Detecting and characterising transmission from legacy collection catalogues

James Baker, Andrew Salway, and Cynthia Roman.

Abstract

Catalogue records underpin the audit, curatorial, and public access functions of collecting institutions. And they are relied upon by many humanities researchers, and increasingly those looking to analyse collection holdings at scale. However, far from being a neutral record of collection holdings, catalogues are the products of cataloguing labour, often spanning many decades, and so are subject to various biases and inequities. Understanding how collection catalogues are shaped by their histories is then crucial for addressing many of the contemporary challenges faced by cataloguing professionals and for enhancing their use in humanities research, as well as for opening up new directions for historical research. This paper contributes a computationally-based approach for generating new and important knowledge about catalogues, in particular for investigating how a catalogue is shaped by an earlier one. We contend that understanding at scale the transmission of records and style from one catalogue to another requires the use of computational techniques to detect and analyse the various ways in which transmission manifests across a catalogue.

Our case study concerns the transmission of Mary Dorothy George's voice through time, across space, and between mediums, from the 1930s to the late-twentieth century and beyond, from the British Museum in London to the Lewis Walpole Library in Farmington, Connecticut, from printed volumes to networked digital data. It aims to show how transmission happens, how it can be found, and how it can be characterised. Understanding transmission is important because cataloguers like George are the interlocuters between us and the pasts they described, legacy voices that refuse to stay in their historical place, and whose raced, sexed, and classed influence on the future should not go unchecked.

Our contributions are relevant both for historical research into catalogues and cataloguing, knowledge organisation and infrastructure, and cultural organisations, and for cataloguing practitioners seeking to rationalise/review their catalogues to improve user experience, address systemic inequalities in object representation, and develop best practice for future work. Furthermore, in broad terms, by contributing to the generation of new knowledge about the biases/inequities of catalogues our work will enable new and better research into the collections that catalogues describe.

1. Introduction

Sick and infirm patients on crutches and in wheelchairs ("Bath chairs") race down a grassy hill as spectators cheer them along. At the top of the hill, the start of the race is labelled "Cripples Corner" and represents The Crescent; the city of Bath is outlined in the distance.[1]

So reads the ‘Summary’ field of the Lewis Walpole Library’s catalogue entry for Thomas Rowlandson’s satirical print *Bath Races* (see Figure 1), published in London by Thomas Tegg in 1810, written in 2009.[2] The description is simple, clear, and informative, and is part of a catalogue record that records a physical description of the object, transcribes its inscriptions, and categorises it by genre and topic. The record, including its free text description, was created by the Lewis Walpole Library but it was not created in a vacuum. Rather - like any catalogue record - it was shaped by the circumstances of its production.

Figure 1: Bath Races (Thomas Rowlandson, published in London by Thomas Tegg in 1810). Courtesy of The Lewis Walpole Library, Yale University.

[1] LWL Orbis Record 9083108. Lewis Walpole Orbis records are accessible at https://findit.library.yale.edu/.

[2] We know that this record was first created November 16, 2009 as indicated by the first six digits of the 008 M ARC field (YYMMDD). We surmise from circumstances in this case that the summary was probably also written then because the practice at that time was to fully catalogue items going out on loan if no digital catalogue record existed. In this case, this Rowlandson print was requested for loan in 2009 for the exhibition “Thomas Rowlandson: Pleasures and Pursuits in Georgian England” organized by Patrician Phagan at the Frances Lehman Loeb Art Center, Vassar College.

For example, cataloguers will typically consult earlier related records when they are available, such as Mary Dorothy George's work on the *Catalogue of Political and Personal Satires Preserved in the Department of Prints and Drawings in the British Museum* (hereafter the *Catalogue of Political and Personal Satires*). For any cataloguer working on "Golden Age" satirical prints this has long been an essential point of reference, whether via the printed volumes published between 1935 and 1954, one of the microfilm copies first published by Chadwyck-Healey in 1978 to facilitate wider access to these out-of-print volumes, or the British Museum Collections Online website, which launched in 2009 and contains lightly edited versions of George's descriptions for over 11,000 prints, including *Bath Races*.

In the case of the Lewis Walpole Library's entry for *Bath Races*, we know that the corresponding record in *Catalogue of Political and Personal Satires* and its later derivations at British Museum Collections Online influenced the content of the record: George's volume is cited and the web version of the object description is quoted.[3] But there is more to the story here than a nod to a respected forbear. In their production of a description of *Bath Races*, the cataloguers behind the Lewis Walpole description were influenced by the *Catalogue of Political and Personal Satires*, or - to use Michael Baxandall's [1985] more active formulations - they drew on, engaged with, reacted to, differentiated from, remodeled, developed, tackled, even subverted George's description. Thus, where George calls the protagonists 'Cripples and invalids', the Lewis Walpole Library entry calls them 'Sick and infirm patients on crutches and in wheelchairs'; where George has the patients 'rush[ing] down a hill', the Lewis Walpole Library entry has them 'rac[ing] down a grassy hill'; where George sees 'Young women cheer[ing] on the competitors', the Lewis Walpole Library entry sees 'spectators cheer[ing] them along'. By taking the structure of George's description but deviating from its now judgemental tone, those who produced the Lewis Walpole Library's catalogue entry transmitted aspects of George's curatorial "voice" across time and space, even as they worked to tackle or subvert George's linguistic choices.

We start from the position that digitised and digital legacy catalogues entries like these should, assembled at scale, be recognised as highly valuable resources for generating important new knowledge about cataloguing and curatorial practices, their histories, and their consequences. At the same time we respond to the calls [e.g. Bowker and Star, 2000] to recover the motivations behind classification and its impact on the fabric of knowledge when classification systems like legacy catalogues become further uncoupled from their circumstances of production as a result of

[3] This was added in 2013, when summaries on British Museum Collections Online were gradually and systematically appended to Orbis records for Lewis Walpole Library collections.

digitisation. Further, we observe that this uncoupling can produce negative effects when “the future that an algorithmic system can predict is limited by the historical data used to train that system” [Agostinho et al, 2019]. Our concerns with historicizing the catalogue thus intersect with important cross-disciplinary and inter-sectoral research on cultural institutions and power [Duncan, 1995; Perez, 2003], histories of anglophone cataloguing [Hill, 2016; Sutherland, 2017], and metadata [Noble, 2018; Thylstrup, 2019].

Catalogue records underpin the audit, curatorial, and public access functions of collecting institutions. And they are relied upon by many humanities researchers: consider the historian searching archival materials, the literary scholar collating a corpus of books, or the art historian identifying and analysing relevant prints. In these contexts, we argue that it is important to recognise that catalogues are the products of curatorial and cataloguing labour, often spanning many decades, and they are shaped by earlier catalogues, by shifting curatorial and cataloguing practices and priorities, and by broader social circumstances and cultures [Johnson, 1990; Kingdon, 2019; Sutherland and Purcell, 2019; Turner, 2020; Yakel, 2003]. Thus many catalogues comprise subsets of records written at different times, by different people, according to different principles and goals. This situation is exacerbated in those cases where parts of an earlier catalogue are incorporated into a new catalogue. This means that far from a neutral record of collection holdings, catalogues are subject to various biases and inequities, which impact in problematic and potentially unknown ways on how they are used.

Crucially, the historically specific labours and practices of catalogue production are all too easily obscured by the presentation of the catalogue as an always-already present unifying entity, and are further obscured when collections are federated for access, into datasets, or as machine readable endpoints. Cataloguers, curators and researchers will be more or less familiar with the histories of the catalogues that they work with. Hence, to some extent, they will be able to account for how specific labours and practices resulted in particular emphases and absences, biases and inequities. However, such knowledge is often held in the heads of individuals or in annotations made to printed catalogues, and is based on interactions with certain parts of a catalogue. And whilst principles and best practices exist for documenting collections and producing catalogues, there is normally a lack of comprehensive documented information about their use in local cataloguing processes. In turn, collecting institutions have not found ways to make this information useable for publics, a move - as recent work has shown - that can foster anti-racist and anti-colonial practice [Pringle 2020]. Similarly, whilst some records of the production processes of a catalogue may exist, it is often necessary to work backwards from the contents of the current catalogue in order to understand its history, and even then the authors of many records will remain unknown and unknowable.

We contend that manifestations of the historically specific circumstances in which cataloguing takes place, from which biases/inequities arise in the form of linguistic and structural traces, are amenable to computational analysis and are more apparent when a catalogue is analysed at scale with support from computational techniques. Further we contend that by approaching catalogues as data we can generate knowledge about them that supports the writing of their histories, documenting their features when reproduced as datasets, and planning revisions that advance equity and social justice [Cox, 2021; Gebru, 2020; Padilla, 2019].

In previous work we demonstrated how linguistic and structural traces in catalogues manifest the "curatorial voices" of cataloguers and curators, and their institutions. By combining archival research and corpus linguistic analysis we analysed curatorial voice in over 9,000 descriptions of printed images from the *Catalogue of Political and Personal Satires*, comprising around one million words [Baker and Salway, 2020; Salway and Baker, 2020]. We were able to make a systematic account of how the catalogue descriptions included or omitted mentions of certain aspects of objects, how they varied in the degree to which they described or interpreted objects, and how they were shaped by the historically specific circumstances in which cataloguing labour took place. We argued that a computationally-derived characterisation of these choices could be usefully applied to current and future cataloguing practice: e.g. to develop or refine guidelines for object description, and to estimate the person time required to edit or enhance catalogue data.

