



Towards a Global Dataset of Digitised Texts: The GDDNetwork

- OCLC Research Mini-Symposium on the Discovery and Use of Open Collections 19/6/2019
- Dr. Paul Gooding (Lecturer in Information Studies, University of Glasgow)
- paul.gooding@glasgow.ac.uk / [@pmggooding](https://twitter.com/pmgooding)

Presentation Overview

1.) Introduction to the GDDNetwork;

2.) Contexts for the Network;

3.) Work to date:

4.) Next steps and conclusion.

- Developing potential use cases;

- Data matching (HathiTrust);



The GDDNetwork Core Partners

- GDDNetwork – Network to investigate the development of a global dataset of digitised texts.
 - AHRC-funded Research Network (Feb 2019– Jan 2020).
 - Investigating the feasibility of a global registry/dataset of digitised texts.
 - More on this later, but first...



The Research and Funding Context



Arts and Humanities Research Council:

Digital Transformations in the Arts & Humanities;
Focus on the implications of the digital shift
Particularly interested in the emergence of "digital scholarship":

- Data Science;
- Digital Humanities;
- Implications of born-digital archiving;
- Open Publishing and Open Data.



This project responds to a specific call:

Research Networking Scheme for "UK-US Collaborations in Digital Scholarship in Cultural Institutions" – announced in October 2018.

Network Overview

Set out to address a key problem:

- Libraries, archives, and many other organisations are digitising collections, but much of it is uncoordinated:
 - Hard for researchers to make the best use of growing collections;
 - Organisations wishing to target their digitisation efforts are unable to easily collaborate;
 - Other forms of collaboration could emerge (co-ordinated digital preservation?).

Identified several potential beneficiaries:

- Digital scholars seeking large corpora of texts, or metadata pertaining to digitised collections.
- Readers wishing to find a digitised text.
- Libraries undertaking digitisation programmes.

Aim to verify the utility of such a resource, and to investigate the feasibility of developing a continuously updated international registry of digitised texts.

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.

Contexts for the Network



- 1.) “The Collective Collection”: “One important trend is that libraries and the organizations that provide services to them will devote more attention to system-wide organization of collections - whether the “system” is consortium, a region, or a country” (Dempsey, 2013).
- 2.) Cross-border challenges to collaborative global efforts: Copyright & Intellectual Property; “Ownership” of library collections.
- 3.) Mass digitisation as a driver of change – for researchers and for libraries.
- 4.) The growth of data-driven research - Data Science, Digital Humanities – that relies upon digital collections from libraries.
- 5.) What does it mean for us to call a resource global? Linguistically, culturally, technologically, practically?

A brief note on our definition of texts...

- This has defined in terms of what we can deliver with the time and resource available, i.e.:
 - Focus primarily on **monographs**.
- But this is flexible – and
- Some questions we might want to consider:
 - What digitised texts would we want to see included?
 - How feasible is their inclusion?
 - What might the limits of such a resource be? Where do we draw the line on what is included, and why?

Work to
Date

1.) Developing Use Cases for a global dataset of digitised texts.

2.) Holdings Analysis.

3.) Community engagement and workshops.



Holdings Analysis (Led by HathiTrust)

- With thanks to HathiTrust for the work and slides – 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.

HathiTrust Datasets - overview

Bibliographic records stored in Zephir

- MARC format
- Contributed by HathiTrust member libraries as part of ingest
- Clustered on OCLC number

HathiFile

- Tab-delimited text file representing every item in the collection
- Derived from Zephir bibliographic records, plus rights and access codes, various HathiTrust-generated administrative identifiers

Library Records received

- MARC records for:
 - Digitised monographs;
 - Print holdings.
- Varied according to format and availability of records:

Organisation	No. of digitised records	No. 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

Approach
#1: Standard
HathiTrust
Overlap
Analysis

	# digitized records	# OCNs	# matching	% matching
British Library	516,212	611	130	0.025
National Library of Scotland	10,919	561	243	2.22
National Library of Wales	2,290	744	101	4.41

- Attempt to match library holdings records to the HathiFile using OCLC number (OCN)
- OCNs present in library record (in the MARC 035 field):
- Digitised items only.

