Proceedings Paper:
Walsh, D., Clough, P., Hall, M. et al. (2019) Analysis of transaction logs from National Museums Liverpool. In: Doucet, A., Isaac, A., Golub, K., Aalberg, T. and Jatowt, A., (eds.) TPDL 2019 Proceedings : Digital Libraries for Open Knowledge. 23rd International Conference on Theory and Practice of Digital Libraries, 09-12 Sep 2019, Oslo, Norway. Lecture Notes in Computer Science (11799). Springer, pp. 84-98. ISBN: 9783030307592. ISSN: 0302-9743. EISSN: 1611-3349.

https://doi.org/10.1007/978-3-030-30760-8_7

This is a post-peer-review, pre-copyedit version of an article published in TPDL 2019 Proceedings. The final authenticated version is available online at: http://dx.doi.org/10.1007/978-3-030-30760-8_7.

Analysis of Transaction Logs from National Museums Liverpool

David Walsh[1,2]*(0000-0003-2972-8233), Paul Clough[2,4] (0000-0003-1739-175X), Mark Hall[3] (0000-0003-0081-4277), Frank Hopfgartner[2] (0000-0003-0380-6088), Jonathan Foster[2] (0000-0002-9439-0884), and Georgios Kontonatsios[1] (0000-0001-5935-4709)

[1] Edge Hill University, Ormskirk, Lancashire, UK
[2] University of Sheffield, Sheffield, UK
[3] Martin Luther University Halle-Wittenberg, Germany
[4] Peak Indicators, Chesterfield, UK

Abstract. The websites of Cultural Heritage institutions attract the full range of users, from professionals to novices, for a variety of tasks. However, many institutions are reporting high bounce rates and therefore seeking ways to better engage users. The analysis of transaction logs can provide insights into users' searching and navigational behaviours and support engagement strategies. In this paper we present the results from a transaction log analysis of web server logs representing user-system interactions from the seven websites of National Museums Liverpool (NML). In addition, we undertake an exploratory cluster analysis of users to identify potential user groups that emerge from the data. We compare this with previous studies of NML website users.

Keywords: digital cultural heritage, museum website, users, survey, transaction log analysis, cluster analysis

1 Introduction

The analysis of transaction logs provides insight into users' searching and browsing behaviours and is a common activity in the design and evaluation of search and website interfaces [1, 2]. Although there is extensive literature on transaction log analysis of user-system interaction logs, such as Online Public Access Catalogues (OPAC's) [3, 4], less attention has been given to digital cultural heritage websites, especially those containing a collection archive. However, with high website bounce rates being reported by cultural heritage institutions [5, 6], there is clearly a need for deeper investigation to be undertaken.

An example of such a cultural heritage institution is the National Museums Liverpool (NML). The NML website[5] is a collection of seven separate sites that cover a wide range of areas from art galleries to natural/world history, maritime and slavery museums. Similar to previous studies of Digital Cultural Heritage

[5] http://www.liverpoolmuseums.org.uk/

(DCH) website usage [5, 7], data from NML transaction logs indicates that approximately 60% of users leave within ten seconds (i.e., a high bounce rate). In 2017, Walsh et al. [7] undertook a large-scale online survey of users visiting the NML website to better understand uses of the website from multiple perspectives. The study investigated characteristics of users with respect to atypical groups that visitors assigned themselves (e.g., student, professional, hobbyist, etc). In this work we continue to study the behaviours of NML website visitors utilising web server logs (weblogs) and transaction log analysis. NML forms an appropriate case study given their wide spread of heritage subject areas leading to a wide range of museum visitors in both physical and online spaces. To the best of our knowledge, this is the first in-depth study of web server logs from National Museums Liverpool. Insights into user groups gained from this study, especially given the multi-site nature of the NML website, may be equally applicable to other digital cultural heritage sites.

The remainder of this paper is structured as follows: Section 2 describes related work in categorising users of digital cultural heritage and transaction log analysis. Section 3 describes the study undertaken. Sections 4 and 5 present and discuss results of the transaction log analysis and presents results of the exploratory cluster analysis. Finally, Section 6 concludes the paper and offers directions for future work.

