PART 3

Archival Limits

CHAPTER 2

Searching for Dr. Johnson: The Digitisation of the Burney Newspaper Collection

Andrew Prescott

As you enter the Rare Books and Music Reading Room of the British Library, the bookshelves on your left-hand side are full of boxes of microfilm which few readers ever touch. These are the microfilms of the Burney Collection of newspapers which twenty five years ago were one of the most frequently used microfilm sets in the British Library. The microfilms are no longer much used because of the release in late 2007 by Gale Cengage Learning in partnership with the British Library of a searchable database of the Burney Newspapers which provides more convenient access to the collection both in the British Library and remotely. The boxes of microfilms of Burney Newspapers in the Rare Books Reading Room now seem like fossils, discarded relics of a superannuated information technology – the British Library collection guide to newspapers only refers to the digital version and does not mention the existence of the microfilms.[1] Yet our use of the digital resource is still profoundly shaped by the technology and limitations of the microfilm set, and whenever you enter the Rare Books Reading Room it is worth glancing at the microfilm as a reminder of the pitfalls inherent in using the digitised Burney Newspapers.

The Burney Collection is the largest single collection of early English newspapers, currently containing nearly a million pages in about 1,290 titles, and it was purchased by the British Museum in 1818 from the estate of Rev. Charles Burney (1757–1817), the son of the music historian Dr. Charles Burney and brother of the novelist Frances Burney. As a young man, fond of high living and carousing, Burney was sent down from Cambridge in disgrace after stealing and defacing over ninety books from the University Library. Refused entry to his father's house, Burney was determined to redeem himself and went to Aberdeen where he became a diligent student.[2] Burney became a successful

1 https://www.bl.uk/collection-guides/burney-collection (accessed 10 October 2016). I am grateful to Michael Alexander, John Goldfinch, Edmund King, Michael Lesk and Simon Tanner for their assistance in completing this essay. Responsibility for errors and misunderstandings is entirely mine.

2 Lars Troide, 'Burney, Charles (1757–1817)', in *The Oxford Dictionary of National Biography* (Oxford: Oxford University Press, 2004), online edition: http://www.oxforddnb.com/view/

schoolmaster and distinguished classical scholar. He was reinstated at Cambridge and awarded a degree by royal mandate, removing the taint of his crime and allowing him to be ordained as a clergyman. Burney resumed his book collecting, using more conventional methods, and was made a Fellow of the Royal Society.

The material purchased by the British Museum in 1818 from Burney's executors comprised five important collections.[3] The first was over 500 manuscripts chiefly of classical texts and Christian texts in Greek and Latin. The second group contained over 13,000 printed volumes including rare early editions of Aeschylus, Sophocles and Euripides. The Museum also purchased over 7,000 prints collected by Burney. The fourth collection acquired by the Museum consisted of nearly 400 volumes of newspaper cuttings, playbills, notes and prints which Burney had compiled towards a projected history of the theatre. The final component of the Burney purchase were the newspapers which he had begun to collect on his return from Aberdeen in 1781 by gathering old papers from Gregg's coffeehouse in York Street, Covent Garden, which was run by his maiden aunts.[4] The bald description by the Museum of the newspaper collection at the time of its purchase as "a collection of early newspapers, filling 700 volumes, more ample than any in existence" has been taken to suggest that the newspapers were considered less important than the classical manuscripts and books,[5] but the price paid for the newspapers, 1,000 guineas, compares favourably with the valuation of the manuscripts which was between £2,500 and £3,000,[6] indicating that the importance of Burney's newspaper collection was recognised at an early date.

The curatorial history of the Burney Collection is complex.[7] Burney did not file his newspapers by title but bound them in a chronological sequence, arranging them day by day so that all the papers collected by him for a given date are together. This was a convenient way of following the reporting of particular

---

article/4079 (accessed 26 October 2016); R.S. Walker, 'Charles Burney's Theft of Books at Cambridge', *Transactions of the Cambridge Bibliographical Society*, 3 (1959–1963), pp. 313–326.

3 P.R. Harris, *A History of the British Museum Library, 1753–1973* (London: The British Library, 1998), pp. 37–38.

4 Lars Troide and Stewart Cooke (eds.), *The Early Journals and Letters of Fanny Burney, Vol. 3* (Oxford: Clarendon Press, 1994), pp. 457–459.

5 Arundell Esdaile, *The British Museum Library: A Short History and Survey* (London: Allen and Unwin, 1946), p. 208.

6 Harris, *History of the British Museum Library*, p. 38.

7 It is helpfully summarised in Moira Goff's note on 'The Burney Newspapers at the British Library' in the 'About' section of the online version of the Burney Newspapers: http://find.galegroup.com/bncn/bbcn_about.htm.

events without having to move from one volume to another, but annoying for anyone trying to establish which issues of a particular title survive or for a reader interested in one particular title. This chronological arrangement became more cumbersome as the number of newspapers increased in the later eighteenth century, when many volumes were needed for a single year.

Following the acquisition of the Burney Newspapers by the British Museum, British Museum curators inserted seventeenth- and eighteenth-century newspapers acquired by the Museum into the Burney Collection, a practice which continued until the 1970s. As a result, perhaps two-thirds of the Burney Newspapers were not collected by Burney himself but added to the collection after its acquisition by the Museum.[8] The first such addition seems to have been newspapers from the collection of Sir Hans Sloane, the founder of the British Museum.[9] The most recent addition to the Burney Newspapers was in 1972 when a bequest by the architect and philatelist Sydney R. Turner filled gaps in runs of titles and added new titles such as the *Corn Cutter's Journal*, a pro-government newspaper heavily subsidised by Walpole.[10] In many ways, the name Burney Newspapers is more of a homage to the founder of the collection than an accurate indicator of their provenance.

