Journal of Documentation

Exploring the information behaviour of users of Welsh Newspapers Online through web log analysis

Paul Gooding

Article information:

To cite this document:
Paul Gooding , (2016),"Exploring the information behaviour of users of Welsh Newspapers Online through web log analysis", Journal of Documentation, Vol. 72 Iss 2 pp. 232 - 246
Permanent link to this document:
http://dx.doi.org/10.1108/JD-10-2014-0149

Downloaded on: 10 November 2016, At: 20:30 (PT)
References: this document contains references to 39 other documents.
To copy this document: permissions@emeraldinsight.com
The fulltext of this document has been downloaded 263 times since 2016*

Received 15 October 2014
Revised 16 July 2015
Accepted 24 August 2015

Exploring the information behaviour of users of Welsh Newspapers Online through web log analysis

Paul Gooding
University of East Anglia, Norwich, UK

Abstract

Purpose – Webometric techniques have been applied to many websites and online resources, especially since the launch of Google Analytics (GA). To date, though, there has been little consideration of information behaviour in relation to digitised newspaper collections. The purpose of this paper is to address a perceived gap in the literature by providing an account of user behaviour in the newly launched Welsh Newspapers Online (WNO).
Design/methodology/approach – The author collected webometric data for WNO using GA and web server content logs. These were analysed to identify patterns of engagement and user behaviour, which were then considered in relation to existing information behaviour.
Findings – Use of WNO, while reminiscent of archival information seeking, can be understood as centring on the web interface rather than the digitised material. In comparison to general web browsing, users are much more deeply engaged with the resource. This engagement incorporates reading online, but users' information seeking utilises website search and browsing functionality rather than filtering in newspaper material. Information seeking in digitised newspapers resembles the model of the "user" more closely than that of the "reader", a value-laden distinction which needs further unpacking.
Research limitations/implications – While the behaviour discussed in this paper is likely to be more widely representative, a larger longitudinal data set would increase the study's significance. Additionally, the methodology of this paper can only tell us what users are doing, and further research is needed to identify the drivers for this behaviour.
Originality/value – This study provides important insights into the underinvestigated area of digitised newspaper collections, and shows the importance of webometric methods in analysing online user behaviour.

Keywords Digital libraries, User studies, Digitization, National libraries
Paper type Case study

1. Introduction

Web log analysis has been utilised in Library and Information Science to analyse online user behaviour for websites (Nicholas *et al.*, 2000), e-journals (Yu and Apps, 2000) and digital resources (Warwick *et al.*, 2008; Meyer *et al.*, 2009). To date, though, it has not been used to analyse user behaviour in digitised newspaper archives. This paper will address this gap, presenting data from a case study into user behaviour with Welsh Newspapers Online (WNO)[1]. WNO is a free online digitised newspaper collection which contained, at the point of this study, 725,000 digitised newspaper pages. It aims to make over one million newspaper pages available from the NLW's own collections. The resource is part-funded by the Strategic Capital Investment Fund[2] and the European Regional Development Fund[3] through the Welsh Government, and created

The author is grateful to the National Library of Wales for providing access to their staff, collections and data sets, and for their encouragement and support in this research.

Journal of Documentation
Vol. 72 No. 2, 2016
pp. 232-246
© Emerald Group Publishing Limited
0022-0418
DOI 10.1108/JD-10-2014-0149

in-house at the NLW following investment in a specialist digitisation studio (National Library of Wales, 2013). It was greeted positively on its launch due to its open approach and site functionality (Tanner, 2013; The Digital Victorianist, 2013).

This paper will focus on the insights into online user behaviour in this collection, utilising webometric data sets covering a period of three months from the launch of WNO in March 2013. In doing so, it will answer two specific questions:

(1) What can we learn about users of WNO through the use of Google Analytics (GA) and web log analysis?

(2) How does this recorded behaviour differ in nature from information seeking in a physical archival context?

First, the paper explores the context for online user behaviour through a review of the scholarly and mainstream literature about online user behaviour. It explains the role of webometrics in exploring these issues in LIS research, and the significant, but extremely different, insights that GA and web log analysis provide. It then utilises data from web logs and GA to explore user behaviour with WNO, demonstrating that, while user engagement in WNO is high, usage of digitised newspaper resources online appears more reminiscent of information seeking behaviour in physical archives. Online, this multifaceted information behaviour centres upon the web interface rather than the archival material, requiring careful consideration of the role of interfaces in shaping user encounters with heritage materials on the web.

2. Literature review

The first section of this review explores how mainstream accounts of the opposition between physical and digital texts are based on a narrow interpretation of information behaviour which prioritises reading over other forms of interaction. The second places where webometric research fits within LIS, and interrogates some of the strengths and flaws of the chosen methods for this research task.

