Sheffield Submission to the CHiC Interactive Task: Exploring Digital Cultural Heritage

Mark Hall[1][2], Robert Villa[2], Sophie A Rutter[2], Daniel Bell[2], Paul Clough[2], and Elaine Toms[2]

[1] m.mhall@sheffield.ac.uk
Department for Computer Science
University of Sheffield
Sheffield, UK

[2] {r.villa,sarutter1,dcbell1,p.d.clough,e.toms}@sheffield.ac.uk
Information School
University of Sheffield
Sheffield, UK

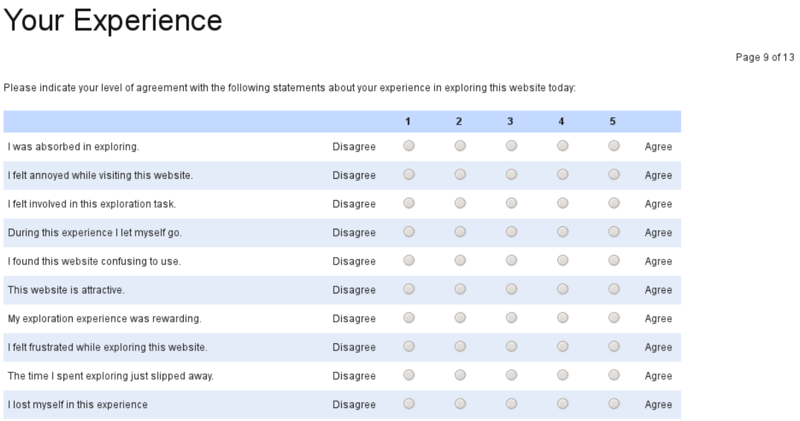
Abstract. The Cultural Heritage in CLEF 2013 (CHiC) interactive task focused on acquiring and analysing interactive information retrieval (IIR) behaviour in a Digital Cultural Heritage collection. The University of Sheffield contributed 120 on-line and 20 in-lab participants to this task. The results of both the on-line and in-lab experiments strongly indicate that when faced with a new, unfamiliar collection and an open-ended task, participants will spend more time using the category hierarchy for exploration, than the search box. However, analysis of the the number of items the on-line participants view in detail and then saved to their workspace indicates that the two access methods fulfil different functions. From this we conclude that the categories are seemingly there to support the development of an initial overview over the collection, while the search is used to locate things in a more focused manner.

1 Introduction

The Cultural Heritage in CLEF 2013 (CHiC) lab included an interactive task aimed at developing a data-set describing undirected exploration and browsing in a collection of approximately 1.1 million English-language Cultural Heritage items. The task required lab-participants to recruit at least 10 experiment participants each for an on-line and an in-lab experiment. In total at the University of Sheffield we contributed 120 on-line and 20 in-lab participants to the pool, all of which were recruited via a Sheffield University volunteers' mailing list. Due to the lab's protocol, the on-line experiment results below are derived from the complete pool of on-line participants (160), while the in-lab analysis uses only the 20 participants recruited at Sheffield.

The two experiments followed the same core research protocol. The difference between them is that the in-lab participants physically attended our usability lab and, after completing the core research protocol, participated in an extended playback and interview session.

The complete research protocol was presented using a web-based system developed at Sheffield as part of the PROMISE[3] project. Figure 1 shows a screenshot of one of the questionnaire pages. Figure 2 shows the main task user interface (UI), consisting of the task instructions, category browser, search box and results, item details display, and a book-bag for saving items. For the main task participants were isntructed to explore the collection using the task UI until they were bored, and to add anything they found interesting to their book-bag.



