DHQ: Digital Humanities Quarterly
2016
Volume 10 Number 4

Digital library search preferences amongst historians and genealogists: British History Online user survey

Adam Crymble <adam_dot_crymble_at_gmail_dot_com>, University of Hertfordshire

Abstract

This paper presents the results of a study of 1,439 users of British History Online (BHO). BHO is a digital library of key printed primary and secondary sources for the history of Britain and Ireland, with a principal focus on the period between 1300 and 1800. The collection currently contains 1,250 volumes, and 120,000 web pages of material. During a website rebuild in 2014, the project team asked its registered users about their preferences for searching and browsing the content in the collection. Respondents were asked about their current search and browsing behaviour, as well as their receptiveness to new navigation options, including fuzzy searching, proximity searching, limiting search to a subset of the collection, searching by publication metadata, and searching entities within the texts such as person names, place names, or footnotes. The study provides insight into the unique and often converging needs of the site's academic and genealogical users, noting that the former tended to respond in favour of options that gave them greater control over the search process, whereas the latter generally opted for options to improve the efficacy of targeted keyword searching. Results and recommendations are offered.

"Is there anything the librarian can do to improve the success of *all* browsers, or at least improve the success of the *average* browser?" – Philip M. Morse [Morse 1970]. 1

Understanding the searching and browsing needs of a digital library's core users is important for anyone building a new online resource or refreshing an existing one [Agichtein et al. 2006a]. In 2014, *British History Online* (BHO) was seeking to rejuvenate its navigation and architecture after a decade in production and sought the direct feedback from its users to shape that process via a voluntary user survey of individuals who had registered accounts with the project. Volunteers were recruited through an email sent in May 2014 to all 6,301 registered users of the site, which included those with paid subscriptions, but also those repeat users who opted to sign up for a free account to take advantage of bookmarking features. 1,439 people responded to the survey call, representing a response rate of 22.8%. This paper discusses the findings of that survey as well as how they provide us with new understanding of the different needs and expectations of academic historians and genealogical users of digital libraries and archives, providing a basis for future conversations on identifying user needs more broadly. 2

BHO is a digital library of key printed primary and secondary sources for the history of Britain and Ireland, with a principal focus on the period between 1300 and 1800. The project is an initiative of the Institute of Historical Research, part of the University of London. The collection currently contains 1,250 volumes, and 120,000 web pages of material (each comprising the equivalent of many "pages" of original printed work). The site has particular strengths in the late medieval and early modern periods up to about 1660, as well as a strong local history collection.[1] As such it is both unique and typical of digitisation initiatives that contain primary or secondary source material. 3

BHO is a product of an evolution in digitisation stretching back to early attempts at microfilming and microfiching in the last century, but also early digital text collections such as *Project Gutenberg* [Hart 1971] as well as multimedia endeavours including the Electronic Beowulf project [Kiernan et al. 1999]. It is also specifically the product of a set 4

of British digitisation initiatives from the turn of the twenty-first century that focused on digitising British historical source material. These included *The Clergy of the Church of England Database*, [Burns et al. 1999-2016], *The Old Bailey Online* [Hitchcock et al. 2002], and the *Charles Booth Online Archive* [Donnelly et al. 2002] as well as *Eighteenth Century Collections Online* [Anon 1999] and *Early English Books Online* [Anon 2000] (see [Hitchcock 2008] for a fuller discussion of the digitisation landscape at the time).

5 The approaches to digitisation, as well as searching, browsing, and access, naturally informed a subsequent wave of British historical digitisation projects, including most notably *London Lives* [Hitchcock et al. 2010] and *Connected Histories* [Hitchcock et al. 2011] and a number of commercial collaborations between the British Library and various partners, resulting in archives such as the digitised *Nineteenth Century Newspapers* [Anon 2007] and *Burney Collection* [Anon 2009]. The project is also an important part of the technical revisionist reactions in the form of discrete re-curated datasets that eschew searchable databases entirely and instead focus on a mutable interpretation of a set of records as the new unit of dissemination [Boulton and Schwarz 2007] [Crymble et al. 2015] [Howard 2016].

6 These and similar projects have been engaged in ongoing conversations about search and browse that are more often expressed as websites or datasets than as journal articles. This paper takes those digital expressions and looks at it in the context of the extensive body of literature on user needs, as well as the findings of BHO's own user survey on the search and browsing desires of its large user-base.

7 After having launched in 2003, by early 2014 the site was beginning to show its age. At the time of writing, readers can find semi-functional archived copies of BHO via the *Internet Archive*.[2] Figure 1 shows a screenshot of the main page of the pre-development site. This old version of the site was navigable via keyword searching, or via browsing. Searching was the most popular option. The search feature compared to those found on most websites. At the time the rebuild began in early 2014, the site was navigable via a search box, powered by a physical Google Mini search server and based on common search algorithms used in 2010. Survey respondents from all three groups rated the search an average of 4 out of 5, suggesting they were happy (or at least comfortable) with the existing search capabilities.

Figure 1. Screenshot of British History Online on 31 December 2013, taken by the *Internet Archive*.

8 The search itself was possible because of the high quality transcriptions available in the BHO collection. The core collection had been built entirely using *double-rekeying*, a process that involved two typists individually transcribing the texts, with the resultant work compared and differences between them reconciled manually. This resulted in a very high level of accuracy, as it is unlikely that two typists would make the same mistakes. The team believes that the content produced through double-rekeying has an accuracy level of more than 99.995 percent.[3] Errors in the original printed volumes have been transcribed verbatim, leaving a level of ambiguity to the number of "errors" in the collection. Because of the double-rekeying approach, the corpus was relatively easy to keyword search - with the limits of keyword searching in mind, such as archaic spelling, abbreviations, and the occasional error in the original volume [Beall 2011]; [Badke 2011].

9 Users also had the separate option of browsing the collection through a series of tags that were manually created by the editors when the content was first uploaded to the website:

- **Places**
  - East, London, Midlands, North, Scotland, South East, South West, and Wales.
- **Subjects**
  - "Administrative and legal", "ecclesiastical and religious", economic, "intellectual, scientific, and cultural", local, parliamentary, "urban and metropolitan".
- **Timelines**
  - Henry VIII, Edward VI, Mary I, Elizabeth I, James I, Charles I, Interregnum, Charles II,

James II, William and Mary, Anne.

- **Centuries**
  - 11th and 12th, 13th, 14th, 15th, 16th, 17th, 18th, 19th.
- **Source Type**
  - Primary sources, Secondary texts, Guides and calendars, Gazetteers and dictionaries, Maps.

10 These tags are indicative of the historical nature of the collection, the British focus, and of early decisions about acquisition strategies, which were tied closely to the interests of some of the early project partners. These include, *The Centre for Metropolitan History* at the Institute for Historical Research, and *The History of Parliament* project.[4] This is why "administrative and legal" has been categorised distinctly from "parliamentary" amongst the subject tags. There was no way to combine searching and browsing in a meaningful way, or to filter results by more than one tag. The options had become what Marcia Bates described as "a hodgepodge of system elements working at cross-purposes" rather than an integrated searching and browsing environment [Bates 2002].

11 In order to improve these navigation options, the BHO team decided to seek feedback from its users on the types of features they wanted to be able to use to identify relevant records within the collection. Though the Internet is relatively young, a number of designers, computer scientists, and library and archive professionals have conducted research and written strategies for improving searching and browsing experiences for users of online collections. The studies include surveys such as the one conducted by BHO, as well as indirect monitoring of users with unobtrusive technologies such as eye-tracking or video recording [Granka et al. 2004]; [Tullis 2007]; [Hill et al. 2011]. There are also a number of studies that implicitly gather feedback, using techniques such as what Agichtein and his colleagues called "clickthrough interpretation" (measuring which links someone chose to click on a given page) [Agichtein et al. 2006b]; [Joachims et al. 2007]. Other forms of discrete monitoring include asking people to use computers fitted with tracking software [Grace-Martin and Gay 2001], analysing server logs [Bates 1996]; [Jansen and Spink 2006], or monitoring subjects in a lab setting [MaKinster et al. 2002]; [Tsai et al. 2012]; [Hsu et al. 2014]. Claire Warwick and her colleagues, for example, conducted this form of discrete monitoring of web logs for the "Log analysis of Internet resources in the Arts and Humanities" project, in an attempt to determine the digital use patterns of scholars in those fields [Warwick et al. 2008]. The growth in this implicit feedback analysis has contributed to the web analytics craze, in which free services such as "Google Analytics" allow website owners to monitor their traffic and gather statistics on user demographics.[5] Whether direct questionnaires or indirect monitoring are the most appropriate way to gather feedback is up for debate. Paul Samuelson argued instead for the power of "revealed preference", which monitors what people want by looking at what they spend their money on [Samuelson 1948]. As Harley and Henke warn, all approaches have their advantages and disadvantages; surveys may be inexpensive, but may not provide representative opinions, whereas analyses of transaction logs are less human and thus make it nearly impossible to infer the goals and intentions of users [Harley and Henke 2007].

12 BHO opted for a survey because it was inexpensive and provided opportunities for a large cohort of users worldwide to offer their opinions in a targeted manner. This mode also allowed users to express preferences for new types of services that did not exist on the current website and may therefore be impossible to monitor by collecting data on current use. We also believed that it was important to listen to our users, many of whom had been using the service for a decade. As each new mode of searching or browsing requires additional metadata to be generated and stored about the records in the collection asking users in advance of their preferences made it possible to prioritise development.

13 Searching and browsing within the context of information retrieval have long roots within discussions about libraries that extend far further back than the advent of the Internet. Should one search or browse to find the most pertinent resources for a project? The limits of the library itself – be that physical or digital – often impose limits on the researcher and make that decision for them. In the fictional fourteenth-century monastic library in Umberto Eco's *The Name of the Rose* [Eco 2004], it was the librarian who decided if you were worthy of finding what you sought.

What you received from the restricted stacks in the labyrinth above the reading room came down to his judgment alone; he was the keeper of the knowledge as well as its search interface. As Barclay argued in 2010, that power relationship between librarian and reader remained until the middle of the twentieth century. He noted that browsable open stacks were not an "ancient scholarly right", but an invention that was "no older than the baby-boomer faculty who so often lead the charge to keep books on campus". Barclay sought to dispel what he called the myths of browsing, noting that the academic value of browsing the stacks was already compromised by the large number of missing and stolen books, books currently in use by other readers, and by the very fact that many people cannot be bothered to look at items on the top or bottom shelves [Barclay 2010]. Barclay believed that new electronic search capabilities would enable scholars to continue to find books held in remote storage.

