Chapter 4: Towards Critically Addressable Data for Digital Library User Studies

Paul Gooding, University of Glasgow

Abstract

 This chapter addresses two key questions: what can the concept of the black box add to our understanding of library catalogues as data in digital library user studies? And how might data-driven approaches help us to increase the transparency of these black boxes and render them critically addressable? Libraries are complex systems comprising a complex interrelationship of staff, space, users and technical infrastructure. However, digital library user studies have not applied the same attention to the creation of large-scale datasets as they have to the ethical and methodological implications of reporting on them. This chapter positions the study of library catalogue data in relation to black box theory and the collections as data imperative. It argues that collaborations between data science, the critical digital humanities, and library and information science can help us to be more transparent in how we reuse catalogue data, and to redefine how this data is created, processed, and documented in the first place.

Introduction

Libraries are information organisations consisting of a complex interrelationship of collections, staff, spaces, users, and technical infrastructures. They have long been conceived of as 'systems', a useful metaphor when viewed in light of Donella H. Meadow's definition: "a system isn't just any old collection of things. A **system** [author's emphasis] is an interconnected set of elements that is coherently organized in a way that achieves something."[1] Huge amounts of administrative data are produced across this system to describe resources and record actions, including: bibliographic metadata about information resources; (meta)data describing actions performed upon those information resources; user data including search queries, borrower records, access requests, and information seeking behaviour; and much more. These datasets have the potential to illuminate our understanding

---

1   Donella H. Meadows, *Thinking in Systems: A Primer*, ed. by Diana Wright London 2008, see 11.

of user behaviour online, and so can be of great value to researchers undertaking digital library user studies.

In general, user studies of this kind do not address the library as a complex system, and instead address a specific resource, user group, or content type. Many such studies utilize catalogue and user data to investigate information-seeking behaviour within a specific digital resource, using aggregated library patron data relating to user interactions with the online interface alongside administrative and bibliographic metadata relating to accessed information resources. Adam Chandler and Melissa Wallace draw attention to several projects that utilize Google Analytics for this purpose,[2] while other studies draw upon a wider range of methodological tools including the analysis of server weblogs, interviews, and surveys.[3] Such studies share common methodological approaches that derive from the multi- and inter-disciplinary nature of library and information studies (LIS), drawn from computational science, the social sciences, and the arts and humanities. In this chapter, I intend to focus on the latter, and specifically on how the epistemologies of the digital humanities (DH) can help us to interrogate the library catalogue as data. In this account, the library catalogue acts as the central point of interaction between digital library users and the resources, standards, technologies, categories and phenomenology that Bowker and Star argue converge in large-scale information systems.[4] To date, there remains a gap in our understanding of the link between library systems, library resources, and user behaviour, and particularly little attention to how the catalogue data that provides these insights is created, processed and received by researchers. In short, the data we rely upon to understand digital library usage too often represents a black box.

The name of this chapter takes inspiration from Bruno Latour's 1987 study of knowledge creation in the physical sciences, where he describes black boxes as devices that take some form of input, and provide some form of output, but which require no knowledge of their internal workings. I will establish the opening of these black boxes in digital library user studies as a humanistic problem, and ar-

---

2   Adam Chandler/Melissa Wallace, Using Piwik Instead of Google Analytics at the Cornell University Library, in: *The Serials Librarian* 71 (3-4/2016), 173-179, see 173-174, doi:10.1080/0361526 X.2016.1245645.

3   For examples, Eric T. Meyer/Kathryn Eccles, *The Impacts of Digital Collections: Early English Books Online & House of Commons Parliamentary Papers*, London 2016, URL: https://papers.ssrn.com/s ol3/papers.cfm?abstract_id=2740299 [last accessed: April 2, 2021]; Claire Warwick et al., If You Build It Will They Come? The LAIRAH Study: Quantifying the Use of Online Resources in the Arts and Humanities, in: *Literary and Linguistic Computing*, 23 (1/2008), 85-102, doi:10.1093/llc /fqm045.

4   Geoffrey C. Bowker/Susan Leigh Star, *Sorting Things Out: Classification and Its Consequences*, Boston, MA, 1999.

gue that the "collections as data" framework[5] provides a rich basis upon which to transform our understanding of library catalogue data as a humanities dataset. I will address two key questions: what can the concept of the black box add to our understanding of library catalogues as data in digital library user studies? And how might data-driven approaches help us to increase the transparency of these black boxes and render them critically addressable? My response is aimed at two audiences: those who wish to be more transparent in their own adoption of existing library catalogue data, and those who might seek to define how this data is created, gathered and processed in the first place. I will therefore finish by suggesting a set of guidelines by which each group might embed a more transparent approach to the adoption of library catalogue data in digital library user studies. My hope is that by making explicit the link between collections as data and catalogues as data, that the black boxes of library catalogue data may be more effectively refined and critiqued in future.

Black Boxes all the Way Down: The Research Context

The observations that inform this chapter were inspired by my work on *Digital Library Futures* (DLF), funded by the Arts and Humanities Research Council in the UK, which ran between 2017 and 2019. DLF set out to investigate the impact of Non-Print Legal Deposit (NPLD) upon UK academic deposit libraries and their users. Legal deposit ensures the systematic preservation of a nation's published output. It has existed in English law since 1662, and in British law since 1710.[6] Until recently the legal deposit regulations gave the libraries the right only to receive print materials including books, periodical, music and maps, but in 2013 the UK extended this right to incorporate digital materials.[7] My interest in this topic arose because the regulations emphasize the bequest value of legal deposit over contemporary usage, creating a tension between posterity-driven preservation and the theoretical accessibility benefits of digital materials. These tensions are represented in access protocols that can be described as "e-reading,"[8] the reading of single items on a screen, and within the reading rooms of the six legal deposit libraries. These

---

5   Thomas Padilla et al., *Santa Barbara Statement on Collections as Data*, 2017, URL: https://collectionsasdata.github.io/statement/ [last accessed: April 2, 2021].

6   Paul Gooding/Melissa Terras/Linda Berube, *Towards User-Centric Evaluation of Non-Print Legal Deposit: A Digital Library Futures White Paper*, URL: http://elegaldeposit.org/resources [last accessed: April 2, 2021].