Your Experience

Page 9 of 13

Please indicate your level of agreement with the following statements about your experience in exploring this website today:

		1	2	3	4	5	
I was absorbed in exploring.	Disagree	○	○	○	○	○	Agree
I felt annoyed while visiting this website.	Disagree	○	○	○	○	○	Agree
I felt involved in this exploration task.	Disagree	○	○	○	○	○	Agree
During this experience I let myself go.	Disagree	○	○	○	○	○	Agree
I found this website confusing to use.	Disagree	○	○	○	○	○	Agree
This website is attractive.	Disagree	○	○	○	○	○	Agree
My exploration experience was rewarding.	Disagree	○	○	○	○	○	Agree
I felt frustrated while exploring this website.	Disagree	○	○	○	○	○	Agree
The time I spent exploring just slipped away.	Disagree	○	○	○	○	○	Agree
I lost myself in this experience	Disagree	○	○	○	○	○	Agree

Fig. 1. Screenshot showing the questionnaire UI used in the experiments.

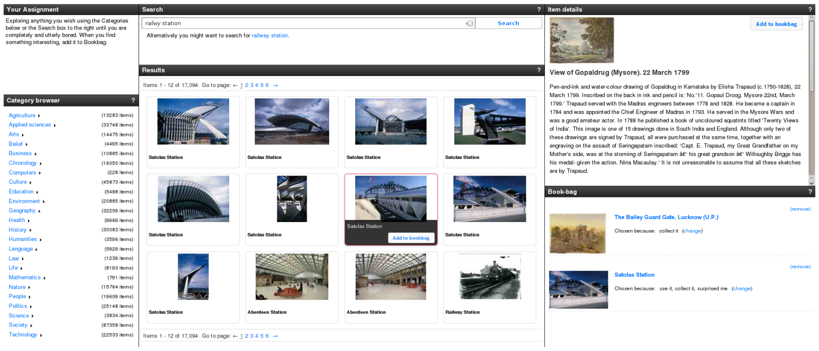
The two experiments are being analysed separately, as figure 3 shows that the in-lab participants spent significantly longer on the task, than the on-line participants.

2 On-line Experiment

For the on-line experiment we analysed the interaction logs gathered by the task UI for all 160 on-line participants. No filtering of participants was performed, thus the analysis includes participants who spent only a few seconds with the task and those who stayed for a few hours. The focus of the analysis is on which of the exploration methods provided by the task UI were used by the participants.

The task UI provided three methods for the participants to explore the collection. On the left there was a category browser, that showed a hierarchical structure into which a sub-set of the items in the collection (approximately 250,000) had been mapped automatically [2]. The second option was to use the search box to type in and run a query. The third method was to click on an

[3] http://www.promise-noe.eu



Exploring anything you wish using the Categories below or the Search box to the right until you are completely and utterly bored. When you find something interesting, add it to Bookbag.

Agriculture ▸	(13283 items)
Applied sciences ▸	(33748 items)
Arts ▸	(14475 items)
Belief ▸	(4495 items)
Business ▸	(10885 items)
Chronology ▸	(16050 items)
Computers ▸	(228 items)
Culture ▸	(45873 items)
Education ▸	(5488 items)
Environment ▸	(20885 items)
Geography ▸	(32209 items)
Health ▸	(9969 items)
History ▸	(30083 items)
Humanities ▸	(3596 items)
Language ▸	(5929 items)
Law ▸	(1239 items)
Life ▸	(8193 items)
Mathematics ▸	(791 items)
Nature ▸	(15764 items)
People ▸	(19609 items)
Politics ▸	(25148 items)
Science ▸	(3834 items)
Society ▸	(87358 items)
Technology ▸	(22533 items)

railwy station

Alternatively you might want to search for railway station.

Items 1 - 12 of 17,094 Go to page: ← 1 2 3 4 5 6 →

Satolas Station	Satolas Station	Satolas Station	Satolas Station
Satolas Station	Satolas Station	Satolas Station	Satolas Station
Satolas Station	Aberdeen Station	Aberdeen Station	Railway Station