2.1 Newspaper digitisation and the cult of deep reading

The critical debate around newspaper digitisation has focused upon the impact of web technologies at a textual level. Mussell (2012) notes that the trend towards article-level representation in digitised newspaper collections foregrounds the partial textual manuscript, even though the article is just one element of a larger textual artefact. Users of digitised newspapers may therefore lose the original context of the material by viewing it online. Brake sounds a note of caution about the negative impact of this shift, noting that creating a digital representation of a historical newspaper "denaturalizes it and transforms the reader […] into a user [author's emphasis] who sees the content inextricably embedded in the matrix of newspaper pages" (Brake, 2012). Her concerns are intimately linked to the perception that digital technologies more generally could have a detrimental impact on intellectual capabilities. In *The Gutenberg Elegies*, Sven Birkets (1994, pp. 3-20) mourns a shift away from deep reading of printed texts caused by digital technologies. He suggests that users are naturally more deeply engaged with physical texts than digital texts. In doing so, he imagines reading as a deep, sustained and intellectually rewarding engagement with individual texts. For Birkets, digital texts not only demand a new form of interaction, but actively erode the reader's capability for deep reading.

In the 20 years since his influential critique, the trope that digital media negatively impact our attentional ability has become an unsubstantiated truism. Accounts can draw on anecdotal experiences of the seemingly insidious effects of screen-based media, often expressed in overwrought terms:

> I'm not thinking the way I used to think. I feel it most strongly when I'm reading. I used to find it easy to immerse myself in a book or a lengthy article […] Now my concentration starts to drift after a page or two. I get fidgety, lose the thread, begin looking for something else to do. I feel like I'm always dragging my wayward brain back to the text (Carr, 2010, pp. 5-6).

Carr sets up the same oppositional relationship as Birkets; physical texts are inherently deep and intellectually rewarding, while digital forms are fundamentally diminishing. This is reinforced in other accounts. Edwards complains about the digitisation of the Codex Sinaiticus, concluding from the increased accessibility of the manuscripts to the untrained public that "if my audience analysis is even broadly correct, the British Library is investing heavily not in scholarship, but in a new branch of the entertainment industry" (Edwards, 2013). His complaint recalls Walter Benjamin (2007), bemoaning the destruction of the "aura" of the original through increase exposure to the masses.

These attacks on digital technologies carry an intrinsic defence of existing intellectual practices; or rather, they defend an idealised version of these practices which flattens complex human behaviours into a binary opposition between physical texts and digital media. Returning to newspapers, Brake's contrast between the value-laden concepts of "reader" and "user" are vital. Yet Birkets and others have effectively created a false sense of the reader, which inevitably frames online interactions negatively. In reality, however, information behaviour is more complex and renders the rivalry between "reader" and "user" flawed. As Levy (1997, p. 209) notes, digital resources are a locus for search, acquisition and reading, and thus all activities are likely to be exhibited simultaneously. Archival research does not follow a simple progression of tasks, but in fact resembles a series of activities that can occur at any time:

> Choosing and refining topics, planning and conducting studies, gathering and interpreting evidence, and writing and revising manuscripts can go on concurrently, both within and across individual projects (Case, 1991, p. 79).

This pattern of archival research is reflected in the online behaviour of users of WNO, as we will see in Section 4.2. The closest corollary with the search-based interface of the resource is with traditional archival finding aids, and web analytics provide important insights into how this applies to users of digitised newspaper collections.

2.2 Web analytics in Library and Information Science

The concept of web analytics is defined as "the measurement, collection, analysis and reporting of internet data for the purpose of understanding and optimizing web usage" (Digital Analytics Association, 2012). It is established in LIS as a method for tracking the impact of web-based resources under the wider heading of "webometrics". One obvious contemporary application is for research into web-based phenomena where other methods may prove inadequate due to remote, poorly defined populations (Thelwall, 2009, p. 1). As Nicholas *et al.* (2004, p. 24) describe, webometrics provide a "direct and immediately available record of what people have done: not what they say they might, or would do; not what they were prompted to say; not what they thought they did". Early webometric techniques necessitated specific expertise in quantitative data analysis, but the launch of GA in 2005 provided an accessible tool for gathering usage and engagement statistics for any website.

Webometric analysis has resultantly become accessible to a wider audience, and GA is a common research tool in LIS studies which evaluate the impact of websites and digital resources[4].

However, studies which utilise GA are often conceptualised narrowly without consideration of other contributors to user behaviour. A limitation of webometric analysis is that it can only reveal how a website is used, and not why (Zuccala and Thelwall, 2006). For this reason, GA in particular can lead to a superficial, metric-driven understanding of user behaviour. Instead, deeper research into online user behaviour in LIS is often done through web log analysis[5]. Although harder to gather and analyse, web logs provide notable benefits as they record all users of a website in a format which can be manipulated by the researcher. Web servers automatically record basic information about each request they receive, including user identifiers, date and time of interaction and the type of content viewed. This allows for deeper analysis of user behaviour which is relatively unobtrusive.

Limitations of webometrics. There are, though, a number of problems with web log analysis. These include difficulties in generating web metrics due to robot traffic[6], and a lack of information about when users leave the website. Additionally, user identification is unreliable because IP numbers can only be traced back to a specific machine, not an individual. The use of proxy servers[7] and point-to-point connections also mean that IP addresses cannot be reliably assumed to relate to use even on a specific machine. This makes tracking return users difficult and, combined with its data-intensive nature, web log analysis is a more complex and time-consuming undertaking than GA. The work involved in collecting and storing web logs means that some scholars report problems in accessing them for research purposes (Warwick *et al.*, 2008; Meyer *et al.*, 2009).