In inserting these additional newspapers into the Burney Collection, the British Museum curators tended to prefer single title volumes, as opposed to Burney's day-by-day arrangement. Consequently, it can be difficult to anticipate where numbers of a particular newspaper may have been placed. For example, Volume 3 for 1758 contains not only the *Universal Chronicle* from 8 April to 30 December 1758, but also numbers of the same title for 1759 and 1760. The complexities caused by the different ways in which newspapers were incorporated into the collection over the years are illustrated by the first volume for 1717. The first volume for 1717 starts with a complete run of the *Daily Courant* for the year, followed by the complete 1717 run of the *London Gazette* and then the complete annual run of the *Weekly Journal or British Gazetteer*. This 1717 volume then continues with one number of the *Original Weekly Journal* for 23 February 1717; numbers of the *St. James's Evening Post* from January, February, November and December; individual numbers of the *Flying Post* and *London*

8 The extent of the additions to the collection is evident from annotations to Burney's handwritten collection of his collection. The estimate of two thirds of the collection being later additions is by Moira Goff, 'Burney Newspapers'.

9 Alison Walker, 'Lost in Plain Sight: Rediscovering the Library of Sir Hans Sloane' in Flavia Bruni and Andrew Pettegree (eds.), *Lost Books: Reconstructing the Print World of Pre-Industrial Europe* (Leiden: Brill, 2016), p. 410.

10 Goff, 'Burney Newspapers'.

Post from November 1717; odd numbers of the *Weekly Journal* or *Saturday's Post* from August to December 1717; and finally one stray number of the *Evening Post* for 28 December 1717.

When the Newspaper Library was established by the British Museum at Colindale in 1905, all post-1801 newspapers were transferred there and assimilated into a general title sequence, including those from the Burney Collection, which was thus split up. However, this division was not cleanly made and the main Burney Collection still retains a few post-1801 London titles – for example, the Burney run of *Lloyd's Evening Post* continues to 18 April 1804. All pre-1801 provincial newspapers were also transferred to Colindale at that time. Although the Burney Collection contains some provincial titles, the vast majority of the titles in the Burney Newspapers came from London and it appears that the British Museum curators regarded it as essentially a London collection, so there do not appear to have been any transfers of pre-1801 provincial newspapers from the Burney Collection.[11]

Until the digitisation project, the primary means of locating material in the Burney Newspapers was by means of Burney's own handwritten catalogue reflecting the chronological arrangement, with later annotations reflecting subsequent additions and changes. Burney's original manuscript was used in the Reading Room until the 1970s, when a photocopy was provided which is still in use in the British Library's Rare Books Reading Room for anyone wishing to consult the microfilms. A project in the 1940s to prepare a new catalogue fizzled out, but a card index was made at that time of titles in the handwritten Burney catalogue. This index was subsequently lost but photocopies of the cards survive. In 1970, the bibliographers John Joliffe and Julian Roberts (afterwards both to move to the Bodleian Library) developed a method to produce an issue-by-issue listing of the collection on computer (to be input using punched tape). A pilot project was carried out to catalogue 200 volumes in the collection using this method, but this visionary proposal was not continued, and the data accumulated has also been lost. The continuing lack of a comprehensive title index and the chronological arrangement posed a direct threat to the preservation of the Burney Newspapers themselves, since readers wishing to trace particular titles had to wade through the chronological volumes, considerably increasing wear and tear on the newspapers.

The origins of microphotography go back to the nineteenth century, but it began to appear in commercial use in the 1920s and the Library of Congress started microfilming material in British libraries. The British Museum became interested in the extent to which microfilming could reduce wear and tear on

11 I am grateful to John Goldfinch for advice on this point.

its collections and in 1935 Eugene Power, who afterwards established University Microfilms International which pioneered the large-scale microfilm publication of primary materials, helped set up a programme at the British Museum for the microfilming of rare books. After the Second World War, the Museum recognised the potential of microfilm for dealing with the problems of preserving and providing access to large volumes of fragile newspapers, and Power was again involved in setting up, with the assistance of the Rockefeller Foundation, a studio to microfilm newspapers at Colindale.[12]

Joliffe and Roberts during their time in the British Museum were alarmed by the deteriorating condition of the Burney Newspapers and in 1971 Roberts suggested that the entire collection should be microfilmed. The American microfilming company Research Publications (later Research Publications International and Primary Source Media, now incorporated into the Gale Group) expressed interest in microfilming the collection in 1972, but it was not until 1977 that filming finally began. It was initially hoped that the microfilm would rearrange the collection into title order rather than reproduce the confusing arrangement of the existing volumes. It was originally intended to do this by splicing the microfilm to rearrange the titles, but this proved impracticable. Instead, it was decided to film title by title, so that each volume went under the microfilm cameras as many times as necessary to assemble all the issues of a particular title. This process created a number of problems. For example, camera operators had to decide what page went with which issue, and a number of the films have odd pages at the end. John Goldfinch has observed that "We don't actually know that all the pages in a given Burney volume made it onto the film".

These films were supplied to Research Publications who in 1979 produced *Early English Newspapers: 1622–1820*, which included both material from the Burney Collection and another large collection of early English newspapers, the newspapers collected by John Nichols in the Hope collection in the Bodleian Library. The Research Publications microfilm set was in title order. However, the title listing published as a guide to *Early English Newspapers* did not distinguish material taken from the Burney and Nichols collection. Moreover, it appears that where the Burney run of a particular title was sporadic, Research Publications did not include it in the hope that fuller holdings of the title might be identified elsewhere. Consequently, the Research Publications set did not provide comprehensive coverage of the Burney Collection, omitting many titles for which Burney only had partial holdings.

12 Harris, *History of the British Museum Library*, pp. 530–531, 601; S. John Teague, *Microform, Video and Electronic Media Librarianship* (London: Butterworths, 1985), pp. 8–9.

By 1981 or so, a complete microfilm of the Burney Collection was available to readers in the British Library's reading rooms. This at least helped ensure that the original volumes would suffer much less wear and tear and, as John Goldfinch observed in 2003, "microfilm has proved its worth in supporting the preservation of the collection without impeding access to the information within it".[13] Nevertheless, it was a tiresome process, sitting beneath the celebrated dome of the Round Reading Room of the British Museum, trying to navigate one's way around the endless reels of microfilm of the Burney Collection. For Ashley Marshall and Robert Hume, the chief benefit of the Gale Burney database was the way it liberated them from the "grim business" of skimming microfilms.[14] According to Marshall and Hume, the "difficult and inefficient" process of searching microfilm meant that "Even those scholars who have spent thousands of hours with the films have mostly stuck to scanning well-known papers for paragraphs on particular subjects".[15] Not only was searching microfilm time-consuming and inconvenient, but only one reader could have access to a single reel of microfilm at any time. This was a particular problem in the case of the Burney Newspapers where the chronological arrangement meant that a reader wanting to read a particular title would require a number of microfilm reels. Above all, heavy use of the single set of microfilms in the reading room meant that the quality of the microfilms rapidly deteriorated as they became scratched and worn.