Items 1 - 12 of 17,094 Go to page: ← 1 2 3 4 5 6 →

View of Gopaldrug (Mysore). 22 March 1799

Pen-and-ink and water-colour drawing of Gopaldrug in Karnataka by Elisha Trapaud (c.1750-1828), 22 March 1799. Inscribed on the back in ink and pencil is: No.'11. Gopaul Droog. Mysore 22nd, March 1799.' Trapaud served with the Madras engineers between 1778 and 1828. He became a captain in 1784 and was appointed the Chief Engineer of Madras in 1793. He served in the Mysore Wars and was a good amateur actor. In 1788 he published a book of uncoloured aquatints titled 'Twenty Views of India'. This image is one of 15 drawings done in South India and England. Although only two of these drawings are signed by Trapaud, all were purchased at the same time, together with an engraving on the assault of Seringapatam inscribed: 'Capt. E. Trapaud, my Great Grandfather on my Mother's side, was at the storming of Seringapatam â€" his great grandson â€" Willoughby Briggs has his medal- given the action. Nina Macaulay.' It is not unreasonable to assume that all these sketches are by Trapaud.

The Bailey Guard Gate, Lucknow (U.P.)

Chosen because: collect it (change)

Satolas Station

Chosen because: use it, collect it, surprised me (change)

Fig. 2. Screenshot showing the task UI used in the experiments. The UI consists of six parts. On the left are the task instructions and the category browser. In the middle are the search box and the search results. On the right are the viewer for an item's details and the book-bag used to store interesting items. Selecting a category on the left, entering search terms, or clicking on a meta-data facet in the item viewer all resulted in the items matching that request being displayed in the central 4 × 3 grid.

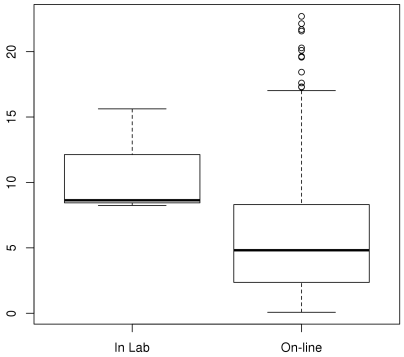
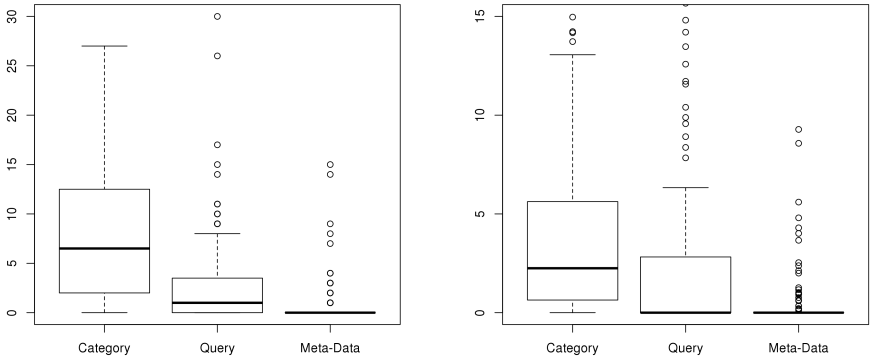


Fig. 3. Box-plots showing the amount of time (in minutes) the participants spent using the task UI. The lab participants clearly spent significantly longer on the task than the on-line participants.

item's meta-data, which would run a search for other items with the same meta-data. In all three cases, the items for the selected category, user-provided query, or meta-data query would be shown in the central grid.

The task UI logged all interactions between the participants and the UI and for each participant we processed the log and determined the number of topics selected (labelled "Category" in the graphs), user-queries run ("Query"), and meta-data queries selected ("Meta-Data"). In addition to the number of interactions with each exploration method, we also calculated how much time each participant spends using each of the methods. This was calculated by measuring the time between when the participant first used that method and the moment when they switched to a different method. If the participants switched back and forth between methods, then the individual times were aggregated.

Figure 4 shows the resulting box-plots. Clearly visible is that participants used the category browser more to explore the collection, than either of the other two methods. Both the number of categories they looked at and also how much time they spent with the task UI after selecting a category are higher (statistically significant in both cases using paired Wilcoxon signed rank test [$p < 0.01$]). Exploring the collection via the "Meta-Data" queries is not frequently used, with only 32 participants making use of the facility at all and of those most only used it once or twice. From the on-line logs we cannot determine whether this is due to people not being interested in this kind of exploration, or whether the ability was not clear from the UI.