7   *The Legal Deposit Libraries (Non-Print Works) Regulations 2013*, URL: http://www.legislation.gov.uk/uksi/2013/777/contents/made [last accessed: April 2, 2021].

8   Georgi Alexandrov, Transformation of Digital Legal Deposit in Terms of Legislation and Public Access, in: *Knygotyra* 70 (2018), 136-153, doi:10.15388/Knygotyra.2018.70.11812.

protocols contrast with libraries' efforts to widen online participation, and with shifting perspectives on the material forms of textual publications, exemplified by N. Katherine Hayles' argument that "the advent of electronic textuality presents us with an unparalleled opportunity to reformulate fundamental ideas about texts."[9] We were motivated to investigate these frictions, and set out to investigate how the new collections were being used by researchers within the UK academic deposit libraries.

My work sits within an established tradition of user studies that span LIS and the digital humanities (DH), and which are united by an explicitly humanistic perspective. Both fields share common methodological characteristics in how they address usage of digital library resources. They aim to inform working practices around digital library collections, and to develop theoretical and practical models of online information behaviour. Previous studies have evaluated the impact of online resources upon users, or developed theoretical and methodological frameworks for impact evaluation.[10] More broadly, DH has much to offer user studies in libraries through a shared interest in problems that are central to both fields, such as "information organization, information behaviour, information retrieval, sociotechnical systems, human-computer interaction, computer supported co-operative work, and information systems."[11] LIS as a field of study has long incorporated multi- and inter-disciplinary perspectives upon the central topic of research into human-recorded information,[12] a tradition that has been enriched by the explicitly humanistic and data-literate epistemology of the digital humanities. In doing so, these authors place digital collections into conversation with conceptions of library usage, and move us towards a rich set of questions around how online delivery of library resources impacts upon access and usage.

To address these questions in DLF, we adopted a mixed methods case study approach, combining expert interview, surveys, and webometric approaches including web analytics and a subject-based analysis of user requests to access NPLD resources. It is these last two methods that occupy me here, as they illuminate how library usage data comes from diverse sources and can resemble black boxes which render their internal workings invisible. Yet this data is imbued with biases

---

9   N. Katherine Hayles, *My Mother Was a Computer: Digital Subjects and Literary Texts*, Chicago 2005.

10   Lorna M. Hughes, ed., *Evaluating and Measuring the Value, Use and Impact of Digital Collections*, London 2012.

11   Tanya Clement/Daniel Carter, Connecting Theory and Practice in Digital Humanities Information Work, in: *Journal of the Association for Information Science and Technology* 68 (6/2017), 1385-1396, doi:10.1002/asi.23732.

12   David Bawden, Organised Complexity, Meaning and Understanding: An Approach for a Unified View of Information for Information Science, in: *ASLIB Proceedings* 59 (4/5/2007), 307-327, doi:10.1108/00012530710817546.

and assumptions that are embedded in their creation, processing and reception by scholars.

This chapter will outline a systemic view of digital libraries that embraces both the complexity and scale of system-level user analysis. To date, studies have focused on individual resources, or particular subsets of users, but addressing libraries as complex systems will increasingly require us to harness machine learning techniques to cope with data that is typologically diverse and large in scale. As I will indicate, the challenges associated with this are shared across many fields wrestling with new forms of data, and emerging forms of computational research. For instance, the debates around the topic of Explainable Artificial Intelligence (XAI) are highly relevant to this analysis. As Matt Turek has noted, work on the concept of XAI aims to create machine learning techniques that produce more explainable models, increase the transparency of machine learning models, and enable human users to understand, trust and manage artificial intelligence operations in their own work.[13] XAI requires the production of an explainable model, and some form of explanation interface that allows a user to understand the decisions that were made. For this reason, the metaphor of the black box provides a common ground with other disciplines, including that on XAI. Adadi and Berrada, for instance, explicitly refer to the black box in their survey of writing on XAI, and note it as a common term within the literature on AI and machine learning.[14]

As Computer Science has developed a theoretical basis for XAI, theorists within DH have developed varied critical theoretical lenses by which to address the study of humanities data. Both disciplines offer models for ethical, transparent utilisation of data, and of machine learning models for data analysis. A similar perspective on the question of library catalogues as data would, I will argue, greatly enrich digital library user studies. It would help to centre questions of data provenance and bias at the heart of our research. In order to establish the rationale for this more humanistic approach to catalogue data, I will explore how library user studies have engaged with the notion of libraries as complex systems before situating this history in relation to black box theory.

---

13   Matt Turek, Explainable Artificial Intelligence, URL: https://www.darpa.mil/program/explain able-artificial-intelligence [last accessed: April 2, 2021].

14   Amina Adabi/Mohammed Berrada, Peeking inside the Black-Box: A Survey on Explainable Artificial Intelligence, in: *IEEE Access*, 6 (2018), 52138-52160, see 52146, doi: 10.1109/ACCESS.2 018.2870052.

Library User Studies and the Library as System

While user studies within LIS have a long history, T. D. Wilson's seminal paper from 1981[15] has been credited as one of the first works to establish a clear conceptual and methodological framework for library user studies. Wilson described his own paper as a "way of thinking about the field" of user studies.[16] The paper sought to differentiate between information needs as a specific need perceived by the user, and the behaviours that were taken to meet that need. Wilson argues that the study of information behaviour on its own does not necessarily address the information needs of users, suggesting that information-seeking behaviour should be placed within a much broader context. The library is positioned as an intermediary between the user's "life world" on the one hand, and the "embodiments of knowledge" that they need in order to meet an information need on the other.[17] In Wilson's model, the information system mediates between the user and the embodiments of knowledge, thus hinting at its overlapping and interlinked components:

> The user will be in contact with a variety of 'information systems', only one of which is shown in the diagram, hence the indicated overlap with the user and his life-world. Within the information system two subsystems are shown: the 'mediator' (generally a living system, i.e., a human being) and the 'technology', used here in the general sense of whatever combination of techniques, tools and machines constitute the information-searching subsystem.[18]

In LIS, a large amount of research exists that focuses on components of this system, primarily in relation to information behaviour rather than the broader contexts relating to user needs. Many of these studies incorporate data-driven methods derived from webometrics and informetrics. However, a gap remains in how we approach the data that underpins these studies, which are often received by researchers as figurative black boxes. This is not to say that there has been no criticism of webometric techniques within the LIS literature. For instance, Michael Thelwall has argued that the insights derived from webometric data are limited in the absence of other contextual factors.[19] This echoes Wilson's warning that early user studies failed to identify the wider context for information-seeking behaviour.[20]

---

15   T. D. Wilson, On User Studies and Information Needs, in: *Journal of Documentation*, 37 (1/1981), 3-15, doi:10.1108/eb026702.

