Proceedings Paper:
Walsh, D., Clough, P., Hall, M.M. et al. (2 more authors) (2021) Clustering and classifying users from the National Museums Liverpool website. In: Berget, G., Hall, M.M., Brenn, D. and Kumpulainen, S., (eds.) Linking Theory and Practice of Digital Libraries: TPDL 2021. 25th International Conference on Theory and Practice of Digital Libraries: TPDL 2021, 13-17 Sep 2021, Virtual conference. Lecture Notes in Computer Science, 12866 . Springer Nature , pp. 202-214. ISBN 9783030863234

https://doi.org/10.1007/978-3-030-86324-1_24

Clustering and Classifying Users from the National Museums Liverpool Website

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

[1] Edge Hill University, Ormskirk, Lancashire, UK
[2] The Open University, UK
[3] University of Sheffield, Sheffield, UK
[4] Peak Indicators, Chesterfield, UK

Abstract. Museum websites have been designed to provide access for different types of users, such as museum staff, teachers and the general public. Therefore, understanding user needs and demographics is paramount to the provision of user-centred features, services and design. Various approaches exist for studying and grouping users, with a more recent emphasis on data-driven and automated methods. In this paper, we investigate user groups of a large national museum's website using multivariate analysis and machine learning methods to cluster and categorise users based on an existing user survey. In particular, we apply the methods to the dominant group - general public - and show that subgroups exist, although they share similarities with clusters for all users. We find that clusters provide better results for categorising users than the self-assigned groups from the survey, potentially helping museums develop new and improved services.

Keywords: Digital Cultural Heritage · Museum Website · User Groups · Cluster Analysis.

1 Introduction

Due to the COVID-19 pandemic, museums and galleries around the world had to temporarily close their physical sites, leading to an increased need to provide online access to their content. This was possible thanks to prior investments in online presences, i.e., websites and the curation of digital collections [1]. Such resources are indeed popular amongst users from diverse backgrounds with increasingly varied goals, tasks and information needs [2]. However, users' individual differences (e.g., age and domain knowledge), search task and context (e.g., location and time), are known to affect the ways in which people search for information [3]. It has, therefore, been long recognised that information systems and services must be developed from the perspective of human actors and their environment [4] and support information seeking behaviours beyond keyword-based search [5, 6]. Since the first museums were made available online, there have been attempts to grow and enhance the use of the online collections, generally based on a categorisation of their users. The diversity in users of digital cultural heritage has resulted in a strategy that simplifies the virtually unlimited possibilities of user-profiles by creating generic groups or categories of users - 'stereotypes' [7]. These groups are sometimes as abstract as *novice* or *expert* [5], but more commonly, user groups are created based on profession (e.g., curator, librarian, researcher, teacher or student). Alternative groups have been based on user interest or motivations (e.g., tourist, explorer, general user) or age group (e.g., adult, child) [8].

Manually defining user groups can be time-consuming and difficult; therefore, approaches to automate the process, such as clustering and automated persona generation [9], must be applied. In this paper, we use multivariate analysis and machine learning methods to study groups at The National Museum Liverpool (NML), a collection of seven museums that cover a wide range of areas from art galleries to natural history and slavery. The NML provide a publicly accessible website, allowing users to access information about the physical museums, as well as digital collections. In a previous study of NML users, Walsh et al. [10, 11] conducted an online survey to gather information about users and their purpose of visiting the museum website. They identified that a large proportion of the NML website users (49% n=253 from 514 respondents) considered themselves as 'General Public' [12], a finding common in other studies [13]. In this paper, as well as studying groups across the population as a whole, we focus on analysing users who describe themselves as General Public to better understand the homogeneity of this group and whether sub-groups exist. This study addresses the following research questions:

RQ1 How do cluster analysis results compare with the self-assigned groups?
RQ2 Do sub-groups exist within the self-assigned General Public group?
RQ3 Can we classify the users based on the identified clusters?