14 Some researchers long before Barclay saw the potential of search, including Bruce Schatz, who in 1997 predicted that by 2010 we would have "concept searching enabling semantic retrieval across large collections" [Schatz 1997]. Schatz was a bit premature in his prediction, but concept or semantic searching has recently become more important, as it becomes clear that what we want from a collection is a relevant document rather than a document that contains a particular keyword. For Dan Cohen, to the question, "is Google good for history", his answer was simply: "of course" [Cohen 2010]. While Jesse Shera, writing in 1964 noted that "the automation of libraries through the use of computers and book-finding robots is 'pure fantasy'", the digital era has certainly needed to meet some of those challenges [Shera 1964]. Users of online archives and libraries need to be in possession of all the tools that they need to help find what they are after.

15 Not everyone agreed with Dan Cohen's enthusiasm; even as early as 1985, Champlin warned of the "perils and pitfalls" of online search, arguing for the importance of print indexes and abstracts for researchers [Champlin 1985]. A survey of historians' preferences by Ian Andersen in 2004 showed that little had changed, with print-based finding aids and "informal leads" from colleagues or librarians the preferred information retrieval strategy of scholars in that study [Anderson 2004]. Many scholars used browsing and shelf proximity as part of their information retrieval strategy. Richard Mott countered Barclay's arguments about the limits of browsing by highlighting the practice of obtaining call numbers as a way of identifying an appropriate shelf to begin browsing. Mott believed that it was "not uncommon for this browsing to yield relevant sources that electronic searching alone misses" [Mott 2010]. This is the principle of serendipity in discovery. Stephen Ramsay calls it "screwing around". For Ramsay, browsing is an important part of the scholarly process, which prevents us from focusing solely on the predefined literary canon. He urges us to take time to wander the library and grab things that seem like they might be interesting, as a means to discovering new knowledge or unexamined connections [Ramsay 2010].

16 Within the world of online libraries, Maxwell argued that simple browsing was more important than keyword searching [Maxwell 2010]. Grey and Hurko echoed those same warnings, suggesting that controlled subject searching was more effective than keyword searching for research, which of course necessitates a good set of controlled subjects built into the collection [Grey and Hurko 2012]. Collins and her colleagues noted that researcher uptake of new digital modes of research depended heavily on what was considered normal and acceptable within their discipline [Collins et al. 2012]. These disciplinary standards in history were well entrenched; in 2003, a survey of 278 historians by Dalton and Charnigo found that while 58% of respondents used websites for their research, only 3% believed it was the most important way they conducted research, lagging far behind traditional finding aids, footnotes, and archival or library catalogues [Dalton and Charnigo 2004]; [Hamburger 2004]. The four million annual users of BHO suggests that acceptance of digital resources may have moved forwards in the past decade. Genealogists, despite tending to be older, have always been more willing to engage with digital sources than their academic historian cousins [Duff and Cherry 2001]. Nevertheless, the question remained: how do these users *want* to be able to explore the collection in ways that they could not do on the pre-development version of the BHO website?

17 Since its launch in 2003, BHO has developed a strong user-base that has been attracted to the historical content on the website. In the project year 2013-14 (August – July), the site received more than 4-million visits from 2.7 million unique visitors. Most visitors are anonymous, however, BHO's users can largely be classified into three categories: academic historians, genealogists, and casual users. Each group comes to the site with different expectations and

needs from the collection.

18 Academic historians use the digital library to conduct historical research that is destined for peer-reviewed publication or postgraduate theses. Many of these users are attracted to the subscription-only materials such as the *Calendar of State Papers*, or the *Calendars of Close Rolls*. This subscription content comprises twenty percent of the material on BHO, and is targeted at these academic users.[6] 354 survey respondents (25%) classified themselves as academics or students. The majority of academic users were between the ages of 55 and 74 years, with people under the age of 55 accounting for most of the outliers (see Figure 2). Nearly six in ten identified as male.

19 Genealogists or family historians are the second main group of users. They tend to use the site to look for details of their family's past, and are generally attracted to the *Ordnance Survey Maps*, as well as the local history resources about the communities in which their relatives lived. 737 respondents (51%) classified themselves as genealogists. Genealogical users were typically older than the academic users. The majority were between 55 and 74, but the largest group was older than 65, and there was a large group older than 75 years. Very few users were under 45. The split between male and female respondents was equal, unlike amongst academics.

20 The third group is best described as casual users, a large number of whom arrive via search engines or sites such as *Wikipedia* [Blaney 2013]. The group includes journalists looking for stories, business owners interested in the heritage of their premises, historical video game makers researching historical context for upcoming projects, and people reading for enjoyment. 348 respondents (24%) decided not to classify themselves under one of the above headings or felt the headings did not apply to them. Their age profile most closely mirrors the academic users, and suggests that most of them are not amongst the genealogical group. Given the fact that most users heard about the survey via an email to the address they used to register with BHO, it is likely that many of the people in this category were repeat users. Because the needs and interests of the casual users are so diverse, they are the most difficult to pin down. While the discussion in this paper will include the views of the respondents from this group, they will not be the study’s core focus.

21 As the average age of BHO users is older than the typical web user, age may be pertinent to a full understanding of the survey results and search and browse preferences. There is a large body of literature that studies the web needs of older web users, or "silver surfers" as they have been called in the past. Much of this research was done in the early years of the 2000s and focused on the cognitive, physical, and sensory decline of the elderly, as well as on ways to address the alleged "fear" older people have of technology [Darin and Kurnaiwan 2000]; [Morrell et al. 2000]; [Zajicek 2001]; [Millward 2003]; [Aula 2005]; [Kurnaiwan and Zaphiris 2005]; [Kurnaiwan et al. 2006]; [Chadwick-Dias et al. 2007]; [Dickinson et al. 2007]; [Gao et al. 2007]; [Tullis 2007]; [Fairweather 2008]; [Hill et al. 2011]. This literature, most of it presumably written by academics under the age of 75, while good intentioned, can at times be patronising and is perhaps less apt for today’s "silver surfer" who may have been in his or her forties when the web first became popular, and thus has decades of experience with the Internet. From the perspective of the BHO project team, it was not so much the age of these users that was important, but how these different groups wanted to navigate online collections like BHO.

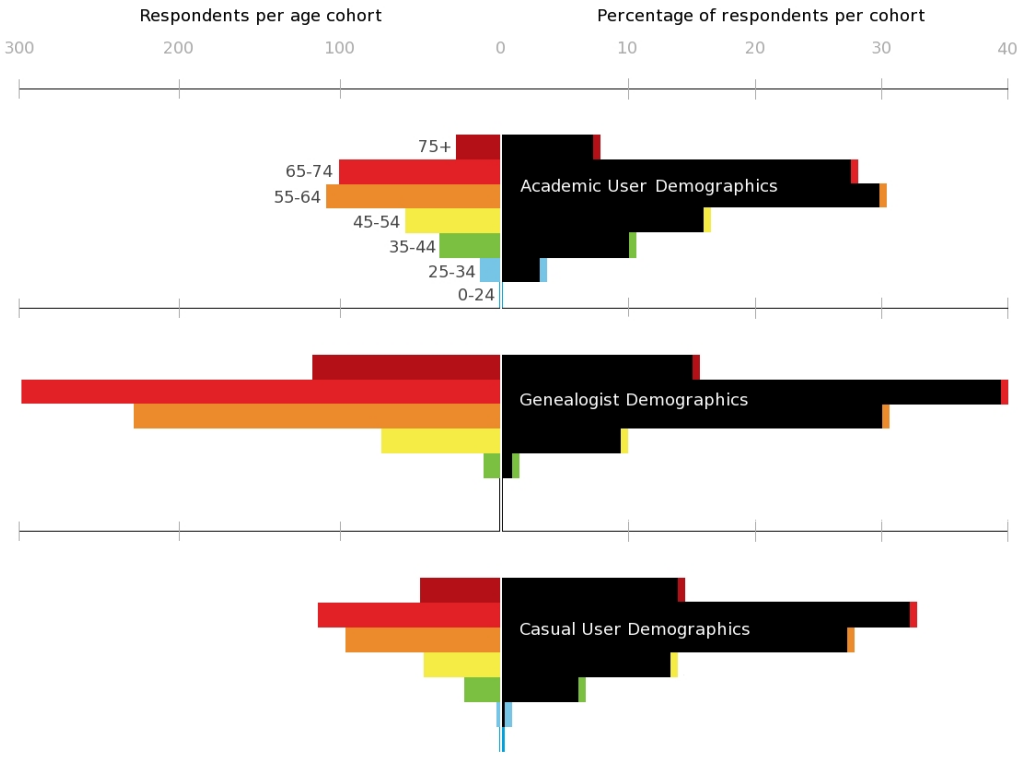

Figure 2. Age profile of survey respondents by type of user.

To get a better understanding of respondents, users were asked how frequently they used the site. Results can be seen in Figure 3. The most common answer in all three groups was that they used the site "occasionally" (31% academics, 46% genealogists, 32% casual users). Academics were considerably more likely to think of themselves as regular users, opting for the "often" or "very often" categories twice for every genealogist respondent (37% academic versus 15% genealogist and 21% casual users). 22

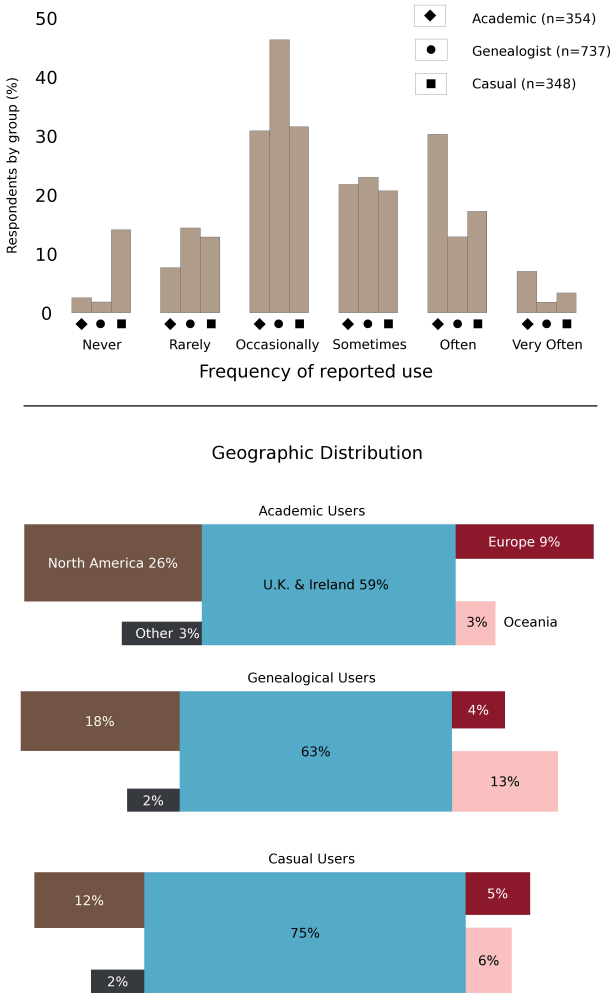

Figure 3. Self-reported frequency of BHO website use and geographic location.

