Initial Steps towards Building a Global Registry of Digitized Works

Mike Furlough, HathiTrust
Stuart Lewis, National Library of Scotland

CNI Spring 2020

6 May 2020

PI: Dr. Paul Gooding (Lecturer in Information Studies, University of Glasgow)

Presentation Overview

1. Introduction to the Global Digitised Dataset Network (GDD Network);
2. Work to date: Exploring use cases
3. Work to date: Data clustering & aggregation
4. What might a sustainable, scalable dataset – and related services – look like?

The GDDNetwork

- GDDNetwork – Network to investigate the development of a global dataset of digitised texts.
  - AHRC-funded Research Network (Feb 2019– Jan 2020).
  - Investigating the feasibility and value of a global registry/dataset of digitised texts.
  - More on this la… …t…

HATHI TRUST

LLYFRGELL GENEDLAETHOL CYMRU
THE NATIONAL LIBRARY OF WALES

University of Glasgow

BRITISH LIBRARY

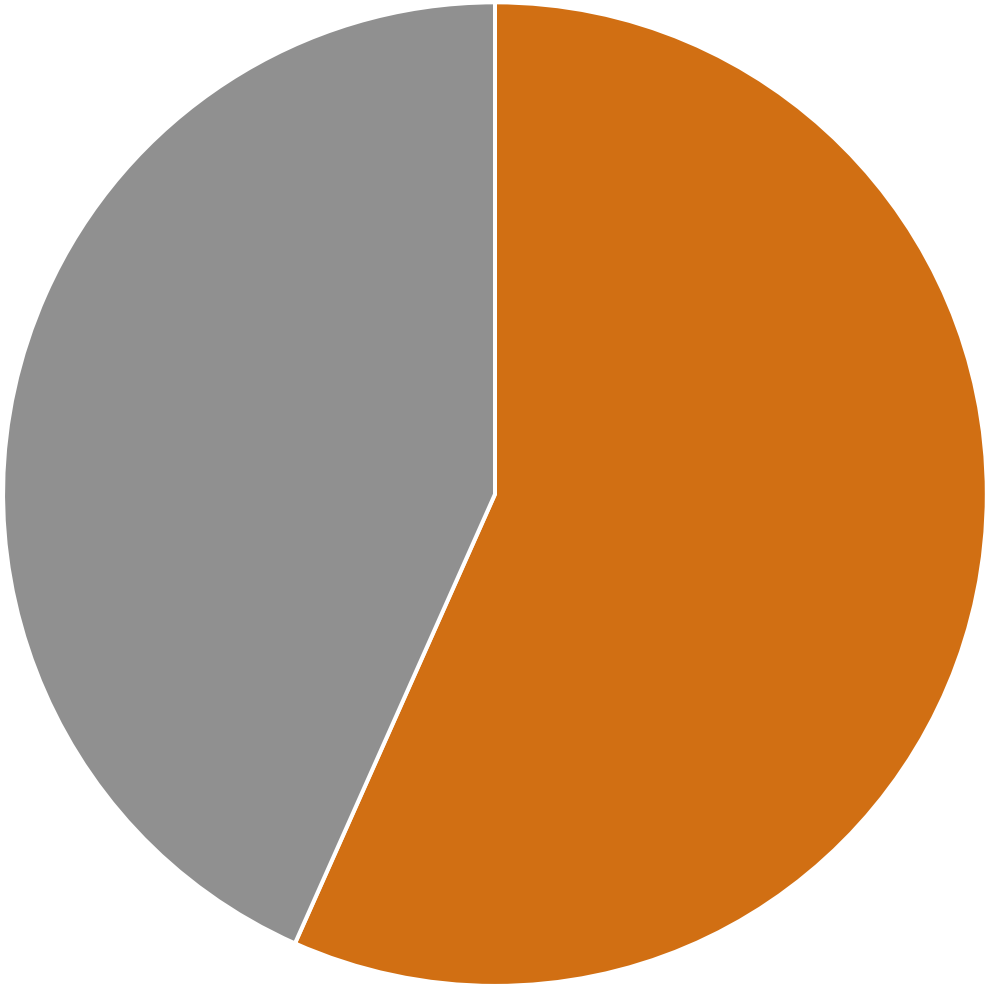
RLUK Research Libraries UK

National Library of Scotland
Leabharlann Nàiseanta na h-Alba

Core question…..