The remainder of this paper is structured as follows: In Section 2, we discuss existing work to understand and classify digital cultural heritage users, particularly using cluster analysis; in Section 3, we describe the study we undertook and our methodology. Sections 4 and 5 present and discusses the results, and Section 6 concludes the paper and provides directions for future work.

2 Related Work

There have been numerous past studies on categorising users of cultural heritage resources (see, e.g., [14]). Often the focus has been on users connected to the museum in either a professional/expert capacity or the lay user/non-expert/novice [5]. Groups, such as the General Public (GP) [10], present an opportunity to explore more nuanced categorisations, thereby expanding the field of study to include potential sub-categories or even new groups.

Fundamentally, the characteristics of professional users have been linked to high levels of training and experience, a good knowledge base of required tasks and systems, and expertise in the field of cultural heritage [15]. Recognition of this particular user group culminated in the term MIP *(Museum Information Professional)* as someone working with information resources, and a goal of meeting user needs both internally and externally to the museum [16]. There are sub-categories within the expert/professional category, often based on role/occupation, such as academic, archivist, student and hobbyist. At the other end of the spectrum are novices/non-experts or lay-users who have limited or no formal training in either the systems [15] or subject knowledge [17, 18] but visit the museum and/or its website for personal interest. Cifter [8] states that "knowledge of the task, information needs and system expectations" are the expert's main distinction.

The hobbyist or non-professional users fit between the extremes of expert and novice [19, 2, 20, 21, 22], sharing with the expert a knowledge of cultural heritage, but mainly in specific domains and being like the lay user with a focus on personal reasons. Casual-leisure users are closely related to novice or hobbyist groups. However, they are typically only "first and short-time visitors" [23] who have stumbled upon [the digital] collection. In this respect, they are similar to Falk's experience seekers [24], who wander into the physical museum just for the experience. Villaespesa [13] studied the Metropolitan Museum of Art's collection users and found that non-specialist users needed better clues to navigate and explore the collection, highlighting thus a lack of knowledge of the collection(s) and the system/website.

In Booth's study of visitors at the London Science Museum [25], the category of 'general visitors' was identified as those seeking general information (e.g., museum opening hours or prices); whilst all other user groups (educational visitors and specialist visitors) were seeking more detailed information. Similarly, the CULTURA project identified 3 groups (professional researchers, apprentice investigators, informed users) who shared some level of domain knowledge. All other visitors were categorised as general public [26].

Although work on identifying user groups has tended to be mostly manual, there has also been use of computational approaches (mainly using cluster analysis) to identify representative users. For example, Krantz et al. [27] used a k-means method to explore and segment a number of museum audiences. Nyaupane et al. [28] identified 3 clusters based on motives for learning cultural heritage of the visitors to Native American heritage sites. These clusters were identified as 'culture-focused', 'culture-attentive' and 'culture-appreciative' with each showing distinct behaviours and experiences. There are many algorithms for doing this, with their use based on the types of data being used, as well as the desired outcome [29, 30, 31, 32]. In our work, we use cluster analysis to group 'similar' users and identify the characteristics of the groups.

3 Methodology

The methodology used in this paper comprised the following main steps (similar to [33]): (i) data collection and preparation; (ii) multivariate analysis; (iii) cluster analysis (assess cluster tendency, run algorithms, validate cluster quality and stability, profile clusters); and (iv) classification of user groups. The steps are described in more detail below.

3.1 Data Collection And Preparation

The dataset was collected in 2016 using an intercept pop-up survey on the NML website. The survey comprised 21 questions to gather information around users' demographics (e.g., age, gender, education, location, cultural heritage knowledge/experience and employment status), interactions with the NML website (e.g., frequency of use), and context of their visit to the website (e.g., purpose and motivation) when answering the survey. More information can be found in [14]. Overall, we obtained 514 complete responses that are used in this study.