Approach
#2: Look for
other usable
identifiers

	# digitized records	# print records	# ISBNs - digital	# ISBNs - print	% ISBNs - print
British Library	516,212	-	34	-	-
National Library of Scotland	10,919	9,640,360	55	2,709,837	28*
National Library of Wales	2,290	3,224,243	17	3,128,171	97**

- Locate other possible identifiers in library records to match against corresponding fields in the HathiFile (or, later, in Zephir)
- For example - ISBN

Exploratory method #1

	# digitized records	# matches	# matches to multiple clusters	% overlap
British Library	516,212	4,559	2,255	0.88
National Library of Scotland	10,919	343	131	3.14
National Library of Wales	2,290	51	9	2.23

- Literal string match of title fields in library datasets (MARC 245|abc) against title fields in the Zephir preferred records (MARC 245|abc)

Exploratory method #2

	# digitized records	# matches	#matches to multiple clusters	% overlap
British Library	516,212	39,815	18,298	7.71
National Library of Scotland	10,919	1,746	837	16
National Library of Wales	2,290	253	83	11.04

- Literal string match of processed title fields in library datasets (MARC 245|ab) against title fields in Zephir preferred records (MARC 245|ab)
- Processing consisted of:
 - Downcasing
 - Removing non-alpha characters

Exploratory Method #3

Attempt A: Word-by-word match on the MARC 245 | abc fields in the BL dataset against the HathiFile (MARC 245 | *)

- For each BL title, lowercase, eliminate stopwords, and produce a “bag of words”
- Search the HathiFile for each of the words in the bag (not a literal string search, word order is not important)
- Determine precision and recall, calculate an average confidence score, rank by score

Output is a list of candidate OCNs for each record, with a corresponding confidence score.

Example

SCORE	PRECISION	RECALL	MATCH_TITLE
0.514	0.600	0.429	The martyrdom of St. Peter and St. Paul; a poem. By George Burgess.
0.514	0.600	0.429	St. Catharine of Siena. The Seatonian prize poem for 1948.
0.857	1.000	0.714	St. Paul at Philippi : a Seatonian poem / by Thomas E. Hankinson.

- ## BL title: "St. Paul at Philippi. A Seatonian poem."
- ## Bag of words: st,paul,philippi,seatonian,poem