A majority of all users were based in the UK or Ireland (66%), with casual users showing even higher levels of local use – See Figure 3. North Americans comprised 19 percent of respondents, with academics proportionately more likely to be North Americans taking advantage of research materials that for them may have been overseas. Only 9 percent of respondents were from Oceania, with a clear majority of them using the site for genealogical reasons. Europeans were rare and generally were academics. Respondents from the rest of the world were very rare. 23

The respondents of the survey are therefore of an older set of Internet users, typically over the age of 55, and mostly British and Irish. Academic users are more likely to be heavy users of the site, and are more male than female. Genealogical users are occasional visitors, and have no discernible gender variance. 24

The rest of this paper discusses the project team's findings derived from the survey. 25

Current Information Retrieval Preferences

Respondents were asked about their current preference for finding information stored online in sites like BHO (for full data see Appendix I), as well as their wishes for the future (see Appendix II). With regards to current preferences, respondents were asked whether they "never", "rarely", "sometimes", "often", or "very often" used the following discovery methods: 26

1. Simple keyword searching
2. Advanced keyword searching
3. Browsing by subject

4. Browsing by publication
5. An external library catalogue or website

For ease of discussion, these have been recombined into three categories: 27

- popular (often + very often)
- occasional use (sometimes)
- unfavoured (rarely, never)

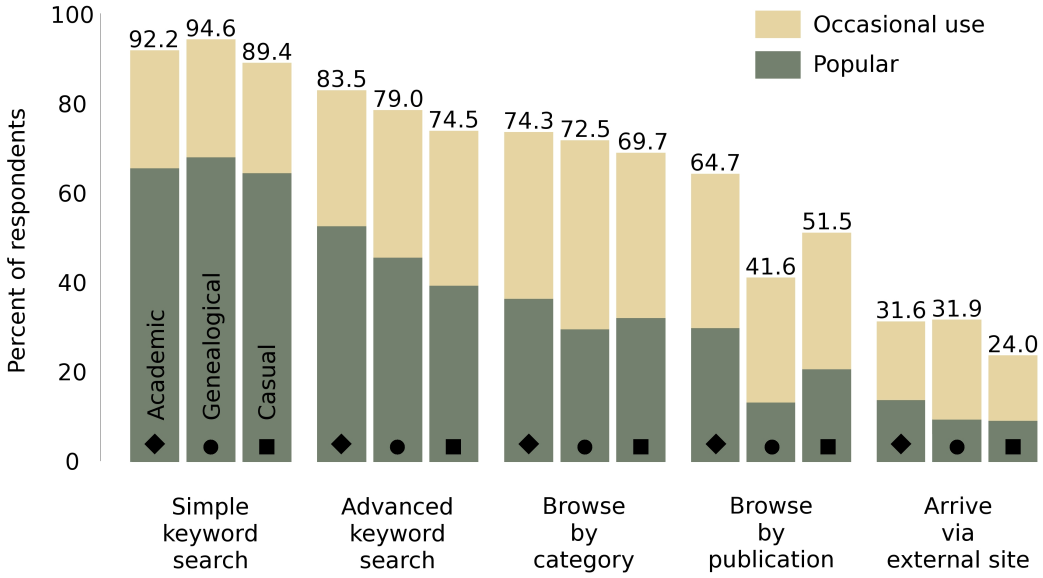

Figure 4. Search and Browse Preferences by User Group

Searching was more popular than browsing for all three user groups (see Figure 4). The ubiquity of search on the Internet means that keyword searching is currently a must for digital libraries. In this case, that means full-text searching whenever possible. Simple keyword searching was more popular than advanced keyword searching for all three user groups. More than half of all users ranked simple keyword searching as a popular option for them. There was very little variation between the groups. Nine in ten respondents claimed at least occasional use of this type of feature. Given the overwhelming level of search on the Internet, the project team members are skeptical of the 44 people (3.25%) who claimed "never" to use simple keyword searching on sites such as BHO. 28

Advanced keyword searching was also popular with most users, but was most popular amongst academic users, of whom 53% claimed to often or very often use this form of searching, and 83.5% at least sometimes did so. Academics are probably more likely than the other groups to overestimate or overemphasize their use of this feature, as it is generally viewed as the more "scholarly" approach, so this finding should probably be viewed as user reporting of their preferences rather than actual usage patterns. Genealogical users were slightly less likely to use advanced searching, and used it considerably less often than simple keyword searching. However, reported use in this group was still high. Casual users were least likely to use advanced keyword searching, but a majority still claimed to do so. Encouraging more uptake may require increased training, as one respondent noted that "the advanced keyword searching has never worked for me. There is not enough explanation". Many academic users receive such training as part of library short courses or induction training for new students, but these options are less likely to find their way into the general public's hands. Despite these slight variations, the overall message is that digital library users overwhelmingly like some form of keyword searching and that both simple and advanced options are used by a majority of users. 29

Though still supported by a majority of users, browsing was less popular than searching and showed considerably more variation between the groups. Academic users were more keen on browsing options, with more than a third 30

declaring browsing by subject a popular option, and marginally fewer stating the same of browsing by publication. This suggests that academic users are interested in using digital libraries to find and read copies of specific books, and that they might be amenable to serendipitous discovery via a well-designed taxonomy system of subject headings. This finding implies that Richard Mott's arguments in favour of using shelf proximity for finding relevant works in physical libraries may also have value for users of digital libraries [Mott 2010]. Academic interest in browsing also suggests a desire for thoroughness, to ensure all possible relevant works have been discovered and exploited, and that a search box cannot be trusted to reach that level of diligence.

Genealogists were slightly more interested in browsing by subject than were casual users, and both were slightly less interested than academics, but not substantially so. However, there were significant differences in preferences for browsing by title, with genealogical users far less interested in this form of discovery (popular: 13.5%) than academics (popular: 30%) and casual users (popular: 21%). This may be explained by the different goals of genealogical users, most of whom are looking for answers or information about their family or the communities in which they lived, and are not as interested in the source of that information. A majority of genealogical users stated that browsing by publication was an unfavoured option. Depending on the type of users a digital library attracts, these differences in preference are important to consider when building a service. 31

Finally, users were asked about their preference for arriving at sites like BHO via a library catalogue or an external site such as *Wikipedia* or a search engine such as *Google*. This option was overwhelmingly unfavoured, with a clear majority in all three groups rejecting this idea. This outcome does not necessarily preclude the importance of external linking, or of integrating content from online libraries into library catalogues, and BHO's anonymised traffic logs suggest that this is an important driver of traffic. Some of the comments left by survey respondents confirm this, with one person noting that they often searched commercial search engines for "keywords that I know are on your site", bypassing the need to arrive first at BHO before diving into the collection. While this may be an important supplementary source of traffic for digital libraries, the survey results do suggest that most users prefer to have the option of navigating a digital library via the site's own navigation system rather than relying on indexing or linking from elsewhere. This suggests that users are interested in maintaining access to what have become known as *siloed* websites containing discretely curated collections, and are perhaps not yet on board with the movement towards aggregated search websites such as *Connected Histories*, which allow users to search a number of repositories from a single search box [Hitchcock et al. 2011]. 32

Rating of New Search Features

In the past decade there have been a number of changes in search technology and entity extraction, allowing people to focus their search results in new ways. The BHO team curated a subset of these options that we felt were appropriate to the historical collection under our management and respondents were asked to rate the likelihood that they would use twelve search features if they were added to the BHO website. The search features can be broken into four categories: 33

1. Set advanced search restraints
    1. Fuzzy searching
    2. Proximity searching

1. Limit search to a subset of the collection
    1. Search free content only
    2. Search within a series of books
    3. Search by content type (maps, texts, etc)

1. Search by publication metadata
    1. Search by title
    2. Search by publication date

1. Search entities within texts
    1. Search by person name
    2. Search by place name
    3. Search by location coordinates (latitude / longitude)
    4. Search footnotes only
    5. Search by date of subject matter

Set advanced search restraints

A number of algorithms have been developed or are under development that seek to improve search results, but also to give users the power to control factors that help determine what matches are returned. Among those increasingly familiar to users of digital libraries and archives are *fuzzy searching* and *proximity searching*. 34

Fuzzy searching is common on websites containing poor quality optical character recognition (OCR), which is an algorithmic means of converting digital scans or photographs of text to machine readable and search engine indexable text that can be searched by web users. A number of genealogical websites and historical databases such as historical newspaper repositories use fuzzy searching because the quality of the OCR is often quite poor. 35

Most commercial databases containing historical source material were produced using OCR and have varying degrees of accuracy.[7] At the time of writing, commercial software packages boast near-perfect accuracy levels, but the tests those companies conduct to make those claims are almost certainly done using modern fonts on crisp white sheets of paper and are perhaps more suited to the needs of a legal office than a library or archives seeking to digitize historical materials en masse. When the *Australian Newspapers Digitisation Program* (also known as *Trove*), undertook one of the world's largest digitisation projects to date in 2007, they discovered their historical newspapers introduced problems that the commercial OCR software at the time had difficulty handling. These included issues with deteriorating inks, highly complex layouts, and narrow space between lines, columns, and gutters, as well as problems with the microfilm that stored the newspapers being digitised. These microfilms may have been second or third generation copies, poorly focused, and dirty or scratched [Holley 2007]. Combined, these issues made OCR accuracy a concern. 36

To test the accuracy of their own project, *Trove* had team members manually check for errors in the OCR of digitised newspaper pages and a representative sample of 30 pages showed accuracy levels could be split into three groups based on average character accuracy, the results of which can be seen in Table 1. 37

Category	Average Character Accuracy (%)
Good group	98.02
Average group	92.61
Bad group	71.00

Table 1. The average character accuracy of newspaper pages from the Australian Newspaper Digitisation Program [Holley 2007].