The current paper is concerned with *transmission* from legacy catalogues to contemporary catalogues and thus builds on and complements the computational analysis of curatorial voice. By "transmission" we refer both to the incorporation of more or less edited catalogue records from a legacy catalogue into a contemporary catalogue, and to the stylistic influence of earlier cataloguing practice in the production of new records.

As a case study we take a catalogue comprising 16,669 records for printed images held by the Lewis Walpole Library, a research centre for eighteenth-century studies in the United States. It was chosen because it is a small-scale and well delimited collection containing free-text descriptions, and particularly because we expected that it would show signs of transmission from the *Catalogue of Political and Personal Satires*. The collections that now form the Lewis Walpole Library began to be assembled by Annie Burr and Wilmarth Sheldon Lewis in the early twentieth century. Documentation of the printed image collections accelerated in the mid-1950s, first with the production of over 12,000 catalogue cards by Annie Burr (officially Curator of Prints from 1957) and her voluntary assistant Elizabeth Creamer [Annie Burr Lewis papers, 1849-1960; Yale University, 1957]. After Annie Burr's death in 1959 cataloguing labour was taken 'vigorously forward' by Genevieve Butterfield [Lewis,

1968]. In 1980 the Lewis Walpole Library became part of Yale University, and from 2003 its collections catalogue was gradually integrated into Orbis, Yale's digital library catalogue. Throughout this period the Lewis Walpole Library's collections were added to and its catalogues updated, requiring particular curatorial and cataloguing expertise to manage and explain the printed image collections, from knowledge of the history of printing to methods for describing visual materials. For one subset of the printed image collections, satirical prints produced in England between the 1770s and 1830s, the Lewises and their staff based their cataloguing on Volumes 5 to 11 of the *Catalogue of Political and Personal Satires*, transforming the entries these volumes contained, all of which were written by George, into index cards that detailed titles, persons, places, events, and keywords [Lewis, 1968; Vermeulen and Carby, 2014]. These volumes were acquired by the Lewis Walpole Library between 1938 and 1970, annotated by the Lewises, and described by Wilmarth as 'invaluable'. Subsequent cataloguers took this practice forward, drawing on the *Catalogue of Political and Personal Satires* and latterly their derivations on British Museum Collections Online.

This paper is about the transmission of Mary Dorothy George's voice through time, across space, and between mediums, from the 1930s to the late-twentieth century and beyond, from the British Museum in London to the Lewis Walpole Library in Farmington, Connecticut, from printed volumes to networked digital data.[4] It aims to show how transmission happens, how it can be found, and how it can be characterised. And it does so because cataloguers like George are the interlocuters between us and the pasts they described, legacy voices that refuse to stay in their historical place, and whose raced, sexed, and classed influence on the future should not go unchecked.

The contributions of this paper are novel computational approaches for detecting records that contain instances of transmission (Section 2) and for analysing these records in order to understand more about how transmission shaped a particular catalogue, and for elaborating general models of transmission (Section 3). The computational approaches are envisioned as part of an overall approach that frames catalogues as data, and that also relies on the expert knowledge of cataloguers and historical research in order to generate knowledge that can form the basis of work to understand, contextualise, and repair collection catalogues.

[4] Note that we do not distinguish between a) transmission from the *Catalogue of Political and Personal Satires* to the Lewis Walpole Library catalogue and b) transmission from the later derivations of *Catalogue of Political and Personal Satires* on British Museum Collections to the Lewis Walpole Library catalogue. These different types of transmission are difficult to disentangle due to cataloguing practices and infrastructural processes that have over-written prior work and not encouraged versioning of catalogue data.

2. Detecting instances of transmission

How can we systematically identify at scale those records in a digital (or digitised) catalogue that are likely to be the products of transmission from earlier catalogues? We present two computational approaches, both of which attempt to identify pertinent subsets of catalogue records based primarily on the content of certain fields in the records, rather than drawing primarily on extant knowledge about the catalogues and the collections they describe. The first approach (described in Section 2.1) should be quite generally applicable to other catalogues in the way it goes about identifying features that may be indicative of transmission. The second approach (described in Section 2.2) pertains to catalogues in which records contain relatively lengthy free-text descriptions of collection items. We show both here as applied to detecting the hypothesised transmission from the *Catalogue of Political and Personal Satires* to the Lewis Walpole Library's catalogue. The outputs from these approaches - i.e. selections of records - are analysed in Section 3 in order to consider to what extent they are instances of transmission, and to explore what we can learn from them about the processes of transmission.

We acquired an export of 16,669 MARC 21[5] records from Orbis, Yale University's online library catalogue,[6] selected by choosing all Lewis Walpole Library records with "k" (two-dimensional nonprojectable graphic) in the "Leader - Type of Record" field.[7] To facilitate subsequent processing the exported XML file[8] was "parsed" such that for each pair of record ID and MARC field, all entries for that field in that record were placed on a tab-separated line comprising record ID, MARC field number, and list of entries for that field in that record.[9] We also had access to a dataset - *the BMSatire Descriptions corpus (BMSat)* - comprising 9,330 descriptions based on entries in Volumes 5 to 11 of the *Catalogue of Political and Personal Satires* that were written by George.[10]

[5] https://www.loc.gov/marc/bibliographic/

[6] https://orbis.library.yale.edu

[7] https://www.loc.gov/marc/bibliographic/bdleader.html

[8] LWL_export.xml in Baker and Salway [2021].

[9] parsed_LWLXML.tsv in Baker and Salway [2021].

[10] Details of the BMSatire Descriptions corpus and its content are available as a Zenodo resource [Baker and Salway, 2019]. This dataset does not associate the descriptions with catalogue record numbers, nor museum registration numbers, which ruled out a straightforward approach to linking some LWL records with BMSat descriptions.

2.1 Using features from various MARC fields to select sets of records

Our first approach to systematically identifying candidate records of transmission from *Catalogue of Political and Personal Satires* to the Lewis Walpole Library catalogue was to automatically generate overviews of the usage and content of the MARC fields across all the records from the Lewis Walpole Library Catalogue and then to manually scan the overviews for signs that particular fields in some records:

(i) explicitly cite the Catalogue of Political and Personal Satires;

(ii) contain evidence that the records could at least possibly be based on the *Catalogue of Political and Personal Satires* in that they relate to objects of the same genre and period as described in the earlier catalogue, i.e. satirical prints from 1771-1832.

For each of the 90 MARC fields used by the Lewis Walpole Library we generated an overview of our dataset that contained: a count of total instances of the field (there may be multiple instances per record); the average "word" count of the content of each field; a sample of up to 20 instances of the field; and eight frequency ordered ngram lists (1<= n <=8) generated from the content of the field after it was split on subfield tags[11]. Figure 2 shows a small part of the overview that was generated for MARC field 600 (Subject Added Entry-Personal Name), with a sample of four entries and the 10 most frequent trigrams. This overview of MARC field 600 shows evidence of point (ii), specifically: persons frequently associated with prints described in the *Catalogue of Political and Personal Satires*; associated date ranges that align with the publication dates (1771 to 1832) for prints George described; and subfields corresponding with the satirical print genre.

```
Tag = "600"

Frequency = 15210

Average number of words, ignoring <>'s (mean, median, mode) = 7.21,
7.0, 6


<marc:subfield   code="a">Telesphorus</marc:subfield><marc:subfield
code="c">(Greek                deity)</marc:subfield><marc:subfield
code="0">http://id.loc.gov/authorities/names/no2013131852</marc:su
bfield>

<marc:subfield       code="a">Leopold</marc:subfield><marc:subfield
code="b">II,</marc:subfield><marc:subfield    code="c">Holy    Roman
Emperor,</marc:subfield><marc:subfield                code="d">1747-
```

[11] See overviewsPart1.txt and overviewsPart2.zip in Baker and Salway [2021].

```
1792</marc:subfield><marc:subfield     code="v">Caricatures     and
cartoons.</marc:subfield>

<marc:subfield                                    code="a">Walpole,
Horace,</marc:subfield><marc:subfield                code="d">1717-
1797</marc:subfield><marc:subfield        code="x">Homes        and
haunts.</marc:subfield>

<marc:subfield             code="a">Barrymore,             Richard
Barry,</marc:subfield><marc:subfield                 code="c">Earl
of,</marc:subfield><marc:subfield                   code="d">1769-
1793</marc:subfield><marc:subfield     code="v">Caricatures     and
cartoons.</marc:subfield>


Caricatures and cartoons.  9714
of Great Britain,     1198
King of Great   1170
Fox, Charles James,   773
Queen, consort of     303
Bute, John Stuart,    295
consort of George     292
Sheridan, Richard Brinsley, 266
Homes and haunts.     259
IV, King of     201
```

Figure 2: A small part of the overview that was generated for the 600 field Subject Added Entry-Personal Name), including four sample entries and the ten most frequent trigrams.

The most frequently occurring MARC field was 500 (General note) with 54,712 instances in the 16,669 records. The MARC fields that occurred more than 16,669 times (and hence potentially more than once in some records) were 655 (Index Term - Genre/Form), 650 (Subject Added Entry - Topical Term), 035 (System Control Number) and 300 (Physical Description). Field 245 (Title Statement) occurred exactly 16,669 times. The fields that occurred more than 8000 times were 040 (Cataloging Source), 043 (Geographic Area Code), 099 (Local Call Numbers), 600 (Subject Added Entry - Personal Name), 700 (Added Entry - Personal Name), 260 (Publication, Distribution, etc. (Imprint)), 100 (Main Entry - Personal Name), 079 (encoding level), 510 (Citation / references note) and 520 (Summary, etc.). Of the remaining MARC fields, 59 of them occurred less than 1000 times and - of those - 48 occurred less than 100 times.