2 Related Work

2.1 Background of this Study

In previous research we identified distinct categories of online NML user based on visitor answers to a large-scale online user survey [7, 6]. We found that user groups could be distinguished by aspects, such as motivation, task, engagement, search strategy, domain knowledge and location. Results showed that the frequently understudied 'general public' and 'non-professional' groups made up the majority of users (approximately 77% of survey respondents). In this research we seek to further our understanding of NML visitor behaviours through the analysis of user-system interactions, where possible automatically deriving characteristics and categories of users. Whilst we are unable to extract log traces of survey respondents directly, we use transaction logs that cover the same time period as the survey was live, thereby offering complementary information. Analysing how people engage with websites and search systems can help identify what information people require, how they find it and potential findability issues [8]. For NML, understanding and identifying user groups will help support users' needs, such as developing personalised and adaptive user interfaces [9].

2.2 Cultural Heritage User Groups

Cultural heritage institutions have made significant efforts to understand visitors to its physical sites. Previous studies have used multiple methods and considered

visitors from various perspectives, such as motivation [10]; engagement [11]; role adopted during the visit [12]; user expertise and profession [13, 14]; information needs [15, 16]; or the chances of wandering in [7]. Similar findings have emerged from studying online users of DCH websites. Clough et al. [17] studied the users of Europeana[6] via an online survey to better understand their search tasks. They devised a scheme for categorising users' search motives and further actions. Walsh et al. [7] conducted a survey on the NML website where existing user groups were identified with five distinguishing characteristics [6]: Motivation, Domain or CH Knowledge, Task, Location and Frequency of visit. The most common group was the "general user" or "general public." This contrasts with much of the prior work that has focused on more specific groups, such as museum professionals, academics, students, and hobbyists.

2.3 Transaction Log Analysis

Whenever users interact with online services their requests are logged in server logs (or weblogs), which can help to provide insights into users' search and navigation behaviours, such as the pages accessed, time spent on pages, and patterns of use [18] (e.g., sequential pages accessed [19] or search queries entered and reformulated [20]). Previous work on transaction log analysis typically relies on grouping requests into *sessions*: all concurrent requests from the same IP address and User-Agent, often within some cut-off period (e.g., 30 minutes). Jones et al. [1] undertook a manual analysis of a digital library log and were able to identify user demographics and the searches and search patterns the users adopted. Chen et al. [21] found a user session could be characterised, based on 47 variables from a library catalogue log, these were grouped into six clusters which characterised users' interactions (mostly reflecting domain knowledge and technical knowledge).

Wang et al. [22] identified clusters of "similar" users by partitioning a similarity graph (where nodes are users and edges are weighted by click stream similarity). The partitioning process leverages iterative feature pruning to capture the natural hierarchy within user clusters to visualise and understand user behaviours. An unsupervised learning method is used to build interaction behaviour models from click stream data. Zhang & Kamps [23] use search logs to identify searcher stereotypes of novice and expert users. By manually analysing user query terms, session lengths, session duration and repeat visits (i.e., frequency of visit), they were successfully able to establish that further groups discovered shared traits with that of the two groups. However, Stenmark [24] states that in order to capture the variety in behaviour known to exist amongst searchers, an automatic clustering technique (based on self-organising maps and search log data) is required. Overall, they identified six clusters and condensed them into three groups: casual seekers (fact seekers - intranet users looking for quickly retrieved answers); a more holistic group (longer sessions and more read-

[6] https://www.europeana.eu/portal/en

ing time); and information-seeking-savvy employees who formulate longer queries and browse through more documents than other groups.

3 Methodology

In this study three months of weblogs from NML (Section 3.1) are analysed. Although Google Analytics provides similar data, for deeper analysis web server logs are needed. To identify potential users groups, cluster analysis is used on features derived from the server logs (Section 3.2). Relationship between the user-system interactions and the groups is also investigated. Clustering is performed on a re-structured version of the logs where each row represents users and features that capture interactions averaged across user sessions.