From the 21 possible questions to use as variables in the study, 9 were deemed important in profiling users based on the results of our previous analysis [12]. All selected variables were categorical (nominal and ordinal): website visit reason (nom, 4 levels), website visit purpose (nom, 9 levels), frequency of website visit (ord, 5 levels), level of domain knowledge (ord, 4 levels), level of general CH knowledge (ord, 5 levels), location (nom, 5 levels), age group (ord, 5 levels), employment status (nom, 8 levels) and user group (nom, 8 levels). The last variable reflects a self-assigned user group: Academic (25), General Public (253), Museum Staff (10), Non-Professional (137), Other (26), Professional (5), Student (33), Teacher (25).

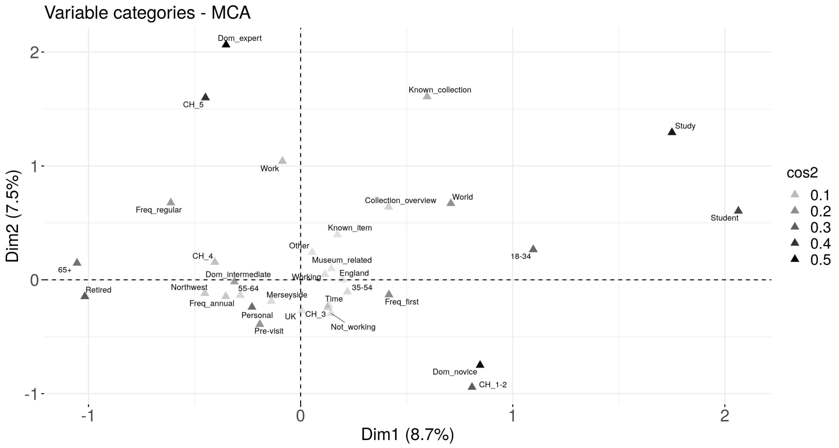
Further preprocessing included removing cases with 'unknown' responses (e.g. for levels of knowledge). We also merged categories (e.g. those with low counts) to reduce the number of variable categories. For example, we combined 'daily', 'weekly' and 'monthly' frequency of visit into a single 'regular' category. The resulting dataset was reduced to 487 cases. For the purposes of cluster analysis, the sample size is adequate, according to Qiu and Joe [34] who suggest that the sample size should be a minimum of 10 times the number of variables.

3.2 Multivariate Analysis

Prior to further analysis, dimensionality reduction was run with categorical variables. In particular, Multiple Correspondence Analysis (MCA), an extension of Correspondence Analysis, was used to identify potential relationships between variables and a lower number of dimensions that can represent the variability in the dataset without losing important information [35]. MCA is similar to PCA; however, it can be used on multiple categorical variables. The use of multivariate analysis enables insight and also helps to confirm our understanding of the data.

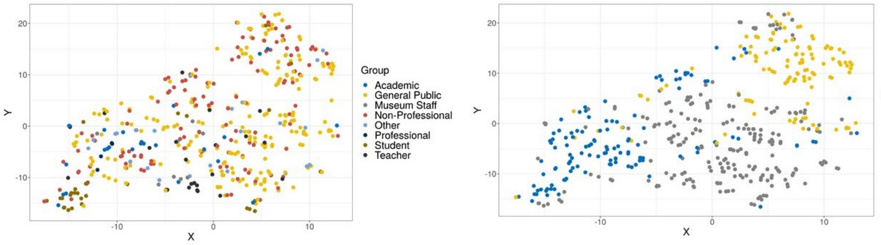
We find that the first 5 dimensions account for 34.8% of the variance in the data. Figure 1 shows an MCA plot for individual variable categories on

Fig. 1: MCA plot showing grouping of individual variable categories on first 2 dimensions



the first 2 dimensions (representing 16.2% of variance), with the shading of the points representing their squared cosine (cos2) score - this measures the degree of association between variable categories and a particular axis. The plot confirms what we might expect to see: that variable categories with a similar profile are grouped (e.g., 65+ and retired), that some variables are well represented on the dimensions (e.g., student, study, retired) and that some variables are negatively correlated and positioned in opposing quadrants (e.g., expert and CH_5 vs novice and CH_1-2). The results of this initial analysis also confirmed the findings of our previous analysis to distinguish characteristics of self-assigned groups [12].