What has been digitized,
by whom,
and where is it?

Mass Book Digitization: Volumes Scanned since c. 2000



**excluding paywalled publisher products

Volumes Deposited in HathiTrust

Volumes NOT deposited in HathiTrust

Investigation questions:

How feasible would it be to aggregate records describing the extent of materials digitized from physical sources?

How useful would this be to scholars, to librarians, and others?

Network Objectives and Deliverables

Undertake a trial matching of data from UK Libraries with the existing HathiTrust dataset of digitised texts.

Hold workshops to explore the range of benefits a global dataset of digitised texts could bring to different groups.

Deliver a dataset that combines HathiTrust and UK Library metadata on digitised texts.

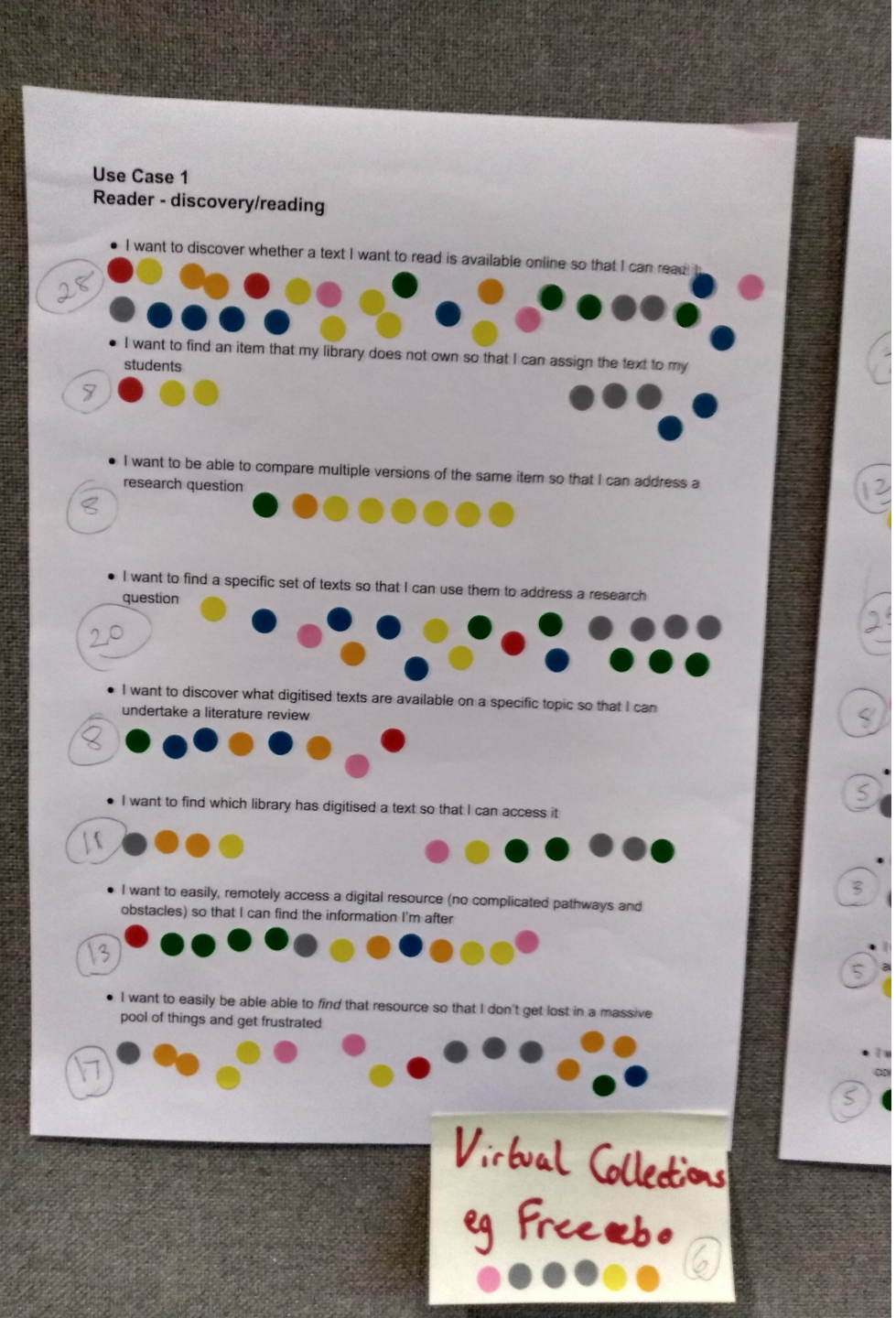
Develop options for an ongoing and sustainable collaborative network of relevant parties that is able to deliver on the ultimate goal of creating a global dataset of digitised texts, along with appropriate services to the scholarly community.