In 1992, at the behest of Professor Robin Alston, the Editor-in-Chief of the Eighteenth Century Short Title Catalogue and bibliographical consultant to the British Library, the British Library purchased a Mekel M400XL microfilm digitiser, and an immediate priority was to see how far this could improve the way in which the Library provided access to its microfilms, by for example allowing multiple reader access to microfilms and above all dealing with the problems caused by the deterioration of single microfilm sets as a result of heavy reading room use. The Library set up a project called 'The Digitisation of Aging Microfilm' (with the uninspiring acronym DAMP) and it is from this microfilm project that the Burney digital resource derives. The DAMP project

13 John Goldfinch, 'The Burney Collection of Newspapers: will digitisation do the trick?' in John Webster (ed.), *'Parallel Lives': digital and analog options for access and preservation: Papers given at the joint conference of the National Preservation Office and King's College London held 10 November 2003 at the British Library* (London: The National Preservation Office, 2004), pp. 49–60 (54).

14 Ashley Marshall and Robert Hume, 'The Joys, Possibilities, and Perils of the British Library's Digital Burney Newspapers Collection', *Papers of the Bibliographical Society of America,* 104 (2010), pp. 5–52 (6).

15 Marshall and Hume, 'Joys, Possibilities, and Perils', pp. 5–6.

formed part of the British Library's first major digitisation and networking programme from 1993 onwards which was called Initiatives for Access.[16]

The original aim of the Burney Newspapers digitisation project was thus not to produce a searchable text at all. It was an experimental project to improve the way in which the Library made its microfilm surrogates available. There was no assumption that digital images would replace microfilm, since microfilm, which tests showed would survive one thousand years in the right conditions, was considered a more stable preservation medium. Details of the early stages of the DAMP project can be found in two articles by the project's manager, Hazel Podmore.[17] She emphasised how the first stages of the project were meant to test the capabilities of the Mekel scanner. She noted that the microfilm was not the best quality the Library has ever produced, both because the microfilm was quite old and because the original newspapers were in such poor condition. However, from the point of view of Podmore and her team, this was ideal since it provided a tough test for the Mekel scanner. Another problem is that it was quickly found that for best results it was necessary to use the master microfilm, but the DAMP team were concerned about the potential for damage to the master microfilms during the scanning process and worked from the reading room set.

The DAMP team's main priority was to establish an efficient workflow for the scanning of the microfilm. They focussed initially on newspapers from the French Revolution period. By the time the project finished in 1996 over 21 gigabytes of images had been produced at a work rate of approximately 6,000 frames per month, suggesting that it would take about eighteen months to scan the 650,000 microfilm frames of the entire Burney Collection. However, while the team had successfully mastered the production workflow, the best way of enabling readers to navigate the thousands of images produced by the system was not immediately clear. Simply dumping hundreds of thousands of digital images on a server was not a practicable way of offering a reading room service, while the production of a special programme for access to the images would have been too expensive. Experiments were made with Optical Character Recognition (OCR) packages to convert the images to machine readable text, but at that time no OCR packages were capable of recognising eighteenth-century type.

---

16 Hazel Podmore, 'Microfilm Revolutionised', *Initiatives for Access News*, 1 (1994), p. 8; Hazel Podmore, 'The Digitisation of Microfilm' in Leona Carpenter, Andrew Prescott and Simon Shaw (eds.), *Towards the Digital Library: the British Library's Initiatives for Access Programme* (London: The British Library, 1998), pp. 68–72.

17 Ibid.

Among the other projects undertaken within the 'Initiatives for Access' programme were experiments with Excalibur PixTex/EFS, a Unix package which offered fuzzy searching, using substantial computing power to recognise shape of letters in images on the fly rather than relying on OCR.[18] Excalibur was used successfully in offering search access to images of a catalogue of medieval seals, but its proprietary nature and use of a Unix platform not generally supported in the Library meant that Excalibur remained only an experimental demonstration of the potential of fuzzy searching. Again, experiments were made with the searching of Burney images using Excalibur but once again these were unsuccessful.

For the time being, the images of the Burney newspapers generated by the DAMP project languished. In a presentation to the Newspapers Section of IFLA in the late 1990s, Graham Jefcoate, then Head of Early Printed Collections at the British Library, declared that it was "melancholy to report that little progress has been made with the digitisation of the Burney newspapers since 1996", but insisted that "The digitisation of the Burney collection must be a priority for us".[19] Jefcoate reported that further tests had been made of scanning the Burney microfilms, this time using a Mekel M500 greyscale production scanner. It was hoped that this slightly more sophisticated scanning would improve image quality and facilitate OCR. Jefcoate cited a technical report which stated that the OCR had achieved an accuracy level of 58% but recommended that manual indexing was advisable.[20]

On the basis of this report and on advice from the National Science Foundation, in 2001 the National Science Foundation in the United States made a grant to the British Library's partner, the Center for Bibliographical Studies at the University of California, Riverside. The grant proposal again emphasised the difficulties of accessing the Burney Collection through a single microfilm set. It also suggested that advances in OCR would now make it possible to produce a searchable version of the Burney Newspapers and "extend the wonders of computer-based text searching to the corpus of texts that form the foundation

18 'Finding the Fuzzy Matches', *Initiatives for Access News*, 1 (May 1994), p. 2; 'Digital Data Retrieval: Testing Excalibur', *Initiatives for Access News*, 2 (December 1994), pp. 6–7; Andrew Prescott and Malcolm Pratt, 'Image – the Future of Text?' in Carpenter, Prescott, Shaw (eds.), *Towards the Digital Library*, pp. 178–189.