GA, on the other hand, provides a powerful business analytics platform which is not tailored for academic research. While increasingly ubiquitous, it has some flaws as a research tool. Its default data bandings are often inappropriate for websites with high levels of engagement. This is exacerbated by the lack of raw, exportable data available to the user. Raw data is hidden from GA users for reasons including data privacy, resulting in a lack of transparent, reproducible data that would make it a truly essential academic resource. This opacity is particularly problematic when dealing with samples of data, which are automatically processed from the inaccessible data set, so researchers are left to trust the representativeness and reproducibility of results. Finally, tracking what happens when a user goes offline remains beyond webometrics. For this reason, it is vital to consider other methods to fill the gaps that webometric analysis leaves, thereby placing webometric data in a qualitative context.

With this in mind, GA provides an adequate replacement for web log analysis when gathering usage and engagement statistics, information on social media visitors and technical and demographic information. However, it provides a weaker source for deep analysis of user behaviour, with implications for the transparency and reproducibility of research data sets. Despite these concerns, GA provides a more robust platform than other webometric techniques which rely on external data sources (Thelwall, 2009, p. 125); for this reason, this study utilises GA to provide baseline usage and engagement metrics, which are then enriched by deeper insights from web log analysis. While this work was undertaken as part of a larger mixed methods study into information behaviour with digitised newspaper collections, the scope of the article is more constrained: it explores the insights into user behaviour which can be gained from webometric approaches.

3. Methodology

The NLW provided the author with two data sets for WNO; GA, gathered and analysed through the GA web platform; and anonymised web logs. Both data sets covered nearly four months starting from the launch date of the resource, from 12 March 2013 to 30 June 2013. This section outlines the methodology for analysing each data set. As the literature review indicates, this study used GA data to provide overall usage and engagement statistics, while more in-depth insights into user behaviour came from analysis of the web logs. The results section is therefore split into sections which maintain this distinction.

3.1 GA

The first data set consisted of GA data harvested from the WNO analytics account, although the underlying data set was inaccessible. One key difference between WNO and other resources is that, in common with all NLW outputs, it is published bilingually in Welsh and English. The resource therefore comprises two structurally identical websites which differ only in language, and each has a separate GA account. Data from both accounts was collated in Excel to facilitate analysis, after separate evaluation to confirm user data were consistent across both websites. The following usage metrics were collected from GA: visitor numbers; user engagement by page visit and visit duration; bounce rate; and mobile and social media usage.

3.2 Web log analysis

The second data source was a set of anonymised processed content logs. These web logs specifically record information about user behaviour on the site, and as a result, they only represent the content-related portion of each user's journey. The logs therefore track the following: searches undertaken by users on the website (henceforth referred to as search queries); instances where users have browsed, filtered or otherwise interacted with search results (search result queries); and instances where users have viewed content (content queries). The web logs record each of these interactions as a single line of plain code text in a file held on the website servers. The following example is a content query, as displayed in the logs:

```
2013-06-02T12:26:50+01: 00 51a5c97c3c8d3 llgc-id:3036868 llgc-id:3039814 llgc-id:3037695
Aberystwyth Observer 21 September 1872[2] ART40.
```

The elements are, in order: date and time of interaction; unique user ID; server ID numbers for website content; title of newspaper viewed; date of newspaper edition; page number viewed; article number on the viewed page. Search queries contain an additional field for input search terms, and search results queries include a field recording the interaction with search results.

This rich data source allowed several metrics relating to user behaviour on WNO to be assessed, including: most viewed newspaper titles; most viewed decades; most commonly viewed page numbers in newspapers; average number of pageviews per visit; and average number of pageviews involving each query category outlined above. Where relevant, results have been presented as a proportion of the total newspaper pages in a given time period, recorded at the time of data analysis (July 2013). While this accounts for the uneven spread of newspaper material across time, it does not indicate what proportion of the material was actually viewed as it does not discern between duplicate views.

4. Results

The results section is presented in two sections, split by data source. The GA section provides an overview of website usage and some insights into engagement from mobile devices and social media, which contextualise the following section. The web logs section provides more detailed insights, and therefore provide a deeper account of user behaviour with WNO.

4.1 Results: GA

Table I shows visitor metrics for the English and Welsh versions of WNO, and collated statistics for both sites.

In structural terms, the websites are identical, and there is little indication of user behaviour varying according to language. They both show a reasonably high volume of traffic, indicating that the resource is already highly visible to interested communities. Visits to WNO are dominated by users from the UK (84.76 per cent). Considering only UK visits, Wales is overrepresented in comparison to its population size, accounting for 37.21 per cent of UK users and 30.98 per cent of all visits, compared to just 4.8 per cent of the total UK population (Office for National Statistics, 2014). The only other nations to account for over 1 per cent of traffic are Australia (5.95 per cent), the USA (3.65 per cent), Canada (2.53 per cent) and New Zealand (1.37 per cent). The majority of users are therefore based in nations with historic or linguistic ties to the UK.

There is also a reasonably deep level of engagement with WNO. The bounce rate[8], for instance is low in comparison to other reported sources (Batra, 2008; Betty, 2009) despite the open access nature of the resource allowing for short, curiosity-driven visits. In addition, roughly 32 per cent of visitors view 20 or more pages per visit, and the average number of pageviews in this group is 55.79, a deep engagement reflected in the visit duration statistics where those visiting for at least 1,801 seconds view an average of 68.79 pages per visit. This represents a significant investment of time and effort, which goes far beyond that associated with many websites: as such, we can surmise that use of a digital resource of this nature is deeper and more sustained than for general web browsing. Section 4.2 will explore the nature of this user behaviour.