3.1 Analysing the Logs

Initially logs were gathered from the Microsoft IIS web server hosting the NML sites. All interactions are logged, including those from 'real' users and those from robots (also known as crawlers or spiders). IIS logs can store referrer details, but this feature has not been enabled for the NML website. The format of logs is standard IIS format:

> *Example log entry*: 2015-07-08 00:00:00 172.17.101.21 GET /onlineshop/graphics/product/thumbnails/sewing-box-new-thumb.jpg - 80 - 66.249.64.146 Googlebot-Image/1.0 304 0 0 15

The logs were prepared by extracting entries for the period of the user survey +/- 1 month (1 Jan - 30 Mar 2017). Using logs for this period allows comparisons between the survey findings and the log analysis to be made. Next, bots, crawlers and developer validation tools (e.g. link checkers etc) were removed using existing seed lists. The location (country, county and city levels) of requests was identified based on the IP address and using the `IP2location` library[7]. The next stage was *sessionisation* - segmenting entries into user sessions based on IP address, User-Agent and a 30 minutes time cut-off, which ensured the full activity for the session was captured (as outlined in [25]). With entries grouped into sessions log data were fully anonymised - all potential user identifying data (e.g., names, emails, credit card numbers, etc.) were removed, IP addresses randomly hashed. Next, data were cleaned: entries without a 200 HTTP status code (successfully displayed pages) were removed; entries that were not GET requests removed; and requests for background files (e.g., images; CSS, JavaScript, font and Adobe Flash files; XML data and linked PDF files) removed. Finally, sessions were grouped by user based on IP address and User-Agent.

[7] https://lite.ip2location.com/ip-address-ranges-by-country

3.2 Clustering the Logs

Previous studies have clustered user sessions [18, 26]; however, in this paper we wanted to establish if potential user groups could be identified from the logs. Therefore, we focus on clustering *users*. To enable this all sessions containing the same IP address and User-Agent were grouped together and features derived from the logs were used to infer user groups (see Table 1). K-means clustering was applied using the features that are likely to distinguish user groups[8].

Table 1: Mapping user group characteristics to log data

User group characteristic	**Log data**
Motivation	Starting level page
Domain / CH knowledge	Page type and queries
Task	Page type and possibly queries
Location	IP (reversed) identifying country, region and city.
Frequency of visits	Repeat visits (sessions), queries, length of session

Features used to characterise users were based on the findings of prior work [7, 6] and can be divided into 3 main categories: *Locational* (country, region and city); *Behavioural* (#sessions; total page requests; #page requests for museum overview, collection overview, item, event, shop and general information page; #museums accessed; average interactions/session); and *Knowledge* (total queries; average queries/session). We also experimented with clustering using the location and session count columns; however, results for both features in all clusters did not result in clear clusters.

As previously mentioned, cluster analysis was used to derive user groupings from the data rather than relying on pre-defined groups. The commonly used k-means algorithm, well-suited to numeric data, was used. Prior to clustering variable values were scaled and to choose the appropriate number of clusters k, the elbow method was used [27]. This identified the suitable number of clusters between 6 and 8. To ensure the number of clusters was valid and stable we repeatedly ran the cluster analysis with different values of k (between 5-9), with $k = 7$ producing the most stable results.

4 Findings

4.1 Website Interactions

Server logs for this study comprised 586,868 rows, each representing a page request. After pre-processing and sessionisation (Section 3), the data reduced to 321,174 rows representing user sessions. Page requests come from 213 different

[8] Alternative algorithms such as k-modes (k-prototypes) and DBScan were also tested, but no stable clusters emerged.

countries[9], with February being the most accessed month across all sites. This aligned to the time period our online survey was running (see Figure 1). Tuesday and Wednesday are the most frequently accessed days with least usage at the weekends (see Table 2). The increased access that can be seen in Figure 1 during February relates to collections that are ending and new ones being announced. This period is also the local school holidays and could explain the increase in accesses for the general information pages.

Table 2: Visits per day (Jan-Mar 2017)

Day	**Mon**	**Tue**	**Wed**	**Thur**	**Fri**	**Sat**	**Sun**	**Total**
Requests	81,450	100,325	101,314	97,311	85,603	54,967	65,898	586,868
Percentage	13.88%	17.09%	17.26%	16.58%	14.59%	9.37%	11.23%	100%

Museum	**Requests**
ISM	97,686
Other pages	92,433
WML	86,516
Walker	73,194
Maritime	68,912
Events	58,273
MOL	54,697
Ladylever	24,607
Shop	21,740
Sudley	8,810
Total	**586,868**