19 Graham Jefcoate, 'The Digitisation of the Burney Collection of Early Newspapers at the British Library', in Hartmut Walravens and Edmund King (eds.), *Newspapers in International Librarianship: Papers Presented by the Newspapers Section at IFLA General Conferences* (Munich: K.G. Saur, 2003), pp. 185, 187.

20 Jefcoate, 'Digitisation of the Burney Collection', p. 186.

of the modern world".[21] This work was not without its difficulties. No count had been kept of the number of images when the microfilming had been done; the only count that had been kept was of feet of film used. While most papers had been filmed one page per frame, others were filmed one opening per frame, and the orientation was variable. As a result, the cost of the scanning was 50% higher than expected, and the OCR work had to be undertaken later, in partnership with Gale Cengage. It was this work in both London and California, taken forward by the determination and pertinacity of curators such as Moira Goff and John Goldfinch and with the enthusiastic support and assistance of Professor Henry Snyder of the University of California and Michael Lesk of the National Science Foundation, building on the pioneering work of Hazel Podmore and her team in the DAMP project, which eventually facilitated the partnership with Gale Cengage and the final release of the digital Burney Newspapers in 2007.

The fundamental driving force of the digitisation of the Burney Newspapers was then not so much the wish to produce a searchable text but rather to find a way of delivering surrogate access to the fragile newspapers that was more convenient and flexible than microfilm. This was recognised by Marshall and Hume in their article on the "Joys, Possibilities and Perils of the British Library's Digital Burney Newspapers Collection". For Marshall and Hume, the joys of this resource consisted chiefly in the way it released them from the drudgery of winding their way through reels of microfilm and in its ability to facilitate off-site access to the collection to a wide audience. Marshall and Hume expressed surprise that the microfilm images of the blotchy eighteenth-century newspapers had cleaned up so well and that the OCR worked as well as it did, but nevertheless noted some serious problems, for example in a search for the German lutenist Sigismund Weiss:

> A search of 'Weys' or 'Wey's' in digital Burney produces fourteen hits (three of them duplicates), which means that in eight cases the search engine failed to spot the target – 10 February, 8, 15, and 22 April, 12, 17, and 18 June, and 1 July. In some cases broken type, creased paper, or bleed through may be responsible, but not in all cases. The appropriate response seems to be Gulp! The error rate is discouraging, but the omission rate is horrifying.[22]

21 National Science Foundation Award Abstract #0219461, available at https://www.nsf.gov/awardsearch/.

22 Marshall and Hume, 'Joys, Possibilities, and Perils', p. 42.

Despite issues such as these, Marshall and Hume remained enthusiastic about the potential of the Digital Burney as an alternative to microfilm, describing it as “pretty fabulous” and “leading to tectonic shifts in the way we do our research and teaching”,[23] largely because the collection had previously been virtually impenetrable when only available on microfilm.

Marshall and Hume suggested that the initial release of the digital Burney Newspapers would have benefitted from greater involvement of subject specialists in the design of the search interface.[24] This may be the case, but it assumes that the project to create this digital resource was focused and self-contained, whereas as has been seen, the evolution of the online Burney Newspapers was a long and complex process involving various partners at different stages and was driven in its initial stages by the Library’s concern to improve management of its microfilm surrogates. Project management textbooks stipulate that digitisation should not be undertaken by libraries in response to such institutional needs as accommodation pressures,[25] but in fact many of the largest digitisation projects are precisely a response to such extraneous issues as accommodation and reader service requirements. For example, the British Library’s need to vacate the Newspaper Library at Colindale was one of the primary drivers in the digitisation of newspapers held there.[26] As a result, digital versions of pre-1801 provincial newspapers in the British Library are generally not to be found in the Burney package but in the *British Newspaper Archive*, a partnership of the British Library with the family history company, findmypast. The *British Newspaper Archive* is available free of charge in the British Library’s reading rooms, but otherwise requires an expensive personal subscription – there is no adequate arrangement for institutional subscriptions.

One of the benefits of digitisation should be that collections are brought together and can be cross-checked, but the roots of many digitisation projects in microform publication and the involvement of commercial partners has meant that digitisation has frequently fragmented access. Given the involvement of separate commercial partners, it is unlikely that it will be possible to undertake remote cross-searching of the Burney Newspapers and the *British Newspaper Archive* in the near future. Moreover, parts of the Hope collection in

23 Marshall and Hume, ‘Joys, Possibilities, and Perils’, pp. 51–52.

24 Marshall and Hume, ‘Joys, Possibilities, and Perils’, p. 50.

25 Lorna M. Hughes, *Digitizing Collections: Strategic Issues for the Information Manager* (London: Facet, 2004), p. 51.

26 See, for example, http://lukemckernan.com/2013/10/09/leaving-colindale/ (accessed 9 November 2016).

the Bodleian Library, another major newspaper collection containing material gathered by the printer John Nichols, have been separately digitised by Adam Matthew for its Eighteenth-Century Journals project which also includes newspapers from such repositories as the Harry Ransom Centre in Texas and Cambridge University Library. It is worth noting that the Eighteenth-Century Journals project offers far better search facilities than either the Burney Newspapers or the *British Newspaper Archive*, since it seems that either the package has been keyboarded or the OCR has been corrected. Although the texts in the Adam Matthews package are not completely accurate, they certainly achieve accuracy rates in excess of 95%.

While the initial concern of the British Library in digitising the Burney Newspaper microfilms was to make reading room access to this material easier, the need to rely on search to navigate the mass of digital images means that, for most users, the value of the resource depends on the quality of the OCR underpinning the search. In 2009, Simon Tanner, Trevor Muñoz and Pich Ros undertook the first detailed analysis of the quality of the Burney OCR and its impact on search results.[27] Modern OCR software is capable of achieving very high levels of accuracy in converting printed text. Commercially available packages such as Omnipage or Adobe Acrobat achieve text accuracy results in excess of 98%. This compares well with the benchmarks used for conversion projects involving double keying and correction, which generally stipulate sampled accuracy rates in excess of 99.5%. However, such high quality OCR can only be achieved with modern print and with high quality images. When OCR packages are used on older typefaces, their accuracy rapidly declines, no matter how assiduously they are trained. Moreover, OCR accuracy depends very much on the quality of the image, and older materials are often only available in quite poor images frequently derived from microfilm.

