

The GDD Network: Towards a Global Dataset of Digitised Texts

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

DCDC2019

paul.gooding@glasgow.ac.uk

Presentation Overview

1. Introduction to the Global Digitised Dataset Network (GDDNetwork);
2. Contexts for the network;
3. Work to date:
 - Developing potential use cases;
 - Data matching (HathiTrust);
4. What might a sustainable, scalable dataset – and related services – look like?

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?

Work to Date

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

2.) Holdings Analysis.

3.) Community engagement and workshops.

Use Case 1
Reader - discovery/reading

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

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

8 • 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

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

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

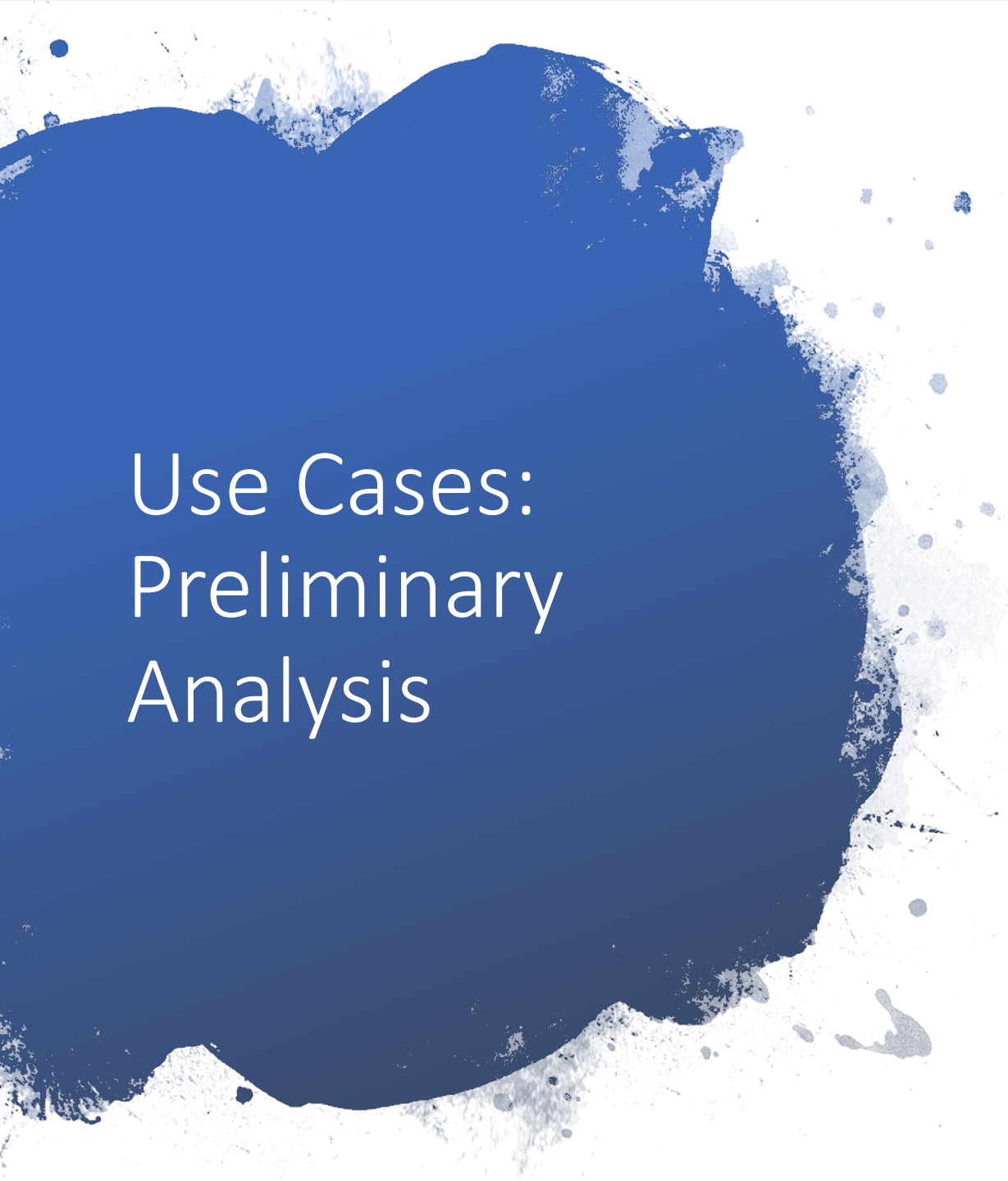
13 • I want to easily, remotely access a digital resource (no complicated pathways and obstacles) so that I can find the information I'm after

