOS="515" WIDTH="62" HEIGHT="37" STYLEREFS="font1"/>
  <SP WIDTH="16" VPOS="515" HPOS="1026"/>
  <String ID="p24_w53" CONTENT="his" HPOS="1042" VPOS="515" WIDTH="50" HEIGHT="37" STYLEREFS="font1"/>
  <SP WIDTH="16" VPOS="515" HPOS="1092"/>
  <String ID="p24_w54" CONTENT="follow-" HPOS="1108" VPOS="515" WIDTH="128" HEIGHT="37" STYLEREFS="font1"/>
</TextLine>
```



National Library of Scotland
Leabharlann Nàiseanta na h-Alba

DIGITISED COLLECTIONS DATASETS

newspapers novels

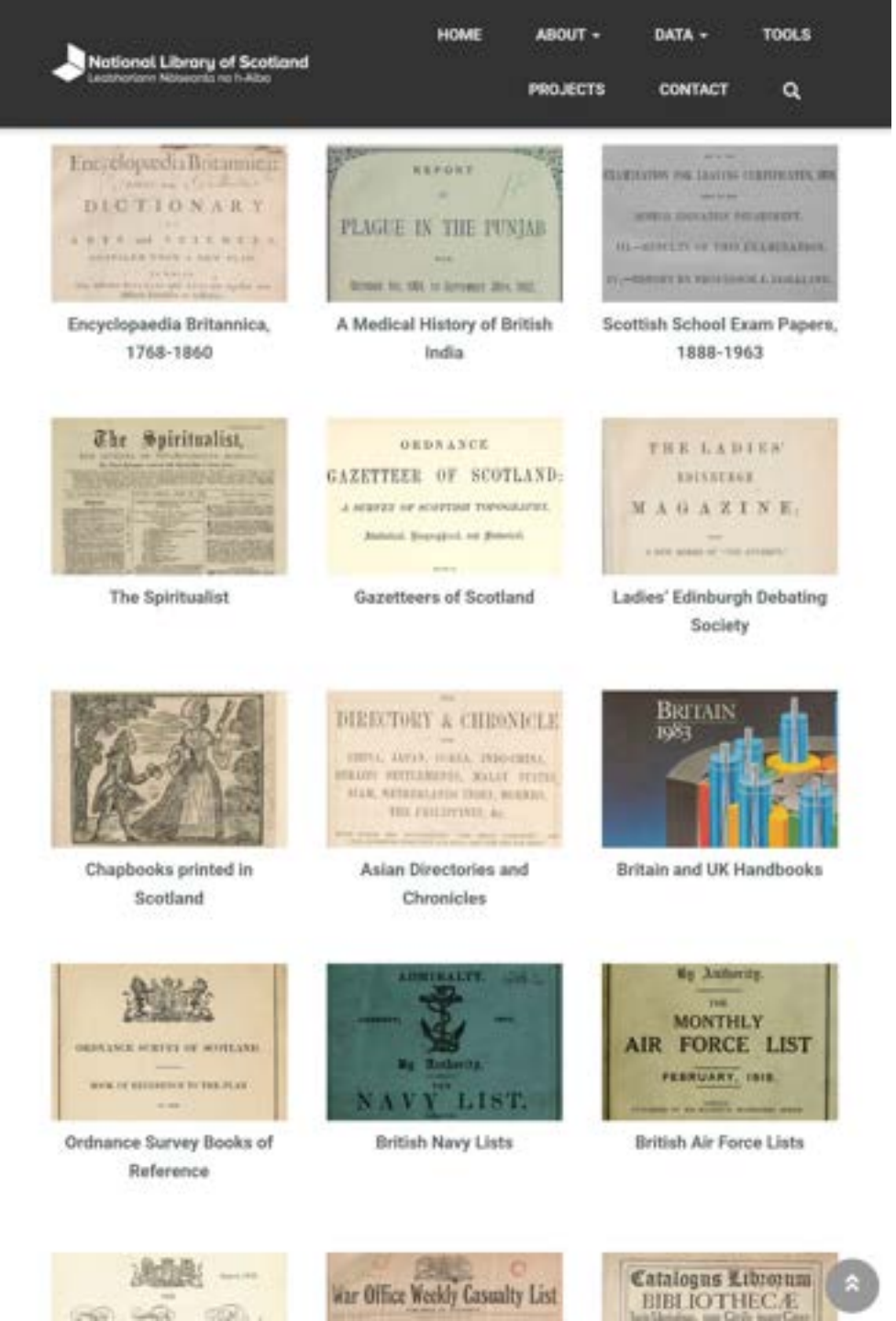
exam papers

gazetteers military lists

encyclopaedia

medical papers

chapbooks



National Library of Scotland
Leabharlann Nàiseanta na h-Alba

DIGITISED COLLECTIONS DATASETS

Largest dataset:

- 117.000 ALTO xml files
- 22.5 million words
- 17.5 GB uncompressed

Smallest dataset:

- 1,700 ALTO xml files
- 970.000 words
- 420 MB uncompressed

National Library of Scotland
Leabharlann Nàiseanta na h-Alba

Metadata collections

Download metadata collections in MARC and Dublin Core.