We examined the overview for each of the fields guided by points (i) and (ii) above, starting with the statistics and sample entries. If these indicated anything potentially relevant, we scanned the most frequent n-grams. When an n-gram of interest was noted we then searched for it as part of longer n-grams and in the lists for other fields. Once we had narrowed our focus to specific “clues” in several fields we then analysed their contents in more detail. For example, having noted several ways in which British Museum registration numbers and BMSat numbers were written in the MARC fields 500 and

510, we wrote regular expressions to count in how many records these numbers were present. Table 1 summarises the fields that were of most interest to us and our observations about them[12].

Field number and name[13]	**Frequency**	**Potential Traces of Transmission**
500 General note	54,712	We saw that British Museum registration numbers and BMSat numbers were given in this field using fairly regular formats, e.g. *'No. 11627 in the Catalogue of prints and drawings in the British Museum'* and *'1868,0808.752'*. Presumably arising from records describing satirical prints without these numbers, the 8-gram "Not in the Catalogue of prints and drawings" occurs 3186 times.
510 Citation / references note	9,068	There were 6742 instances of "no." (for number), and frequent trigrams included "v. 5, no.", "v. 6, no." and "v. 7, no." suggesting that - for records describing satirical prints that are in BMSat - BMSat numbers are often recorded in a regular format.
520 Summary, etc.	8,635	When used this field tended to contain lengthy free text descriptions (mean length 83 words) of the pictorial content of the printed image. We observed the British Museum is cited with the 4-gram "British Museum online catalogue" around 3,000 times. We also observed cases where descriptions are contained within double quotation marks and indicate that they are also taken from another source.
600 Subject added entry -- personal name	15,210	3,078 records contain 'Caricatures and cartoons' in this field.
655 Index term -- genre/form	43,661	10,033 records contain 'Satires (Visual works)' in this field.

Table 1: A summary of the fields that became the basis for selecting sets of records that were considered to be likely cases of transmission.

[12] Separately we also noted that the control field 001 (Control number) occurs 16,669 times - once per record - and this gives us a unique reference for each record. In the Lewis Walpole Library this number reflects the order in which records were added to the catalogue which is potentially useful information, although they may have been edited subsequently. As above, the exact date of creation is also indicated by the first six digits of the 008 MARC field.

[13] Names taken from https://www.loc.gov/marc/archive/2002/concise/bibliographic/ecbdlist.html

Based on some of the observations that are noted in Table 1, we then selected two subsets of records for investigating transmission, see Sections 2.1.1 and 2.1.2, see Sections 2.1.1 and 2.1.2.

2.1.1 Selection of records presumed to be written by the Lewis Walpole Library

In order to investigate the influence of the *Catalogue of Political and Personal Satires*, and in particular Mary Dorothy George's voice on later cataloguing at the Lewis Walpole Library, for our first selection of records we created a corpus of descriptions from MARC field 520 that - based on evidence from other MARC fields in the record - could not have been based directly on the *Catalogue of Political and Personal Satires*. Specifically we aimed to select those Lewis Walpole Library records that describe satirical prints that are not described in BMSat. Hence we selected all records that met all the following criteria: (i) the string 'Satires (Visual Works)' appears in the MARC 655 field, and/or 'Caricatures and Cartoons' appears in the MARC 600 field; (ii) in the MARC 500 and 510 fields there are no string matches for patterns that characterise the most common ways in which BMSat and British Museum registration numbers are written; (iii) the string 'not in the catalogue of prints and drawings' appears in the MARC 500 field (case insensitive matching); (iv) there are no matches for the string '--british museum online catalogue' in the MARC 520 field (case insensitive matching); and, (v) there is free text in the MARC 520 field but that free text is not enclosed in quotation marks.

This gave 543 records. Based on subject knowledge and curatorial expertise we judged that 543 is reasonable as the number of satirical prints held by the Lewis Walpole Library that have narrative descriptions and no equivalent in the British Museum. Further, we checked if any of the 543 records indicated reuse of a description in a catalogue other than the *Catalogue of Political and Personal Satires* or of a dealer description. Finally, we checked if any of the 543 records do in fact relate to printed images held by the British Museum, but escaped our tests, therefore making its MARC 520 field entry a potentially unattributed adaptation from the *Catalogue of Political and Personal Satires* or British Museum Collections Online. These checks revealed no issues, and so we subsequently made a corpus of descriptions that we are confident were written by staff at the Lewis Walpole Library without reference to the *Catalogue of Political and Personal Satires;* see Section 3.2 for analysis of this corpus.[14]

[14] The 543 selected records are listed in 543Records_forCorpusOfWrittenByLWL.xlsx [Baker and Salway, 2021]; they were selected by filtering a larger set of records according to features given in forCorpusOfWrittenByLWL.csv. The criteria for filtering may have been stricter than necessary and led to some 'valid' records being excluded, but we were prioritising the precision of the results over recall.

2.1.2 Selection of records that refer to different impressions of the same print

The second selection of records that we made was based on the observation that different Lewis Walpole Library records mentioned the same BMSat number in either the MARC 500 or 510 fields. This indicates a phenomenon particular to catalogues of collections of printed images, whereby institutions hold multiple unique impressions, versions, states, or copies of the same printed image.[15] Because these multiple impressions, versions, states, or copies of the "same" printed images may arrive at collecting institutions at different times, be catalogued by different individuals working in different institutional contexts, and are subject to different professional and scholarly influences, this raises the possibility that variation between catalogue records for the "same" image could support the investigation of transmission from BMSat to the Lewis Walpole Library catalogue by contrasting those records. For example, we might consider adjustments made to legacy descriptions in order to account for slight differences between collection objects, observe differing uses of BMSat as a source, or think through the possible rationale for and historical specificity of widely divergent descriptions.

We made a dataset that grouped descriptions from each set of Lewis Walpole Library records that refer to the same BMSat number.[16] There are three groups with six Lewis Walpole Library records; four groups with five records; 11 groups with four records; 59 groups with three records; and 441 groups with two records. For an example, see Figure 3 which shows the contents of the MARC 520 field in four different Lewis Walpole Library records that all appear to refer to different impressions and copies of the same satirical print (BMSat 4050).[17] See Section 3.1 for the analysis of the selection of 1123 records.

```
Sketch of John Wilkes holding a "Staff of Maintenance" wtih the cap of
Liberty on top, drawn at the time of Wilkes' second trip to Westminster
Hall for slander. On the table beside him are two newspapers -- North
Briton Number 45 and North Briton Number 15 -- which allude to Wilkes'
attack on Hogarth and King George III.
```

[15] During the "Golden Age" of British satirical prints, print reproduction used craft-like technologies that made exact reproduction unachievable: for example, each copper plate engraving or etching was inked by hand, lines incised on copper plates lost definition each time they we used in printing, and colouring was completed by outworkers by hand. As a result, whilst multiple impressions from the same plate often constituted part of the same print run and would have contained the same narrative content, the impressions often sufficiently differ so as to require divergent description. For the history of "Golden Age" British satirical prints and their production see Baker [2017], Donald [1996] and Gatrell [2006]. For the production of printed images in long-eighteenth century Europe, see Stijnman [2012] and Griffiths [2016].

[16] withBMSatNumbers_multipleRecords.xlsx [Baker and Salway, 2021].

[17] There is a fifth Lewis Walpole Library record referring to this print but its description field is empty. BMSat records are accessible at https://www.britishmuseum.org/collection.

```
John Wilkes is shown holding a "Staff of Maintenance" wtih the cap of
Liberty on top, drawn at the time of Wilkes' second trip to Westminster
Hall for slander. On the table beside him are two newspapers -- North
Briton Number 45 and North Briton Number 15 -- which allude to Wilkes'
attack on Hogarth and King George III.


Caricatural portrait of John Wilkes sitting on a chair holding stick
topped by a cap of Liberty.  On the table beside him are two issues of
the newspaper North Briton, nos. 17 and 45 as well as a box with a
feather pen in an inkwell.


Caricatural portrait of John Wilkes sitting on a chair holding stick
topped by a cap of Liberty.  On the table beside him are two issues of
the newspaper North Briton, nos. 17 and 45 as well as a box with a
feather pen in an inkwell.
```

Figure 3: The contents of the 520 field in four different LWL records that all seem to refer to different impressions and copies of the same satirical image (BMSat number 4050).

2.2 Using a text distance metric to find pairs of related descriptions in two catalogues

Our second approach to systematically identifying candidate records of transmission from *Catalogue of Political and Personal Satires* to the Lewis Walpole Library catalogue addresses cases where the writing of descriptions in records has involved copying (parts of) descriptions from earlier catalogues with some degree of subsequent editing and/or addition of new text, but there is insufficient metadata to match up pairs of records from the two catalogues. To find these records we made pairwise comparisons of all the descriptions from one catalogue with all descriptions from the other using a text distance metric, and ranked the pairs of descriptions according to the metric. From this the pairs ranked as most similar (lowest text distance) can be manually inspected as candidate examples of transmission.[18]

There are many different text distance metrics belonging to several families. We conducted trials to determine which is the most effective metric for our purposes, i.e. which most consistently ranks pairs of descriptions that are examples of transmission more highly than other pairs. We also considered how resource intensive the metrics are.[19]

[18] Such an approach is appropriate to a case like ours where descriptions for one catalogue (BMSat) are not associated with catalogue or registration numbers; and, as it happens we cannot be sure that all LWL records refer to BMSat catalogue/registration numbers when they should.