Fig. 2: 2D t-SNE plots showing users by self-assigned groups (left) and PAM clusters (right)



We also applied non-linear dimensionality reduction using t-SNE that allows the visualisation of data in a lower-dimensional space, such as 2D, to identify patterns and trends [36]. Figure 2 shows example t-SNE plots for users by self-assigned group (left) and assigned cluster (right). We observe that points for the self-assigned groups are not as clearly separated as compared to those based on the cluster number.

3.3 Cluster Analysis

The overall approach to clustering followed the 4 steps in [37]: (i) data preprocessing; (ii) clusterability evaluation; (iii) select and run algorithms; and (iv) cluster evaluation. In addition to the data preprocessing already described, we also computed a dissimilarity matrix using Gower distance. For assessing clusterability, we computed the Hopkins statistic and manually inspected a visualisation of the dissimilarity matrix, looking for blocks of similar colour.

To perform cluster analysis, we reviewed approaches that could be used on nominal and ordinal categorical data (e.g., [38]). One group of methods is based on using a dissimilarity matrix (distance-based); another group can be applied directly to the data and model class probabilities (model-based). We opted for the simpler first approach whereby a dissimilarity matrix is first computed using the Gower distance that can handle multiple data types (in this case, using an adapted version of Manhattan distance for ordinal and the Dice coefficient for nominal categories). The dissimilarity matrix can then be used with standard clustering methods. In our case, we used PAM (Partitioning Around Medoids) [39], a partition-based algorithm that works similarly to k-means clustering, but cluster centres are restricted to be the observations themselves (i.e., medoids). Compared to k-means, the algorithm is more robust to noise and outliers and also has the benefit of having an observation serve as the exemplar for each cluster, thereby making cluster interpretation easier. Different approaches were used to determine the optimum number of clusters and evaluate cluster quality.

3.4 Classification

In this work for classification, we use a Random Forest (RF) classifier, a popular learning algorithm that builds many trees on bootstrapped copies of the training data [40]. Bagging aggregates the predictions across all trees and commonly gives good predictive performance with little hyperparameter tuning. For training the RF model, we split the data into train and test sets (70:30) and apply 5-fold cross-validation repeatedly (3 times). The Gini coefficient is used as the split rule for building the trees, and we also apply tuning using a grid search over the hyperparameter space, varying the *mtry* and *min.node.size* parameters. Feature importance is assessed using impurity. We also experimented with varying the *num.trees* parameter (starting with the suggested 10 x number of features) and settled on 400. The trained model is applied to the test data, and reported accuracy scores are based on these predictions.

4 Results And Analysis

4.1 Analysing The Self-assigned Groups

We first analysed the data to determine whether the self-assigned groups are separable. By inspecting the left-hand t-SNE plot in Figure 2 we can see that the self-assigned groups, on the whole, tend to spread across the plot, suggesting high overlap (including the General Public group). To further test this, we compared the output of PAM clustering (with $k = 8$) against the self-assigned groups using the Adjusted Rand Index (ARI). The result is a very low score of 0.021, indicating almost no overlap. Classifying the users by their self-assigned group results in an overall accuracy of 0.5315 (No Information Rate, NIR = 0.4895) on the test data, which is not that high. However, since a significant fraction is correctly classified, we hypothesised that using clustering it should be possible to create a potentially, smaller set of more distinguishable user groups, which would be easier to cater to.

4.2 Clustering And Classifying All Users

We first checked for cluster tendency using the Hopkins statistic. A resulting score of 0.1826 is well below 0.5 suggesting the data is clusterable and therefore suitable for cluster analysis. We next perform cluster analysis on all users to investigate what groups emerge from the data. Using the *fviz_nbclust* package, we compute total WSS, average silhouette width and the gap statistic to identify the optimum number of clusters for PAM. The metrics suggest 3, 3, and 9 cluster solutions respectively. Opting for the majority solution we perform PAM with $k = 3$. The average silhouette width is 0.16 and cluster medoids are shown in Table 1. The representatives reflect the mode value for each of the categories and therefore hide some of the variation within the groups.