The National Bibliography of Scotland


Moving Image Archive


National Library of Scotland catalogue of published material (version 1)


John Adair – county maps 1680-1720


Tay Bridge Enquiry

English Ballads

Map and spatial data

Explore map and spatial data from the collections.


GB1900 gazetteer


Edinburgh Boundaries


Stevenson Maps and Plans of Scotland


LiDAR point-cloud: National Library of Scotland George IV Bridge


Roy Gazetteer


Edinburgh OS 25 inch transcriptions


Historic Footpaths


Living with Machines: railspace and building datasets

Organisational data

Our organisational data is available as CSV files.


Transactions over £25,000


Government procurement card spend


Collections size (coming


Environmental data

CHALLENGES

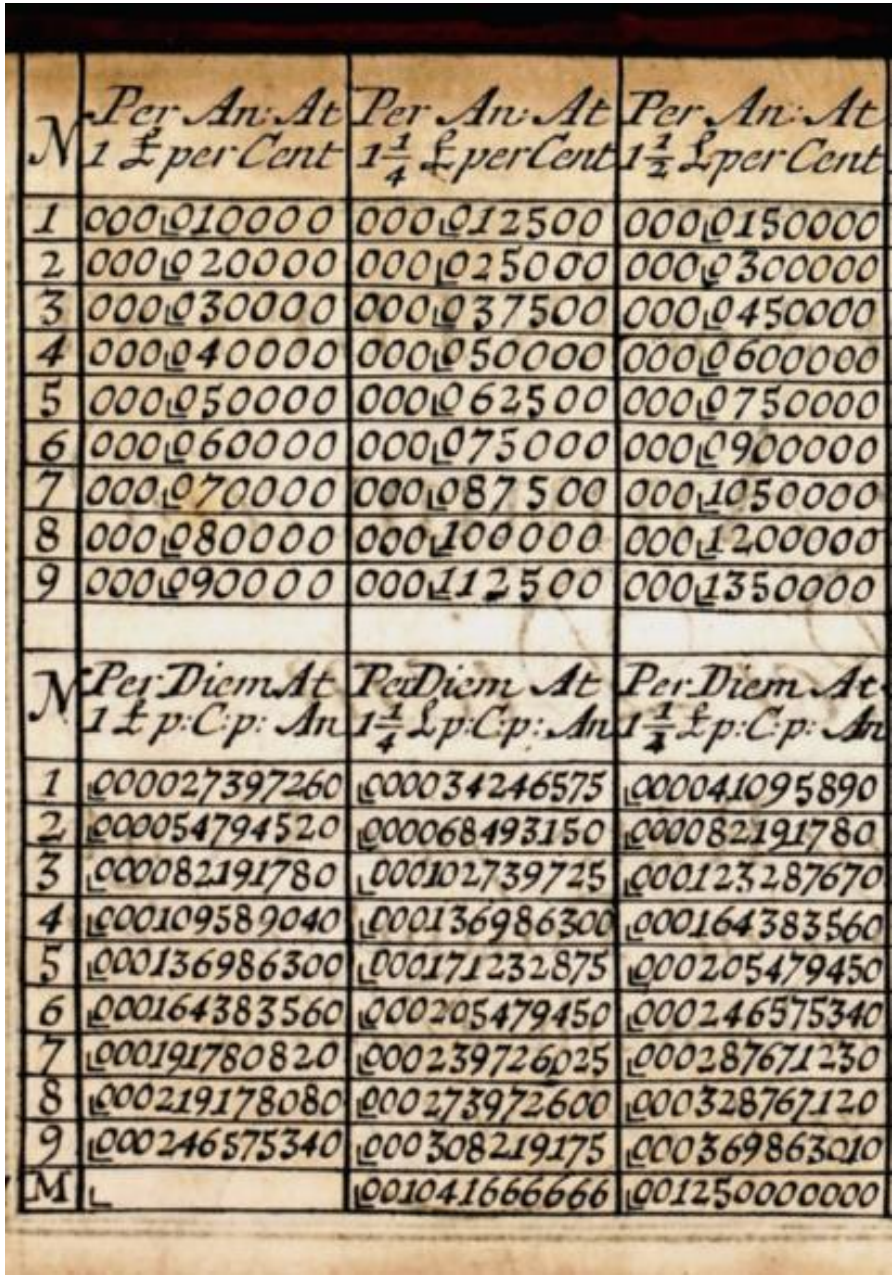
WORKFLOWS / PROCESSES

OCR

COPYRIGHT

COLLECTIONS AS DATA – DATA AS COLLECTIONS

PROVENANCE



N	Per An: At 1 £ per Cent	Per An: At 1¼ £ per Cent	Per An: At 1½ £ per Cent
1	000 £ 01000	000 £ 01250	000 £ 01500
2	000 £ 02000	000 £ 02500	000 £ 03000
3	000 £ 03000	000 £ 03750	000 £ 04500
4	000 £ 04000	000 £ 05000	000 £ 06000