Tanner and his colleagues found that the OCR for the Burney newspapers offered character accuracy of 75.6% and word accuracy of 65%. This means that, given a notional newspaper page of 1,000 words with 5,000 characters, in each page of the converted Burney newspaper text, there are about 1,200 incorrect characters or say about 350 wrongly rendered words, depending on word length. This need not necessarily be a big problem if significant words are accurately rendered by the OCR. If all the inaccuracies occurred in insignificant words such as 'and' or 'the', but words like 'British Museum' were

27 Simon Tanner, Trevor Muñoz and Pich Ros, 'Measuring Mass Text Digitization Quality and Usefulness: Lessons Learned from Assessing the OCR Accuracy of the British Library's 19th Century Online Newspaper Archive', *Digital Library Quarterly*, 15: 7–8 (2009): http://www.dlib.org/dlib/july09/munoz/07munoz.html (accessed 9 November 2016).

correctly rendered, then it would be possible to achieve 90% or more hit rates in a search for 'British Museum'. However, Tanner and his colleagues found that this was not the case. Significant word accuracy for the Burney Collection is only 48.4%. In other words, a search of the Burney Collection has a less than evens chance of finding the desired word – indeed, the situation may be even worse than that. Tanner *et al.* point out that "At below 80% word accuracy, the search capacity will drop off steeply as the word accuracy drops. Thus, the BL's Burney Collection loses a lot of search capacity because the long 's' character reduces word accuracy to such a low point that searching can become very difficult".

The implication of the findings of Tanner and his team is that searchers would be lucky to find 20–30% of the references for any particular search term, which echoes the experience of Marshall and Hume in searching for Sigismund Weiss. The Burney Newspapers interface conceals the raw OCR from the user, but it is available via other routes such as the *Connected Histories* website, and one glance at this raw OCR quickly reveals the difficulty of accurate searching, as this example illustrates:

> It is frippofed that we have actually 40 Frigates at Sea for the ProteEaion of our Trade. r The St. Domingo Fleet is fafely arrived at La Rochelle vith a rich t Cargo. ltsArrival is the more agreeable to the Merchants, as there was not a Ship of tde whole Fleet infui ed." Extrolg of a Lette?-from Mr, Coxvwood, Afate of tThe Wfter, ViWlsaller, from Cork fir Ne-w-rork, dated P3qlon, March So. &quot.[28]

The potential problems of faulty OCR have since been further highlighted by Laura Mandell in her discussion of Gale Cengage's *Eighteenth Century Collections Online* (ECCO)[29] and by scholars such as Tim Hitchcock.[30] However, it is important to bear in mind that in the case of the Burney Newspapers the role of the OCR was not so much to facilitate search as to provide a means of conveniently moving around a mass of scanned microfilm images. The digital interface to the Burney Newspapers is in many ways a replacement for the microfilm reader.

---

28 T. Hitchcock, 'Confronting the Digital, or How Academic History Writing Lost the Plot', *Cultural and Social History*, 10 (2013), p. 13.

29 L. Mandell, 'Brave New World: A Look at 18thConnect', *Age of Johnson*, 21 (2012). Web: http://earlymodernonlinebib.files.wordpress.com/2012/10/mandell-fixed-final-oct-2012.pdf.

30 Hitchcock, 'Confronting the Digital', pp. 12–14.

Nevertheless, the deceptive simplicity of the search interface for the Burney Newspapers seduces scholars into assuming that hit rates for particular terms can reveal historical trends, regardless of the fact that many references will probably be missed. In a 2014 lecture at the London Guildhall, Linda Colley described the development of what she called the cult of Magna Carta.[31] Colley argued that the eighteenth and nineteenth centuries saw a substantial growth in the reputation and resonance of Magna Carta. To illustrate this point, Colley listed the number of hits returned by a search for 'Magna Carta' in what she described as an index of British newspapers, by which she apparently meant the digital collection of British Library Historic Newspapers produced by Gale Cengage (incorporating the Burney Collection). Colley states that there were 30 references to Magna Carta in the first half of the eighteenth century, 450 articles in the second half, 4,000 articles in the first half of the nineteenth century, and 13,500 articles in the second half. While this increase in references to Magna Carta might be partly explained by the growth in the number of newspapers, Colley suggests that it also reflects a greater national engagement with Magna Carta, stimulated by such factors as the reprinting of the charter in school text books. However, given the problems with the OCR in the Burney Newspapers, doubt must be felt as to whether the growth was as dramatic as she suggests – simply searching for 'Magna Charta' produces over 450 hits in the first half of the eighteenth century (when OCR errors are far more likely), suggesting that the changes are far less dramatic than suggested by Colley, and that the trend described by her might be a mirage, produced by poor OCR.

Perhaps the most extended and disturbing illustration of indiscriminate reliance on search is Peter de Bolla's 2013 book *The Architecture of Concepts* which uses very crude searching of ECCO to illustrate the development of the discourse of human rights in the eighteenth century. This book, which has been widely acclaimed, uses an approach very similar to that of Colley, simply searching ECCO for references to rights and looking for words that occur nearby. By these means, de Bolla seeks to demonstrate the development of the discourse of rights during the eighteenth century, arguing that figures like Thomas Paine had a more crucial influence on the emergence of ideas about rights than has hitherto been allowed. The issues for OCR in ECCO are very similar to those in the Burney Newspapers, but for de Bolla these OCR problems are a minor concern. He acknowledges that the extraction of date information from ECCO is "beset with problems", adding that "as is well-known the OCR

31 The lecture is available at: https://www.youtube.com/watch?v=cFTDUtK2a6Y (accessed 9 November 2016). Colley's figures are repeated in D. Hughes, 'A Brief History of Magna Carta', *House of Lords Library Note*, 1 (2015), pp. 1–27 (23).