Cluster	**Visit Reason**	**Visit Purpose**	**Freq**	**Domain Know**	**CH Know**	**Location**	**Age**	**Emp Status**
1	personal	other	first	intermediate	3	world	35-54	working
2	personal	pre-visit*	first	intermediate	4	northwest	65+	retired
3	personal	pre-visit*	annual	intermediate	3	merseyside	35-54	working

**Pre-visit refers to website visitors preparing or planning for a physical museum visit.*

Table 1: Cluster representatives for PAM clustering ($k = 3$) of all users

However, inspecting the distribution of individual categories and exemplars within the clusters, we can summarise the clusters as follows:

- **Cluster 1 - Online Researchers**: part- and full-time workers (including students) visiting the website for a wide mix of reasons (including work or study), mainly seeking known items or collections and information about the museum, often first-time visitors with a range of domain knowledge (mostly intermediate) but higher CH knowledge, mostly aged between 18 and 54 and from outside the UK, therefore less likely to visit the museum in person (125 users).
- **Cluster 2 - CH Enthusiasts**: mostly first-time and annual visitors to the website for personal reasons, perhaps preparing for a physical visit but also a range of other museum-related and other purposes, generally intermediate levels of domain and CH knowledge, predominately working or retired and aged 55+ and located in the Northwest of England and Merseyside (126 users).

- **Cluster 3 - Local Visiting Workers**: Typically, regular users, visiting the website for personal reasons and to pass time (although in a working capacity), mostly preparing for a visit to the museum, generally lower level of domain knowledge but an intermediate level of CH knowledge, mainly in the 35-64 age range and working from the local Merseyside area (236 users).

Inspecting the t-SNE plot (right) for the PAM clustering in Figure 2 would suggest that the clustering forms clear groups - the top right set of points is clearly representing Cluster 2, which seems to map onto mostly the retired and CH enthusiasts user group. Cluster 1 at the bottom left includes the student user group (amongst others). To check the stability of the clusters, we apply bootstrapping using the R *clusterboot* package. This runs PAM multiple times on samples of the data and compares the cluster outputs to determine how many points remain in the sample. The mean scores (1=all points remain in the same cluster) for the 3 clusters are 0.7432, 0.6770 and 0.8044, suggesting the first and third clusters are the most stable.

Training the Random Forest classifier on clusters from PAM, we obtain an accuracy of 0.9306 (NIR=0.4861) on the test data (0.9086 on the training data). Assessing global feature importance, variables are ranked as follows (by impurity): location (100), age (66.4), visit purpose (58.6), frequency (52.6), employment (52.5), CH knowledge (14.8), visit reason (6.2) and domain knowledge (0).

4.3 Clustering And Classifying General Public Users

In this section, we focus on the users who have identified themselves as General Public (or General Users) for the purposes of the survey. As this is a dominant group for NML (and DCH more generally [13, 25]), we wanted to establish the homogeneity of this group, any sub-clusters and their defining characteristics. In prior work [12], we found that general users could often be distinguished as using the museum for personal use, often visiting for the first time, novice/intermediate domain knowledge, medium levels of CH knowledge, mainly from Merseyside/Northwest and generally in the mid-life age range.

In Section 4.1 we find that when classifying based on all self-assigned groups, the overall classification accuracy is low (0.5315). Furthermore, performing binary classification for GP vs Other, we obtain an overall accuracy of 0.6966 (NIR=0.5172) on the test data. Inspecting the GP class only, we obtain an accuracy of 0.83. Overall, we find the GP group shares similarities with other groups (see the t-SNE plot in Figure 2), although there are still potential differences that can be used to automatically distinguish this group, suggesting that the group is fairly homogeneous.