5	000 £ 05000	000 £ 06250	000 £ 07500
6	000 £ 06000	000 £ 07500	000 £ 09000
7	000 £ 07000	000 £ 08750	000 £ 10500
8	000 £ 08000	000 £ 10000	000 £ 12000
9	000 £ 09000	000 £ 11250	000 £ 13500

N	Per Diem At 1 £ p: C p: An	Per Diem At 1¼ £ p: C p: An	Per Diem At 1½ £ p: C p: An
1	£ 00002 73972 60	£ 00003 42465 75	£ 00004 10958 90
2	£ 00005 47945 20	£ 00006 84931 50	£ 00008 21917 80
3	£ 00008 21917 80	£ 00010 27397 25	£ 00012 32876 70
4	£ 00010 95890 40	£ 00013 69863 00	£ 00016 43835 60
5	£ 00013 69863 00	£ 00017 12328 75	£ 00020 54794 50
6	£ 00016 43835 60	£ 00020 54794 50	£ 00024 65753 40
7	£ 00019 17808 20	£ 00023 97260 25	£ 00028 76712 30
8	£ 00021 91780 80	£ 00027 39726 00	£ 00032 87671 20
9	£ 00024 65753 40	£ 00030 82191 75	£ 00036 98630 10
M	£	£ 00104 16666 66	£ 00125 00000 00

Image: Marion Carré
https://www.marioncarre.com/living-organism

What's in it for AI ?

Dr Giles Bergel: chapbook image recognition

Dr Rosa Filgueira: *frances* AI toolbox trained on Encyclopaedia Britannica dataset

Living with Machines using our maps as training data

Joe Nockels: HTR research on manuscripts and maps images. Our first dataset generated by AI.