[19] Whilst the resource intensivity of our work is not equivalent to that of, say, training a large language model, we contend that all DH work should be mindful of its resource intensivity. This is particularly true for any DH work that has justice-oriented goals, because without attending to how research takes place, that research can

First, we hand-picked a set of fifteen pairs of descriptions that exhibited varying degrees of copying, editing, and addition, including pairs of descriptions that related to different collection items with similar content, and dissimilar pairs. We then generated results (in most cases text distance metrics) for each pair using sixteen of the functions that are implemented in the Python *textdistance* package,[20] with several from each of four families of function.[21] Broadly speaking we observed the following.

Edit-based metrics (Hamming and Levenshtein): these metrics count how many character-level changes would be needed to change one string into another, so the higher the value of the metric the greater the difference between the two strings, i.e. values closer to zero would be suggestive of transmission. In most cases these metrics are not suitable for our purposes because the editing of descriptions often entails inserting or deleting words that shift the positions of many other characters, causing the edit distance to increase greatly even if the strings share much in common. For example, see Table 2 for a pair of descriptions from the Lewis Walpole Library and BMSat respectively, which gave distances of 408 (Hamming) and 133 (Levenshtein) even though it would appear that the Lewis Walpole description is very much based on the one in BMSat.[22] It seems that these metrics would only be effective in cases where very few edits were made or edits were direct replacements of single characters, e.g. if single quotation marks were replaced with double quotation marks throughout a description.

LWL Orbis Record 13201977	**BMSat 15731A**
A satire on the Duke's pressure on the King to accept Emancipation. **Wellington stands in profile to the right, dressed as the driver of a mail-coach, holding his whip and (as way-bill) a paper resembling the 'Gazette', headed 'Bill' [i.e. for Catholic Relief]. His (gloved) left hand touches the broad brim of his hat. He wears a**	**Wellington stands in profile to the right, dressed as the driver of a mail-coach, holding his whip and (as way-bill) a paper resembling the 'Gazette', headed 'Bill' [i.e. for Catholic Relief]. His (gloved) left hand touches the broad brim of his hat. He wears a triple-caped greatcoat, tight at the waist, over tightly**

cause unintended harms. For how we frame our computational work see Sussex Humanities Lab Carbon Use and Environmental Impact Working Group et al [2020]. The first author in particular would like to recognise Emily M. Bender, Timnit Gebru, Max Liboiron, Angelina McMillan-Major and Shmargaret Shmitchell for inspiring their ongoing (often unsuccessful) attempts to ground their DH research methods in practices that reduce environmental and ecological harms without amplifying colonial harms, in an environmentalism that is intersectional [Bender et al., 2021; Liboiron, 2021].

[20] https://pypi.org/project/textdistance/

[21] The file 15pairs_16text-distance-metrics.txt [Baker and Salway 2021] shows the results from the sixteen functions for fifteen pairs of BMSat-LWL descriptions which were chosen to reflect a range of similarity/transmission.

[22] This pair of examples and the other pairs presented subsequently come from the previously mentioned results file.

triple-caped greatcoat, tight at the waist, over tightly strapped white trousers, and is smart and erect.	**strapped white trousers, and is smart and erect**, in contrast with his rival, see BM Satires No. 15736. April 1829

Table 2: A pair of apparently related descriptions with common text bolded.

Token-based metrics (Jaccard, Tversky, Tanimoto and cosine): these metrics treat the two strings as bags of words and measure how many words they have in common - that is, they are not concerned with the order or sequences of words. The Tanimoto function gives a negative value, where the lower the value the greater the distance between the strings. The other three functions give values from 0 (no words in common) to 1 (the counts of words in the strings are identical). We observed that all four functions tended to give results that were proportional to one another, and they were mostly effective at separating cases of transmission (high similarity) from non-transmission (low similarity). However, they were susceptible to cases in which pairs of descriptions shared much vocabulary whilst not being examples of transmission: this happens when descriptions of different impressions of the same printed image are written independently but refer to many of the same things in the printed image, or when two descriptions relate to two different printed images that depict many similar kinds of things. For the example in Table 3, which shows descriptions of two different impressions of the same printed image which appear to us not to be based on one another, the Jaccard, Tversky and cosine functions all returned scores greater than 0.8, and the Tanimoto function returned -0.31.

LWL Orbis Record 11648648	**BMSat 9672**
Seven men are gathered around a gambling table in a tavern, two of them playing at cards, others watching. The man on the far right is fast asleep, his dog's head resting on his knee. In the background, a barmaid tallies up the drinks inside a bar. The game is between a shrewd looking man on the left and a tallow youth on the right who is receiving bad advice from a man to his right, with a glass in hand. Behind the youth a broken mirror hangs tilted on the wall. Below it, one of the onloookers is leaning over the back of the settee peeking at the youth's cards. Standing in the center is an obese man holding a bowl and smoking a pipe.	Seven men (three-quarter length) are grouped round a card-table in a Smithfield tavern. One (right), young and innocent, inspects his cards; beside him an older countryman lies back asleep (right), his dog resting his head on his knee. The other gambler (left), holding his cards, looks at his victim. Three onlookers have crafty expressions. A fat man, smoking, approaches with a bowl of punch. In the bar (left) a fat woman chalks up a score. Coins, a watch, and pocket-book are on the table. A broken mirror and a picture of a horse decorate the walls. Beneath the table are twelve lines describing the sleep of 'Old Trusty' while his son is cheated by 'the Harpy-Tribe'.

Table 3: A pair of descriptions for the same print that do not appear to be directly related, whilst sharing much common vocabulary.

Sequence-based functions (Ratcliff Obershelp (RO), Longest Common Substring (LCS)): the LCS function simply returns the longest substring that the two strings share in common[23]. For the pair of descriptions in Table 3 this is 'broken mirror' which quickly suggests that the descriptions were written independently, whereas a long substring could indicate copying. We could count the number of characters in the longest common substring in order to rank pairs of descriptions as a percentage of the length of the description. However, it would give a misleading low score on cases where a copied description has been lightly edited in several places throughout. The RO function accounts for multiple common substrings, and returns a value between 0-1 which gets closer to 1 as the two strings share more and longer sequences of words. Unlike the token-based metrics, it seems to be effective at distinguishing pairs of descriptions that are partial copies from pairs that just happen to share a lot of common vocabulary: for the pair in Table 3, RO gave a score of 0.37, and for the pair in Table 2 it gave 0.84. The only problem we noted with RO was that, like all the other functions we considered, it is vulnerable to pairs of short descriptions that share multiple multiword technical and indexing terms, whilst not being examples of transmission.

Normalised compression distance (bwtrle_ncd, sqrt_ncd, bz2_ncd, lzma_ncd, zlib_ncd): in principle the application of a text compression algorithm to two strings serves to amplify their similarities or differences; the five functions we tried correspond to five different compression algorithms. However, we struggled to understand how these functions work and how to interpret their results. For example in a preliminary trial they gave the following values for pair of identical descriptions when, intuitively we might have expected either 0 or 1: bwtrle_ncd (0.67), sqrt_ncd (0.41), bz2_ncd (0.23), lzma_ncd (0.02), and zlib_ncd (0.03). This discouraged us from exploring them further, but this is not to say that they do not warrant further consideration for the task of detecting transmission.

An optimal approach might combine several text distance functions in order to detect the broadest possible range of copying/editing/addition, e.g. direct complete copy, with minor formatting or other changes; incorporation of (part) of one description into another; and, paraphrasing of another description. However it would require much more research (beyond the scope of this paper) to understand how to effectively combine the functions and optimise their use, so we proceeded using just one.

[23] Related to this are the prefix and postfix functions which return the substrings that the two given strings share at the beginning and at the end.

As noted, the Ratcliff Obershelp (RO) function looked to be the most suitable for our purposes, however we realised that it would be prohibitively resource intensive for the task at hand. We estimated that it would take up to two years to run a pairwise comparison of 4545 LWL x 9330 BMSat descriptions, using a relatively high-powered workstation and without exploiting parallelisation. Usefully in the standard Python library the Sequence Matcher class includes an implementation of a text distance function which is considered to be similar to RO. With a further trial we observed that this function produces results that are similar to or better than RO, i.e. it makes similar, and sometimes sharper, distinctions between transmission and not transmission.[24] Importantly it executes significantly more quickly than RO - according to its documentation[25] it executes in linear time in the best case and in quadratic time at worst; this compares with cubic time in the worst case for RO. Furthermore, the function can very quickly calculate a highest-possible value ('sequenceMatcher.realquickratio'), allowing for many pairs to be discounted before a more expensive and accurate calculation is done ('sequenceMatcher.ratio').