Cluster	**Visit Reason**	**Visit Purpose**	**Freq**	**Domain Know**	**CH Know**	**Location**	**Age**	**Emp Status**
1	personal	pre-visit	annual	intermediate	4	merseyside	35-54	working
2	personal	pre-visit	first	intermediate	3	northwest	65+	retired
3	personal	pre-visit	first	novice	3	england	35-54	working

Table 2: Cluster representatives for PAM clustering ($k = 3$)

Prior to performing clustering to identify potential subgroups within the general public users (236 users), we first check for clusterability using the Hopkins statistic (0.20) and visual inspection of a visualisation of the dissimilarity matrix. We conclude that this sub-group may contain clusters. Similarly to clustering all users with PAM, we start by computing a dissimilarity matrix using the Gower distance (using *daisy*). We then seek to identify the optimum number of clusters using the total within-cluster sum of square (WSS), average silhouette score and gap statistic. This time all metrics output $k = 3$, which we use for clustering with PAM. The resulting clustering has an average silhouette width of 0.18. The cluster medoids are shown in Table 2. We might summarise the groups as follows:

- **Cluster 1 - Regular Website Visiting Local Workers**: generally users mainly visiting the website on a regular or annual basis for personal reasons, including preparing for a visit and seeking museum-related information, mostly intermediate and higher levels of domain and CH knowledge, working, aged 35-64 and local to the Merseyside area (99 users).
- **Cluster 2 - Local Enthusiasts**: also mainly using the site for personal use (and pass time) and preparing for a visit; however, mostly first-time and annual website visitors with intermediate levels of domain and CH knowledge, mostly 55+, retired and from the Northwest and Merseyside (64 users).
- **Cluster 3 - First-time Non-local Workers**: mostly first-time users of the website using the website for personal use and to pass time, mostly in preparation for a visit; generally working with lower levels of domain and CH knowledge, mostly middle-aged 35-64 and from England but outside the Merseyside area (73 users).

The RF classifier is trained, with the target variable being the cluster number from PAM. Using a similar experimental setup as before, we obtain an accuracy of 0.8841 (NIR=0.4203) on the test data (0.8980 on training data). Again, using impurity to calculate global feature importance, the variables are ranked as follows: employment (100), frequency (68.9), location (62.6), domain knowledge (50.9), age (50.4), CH knowledge (24.4), visit purpose (7.2) and visit reason (0).

5 Discussion

In this section, we summarise our findings and revisit the questions posed in Section 1.

[RQ1] How do cluster analysis results compare with the self-assigned groups? The results of the self-assigned groupings differ from the clusters based on the cluster analysis. This is evident from our analysis in that the self-assigned groups (using features collected from the users) tend to overlap. This is most clearly seen in the classification with self-assigned groups as the target feature where the classification accuracy is very low. In comparison, the groups based on clustering are fewer in number (using the simplest solution) and more distinct, resulting in far higher overall classification scores. We deduce three main categories of users from the overall data: online researchers, CH enthusiasts and local visiting workers that may provide much simpler (and more distinct) categories of users for NML to cater for.

[RQ2] Do sub-groups exist within the self-assigned General Public group? Much previous work has analysed the rather mystical 'general public' user that for NML provide a dominant user category. We find this group seems largely homogeneous but distinct and separable from the other groups - this is seen in the far higher classification accuracy for the GP class. However, potential sub-groups within the GP user group are identifiable, which we have labelled as: regular website visiting local workers, local enthusiasts, and first-time non-local workers. The first group may represent off duty teachers from the local area preparing for a personal visit or searching family history in their own time; the second group mainly reflects hobbyists and enthusiasts (e.g. interested in local history and genealogy). Finally, the third group may represent groups such as culture tourists, who are from outside the local area, arrive at the website and only view one or two pages before leaving. There are clearly similarities between the GP sub-groups and clusters obtained using all users (Tables 1 and 2). This may suggest that we do not need to cater for this group separately as they may be implicit within all groups already.