16   Wilson, On User Studies and Information Needs, 4.

17   Wilson, On User Studies and Information Needs, 5-6.

18   Wilson, On User Studies and Information Needs, 6.

19   Mike Thelwall, *Introduction to Webometrics: Quantitative Research for the Social Sciences*, San Rafael, CA, 2009.

20   Wilson, On User Studies and Information Needs, 5.

Researchers have engaged with several aspects of this broader context. Mixed methods are extremely common in digital resource impact evaluation, while a large body of work exists to address the ethical concerns around using library catalogue data within research. These studies have addressed the reuse of library user data, addressing difficult questions around surveillance,[21] the role of research libraries in university-wide learning analytics programs,[22] student privacy,[23] and ethical data practices for librarians.[24] However, such debates focus largely upon the end use of this data, leaving a gap in our understanding of the contexts around the creation, processing, and management of library catalogue and user data. This chapter aims to address that context, the information systems that sit at the centre of Wilson's information-seeking model. In doing so, I hope to support critique of the black boxes that sit within these systems, and which constitute major data sources for LIS scholars. I will draw on existing concepts of the library as a system, in order to understand how the problem of black boxes arises in digital library user studies.

The Digital Library as a Complex System

In the introduction, I proposed that we might understand the library as a coherently organized system of interconnected elements. The idea that libraries can be understood as interoperable systems is by no means new, and indeed significantly predates Wilson's model. Indeed, Merrill M. Flood argued in 1964 that the services and collections offered by a single library are dependent upon a complex infrastructure of linked collections. These collections can be hosted locally, or be physically distributed across a geographic area; Flood, for instance, refers to the University of California, whose collections are dispersed across an entire state.[25] While Flood considers the collections as the key component of the library system, he does note that collections alone are not likely to be considered a library and that the system

---

21   Martin Patrick, Patron Data and the Fear of Surveillance: Some Thoughts, in: *Medium* 28.03.2016, URL: https://medium.com/@martinpatrick/patron-privacy-and-freedom-b6ebc62 5021a [last accessed: April 2, 2021].

22   Kyle M. L. Jones/Dorothea Salo, Learning Analytics and the Academic Library: Professional Ethics Commitments at a Crossroads, in: *College & Research Libraries* 79 (3/2018), 304-323, doi:1 0.5860/crl.79.3.304.

23   Alan Rubel/Kyle M. L. Jones, Student Privacy in Learning Analytics: An Information Ethics Perspective, in: *The Information Society*, 32 (2/2016), 143-159, doi:10.1080/01972243.2016.11305 02.

24   Andrew Asher, Risk, Benefits, and User Privacy: Evaluating the Ethics of Library Data, in: Bobbi Newman/Bonnie Tijerina (eds.), *Protecting Patron Privacy: A LITA Guide*, Lanham, MD, 2017, 43-56.

25   Merrill M. Flood, The Systems Approach to Library Planning, in: *The Library Quarterly: Information, Community, Policy* 34 (4/1964), 326-338, doi:10.1086/619267.

also needs to incorporate discovery mechanisms and collections management processes: "in a sentence, a library is a collection of records together with a retrieval procedure. The total library system includes the libraries and the persons responsible for them, along with the concepts and procedures that guide the planning and operation of the entire system."[26]

The scope of this library system must inevitably be expanded somewhat to include the technical and physical infrastructures of digital libraries. The digital shift has seen the library system atomizes even further than in Flood's account; libraries now have physical and digital collections, with information resources and related data spread across multiple platforms. A large research library is likely to hold hundreds or thousands of subscriptions to online resources, spread across several proprietary data platforms provided by multiple vendors, each with their own data structures, usage metrics, and licensing conditions. A brief glance at OCLC, an American non-profit organisation that develops services to support access to information resources, shows the level of dispersion in digital library systems. OCLC works with thousands of partners including software vendors, library service providers, publishers, and major consumer services such as Google Books and Google Scholar. These partnerships allow them to provide access to millions of pages of digitized content, to hold descriptive metadata and cover art for over 1.2 billion resources, and make it accessible via a simple library discovery platform.[27] OCLC's services depend upon a myriad of dependencies, and yet as with all library service providers still do not encompass anywhere near the full range of resources available to libraries. The proliferation of platforms, stakeholders, and data creators means that many unseen assumptions are embedded into both library catalogues and the administrative data that structures and records user interactions.

The study of user data is also the study of interaction between users, information resources, and the library systems that act as intermediaries. Bowker and Star explore the role of classifications, in the broadest sense, in shaping our lives. They propose a helpful definition of infrastructure as an embedded structure "sunk into, inside of, other structures, social arrangements and technologies"[28] that has a temporal and physical reach. They identify as a key characteristic the way that infrastructure acts as an embodiment of standard that is plugged into other infrastructures and tools in a standardized way. This helps us to understand library systems as infrastructural representations of the standards, classifications and working practices that shape knowledge within that system. As Wilson argued, in order to fully understand usage of these systems we must also engage with the contexts,

---

26   Flood, The Systems Approach to Library Planning, 327.

27   OCLC, Examples of OCLC Partnerships, 2021, URL: https://www.oclc.org/en/partners-for-libra
      ries/partnerships_examples.html [last accessed: April 2, 2021].