For many, these numbers may seem high and are perhaps good enough. This is particularly the case when we consider the scale of the task involved. However, it is worth considering the following sentence has an 80% average character accuracy: 38

Tnis Is whot eigkty percemt accurecy louks lilce 39

At that level of accuracy, a human reader should be able to make out the sentence. But which keywords could you use to find it in a larger text? The problem is not merely with the level of accuracy, but also with the level of accuracy in key places; specifically, as a user you want to know that the keywords you seek have been accurately transcribed by the software. In the example above, using another measure known as *average word accuracy*, the accuracy 40

drops to 0%, making the transcription nearly useless for anyone attempting a keyword search.

This is what ninety-eigkt percent accuracy looks like 41

This second example, at 98% average character accuracy, is certainly an improvement, but still leaves those searching for “ninety-eight” without the level of accuracy needed for their project. Fuzzy searching works by incorporating spelling variations within its algorithm, focusing specifically on combinations of characters for which there are known OCR problems, such as "rn" (R N), which can often be interpreted as a lower case "m" by the software. This increases the chance, but does not guarantee that a user can overcome errors in automatic transcription (see [Myka and Güntzer 1996]; [Rares and Chen 2009]). 42

The option is also helpful for users seeking family or place names that may not have been spelled properly, were spelled phonetically, or may have used archaic spelling. For example, the surname "Kedgley" might also be spelled "Kegley", "Kedglee", or "Kidglie" and may still refer to the same historical individual [Titford 2005]. A fuzzy search algorithm trained using the *Soundex Algorithm* for converting English words to their composite phonemes can make it possible to fuzzy search for words that most likely sound like the search term but that may look quite distinct [Russell 1918]. It is also particularly helpful for early modern and medieval texts that were created before the push to standardise the English language in the eighteenth century. Techniques such as this often use Levenshtein distance, which can aid in this process of identifying similar words. Levenshtein distance calculates the number of changes it takes to turn one word into another. For example, changing *Water* into *Wine* takes three changes (change "a" to "i", "t" to "n", and remove "r"). A word with a Levenshtein distance of three is almost certainly a distinct English word, but as in the example above of "Kedgley" and "Kegley", could be a useful way of identifying good alternative results. 43

Though the texts in BHO were not transcribed via OCR and contain very high levels of accuracy, many of them use medieval and early modern language, and thus users could potentially benefit from a fuzzy search option. A large majority of survey respondents agreed, with very little variation between the three groups of users. Academics were slightly more likely than other users to really want this feature, but support was high in all groups, with more than half of respondents rating it "quite useful" or "I’d like to have that" and only a quarter unsure about its benefits (see Figure 5). The levels of support certainly suggest that builders of online libraries should experiment with the value that fuzzy searching could provide for their collection, or at least make the limits of their search options clear if this feature is not available so that users are aware of the potential pitfalls of relying on the search box. 44

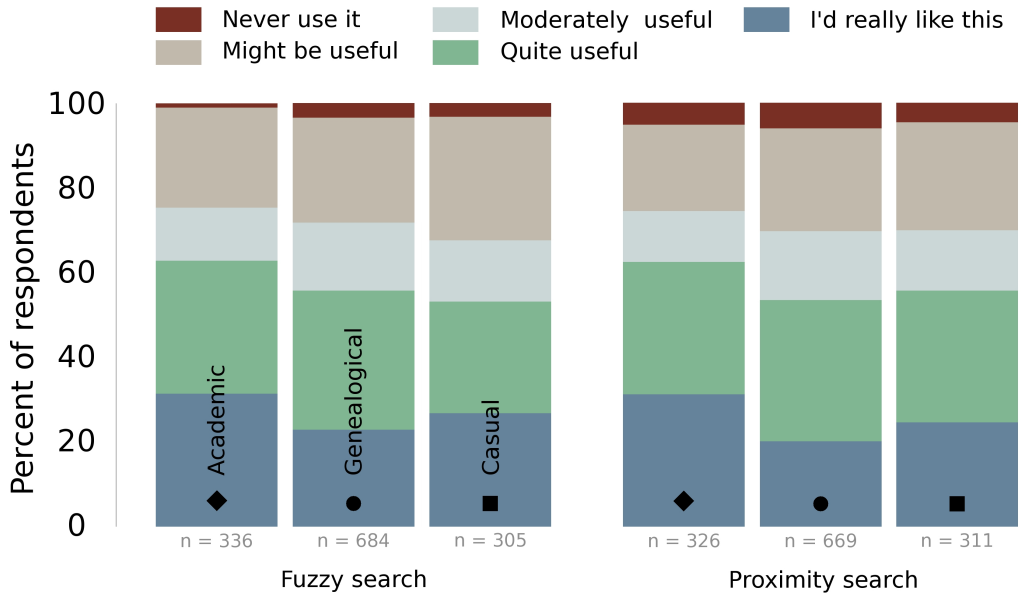

Figure 5. Advanced Search Restraint Option Preferences by Group.

Proximity searching is another algorithm designed to give users the power to define search parameters [Schenkel et al 2007]. It is as yet less common than fuzzy searching. The option allows users searching for multiple keywords to limit the distance between them, ignoring results that appear too far apart as well as helps searchers avoid overly specific search parameters that might otherwise miss pertinent information. For example, someone interested in Caroline of Brunswick, the wife of King George IV, might search for "Caroline of Brunswick" as she was commonly known. However, her full name was "Caroline Amelia Elizabeth of Brunswick-Wolfenbüttel", and a strict search for "Caroline of Brunswick" would miss that result. By using a proximity search for "Caroline" and "Brunswick" limited to within five or perhaps ten words of one another, the match would find both patterns. 45

At the other extreme, proximity search is useful for preventing irrelevant matches of unrelated words that appear far from one another in a text. This is particularly useful for web pages containing whole chapters of text, or perhaps even full books. These particularly long pages will be inordinately likely to match a set of two or more keywords than will a short page, purely based on the number of unique words. This means a user searching for "blue dog" (without quotes) in the BHO website will currently be directed to a chapter on the London village of Rotherhithe in *Old and New London, Volume 6*, where both the words "blue" and "dog" appear, but not in a context in which one is related to the other. Instead, blue refers to "Blue Anchor Road", and "dog" refers to the "Dog and Duck" tavern, much further down the page [Walford 1878]. This masks the real entry of interest: a "blue dog" allegedly given to Thomas Cave by Thomas Cromwell on the 8th of July 1528 [Henry VIII 1528]. 46

As much of the collection on British History Online includes chapter length works on a single webpage, this is a particular problem for users searching for multiple terms, especially if one of those terms happens to be a fairly common word in the collection, such as the name of a British place, or a common noun. To ultimately solve these challenges, the makers of search algorithms are turning towards semantic search options that looks for what the user is actually interested in finding rather than the sometimes ambiguous keyword they happened to type into the search box. Instead of asking for documents containing a keyword, you could look for documents about the concept embodied by that keyword, which may not contain the word at all [Crymble 2015]. These techniques rely on gazetteers, or lists of words that are associated with certain concepts. This leverages the idea of the thesaurus: that there is more than one way to express an idea with different words. The "Semantic Annotation and Mark-Up for Enhancing Lexical Searches" or SAMUELS project currently underway at the University of Glasgow is taking this approach as it seeks to "deliver a system for automatically annotating words in texts with their precise meanings, disambiguating between possible meanings of the same word" [Alexander et al. 2015]. Instead of searching for one or two keywords, the algorithm can search for dozens or hundreds of related terms at the same time, and present the user with a wider range of potentially relevant materials. The SAMUELS project relies on historical thesauri, putting the word in the context of its surrounding words to assign it to a metaphor: a word or series of words with an abstract meaning that captures the intent of the person expressing that idea. 47

While these more advanced options are in development, in the meantime, proximity searching gives users the power to set some limits on the matches the search engine will find. Proximity searching saw very similar levels of support amongst survey respondents to those expressed about fuzzy searching. More than half chose the top two categories of preference, and again roughly a quarter were unconvinced. Like fuzzy searching, academic users were slightly more likely to say they would "really like to have" the feature (31% academic; 20% genealogist; 24% casual users). This suggests that some academic users in particular are looking to be given more control over the types of results they receive from a search engine. 48

Whether or not it is feasible to integrate this type of advanced searching option into an online library website largely depends upon its availability as a setting choice on the search engine software employed by the project. Open source search engine Apache Solr currently allows this type of feature to be turned on by administrators, rather than built from scratch, so parties developing online libraries should both interrogate the needs of their users, but also the features available on various search engine packages before investing in a solution.[8] 49

Limit search to a subset of the collection

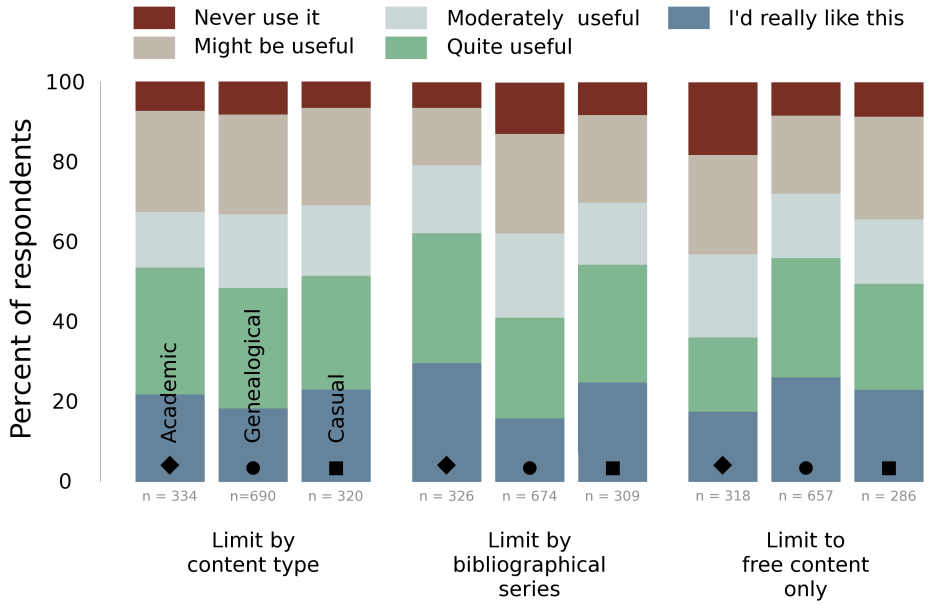

Figure 6. Preferences on Limiting Search to a subset of the collection.