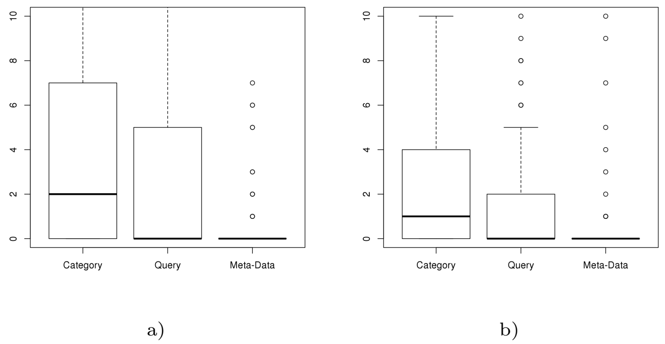


a)

b)

Fig. 4. Box-plots showing: a) the number of interactions in each of the three exploration methods; b) the amount of time (in minutes) spent on each of the three exploration methods. "Category" refers to using the category browser, "Query" to typing a query into the search box, and "Meta-Data" to exploring the collection by clicking on the meta-data facets shown in the item viewer.

We also counted how many items participants viewed for the three methods and how many they then saved to their book-bag. Figure 5 shows that while participants tended to find more items via the categories, the differences are no longer statistically significant using a Wilcoxon paired signed-rank test.



a)

b)

Fig. 5. Box-plots showing: a) the number of items viewed ; b) the number of items saved to the book-bag when using each of the exploration methods. "Category" refers to using the category browser, "Query" to typing a query into the search box, and "Meta-Data" to exploring the collection by clicking on the meta-data facets shown in the item viewer.

Finally we used affinity propagation clustering [1] (using the scikit learn package[4]) to see if we could identify clusters of participants with similar behaviour (default parameters used except for a damping value of 0.88). Participants were clustered based on the time they spent using each of the three exploration methods. The times were normalised for each participant using the total time the participant spent on the task. Thus a value of 0 means that the participant spent no time using that exploration method, while a score of 1 indicates that they only used that method.

Table 1 shows the six clusters that are identified in the data, including the participant that has been identified as the exemplar for the cluster, and a classification label that was manually attached. Clearly visible is that the majority of participants use a mixed-mode approach, spending slightly more time exploring the collection with the categories. Approximately a quarter of the participants used the hierarchy very briefly and spent most of their time searching for things. Interestingly, only four users restricted themselves to exploring only via the category browser. The final cluster #6 is somewhat misleading, as it represents not people who have done nothing, which the sample data implies, but instead contains 12 participants who don't fit into any of the other clusters. The sample is thus not representative in this case.

[4] http://scikit-learn.org

Cluster #	Participants	Category	Query	Meta-Data	Classification
1	4	1.	0.	0.	Browser
2	83	0.687	0.313	0.	Mixed-mode, browsing preferred
3	19	0.506	0.494	0.	Mixed-mode
4	26	0.107	0.893	0.	Mostly search
5	16	0.011	0.989	0.	Searcher
6	12	0.	0.	0.	Other

Table 1. The six clusters found using affinity propagation clustering. The "Category", "Query", and "Meta-Data" columns show the fraction of their task time each cluster's exemplar participant spent on each of the three exploration methods. Cluster #6 is a catch-all cluster for participants that fit into none of the other clusters, thus the exemplar participant data should be ignored.

2.1 Discussion

The results strongly indicate that when faced with a new, unfamiliar collection and an open-ended task, participants will spend more time using the category hierarchy for exploration, than the search box. However, the number of items participants viewed in detail and then saved to their workspace indicates, that the two access methods fulfil different functions. The categories are seemingly there to support the development of an initial overview, while the search is used to locate things in a more focused manner. We intend to investigate this in more detail by analysing the participants qualitative descriptions of why and how they completed their task, and also by performing a temporal analysis of the log to see whether participants are more likely to use the category hierarchy at the beginning of their task and then switch to using the search when they are more familiar with the collection.