For the pairwise comparison of 4545 Lewis Walpole Library descriptions with 9330 BMSat descriptions we first pre-processed the Lewis Walpole Library descriptions to mirror how the BMSat descriptions had been prepared. This involved some text clean-up and substituting bracketed and quoted text with the strings 'BRACKETED' and 'TRANSCRIBED'.[26] Then we executed a script which for each LWL-BMSat pair: (i) computed sequenceMatcher.realquickratio and skipped to the next pair if this was < 0.9; then, (ii) computed sequenceMatcher.ratio[27] and filtered results to remove pairs for which SM.ratio < 0.5 which, from inspection of trial results, we are confident is a suitable threshold to remove only pairs that are not examples of transmission. The remaining 1649 pairs were written to a file with their scores[28] for subsequent analysis (see Section 3.X). The script took about 5 hours to execute.

[24] Results from our trial comparing Ratcliff Obershelp with the Sequence Matcher functions are given in RatcliffObershelp_sequenceMatcher_comparison.txt [Baker and Salway 2021].

[25] https://docs.python.org/3/library/difflib.html#difflib.SequenceMatcher

[26] We refer the reader to the earlier paper [Salway and Baker, 2020] for details of the pre-processing of BMSat. Here and in LWL 543 transcribed text and bracketed text were substituted with the strings 'TRANSCRIBED' and 'BRACKETED', so that this text - which is not original to the curator/cataloguer's description - is ignored in the analysis.

[27] Two parameters were set: autoJunk = True (default value); and, isJunk = None. Note the sequenceMatcher.ratio function is not commutative and we only generated results with the LWL descriptions passed as the first text string, and the BMSat descriptions passed as the second.

[28] pairwiseComparisonResults.tsv [Baker and Salway 2021].

3. Analysing transmission

The three selections of records made in Section 2 were analysed in order to consider to what extent they contain instances of transmission, and to explore what we can learn from them about the processes of transmission. Here we describe methods for characterising transmission and elaborate on different kinds of transmission from BMSat to records at the Lewis Walpole Library. It is important to note that this analysis is situated in our shared knowledge of and familiarity with the catalogues at hand.[29] As such then, our case study and our analysis of it provides a way into developing a conceptual framework for thinking about the different ways in which catalogues are shaped and constructed over time. But other case studies of transmission between catalogues, that - we contend - could follow our approaches to select candidate records and analyse those selections, will provide different contributions to that framework.

3.1 Computer-assisted close reading

In this sub-section we focus on the close reading of computationally selected groups of descriptions, a method akin to what Martin Paul Eve calls *Close Reading with Computers* [2019]. The selection reported in Section 2.1.2 consists of 518 instances in which the Lewis Walpole Library catalogue contained multiple records for different impressions of the “same” printed image; there were 1141 records in total. We read the MARC 520 ‘Summary’ Fields (used by the Lewis Walpole Library for object descriptions) for these records. In the majority of instances the linked records contained either identical descriptions or a combination of an empty field and a description. In the remaining instances, we observed from the MARC 520 ‘Summary’ Field signs of transmission that fell broadly into one of three categories:

1. Descriptions which are copies or near copies of descriptions of different impressions of the “same” printed image in other catalogues, either a) with the copied text quoted and with an accompanying citation in the MARC 520 Field (most of which appear to come from British Museum Collections Online) or b) with the text unquoted and with an accompanying citation in the MARC 500 or 510 Fields. For example, in the case of *A Nincompoop, or hen peck'd husband* (1807) there are two Lewis Walpole Library descriptions. In one of these Lewis

[29] For example, even though the methods used in Section 2.2 do not account for directionality of influence, our analysis of the pairs selected in Section 2.2 assumes that since the mid-twentieth century cataloguing of satirical prints at the Lewis Walpole Library was a task potentially influenced by - drawing on, engaging with, reacting to, etc - cataloguing that took place at the British Museum in the early- to mid-twentieth century, and not vice versa.

Walpole Library descriptions the opening line of the British Museum Online description is edited from 'A small ugly man trots (l. to r.)' to 'A small ugly man trots (walking left to right)', the British Museum Online description[30] is placed within double quotation marks, and '-- British Museum catalogue' is inserted after the quotation marks.[31] In the second of these descriptions the British Museum Online description is again placed within double quotation marks, but on this occasion it is followed by the wording 'British Museum online catalogue, description of an earlier state'.[32]

2. Descriptions which take narrative or linguistic elements from descriptions of different impressions of the "same" printed image in other catalogues, usually British Museum Collections Online or the *Catalogue of Political and Personal Satires*, but contain significant differences or adaptations. For example, we find Lewis Walpole Library descriptions that are summaries of corresponding descriptions in another catalogue. This is the case with *A fashionable suit!* (1800), for which the Lewis Walpole Library catalogue description retains the structure of the British Museum Collections Online description but in truncated form (see Table 4).

LWL Orbis Record 8212182	**BMSat 9625**
A satire on the new fashion of Jean Debry coats: A tailor holds a mirror to a customer who looks at his image with horror. The customer complains that he has put a hump upon each shoulder. The tailor replies that the coat has been made to his wife's specifications.	A tailor (left) holds out a mirror to an ugly and disgusted customer, who wears a coat of 'Jean de Bry' pattern, see BMSat 9425, with short bulky breeches and slippers. The customer says: "Why you have put me a hump upon each shoulder - and here's a pair of Dutchman's breeches that would hold provision for a marching regiment - well I tell you what Master Taylor D------m me if I would go to our Club such a figure for fifty Pounds!" The tailor, alarmed, answers: "Made entirely to your Lady's Orders your Honor I asshure you - she said now you was married you should look like the rest of the world."

Table 4: An example of a Lewis Walpole Library description adapting a BMSat description.

[30] BMSat 10909

[31] LWL Orbis Record 8638085

[32] LWL Orbis Record 8638683

3. Descriptions which do not appear to copy or take direct narrative or linguistic elements from descriptions of different impressions of the "same" printed image in other catalogues, and so are believed to be Lewis Walpole Library originals. These are few in number, and include the description of *Bath Races* explored in Section 1.

The selection made in Section 2.1.2 enabled us to observe the presence of transmission from BMSat to the Lewis Walpole Library, and to begin to understand the variability of that transmission, but the analysis is insufficient alone to form the basis of a conceptual framework for thinking about the different ways in which catalogues are shaped and constructed over time. The selection of LWL-BMSat 1649 pairs ranked by a text distance metric (described in Section 2.2) provided then a usefully different set of descriptions to read closely.[33]

In an echo of the Section 2.1.2 selection, the majority of pairs ranked highly as candidate examples of transmission by Sequence Matcher are very close matches: roughly 3 in 4 pairs have a SM.ratio greater than 0.9, and these pairs differ only by single word replacements (e.g. "pig which" to "pig who"), the removal of a series number, or the expansion of an acronym. The pairs with a lower SM.ratio score exhibit different phenomena. One pair (SM.ratio=0.5) shows transmission from BMSat to Lewis Walpole Library that includes removing an opening line ("Heading to printed verses *BRACKETED*") in favour of placing that information in a different MARC field, and a conclusion that adds additional descriptive detail ("They raise glasses whose stems have been broken").[34] A second pair (SM.ratio=0.58) indicates direct transmission from BMSat to Lewis Walpole Library, but with a new line inserted in the middle of the description that emphasises an aspect of the print omitted from the BMSat description.[35] We also find in this selection the description of a Rowlandson border fragment published in 1799 that is not in the *Catalogue of Political and Personal Satires* but is written in the style so closely resembling descriptions in the *Catalogue of Political and Personal Satires* that it is paired (SM.ratio=0.5) with a description for a different print, see Table 5.

LWL Orbis Record 8186561	**BMSat und (1876,1014.67)**
One image only. An elderly woman and a young man face an obese parson who is apparently about to marry them. The young man seems to	Satire; a woman stands, raising a whip over a man kneeling in supplication, next to a gallows;

[33] One in five of the LWL records in the selection contain no mention of BMSat, and so would have evaded methods described in Part 2.1

[34] LWL Orbis Record 8782502; BMSat 10958.

[35] LWL Orbis Record 7949892; BMSat 8408.

be moving away from his smiling bride, saying: *TRANSCRIBED*	behind the man stands a devil with a fork, smiling and saying *TRANSCRIBED* .

Table 5: An example of a Lewis Walpole Library description that shares the style of a BMSat description whilst not being based on it directly.

But these are edge cases. As we read across all the selected pairs, more regular patterns of transmission start to emerge. Lewis Walpole Library entries expand on abbreviations in BMSat, for example, "BMSat 4922" becomes "British Museum satire no. 4922".[36] Lewis Walpole Library entries systematically revise quirks of BMSat descriptions: for example, they remove brackets from spatial vocabulary and quotation marks from eighteenth-century terminology.[37] And in Lewis Walpole Library descriptions leading and trailing details from BMSat descriptions are removed and distributed to other parts of the catalogue entry. These patterns of revision are not always consistent. For example, the Lewis Walpole Library description for *The rake's progress at the University* (1806) is a lossy transmission in that it omits from the description the leading text "See BMSat 10639", a reference to another print in the *Catalogue of Political and Personal Satires*, without finding it a place elsewhere in the record.[38] More often, however, these details are retained, such as in the Lewis Walpole Library entry for Rowlandson's 1807 print *Mrs. Showwell* which removes from the close of the description "26 February 1807. Hand-coloured etching.", and places those details appropriate to the object at hand (it is not coloured) in the MARC 260 and 300 Fields respectively.[39] Here transmission intersects with temporally specific cataloguing infrastructures. Printing methods, dates of production, and references to other prints were details that it made sense for Mary Dorothy George to include in the main body of her object descriptions: in a printed catalogue, the consistent placement of information on the page and relative to other details was a proxy for record structure. When these records were moved to British Museum Collections Online, they were parsed in largely unadapted form so as to create unmarked sub-fields in a flexible collection database designed for a large and varied museum collection [Griffiths, 2010]. But when parsed to fit first a card catalogue system devised by the Lewis's

[36] LWL Orbis Record 8362224

[37] LWL Orbis Record 7951545; BMSat 7467.