Table 3: Requests by Museum

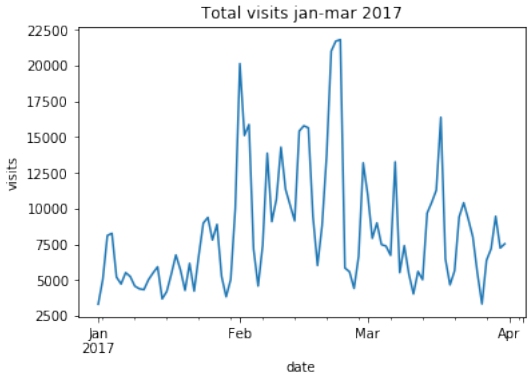


Fig. 1: Total visits (Jan-Mar 2017)

4.2 Museum and Gallery Access

The International Slavery Museum (ISM) is the most accessed site of the NML website with 97,686 requests. This is followed by the World Museum Liverpool, Walker art gallery, Maritime Museum, Museum of Liverpool, Ladylever art gallery and lastly Sudley House. The overall request breakdown between the main sections of the website can be seen in Table 3.

4.3 Page-level Access

Throughout the NML website(s), there are 8 main types of pages: **General** pages (G), which include the highest level NML home page, about page, visit

[9] Based on the `IP2Location` IP4 allocated IP address ranges; however, it is noted that the United Nations only identifies 195

pages, floor plans, policy pages and contact pages etc; the individual **Museum/gallery** home pages and the general overview pages (M); the **Collection overview** pages (C), which introduce the collections the museum is displaying on the site; the **Item** pages (I), which contain item/artefact detail pages for collections; the **Archive** pages (A) for the maritime museum only, separate to the collections but act as a more traditional archive of artefacts; the **Events** (E) pages, including 'Whats-on' pages regarding attractions visitors can view/book; the **Kids** (K) pages, which is a dedicated children's area; and **Undefined** pages: pages that do not fit other categories or occasional pages in new sections, such as Christmas e-cards or 'picture-of-the-month'. The general pages are the most requested (31%), followed by the item pages (24%) - see Table 4. The most frequent (requested) page is the NML homepage (/) with 44,247 requests, followed by the world museum home page (/wml/) with 27,015 requests. The next is (/wml/collections/zoology/herptile/item.aspx) with 23,788 requests.

Table 4: Page type usage

Page Type	**Count**	**Percentage**
General	182,185	31%
Item	139,163	24%
Event	58,273	10%
Museum overview	56,675	10%
Collection overview	40,573	7%
Undefined	40,546	7%
Kid	36,694	6%
Shop	21,740	4%
Archive	11,019	2%

With the page types extracted from the `cs-uri-stem` (uristem) we see that the UK provides the most visitors to the website (see Table 5). We also note that the top two page types are the general pages and the item pages. Surprisingly, the events pages are third (and less surprising is the shop, which is the lowest accessed page type). When exploring only the UK page requests, the general pages are still the most requested with the events a close second. The events pages seem to be high for a number of the more local areas to Liverpool. This might be an indication of a possible physical visit. For those areas outside of the Northwest UK the pattern of requests focuses more on the general pages and the item pages (see Table 6).

Findings also show that single page visits occur when users leave the site, without exploring more than the initial page. There are 586,868 total requests over the 3 months with 238,061 (40.56%) of the requests for single page accesses (a breakdown of this can be seen in Table 7). The country with least accesses is Denmark (2.75%); the highest Brazil (70.73%). Looking at the UK's single page request, which is the largest proportion of requests, the majority of requests are from areas within the locality of physical museums in Liverpool (see Table 8).

Table 5: Page types by country

Countries	**G**	**M**	**C**	**I**	**S**	**E**	**Requests**	**Queries**
United Kingdom	**45,242**	35,799	16,972	38,500	9,235	43,245	307,347	181,903
United States	12,658	6,033	3,768	**14,153**	6,132	2,768	120,584	43,062
Denmark	2,279	491	546	**17636**	3,111	599	32,012	9,098
Germany	**5,158**	833	934	3,036	291	488	16,878	7,846
Australia	**731**	250	171	507	86	48	15,805	4,012
Ireland	1,177	447	620	1,384	557	**7,018**	15,306	5,527
France	**2,484**	961	447	1,802	454	732	11,671	5,630
Canada	**979**	256	190	896	417	168	6,418	3,592
...	...	...	...	...	...	...	...	
Grand Total	**84,133**	49,849	25,974	**87,712**	21,740	58,273	586,868	297,492