The opportunity to exclude certain parts of a collection from a search could reduce the number of unwanted results rather dramatically for a user. Respondents were asked how useful it would be to limit their searches to certain types of content (only text, or only maps), to search only within a series of works (such as the *Survey of London* or *Calendar of Treasury Books* series), or to search only "free content" (non-subscription material)[9]. These particular questions were chosen with the BHO website in mind, but the project team believes that they are representative of the broader desire to have the ability to ignore part of the collection in the searching process. Respondents to the survey generally viewed these limiting options favourably. A majority of users in all user categories chose to rank all of these options "moderately useful" or better (see Figure 6). 50

Searching by type of content was considered useful for all types of users, with very little variation between them. One in five users in all categories stated that they would "really like to have" the option to exclude content of certain media types. Nearly seven in ten felt it was at least moderately useful. The BHO collection contains a large textual corpus which is split into "primary sources", "secondary texts", "guides and calendars", and "datasets".[10] All of this material is easily keyword searchable as it is primarily text based, but it is not necessarily useful to all types of users, with "guides and calendars" primarily of interest to academic users, for example. The collection also contains a substantial set of historical maps, principally the *Ordnance Survey* maps of Britain, as well as some early modern maps of London, which are necessarily more difficult to search with keywords.[11] A user only interested in primary source material might therefore benefit from the ability to omit secondary texts from the search. 51

Searching within a bibliographic series was less popular than searching by content type, and was considerably more popular amongst academic users than genealogists. More than six in ten academics ranked this option "quite useful" or better, compared to only four in ten genealogists. Nearly thirteen per cent of genealogists said they would "never" use this feature, again suggesting that for a genealogist, finding information about a person or place of interest far exceeds the importance of where that match came from. Casual users were much more like academics in their preferences, suggesting that it is the genealogists who are unique in their attitudes towards this option. 52

Limiting searches to "free" or non-subscription content was more polarising than the other options. Academics were generally uninterested in this option, with 17.5% claiming that they would never use this feature – more than double the rate expressed by the other two groups. Genealogists were far more likely to want this option, with nearly seven in ten rating it "moderately useful" to "I'd really like to have this". Casual users' preferences fell in between those of the other two groups. The reactions to a "free content search" are perhaps the most interesting of the options in this category, suggesting that academic users think that they are prepared to access material even if it will incur a cost, 53

as thoroughness is of the utmost importance. Another explanation may be known trends in BHO subscribers. A number of academic libraries around the world subscribe to BHO on behalf of their staff and students, meaning that a higher proportion of academic respondents may already have access to the subscription material and therefore see no value in restricting their search thus. The results could also suggest that many genealogical users may prefer to remain in the dark about matches they cannot see, rather than find themselves frustrated by a tantalizing match that will cost them money. Without the option of following up with users, this is merely speculation, but the clear differences in preference for this type of feature suggests digital libraries and archives need to consider their audiences carefully when deciding if they will integrate it.

Search by publication metadata

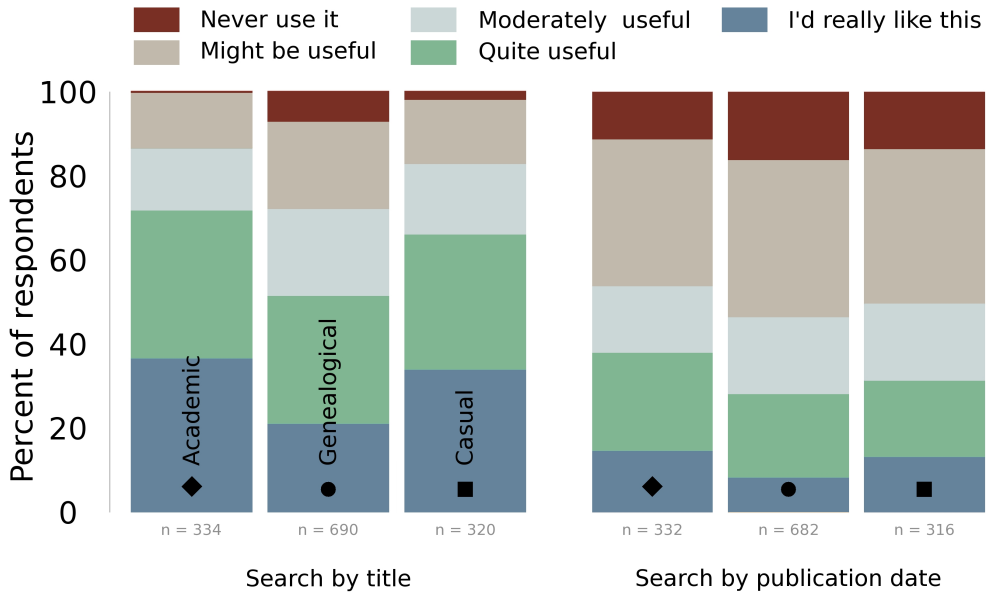

Figure 7. Preferences on searching by metadata fields

Most library catalogues contain extensive metadata about a publication. The types of metadata available on books are nearly endless, with standards such as Machine Readable Cataloguing (MARC), the Dublin Core Metadata Initiative, the Metadata Object Description Schema (MODS XML), and Context Objects in Spans (COinS) widely used in the library world to store data on everything from the author of a work, to its title, publisher, or date of publication.[12] If good metadata standards are adhered to and good data collected, these fields could be useful to people interested in searching the collection in particular ways. For example, someone might want to search only the title of a book rather than the full text of the entire collection. Or they may be looking for works by a particular author, or published in a certain city, or by a certain publisher. For a typical user, it may prove unhelpful to find only work published by eighteenth century bookseller Thomas Cadell, but for the right research project that ability may prove invaluable. 54

As extra metadata fields for a publication add extra time and therefore extra cost to the cataloguing process, the value of this for a particular site should be weighed carefully. Extra fields in the search interface also require a way of integrating the option into the design of the site, and too many such options may prove for a less effective user interface. To test the appetite for this type of metadata field searching, the survey asked respondents to rank the usefulness of being able to search by two such fields: title, and publication date. 55

As can be seen in Figure 7, searching by title was a considerably more popular option than searching by publication date. This is perhaps not surprising, as the ability to search by title is consistent with options available in many digital library catalogues. A large majority of academic users, but also casual users, wanted the ability to search by title, while far fewer genealogists wanted the option, which is again consistent with genealogical needs to find 56

content rather than to read widely. While the survey could have asked about many other metadata fields, we felt that these two represented an obviously and less obviously useful option, and the user responses reflected that. This suggests that developers will want to choose carefully which if any metadata fields are exposed to search options for users, as some are unlikely to find wide user-bases.

Search entities within texts

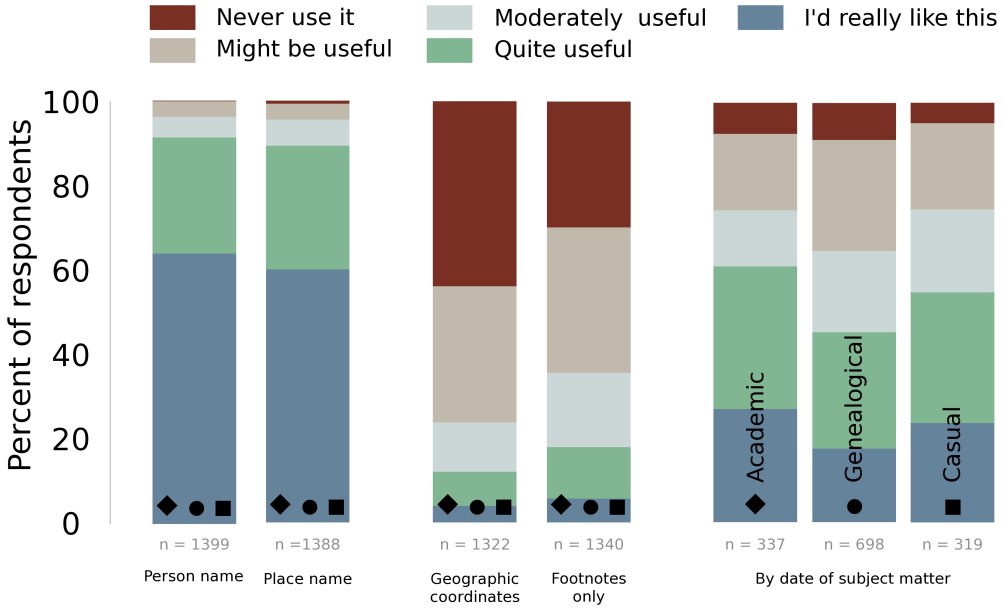

Figure 8. Preferences on searching by entities within the text

Tagging entities within the text also makes it possible to allow users to search only those entities. For example, many genealogical websites have tagged the names of people and places, knowing that these variables are often useful for finding relatives. People with names that are the same as common English words can be difficult to find without this option. This requires identifying the entities in the corpus and annotating them or indexing them first. This can be done in a semi-automated fashion using named entity recognition, also known as *term extraction*, but to achieve a high degree of accuracy, it often requires extensive manual markup and editing [van Hooland et al. 2013]; [Freire et al. 2012]. Once all instances of names are extracted, this information can be opened up for searching by the search engine administrator, but as it requires a significant investment of time and money, it is important to determine whether or not a community would find such a feature useful before proceeding. Respondents to the survey rated five different entity extraction options: searching by person name, searching by place, search by the date of the subject matter covered in a work (as opposed to the publication date), search using geographic coordinates (latitude or longitude), and search only the footnotes (possibly useful for looking for references). 57

As can be seen in Figure 8, these were not universally popular, though apart from searching by the date of the subject matter, opinions about each option were almost universally consistent between the three user groups and have thus been depicted together in the figure (full data in Appendix II). Searching by person name and place name were incredibly popular options across the board, with six in ten people stating they would "really like to have this" option, and only 1 in ten showing little or no interest. It was not surprising to see high levels of enthusiasm amongst genealogical users, but it was more surprising that academics and casual users were also interested in these features. This result suggests that people are not "reading" books in these online environment as they might in a physical library, but are instead looking for references to specific people or places – though this cannot be fully substantiated from the data available in this survey. 58

Searching by geographic coordinates and searching footnotes only were both unpopular options. More than four in ten people said they would never use the "geographic coordinates" search, and only about one in ten thought it was 59

a particularly good idea. Users are perhaps too used to the value of keywords, and are not yet thinking in terms of being able to search for works about a place that may not be explicitly mentioned in a text – eg, within five miles of X. The footnotes only searching was only marginally more popular with about fifteen percent of users wanting to see this feature. Not surprisingly, academics were more likely to want to search just the footnotes, presumably to identify other potential references worth pursuing. About 43 percent of academics ranked the option "moderately useful" or better, compared to only 30 percent of genealogists.