[38] LWL Orbis Record 8545613; BMSat 10641.

[39] LWL Orbis Record 8626750; BMSat 10786.

and later a digital catalogue built on a granular bibliographic standard capable of handling hundreds of unique fields, it appears that it made greater sense to move these details to dedicated fields.[40]

If this reading provides insight into how cataloguing infrastructures shaped the transmission of catalogue data over time, the selection made with a text distance metric in Section 2.2 also underscores the role of objects in shaping transmission: in short, on numerous occasions Lewis Walpole Library catalogue entries diverges from BMSat equivalents so as to better represent the object at hand. For example, in the Lewis Walpole Library catalogue entry for *Mrs. Showwell* the opening line - "Below the title: 'The Woman who shews General Guise collection of Pictures at Oxford'" - is removed as the Lewis Walpole Library version is trimmed and thus contains no title. In the Lewis Walpole Library description of a Peter Pinder print from 1787, the opening to the corresponding BMSat entry - "Proof without letters" - is removed as the Lewis Walpole Library object is not a proof.[41] And in the Lewis Walpole Library catalogue record for a 1828 print depicting chess players, the MARC 520 field repeats the description found in BMSat for a near identical print from circa 1788, but removes the opening line - "Title perhaps cut off" - because the Lewis Walpole Library object contains the full print and shows that no title was ever present.[42]

The text distance selection thus enabled us to build on model of transmission discussed above and to identify three categories of divergent transmission:

1. **procedural divergence**, in which BMSat records were adapted to align with cataloguing practices at and/or infrastructural choices made by the Lewis Walpole Library;
2. **divergence due to variant objects**, in which BMSat records were adapted to align with material characteristics of objects held by the Lewis Walpole Library;
3. **revisions to descriptions**, whether in the form of truncated prose, additional detail, or word switches, the motivations for which are hard to ascertain.

Taken together, close reading suggests that many Lewis Walpole Library records projected Mary Dorothy George into the future, lightly editing her on the way, but doing little to arrest her historically specific interlocution between us and the catalogued past. Where George was revised for reasons other than cataloguing procedure or variant objects, we glimpse the aspects of her descriptions that

[40] The British Museum migrated entries from the *Catalogue and Political and Personal Satires* to, first, fields used in their internal database and, later, fields used in their British Museum Collections Online [Griffiths, 2010].

[41] LWL Orbis Record 7780448; BMSat 7188.

[42] LWL Orbis Record 9760098; BMSat 7400.

most troubled subsequent cataloguers. Some of these revisions suggest an attentiveness to the identities that George saw as normative. For example, the Lewis Walpole Library entry for *A West India sportsman* adapts the opening "The sportsman sits in a chair.." to "The English sportsman sits in a chair..", in recognition of the fact that in BMSat men are Englishmen unless described otherwise.[43] And whilst the retention of a racial epithet in the Lewis Walpole Library entry for *A West India sportsman* reminds us that such revisions were themselves historical acts, and whilst as Vermeulen and Carby [2014] note in the Lewis Walpole Library catalogue '[t]here is no subject "White"; white people are, quite literally, the unmarked bodies of the archive', other revisions suggest some recognition of George's historically specific choice of language. For example, the Lewis Walpole Library entry for Woodward and Tegg's 1807 *A riddle expounded, or, The dignity of a parsons horse* - an interaction between a parson and "a jovial countryman" that sends up monarchical power - shows transmission intersecting with changing curatorial sensibilities (see Table 6). This print features a heavily caricatured parson, a hybrid of two long-eighteenth century stereotypes: the money-grubbing tithe pig and the drunken clergyman popularised by the ballad "The Vicar and Moses" [Virgin, 1989]. In turn George described him as "drink-blotched and prosperous-looking",[44] language revised by a Lewis Walpole Library cataloguer(s) to "red-faced and freckled and prosperous looking, with a round belly".[45] This revision performs three functions: first, it delegitimises pejorative associations of facial markings with alcoholism; second, it provides an example of what "prosperous looking" might look like to long-eighteenth century British satirical audiences; and third, it expands - if only a little - the assumed readers of the description from a narrow group of experts in long-eighteenth century British history to a more general public. Taken together with other revisions made to George's original - a correction to the mounting point of a railing, the creation of greater motion in spatial terminology, the assertion that to have a sermon in one's pocket is to carry it - the Lewis Walpole Library description of *A riddle expounded* transmits George for the present: for search, for diverse audiences, and for a context in which collecting institutions recognise and address their roles in maintaining structural inequalities.

LWL Orbis Record 8712571	**BMSat 10904**
A jovial countryman leans on a rustic railing next to a tree, to address a fat elderly parson on	A jovial countryman leans on a rustic railing nailed to a tree, to address a fat elderly parson

[43] LWL Orbis Record 11990160; BMSat 10804.

[44] BMSat 10904.

[45] LWL Orbis Record 8712571.

horseback (riding to the left). He asks, “Ha! Ha, the knaust Doctor I be a rum fellow, Canst thee tell me why a parsons horse be like a king?” The parson answers with a grin, “Why you rogue, because it is guided by a minister.”; He is red-faced and freckled and prosperous looking, with a round belly; he carries a sermon in his pocket whose title is “Sermon to be prea[ched] …”	on horseback (r.). He asks "Ha! Ha - the knaust Doctor I be a rum fellow, - Canst thee tell me - why - a Parsons Horse be like a King?" The parson answers with a grin: "Why you rogue, because it is guided by a Minister." He is drink-blotched and prosperous-looking; in his pocket is a 'Sermon to beprea[ched] ...'. Copied in BMSat 10916. 1807

Table 6: An example of transmission from BMSat to the Lewis Walpole Library catalogue which reflects changing sensibilities.

The case of *A riddle expounded* is also a validation of our methods. From George’s 80 word description of the print, thirteen words are edited or replaced by the Lewis Walpole Library and sixteen words are added. These edits are not clustered together, they change every line and most clauses, such that a 58 character string - “to a tree, to address a fat elderly parson on horseback (r” - is the longest shared by both the BMSat and Lewis Walpole Library descriptions. By using a suitable text distance metric we were able to surface transmission like this, to read it, and to report on it. In so doing we expanded our knowledge of BMSat to Lewis Walpole Library transmission beyond MARC fields that record transmission, revisions between descriptions of different impressions of the “same” printed image, and descriptions truncated by the omission of opening or closing details. This understanding of transmission was complemented by analysis, reported in the next sub-section, of those descriptions presumed to be written by the Lewis Walpole Library.

3.2 Corpus linguistic analysis

In this sub-section our analysis switches from the close reading of individual descriptions to a corpus-level comparison of descriptions from the two catalogues at hand. We previously conducted a corpus linguistic analysis of the BMSat corpus which identified a characteristic curatorial voice in its descriptions, in terms of what things were typically referred to or ignored in descriptions, and in terms of the degree of description/interpretation/evaluation [Salway and Baker, 2020]. In Section 2.1.2 of the current paper we described the creation of a set of 543 Lewis Walpole Library descriptions which we believe were not based directly on any BMSat descriptions. If this set of Lewis Walpole Library descriptions exhibits similar characteristics to the BMSat corpus then that would support the idea of transmission of style, although this alone would not be sufficient evidence to say that the Lewis Walpole Library style was influenced by George’s style rather than their both being influenced by something else.

We applied the method from Salway and Baker [2020] to analyse the 543 Lewis Walpole Library descriptions (hereafter "LWL 543"). Here we show and discuss the results alongside the earlier results for the BMSat corpus;[46] given that the LWL 543 comprise only about 33,000 words we recognise that any conclusions drawn from the analysis can only be tentative.

Table 7 shows the 100 most frequent words in BMSat alongside the 100 most frequent in LWL 543 (cf. Figure Table 2 in Salway and Baker [2020]). Words that occur in both top 100's are shown in bold, whilst the remaining words are underlined and their rank position in the other corpus is given in brackets[47]. There are seventy words in the top 100's for both corpora, and many of these appear in similarly ranked positions. Furthermore, many of the thirty remaining words in each list occur in rank position 101-200 in the other list, i.e. sixteen of the BMSat top 100 are in positions 101-200 in LWL 543, and nineteen of the LWL 543 top 100 are ranked 101-200 in BMSat.[48]

Whilst some words would be expected to feature high in the frequency list for most English-language corpora, e.g. 'the', 'a', 'of' etc., many words that are prominent in both lists are particular to the task of describing visual objects, and satirical prints specifically, e.g. 'left', 'right', 'wearing', 'hat', 'dressed', 'background'. Thus the results shown in Table 7 suggest that the LWL catalogue contains descriptions that, whilst not based directly on specific descriptions from BMSat, share a style or curatorial voice.