Table 6: Page types by UK city

UK City/Town	**G**	**M**	**C**	**I**	**S**	**E**	**Requests**	**Queries**
Manchester	5,157	4,661	2,405	**5,590**	1,375	**5,899**	40,992	20,696
Liverpool	**7,304**	4,146	2,173	4,278	1,242	**7,383**	37,804	23,014
London	**3,656**	2,784	1,304	**3,557**	778	2,600	26,513	14,798
Runcorn	1,398	1,220	698	1,200	275	**1,994**	10,070	6,468
Sheffield	**973**	**838**	324	**889**	207	738	7,151	4,473
Birmingham	**1,014**	**908**	302	804	198	831	7,148	4,132
Warrington	**940**	**905**	364	616	193	844	6,114	3,974
Chesterfield	**1,043**	**852**	294	634	137	681	5,097	2,743
...	...	...	...	...	...	...	...	...
Grand Total	**45,242**	35,799	16,972	38,500	9,235	**43,245**	307,347	181,903

Table 7: Single page sessions

Country	**Single Sessions**	**Total Requests**	**Percentage**
United Kingdom	130,017	307,347	42.30%
United States	50,493	120,584	41.87%
Germany	11,133	16,878	**65.96%**
Australia	5,174	15,805	32.74%
France	4,086	11,671	35.00%
Canada	2,874	6,418	44.78%
Ireland	2,623	15,306	17.14%
Italy	2,049	5,822	35.19%
...	...	...	...

Table 8: Single page session by UK city

UK City/Town	G	M	C	I	S	E	Total	Queries
Liverpool	1,714	1,257	936	1,452	222	**1,951**	7,532	11,934
Manchester	**1,491**	1,242	798	1,381	164	1,279	6,355	10,566
London	1,287	868	512	**1,271**	158	638	4,734	8,300
Runcorn	479	381	342	434	54	570	2,260	3,584
Warrington	388	303	160	283	45	237	1,416	2,367
Sheffield	367	280	147	370	42	232	1,438	2,658
Birmingham	358	277	118	289	40	204	1,286	2,255
Chesterfield	257	198	100	163	34	193	945	1,407
...	...	...	...	...	...	...	...	...
Grand Total	**14,915**	11,092	7,020	13,562	1,782	10,998	59,369	99,398

4.4 Sessions

A total of 321,174 sessions were identified when the 30 minute cut-off time limit was applied (265,776 sessions without applying the cut-off). Looking at sessions comprising single pages, we see general pages (G) are accessed the most (see Table 8). These are closely followed by item pages (I). In the case of NML item level pages, bounces are likely cases when the visitor has obtained the information needed or desired. Also the item level pages do not promote other items; rather highlight the collection that the item belongs (but the links are not obvious).

Session entry pages: Understanding users' entry points can offer insights regarding how a site is used, as well as the level of knowledge being sought. The logs show that the NML group homepage (G) and the item pages (I) are the most initially requested pages. This is followed by the museum homepages, the collection overview pages, the event pages and then the kids area pages, the shop and only a small number for the archive pages (See Table 9).

Session exit pages: Just as interesting, and potentially as informative as entry pages, are the exit pages. The closeness of the entry and exit results clearly highlight the amount of single page sessions (see Table 9). There is an indication of some movement around the site from the Museum overview pages and some of the undefined pages to all other page types, from which they then leave.

Length of session: Results show that 246,643 sessions ended within 10 seconds, only 7,637 of these sessions ended within 10 seconds after visiting more than one page - a 43% drop-out within the 10 second period. Google Analytics reports a 52.6% bounce rate for the same period. The drop-out rates over time clearly identify that there are very few users during this three month period who spend significant time on the site (See Table 10).