[RQ3] Can we classify the users based on the identified clusters? Overall, we are able to classify the users based on the features derived from the online survey and using the groups derived from cluster analysis. We have shown that the results on the clusters are far higher than the self-assigned groups, although we do find that the General Public user group can be distinguished using classification. This suggests we could automatically identify this group and then apply cluster analysis to further segment the group if desired.

6 Conclusions And Future Work

In this paper, we have extended prior work on studying users of the National Museums Liverpool websites using cluster analysis and classification. Based on a sample of users taken from an online survey, We have shown that a smaller set of more distinct groups exists, which may be easier to cater for than using self-assigned groups that commonly overlap and share characteristics. We also find that the General Public group (often treated as one group) may contain sub-groups. However, these reflect clusters from all users and may alleviate the need to model them separately. In future work we plan to experiment with further approaches for automatically profiling users, such as automated persona generation, comparing clustering methods and identifying ways of automating the process to alleviate the need for gathering data from user surveys for categorising online visitors (e.g., using relevant features from transaction logs).

References

1 Hadley, W.: Covid-19 impact museum sector research findings (2020) [Accessed March 23rd, 2021].

2 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

3 Case, D.: Looking for information: a survey of research on information seeking, needs, and behavior. Library and information science (2007)

4 Taylor, R.S.: Value-added processes in the information life cycle. Journal of the American Society for Information Science **33**(5) (1982) 341–346

5 Johnson, A.: Users, use and context: supporting interaction between users and digital archives. What Are Archives?: Cultural and Theoretical Perspectives: A Reader (2008) 145–64

6 Whitelaw, M.: Generous interfaces for digital cultural collections. Digital Humanities Quarterly **9**(1) (2015) 38

7 Allen, R.B.: Chapter 3 - mental models and user models. In Helander, M.G., Landauer, T.K., Prabhu, P.V., eds.: Handbook of Human-Computer Interaction (Second Edition). Second edition edn. North-Holland, Amsterdam (1997) 49–63

8 Cifter, A.S., Dong, H.: User characteristics: Professional vs. lay users (2009)

9 Salminen, J., Jung, S.G., Chowdhury, S., Robillos, D.R., Jansen, B.: The ability of personas: An empirical evaluation of altering incorrect preconceptions about users. International Journal of Human-Computer Studies (2021) 102645

10 Walsh, D., Hall, M.: Just looking around: Supporting casual users initial encounters with digital cultural heritage. In Gade, M., Hall, M., Huurdeman, H., Kamps, J., Koolen, M., Skov, M., Toms, E., Walsh, D., eds.: Proceedings of the First International Workshop on Supporting Complex Search Tasks co-located with the 37th European Conference on Information Retrieval (ECIR 2015). Volume 1338 of CEUR Workshop Proceedings., CEUR-WS.org (March 2015)

11 Walsh, D., Hall, M., Clough, P., Foster, J.: The ghost in the museum website: Investigating the general public's interactions with museum websites. In Kamps, J., Tsakonas, G., Manolopoulos, Y., Iliadis, L., Karydis, I., eds.: Research and Advanced Technology for Digital Libraries, Cham, Springer, Springer International Publishing (2017) 434–445

12 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 **21**(1) (Mar 2020) 75–87

13 Villaespesa, E.: Museum collections and online users: Development of a segmentation model for the metropolitan museum of art. Visitor Studies **22**(2) (2019) 233–252

14 Walsh, D., Clough, P., Foster, J.: User categories for digital cultural heritage. In Clough, P., Goodale, P., Agosti, M., Lawless, S., eds.: Proceedings of the

First International Workshop on Accessing Cultural Heritage at Scale co-located with Joint Conference on Digital Libraries 2016 (JCDL 2016). Volume 1611 of CEUR Workshop Proceedings., CEUR-WS.org (June 2016)

15 Pantano, E.: Virtual cultural heritage consumption: a 3d learning experience. International Journal of Technology Enhanced Learning **3**(5) (2011) 482–495

16 Marty, P.F.: Meeting user needs in the modern museum: Profiles of the new museum information professional. Library & information science research **28**(1) (2006) 128–144

17 Hogg, C., Williamson, C.: Whose interests do lay people represent? towards an understanding of the role of lay people as members of committees. Health Expectations **4**(1) (2001) 2–9

18 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

19 Kelly, L.: The interrelationships between adult museum visitors‘ learning identities and their museum experiences. chapter 3. Methodology (2007) 3–46

20 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.