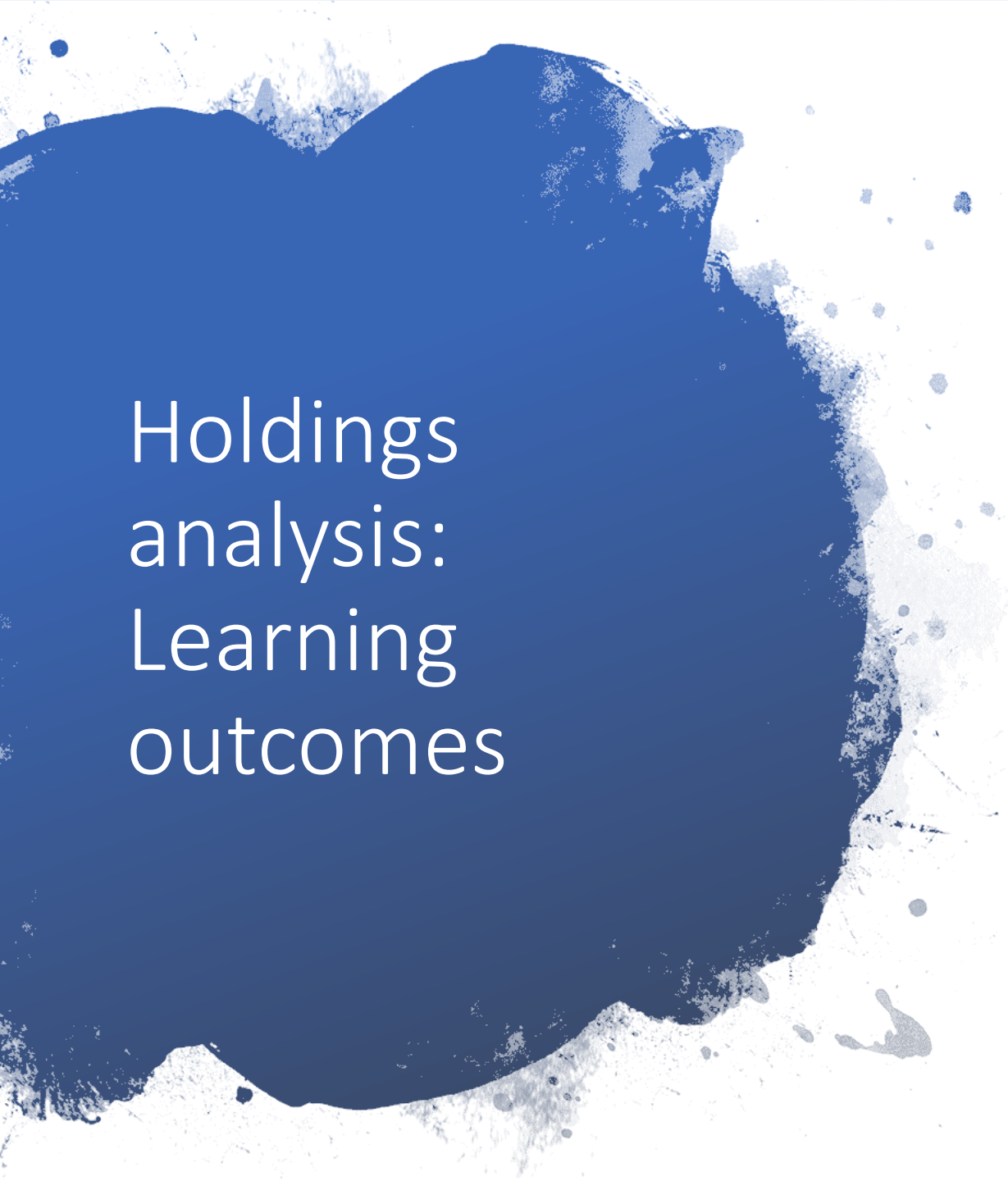


Exploratory Method #4

- Continues the work of Michael Morris-Pearce, a former HathiTrust colleague at CDL
- Query: Can you train a support vector machine (SVM) classifier to distinguish between title matches and non-matches?
- Machine Learning process:
 - Setup
 - Training/Iteration phase
 - Implementation phase

Results (Take these with a BIG pinch of salt)

	# of clusters in test set	# of predicted clusters	Precision	Recall
Polynomial	5989	5558	0.982	0.911
Gaussian RBF	5989	5948	0.975	0.968



Holdings analysis: Learning outcomes

- Duplicate detection is hard...
 - Short titles, long titles, common titles;
 - Different manifestations of the same work.
- ...Involves tradeoffs:
 - Resource-intensive methods yield better results.
- Implications for aggregation:
 - Duplicate detection (overlap) vs. Clustering – how to express relationships to registry users?

Use Cases for a Global Dataset of Digitised Texts

Use Case 1 Reader - discovery/reading

29 • I want to discover whether a text I want to read is available online so that I can read it

30 • I want to find an item that my library does not own so that I can assign the text to my students

31 • I want to be able to compare multiple versions of the same item so that I can address a research question

20 • I want to find a specific set of texts so that I can use them to address a research question

32 • I want to discover what digitised texts are available on a specific topic so that I can undertake a literature review

11 • I want to find which library has digitised a text so that I can access it

15 • I want to easily, remotely access a digital resource (no complicated pathways and obstacles) so that I can find the information I'm after

17 • I want to easily be able to find that resource so that I don't get lost in a massive pool of things and get frustrated