Mobile traffic and social media. Mobile devices are an important traffic source for WNO, accounting for 10.79 per cent of all traffic. However, as Table II shows, mobile visits are shorter (00:10:03) than non-mobile visits (00:18:31), and mobile users view significantly fewer pages on average (13.97 pages) than non-mobile users (22.58 pages).

WNO utilise HTML5 to provide an adaptive website layout which adapts automatically to screen size, but there is no simple solution for presenting large-format digitised material on smaller mobile screens. Indeed, even ubiquitous features such as

Table I. Metrics for all traffic to Welsh Newspapers Online

Site	Total visits	Unique visitors	Pageviews	Pages/ visit	Avg. visit duration	Bounce rate (%)	% of new visits
English site only	34,898	13,944	765,356	21.93	00:17:41	22.63	34.75
Welsh site only	17,869	5,861	377,021	21.1	00:17:25	24.15	28.35
Total across both sites	52,767	19,805	1,142,377	21.65	00:17:36	23.14	32.58

Note: The high-engagement levels, as indicated by the metrics for pages/visit and averaged visit duration search pose greater difficulty on mobile devices. The pattern of user engagement correlates closely with screen size: desktop and laptop users exhibit the highest engagement levels, followed by tablet users and lastly mobile users. Addressing the extent to which this is structural, or a reflection of engagement being driven by the tasks which are attempted on different devices, is an important questions that warrant further attention in future.

Social media referrals account for 4,639 visits (8.8 per cent) to WNO, and they exhibit low-engagement levels most similar to mobile users. The exception is traffic referred from blogging platforms: referrals from WordPress and Blogger exhibit much higher engagement than other social media sources (Table III). We would note, though, that only the top four sources returned over 100 visits in our sample:

Overall it appears that an engaged user community has spent significant time using the resource. With the exception of mobile and social media traffic, this is well beyond the engagement levels exhibited by general web browsers. We can therefore say with some confidence that usage is more likely to represent information seeking behaviour than curiosity-driven web browsing, and that many users are utilising the resource for specific knowledge attainment or research tasks. The following section provides insights into how this behaviour is manifested.

4.2 *Results: web log analysis*

This section presents the findings from an analysis of over 300,000 separate page impressions recorded in web logs for the Welsh and English language versions of WNO. During the time of this study, the 1840s and 1850s proved the most popular

Table II. Comparison of engagement for mobile and non-mobile visits to Welsh Newspapers Online (52,767 visits)

Mobile (including tablets)	Visits	Pages/visit (%)	Avg. visit duration	% of new visitors	Bounce rate (%)
No	47,076	89.21	22.58	00:18:31	31.62
Yes	5,691	10.79	13.97	00:10:03	42.45

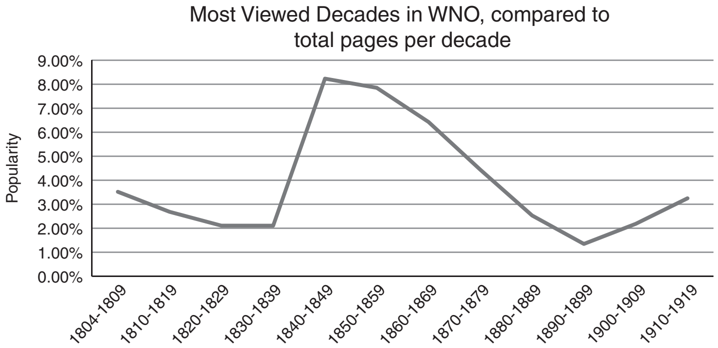
Note: Engagement metrics are considerably lower for visits from mobile devices

Table III. Referral statistics for visits to Welsh Newspapers Online from social networks (4,639 visits)

Total	Visits	Pageviews	Avg. visit duration	Pages/visit
Facebook	2,320	36,971	00:13:32	15.94
Twitter	1,223	8,931	00:05:16	7.30
WordPress	575	15,126	00:21:50	26.30
Blogger	398	7,826	00:14:41	19.66
Ravelry	55	193	00:01:19	3.51
tinyURL	20	259	00:06:17	12.95
Hootsuite	19	125	00:00:09	2.63
Flickr	19	272	03:07:00	14.31
Google+	6	16	00:06:55	2.67
Netvibes	4	14	00:01:15	5.50
Total	4,639	69,733	00:12:56	15.02

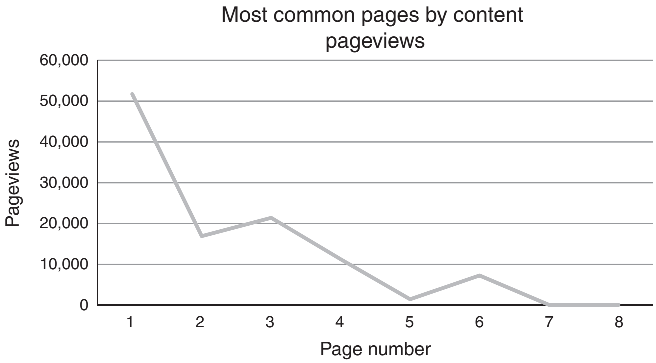
Note: Referrals from blogging platforms exhibited higher levels of engagement than other social traffic period for researchers (Figure 1). It is difficult to say why; it may be caused by specific heavy users having a particular interest in this period, or it may correlate with a period of interest in Welsh history which saw the beginning of heavy industrialisation and a growing population (Davies, 1994, pp. 366-391). A longer analysis period would help to identify whether this is a permanent or short-term trend, and could help to identify priority periods for future digitisation.