Finally, date of subject matter, which has been split into the three distinct user groups in Figure 8, shows the academics and casual users with remarkably similar preferences, as has often been the case throughout this study. Both academics and casual users were fairly supportive of the ability to search by the date of events discussed in the texts, with a clear majority in favour, while only 45 percent of genealogists felt the same way. 60

The outcome of this question shows that entity extraction is an incredibly popular option for digital library users, but that not all forms are as popular as others, so understanding the project's user base is important. 61

Conclusions

The results of this survey show that genealogists are primarily interested in names of people and names of places related to their search. They do not seem to be interested in reading material, or even what the source of the material is, as long as it contains the details they are after. Searching free content only was more important to genealogists than it was to academics and casual users. 62

Academic users are in many respects very similar to casual users. They, perhaps unsurprisingly, tend to be drawn to more control over aspects of searching and browsing that will allow them to scour a collection systematically. They seem less willing to trust the search and browse features, so they claim to like options that put them in control. 63

Overall, all three groups have clear preferences for searching by name and by place, and they are generally in favour of searching by the date of the subject matter. They strongly support fuzzy searching and proximity searching. Whenever possible, they prefer to conduct simple keyword searches, but they prefer advanced searching to browsing – again implying they are after snippets rather than a traditional reading experience. Being able to limit the search to predefined subsets of the collection, including by content type or by bibliographic series was popular. 64

By contrast, they do not like to have to rely on external search engines or library catalogues, and would rather interact directly with the navigation features of the digital library. They saw little value in being able to search by geolocation coordinates or to search within the footnotes of the collection only. While they were in favour of searching by title, they were less interested in searching by publication date, suggesting not all publication metadata was of interest to users as an information retrieval strategy. While the age of the BHO user-base was typically older than the average Internet user, age does not seem to be an inhibiting factor in the preferences of these web users. Apart from their self-reporting of their age, nothing in the findings suggests these respondents are suffering from age-related disabilities that influence their preferences, or that they are in any way less advanced Internet users than someone much younger. Instead, the types of tasks they seek to undertake seem to be the primary drivers of preference. 65

These specific findings contribute to ongoing discussions of the needs of users of digital collections more broadly. Because all collections have curators and users, the findings are equally applicable to any field, from cultural heritage, to private enterprise, to institutional repositories. For project managers, the challenge is to ensure that the services for accessing the collection material meet the needs and expectations of those users. The survey tells us the specific wishes of distinct groups of BHO users, but it also points us to the underlying theme: what do users want to do with the material in the collection? Genealogists in the sample group were less concerned with where or how they found something as long as they did find it. They were not using the site to read as one might in a traditional library, nor were they conducting any form of digital analysis. For them, the collection was a potential source of specific information in their wider search for details about their family history. Academics, on the other 66

hand, wanted control over their research processes, and were looking for tools that would let them take a systematic approach to both search and browse. Therefore, it is not the fact that one group of historians is academic and the other amateur that underlies the distinction between them. The important factor is that the two groups have different goals and different reasons for accessing the material in the first place. Understanding those needs is a fundamental first step in designing a search/browse facility to let both groups make the most of a collection, no matter what that collection contains. This is something that any collection manager can take away, regardless of their own target audiences. A self-assessment of how a typical user plans to use one's site is an important part of allocating energy and resources in the building and re-building process.

The specific results of this survey will also prove useful for collection managers looking to discern which approaches might be most suitable for their own audiences. Despite this value, the findings are not static; technology and user expectations on the web are constantly shifting. As users increasingly expect to be able to use digital content in digital analyses, be that data mining, or linked data, the pressures on project teams to provide increasingly open access and advanced search and browse options will continue to mount. In the meantime, it is the project team's hope that providing the findings of this survey will prove useful for those building or redeveloping digital archives and libraries, or digital collections more broadly. 67

These quantitative results give a clearer picture of what users of digital libraries are looking for and hoping to see in the searching and browsing options available on websites such as BHO. However, the most important advice received via the survey is much more qualitative and came from one of the respondents, who pleaded: "Keep it simple please I don't want 100's of options when I search for specific topics". It is all down to usability, after all. 68

Acknowledgments

The author would like to thank Jonathan Blaney and Jane Winters for reading and commenting on earlier drafts of this article, and to Jonathan Blaney, Jane Winters, Danny Millum, and Martin Steer, of the Institute of Historical Research in London, the home of *British History Online*, and the core project members who were also involved in the redevelopment of the project in 2014 – a process of which this survey was a part. 69

Appendix I: Current search preferences of survey respondents

		Academics		Genealogists		Casual Users		Total	
		#	%	#	%	#	%	#	%
Simple Keyword Search	Never do this	12	3.6	8	1.2	24	7.5	44	3.3
	Rarely	14	4.2	30	4.3	10	3.1	54	4.0
	Sometimes	88	26.5	184	26.4	80	24.8	352	26.0
	Often	137	42.3	282	40.4	117	36.3	536	39.6
	Very Often	81	24.4	194	27.8	91	28.3	366	27.1
Advanced Keyword Search	Never do this	16	5.1	42	6.4	43	14.4	101	8.0
	Rarely	36	11.4	95	14.6	33	11.1	164	13.0
	Sometimes	97	30.8	216	33.1	104	34.9	417	33.0
	Often	118	37.5	201	30.8	68	22.8	387	31.6
	Very Often	48	15.2	98	15.0	50	16.8	196	15.5
Browsing by Category	Never do this	21	6.8	47	7.5	30	10.6	98	8.1
	Rarely	58	18.9	125	20.0	56	19.7	239	19.7
	Sometimes	115	37.5	266	42.6	106	37.3	487	40.1
	Often	88	28.7	143	22.9	67	23.6	298	24.5
	Very Often	25	8.1	44	7.0	25	8.8	94	7.7
Browsing by Publication	Never do this	37	12.2	129	22.1	66	24.1	232	20.0
	Rarely	70	23.1	212	36.3	67	24.5	349	30.1
	Sometimes	105	34.7	164	28.1	84	30.7	353	30.4
	Often	60	19.8	59	10.1	45	16.4	164	14.1
	Very Often	31	10.2	20	3.4	12	4.4	63	5.4
Arrive via External Site	Never do this	124	44.0	249	44.2	131	49.8	504	45.5
	Rarely	69	24.5	135	23.9	69	26.2	273	24.6
	Sometimes	50	17.7	127	22.5	39	14.8	216	19.5
	Often	32	11.4	31	5.5	19	7.2	82	7.4
	Very Often	7	2.5	22	3.9	5	1.9	34	3.1

Table 2.

Appendix II: Rating of future search and browse options by respondents

		Academics		Genealogists		Casual Users		Total	
		#	%	#	%	#	%	#	%
Set advanced search restraints									
Fuzzy keyword searching	I don't understand	7	2.0	22	3.1	20	6.2	49	3.6
	Would never use	4	1.2	24	3.4	10	3.1	38	2.8

	Might be useful	79	23.0	169	23.9	89	27.4	337	24.5
	Moderately useful	42	12.2	110	15.6	44	13.5	196	14.3
	Quite Useful	105	30.6	223	31.6	80	24.6	408	29.7
	I'd really like this	106	30.9	158	22.4	82	25.2	346	25.2
Proximity searching	I don't understand	6	1.8	20	2.9	9	2.8	35	2.6
	Would never use	17	5.1	40	5.8	14	4.4	71	5.3
	Might be useful	66	19.9	162	23.5	79	24.7	307	22.9
	Moderately useful	39	11.8	108	15.7	44	13.8	191	14.2
	Quite Useful	101	30.4	222	32.2	96	30.0	419	31.3
	I'd really like this	103	31.0	137	19.9	78	24.4	318	23.7
Limit search to a subset of the collection									
Limit by content type	I don't understand	5	1.5	8	1.2	7	2.2	20	1.5
	Would never use	24	7.3	56	8.2	20	6.4	100	7.6
	Might be useful	82	24.9	167	24.5	75	23.9	324	24.5
	Moderately useful	45	13.6	124	18.2	54	17.2	223	16.8
	Quite Useful	103	31.2	202	29.7	87	27.7	392	29.6
	I'd really like this	71	21.5	124	18.2	71	22.6	226	20.1
Limit by bibliographical series	I don't understand	2	0.6	9	1.3	4	1.3	15	1.1
	Would never use	21	6.4	88	12.9	26	8.3	135	10.2
	Might be useful	47	14.3	167	24.5	67	21.4	281	21.2
	Moderately useful	55	16.8	142	20.8	48	15.3	245	18.5
	Quite Useful	106	32.3	169	24.7	91	29.1	366	27.6
	I'd really like this	97	29.6	108	15.8	77	24.6	282	21.3
Limit to free content only	I don't understand	13	3.9	25	3.7	23	7.4	61	4.6
	Would never use	58	17.5	55	8.1	25	8.1	138	10.4
	Might be useful	79	23.9	128	18.8	73	23.6	280	21.2
	Moderately useful	66	19.9	106	15.5	46	14.9	218	16.5