Top 100 words in BMSat	Top 100 words in 'LWL 543'
the, **a**, **of**, **and**, **is**, **in**, **his**, **with**, **on**, **to**, **The**, **are**, **A**, **by**, inscribed (225), **from**, **which**, **left**, **right**, **an**, **at**, **He**, **stands**, **her**, says (170), **who**, **On**, **he**, **man**, **two**, **him**, **hand**, **head**, **holding**, **one**, **holds**, **as**, **In**, **behind**, **large**, No (---), **other**, profile (134), **wearing**, **hat**, saying (458), **up**, wears (101), Behind (108), **has**, **back**, **sits**, **it**, **for**, **out**, **over**, **table**, **woman**, **three**, towards (119), or (103), **their**, design (503), **small**, **that**, paper (183), **arm**, but (140), Fox (1002), **hands**, **each**, title (145), BMSat (---), round (370), **men**, **dressed**, **them**, **background**, extreme (288), **long**, **His**, its (106), **looks**, ground (161), stand (113), Lord (253), **under**, **Two**, See (1483), **wall**, John (424), &c	**a**, **the**, **of**, **in**, **and**, **on**, **his**, **with**, **is**, **A**, **to**, **her**, **man**, **The**, **an**, **right**, **left**, **at**, **as**, **from**, **woman**, **are**, **In**, **hand**, **two**, **large**, young (126), **by**, **one**, **table**, **On**, **who**, **he**, **stands**, **holds**, **head**, **him**, **behind**, **sits**, **hat**, **other**, **their**, **which**, **wall**, **holding**, **dressed**, **men**, **up**, image (4165), **them**, **back**, face (101), while (153), **over**, **He**, chair (138), another (164), **it**, **looks**, **above**, **has**, **small**, **under**, **Two**, **that**, **arm**, **for**, wig (106), **background**, An (119), front (116), **each**, looking (110), dog (203), **out**, top (500), figure (165), room (215), she (139), **wearing**, floor (223), side (108), around (2654), look (314), **long**, **three**, scene (193), very (158), before (188), stick (436),

[46] We refer the reader to the earlier paper [Salway and Baker, 2020] for details of motivation, methodology, results and discussion of curatorial voice.

[47] Capitalised words are counted separately from their non-capitalised versions, this proved to be fruitful in the previous analysis.

[48] Some cases of bigger differences may be due to varying content in the prints being described, e.g. 'Lord', 'Fox' and 'John' being in the BMSat top 100.

(2221), have (204), arms (130), **seated**, **above**, She (129), beside (115), I (153), Below (324)	be (121), door (131), **hands**, walking (560), women (213), **His**, glass (270), hangs (159), open (103), **seated**

Table 7: The 100 most frequent words in BMSat and in 'LWL 543', listed in frequency order. Bolded words occur in both top 100s. Numbers in brackets are the rank position of that word in the frequency list for the other corpus.

The next main step in the analysis of BMSat descriptions was to group the 300 most frequent words according to the kind of information they provide, see Table 3 in Salway and Baker [2020]. Table 8 in the current paper uses the same categories to group the top 100 words in LWL 543[49]. The fact that these words fit well in the same categories suggests that the LWL descriptions are providing similar informational content as the BMSat descriptions. For example, 10 of the 18 nouns listed in Table 8 appear in the corresponding list for BMSat: the remaining 8 (chair, wig, door, dog, women, room, floor, stick) are similar in referring to people and everyday things in quite generic ways.

Content descriptors	**Nouns (mostly people, body parts, objects, clothes)**	man, woman, hand, table, head, hat, wall, men, chair, arm, wig, dog, room, floor, stick, door, hands, women
	Verbs (mostly physical actions)	stands, holds, sits, holding, dressed, looks, looking, wearing, look, walking, hangs, seated
	Adjectives (physical properties, appearance)	large, young, small, long
	Names (mostly people)	---
Meta/special	**Art terms**	image, figure, scene
	Prepositions for spatial organisation	right, left, behind, above, under, background, front, side
	Misc.	---
Function words		a, the, of, in, and, on, his, with, is, A, to, her, The, an, at, as, from, are, In, two, by, one, On, who, he, him, other, their, which, up, them, while, over, He, another, it, has, Two, that, for, An, each, out, she, around, three, very, before, be, His, open

[49] Given the much smaller corpus it did not seem meaningful to go beyond the top 100.

Polysemous words	back, face, glass, top

Table 8: The top 100 words in 'LWL 543' grouped by informational content, following the process and categories from Salway and Baker [2020].

Turning from informational content to consider the extent to which BMSat descriptions interpret/evaluate visual content, rather than describe it, Salway and Baker [2020] analysed the 100 most frequent words ending in -ly. We noted many examples of words that were likely used to interpret something about somebody's actions in terms of their mental state, e.g. 'angrily', 'delightedly', 'derisively', 'contemptuously', such as to give a sense of a story playing out (see Table 6 in Salway and Baker [2020]). We also noted words used to make an evaluative judgement such as about somebody's appearance, e.g. 'fashionably' and 'grotesquely'. For comparison, a list of -ly words occurring three or more times in the LWL 543 descriptions is given in Figure 4. Though the small number of examples means we cannot draw firm conclusions, it does seem that whilst the LWL 543 descriptions show signs of evaluating appearances (see the bolded words), there is only one example of a word that could be used to interpret an action such as to give a sense of a story playing out, i.e. 'intently'. Whether this is due to a different curatorial style, or simply the contents of the prints being described, would require a different line of investigation.

fashionably, elderly, probably, only, **elegantly**, heavily, partially, similarly, **ugly**, belly, elaborately, equally, family, fly, possibly, presumably, assembly, enormously, highly, intently, mostly, **portly**, slightly

Figure 4: Words ending in -ly and occurring more than three times in 'LWL 543'.

By basing our comparison on existing results for one corpus there is perhaps a danger of us observing similarities and missing differences. To actively seek differences, we conducted a keyness analysis to identify words that occur relatively more often in one corpus than the other, with statistical significance. Tables 9a and 9b show the keywords in BMSat and LWL 543 respectively[50].

Various differences stand out in the keyword lists. First, proper names (Mrs, Wellington, Fox, Napoleon, Duke, John) are keywords in BMSat relative to LWL 543, whilst unnamed people (woman, man) and everyday scenes (dog, chair, table, wall) are keywords in LWL 543 relative to BMSat. This

[50] The keyness analysis used 'Ratio of relative frequencies' for effect size, and 'Log-likelihood' for statistical significance. The selected keywords had an effect size >=2, and statistical significance $p < 0.05$. A frequency threshold was also applied so the BMSat results show only words occurring 500 or more times in BMSat, and the LWL results show words occurring 50 or more times in LWL. We acknowledge that these are all rather arbitrary choices and could be relaxed to show more keywords and hence potentially more difference.

may be explained by the LWL 543 selection containing a higher proportion of records for social satires than BMSat, wherein descriptions of political satires - featuring men like Charles James Fox, Napoleon Bonaparte and the Duke of Wellington - account for roughly three-fifths of all descriptions. Second, when compared with BMSat, the LWL 543 descriptions do not contain references such as 'See BMSat' 'BM No'. This supports that observation in Section 3.1 that Lewis Walpole Library records either expand on these abbreviations or move such abbreviations to dedicated data fields.[51] The BMSat list also contains words that we might consider to be part of a general "language of description" and it might seem surprising that these are not used so much in the LWL 543. However, the presence of some may be due to the content of prints, i.e. more prints depicting speech (answers, saying and says), and another due to spelling variation (centre). The remaining BMSat keywords (inscribed, cocked, design, round, extreme, profile, Behind, He, which, wears) may point to minor stylistic differences or may simply be due to the small amount of text in the LWL 543 descriptions.

No, BMSat, answers, &c, Mrs, BM, Satires, Wellington, centre, See, Fox, inscribed, see, cocked, saying, Napoleon, design, Duke, says, John, round, extreme, profile, Behind, He, which, wears

Table 9a: Keywords in BMSat, compared with 'LWL 543'.

image, young, woman, another, while, dog, chair, her, man, table, wall, as

Table 9b: Keywords in 'LWL 543', compared with BMSat.

In summary, for all the points of comparison relating to informational content - i.e. what is described, and what vocabulary is used - the results suggest that overall the BMSat and LWL 543 descriptions are very similar. The comparison in terms of degree of description/interpretation/evaluation suggested some difference, but was inconclusive given the small number of examples available in the LWL 543. Thus, we may tentatively conclude that a common language of cataloguing is used for descriptions in both catalogues, even for those records where there cannot have been direct transmission of records from one to the other. This leaves open the possibility that the LWL 543 descriptions were influenced by the transmission of curatorial voice from BMSat: however, to make any claims about direct influence from one catalogue to another - as distinct from two catalogues sharing an influence from elsewhere - would require a different kind of investigation, i.e. a parallel historical investigation into the circumstances in which catalogue records were produced.

[51] Note that given the importance of the *Catalogue of Political and Personal Satires* even records not based on it often still refer to it.

4. Discussion

We have shown how a mixed methods approach can be used to systematically identify those catalogue records that are likely to be the products of transmission from an earlier catalogue and to analyse them such that models of transmission can be elaborated. Whilst we stopped short of asserting that these models form the beginnings of a conceptual framework for thinking about the different ways in which catalogues are shaped and constructed over time, we expect that the findings from our case study will resonate with both cataloguing professionals and researchers who rely on catalogue records, especially those who work closely with visual materials, domains where cataloguers like Mary Dorothy George are canonical influences.