Figure 2 evaluates which pages users view, judged by their position in the physical edition: page 1 represents the title page, and the highest page number was eight. Although the collection does contain newspapers with 16 pages, there were an average of 6.4 pages per edition across. We found that users view the title page more than any other by a large margin. This is likely to reflect the way that the formal significance of the front page has been strengthened by the browsing interface: users accessing newspapers by browsing, for instance, will be taken to the title page by default, with no guarantee that they will browse further pages in the edition.



Notes: There is a peak in popularity in the 1840s and 1850s, and a notable rise in the period directly leading to First World War. Results are represented as a percentage of total pages for each decade

Figure 1. Relative popularity of decades in Welsh Newspapers Online



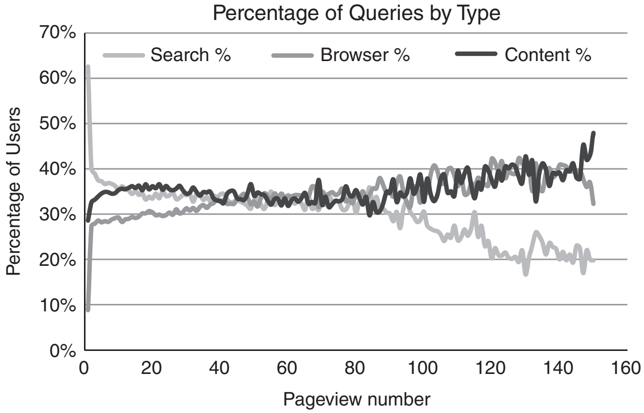
Note: Page number one is the front page of each newspaper, and is viewed far more frequently than all other pages

Figure 2. Most commonly viewed page number in newspaper titles

What is clear is that users do not view later pages as frequently as the title page. We interpret these findings as evidence that users engage differently with newspapers online, indicating a reliance on web technologies rather than manual browsing of material for discovery. This reliance on automated filtering tools is inevitable in a large-scale resource, where users must automate elements of the research process to make the most of the collection. The fact that users do not seem to browse through editions sequentially in no way suggests an impoverishment of attention, not least because we are viewing usage which combines search, browsing and reading in one web platform.

The following chart shows how this information seeking takes place in WNO. It shows the proportion of users engaged in search, search result or content queries at any pageview in their visit to WNO, with the pageview number ascertained by access times for each unique user ID (Figure 3).

We can see that content queries remain an important element of user activity regardless of the visit duration. In fact, once a visit incorporates over 100 pageviews, users view increasingly large amounts of newspaper content. By contrast, the longer a user spends on the website, the less likely they are to be engaged in searching. Instead, search result queries replace search queries, indicating that the average user becomes less reliant on searching the longer their visit, slowly moving towards browsing search results or viewing content. Effectively, the percentage of users performing each query category is indicative of general patterns of user behaviour. Despite the above observations, over 50 per cent of users are engaged with search or search result queries at any point: thus over half of all pageviews are dedicated to interacting with the web interface rather than the historical sources. This leads to two important observations: first, that users interact with WNO extensively while engaging with large amounts of content, in a manner reminiscent of the multifaceted information seeking behaviour of researchers in archives; and second, that while this engagement is deep, it occurs primarily with the web interface and secondarily with the material. Thus, while



Note: This shows the filtering process in action, moving from search to browsing and content views as users spend longer on the website

Figure 3.
Percentage of queries in Welsh Newspapers Online by type of query

users of physical resources must engage directly with the textual object for filtering, users of online resources rely primarily on automated processes facilitated by the web interface.

5. Discussion

In comparison to the arguments addressed in the literature review, the impacts on user behaviour outlined here are relatively modest. This is partly due to the methodology: tracking user interactions online inevitably excludes any research behaviour which bypasses the web interface. Many innovative methodologies do precisely this, relying on quantitative analysis of derivative data sets (see e.g. Liddle, 2012; Nicholson, 2012). The continuing growth of Digital Humanities will make such work more common, but it is more likely to be discovered through literature searches and direct contact with researchers.

This discussion will therefore focus on where the points of difference lie between online user behaviour with digital newspapers, and accounts of information behaviour with physical formats. We previously mentioned the stereotypes which define representations of reading in mainstream media, where it is commonly characterised as deeper and more intellectually engaged than reading online (Birkets, 1994; Scarry, 2001; Carr, 2010). The findings do indicate behaviour distinct from deep engagement with the printed text: users rarely browse through specific newspapers in WNO, instead searching and browsing through the web interface to discover material. More than half of all pageviews are spent not viewing digitised material, but engaged in this iterative information seeking. While we would reject the stereotype of deep reading as a flawed stereotype which accounts for just one part of a multifaceted information behaviour, it is certainly true that digitised online resources assert the centrality of the web interface to the user's experience. This defines and shapes the user's experience, increasing their reliance on automated search, Optical Character Recognition, and online information literacy to aid discoverability. In doing so, digital resources place an increasing demand upon resource creators to ensure that these functions meet the requirements of users, who effectively place trust in algorithmic discovery, metadata production and digitisation technologies to ensure the quality of the resources they discover.