software used by Gale, the publisher, compromises the reliability of the data extracted".[32] But as far as de Bolla is concerned this is a mere technical glitch which does not challenge the validity of the new kind of conceptual history he seeks to present in his book. He is blithely confident that when these annoying technical gremlins are resolved his conclusions will remain valid, declaring that "I doubt there will be significant changes to the profiles I have created for the concepts here studied, the revisions of precise numerical values will be unlikely to lead to different conclusions".[33]

Is the confidence of scholars like Colley and de Bolla that comparatively uncritical searching and counting within digital resources such as the Burney Newspapers can produce valid historical insights justified? Or are the problems with OCR such that we should be far more cautious in presenting the results of our searching of these packages, as Tanner's study suggests? One method of assessing the issues in using the Burney Collection is to compare the results of searches with checklists of newspaper articles compiled manually. In 1976, Helen McGuffie published a *Chronological Checklist of Samuel Johnson in the British Press 1749–84*.[34] McGuffie compiled her list manually, searching through thousands of pages and concentrating on collections in the Bodleian Library, the British Library, the National Library of Scotland and the British Library, checking each item listed personally. The checklist focuses on the London and Edinburgh press, and there was no systematic search of provincial publications. The listing runs to over 325 pages and confirms that Johnson's comment to Boswell in 1781 that "I believe there is hardly a day in which there is not something about me in the newspapers" was not far off the truth.[35] McGuffie admitted that "Even among the thousands of pages that I did turn over, there no doubt lurk unrecorded items that my eye did not catch". Nevertheless, McGuffie felt confident that "It is unlikely that many new or significant items will be found in the future".[36]

McGuffie's checklist thus provides an excellent yardstick with which to measure the performance of the Burney Newspapers search engine. I have made a comparison between the results of Burney searches for Dr. Johnson

---

32 Peter de Bolla, *The Architecture of Concepts: the Historical Formation of Human Rights* (New York: Fordham University Press, 2013), p. 8. For a detailed critique of de Bollas's methods, see further James Baker, *Interfaces between Us and Our Digital Resources*, https://cradledincaricature.com/category/research/ (accessed 11 November 2016).

33 de Bolla, *Architecture of Concepts*, p. 8.

34 Helen McGuffie, *Samuel Johnson in the British Press: A Chronological Checklist* (New York and London: Garland Publishing, 1976).

35 McGuffie, *Johnson in the British Press*, p. 5.

36 McGuffie, *Johnson in the British Press*, p. 5.

and McGuffie's checklist for four sample years: 1765, 1766, 1773 and 1782. These confirm that online searching of the Burney Newspapers misses the majority of target references but suggest that the problems may be less acute for certain types of material, such as adverts. It also seems that the performance of the Burney searching is significantly improved by the use of fuzzy searching. Moreover, McGuffie's listing reminds us that, even if it was possible to have OCR with an accuracy in excess of 99%, many key references to a figure such as Samuel Johnson will still not be retrieved by simple searches. This has implications for the way we document and report research into the contents of newspapers.

There are a number of important preliminary points to bear in mind before considering the results of these comparisons. First, McGuffie only reports news items – she ignores advertisements, whereas, as we will see, one of the valuable aspects of the online Burney is the ability to explore newspaper advertisements. Second, a glance at McGuffie quickly shows how, although the coverage of newspaper titles by the Burney Collection coverage is impressive, it is far from comprehensive. For the four sample years, McGuffie lists references to Johnson in 78 different publications. Of these, 36 (just under half) are in the Burney Collection. These are mostly London publications, but they do include some provincial and Scottish titles. The third major point to bear in mind when comparing the McGuffie checklist with the Burney newspapers is that McGuffie includes not only newspapers but also items from the periodical press, such as the *Annual Register*, *Gentleman's Magazine* and *Monthly Review*. One of the striking features of McGuffie's listing is the way in which she demonstrates the porosity between these more occasional publications and the weekly newspapers, with information about Johnson shared freely between the various publications.

From a library point of view, periodicals such as the *Gentleman's Magazine* or *Monthly Review* have always been treated differently to newspapers, but the case of Dr. Johnson illustrates that the distinction between the two has been overstated, at least in the eighteenth century. The difference in library approaches to monthly and weekly publications have been carried over into the digital sphere. Annual and monthly publications have tended to fall between the two stools of books and newspapers. Digital coverage of eighteenth-century periodical publications is patchy. A few annual and monthly publications have been included in ECCO, while others have been covered by book scanning programmes such as Google Books, but not consistently. Above all, metadata for scanned periodical publications is poor, whether on Google Books, the Internet Archive or in more sophisticated metadata presentations such as the HATHI Trust catalogue. McGuffie's checklist reveals starkly how the fragmentary and inconsistent coverage of periodical publications is a significant deficiency in digital coverage of the eighteenth century. By comparison with the extensive coverage of newspapers and the systematic coverage of book publications in ECCO, the patchy coverage of periodical publications is surprising. Pulling together existing scans of these periodical publications, improving metadata and facilitating search across these titles should be a priority for future activity.

In comparing the results of a basic search for 'Johnson' in the Burney Newspapers with McGuffie's checklist, one is tempted to share the reaction of Marshall and Hume at the omission rate: 'Gulp!'. For 1765, McGuffie's checklist itemises 122 references to news items about Dr. Johnson in newspapers and periodical publications. A basic search for 'Johnson' in the Burney newspapers for 1765 produces just eight hits under 'news'. Just five of these hits relate to Dr. Johnson. Two of these are advertisements; so a simple Burney search finds just 3 of the 122 items in McGuffie – a success rate of just 2.4%. The results for 1766 are even worse. McGuffie lists 59 news items relating to Johnson for 1766. Of these, 39 items are from titles in the Burney Collection. A basic search on Johnson in the Burney Collection for 1766 retrieves 3 news items, all of which are actually advertisements. Two of these are for the third edition of Johnson's *Dictionary*, but the search retrieves none of the news items listed by McGuffie for 1766 – the search is a complete failure.