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

Use Cases for a Global Dataset of Digitised Texts

- Team meeting in Chicago:
 - Brainstorming agile user stories:
 - “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.

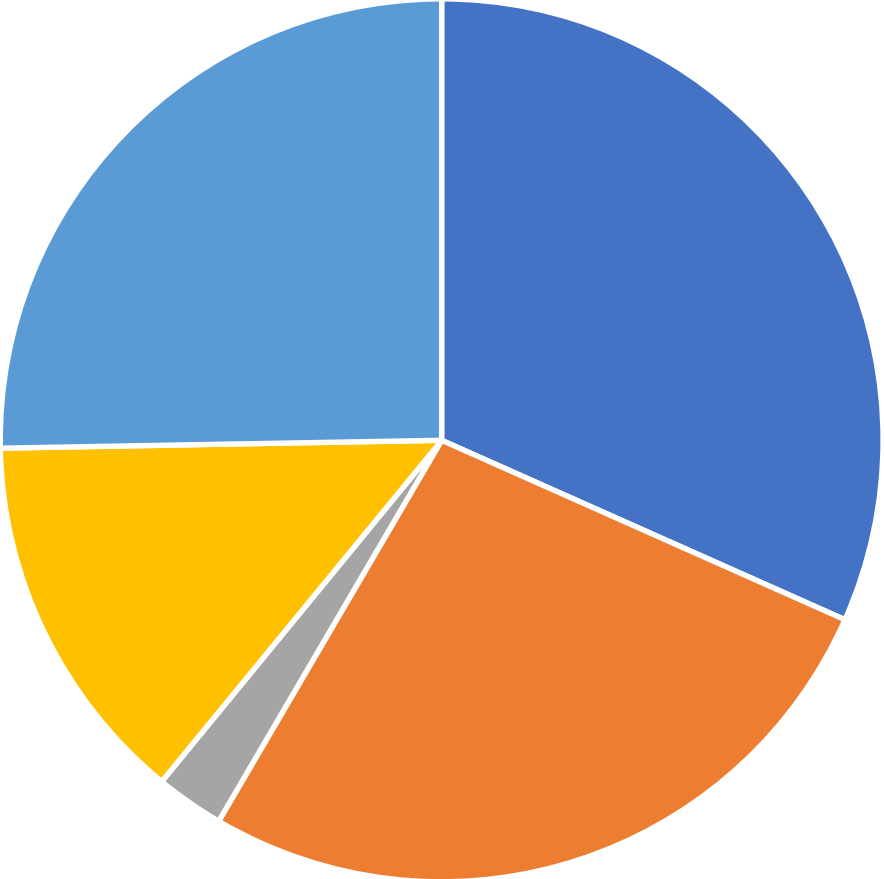


Use Cases: Preliminary Analysis

- 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 that I 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.”

Use cases: Expert Workshop

Most Popular Use Cases: June 2019 Expert Workshop (878 votes)



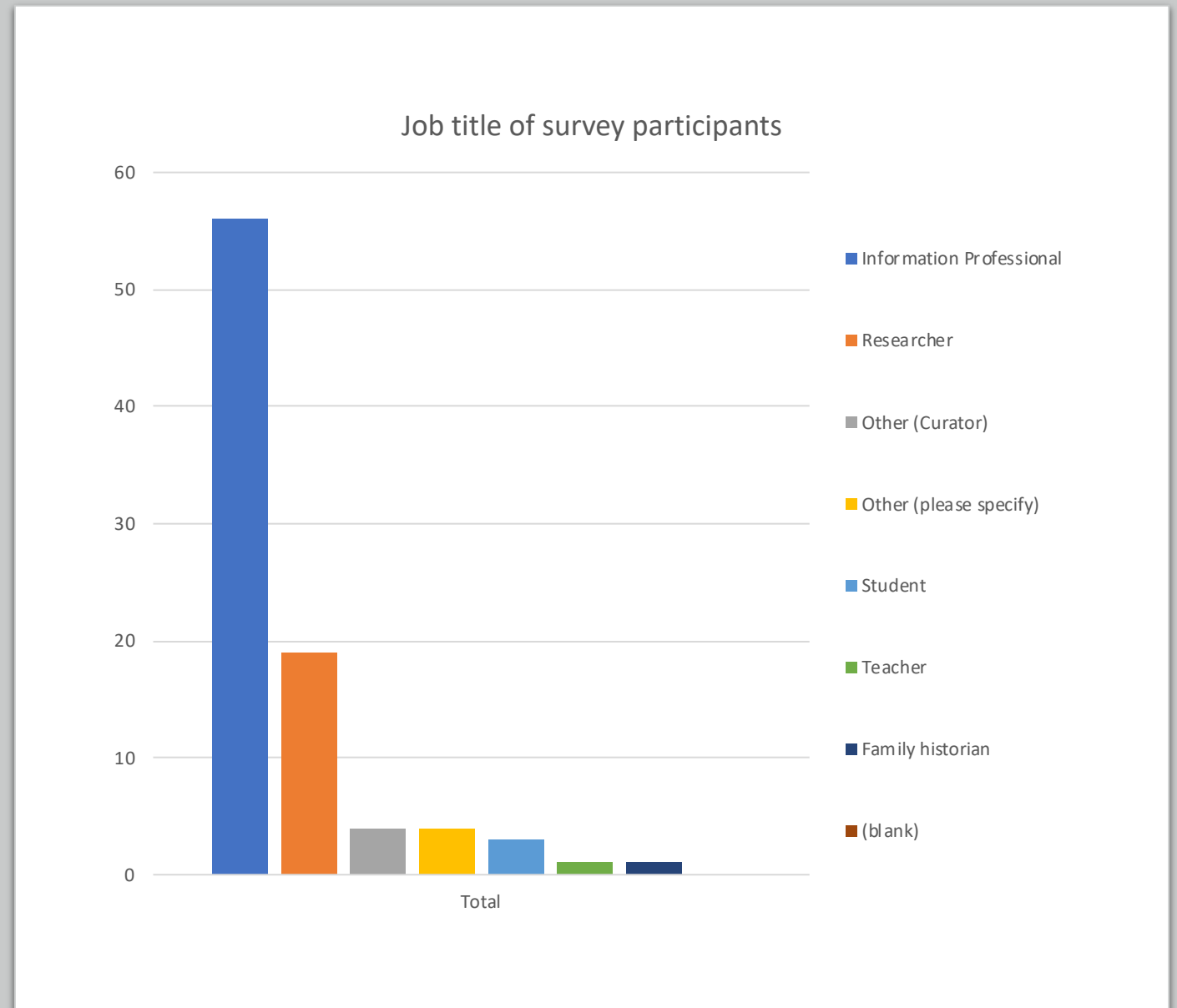
■ discovery & access ■ efficiency, cost, impact, value ■ product / service development ■ provenance ■ research

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.

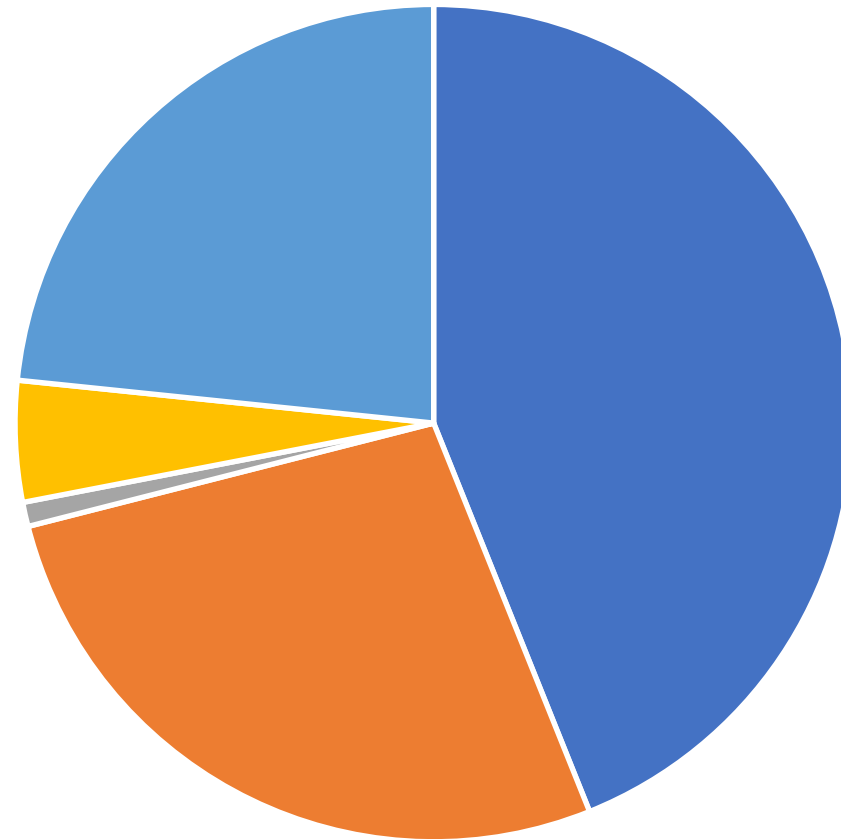
Use cases: Survey

- Lightweight survey designed to ensure stakeholder communities were targeted (right)
- Survey allowed people to type their own answers, which were then coded in line with categories used at expert workshop.
- Focus on this stage on “benefits” – some responses queried this, but we wanted to know what use cases would benefit users the most, should a global digitised dataset exist.
- 86 usable responses gathered between 28th July 2019 and 11th October 2019.



Q.1) How would it help you improve your current work practices (n=86)?