5 Discussion

The results clearly show that the majority of pages accessed by users fall under the categories of general and item (see Table 4). This aligns with the findings identified in [6] for understanding users' purposes for using the website,

Table 9: Session entry and exit pages

Page type	Entry	Exit
General	110,322	114,884
Item	62,576	65,922
Museum overview	37,698	28,432
Undefined	29,478	28,840
Collection overview	26,322	26,418
Event	25,856	26,647
Kids	14,125	14,087
Shop	7,950	8,983
Archive	6,847	6,961

Table 10: Session drop-out rates

Length	Count	Percentage
<10 sec	246,643	76.8%
<20 sec	252,239	78.5%
<30 sec	256,571	79.9%
<1 min	265,991	82.8%
30 mins	321,174	100%

which showed that the major reason for visiting by the general public and non-professional users was to prepare for a visit.

The high number of single page sessions (see Tables 7 and 8) may indicate that users are being provided with the correct information to meet their needs (e.g., finding a known-item). These numbers also align with [6], where the majority of users (especially the general public and non-professional users) were visiting for the first time. Linking to the number of sessions that end within 10 seconds (Table 10) and little movement around the site indicated by session entry and exit pages (Table 9) highlights limited engagement with the site(s).

The rareness of search results (only 7,121 from 586,868 requests) also aligns with the [6] study where the survey identified that the use of navigational links was the preferred method of access by the majority of users (general public and non-professionals). A 2015 survey conducted by NML also indicated that only a small fraction of visitors were museum professionals or academics. Taking into consideration their higher CH and domain knowledge levels this could indicate that the searchers of known-items or collections are from these user groups.

The NML websites are being accessed from all over the world (Table 5). When IP addresses are reversed we see the largest number of visits are from the UK (52.3%). Whilst this is not as high as the survey results from [6] indicate (75%), it is still the largest proportion of requests. The bulk (12.7%) of these UK users are local to the Liverpool area (Table 6) and thus are within in a reasonable distance to physically visit the museums/galleries. This also aligns with findings of [6]: location is a good indicator of user group. In our case the high number of local users could indicate general public (pre-visit) users.

5.1 Clustering Results

The logs provide a data set of 225,796 total users, which reduces to 225,730 when reducing the noise in the clusters by removing rows with over 100 in the columns of Museum overview pages (M), Collection overview pages (C), artefact Item pages (I), Event pages (E), Shop pages (S), General information pages (G), and queries.

Table 11: Cluster results with assigned labels

Cluster	# Users	M	C	I	E	S	G	Queries	Label
1	172,692 (76.6%)	0.17	0.09	0.21	0.15	0.04	0.00	1.07	Single page viewers
2	46 (0.02%)	2.02	4.33	31.48	9.91	26.85	11.04	56.91	High all round searchers
3	4,162 (2.1%)	0.80	0.30	0.34	3.77	0.15	0.57	3.39	Event visitors
4	45,282 (20%)	0.20	0.05	0.07	0.08	0.02	1.29	1.15	Single query general page visitors
5	292 (0.1%)	1.10	3.48	23.35	4.73	7.48	5.90	21.84	Deep level browsers
6	290 (0.1%)	4.96	1.69	2.02	1.68	0.76	17.58	3.78	General museum visitors
7	2,966 (0.8%)	0.42	0.63	4.57	0.37	1.11	0.74	5.25	Known-item searchers

5.2 Describing the Clusters

As discussed previously 7 groups of users were identified from the cluster analysis and assigned labels, based on their characteristics (see Table 11):

Cluster 1: The cluster labelled "Single page viewers" is by far the largest grouping (76.6%). This cluster is comprised of users who look at only one page and then leave. The pages are at various levels (general to detailed item pages) and on average the users issue 1-2 queries.

Cluster 2: The smallest cluster (0.02%) we label as "High all round searchers" as they view high numbers of all level pages and also issue many queries. Hobbyists have been identified as undertaking research for personal reasons, but are very knowledgeable in their particular areas of interest and as such are able to use search successfully. They are also interested in both the overview as well as the deeper item level pages [28, 29, 30].

Cluster 3: This cluster (2.1% of users) are labelled as "Event visitors" and reflect users who predominantly visit event pages and tend to query for them.