In the case of George and the hypothesised transmission from the *Catalogue of Political and Personal Satires* to catalogue records for printed images held by the Lewis Walpole Library, we found - *per* Baxandall - that influence was varied. Transmission through time, across space, and between mediums took many forms. There were cases where catalogue descriptions were copied, quoted, and cited, and there were cases where transmission was stylistic. Staff at the Lewis Walpole Library drew on and engaged with the *Catalogue of Political and Personal Satires*, and latterly their derivations on British Museum Collections Online, as a source of expertise. Individual cataloguers reacted to the normative assumptions of George’s early- to mid -twentieth century British worldview, developed her thinking for modern audiences, and tackled her prejudices. When their objects demanded it, the Library staff differentiated their records from those in the *Catalogue of Political and Personal Satires*. And records made for a paper catalogue were remodelled for digital cataloguing infrastructures.

Whilst we can speculate, often with some confidence, on the reasons why a given pattern of transmission occurred, the methods we present are less able to get at purpose and process, at why and how transmission occurred the way it did. For example, we may know that descriptions written by the Lewis Walpole Library are linguistically similar to those in BMSat (Section 3.2), but we do not know why cataloguers produced each George-like descriptions, how their access to the *Catalogue of Political and Personal Satires* and latterly their derivations on British Museum Collections Online shaped the production of catalogue records, or the extent to which George-like precedents set early in the Library’s life shaped later behaviour. Equally, we know that, for example, the Lewis Walpole Library record for *Bath Races* tackles and subverts the linguistic choices in George’s entry, but only by working in close proximity to institutional histories and processes are we able to know why the decision was made to produce this description and how and when that occurred. It is a cliche to say that ‘more research is needed’, but when historically specific labours and practices are obscured both by the presentation of the catalogue as an always-already present unifying entity, and when

collections are federated for access, into datasets, or as machine readable endpoints, knowing our data is vital; we may not be able to fix or eliminate the biases in a given catalogue or dataset, but we can and should enquire into their ‘deeper structural issues, historical antecedents, and power asymmetries’ [Birhane, 2021], of which - in the domain of cultural heritage - transmission of legacy voices across time and outside their historical place is one.

Our methods can advance these imperatives in that they provide a relatively collection agnostic approach to characterising types of transmission between collection catalogues at scale. We were able to produce a series of findings that can form the basis for a variety of actions to be taken forward depending on institutional context and priorities: further archival research into cataloguing processes to correlate and enrich data driven findings; oral histories with staff to create data on why choices were made; revisions to records to highlight the source of their information; rationalisation between records to even out parsing during transmission; allocation of staff time towards plans to repair records, document common types of transmission, and/or write business cases for recataloguing. We note here that the age or size of an institution may create variability in how far our methods can be usefully redeployed or enable opportunities not presented by our case study: for example, departments that are (or once were) responsible for managing their own catalogues may have retained rich data on the editing and versioning of their catalogue records; alternatively, catalogues may have been produced in distinct and known phases or batches to support the use of a new cataloguing infrastructures (e.g. card catalogue, early database, online collections portal), such that sub-divisions of records can be made prior to analysis. Nevertheless, for collecting institutions of all ages and sizes, maintaining a catalogue is central to their operation. Computational methods for characterising the multi-faceted influences that produced their catalogue can, we hope, usefully support that operation and in turn enrich understanding of the collections those catalogues describe.

References

[Agostinho et al., 2019] Agostinho, D., D’Ignazio, C., Ring, A., Thylstrup, N.B., Veel, K., “Uncertain Archives: Approaching the Unknowns, Errors, and Vulnerabilities of Big Data through Cultural Theories of the Archive”, Surveillance & Society, 17 (2019).

[Annie Burr Lewis papers, 1849-1960] Annie Burr Lewis papers, Lewis Walpole Library, LWL MSS 21 (1849-1960)

[Baker, 2017] Baker, J., The Business of Satirical Prints in late-Georgian England. Palgrave Macmillan, London (2017]

[Baker and Salway, 2020] Baker, J., Salway, A., “Curatorial labour, voice and legacy: Mary Dorothy George and the Catalogue of Political and Personal Satires, 1930–54”, Historical Research, 93 (2020).

[Baxandall, 1985] Baxandall, M., Patterns of intention: on the historical explanation of pictures. Yale University Press, New Haven (1985).

[Bender et al., 2021] Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S., “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?”, FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (2021).

[Birhane, 2021] Birhane, A., “Algorithmic injustice: a relational ethics approach”, Patterns, 2 (2021).

[Bowker and Star, 2000] Bowker, G.C., Star, S.L., Sorting things out: classification and its consequences. MIT Press, Cambridge, Mass (2000).

[Cox, 2021] Cox, A.M., Research report: The impact of AI, machine learning, automation and robotics on the information profession. CILIP (2021)

[Donald, 1996] Donald, D., The Age of Caricature: Satirical Prints in the Reign of George III. Published for the Paul Mellon Centre for Studies in British Art by Yale University Press, New Haven (1996).

[Duncan 1995] Duncan, C., Civilizing rituals: inside public art museums. Routledge, London (1995)

[Eve 2019] Eve, M.P., Close Reading with Computers: Textual Scholarship, Computational Formalism, and David Mitchell’s Cloud Atlas. Stanford University Press, Stanford (2019).

[Gatrell, 2006] Gatrell, V.A.C., City of laughter: sex and satire in eighteenth-century London. Atlantic Books, London (2006).

[Gebru et al, 2020] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., Daumé III, H., Crawford, K., “Datasheets for Datasets”, arXiv:1803.09010 (2020).

[Griffiths, 2010] Griffiths, A., “Collections Online: The Experience of the British Museum”, Master Drawings, 48 (2010).

[Griffiths, 2016] Griffiths, A., The print before photography: an introduction to European printmaking, 1550-1820. The British Museum Press, London (2016).

[Hill, 2016] Hill, K., Women and museums 1850-1914: modernity and the gendering of knowledge. Manchester University Press, Manchester (2016).

[Johnson, 1990] Johnson, W.M., Art History: Its Use and Abuse. University of Toronto Press, Toronto (1990)

[Kingdon, 2019] Kingdon, Z., Ethnographic collecting and African agency in early colonial West Africa: a study of trans-imperial cultural flows, Contextualizing art markets. Bloomsbury Visual Arts, London (2019).

[Lewis 1969] Lewis, W.S., One Man’s Education. A. A. Knopf, New York (1968)

[Liboiron, 2021] Liboiron, M., Pollution is colonialism. Duke University Press, Durham, N.C. (2021).

[Noble, 2018] Noble, S.U., Algorithms of oppression: how search engines reinforce racism. New York University Press, New York (2018).

[Padilla, 2019] Padilla, T., Responsible Operations: Data Science, Machine Learning, and AI in Libraries. OCLC Research, Dublin, OH. (2019).

[Perez, 2003] Perez, E., “Queering the Borderlands: The Challenges of Excavating the Invisible and Unheard”, Frontiers: A Journal of Women Studies, 24 (2003).

[Pringle, 2020] Pringle, E., Provisional Semantics: Addressing the challenges of representing multiple perspectives within an evolving digitised national collection (Interim Report), Towards a National Collection (2020).

[Salway and Baker, 2021] Salway, A., Baker, J., Analysis of transmission from BMSat to LWL. Zenodo (2021) https://doi.org/10.5281/zenodo.5148228

[Salway and Baker, 2020] Salway, A., Baker, J., “Investigating Curatorial Voice with Corpus Linguistic Techniques: the case of Dorothy George and applications in museological practice”, Museum and Society, 18 (2020).

[Stijnman, 2012] Stijnman, A., Engraving and etching, 1400-2000: a history of the development of manual intaglio printmaking processes. Archetype Publications ; in association with HES and DE GRAAF Publishers, London; Houten, Netherlands (2012).

[Sussex Humanities Lab Carbon Use and Environmental Impact Working Group et al., 2020] Sussex Humanities Lab Carbon Use and Environmental Impact Working Group, Walton, J., Eldridge, A., Baker, J., Banks, D., Hitchcock, T., The Sussex Humanities Lab Environmental Strategy. Zenodo (2020). https://doi.org/10.5281/zenodo.3776161

[Sutherland, 2017] Sutherland, T., “Archival Amnesty: In Search of Black American Transitional and Restorative Justice”, Journal of Critical Library and Information Studies, 1 (2017).

[Sutherland and Purcell, 2021] Sutherland, T., “A Weapon and a Tool: Decolonizing Description and Embracing Redescription as Liberatory Archival Praxis”, The International Journal of Information, Diversity, & Inclusion, 5 (2021).

[Thylstrup, 2019] Thylstrup, N.B., The politics of mass digitization. MIT Press, Cambridge, Massachusetts (2019).

[Turner, 2020] Turner, H., Cataloguing culture: legacies of colonialism in museum documentation. UBC Press, Vancouver (2020).

[Vermeulen and Carby, 2014] Vermeulen, H.V., Carby, H.V., Prospects of Empire: Slavery and Ecology in Eighteenth-Century Atlantic Britain. Yale University Library (2014).

[Virgin, 1989] Virgin, P., The Church in an age of negligence: ecclesiastical structure and problems of Church reform, 1700-1840. James Clarke, Cambridge (1989).

[Yakel, 2003] Yakel, E., “Archival representation”, Archival Science, 3 (2003).

[Yale University, 1957]. Resolution for Annie Lewis’s appointment as Curator of Prints at the Lewis Walpole Library, Lewis Walpole Library, LWL MSS 21 Series I; Box: 12, Folder: 12 (1957).