28   Bowker and Star, *Sorting Things Out: Classification and Its Consequences*, 35.

practices and assumptions that inform them. The process of opening these contexts up to scrutiny, of making visible the hidden, is therefore an important tool in the study of the interactions between complex systems and their users:

> Standards and classifications, however imbricated in our lives, are ordinarily invisible. The formal, bureaucratic ones trail behind them the entourage of permits, forms, numerals, and the sometimes-visible work of people who adjust them to make organizations run smoothly. In that sense, they become more visible, especially when they break down or become objects of contention.[29]

As this broad overview of library systems and user studies demonstrates, the context in which information seeking occurs in libraries is dependent upon an interwoven system of linked resources, datasets, standards, classifications and working practices. However, we often receive these datasets as if they were unmediated, raw data. Rosenberg, Jackson and Gitelman argue that this is rarely the case:

> At first glance, data are apparently before the fact: they are the starting point for what we know, who we are, and how we communicate. This shared sense of starting with data often led to an unnoticed assumption that data are transparent, that information is self-evident, the fundamental stuff of truth itself.[30]

If we treat library catalogues as data without understanding the context of how that data was actually created, then we risk misunderstanding the biases, assumptions and practices that inform its creation. As a result, user studies are at risk of treating the datasets that form the basis of quantitative research into library users as "black boxes" rather than embodiments of intricate human and technological infrastructures.

The Black Box in Science

The black box, as Philip von Hilgers and William Rauscher observe in their history of the concept, is both "word and thing."[31] The black box as a physical object became an idea of great interest to the scientific community in the 1950s and 1960s, leading to theorisation of its relevance in scientific work.[32] In the cybernetics community,

---

29 Bowker and Star, *Sorting Things Out: Classification and Its Consequences*, 2-3.

30 Dan Rosenberg/Virginia Jackson/Lisa Gitelman, Introduction, in: Lisa Gitelman (ed.), "*Raw Data" Is an Oxymoron*, Cambridge, MA, 2013.

31 Phillip Von Hilgers/William Rauscher, The History of the Black Box: The Clash of a Thing and Its Concept, in: *Cultural Politics*, 7 (1/2011), 41-58, see 45, URL: https://muse.jhu.edu/article/58 4290 [last accessed: April 2, 2021].

32 Von Hilgers/Rauscher, The History of the Black Box, 46-51.

researchers developed Black Box Theory as a means for understanding and manipulating complex systems that were too large to understand in other ways. In her thorough account of the history of Black Box Theory, Elizabeth Petrick argues that cyberneticians modelled the concept of black boxes to address two related desires: first, they wanted to be able to model complex systems using other systems, such as modeling the human brain using an electronic computer; and second, they wanted a way to understand inputs and outputs for modeling systems that were otherwise closed from inquiry.[33] To this end, the black box was theorized as a tool or system where only the inputs or outputs are known, and the inner workings obscured.

Since this foundational work, black box theory has been adopted broadly across the computer sciences, social sciences, and humanities disciplines such as philosophy. The characteristics of the black box remain relatively stable across this disciplinary spectrum: it continues to represent an object, method, or tool for which only the inputs and outputs are known. The questions posed by William Ashby in his *Introduction to Cybernetics* therefore continue to hold great resonance for digital library user studies:

> How should an experimenter proceed when faced with a Black Box?
> What properties of the Box's contents are discoverable, and what are fundamentally not discoverable?
> What methods should be used if the Box is to be investigated efficiently?[34]

Ashby introduces the possibility that some elements of the black box are fundamentally undiscoverable, a largely conceptual point that is nevertheless worth considering in a more literal sense in relation to questions of research integrity, data privacy, and commercially sensitive data sources.

Bruno Latour's investigation into scientific knowledge construction establishes black boxes as a metaphor for the progress of scientific research. He refers to black boxes as a way of compartmentalising knowledge that is no longer open to questioning, that is made more solid by using it without further questioning. In Latour's account, black boxes serve a particular purpose, acting as a form of shorthand that allows researchers to modularize particular tools, or knowledge, so that new ideas can be more efficiently developed.[35] Latour views the black box as a sign of success, an indication that a particular module of scientific knowledge is working ef-

---

33   Elizabeth R. Petrick, Building the Black Box: Cyberneticians and Complex Systems, in: *Science, Technology, & Human Value*, 45 (4/2019), 575-595, doi:10.1177%2F0162243919881212 [last accessed: April 2, 2021].

34   William R. Ashby, *An Introduction to Cybernetics*, London 1956, see 87.

35   Victoria Stodden, The Scientific Method in Practice: Reproducibility in the Computational Sciences, in: *MIT Sloan Research Papers*, 477 (3-10/2010), doi:10.2139/ssrn.1550193.

ficiently to the point that we only need to focus upon its inputs and outputs.[36] The construction of a black box in a certain community becomes a collective process, and as it becomes more integrated in the practices of that community it becomes increasingly difficult to revise and discourages questioning of results.[37] The black box therefore serves a particular purpose, allowing new ideas to be more efficiently built upon old, and saving the need to rehearse existing arguments around it. As a result, the black box becomes a shorthand for a complex set of "commands, machinery, or a methodology underlying a result."[38]

However, as I have argued above, the library catalogue as data is often received by researchers without this process of conceptual stabilisation within the community, and thereby we require additional labour to make the data transparent and open to critique. Critical addressability refers to the notion that one should be able to evaluate the technical and social forces that shape data, through data documentation and transparent workflows. Thomas Padilla describes the minimum requirements for data to be critically addressable:

> A researcher should be able to understand why certain data were included and excluded, why certain transformations were made, who made those transformations, and at the same time a researcher should have access to the code and tools that were used to effect those transformations.[39]

In the following section, I will address the concept of "collections as data",[40] which has emerged in response to a particular moment when cultural heritage organisations are wrestling with how to operationalize their collections for digital scholarship in the humanities. I will explore the humanistic roots of the collections as data imperative, before turning to consider how collaborations between LIS and DH researchers and data scientist might enhance the field of digital library user studies.

The Collections as Data Imperative

The Santa Barbara Statement on Collections as Data sets out ten principles for thinking of cultural heritage collections as data, imagining it as an ongoing process of making collections more accessible, transparent, interoperable and, ultimately,

---

36   Bruno Latour, *Science in Action: How to Follow Scientists and Engineers through Society*, Cambridge, MA, 1987, see 304.

37   Stodden, The Scientific Method in Practice, 5.

38   Stodden, The Scientific Method in Practice, 5.

39   Thomas Padilla, Humanities Data in the Library: Integrity, Form, Access, in: *D-Lib Magazine* 22 (3-4/2016), doi:10.1045/march2016-padilla.