Understanding the use cases: the initial three

Readers wishing to find a digitised text would be able to search quickly and efficiently across all potential sources.

Digital scholars seeking large corpora of texts could easily search and compile links to items across many sources, creating new or bespoke collections.

Libraries undertaking digitisation programmes would be able to discover already digitised texts, and thereby make their own digitisation efforts more efficient by avoiding duplication.



Use Cases for a Global Dataset of Digitised Texts

- Team meeting in Chicago:
  - Brainstorming agile user stores:
    - "As a *…* I want to *…* so that I can *…*"

- London Workshop (June 2019):
  - Further brainstorming to identify additional user stories;
  - "Investment" exercise: voting for preferred use cases in order to suggest priority investment areas;
  - Group discussions around feasibility, key stakeholders, ways forward.

Use Cases: Preliminary Analysis

Five themes emerged:

- **Efficiency, Cost, Impact, Value:**
  - "As a collections manager I want to know what has already been digitised so that I can avoid duplication of effort".
- **Discovery & Access:**
  - "As a reader I want to easily, remotely access a digital resource so that I can find the information I'm after."
- **Provenance:**
  - "As a digital scholar I want to understand the provenance of the dataset so that I can put the digitised materials in context and apply my own relative score to the source (e.g. how much I trust it)."
- **Research:**
  - "As a digital scholar I want to download a list of links to digitised texts from different libraries so thatI can create a corpus specific to my needs."
- **Product/Service Development:**
  - "As a vendor I want to know what libraries have digitised so that I can include a new discovery channel in my product."

Survey learning outcomes

- Some respondents were confused by the concept:
  - Need to scope and, ultimately, explain the service as it is developed.
- Additional categories of interest emerged:
  - Teaching;
  - Positioning of the dataset in relation to other services;
  - Support for library users, and collaboration in research, digitisation and infrastructure;
  - Clear interest in the project among stakeholders.
- But caution advised:
  - Small, self-selecting sample – not necessarily representative of broader stakeholder community
  - Further work needed on areas of concern:
    - Language and geographical reach;
    - Balance between larger and smaller organisations;
    - Data quality and holdings analysis.

Holdings Analysis and Aggregation

- With thanks to the HathiTrust team – Natalie Fulkerson, Josh Steverman, Martin Warin, and Heather Christenson.
- Partner libraries effectively went through a trial “onboarding” process similar to that undertaken by new HathiTrust members.
- Key goals:
  - To identify the extent of overlap between Partner Libraries and HathiTrust;
  - To identify an effective methodology for matching data across the library catalogues – essential to allow accurate deduplication.

HathiTrust's Holdings/ Bibliographic Analysis

Print Holdings overlap analysis

- Supports collection development, shared print program
- Supports some access services
- Informs fee calculation
- Relies on OCLC record number for matching

Bibliographic records stored in Zephir

- MARC format, contributed by depositing libraries
- Multiple copies of the same work means multiple records that describe the same instantiation
- Clustered on OCLC number

Library Records Received

	# of digitized records	# of print records
British Library	516,212	–
National Library of Scotland	10,919	9,640,360
National Library of Wales	2,290	3,224,243
HathiTrust	16, 987, 842	–

BRITISH LIBRARY

nls National Library of Scotland Leabharlann Nàiseanta na h-Alba

HATHI TRUST

LLGC NLW LLYFRGELL GENEDLAETHOL CYMRU THE NATIONAL LIBRARY OF WALES

Matching using identifiers

⇒ Match library holding records to the HathiTrust collection using OCLC number.

- Over 90% of HathiTrust records have OCNs
- But OCNs are *rare* in the records provided to us (less than 1% to 5% to 30%)
- More common in records for undigitized materials

⇒ Match library holding records to the HathiTrust collection using other identifiers, such as ISBN, LCCN, ISSN

- ISBNs date back only to the 1970s:
- ~only 15% of HathiTrust collection….