But how different is this reliance on surrogate technologies to information discovery in the physical archive? Although Edwards' (2013) assertion that "a willingness to trust surrogates is a willingness to abandon scholarly responsibility" was directed against digital technologies, it could equally be applied to traditional information discovery. In both scenarios, users are largely reliant upon, and appreciative of, finding aids such as indexes or keyword searches (Duff and Johnson, 2002). Research in physical archives does not solely consist of deep engagement with historical sources for extended periods, though this certainly does occur. Researchers frequently scan physical archival materials for keywords, rather than reading them in their entirety, and then use this information to diversify their search (Duff and Johnson, 2002). Online, this is evident in the way that users of WNO rely on filtering rather than search to identify relevant material. This returns us to Levy's (1997, p. 209) point: both digitised newspaper resources and physical archives provide a locus for search, acquisition and reading. His observation is reinforced in a variety of studies which note that research consists of multiple consecutive activities rather than a series of discrete stages (Uva, 1977; Case, 1991; Duff and Johnson, 2002; Sinn and Soares, 2014). This closely resembles the model in this study, where searching and browsing sit alongside reading. As such, usage of online digitised newspapers can be more closely mapped to the multifaceted approach of archival research than the stereotype of deep reading, allowing us to conclude that researchers can benefit from the automation of filtering and discovery offered by digitised collections.

In summary, we believe that digitised newspapers have not, for the majority of users, supported a shift away from the "scholarly primitives" (Unsworth, 2000) that constitute humanities research. Instead, they facilitate the acceleration of existing information behaviour, allowing researchers to undertake complex information seeking tasks via an online resource rather than incurring financial and time costs associated with archival research. This occurs at a layer of abstraction from the original newspapers, centred upon the web interface and recalls Brake's (2012, p. 223) matrix-embedded "user". While the theoretical implications of this shift in focus should be incorporated in the wider debate around the impact of digital technologies, we would point out that a differently remediated experience is not necessarily any less rich. Ranganathan's (1931) conception of the term "reader" as a user of library collections, rather than an individual engaged directly in reading, remains relevant: Ranganathan situates the reader as an individual with the need for library services, and emphasises the importance of meeting user needs regardless of the way the reader chooses to user library materials. It thus reinforces the idea of library as service at a time when library user communities are increasingly diffuse and anonymous, incorporating both traditional and emerging technologies which should wherever possible fit the diverse requirements of its users. It is therefore the job of digital resource creators to guard against any negative impacts from the inevitable reshaping of historical materials which occurs through online digitised collections. This can be manifested through a critical, user-centric approach to interface design for digital resources.

6. Conclusion

This paper demonstrates that content log analysis can enrich our understanding of users of digitised newspaper collections. Web logs provide a more nuanced data source than web analytics platforms for understanding how users interact with digital resources. Because webometrics cannot interpret the reasons for this behaviour, this paper steers away from hypothesising user motivations, instead interpreting their behaviour in comparison to existing models of information behaviour. These users, although abstracted from the newspapers by a web interface which shapes their interactions to a large extent, are still engaged in a research process which is recognisable from physical archives. Archival researchers have previously engaged in information seeking behaviour which closely matches that for digitised collections, particularly in their reliance on discovery tools and filtering techniques to navigate large archival collections. Multiple strands of search, discovery and reading overlap online, much as they do in the physical archive, and we would therefore characterise usage of digitised newspapers as an accelerated version of existing information seeking behaviour.

There are some limitations to this study which suggest the need for further work. As a medium, digitised newspapers have complex formal arrangements that are profoundly influenced by digital remediation. We have not considered how this editorial process affects usage, and this would extend the scope of the research considerably. Furthermore, while webometrics provide a strong statistical base for analysing user behaviour, they cannot interpret the reasons for this behaviour. In order to discern how this user behaviour fits into online information behaviour more generally, future work should incorporate qualitative methods which involve direct contact with scholars; indeed, we believe the time and cost pressures of mounting such studies make it an underexplored area. Finally, the editorial and political implications of interface design are being interrogated by a number of scholars (Mussell, 2012; Baker, 2013; Wragge, 2015): given the findings of this paper, an empirical study into the impact of interface design upon user experiences of cultural heritage materials would be particularly timely. The centrality of web interfaces to effective information discovery online make it essential that, alongside the widespread digitisation of cultural materials, priority should be given to ensuring that these interfaces are suitable for supporting the contemporary needs of connected researchers.