40   Padilla et al., *Santa Barbara Statement on Collections as Data*.

readable as humanities datasets.[41] It takes as a starting point the idea that the digital shift requires us to reframe all digital objects as data, and to thereby address the shift in collections management workflows that must occur in response. The principles have helped to develop a clear community of practice spanning LIS and DH, with a common goal of increasing the usability of library collections within digital scholarship. The digital humanities contribute a strong critique of the notion of so-called "raw data"[42] through critical theoretical lenses from fields as diverse as feminist studies, critical race studies, sexuality studies, queer theory, and class studies. These scholars share a desire to make legible the assumptions that underpin the creation, manipulation and analysis of humanities data. Roopika Risam applies an intersectional frame, addressing the ways in which race, class, gender and other aspects of identify overlap with each other, to the practices of producing digital humanities data. Risam notes that:

> Existing digital humanities projects provide examples of how, in small and large ways, theory and method can be combined to address recurring questions of the role of race, class, gender, ability, sexuality, nationality, and other categories of difference within the field. These phenomena subtend the development and production of digital humanities projects but they may not be evident. Therefore, it is incumbent on us, as digital humanities practitioners, to make them legible, to move them beyond the margins.[43]

Risam's work, along with that of Miriam Posner[44] and Victoria Stodden,[45] has inspired Padilla to propose a simple rubric for evaluating the readiness of humanities collections for digital forms of scholarship:

- **Posner:** to what extent is information about Humanities data collection provenance, processing, and method of presentation available to the user?
- **Stodden:** to what extent are data and the code that generates data available to the user?
- **Risam:** to what extent are the motivations driving all of the above available to the user?[46]

---

41   Padilla et al., *Santa Barbara Statement on Collections as Data*.

42   Rosenberg, Jackson, and Gitelman, Introduction.

43   Roopika Risam, Beyond the Margins: Intersectionality and the Digital Humanities, in: *Digital Humanities Quarterly*, 9 (2/2015), URL: http://www.digitalhumanities.org/dhq/vol/9/2/000208/000208.html [last accessed: April 2, 2021].

44   Miriam Posner, *How Did They Make That?*, 29.08.2013, URL: http://miriamposner.com/blog/how-did-they-make-that/ [last accessed: April 2, 2021].

45   Stodden, The Scientific Method in Practice.

46   Padilla, Humanities Data in the Library: Integrity, Form, Access.

In this respect, collections as data link to wider imperatives for research data to be open, accessible and interoperable. For instance, the FAIR Data Principles provide a general set of guidance for making scientific data "Findable, Accessible, Interoperable, and Reusable."[47] The FAIR Data Principles have been influential in the library sector, and encompass specific principles aimed at ensuring richly described, transparent, and interoperable data that is suited to both human-driven and machine-driven activities. For humans, the principles focus upon the semantics, or contexts, of data and digital objects, while for machines the focus is upon developing "steps along a path" towards data that is more easily machine-actionable.[48] Collections as data, with its foundations in the critical theory of the digital humanities, develops principles that are more explicitly humanistic in nature.

Padilla notes that the whole library sector is already rising to the challenge of reframing humanities information resources as data, and what I propose here is closely aligned but subtly different. Whereas collections as data reimagines existing humanities information resources as data, emphasising considerations of form, integrity, and access,[49] here I propose that the first step in imagining library catalogues as data is to address them as a humanities information resource. Johanna Drucker usefully distinguishes between data as a "given" that is able to be recorded and observed, and "capta" that is actively taken.[50] By viewing data as something that is created, and constituted, she identifies several humanistic principles for creating and analysing data:

> Humanistic Inquiry acknowledges the situated, partial and constitutive character of knowledge production, the recognition that knowledge is constructed, *taken*, not simply given as a natural representation of pre-existing fact.[51]

Based on these foundations, I will argue that it is precisely this conceptual shift that is required in relation to library catalogues as data. While humanistic perspectives embed new forms of interpretation, it is their combination with data science approaches that offers us the opportunity to fully address the mode of capture and analysis of catalogue data.

---

47   Mark D. Wilkinson et al., The FAIR Guiding Principles for Scientific Data Management and Stewardship, in: *Scientific Data*, 3 (1/2016), doi:10.1038/sdata.2016.18.

48   Wilkinson et al., The FAIR Guiding Principles, 3.

49   Padilla, Humanities Data in the Library: Integrity, Form, Access.

50   Johanna Drucker, Humanities Approaches to Graphical Display, in: *Digital Humanities Quarterly* (2011) URL: http://www.digitalhumanities.org/dhq/vol/5/1/000091/000091.html [last accessed: April 2, 2021].

51   Drucker, Humanities Approaches to Graphical Display.

Interdisciplinary Collaborations: Towards Humanistic Catalogue Datasets

The previous section has dealt with the humanistic foundations of the collections as data imperative, and how they overlap with the objectives of the FAIR Data Principles to emphasize the core values that would help us to break out of catalogue datasets as black boxes. The overall aim is for research into library catalogues as data that aspire to greater transparency, and that account for the way that broader professional, technical, social and user contexts embed particular forms of meaning within administrative and bibliographic datasets relating to information resources and their usage. It is certainly the case that data within the library catalogue system represents a black box, in many cases one that has never truly been opened up to scrutiny. Proprietary data analytics platforms present one particular example: web analytics have become an almost ubiquitous data source in digital library user studies. The field of web analytics aims to collect, analyse and report data relating to web traffic, with a particular focus on improving website effectiveness. However, these platforms are often proprietary in nature and do not provide researchers with unprocessed data. In many cases, the researcher will also be unable to access the tools and code that were used to effect transformations and present results to users. Google Analytics (GA), due to its extremely high profile, has been the brunt of just such criticism; it is very obviously a black box, with its raw data inaccessible and hidden from users for reasons including data privacy. Adam Chandler and Melissa Wallace provide a useful introduction to the adoption of GA in library studies, addressing its limitations in relation to user privacy. They criticize the inability to undertake local data collection, and the willingness of libraries to accept a loss of control of user privacy by allowing patron data to be owned by Google and stored on its US-based servers.[52] This is highly problematic for researchers in light of the UK and EU General Data Protection Regulations, which remain largely aligned in 2021 and mandate that data should not be transferred to legislatures without similar data protection requirements. Google has taken several actions to comply with GDPR, but the situation around data residency is evolving and fluid, and largely outside the control of libraries. While organisations can take actions to ensure their implementation of GA is GDPR compliant, then, users face the problem that data processing is offsite and opaque. The inaccessibility of the underlying data renders GA a black box, with only the inputs and outputs known.