For 1773, McGuffie lists 190 news items relating to Johnson. A basic Burney search on Johnson for 1773 retrieves just 16 news items, of which 10 relate to Dr. Johnson and are also found in McGuffie. So, for 1773 the success rate of a Burney search is slightly better than in 1765 and 1766, but still unimpressive: 5.2%. Finally, for 1782, McGuffie lists 195 news items relating to Dr. Johnson. A basic Burney search retrieves 18 news items for Johnson; of these just three relate to Dr. Johnson – most of what is retrieved are actually adverts. For 1782, then the success rate of a Burney search on Johnson is 1.5%. These are very crude initial results. Account needs to be taken of the fact that the Burney Collection only covers about a half of the titles used by McGuffie, and also of the fact that a basic 'Johnson' search does not cover some items listed by McGuffie, but even so it is evident that results of basic Burney searches are just as bad as Tanner's analysis indicates, with a success rate probably barely in excess of 5–6% and frequently much worse. Each of our basic Burney searches probably misses well over 90% of the information we are seeking.

The extremely poor quality of basic Burney Newspaper search results not only suggests that quantitative conclusions of the sort offered by Colley and de Bolla are very hazardous but it also raises questions about other methods derived from this data. For example, it appears that the linking of other data sets to Burney newspaper data in the *Connected Histories* resource (www.connectedhistories.org) relies on a similar search method to the basic Burney search, suggesting that the likelihood of establishing worthwhile linkages using the Burney data is extremely low. However, all is not complete doom and gloom, and there are some points to make about the Burney package which may suggest fruitful future lines of approach.

The first is that the success rate of the search apparently varies according to whether the search covers news items, advertisements and so on. The Burney package attempts to segment eighteenth-century newspapers into different sections such as news, business news and advertisements. Inevitably, this segmentation is not entirely successful, so that for example an advertisement will sometimes be treated as a news item. However, it seems that the Burney package is far more successful at searching for text in advertisements than it is at retrieving text in news items. It is also striking how, although a Burney search only retrieves a very low proportion of news items, those which it does find are often similar in format to advertisements, such as extracts in the *St. James's Chronicle* from the Johnson and Steevens edition of Shakespeare in 1773.[37] McGuffie does not include advertisements for Johnson's work, so we do not have a benchmark against which to measure the Burney search performance with advertisements. It is unlikely to be very high, but it does seem to be significantly better than with news items. A method which blended McGuffie's comprehensive checklist and the ability of the Burney newspapers to open up the wealth of information in the advertisements would potentially provide a valuable new resource for Johnson studies.

The other important point which the comparison of Burney searches with McGuffie's checklist reveals is the extent to which fuzzy searching improves matters. A search on 'johnson' for 1765 with a low fuzzy search setting produces 305 items under 'news', a sixty-fold increase. Moreover, there are no false hits – all of these 305 items refer to a 'johnson' of one sort or another, most of whom had been completely ignored by the basic search. Taking account of the fact that the Burney Collection only covers less than half of the titles in McGuffie, a low fuzzy search produces a success rate of about 50% – a dramatic improvement on the performance of the basic search. Moreover, what is particularly exciting is the way in which fuzzy searching allows us suddenly to start finding items which are not in McGuffie's checklist. For 1765, a fuzzy search identifies four items which are not in McGuffie, ranging from a copy of a Jacobite toast,

37 *St. James's Chronicle*, 5–7 October, 12–14 October, 16–18 November, 18–20 November, 4–7 December, 11–14 December, 18–21 December 1773. These are not included in McGuffie's checklist, presumably because they are simply extracts from the prefaces and do not provide any biographical information about Johnson.

allegedly found in a copy of Johnson's works,[38] to an attack on Johnson's edition of Shakespeare in a series called 'The Babler' in *Owen's Weekly Chronicle*.[39] Similarly, for 1773, fuzzy searching adds enormously to the material returned, creating a tenfold increase in the number of hits relating to Dr. Johnson. Once again, use of fuzzy searching enables entries not in McGuffie to be found, including for the later part of 1773 reports of the movements of Johnson and Boswell during their tour of the Highlands.[40]

Of course, as Tanner observes, the problems with the OCR are such that fuzzy searching cannot completely compensate for them. Fuzzy searching still only retrieves about half of the entries found manually by McGuffie, but nevertheless the improvement is such as to suggest that it would be worth using low fuzzy searches by default. Moreover, the hit rate of fuzzy searches may be better than 50%, since I have so far only experimented with simple word searches, and many of the items listed by McGuffie would not be found by simple word searches, even if the OCR was very accurate. Some of the references to Johnson are allusive, such as the attack on obscure writers published in the *London Chronicle* on 26 February 1765 which does not mention Johnson by name.[41] In other cases, Johnson's name has been obscured as in the references to 'Pensioner J——', 'Learned Pensionary J—n—n' or 'Dr J——'.[42] Many of the newspaper references to Johnson found by McGuffie are only evident because the piece parodies Johnson's literary style.[43] Other entries do not refer to Johnson by name but refer to works of his, such as the many attacks on *Taxation No Tyranny* in 1773. Such references are unlikely ever to be retrieved by simple searching unless you already know they are there.

38 *St. James's Chronicle or British Evening Post*, 19–22 January 1765.

39 *Owen's Weekly Chronicle and Westminster Journal*, 21 December 1765.

40 For example, the following reports of Johnson and Boswell in the Highlands are not noted by McGuffie: *General Evening Post*, 14–16 September 1773; *Lloyd's Evening Post*, 15 September 1773; *London Evening Post*, 16–18 September 1773; *Lloyd's Evening Post*, 3–5 November 1773; *Daily Advertiser*, 5 November 1773; *London Chronicle or Universal Evening Post*, 6–9 November 1773; *The Craftsman or Say's Weekly Journal*, 13 November 1773; *St. James's Chronicle or the British Evening Post*, 13–16 November 1773; *London Evening Post*, 16–18 November 1773.

41 McGuffie, *Johnson in the British Press*, p. 34.

42 For example, *Gazetteer*, 28 March 1769 (McGuffie, *Johnson in the British Press*, p. 60); *Middlesex Journal*, 6 February 1770 (McGuffie, *Johnson in the British Press*, p. 68); *London Evening Post*, 13 March 1770 (McGuffie, *Johnson in the British Press*, p. 72); *Caledonian Mercury*, 6 February 1775 (McGuffie, *Johnson in the British Press*, p. 143).

