Actually Existing AI Applications for Personal Digital Archives

Callum McKean, Digital Lead Curator,
Contemporary Archives and Manuscripts

BRITISH LIBRARY

Remarks on ‘unlocking’ and ‘the future’

- It is not inevitable that these collections will be made widely available in the future

- Our ability to leverage in any future ‘AI revolution’ relies upon our ability to build capacity now

- Building this capacity relies, first and foremost, upon experimentation and failure

Three (partial) failures…

- Writers’ Lives PhD Placement Project (2022-3)

- Data Analytics and Network Visualisation for Hybrid Correspondence Collections (2023)

- Automated Migration and Cataloguing Workflow (2024-)

Writers’ Lives (2022-3)

- Used KNIME to parse, filter, modify and visualise collection metadata from a csv file created using DROID

- Manually added tags to metadata in order to enrich the metadata for visualisation

- Outputs: visualisations from Andrea Levy and Will Self collections; enriched tagged metadata for subsets of both collections; KNIME workflow for use with other collections.

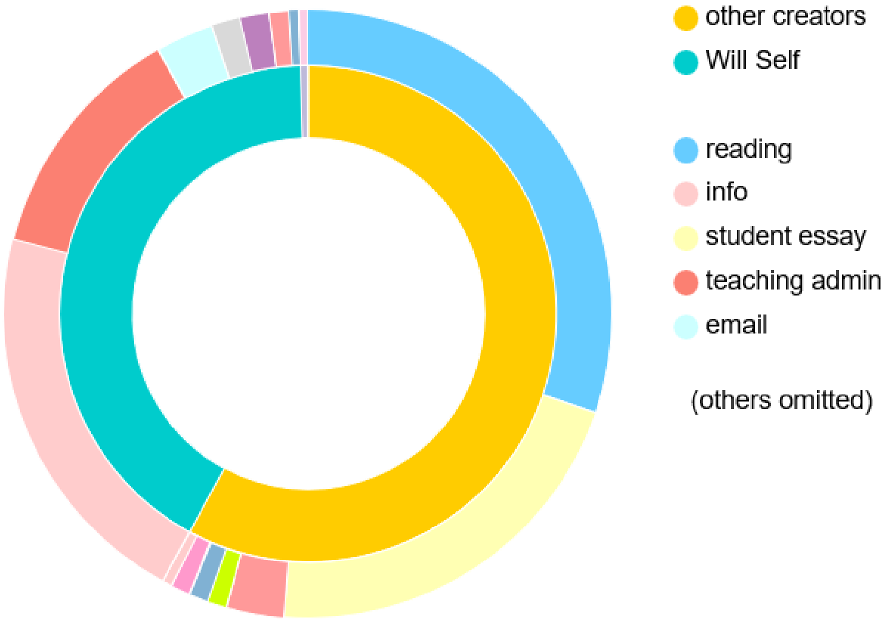
Andrea Levy's notes on Mary Seacole brought to light by IT experts

The writer's scripts for a TV series about the nurse were among those recovered from her old computer by the British Library

BRITISH LIBRARY

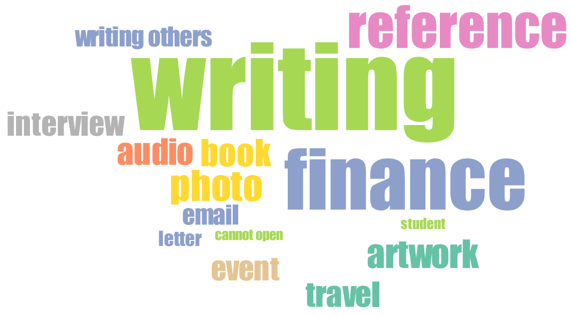
Author& file breakdown

Will Self



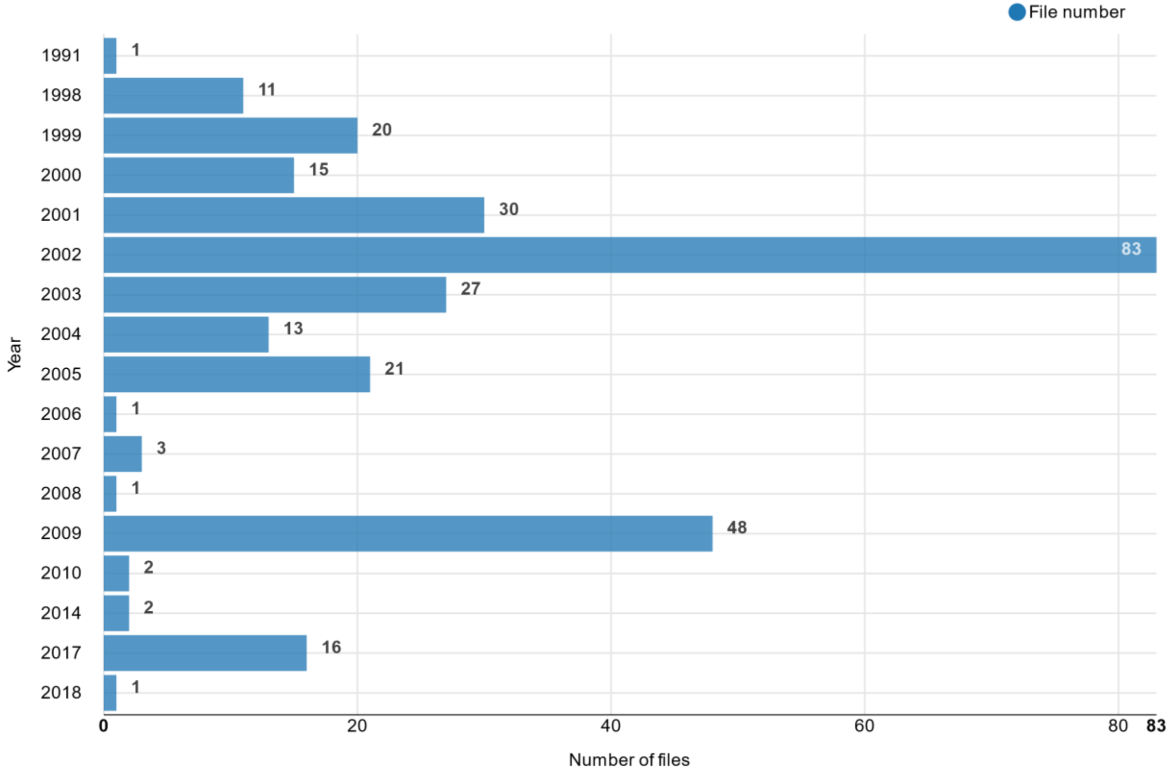
Type of file

Will Self



Small Island/ year

Andrea Levy



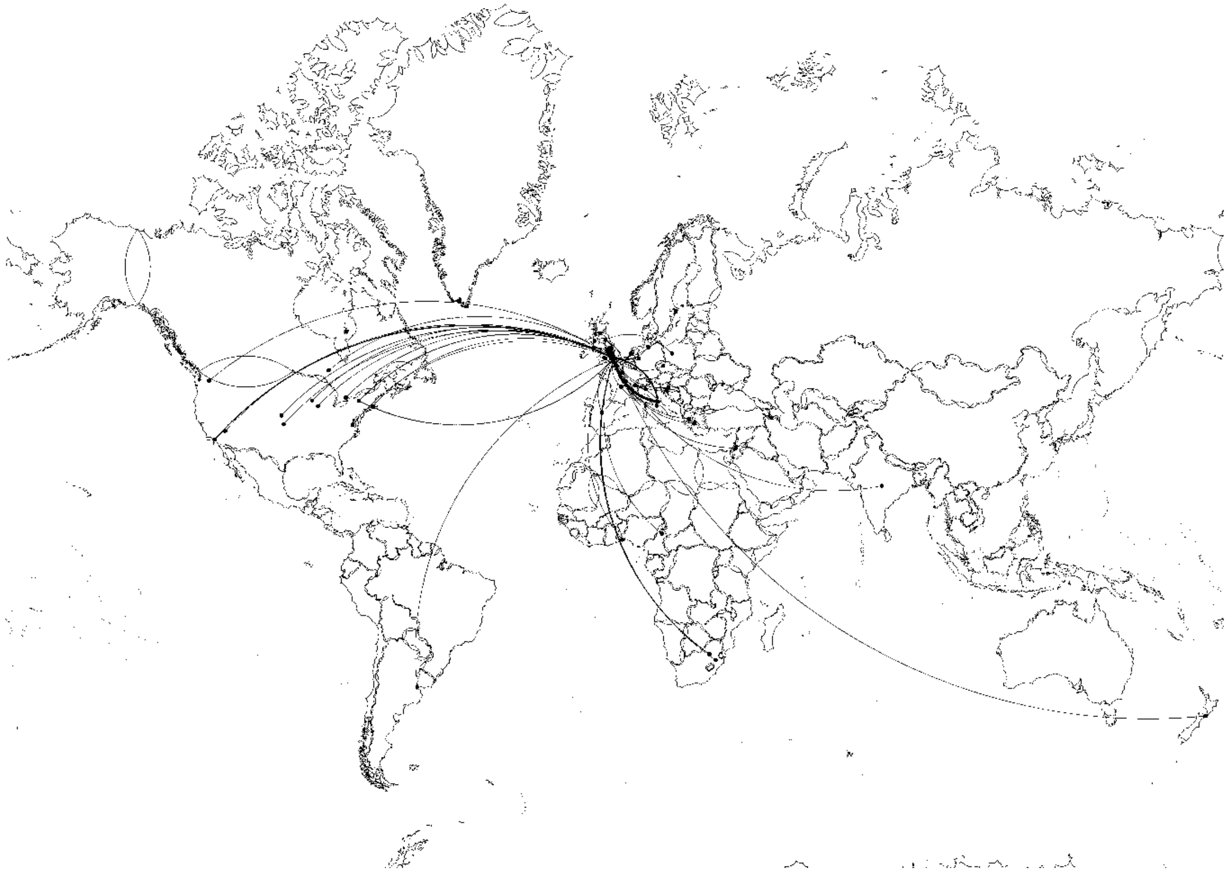
Hybrid Correspondence Collections (2023)