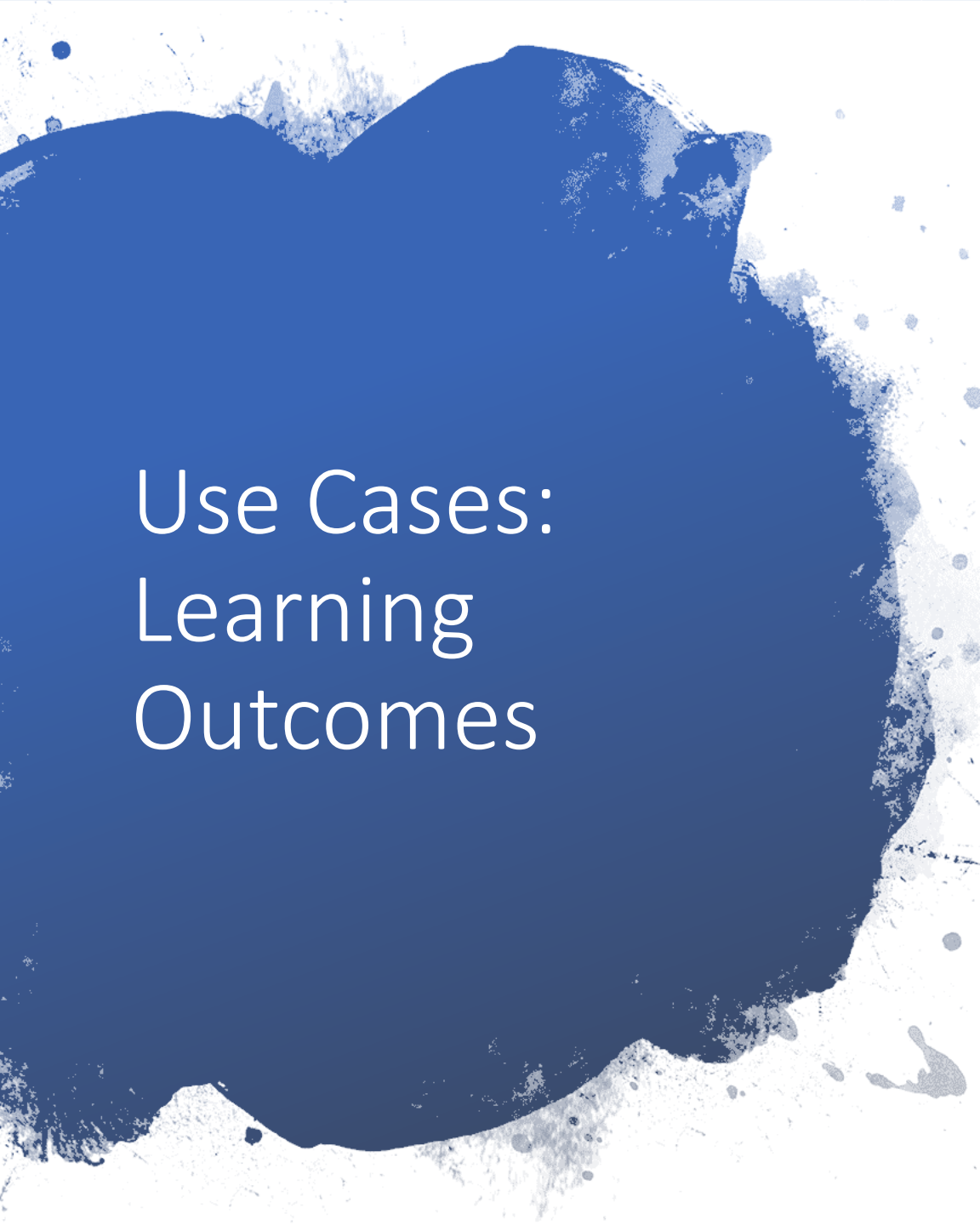
Virtual Collections
eg Freeabo

- Team meeting in Chicago:
 - Brainstorming agile user stories:
 - “As a *…* I want to *…* so that I can *…*”
- London Workshop:
 - 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 emerge, and most popular in each theme:
 - 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 that I can create a corpus specific to my needs.”
 - Product/Service Development:



Use Cases: Learning Outcomes

- Bias towards “Research” due to presence of several involved in digital scholarship.
- Library service providers underrepresented in network to date, reflected in lack of use cases for vendors.
- Need to identify and reach out to new stakeholder groups in remaining project time.
- BIG ONE: the scope and extent of the dataset needs careful definition:
 - Many assumed case studies were built upon the idea that it would provide direct access to digitised full text.
 - Focus to date has been primarily on unifying metadata, NOT aggregating full text.



Community Engagement and Workshops

- Two workshops over the course of the year:
 - 1.) Workshop held at the British Library, 10th June 2019.
 - Objectives:
 - To refine, test, and validate our assumptions about the project;
 - To identify potential valuable uses of such a resource;
 - To identify possible paths to developing this resource.
 - 2.) Workshop to be held at the University of Glasgow, December 2019.
 - Objectives:
 - To publicly present the prototype dataset, and the results of our overlap analysis;
 - To identify possible next steps towards a global dataset of digitised texts;
 - To establish feasibility of the various options.
- Ongoing programme of awareness raising:
 - Today!
 - HathiTrust and RLUK membership engagement;
 - DCDC Conference (Nov 2019);
 - More to follow.



Future Plans

- Two key deliverables (Dec 2019):
 - Prototype dataset and results of overlap analysis and data matching;
 - Accompanying report into broader context for, and benefits of, a potential global dataset for digitised texts.
- December Workshop (University of Glasgow):
- December 2019/January 2020: Scope potential future work, feasibility study based on functionality of prototype dataset:
 - Identify gaps (expertise, data, infrastructure, funding) necessary to scale up to meet a broader range of use cases.
 - Clearly identify and express benefits and scope of follow-up work.

Thank you for listening!

Any questions?

Contact:

paul.gooding@glasgow.ac.uk

[@pmgooding](#)

Project website:

<https://gddnetwork.arts.gla.ac.uk/>