43 For example, *Public Advertiser*, 21 May 1779 (McGuffie, *Johnson in the British Press*, p. 233).

Search is the paradigm of the age and our engagement with knowledge becomes increasingly bound up with searching. It seems to be the only way to cut through the mass of information with which we are now confronted. But search is not a simple or consistent process. While it might be possible to develop quantitative findings from fully curated and consistently presented resources like the *Proceedings of the Old Bailey* (www.oldbaileyonline.org) or the Text Creation Partnership (www.textcreationpartnership.org), it is not possible when the resource is essentially intended to assist navigation of unstructured data such as video images or scans from microfilm. Digitisation and search is frequently presented as having a consistent aim and structure but different projects develop in different ways for a variety of reasons, and to make critical use of a digital resource it is necessary to have some background information on the nature of the digitised material and on the evolution of the project.

Our dependence on search also assumes a particular relationship of information object to intellectual outcome. We hew out the raw material from newspapers, periodicals, books and then by a process of intellectual alchemy transmute it into an academic discourse, still largely presented in book or article form. Search has become a fundamental way by which we extract that raw material. The impression given by scholars like Colley or de Bolla is that search is an objective and scientific process, but actually the way in which we engage with a resource like the Burney Newspapers is much more iterative than lists of number of hits suggests. We try a variety of searches, assemble different fragments of information, and gradually try to piece the story together. In this backwards and forwards mixture of searching, reading, browsing and cross-checking, we will probably use a mixture of digital and printed resources, in much the way I have moved backwards and forwards between the digital Burney Newspapers, other digital resources and McGuffie's list in writing this essay.

In this context, how far do the deficiencies of search matter? Not a great deal. For all the problems with the OCR in the Burney Collection, it has still established itself as a fundamental and indispensable resource for the study of the eighteenth century. It would clearly be desirable for Gale to re-run the OCR in the Burney package which would lead to some improvements in the hit rates,[44] but we would still then be puzzling as to how we could trace those

44 The British Library reported recently that, in connection with the *Black Abolitionist Performances and their Presence in Britain* project, it had experimented with new OCR of some text material from the nineteenth-century newspapers collection with good results: http://blogs.bl.uk/digital-scholarship/2016/11/black-abolitionist-performances-and-their-presence-in-britain-an-update.html (accessed 11 November 2016).

allusive references to Johnson. Crowdsourcing, following the model of the Australian Trove package, might provide another solution.[45] But perhaps more important is to think about how we structure processes of reading, browsing and annotation around the use of a resource like the Burney newspaper. In his remarkable recent book, *Metaphors of the Mind: An Eighteenth-Century Dictionary*, Brad Pasanek describes how a process of search, reading and browsing enabled him to explore metaphors of the eighteenth century and to use them to create around headwords an encyclopaedic dictionary view of Enlightenment mentalities.[46] Pasanek calls this process "desultory reading" and I suggest it is this kind of process we want to capture in exploring resources like the Burney newspapers. Incidentally, it is perhaps worth remarking how Pasanek's use of digital techniques to explore enlightenment mentalities is far more sophisticated and useful than that of de Bolla. Pasanek bases his analysis on higher quality data – the structured texts in *Literature Online* rather than the poor quality OCR of ECCO. Pasanek also avoids quantitative methods rather looking at textual relationships to shape his discussion around a series of headwords. In general, Pasanek's book is an exemplar of how methodology needs to respond to data.

Ideally, I would like to be able to mark all the entries listed in McGuffie's checklist in the digital version of the Burney Newspapers. I would then like to be able to do the same with other digitised resources covered by McGuffie but not in the Burney Collection. This would give me a consolidated view of the resources listed by McGuffie. I would then like to start adding other material not covered by McGuffie, particularly the advertisements, but also the new material found in experimenting with fuzzy searching. I might want to mix a process of reading and annotation with one of search. I could produce these annotations in a variety of ways – I might develop a shared spreadsheet with links using something like Google Drive. One method of doing this might be by annotating digital resources, using a tool such as hypothes.is, an open annotation tool. This clearly has potential for recording the type of information given in Helen McGuffie's checklist of references to Johnson, but how it might perform with more complex types of listing is not yet clear. However, open annotation tools such as hypothes.is do suggest that there are other ways of

45 Marie-Louise Ayres, '"Singing for their Supper": Trove, Australian Newspapers and the Crowd', paper presented at IFLA WLIC Singapore 2013, available at http://library.ifla.org/245/1/153-ayres-en.pdf (accessed 11 November 2016).

46 Brad Pasanek, *Metaphors of Mind: an Eighteenth-Century Dictionary* (Baltimore: Johns Hopkins University Press, 2015).

engaging with digital resources such as the Burney Newspapers beyond the erratic and unpredictable simple search.

I am not advocating annotation of this form as a replacement for search. It seems to me rather to be valuable as a means of recording the kind of mixed methodologies that we all use. Its openness, and the potential for sharing, is also very important. In producing resources like the Burney newspapers, it seems to me that the model we all had in mind during the 1990s and first decade of this century was a closed one – that resources like the Burney Newspapers would be similar to big books, stand-alone finite resources whose contents could be navigated by search functioning in a similar way to an index. This closed approach was blown apart by the Text Creation Partnership, which first illustrated how a package like Early English Books Online could be extended, developed and effectively repurposed. EEBO/TCP are now an open-ended process, rather than a package, and this is a model of one way our use of the Burney Newspapers might develop. A good exemplar for this process is the way in which library catalogues have developed into open-ended and linked resources. I hope that we will see a similar process develop with major resources like the Burney Collection.

In the preface to her checklist, Helen McGuffie quoted Boswell to the effect that "from the diversity of dispositions it cannot be known with any certainty beforehand whether what may seem most trifling to some, and perhaps to the collector himself, may not be most agreeable to many".[47] As we develop online resources, it is our ability to build collaborative frameworks allowing us to explore and mix different configurations of search and annotation which will enable us to match this ideal of Boswell.

47 McGuffie, *Johnson in the British Press*, p. 6.