Data normalization will be a significant undertaking for the

Four more exploratory methods

1. Literal string match of raw title fields in library datasets (MARC 245|abc) against HathiTrust records (MARC 245|abc)
2. Literal string match of normalized title fields in library records against HathiTrust (downcasing, removed non-alpha)
3. Word-by-word match of BL titles against HathiTrust (bag of words approach)
   - Output is a list of candidate OCNs for each record, with a corresponding confidence score.
4. Machine learning: training a support vector (binary) classifier to distinguish between title matches and non-matches.

See blog post https://gddnetwork.arts.gla.ac.uk/index.php/2019/10/05/matching-bibliographic-records/

Analysis outcomes

Duplicate detection/clustering across heterogeneous metadata sources is challenging...

- Short titles, long titles, common titles
- Different manifestations of the same work

...involves tradeoffs…

- Resource-intensive methods yield better results

…and has implications for aggregation:

- Duplicate detection vs. Clustering – how to express relationships to registry users?

Proto-registry: Aggregating Records

"Proto-registry" datafile was specified in grant proposal, but no functionality promised.

Scoped to ensure that records were reasonably complete and comparable between institutions.

Using the HathiFile as a model:

- Identify common fields
- Assess prevalence across project partner records

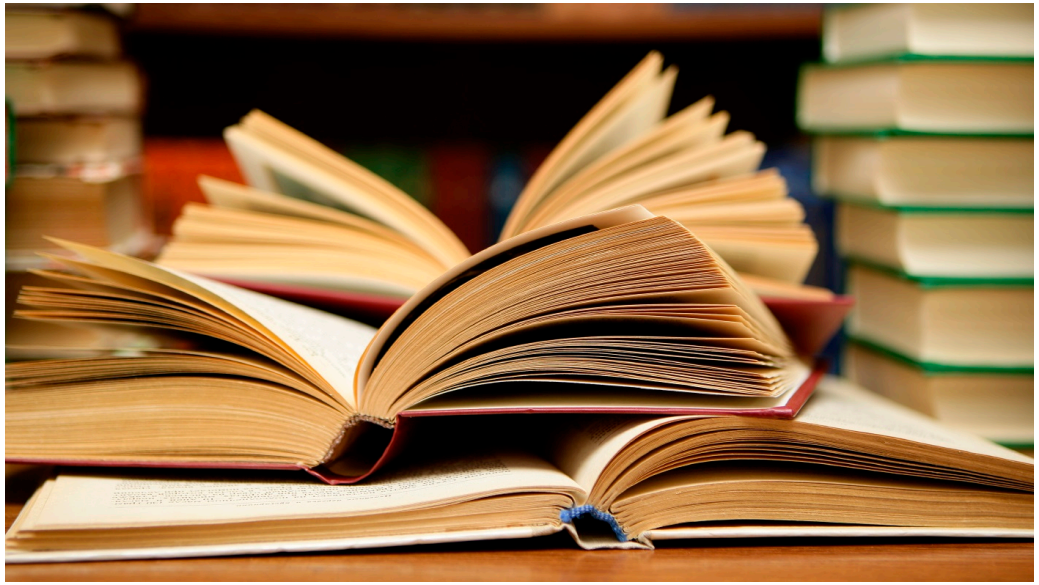
Proto-Registry fields

HathiFile Data Element	Description - Brief	MARC field/s
Volume Identifier	Permanent item identifier	
Title		245|a
Imprint	Publisher + Date of publication	260|bc
Publication Place		008 (bytes 15-17)
Author	Aggregated dataset: https://doi.org/10.34812/fda4-5336	100|abcd, 110|abcd

Conclusions and future work

- Registries are critical but undervalued infrastructure
- Engagement work – sense that there is a need for a resource that specifically addresses digitized texts
- What might a sustainable project look like?
  - Address Metadata costs / what is ‘good enough’
  - Understand metadata requirements and the cost/value of data matching (or not)
  - Scalability
  - Business models, funding, sustainability, community buy-in and participation
  - Continuously updated and accurate
  - Beyond Anglophone collections

Thank you for listening!



GDD Network:
https://gddnetwork.arts.gla.ac.uk/

Final report:
http://eprints.gla.ac.uk/211898/

Aggregated dataset:
https://doi.org/10.34812/fda4-5336

Stuart Lewis: stuart.lewis@nls.uk

Mike Furlough: