DIGITAL
LIBRARY
FUTURES

elegaldeposit.org

Subjectifying Library Users to the Macroscope Using Automatic Classification Matching

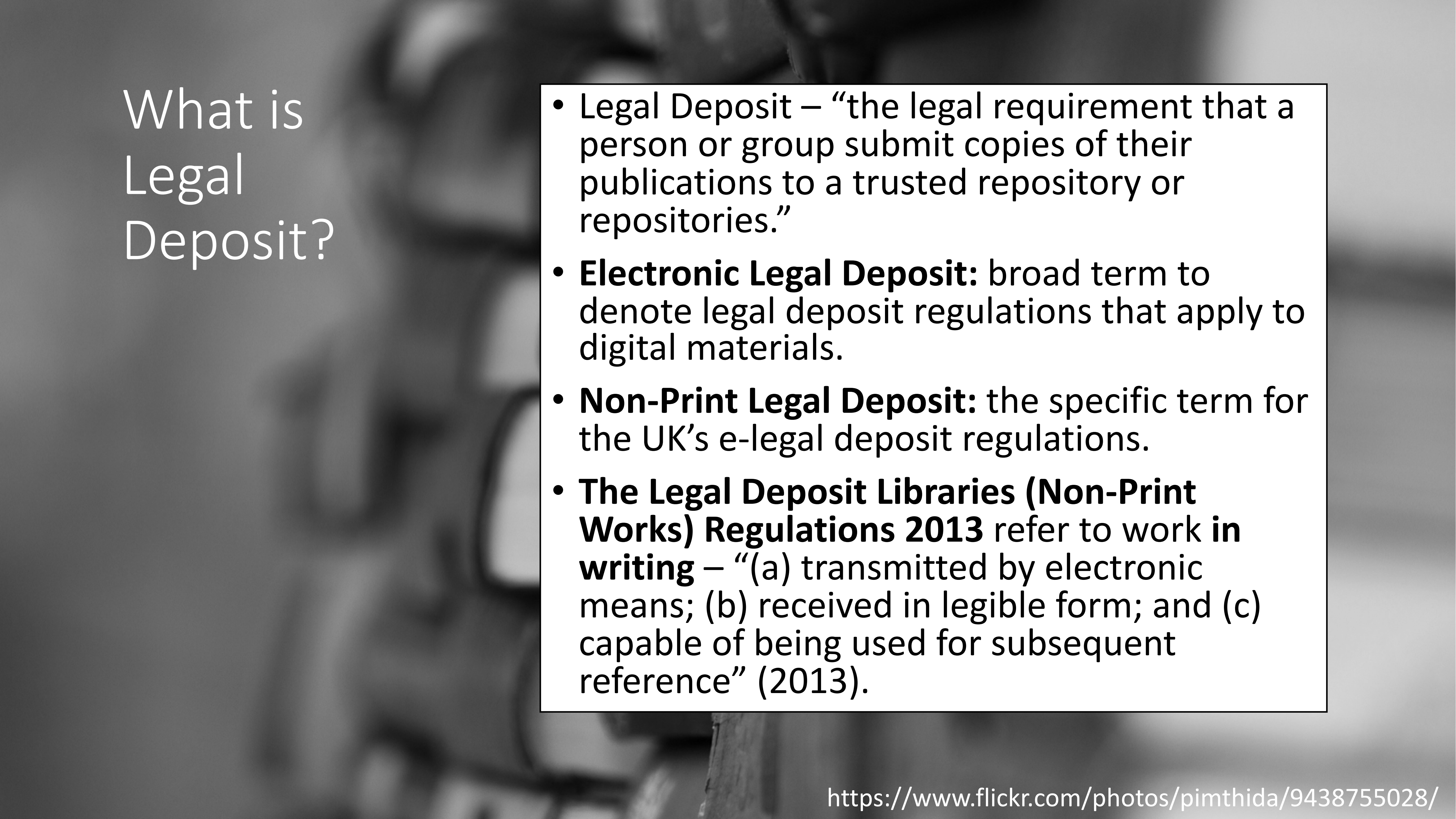
Paul Gooding (University of Glasgow); Melissa Terras, Mike Bennett, Richard Hadden (University of Edinburgh); Linda Berube (University of East Anglia); @pmgooding / paul.gooding@glasgow.ac.uk

Talk Outline

- Introduction:
  - What is Legal Deposit?
  - Studying users of Non-Print Legal Deposit in the UK
- Methodology
- Findings and Discussion:
  - Results of the Subject-based analysis.
  - The problem at the heart of the method: is there a better way to classify the material than existing library classification?

What is Legal Deposit?

- Legal Deposit – “the legal requirement that a person or group submit copies of their publications to a trusted repository or repositories.”
- **Electronic Legal Deposit:** broad term to denote legal deposit regulations that apply to digital materials.
- **Non-Print Legal Deposit:** the specific term for the UK’s e-legal deposit regulations.
- **The Legal Deposit Libraries (Non-Print Works) Regulations 2013** refer to work **in writing** – “(a) transmitted by electronic means; (b) received in legible form; and (c) capable of being used for subsequent reference” (2013).



https://www.flickr.com/photos/pimthida/9438755028/

The Origins of UK Legal Deposit

- 1610: Informal agreement between Sir Thomas Bodley (founder of the Bodleian Library) and the Stationer's Company:
  - Bodleian could claim a copy of everything printed under Royal License.
- 1662: First legal framework for legal deposit in the UK – extended Royal License to Cambridge University Library.
- 1709/1710: Copyright Act under Queen Anne.
- 1753: Establishment of British Museum;
  - Until this date the Bodleian Cambridge University Libraries were the de facto national libraries of the United Kingdom.
- 1753-1911: Various minor changes, but…

“Non-Print Legal Deposit” in the United Kingdom

- “Legal Deposit Libraries (Non-Print) Regulations 2013”:
  - Bring electronic publications into line with printed materials, and cover:
    - Websites;
    - e-Journals;
    - e-Books;
    - Digital Newspapers;
    - Digital Maps.
- Users can access electronic materials within the six legal deposit libraries.
- But what does this mean for us? We are attempting to investigate the following key research problems:

Access to NPLD Materials in the UK

1.) Reader access to NPLD materials is limited to computer terminals located on premises controlled by the legal deposit libraries (part 1, regulation 2).

2.) Materials must only be accessible concurrently to readers via one computer at each legal deposit library (part 4, regulation 23).

3.) For materials published online, seven days must elapse between the date of delivery of that material, and the date on which it is made available (part 4, regulation 24).

4.) A copyright owner may request in writing that certain materials should be embargoed for a specific period. Deposit libraries are bound to comply with such requests, provided that:

- The period for which materials are withheld is limited to three years from the date of the request;
- The deposit library is satisfied that, during the requested timeframe, viewing of the relevant materials by readers would, or would be likely to, “unreasonably prejudice the interests of the person making the request” (part 4, regulation 25).

5.) Deposit libraries are permitted to produce and allow access to copies of non-print work on their premises for a visually impaired person, if copies of the relevant material are not commercially available in an accessible form (part 4, regulation 26).

Methodology

- Marcia Bates observes that scholarly communication practices function differently across domains, and that "these differences *do* make a difference" (1998: 1,200).
  - So we should be able to identify differences in behaviour by studying which subjects are requested by users.
- Access restrictions make it easy to ensure that we get hold of a complete dataset of NPLD usage statistics.
- This study is part of an established tradition of user studies in Digital Humanities:
  - Focus on user behaviour with digital resources.
  - Web log analysis used commonly for over twenty years.
  - Fewer studies have engaged with critical humanistic perspectives to inform approaches to the data.

Research Questions

- What insights into users of Non-Print Legal Deposit Collections can be derived from automatic classification matching?
- What limitations are created through the use of existing classification schemes, and how might DH/LIS scholars collaborate to further develop ethical analysis of large-scale library datasets?

What insights into users of Non-Print Legal Deposit Collections can be derived from automatic classification matching?

The datasets

- Two datasets: both contained all NPLD requests within UK Legal Deposit Library reading rooms – 31st July 2015 to 31st March 2017:
- Bibliographic metadata relating to titles requested from fixed terminals:
  1. Metadata for all **eBook** title requests – total 91,809 requests (title-level).
  2. Metadata for all **eJournal** article requests – total 36,505 requests (article-level).
- Metadata provided: date and time of access request; originating legal deposit library; title of book or article; journal title (where applicable); publisher; and ISBN or ISSN.
- Provided as CSV file and cleaned in OpenRefine.

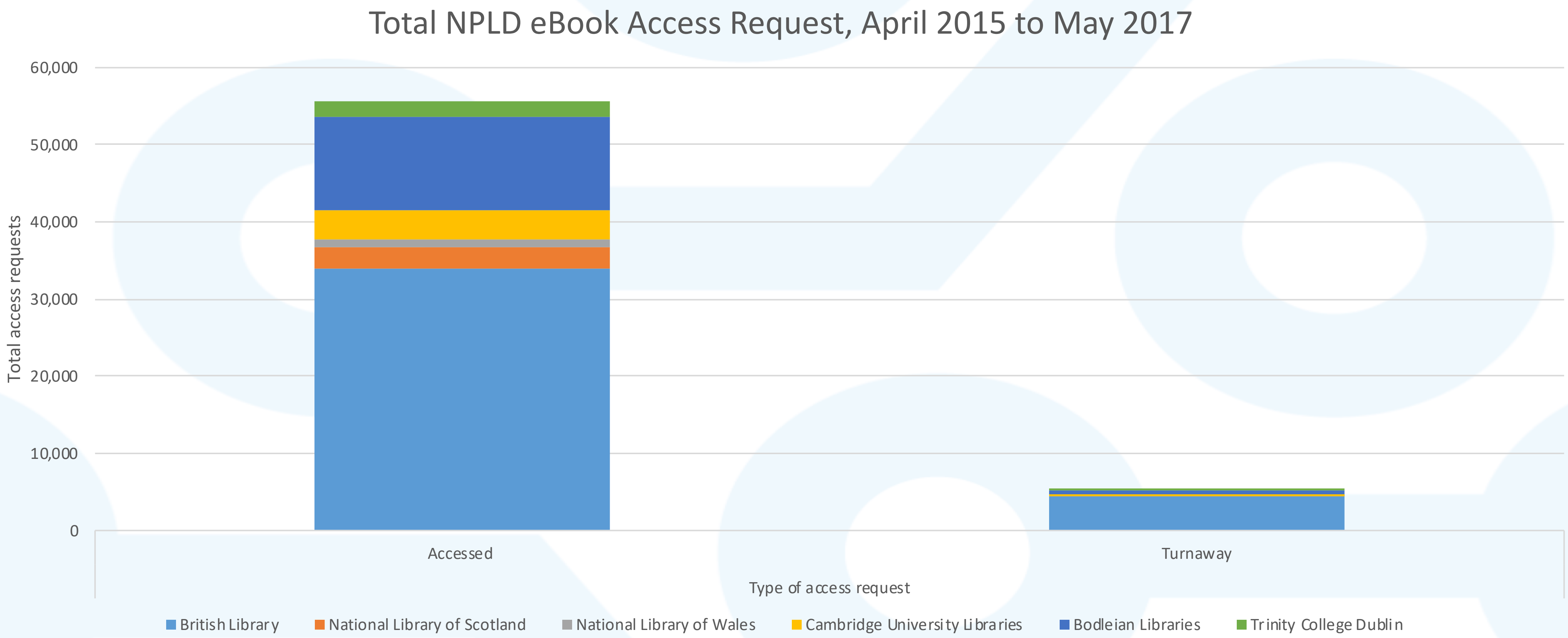
The *Subjectify* Tool

- *Subjectify* is a Python-based tool:
  - Required to preserve anonymity: no identifiable information about users, but possible to infer information about users from the works they consulted.
  - Ethically necessary to consider microanalytic approaches – and to identify meaningful usage patterns.
- Uses the OCLC Classify2 API Service to automatically obtain Dewey Decimal (DDC) and Library of Congress (LCC) classmarks:
  - Author, title, ISBN, and ISSN data taken from a provided CSV file;
  - Tool designed to work on varied data sources, with different options for how to locate relevant fields;
- Discarded unclassified records and used the remaining records to identify subject-based patterns of usage of NPLD materials.

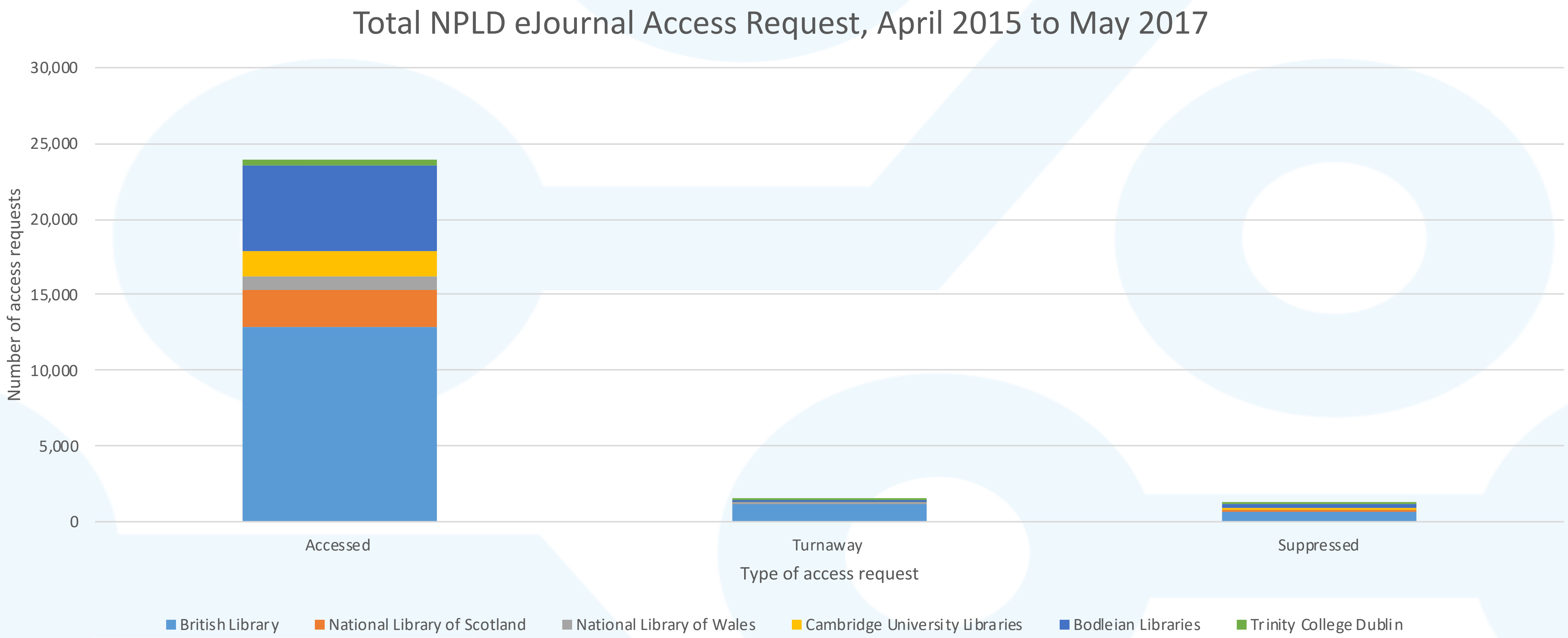
Tool Accuracy

- Subjectify found a matching classmark for:
  - 76.42% of eBooks;
  - 55.53% of eJournals.
- Two main reasons for this:
  - eJournal records often missed key data fields that aided recall in comparison to eBooks;
  - Many records did not have a corresponding classmark in OCLC.
- Manual classification unlikely to be significantly more accurate:
  - Manual spot check showed similar rates of accuracy for eBooks.

Total Usage Statistics for NPLD eBooks

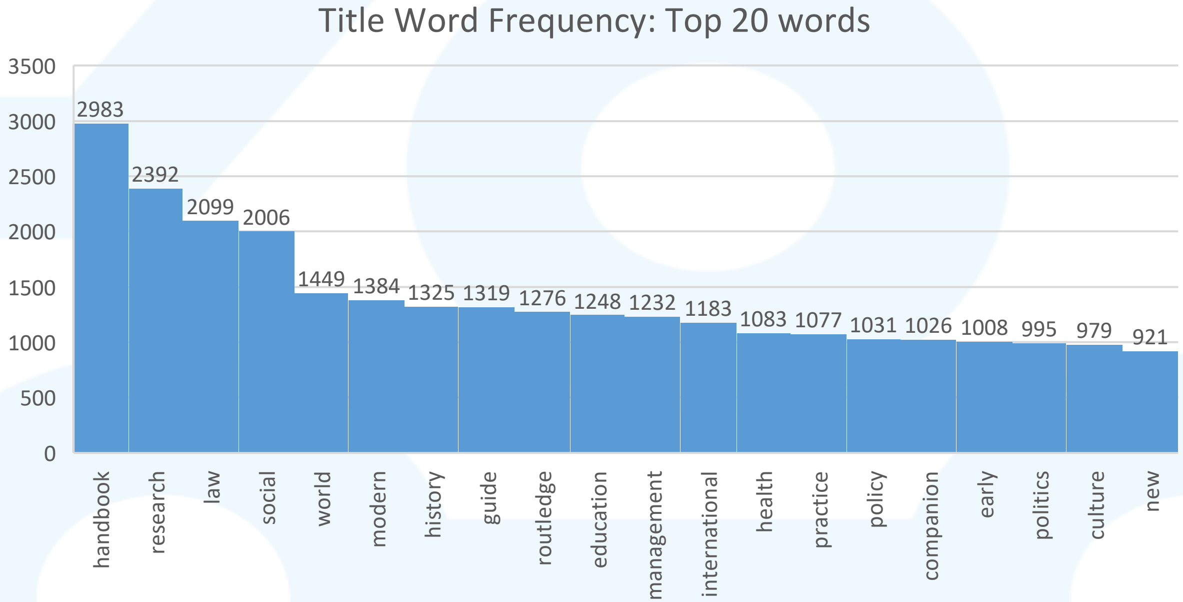


Total Usage Statistics for NPLD eJournals



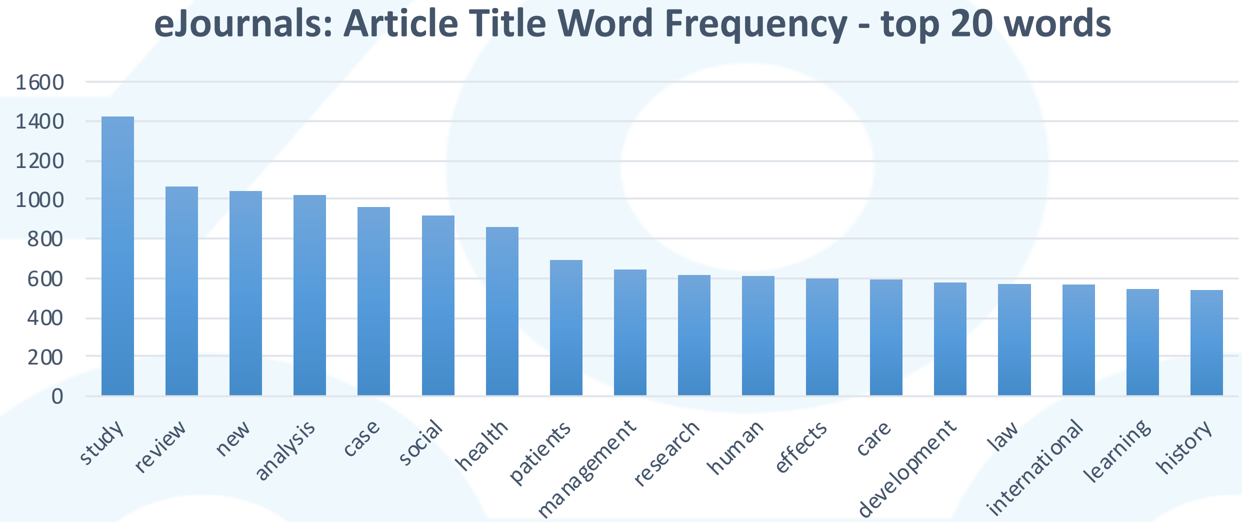
Most Frequent Words: NPLD eBooks

Classification	Words (Word Frequency)
Publisher	Routledge (1,276)
Type of book	Handbook (2,983); Research (2,392); Guide (1,319); Companion (1,026);
Subject Area	Law (2,099); Social (2,006); History (1,325); Education (1,248); Management (1,232); Health (1,083); Policy (1,031); Politics (995); Culture (979).
Scope/Focus of book	World (1,449); Modern (1,384); International (1,183); Practice (1,077); Early (1,008); New (921).

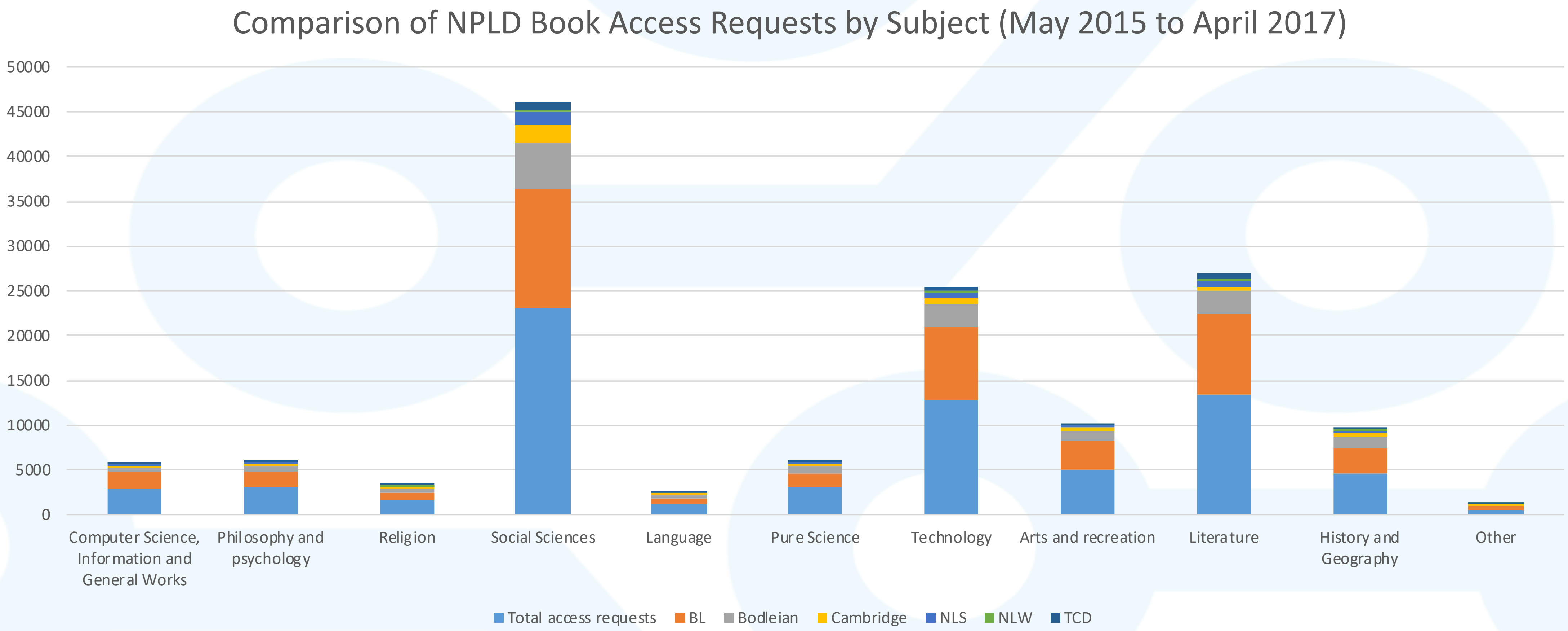


Most frequent words: NPLD eJournals

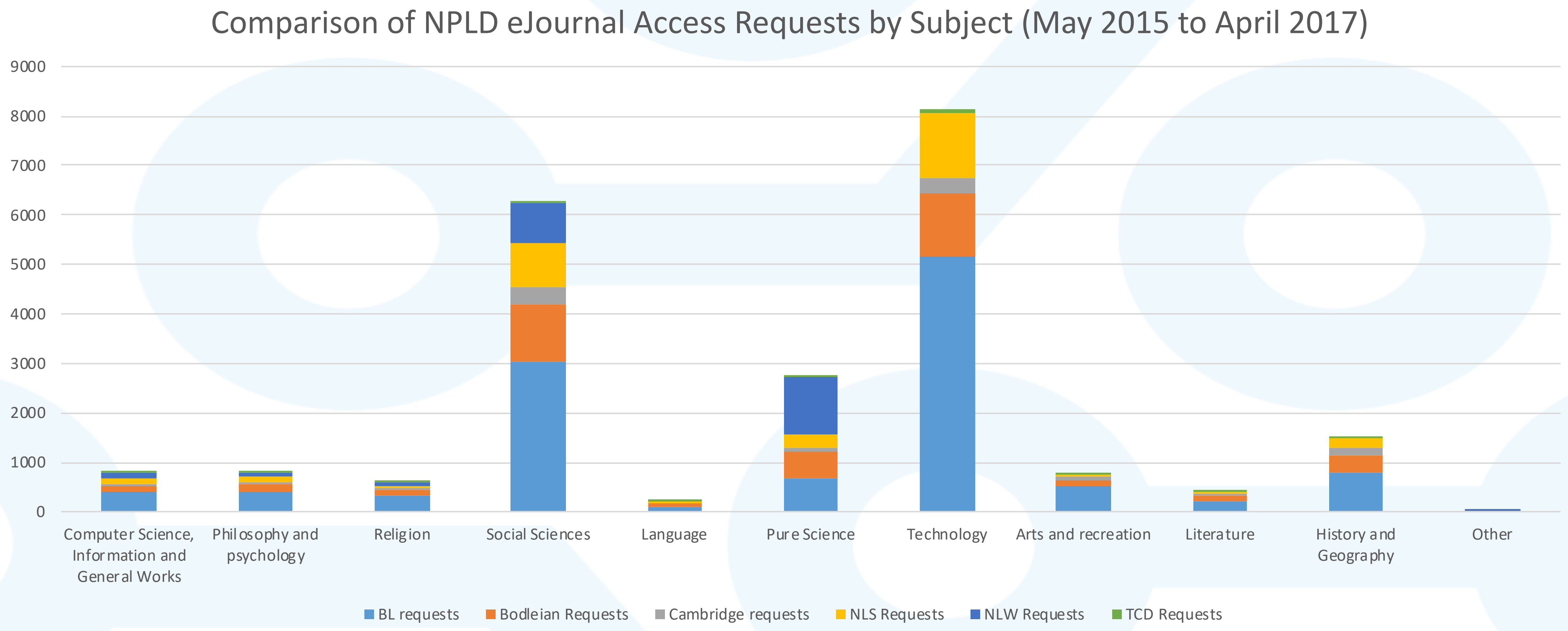
Classification	Words (Word Frequency)
Type of Article	Study (1,422); Review (1,065); Analaysis (1,022); Case (961) Research (615);
Scope/Focus of Article	New (1,042); International (567);
Subject Area	Social (919); Health (859); Management (644); Care (594); Development (578); Law (569); Learning (546); History (539); Education (520);
Community of Study	Patients (691); Human (610);
Multiple/Uncertain Meaning	Performance (516); Effects (599)



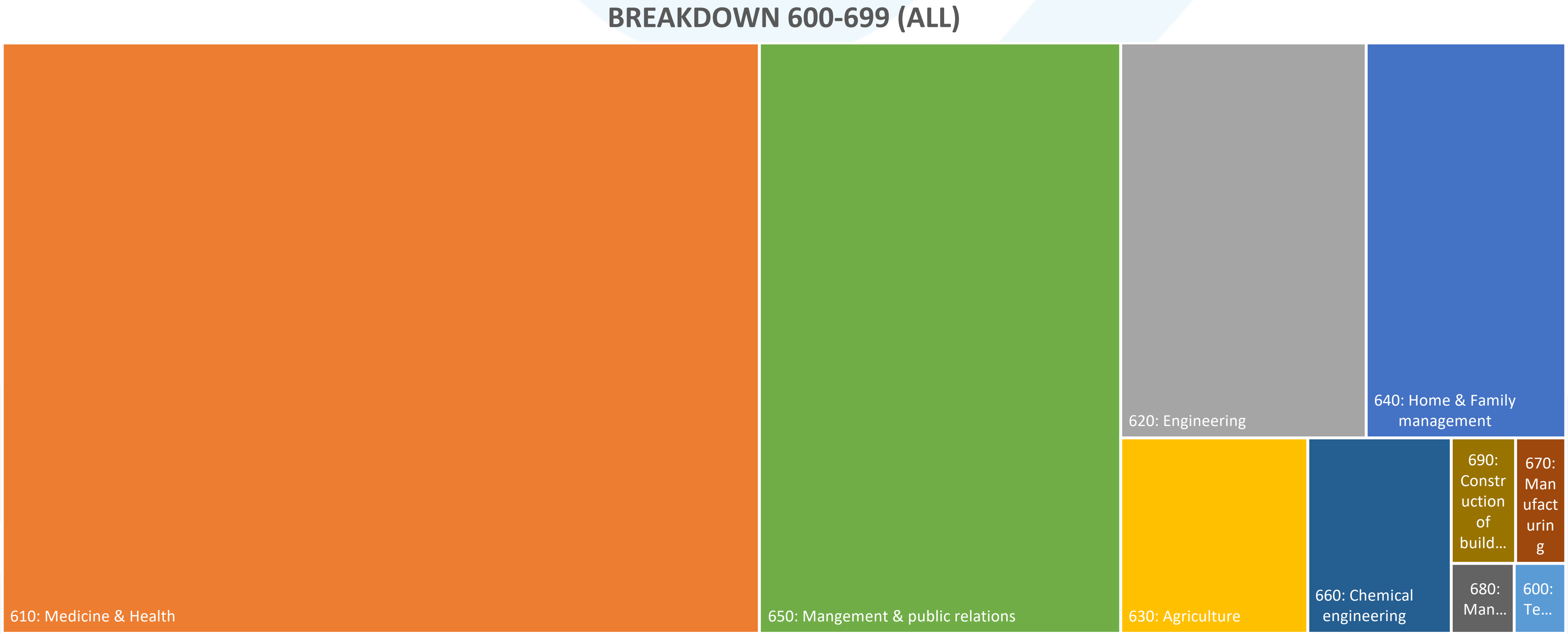
NPLD Reflects Long-Established Disciplinary Usage (eBooks)



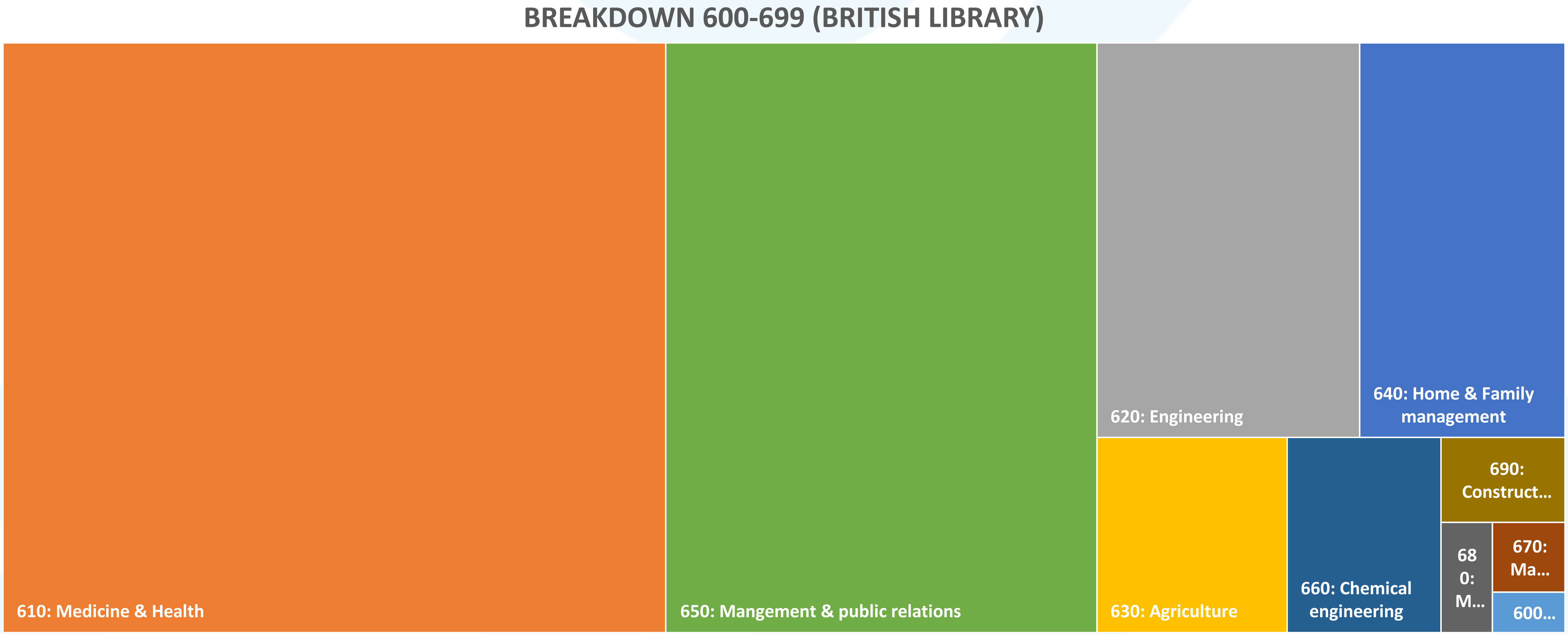
And eJournals



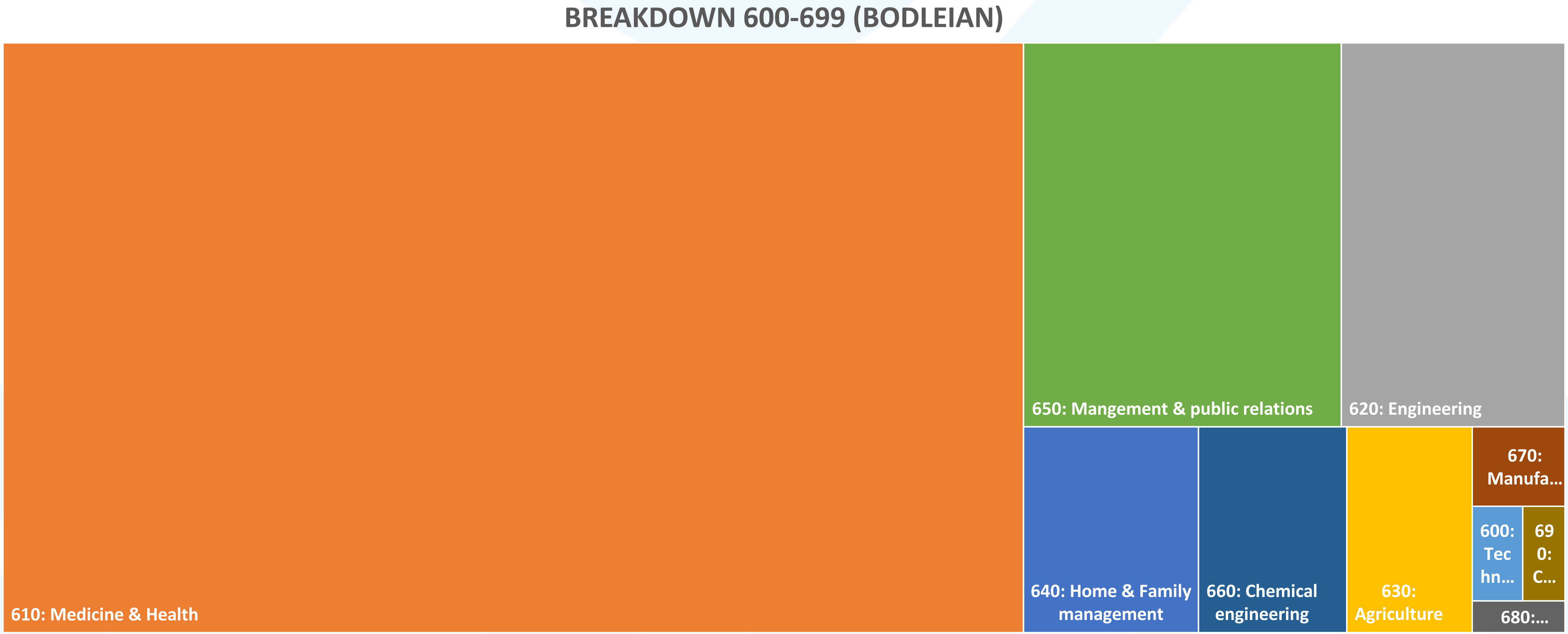
Zooming into the data: eBook usage for 600-699 in DDC (Technology/Applied Sciences)



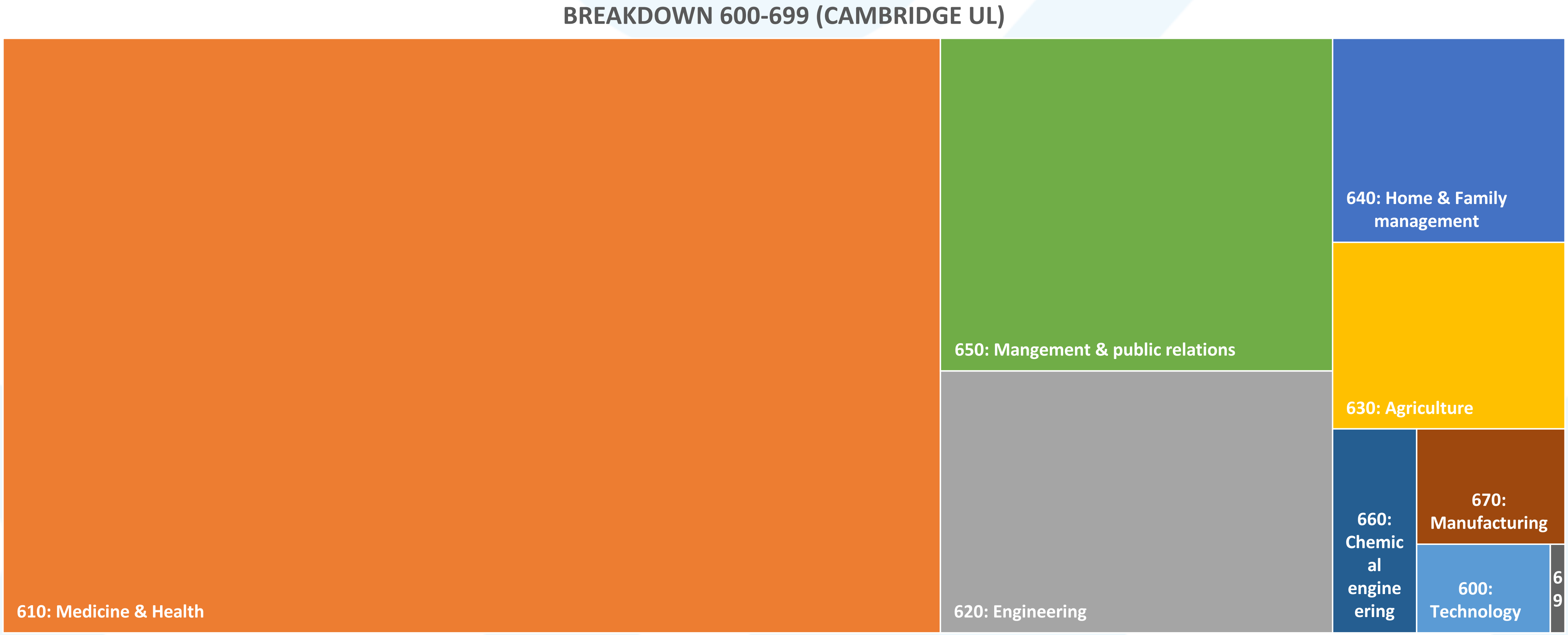
Usage of 600-699 in the British Library



Usage of 600-699 in the Bodleian Libraries



Usage of 600-699 in the Cambridge University Library



Usage of 600-699: Table for Comparison

DDC Category	Cambridge request	Cambridge (%)	Bodleian requests	Bodleian (%)	BL requests	BL (%)
600: Technology	9	1.28%	13	0.52%	26	0.32%
610: Medicine & Health	423	60.00%	1637	65.32%	3454	42.44%
620: Engineering	78	11.06%	234	9.34%	916	11.25%
630: Agriculture	33	4.68%	70	2.79%	328	4.03%
640: Home & Family management	36	5.11%	98	3.91%	717	8.81%
650: Mangement & public relations	99	14.04%	332	13.25%	2245	27.58%
660: Chemical engineering	13	1.84%	83	3.31%	265	3.26%
670: Manufacturing	13	1.84%	20	0.80%	45	0.55%
680: Manufacture for specific uses	0	0.00%	8	0.32%	50	0.61%
690: Construction of buildings	1	0.14%	11	0.44%	93	1.14%
	705		2506		8139	

What limitations are created through the use of existing classification schemes, and how might DH/LIS scholars collaborate to further develop ethical analysis of large-scale library datasets?

The big problem with Dewey Decimal Classification

- Library classification is a subjective process undertaken by humans that reflects existing biases (Mai, 2010).
- DDC provides distinct categories for English, American, and classical European schools of literature, while lumping the rest of the world under "other literatures":
  - Bias emerges from the 19th Century North American perspective of DDC (Kua, 2008).
- Automatic matching of this kind embeds existing bias into our data, and problematic perspectives:
  - This bias works well for UK-centric library collections (NPLD is a record of UK publications);
  - But what about the wider applicability of this method? How do we become "ethical stewards" (Weingart, 2014) of library usage data?

Possible next steps?

- Compare these findings to subject usage of non-NPLD materials.
- Investigate other ways to derive subject data from metadata records.
- How might a fruitful conversation between DH and Information Science develop more nuanced approaches to “representing” (Unsworth, 2000) library data?

Project Partners and Funding

- This research was supported by a grant from the Arts and Humanities Research Council under grant number AH/POO5845 between June 2017 and May 2019.
- The project white paper is available to download here: http://elegaldeposit.org/dlf-white-paper

DIGITAL LIBRARY FUTURES

University of Glasgow

UCL

Arts & Humanities Research Council

THE UNIVERSITY of EDINBURGH

UEA University of East Anglia

DIGITAL
LIBRARY
FUTURES

elegaldeposit.org