Platforms such as GA fall well short of the FAIR Data Principles, as well as the test of their critical addressability, even if their speed and convenience makes them an invaluable tool for user studies. However, other less obvious forms of black box

---

52   Chandler and Wallace, Using Piwik Instead of Google Analytics, 173-179.

also exist. This was a problem we faced in the DLF project, where we chose to utilize a subject-based approach to the analysis of users of NPLD. We were inspired by Marcia Bates' observation that scholarly communication practices function differently across domains, and set out to see the extent to which subject-based practices were visible in access requests for NPLD materials. We therefore developed a methodology that applied Dewey Decimal Classifications (DDC) to access logs which recorded all requests for NPLD eBooks and eJournals in the legal deposit library reading rooms. This made it possible to infer information about users at an aggregate level, and to explore the frequency with which information resources from each discipline were accessed. This method gave us clear and valuable insights to compare usage of NPLD resources against established information-seeking behaviour in various academic disciplines, and further details were reported in our white paper.[53] Here, though, I will focus briefly upon the problems of adopting a particular classification scheme for this purpose. While the insights that we gained were highly relevant to NPLD in the United Kingdom, the wider contexts informing the DDC classification scheme suggest limits to this method's broader applicability.

The data we used, of course, was not neutral. This is a general feature of library classifications that attempt to provide a universal view of the world's knowledge, which is a subjective process that is undertaken by humans and reflects existing biases.[54] The Dewey Decimal Classification which we used was conceived by Melvil Dewey in 1873 and published in 1876 as a "general organizational tool that is continuously revised to keep pace with knowledge."[55] It provides notations in Arabic Numerals from 000 to 999, and utilizes a hierarchy based upon ten top level disciplinary classes. According to OCLC, who have administrative responsibility for publishing and maintaining DDC, it is the world's most widely used library classification, and is used by libraries in at least 138 countries.[56] Despite its widespread adoption, several accounts exist that address the biased perspective that arises from its historic origins. For instance, Kua has criticized its poor representation of non-Western languages and literatures. It provides categories for "literature, rhetoric and criticism" that use national boundaries informed by a nineteenth century North American perspective of the most important schools of literature.[57] As a result, the rest of the world is lumped under a single class of "other literatures,"

---

53   Gooding/Terras/Berube, *Towards User-Centric Evaluation of Non-Print Legal Deposit*.

54   Jens-Erik Mai, Classification in a Social World: Bias and Trust, in: *Journal of Documentation*, 66 (5/2010), 627–642, doi:10.1108/00220411011066763.

55   OCLC, Introduction to the Dewey Decimal Classification, 2019, URL: https://www.oclc.org/content/dam/oclc/dewey/versions/print/intro.pdf [last accessed: April 2, 2021].

56   OCLC, Introduction to the Dewey Decimal Classification, 2.

57   Eunice Kua, Non-Western Languages and Literatures in the Dewey Decimal Classification Scheme, in: *Libri* 54 (4/2008), 256–265, doi:10.1515/LIBR.2004.256.

occluding much of the granularity that is afforded classical and canonical literary traditions. Further criticisms have included its marginalization of certain sexual orientations and sexual preferences as forms of perversion,[58] and its lack of flexibility that amplifies existing inaccuracies and limitations.

These examples both support Latour's assertion that it is far simpler to understand how a black box works if we are able to witness, or to be involved in, its creation: "instead of black boxing the technical aspects of science and *then* looking for social influences and biases, we realised... how much simpler it was to be there *before* the box closes."[59] How, then, is library classification a black box? It is much less opaque than something like Google Analytics, after all. Yet it still shares features that indicate a lack of transparency. We used an automated tool to apply classmarks to the dataset, due to its size. As such, it shares several features that indicate its status as a black box: it still takes an input, in the form of an information resource, and provides an output, in the form of a library classmark. The process is only transparent insofar as we are open to engaging with the biases of the classification, and automated tools, used in our research. Indeed, Mai usefully distinguishes between "administrative authority" – those individuals trusted to design and edit classifications as a technical task – and "cognitive authority" – those who can be trusted to make ontological statements about the relationships between entities:

> Cognitive authorities are those people we turn to for knowledge, insight, and advice on particular matters. These people have gained their authority not by being chosen by some body but because we let them influence our thinking. We recognize that there are certain people whom we trust on particular matters and other people we trust on other matters.[60]

Mai argues that transparency is a key criterion for granting cognitive authority to particular classifications.[61] This does not necessarily entail agreeing with the structure of the data to find it trustworthy: rather, closed systems that rely upon their administrative authority appear less trustworthy than those that acknowledge their partiality and bias in a transparent, documentable manner.

This final point is an important one, because it hints at the role of data science in collaborations around library catalogues as data. Analytics platforms simply do not meet most of the FAIR Data Principles, or the rubric for humanities data proposed by Padilla, and I would argue that the user studies community has never

---

58   Melissa Adler, *Cruising the Library: Perversities in the Organization of Knowledge*, New York 2017.

59   Latour, *Science in Action: How to Follow Scientists and Engineers through Society*, 21.

60   Mai, Classification in a Social World: Bias and Trust, 635-636.

61   Mai, Classification in a Social World: Bias and Trust, 639.

opened these black boxes up to full scrutiny. However, their ubiquity within library systems and digital library user systems suggests they fill a clear and obvious need within the research community. What I therefore propose is that we approach transparency around library catalogue data in a layered and reflective manner, with data literacy at one end and the definition and creation of new data structures at the other. As a starting point, I envisage three layers: first, datasets that are unable to be critically addressed due to the inaccessibility of unprocessed data; second, datasets that are "received" but allows a degree of critical addressability due to documentation, code, accessible data, or other indicators of transparency; and third, datasets that are "captured" by researchers to *be* critically addressable. The latter type represent an ontological argument that attempts to embed a particular critical theoretical perspective into the act of data creation, processing, and analysis, and is the category where I foresee a major role for data scientists and digital humanists. I would therefore propose differing approaches, based upon the degree of critical addressability of the available data:

1. For data that is not critically addressable (e.g., Google Analytics, where "raw" data and code are both hidden), researchers should:
   o   Be open in their methodology about what aspects of the dataset are not critically addressable.
   o   Investigate whether there are alternative data sources that would provide a greater degree of transparency.
   o   Transparently record the extent to which their dataset corresponds to the FAIR data principles.
2. For data that is received, and relies on existing classifications and infrastructures (e.g., subject-based analysis that utilizes existing library classifications), researchers should:
   o   Engage with the broader literature to identify the broader epistemological and methodological contexts that inform the creation of the data.
   o   Consider the extent to which those contexts shape their results.
   o   Transparently record both the actions that they take in their own data analysis, and the assumptions and processes behind its creation.
3. For data that is captured by the researcher, in a manner that defines how it is created, organized and processed (e.g., subject-based analysis with categories defined by the researcher), researchers should:
   o   Engage in interdisciplinary collaboration to ensure that data is critically addressable.
   o   Transparently record the actions that they take in the entire data lifecycle.

Data scientists have begun to intervene already, with proposals for bias-aware methodologies in natural language processing research,[62] and interventions that seek to apply data visualisation approaches to library catalogue data. However there remains a great opportunity for work that combines these approaches to address the library catalogue as a humanistic data source. In particular, there is a rich opportunity to investigate the application of NLP methodologies to the study of bibliographic metadata, and to contextualize it in relation to datasets representing the wider library system. Furthermore, informed by Victoria Stodden's argument of how black boxes discourage dissent, we should apply two generalized criteria: first, that the community should have collectively, and adequately, interrogated a tool before it becomes a black box; and second, that researchers should be transparent about the biases and assumptions that go into the creation and analysis of their own data, to the extent that this is practicable.

Conclusion

I have argued in this chapter that researchers engaged in digital library user studies must take further steps to ensure the critical addressability of library catalogue and user data. Library practitioners and researchers have defined good practice in the ethical reuse of catalogue and user data, but I believe we still have more work to do to understand the data itself. Many datasets are received as black boxes, with only the input and output known, and therefore fall short of the requirements for transparency and addressability of the FAIR Data Principles and the rubrics required to ensure critical addressability. The library catalogue, and the data on collections, library work, infrastructural contexts, and users, are where I believe that interdisciplinary collaboration can support us to begin to open these black boxes: both to more transparently interrogate the data that we receive, and to look towards DH and data science to actually redefine how it is created and processed. Furthermore, as the complexity of library data increases, it is likely that the field will move towards forms of analysis that incorporate Machine Learning to support the analysis of linked, typologically diverse datasets. The field therefore has much to learn from how Computer Science has theorized Explainable Artificial Intelligence, and how Data Science and the Digital Humanities have addressed the notion of critically addressable data. I propose that this work should proceed on a humanistic basis, and have laid out a broad critical context for doing so based on the complementary theoretical lenses of black boxes and collections as data which speak to this broad

---

62   Lucy Havens et al., Situated Data, Situated Systems: A Methodology to Engage with Power Relations in Natural Language Processing Research, in: *The 28th International Conference on Computational Linguistics*, URL: https://arxiv.org/abs/2011.05911 [last accessed: April 2, 2021].

interdisciplinary framework. The key intervention of this chapter is to position catalogue data within an explicitly humanistic framework, thus opening it up to the same forms of critique as collections data.

I have proposed some preliminary ideas for how we might go about critically addressing catalogues as data, but intend this chapter to be a starting point for further discussion rather than a definitive perspective on the subject. The proposals recognize the complexity of the digital library "system" that sits at the interface between users and information resources, and that this will influence the adoption of opaque datasets by individual research practitioners. In all cases, my suggestions represent an attempt not only to think of the contexts that surround information seeking behaviour, but the data that we use to study this behaviour. Black box theory helps to underline suitable approaches to the development of knowledge on catalogues as data, while the collections as data principles can aid us to establish a humanistic perspective to the resultant interrogation of data creation, management, and analysis. The challenge that remains unanswered by this chapter, and the focus of potential collaborative activity, is what might it look like in practice to investigate library catalogues as data? And just as importantly, how might data-driven approaches help us to understand, and even to address, the contexts for data capture in digital library systems?

A focus for future work should be on developing prototypical work for putting these principles into practice. These interventions must develop workflows that address large-scale, fragmented, library ecosystems. Data-driven approaches, informed by cutting edge critical humanistic theory, allow researchers to embrace pluralism in the ontological statements that become embedded in our library datasets. By adopting a position of critical addressability, we can account for the assumptions, decisions, and labour that underpin the order and usage of library collections as data. In doing so, I believe we can begin to imagine a more explicitly humanist basis by which to analyse the creation and capture of library user data that has traditionally been treated as largely administrative.

Bibliography

Adabi, Amina/Berrada, Mohammed, Peeking inside the Black-Box: A Survey on Explainable Artificial Intelligence, in: *IEEE Access*, 6 (2018), 52138-52160, doi: 10.1109/ACCESS.2018.2870052.

Adler, Melissa, *Cruising the Library: Perversities in the Organization of Knowledge*, New York 2017.

Alexandrov, Georgi, Transformation of Digital Legal Deposit in Terms of Legislation and Public Access, in: *Knygotyra* 70 (2018), 136-153, doi:10.15388/Knygotyra.2018.70.11812.

Ashby, William R., *An Introduction to Cybernetics*, London 1956.

Asher, Andrew, Risk, Benefits, and User Privacy: Evaluating the Ethics of Library Data, in: Bobbi Newman/Bonnie Tijerina (eds.), *Protecting Patron Privacy: A LITA Guide*, Lanham, MD, 2017, 43-56.

Bawden, David, Organised Complexity, Meaning and Understanding: An Approach for a Unified View of Information for Information Science, in: *ASLIB Proceedings* 59 (4/5 /2007), 307-327, doi:10.1108/00012530710817546.

Bowker, Geoffrey C./Star, Susan Leigh, *Sorting Things Out: Classification and Its Consequences*, Boston, MA, 1999.

Chandler, Adam/Wallace, Melissa, Using Piwik Instead of Google Analytics at the Cornell University Library, in: *The Serials Librarian* 71 (3-4/2016), 173-179, doi:10.1080/0361526X.2016.1245645.