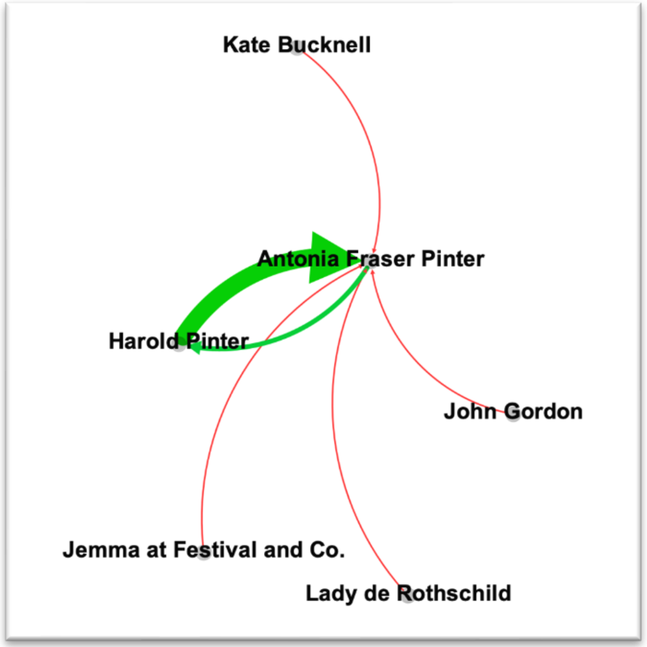
- Used Python and Gephi to visualise paper and digital correspondence collection of Harold Pinter as networks

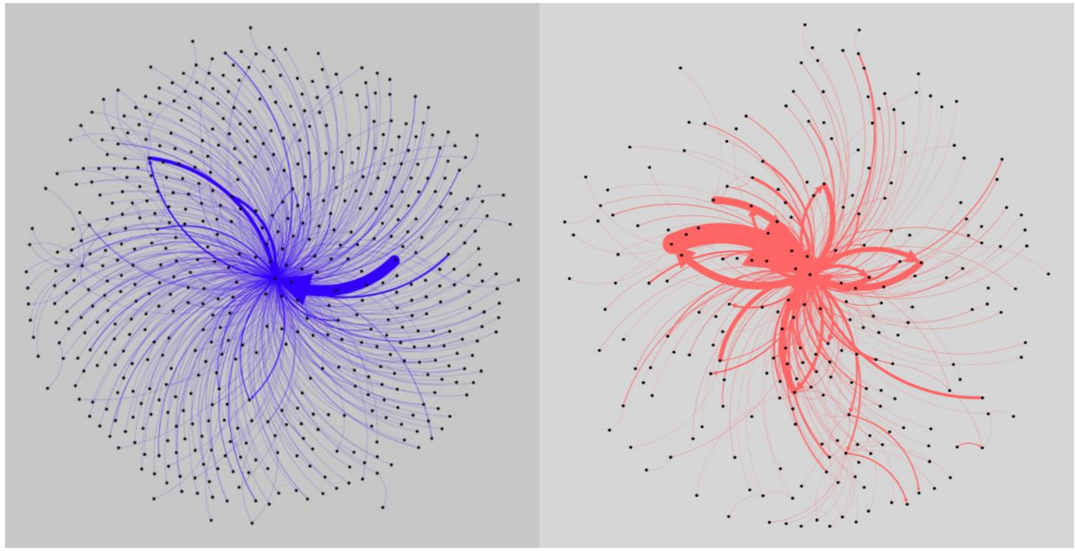
- Manually created metadata sheets for subsection of paper correspondence, Python script for extracting equivalent data from Header and Body of emails.

- Enriched metadata (geo-coded IP addresses, NER for email bodies to extract works by Pinter)

- Outputs: Python script for creation of GDPR compliant email metadata; visualisations in Gephi




Kate Bucknell
Antonia Fraser Pinter
Harold Pinter
John Gordon
Jemma at Festival and Co.
Lady de Rothschild



Automated Migration Workflow (2024-)

- Uses Python and Aspose Total to sort, de-duplicate, arrange, re-name and migrate relevant files to PDF/A for further description and access

- Uses technical metadata, archival principles and limitations of our cataloguing system to make decisions about how to treat material using a strict if/else logic

- Prepares collection material and metadata for manual description and sensitivity review by Cataloguer

Some concluding thoughts

- Transformer models (like Chat GPT) are most useful when connected to the internet.
- With rules based if/else approaches we write and understand the rules.
- Labelling datasets is time and resource intensive (and often complex)
- We need to build up human best practice before we can train good models
- We need resource beyond fixed, time-bound projects to allow us to integrate AI into our work

Thank you

BRITISH LIBRARY