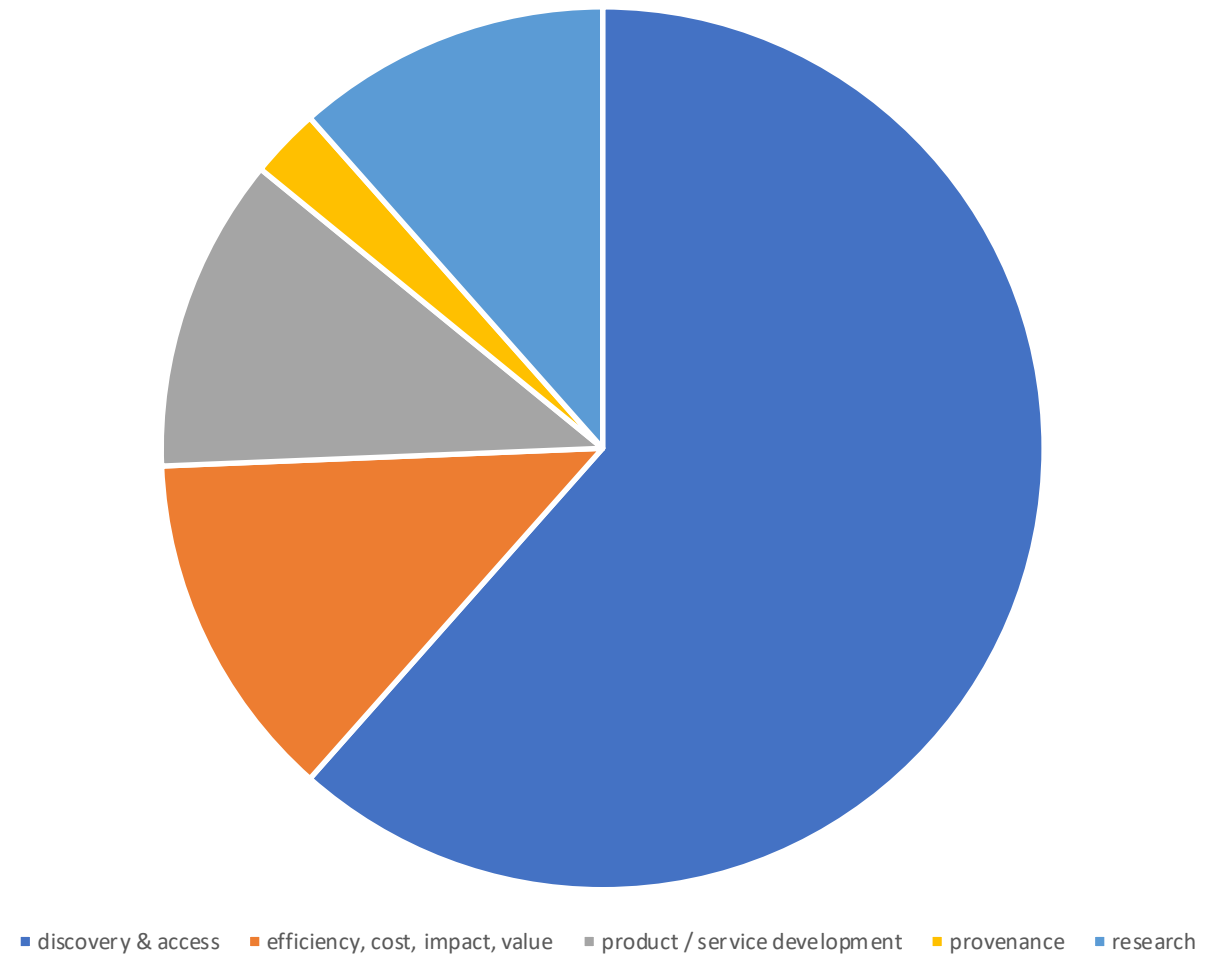
Q.1) How would it help you improve your current work practices (n=86)?



■ discovery & access ■ efficiency, cost, impact, value ■ product / service development ■ provenance ■ research

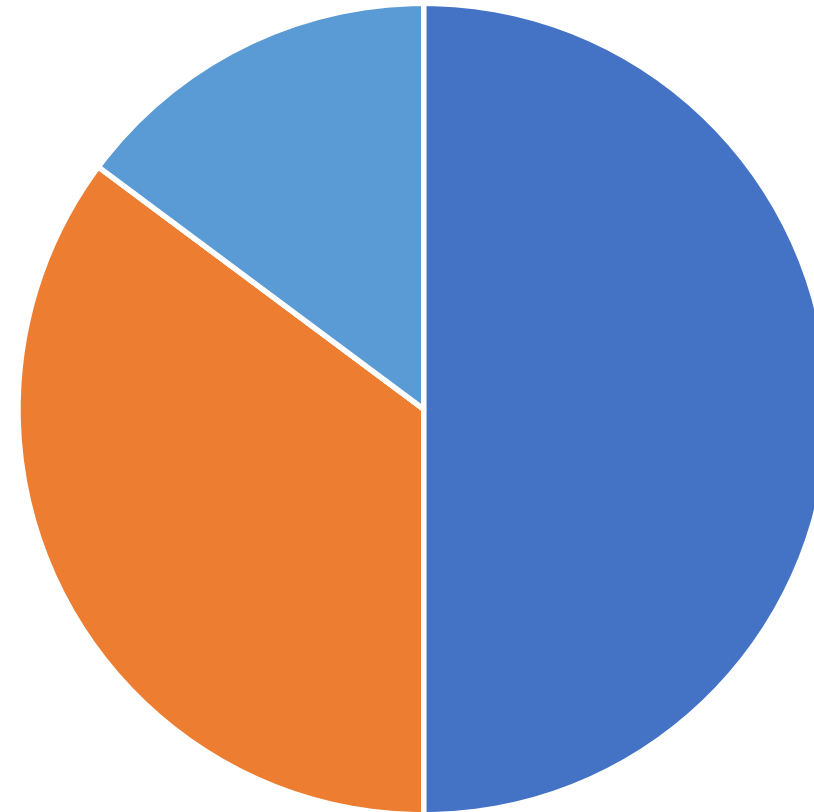
Q.2) What would it enable you to do that you can't do now (n=81)?