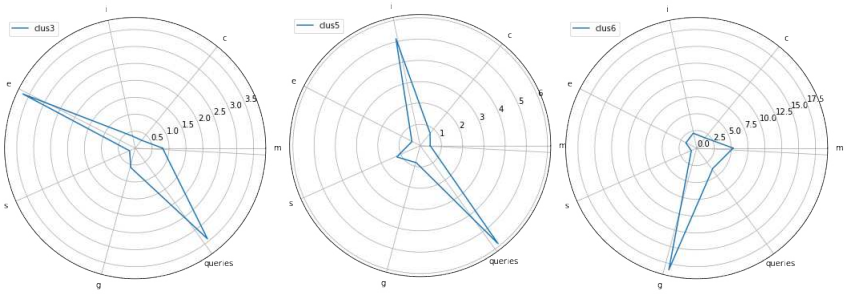
Cluster 4: The second biggest cluster (20%) of users has been labelled as "Single query general page visitors" with the pages mostly being viewed being general and museum overview pages. This aligns with Booth's [15] findings of the general user - someone seeking general information about opening hours, prices, facilities, etc.

Cluster 5: This cluster (0.1% of users) we label as "Deep level browsers" as they commonly use queries, view many item level pages, but very few general or museum overview pages. Similarities can be drawn with Vilar et al. [13] who define professional users (museum staff) as those who have good knowledge of the task, are trained in the systems and have deep understanding of the context.

Cluster 6: This cluster (0.1% of users) is labelled as "General museum visitors" who mostly view general and museum overview pages, with the occasional "deep dive". This aligns with Booth's [15] general user.

Cluster 7: This cluster (0.8% of users) we label as "Known item searchers" who view only item level pages and query frequently. This behaviour may arise as the majority of item level pages in the NML site(s) can only be accessed by the search feature. This behaviour aligns with Marchionini et al's [16] professional user (including academics) who are described as being highly motivated, having high domain knowledge, system knowledge, time available and focus.

The differences in cluster profiles can be clearly seen from the example polar/spider charts shown in Figure 3. The differences in the shapes and the scales are some of the most striking differences.



Cluster 3
Event visitors

Cluster 5
Deep level browsers

Cluster 6
General museum visitors

Fig. 3: Example cluster polar charts

Some of the clusters produced as part of this study show some potential as being candidates for mapping to the known user groups. The potential candidates are shown in Table 12.

Table 12: Clusters mapped to potential user groups

Cluster	# Users	Label	Potential user group
1	172,692	Single page viewers	Currently un-documented user-group called "Bouncers"
2	46	High all round searchers	Non-professionals (hobbyists)
3	4,162	Event visitors	Teachers / General Public
4	45,282	Single query general page visitors	General public (Pre-visit) / Teachers
5	292	Deep level browsers	Museum Staff
6	290	General museum visitors	General public / Students
7	2,966	Known item searchers	Academics (experts) / Non-professionals

6 Conclusions and Future Work

In this paper we have extended prior work on studying users of the National Museums Liverpool (NML) websites through the use of transaction log analysis on web server logs. The analysis has identified a number of documented activities in the server logs that align with known user group activities and behaviours. Using only the types of pages visited as features for cluster analysis the resulting groups are similar to user categories previously discovered in the literature and previous studies of NML users based on user surveys. In future work we plan to further analysis the session data and cluster the logs using additional features of user-system interaction and user characteristics.

Acknowledgements We would like to thank National Museums Liverpool for providing access to the web server transaction logs.