	Quite Useful	59	17.8	196	28.7	76	24.6	331	25.0
	I'd really like this	56	16.9	172	25.2	66	21.4	294	22.2
Search by publication metadata									
Search by title	I don't understand	0	0	3	0.4	3	0.9	6	0.4
	Would never use	2	0.6	51	7.4	7	2.2	60	4.4
	Might be useful	44	13.2	143	20.6	49	15.2	236	17.5
	Moderately useful	49	14.7	142	20.5	53	16.4	244	18.1
	Quite Useful	117	35.0	208	30.0	102	31.6	427	31.6
	I'd really like this	122	36.5	146	21.1	109	33.8	377	27.9
Search by publication date	I don't understand	0	0	6	0.9	4	1.3	10	0.8
	Would never use	38	11.5	111	16.1	43	13.4	192	14.3
	Might be useful	116	34.9	254	36.9	116	36.3	486	36.3
	Moderately useful	51	15.4	125	18.2	58	18.1	234	17.5
	Quite Useful	78	23.5	134	19.5	57	17.8	269	20.1
	I'd really like this	49	14.8	58	8.4	42	13.1	149	11.1
Search entities within texts									
Search by person name	I don't understand	1	0.3	0	0	2	0.6	3	0.2
	Would never use	1	0.3	0	0	3	0.9	4	0.3
	Might be useful	20	5.8	17	2.3	13	4.0	50	3.6
	Moderately useful	20	5.8	27	3.7	22	6.7	69	4.9
	Quite Useful	101	29.4	183	25.1	98	30.0	382	27.3
	I'd really like this	201	58.4	501	68.8	189	57.8	891	63.7
Search by place name	I don't understand	2	0.6	0	0	2	0.6	4	0.3
	Would never use	0	0	7	1.0	2	0.6	9	0.7
	Might be useful	19	5.6	24	3.3	12	3.7	55	4.0
	Moderately useful	30	8.9	38	5.3	17	5.2	85	6.1
	Quite Useful	105	31.0	207	28.8	94	28.6	406	29.3
	I'd really like this	183	54.0	444	61.7	202	61.4	829	59.7

Search by geo-coordinates	I don't understand	9	2.7	13	1.9	6	1.9	28	2.1
	Would never use	148	44.6	272	40.1	150	48.2	570	43.1
	Might be useful	98	29.5	237	34.9	85	27.3	420	31.8
	Moderately useful	39	11.8	83	12.2	29	9.3	151	11.4
	Quite Useful	26	7.8	53	7.8	25	8.0	104	7.9
	I'd really like this	12	3.6	21	3.1	16	5.1	49	3.7
Search footnotes only	I don't understand	10	3.0	26	3.9	10	3.2	46	3.5
	Would never use	59	17.6	232	34.8	89	28.7	380	29.0
	Might be useful	123	36.7	207	31.0	108	34.8	438	33.4
	Moderately useful	66	19.7	114	17.1	44	14.2	224	17.1
	Quite Useful	50	14.9	64	9.6	40	12.9	154	11.7
	I'd really like this	27	8.1	24	3.6	19	6.1	70	5.3
Search by date of subject	I don't understand	5	1.5	9	1.3	4	1.3	18	1.3
	Would never use	25	7.4	62	8.9	16	5.0	103	7.6
	Might be useful	61	18.1	181	25.9	65	20.4	307	22.7
	Moderately useful	44	13.1	135	19.3	62	19.4	241	17.8
	Quite Useful	113	33.5	192	27.5	98	30.7	403	29.8
	I'd really like this	89	26.4	119	17.1	74	23.2	282	20.8

Table 3.

Notes

[1]"About British History Online", *British History Online (version 5.0)* https://www.british-history.ac.uk/about.

[2]"British History Online, 31 December 2013", *The Internet Archive* (https://web.archive.org/web/20131231194614/http://www.british-history.ac.uk/)

[3]"About British History Online", *British History Online (version 5.0)* https://www.british-history.ac.uk/about.

[4]The Centre for Metropolitan History, *Institute of Historical Research* (http://www.history.ac.uk/cmh/main); *The History of Parliament* (http://www.historyofparliamentonline.org/).

[5]"Google Analytics", *Google* http://www.google.com/analytics.

[6]"Premium content and premium page scans", *British History Online*. https://www.british-history.ac.uk/premium-content

[7]For more on OCR and historical texts, see Michael Piotrowski, "Chapter 4: Acquiring Historical Texts" in *Natural Language Processing for Historical Texts* (2012), 25-52.

[8]"Apache Solr", version 5.1.0 (April 2015) http://lucene.apache.org/solr/.

[9]"Survey of London", *British History Online* (2015) http://www.british-history.ac.uk/search/series/survey-london; "Calendar of Treasury Books", *British History Online* (2015) http://www.british-history.ac.uk/search/series/cal-treasury-books.

[10]"Browse Catalogue", *British History Online* (2015) http://www.british-history.ac.uk/catalogue.

[11]"Maps", *British History Online* (2015) http://www.british-history.ac.uk/catalogue/maps.

[12]"MARC 21", (1991) http://www.loc.gov/marc/bibliographic/; "Dublin Core Metadata Initiative" (1995-2015) http://dublincore.org/; "Context Objects in Spans", (2005-2009) http://ocoins.info/; "Metadata Object Description Schema", (2015) http://www.loc.gov/standards/mods/.

Works Cited

Agichtein et al. 2006a Agichtein, E., Brill, E., and Dumais, S. (2006) "Improving web search ranking by incorporating user behavior information". *SIGIR '06*, 19-26.

Agichtein et al. 2006b Agichtein, E., Brill, E., Dumais, S., Ragno, R. (2006) "Learning User Interaction Models for Predicting Web Search Results Preferences". *SIGIR '06 Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, 3-10.

Alexander et al. 2015 Alexander, M., Anderson, J., Archer, D., Baron, Al., Hope, J., Jeffries, L., Kay, C., Rayson, P., Walker, B. (2014) "SAMUELS (Semantic Annotation and Mark-Up for Enhancing Lexical Searches)". *School of Critical Studies, University of Glasgow.* Available from http://www.gla.ac.uk/schools/critical/research/fundedresearchprojects/samuels/.

Anderson 2004 Anderson, I.G. (2004) "Are you being served? Historians and the search for primary sources". *Archivaria*, 58, 81-130.

Anon 1999 *Eighteenth Century Collections Online*. http://quod.lib.umich.edu/e/ecco/

Anon 2000 *Early English Books Online*. http://eebo.chadwyck.com/home

Anon 2007 *Nineteenth Century Newspapers Online*. http://www.bl.uk/reshelp/findhelprestype/news/newspdigproj/database/

Anon 2009 *17th and 18th Century Burney Collection Newspapers*. http://www.bl.uk/reshelp/findhelprestype/news/newspdigproj/burney/

Apache Solr Apache Solr, version 5.1.0 (April 2015) http://lucene.apache.org/solr/.

Aula 2005 Aula, A. (2005) "User Study on Older Adults' Use of the Web and Search Engines". *Universal Access in the Information Society,* 4(1), 67-81.

Badke 2011 Badke, W. (2011) "The treachery of keywords". *Online*, 35(3), 52-54.

Barclay 2010 Barclay, D.A. (2010) "The Myth of Browsing". *American Libraries*, 41(6/7), 52-54.

Bates 1996 Bates, M.J. (1996) "The Getty End-User Online Searching Project in the Humanities: Report No. 6: Overview and Conclusions". *College and Research Libraries*, 57(6), 514-523.

Bates 2002 Bates, M.J. (2002) "The cascade of interactions in the digital library interface". *Information Processing and Management,* 38, 381-400.

Beall 2011 Beall, J. (2011) "The weakness of full-text searching". *Journal of Academic Librarianship*, 34(5), 438-444.

Blaney 2013 Blaney, J. (2013) "British History Online: assessing the impact of a successful resource". *Digital Humanities@Oxford Summer School 2013*, pages 1-6. Retrieved from http://digital.humanities.ox.ac.uk/dhoxss/2013/materials/cc-blaney_bho-assessing.pdf.

Boulton and Schwarz 2007 Boulton J., Schwarz L. (2007) "St. Martin-in-the-Fields Workhouse Admission and Discharge Registers, 1725-1819". *Pauper Lives Project*. Available from http://www.londonlives.org/static/SMDSWHR.jsp

British History Online *British History Online (version 5.0)* https://www.british-history.ac.uk/about.

Burns et al. 1999-2016 Burns, A., Fincham, K., Taylor, S., Yorke, P, Clayton. M., Wales, T. *Clergy of the Church of England Database* http://theclergydatabase.org.uk.

Calendar of Treasury Books "Calendar of Treasury Books (1904-1962)". *British History Online*. Available from http://www.british-history.ac.uk/search/series/cal-treasury-books.

Chadwick-Dias et al. 2007 Chadwick-Dias, A., Bergel, M., and Tullis, T.S. (2007) "Senior Surfers 2.0: A Re-examination of the Older Web User and the Dynamic Web". *Universal Access in Human Computer Interaction: Coping with Diversity*, 4554, 868-876.

Champlin 1985 Champlin, P. (1985) "The Online Search: Some Perils and Pitfalls". *RQ*, 25(2), 213-217.

Cohen 2010 Cohen, D. (2010) "Is Google good for history?" *Dancohen.org*. Retrieved from http://www.dancohen.org/2010/01/07/is-google-good-for-history.

Collins et al. 2012 Collins, E., Bulger, M.E., and Meyer E.T., (2012) "Discipline matters: technology use in the humanities". *Arts and Humanities in Higher Education*, 11(1-2), 76-92.

Crymble 2015 Crymble, A. (2015) "A Comparative Approach to Identifying the Irish in Long Eighteenth Century London". *Historical Methods*, 48(3).

Crymble et al. 2015 Crymble, A., Falcini, L., Hitchcock, T. (2015) "Vagrant Lives: 14,789 Vagrants Processed by the County of Middlesex, 1777-1786". *Journal of Open Humanities Data*. 1, p.e1 DOI: http://doi.org/10.5334/johd.1

Dalton and Charnigo 2004 Dalton, M.S. and Charnigo, L. (2004) "Historians and their information sources". *College and Research Libraries*, 65(5), 400-425.

Darin and Kurnaiwan 2000 Darin, E.R. and Kurniawan, S.H. (2000) "Increasing the Usability of Online Information for Older Users: A Case Study in Participatory Design". *International Journal of Human-Computer Interaction*, 12(2), 263-276.