Martin Disley: training GAN systems to generate new works of art

Marion Carré: training AI algorithm to produce fake text

National Library of Scotland
Leabharlann Nàiseanta na h-Alba

TAKE AWAYS

The journey of a collection

National Library of Scotland
Leabharlann Nàiseanta na h-Alba

TAKE AWAYS

IT’S A MIND SHIFT AND A JOURNEY

BE TRANSPARENT AND CONSISTENT

ANALYSIS SKILLS SIT EXTERNALLY

STARTING SMALL IS ALLOWED

DON’T WAIT FOR PERFECTION

National Library of Scotland
Leabharlann Nàiseanta na h-Alba

READING

GUSTAVO CANDELA: A Checklist to Publish Collections as Data in GLAM Institutions
https://arxiv.org/abs/2304.02603

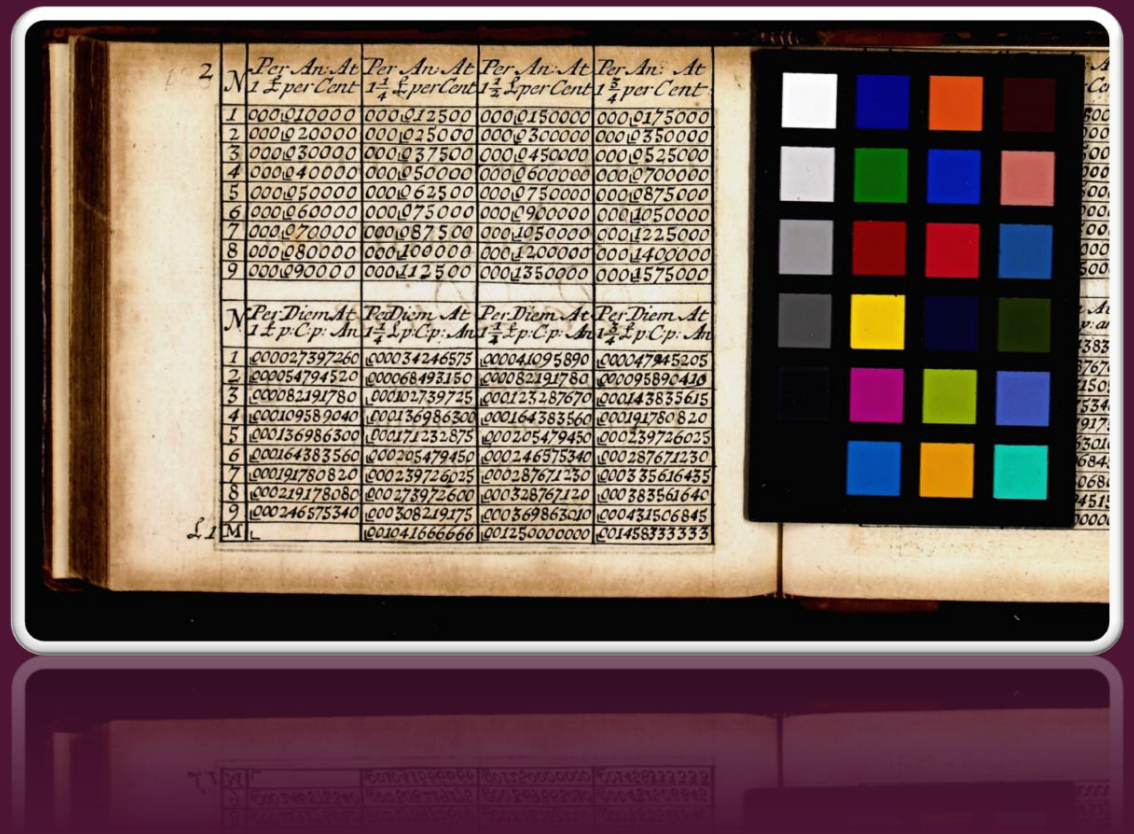
THOMAS PADILLA: Vancouver Statement on Collections as Data
https://zenodo.org/records/8342166

TIMNIT GEBRU et al.: Datasheets for Datasets

https://arxiv.org/abs/1803.09010

National Library of Scotland
Leabharlann Nàiseanta na h-Alba

THANK YOU



INES BYRNE

DIGITAL TRANSITION MANAGER

NATIONAL LIBRARY OF SCOTLAND

i.byrne@nls.uk