21 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

22 Spellerberg, M., Granata, E., Wambold, S.: Visitor-first, mobile-first: Designing a visitor-centric mobile experience. In: Museums and the Web. (2016)

23 Ardissono, L., Kuflik, T., Petrelli, D.: Personalization in cultural heritage: The road travelled and the one ahead. User Modeling and User-Adapted Interaction **22**(1-2) (April 2012) 73–99

24 Falk, J.H.: Identity and the museum visitor experience. Left Coast Press (2009)

25 Booth, B.: Understanding the information needs of visitors to museums. Museum Management and Curatorship **17**(2) (1998) 139–157

26 Sweetnam, M., Siochru, M., Agosti, M., Manfioletti, M., Orio, N., Ponchia, C.: Stereotype or spectrum: Designing for a user continuum. In: the Proceedings of the First Workshop on the Exploration, Navigation and Retrieval of Information in Cultural Heritage, ENRICH. (2013)

27 Krantz, A., Korn, R., Menninger, M.: Rethinking museum visitors: Using k-means cluster analysis to explore a museum's audience. Curator: The Museum Journal **52**(4) (2009) 363–374

28 Nyaupane, G.P., White, D.D., Budruk, M.: Motive-based tourist market segmentation: An application to native american cultural heritage sites in arizona, usa. Journal of Heritage Tourism **1**(2) (2006) 81–99

29 Ackerman, M., Ben-David, S., Loker, D.: Towards property-based classification of clustering paradigms. In: Advances in Neural Information Processing Systems. (2010) 10–18

30 Ackerman, M., Ben-David, S., Loker, D.: Characterization of linkage-based clustering. In: COLT. (2010) 270–281

31 Ackerman, M., Ben-David, S.: Discerning linkage-based algorithms among hierarchical clustering methods. In: Twenty-Second International Joint Conference on Artificial Intelligence. (2011)

32 Ackerman, M., Ben-David, S., Brânzei, S., Loker, D.: Weighted clustering. In: Twenty-Sixth AAAI Conference on Artificial Intelligence. (2012)

33 Brida, J.G., Meleddu, M., Pulina, M.: Understanding museum visitors' experience: a comparative study. Journal of Cultural Heritage Management and Sustainable Development **6**(1) (2016) 47–71

34 Qiu, W., Joe, H.: Generation of random clusters with specified degree of separation. Journal of Classification **23**(2) (2006) 315–334

35 Brickey, J., Walczak, S., Burgess, T.: A comparative analysis of persona clustering methods. In: AMCIS. (2010) 217

36 Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)

37 Adolfsson, A., Ackerman, M., Brownstein, N.C.: To cluster, or not to cluster: An analysis of clusterability methods. Pattern Recognition **88** (2019) 13–26

38 Preud'homme, G., Duarte, K., Dalleau, K., Lacomblez, C., Bresso, E., Smaïl-Tabbone, M., Couceiro, M., Devignes, M.D., Kobayashi, M., Huttin, O., et al.: Head-to-head comparison of clustering methods for heterogeneous data: a simulation-driven benchmark. Scientific reports **11**(1) (2021) 1–14

39 Kaufman, L., Rousseeuw, P.J.: Finding groups in data: an introduction to cluster analysis. Volume 344. John Wiley & Sons (2009)

40 Breiman, L.: Random forests. Machine Learning **45**(1) (2001) 5–32