Dickinson et al. 2007 Dickinson, A., Smith, M., Arnott, J., Newell, A., and Hill, R. (2007) "Approaches to Web Search and Navigation for Older Computer Novices". *CHI Proceedings*, 281-290.

Donnelly et al. 2002 Donnelly S., Etherton J., Shaw, C., Ferris C., Morrison, A. (2002), *Charles Booth Online Archive* http://booth.lse.ac.uk/static/d/index.html

Duff and Cherry 2001 Duff, W.M. and Cherry, J.M. (2001) "Use of historical documents in a digital world: comparisons with original materials and microfiche". *Information Research*, 6(1).

Eco 2004 Eco, U. (2004) *The Name of the Rose*, (Vintage).

Fairweather 2008 Fairweather, P.G. (2008) "How Older and Younger Adults Differ in their Approach to Problem Solving on a Complex Website". *Proceedings of the 10th Annual ACM SIGACCESS Conference on Computers and Accessibility*, 67-72.

Freire et al. 2012 Freire, N., Borbinha, J., and Calado, P. (2012) "An Approach for Named Entity Recognition in Poorly Structured Data". *The Semantic Web: Research and Applications, Lecture Notes in Computer Science*, 7295, 718-732.

Gao et al. 2007 Gao, Q., Sato, H., Rau, P.P., and Asano, Y. (2007) "Design Effective Navigation Tools for Older Web Users". *Human-Computer Interaction: Interaction Design and Usability*, 4550, 765-773.

Google Analytics "Google Analytics", *Google*. http://www.google.com/analytics.

Grace-Martin and Gay 2001 Grace-Martin, M. and Gay, G. (2001) "Web Browsing, Mobile Computing and Academic Performance". *Educational Technology & Society*, 4(3), 95-107.

Granka et al. 2004 Granka, L.A., Joachims, T., and Gay, G. (2004) "Eye-tracking analysis of user behavior in WWW search". *SIGIR '04, Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, 478-479.

Grey and Hurko 2012 Grey, A. and Hurko, C.R. (2012) "So you think you're an expert: keyword searching vs. controlled subject headings". *Codex*, 1(4), 15-26.

Hamburger 2004 Hamburger, S. (2004) "How researchers search for manuscripts and archival collections". *Journal of Archival Organization*, 2(1-2), 79-102.

Harley and Henke 2007 Harley, D. and Henke, J. (2007) "Towards an Effective Understanding of Website Users: Advantages and Pitfalls of Linking Transaction Log Analyses and Online Surveys". *D-Lib Magazine*, 13(3/4).

Hart 1971 Hart, M.S. (1971) *Project Gutenberg*, http://www.gutenberg.org/wiki/Main_Page

Henry VIII 1528 *Henry VIII* (1528) July 1528, 1-10. In *Letters and Papers, Foreign and Domestic, Henry VIII, Volume 4, 1524-1530*, ed. J S Brewer (London, 1875), 1948-1965. Available from http://www.british-history.ac.uk/letters-papers-hen8/vol4/pp1948-1965 [accessed 9 May 2015].

Hill et al. 2011 Hill, R.L., Dickinson, A., Arnott, J.L., Gregor, P., and McIver, L. (2011) "Older Web Users' Eye Movements: Experience Counts". *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 11, 1151-1160.

Hitchcock 2008 Hitchcock, T. (2008) "Digitising British history since 1980". *Making History: The changing face of the profession in Britain*. Available at http://www.history.ac.uk/makinghistory/resources/articles/digitisation_of_history.html

Hitchcock et al. 2002 Hitchcock, T., Shoemaker, R., Emsley, C., Howard, S., McLaughlin, J., et al. (2002), http://www.oldbaileyonline.org

Hitchcock et al. 2010 Hitchcock, T., Shoemaker, R., Howard, S., McLaughlin, J., et. al. (2010), http://www.londonlives.org

Hitchcock et al. 2011 Hitchcock, T., Shoemaker, R., Winters, J., McLaughlin, J., Rogers, K., Howard, S. et al. (2011) "Connected Histories". Available at http://www.connectedhistories.org/

Holley 2007 Holley, R. (2007). "Optical Character Recognition (OCR) on Newspapers – An Overview. Version 1.0". (National Library of Australia).

Howard 2016 Howard S. (2016). *The London Lives Petitions Project: a dataset of eighteenth-century petitions to London magistrates*, version 1.2 http://sharonhoward.github.io/llpp/

Hsu et al. 2014 Hsu, C.-Y., Tsai, M.-J., Hou, H.-T., and Tsai, C.-C. (2014) "Epistemic beliefs, online search strategies, and behavioral patterns while exploring socioscientific issues". *Journal of Science Education and Technology*, 23, 471-480.

Jansen and Spink 2006 Jansen B.J., and Spink, A. (2006) "How are we searching the World Wide Web? A comparison of nine search engine transaction logs". *Information Processing & Management*, 42(1), 248-263.

Joachims et al. 2007 Joachims, T., Granka, L., Pan, B., Hembrooke, H., Radlinski, F., and Gay, G. (2007) *ACM Transactions on Information Systems*, 25(2), article no. 7.

Kiernan et al. 1999 Kiernan, K.S., Szarmach, Prescott, A., Solopova, E., French, D., Cantara, L., Ellis, M., and Yuan, C.J. (1999) *Electronic Beowulf 1.0*.

Kurnaiwan and Zaphiris 2005 Kurnaiwan, S. and Zaphiris, P. (2005) "Research-Derived Web Design Guidelines for Older People". *Proceedings of the 7th International ACM SIGACCESS Conference on Computers and Accessibility*, 129-135.

Kurnaiwan et al. 2006 Kurnaiwan, S.H., King, A., Evans, D.G., and Blenkhorn, P.L. (2006) "Personalising Web Page Presentation for Older People". *Interacting with Computers*, 18(3), 457-477.

MaKinster et al. 2002 MaKinster, J.G., Beghetto, R.A., and Plucker, J.A. (2002) "Why Can't I Find Newton's Third Law? Case Studies of Students' Use of the Web as a Science Resource". *Journal of Science Education and Technology*, 11(2), 155-172.

Maxwell 2010 Maxwell, A. (2010) "Digital archives and history research: feedback from an end-user". *Library Review*, 59(1), 24-39.

Millward 2003 Millward, P. (2003) "The 'Grey Digital Divide': Perception, Exclusion and Barriers of Access to the Internet for Older People". *First Monday*, 8(7).

Morrell et al. 2000 Morrell, R.W., Mayhorn, C.B., and Bennett, J. (2000) "A survey of World Wide Web use in middle-aged and older adults". *Human Factors*, 42(2), 175-182.

Morse 1970 Morse, P.T. (1970) "Search Theory and Browsing". *The Library Quarterly*, 40(4), 391-408.

Mott 2010 Mott, R. (2010) "Serendipity through Browsing". *American Libraries*, 41(8), 6.

Myka and Güntzer 1996 Myka, A. and Güntzer, U. (1996) "Fully Full-Text Searches in OCR Databases". In *Digital Libraries Research and Technology Advances: Lecture Notes in Computer Science*, 1082, 131-145.

Piotrowski 2012 Piotrowski, M. (2012) "Chapter 4: Acquiring Historical Texts". In *Natural Language Processing for Historical Texts*, 25-52.

Ramsay 2010 Ramsay, S. (2010) "The Hermeneutics of Screwing Around". In *Pastplay: Teaching and Learning History with Technology*, ed. Kevin Kee (University of Michigan Press), 111-120.

Rares and Chen 2009 Rares, V. and Chen, L. (2009) "Efficient top-k algorithms for fuzzy search in string collections". *KEYS '09 Proceedings of the First International Workshop on Keyword Search on Structured Data* (New York), 9-14.

Russell 1918 Russell, R.C. (1918) "US Patent 1,261,167", (April 2).

Samuelson 1948 Samuelson, P.A. (1948) "Consumption Theory in Terms of Revealed Preference". *Economica*, 15(60), 243-253.

Schatz 1997 Schatz, B.R. (1997) "Information retrieval in digital libraries: bringing search to the net". *Science*, 275, 327-334.

Schenkel et al 2007 Schenkel, R., Boschart, A., Hwang, S., Theobald, M., and Weikum, G., (2007) "Efficient Text Proximity Search". *String Processing and Information Retrieval: Lecture Notes in Computer Science*, 4726, 287-299.

Shera 1964 Shera, J. (1964) "Automation and the Reference Librarian". *RQ*, 3(6), 3-7.

Survey of London (1894-2015) "Survey of London (1894-2015)", *British History Online*. Available from http://www.british-history.ac.uk/search/series/survey-london.

The Centre for Metropolitan History "The Centre for Metropolitan History", *Institute of Historical Research*. http://www.history.ac.uk/cmh/main

The History of Parliament "The History of Parliament". http://www.historyofparliamentonline.org/.

Titford 2005 Titford, J. (2005) "Kedgley of north London: a case study". *Family Tree Magazine*, (December), 63-65.

Tsai et al. 2012 Tsai, M.-J., Hsu, C.-Y., and Tsai, C.-C. (2012) "Investigation of High School Students' Online Science Information Searching Performance: The Role of Implicit and Explicit Strategies". *Journal of Science Education and Technology*, 21, 246-254.

Tullis 2007 Tullis, T.S. (2007) "Older Adults and the Web: Lessons Learned from Eye-Tracking". *Universal Access in Human Computer Interaction: Coping with Diversity*, 4554, 1030-1039.

Walford 1878 Walford, E. (1878) "Rotherhithe". In *Old and New London: Volume 6* (London), 134-142. Available from http://www.british-history.ac.uk/old-new-london/vol6/pp134-142 [accessed 8 May 2015].

Warwick et al. 2008 Warwick, C., Terras, M., Galina, I., Huntington, P., and Pappa, N. (2008) "Library and information resources and users of digital resources in the humanities". *Program*, 42(1), 5-27.

Zajicek 2001 Zajicek, M. (2001) "Interface Design for Older Adults". *Proceedings of the 2001 EC/NSF workshop on Universal Accessibility of Computing*, 60-65.

van Hooland et al. 2013 van Hooland, S., De Wilde, M., Verborgh, R., Steiner, T., and Van de Walle, R. (2013) "Exploring entity recognition and disambiguation for cultural heritage collections". *Digital Scholarship in the Humanities*. DOI: http://dx.doi.org/10.1093/llc/fqt067.