References

1 Jones, S., Cunningham, S.J., McNab, R., Boddie, S.: A transaction log analysis of a digital library. International Journal on Digital Libraries **3**(2) (2000) 152–169
2 McKay, D., Buchanan, G., Chang, S.: It ain't what you do, it's the way that you do it: Design guidelines to better support online browsing. Proc. of the Association for Information Science and Technology **55**(1) (2018) 347–356
3 Peters, T.A.: The history and development of transaction log analysis. Library hi tech **11**(2) (1993) 41–66
4 Jansen, B.J., Spink, A., Saracevic, T.: Real life, real users, and real needs: a study and analysis of user queries on the web. Information processing & management **36**(2) (2000) 207–227
5 Ciber: Europeana 2012-2013: usage and performance update. Technical report, CIBER Research (july 2013)
6 Walsh, D., Hall, M.M., Clough, P., Foster, J.: Characterising online museum users: a study of the national museums liverpool museum website. International Journal on Digital Libraries (Jul 2018)
7 Walsh, D., Hall, M., Clough, P., Foster, J.: The ghost in the museum website: investigating the general public's interactions with museum websites. In: International Conference on Theory and Practice of Digital Libraries, Springer (2017) 434–445
8 Farrell, S.: Search-log analysis: The most overlooked opportunity in web ux research. https://www.nngroup.com/articles/search-log-analysis/ (July 2017) (Accessed on 03/14/2019).
9 Eirinaki, M., Vazirgiannis, M.: Web mining for web personalization. ACM Transactions on Internet Technology (TOIT) **3**(1) (2003) 1–27
10 Falk, J.H.: Identity and the museum visitor experience. Left Coast Press (2009)
11 Templeton, C.A.: Museum visitor engagement through resonant, rich and interactive experiences. (2011)
12 Spellerberg, M., Granata, E., Wambold, S.: Visitor-first, mobile-first: Designing a visitor-centric mobile experience. In: Museums and the Web. (2016)
13 Vilar, P., Šauperl, A.: Archival literacy: Different users, different information needs, behaviour and skills. In: Information Literacy. Lifelong Learning and Digital Citizenship in the 21st Century. Springer (2014) 149–159
14 Pantano, E.: Virtual cultural heritage consumption: a 3d learning experience. International Journal of Technology Enhanced Learning **3**(5) (2011) 482–495
15 Booth, B.: Understanding the information needs of visitors to museums. Museum Management and Curatorship **17**(2) (1998) 139–157
16 Marchionini, G., Plaisant, C., Komlodi, A.: The people in digital libraries: Multifaceted approaches to assessing needs and impact. Digital library use: Social practice in design and evaluation (2003) 119–160

17 Clough, P., Hill, T., Paramita, M.L., Goodale, P.: Europeana: What users search for and why. In: International Conference on Theory and Practice of Digital Libraries, Springer (2017) 207–219
18 Russell-Rose, T., Clough, P.: Mining search logs for usage patterns. Text Mining and Visualization: Case Studies using Open-Source Tools **40**
19 Kachhadiya, B.C., Patel, B.: A survey on sequential pattern mining algorithm for web log pattern data. In: 2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI), IEEE (2018) 1269–1273
20 Lau, T., Horvitz, E.: Patterns of search: analyzing and modeling web query refinement. In: UM99 User Modeling. Springer (1999) 119–128
21 Chen, H.M., Cooper, M.D.: Using clustering techniques to detect usage patterns in a web-based information system. Journal of the American Society for Information Science and Technology **52**(11) (2001) 888–904
22 Wang, G., Zhang, X., Tang, S., Zheng, H., Zhao, B.Y.: Unsupervised clickstream clustering for user behavior analysis. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, ACM (2016) 225–236
23 Zhang, J., Kamps, J.: Search log analysis of user stereotypes, information seeking behavior, and contextual evaluation. In: Proceedings of the third symposium on Information interaction in context, ACM (2010) 245–254
24 Stenmark, D.: Identifying clusters of user behavior in intranet search engine log files. Journal of the American Society for Information Science and Technology **59**(14) (2008) 2232–2243
25 He, D., Göker, A.: Detecting session boundaries from web user logs. In: Proceedings of the BCS-IRSG 22nd annual colloquium on information retrieval research. (2000) 57–66
26 Bogaard, T., Hollink, L., Wielemaker, J., Hardman, L., van Ossenbruggen, J.: Searching for old news: User interests and behavior within a national collection. In: Proceedings of the 2019 Conference on Human Information Interaction and Retrieval, ACM (2019) 113–121
27 Bholowalia, P., Kumar, A.: Ebk-means: A clustering technique based on elbow method and k-means in wsn. International Journal of Computer Applications **105**(9) (2014)
28 Skov, M., Ingwersen, P.: Exploring information seeking behaviour in a digital museum context. In: Proceedings of the Second International Symposium on Information Interaction in Context. IIiX '08, New York, NY, USA, ACM (2008) 110–115
29 Skov, M.: The reinvented museum: Exploring information seeking behaviour in a digital museum context. PhD thesis, Københavns Universitet'Københavns Universitet', Faculty of Humanities, School of Library and Information Science, Royal School of Library and Information Science (2009) unpublished thesis.
30 Elsweiler, D., Wilson, M.L., Lunn, B.K.: Chapter 9 understanding casual-leisure information behaviour. New Directions in Information Behaviour (Library and Information Science, Volume 1) Emerald Group Publishing Limited **1** (2011) 211–241