Notes

1. The collection can be freely accessed online via the following URL available at: http://papuraunewyddcymru.llgc.org.uk/en/home
2. This fund is managed and administered by the Welsh Government. It was established in 2008 to take decisions on investment proposals within Wales and to oversee the delivery of capital investment programmes. Further details are available at: www.assemblywales.org/qg10-0011.pdf
3. The European Regional Development Fund aims to correct imbalances between its regions, in order to strengthen economic and social cohesion in the European Union. Further details can be available at: http://ec.europa.eu/regional_policy/thefunds/regional/index_en.cfm
4. For examples of studies which have used Google Analytics as a data source see Fang (2007), Betty (2008, 2009), Turner (2010) and Way (2010).
5. Studies which have taken this approach notably include Warwick *et al.* (2008) and Meyer *et al.* (2009).
6. A robot, or internet bot, is a computer that does automated tasks. In the case of web logs, this refers to web crawlers which are deployed by major search engines to automatically and systematically browse the World Wide Web, most commonly for the purpose of web indexing. These are tracked by web servers, and therefore cause a great deal of noise which must be cleaned before web server logs are usable. By contrast, GA automatically ignores web crawlers which do not execute the Javascript embedded in each page to collect data (Pinto, 2012).
7. A proxy server acts as an intermediary for a client computer seeking resources from other servers: the client connects to the server, which then connects to the other server to provide whichever service the client has requested. This means that web logs will record the IP address of the proxy server rather than the client computer (Rouse, 2008).
8. Bounce rate is a measure of the percentage of single page sessions on a website, defined as session in which a user leaves the site without interacting with more than one page.

References

Baker, J. (2013), "Digital research in the wild", British Library Digital Scholarship Blog, 19 November, available at: http://britishlibrary.typepad.co.uk/digital-scholarship/2013/11/digital-research-in-the-wild.html?utm_content=buffer2d359&utm_source=buffer&utm_medium=twitter&utm_campaign = Buffer (accessed 20 November 2013).

Batra, A. (2008), "Typical bounce rates: survey results", Web Analytics, Behavioural Targeting and Optimization by Anil Batra, 10 March, available at: http://webanalysis.blogspot.co.uk/2008/03/typical-bounce-rates-survey-results.html (accessed 6 January 2014).

Benjamin, W. (2007), "The work of art in the age of mechanical reproduction", *Illuminations*, Schocken Books (Penguin Great Ideas), New York, NY.

Betty, P. (2008), "Creation, management, and assessment of library screencasts: the regis libraries animated tutorials project", *Journal of Library Administration*, Vol. 48 Nos 3-4, pp. 295-315.

Betty, P. (2009), "Assessing homegrown library collections: using Google Analytics to track use of screenshots and flash-based learning objects", *Journal of Electronic Resource Librarianship*, Vol. 21 No. 1, pp. 75-92.

Birkets, S. (1994), *The Gutenberg Elegies: The Fate of Reading in an Electronic Age*, Ballentine Books, New York, NY.

Brake, L. (2012), "Half full and half empty", *Journal of Victorian Culture*, Vol. 17 No. 2, pp. 222-229, available at: http://dx.doi.org/10.1080/13555502.2012.683151

Carr, N. (2010), *The Shallows: How the Internet is Changing the Way We Read, Think and Remember*, Atlantic Books, London.

Case, D.O. (1991), "The collection and use of information by some American historians: a study of motives and methods", *Library Quarterly*, Vol. 61 No. 1, pp. 61-82.

Davies, J. (1994), *A History of Wales*, Penguin Books Ltd, London.

Digital Analytics Association (2012), "About us", Digital Analytics Association, available at: www.digitalanalyticsassociation.org/?page = aboutus (accessed 20 April 2012).

Duff, W.M. and Johnson, C.A. (2002), "Accidentally found on purpose: information-seeking behavior of historians in archives", *The Library Quarterly*, Vol. 72 No. 4, pp. 472-496.

Edwards, A.S.G. (2013), "Back to the real?", *The Times Literary Supplement*, 7 June, available at: www.the-tls.co.uk/tls/public/article1269403.ece (accessed 10 June 2015).

Fang, W. (2007), "Using Google Analytics for improving library website content and design: a case study", *Library Philosophy and Practice*, June, pp. 1-17, available at: http://works.bepress.com/wfang/1/ (accessed 20 April 2012).

Levy, D.M. (1997), "I read the news today, oh boy: reading and attention in digital libraries", *Proceedings of the Second ACM International Conference on Digital Libraries, New York, NY*, pp. 202-211, available at: http://renu.pbworks.com/f/p202-levy.pdf (accessed 18 August 2014).

Liddle, D. (2012), "Reflections on 20,000 Victorian newspapers: 'distant reading' the times using the times digital archive", *Journal of Victorian Culture*, Vol. 17 No. 2, pp. 230-237, available at: http://dx.doi.org/10.1080/13555502.2012.683151

Meyer, E.T., Eccles, K., Thelwall, M. and Madsen, C. (2009), "Usage and impact study of JISC-funded phase 1 digitisation projects & the toolkit for the impact of digitised scholarly resources (TIDSR)", Oxford Internet Institute, University of Oxford, Oxford, available at: http://microsites.oii.ox.ac.uk/tidsr/sites/microsites.oii.ox.ac.uk.tidsr/files/TIDSR_FinalReport_20July2009.pdf (accessed 9 May 2012).

Mussell, J. (2012), *The Nineteenth-Century Press in the Digital Age*, Palgrave MacMillan, Basingstoke.