Clement, Tanya/Carter, Daniel, Connecting Theory and Practice in Digital Humanities Information Work, in: *Journal of the Association for Information Science and Technology* 68 (6/2017), 1385-1396, doi:10.1002/asi.23732.

Drucker, Johanna, Humanities Approaches to Graphical Display, in: *Digital Humanities Quarterly* (2011) URL: http://www.digitalhumanities.org/dhq/vol/5/1/000091/000091.html [last accessed: April 2, 2021].

Flood, Merrill M., The Systems Approach to Library Planning, in: *The Library Quarterly: Information, Community, Policy* 34 (4/1964), 326-338, doi:10.1086/619267.

Gooding, Paul/Terras, Melissa/Berube, Linda, *Towards User-Centric Evaluation of Non-Print Legal Deposit: A Digital Library Futures White Paper*, URL: http://elegaldeposit.org/resources [last accessed: April 2, 2021].

Havens, Lucy, et al., Situated Data, Situated Systems: A Methodology to Engage with Power Relations in Natural Language Processing Research, in: *The 28th International Conference on Computational Linguistics*, URL: https://arxiv.org/abs/2011.05911 [last accessed: April 2, 2021].

Hayles, N. Katherine, *My Mother Was a Computer: Digital Subjects and Literary Texts*, Chicago 2005.

Hughes, Lorna M., ed., *Evaluating and Measuring the Value, Use and Impact of Digital Collections*, London 2012.

JONES, Kyle M. L./SALO, Dorothea, Learning Analytics and the Academic Library: Professional Ethics Commitments at a Crossroads, in: *College & Research Libraries* 79 (3/2018), 304-323, doi:10.5860/crl.79.3.304.

KUA, Eunice, Non-Western Languages and Literatures in the Dewey Decimal Classification Scheme, in: *Libri* 54 (4/2008), 256-265, doi:10.1515/LIBR.2004.256.

LATOUR, Bruno, *Science in Action: How to Follow Scientists and Engineers through Society*, Cambridge, MA, 1987.

MAI, Jens-Erik, Classification in a Social World: Bias and Trust, in: *Journal of Documentation*, 66 (5/2010), 627-642, doi:10.1108/00220411011066763.

MEADOWS, Donella H., *Thinking in Systems: A Primer*, ed. by Diana Wright, London 2008.

MEYER, Eric T./ECCLES, Kathryn, *The Impacts of Digital Collections: Early English Books Online & House of Commons Parliamentary Papers*, London 2016, URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2740299 [last accessed: April 2, 2021].

OCLC, Examples of OCLC Partnerships, 2021, URL: https://www.oclc.org/en/partners-for-libraries/partnerships_examples.html [last accessed: April 2, 2021].

OCLC, Introduction to the Dewey Decimal Classification, 2019, URL: https://www.oclc.org/content/dam/oclc/dewey/versions/print/intro.pdf [last accessed: April 2, 2021].

PADILLA, Thomas, Humanities Data in the Library: Integrity, Form, Access, in: *D-Lib Magazine* 22 (3-4/2016), doi:10.1045/march2016-padilla.

PADILLA, Thomas, et al., Santa Barbara Statement on Collections as Data, 2017, URL: https://collectionsasdata.github.io/statement/ [last accessed: April 2, 2021].

PATRICK, Martin, Patron Data and the Fear of Surveillance: Some Thoughts, in: *Medium* 28.03.2016, URL: https://medium.com/@martinpatrick/patron-privacy-and-freedom-b6ebc625021a [last accessed: April 2, 2021].

PETRICK, Elizabeth R., Building the Black Box: Cyberneticians and Complex Systems, in: *Science, Technology, & Human Value*, 45 (4/2019), 575-595, doi:10.1177%2F0162243919881212 [last accessed: April 2, 2021].

POSNER, Miriam, *How Did They Make That?*, 29.08.2013, URL: http://miriamposner.com/blog/how-did-they-make-that/ [last accessed: April 2, 2021].

RISAM, Roopika, Beyond the Margins: Intersectionality and the Digital Humanities, in: *Digital Humanities Quarterly*, 9 (2/2015), URL: http://www.digitalhumanities.org/dhq/vol/9/2/000208/000208.html [last accessed: April 2, 2021].

ROSENBERG, Dan/JACKSON, Virginia/GITELMAN, Lisa, Introduction, in: Lisa Gitelman (ed.), *"Raw Data" Is an Oxymoron*, Cambridge, MA, 2013.

RUBEL, Alan/JONES, Kyle M. L., Student Privacy in Learning Analytics: An Information Ethics Perspective, in: *The Information Society*, 32 (2/2016), 143-159, doi:10.1080/01972243.2016.1130502.

STODDEN, Victoria, The Scientific Method in Practice: Reproducibility in the Computational Sciences, in: *MIT Sloan Research Papers*, 477 (3-10/2010), doi:10.2139/ssrn.1550193.

The Legal Deposit Libraries (Non-Print Works) Regulations 2013, URL: http://www.legislation.gov.uk/uksi/2013/777/contents/made [last accessed: April 2, 2021].

THELWALL, Mike, *Introduction to Webometrics: Quantitative Research for the Social Sciences*, San Rafael, CA, 2009.

TUREK, Matt, Explainable Artificial Intelligence, URL: https://www.darpa.mil/program/explainable-artificial-intelligence [last accessed: April 2, 2021].

VON HILGERS, Phillip/RAUSCHER, William, The History of the Black Box: The Clash of a Thing and Its Concept, in: *Cultural Politics*, 7 (1/2011), 41-58, URL: https://muse.jhu.edu/article/584290 [last accessed: April 2, 2021].

WARWICK, Claire, et al., If You Build It Will They Come? The LAIRAH Study: Quantifying the Use of Online Resources in the Arts and Humanities, in: *Literary and Linguistic Computing*, 23 (1/2008), 85-102, doi:10.1093/llc/fqm045.

WILKINSON, Mark D., et al., The FAIR Guiding Principles for Scientific Data Management and Stewardship, in: *Scientific Data*, 3 (1/2016), doi:10.1038/sdata.2016.18.

WILSON, T. D., On User Studies and Information Needs, in: *Journal of Documentation*, 37 (1/1981), 3-15, doi:10.1108/eb026702.