3 In-Lab Experiment

One of the focuses of the in-lab CHiC study is how individuals explore a cultural heritage collection when given no task. The results may be used both to contrast with studies which have used explicit tasks, and to motivate changes to cultural heritage systems to better support a diverse range of user tasks.

The work reported here is based on initial results. Here, our focus is on how users explored the collection and in particular how search and browse were used in this exploration. We consider three research questions:

- RQ1. How do participants initiate their exploration?
- RQ2. Do participants have a preference for browse or search in their exploration of the collection?
- RQ3. Can reasons be identified for why users chose to search or browse?

With RQ1 we are particularly interested whether users start their exploration by browsing categories, or by search. RQ2 then considers how users access the collection over their whole session. For RQ3 we will present some initial qualitative data from our lab-based interactive study, where the aim is to identify reasons for the use of either the search or browse functions.

3.1 Experimental setup

All data reported here is from an in-lab study, which allowed a follow up interview to be carried out, during which each participant reviewed his or her search session. To enable this reviewing, Morea screen recording software was used to record the users activity, and during the interview, an audio recording was made of the users comments.

An important aspect of the interactive CHiC experimental design was that no explicit task was provided to users. Instead instructions asked the user to explore freely as they wished, until they were bored. Users were informed after they had been active for 10 minutes, and could then continue for a further 5 minutes if they wished, at which point they would be asked to stop (these timings were carry out by hand, and were approximate). Once this was finished, the users search session would be replayed to them, and an interview conducted to investigate the users search process. Participants were paid 10 pounds for taking part. In total 20 participants were recruited for the study, 11 male and 9 female. Eight participants were in the 18-25 year age band, nine in the 26-35 band; the other 3 between 36-45. The majority were students (13), with 5 employed, one unemployed, and one other. 13 had completed a higher education degree, while six were currently studying an undergraduate degree.

3.2 In-lab observations

Before presenting initial results of the in-lab study, we first present some observations gathered from the experimenters who carried out the study. In total, four different individuals were involved in running the in-lab experiment (it should be noted that a detailed procedure was developed to ensure consistency, and an initial pilot was carried out). The list below is therefore informal, being the perceptions of the experimenters from observing and interviewing the participants. Some of these observations (e.g. preference for the categories) are backed up by the subsequent analysis, other items will be investigated more fully in future work:

- The categorisation of items was of a surprise to many participants, with many items being apparently “miss-classified” by the system. Categories often had confusing or incorrect content compared to what the participants expected to find (for example the categories “Computers”, “people”, and “education”). Several participants expressed confusion/dissatisfaction upon clicking on the “People” category and then being presented with coins. Generally participants felt that the items seemed ill fitted to the categories that they were in.

- Some categories had items with a lot of missing information (e.g. the titles were unintelligible, a lack of images etc.) that annoyed or confused participants.
- Some participants often failed to notice that there were subcategories or only found them by accident. The subcategories were accessed via a small arrow to the right of the category name.
- Some participants didn't notice the search box at the top of the page
- Some participants did not spot the "more like this" link till later on in the search. Some did think this was a useful feature and would like to have seen it earlier. Often scrolling was required to see these recommendations (one participant asked if this requirement to scroll was a bug)
- Larger pictures and more information for the pictures was a common request. Some participants asked to be able to zoom in on the images. The general view was that the thumbnails used did not provide enough detail for the participants to really engage with the item.
- Participants were often disconcerted when pictures did not display, thinking there was a problem with the system.
- Some participants encountered multimedia objects such as video and audio clips in the collection, but the interface seemed to be unable to play them.
- Most participants would have liked more detail/description for the items. Often the additional information provided about items (topic/type/country etc.) was not what participants expected. They either wanted to find similar items through the information (the topic information was often too specific and only showed the same item again when clicked) or they expected more information relating to that classifier (e.g.. more information on the topic or type of item).
- Most participants mentioned that they were unsure of the purpose of the book bag. Several used it like a bookmark system, but some used it in a similar manner to an e-commerce basket, where they hoped to buy copies/prints of the items later.
- Some participants mentioned that the site did not provide a clear overview of the kinds of objects that the collection contained.
- Some participants would have preferred the categories/structure to have followed a more narrative structure similar to that found in a museum. They thought that the objects lacked context and background story.