National Library of Wales (2013), "Welsh Newspapers Online", Digitisation Projects, National Library of Wales, Wales, available at: www.llgc.org.uk/index.php?id = 4723 (accessed 15 January 2014).

Nicholas, D., Huntington, P., Lievesley, N. and Wasti, A. (2000), "Evaluating consumer website logs: a case study of the times/the sunday times website", *Journal of Information Science*, Vol. 26 No. 6, pp. 399-411. doi: 10.1177/016555150002600603.

Nicholas, D., Huntington, P., Williams, P. and Dobrowolski, T. (2004), "Re-appraising information seeking behaviour in a digital environment: bouncers, checkers, returnees and the like", *Journal of Documentation*, Vol. 60 No. 1, pp. 24-39.

Nicholson, B. (2012), "Counting culture; or, how to read Victorian newspapers from a distance", *Journal of Victorian Culture*, Vol. 17 No. 2, pp. 238-246. doi: 10.1080/13555502.2012.683331.

Office for National Statistics (2014), "Statistical bulletin: annual mid-year population estimates, 2013", Office for National Statistics, Newport, available at: www.ons.gov.uk/ons/rel/pop-estimate/population-estimates-for-uk–england-and-wales–scotland-and-northern-ireland/2013/stb—mid-2013-uk-population-estimates.html#tab-What-do-the-mid-2013-UK-population-estimates-show- (accessed 10 June 2015).

Pinto, A. (2012), "Google Analytics: how to segment and filter robot traffic", Yottaa's Site Performance and Optimization Blog, 25 September, available at: www.yottaa.com/blog/bid/223629/Google-Analytics-How-to-Segment-and-Filter-Robot-Traffic (accessed 6 January 2014).

Ranganathan, S.R. (1931), *The Five Laws of Library Science*, Asia Publishing House, Bombay.

Rouse, M. (2008), "What is proxy server?, search networking", available at: http://whatis.techtarget.com/definition/proxy-server (accessed 20 February 2014).

Scarry, E. (2001), *Dreaming by the Book*, Princeton University Press, Princeton, NJ.

Sinn, D. and Soares, N. (2014), "Historian's use of digital archival collections: the web, historical scholarship, and archival research", *Journal of the Association for Information Science and Technology*, Vol. 65 No. 9, pp. 1794-1809, available at: http://onlinelibrary.wiley.com/doi/10.1002/asi.23091/abstract (accessed 26 June 2014).

Tanner, S. (2013), "The value of Welsh newspapers online", When the Data Hits the Fan: The Blog of Simon Tanner, 28 March, available at: http://simon-tanner.blogspot.co.uk/2013/03/the-value-of-welsh-newspapers-online.html (accessed 28 March 2013).

The Digital Victorianist (2013), "Welsh Newspapers Online", The Digital Victorianist, 15 March, available at: www.digitalvictorianist.com/2013/03/welsh-newspapers-online/ (accessed 15 March 2013).

Thelwall, M. (2009), *Introduction to Webometrics: Quantitative Research for the Social Sciences*, Morgan and Claypool Publishers, San Rafael, CA, available at: www.morganclaypool.com/doi/pdf/10.2200/S00176ED1V01Y200903ICR004 (accessed 15 February 2012).

Turner, S.J. (2010), "Website statistics 2.0: using Google Analytics to measure library website effectiveness", *Technical Services Quarterly*, Vol. 27 No. 3, pp. 261-278.

Unsworth, J. (2000), "Scholarly primitives: what methods do humanities researchers have in common, and how might our tools reflect this?", Humanities Computing: Formal Methods, Experimental Practice, King's College London, London, available at: http://people.brandeis.edu/~unsworth/Kings.5-00/primitives.html (accessed 5 August 2014).

Uva, P.A. (1977), "Information-gathering habits of academic historians: report of the pilot study", ERIC ED 142 483, Upstate Medical Center, State University of New York, Syracuse.

Warwick, C., Terras, M., Huntington, P. and Pappa, N. (2008), "If you build it will they come? The LAIRAH study: quantifying the use of online resources in the arts and humanities", *Literary and Linguistic Computing*, Vol. 23 No. 1, pp. 85-102.

Way, D. (2010), "The impact of web-scale discovery on the use of a library collection", *Serials Review*, Vol. 36 No. 4, pp. 214-220.

Wragge, T. (2015), "Unremembering the forgotten", Digital Humanities 2015, Sydney, available at: http://discontents.com.au/unremembering-the-forgotten (accessed 16 July 2015).

Yu, L. and Apps, A. (2000), "Studying e-journal user behaviour using log files: the experience of superjournal", *Library and Information Science Research*, Vol. 22 No. 3, pp. 311-338. doi: 10.1016/S0740-8188(99)00058-4.

Zuccala, A. and Thelwall, M. (2006), "LexiURL web link analysis for digital libraries", *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries, 2006. JCDL '06, Chapel Hill, NC, 11-15 June*. doi: 10.1145/1141753.1141867.

Corresponding author
Paul Gooding can be contacted at: p.gooding@uea.ac.uk

This article has been cited by:

1. Jiří Štěpánek, Vladimír BurešGeneric Model for Adaptable Caching in the Knowledge-Oriented Web Engineering 251-261. [CrossRef]