Q.2) What would it enable you to do that you can't do now (n=81)?



Q.3) What would be the broader benefits for others in your role (n=79)?

Q.3) What would be the broader benefits for others in your role (n=79)?



■ discovery & access ■ efficiency, cost, impact, value ■ product / service development ■ provenance ■ research



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.



HATHI
TRUST

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 – essential to allow accurate deduplication.

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

Approaches to Data Matching

1. Standard HathiTrust overlap analysis:
 - 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.
 2. Look for other usable identifiers:
 - Locate other possible identifiers in library records to match against corresponding fields in the HathiFile (or, later, in Zephir)
 - For example – ISBN
- Lower than expected matching rates, and inconsistent coverage of ISBNs:

	# 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**

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

	# 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

Results (Take these with a BIG pinch of salt)

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.

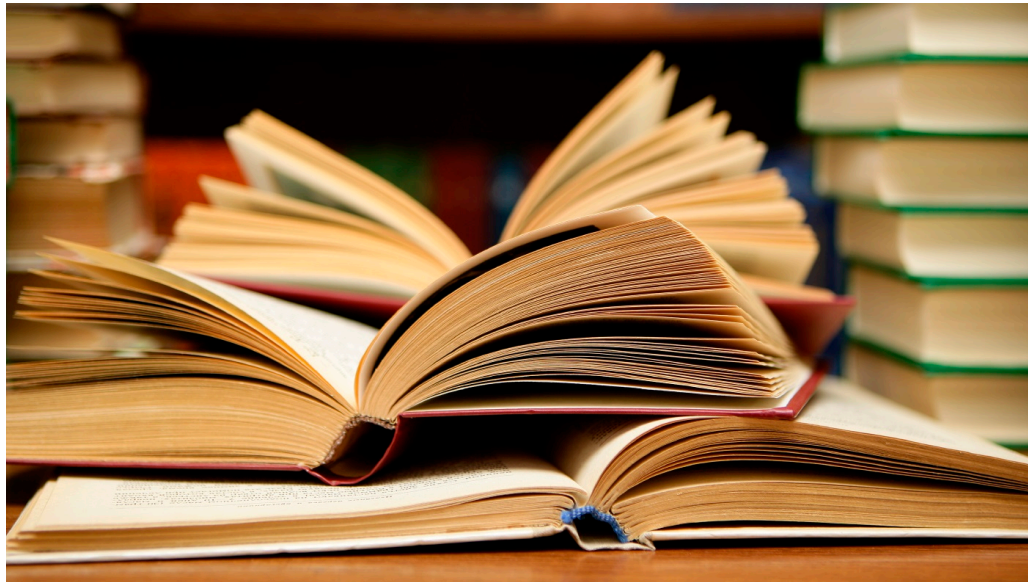
Implications for aggregation:

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

Conclusions and future work

- Engagement work – sense that there is a need for a resource that specifically addresses digitised texts:
 - Follow-on work required to define scope of: materials / participants / functionality / geographic reach / language coverage.
- What might a sustainable project look like?
 - Scalability
 - Requirement for effective business model / planning to ensure regular updating.
 - Limited value if this is a one-off ingest of materials – must be on an ongoing service.
 - Must extend beyond the US/UK context of the original funding call to be considered truly “global”.
- Future work:
 - Identify possible paths and business models;
 - Identify overlap with existing resources, and potential routes to collaboration;
 - Address key research challenges around provenance, data quality and version control.

Thank you for listening!



Any questions?

Contact:

paul.gooding@glasgow.ac.uk

[@pmgooding](#)

Project website:

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