3.3 Results

Initiation of exploration RQ1 asks how users initiate their exploration of the collection. To investigate this, we first looked at how users started their session, and in particular, their searching. For example, did they select a category or enter a query?

Over the whole data set four different actions were used by participants to initiate their session (Table 2, column 2). For the majority of users, the first action was to select one of the categories (15 out of the 20 users). It should be noted that the interface, on startup, showed a set of default results to all users. For three users, the first action was to display one of these default results, another user clicked the next page to view the next page of default results, while the final users first action was to bookmark one of the default result items.

We also investigated the logs to find out each users first search or browse action, which could be one of category select, text query, or metadata/more like this select. As shown in Table 2 (column 3), for all users this was a category select. In addition to counting the first actions, we also investigated how long each user spent before either clicking the interface, or starting a new search/browse using the three previously listed methods. These results are shown in Table 3, along with the overall length of time of each session.

Action	#Users first action	#Users first search/browse action
Category select	15	20
Display item	3	-
Next search result page	1	-
Add to book-bag	1	-

Table 2. Number of users whose first action/first search or browse action were as column one.

	Min	1^{st} Qu.	Median	Mean	3^{rd} Qu	Max
First action	7.00	19.00	25.00	30.50	38.75	90.00
First search/browse	7.00	22.75	38.00	57.50	81.75	204.0
Total time	129	631.8	783.5	787.8	918.0	1544

Table 3. Time to first action, time to first search/browse action, and overall session time (all times in seconds)

There was a considerable variance in the length of time users spent on the task. The median time taken by users was 783.5 seconds (just over 13 minutes), with an interquartile range of 286.2 seconds (approximately 4 minutes, 45 seconds). The minimum time was 129 seconds, and maximum 1544 seconds (over 25 minutes).

Most users spent some time at the start of their session before either clicking on an interface element (median time 25 seconds) or initiating a search (median 38 seconds).

Search vs. browse RQ2 asks whether users have a preference for search or browse. Figure 6 presents query and category counts across all users (i.e. counts of how often either text queries were executed or categories selected). Item select and the “more like this” functionality is not included here, due to the relative rarity of these events (across the whole data set this functionality was used only 15 times, by 7 different users).

A non-parametric Wilcoxon rank-sum test indicated that there was a significant difference between queries executed and categories selected (W = 50.5, $p \leq 0.001$). As can be seen from the box-plots, categories were selected far more than queries entered, the median number of queries executed being 2, compared to a median of 11 for category selects. All but three users selected more categories than executed queries, and 8 users did not enter a text query at all.

A similar situation exists when the time querying vs. browsing categories is estimated (Figure 7). Such times were estimated by starting a timer when a query or category was selected, and taking all activity between this point and the next query or category select as the user "querying" or "browsing categories". As might be expected, the trend is similar to that of Figure 2, with users spending more time browsing categories when compared to executing queries. All but five participants spent more time browsing using the categories than spent querying.

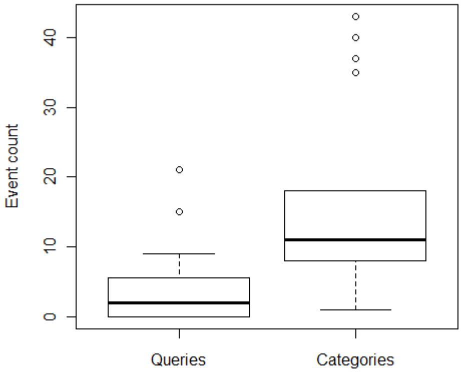


Fig. 6. Comparison of query and category select counts

"How did you start?" In addition to the quantitative data above, in the post-session interview two questions were asked of users: "how did you start?" and "Why did you choose to start with a [category/search query]?" It was intended to alter this latter question depending on how the user initiated their exploration, although in the event all users started by using the category browser. At the time of writing twelve of the twenty user interviews have been transcribed; initial results will be presented here.

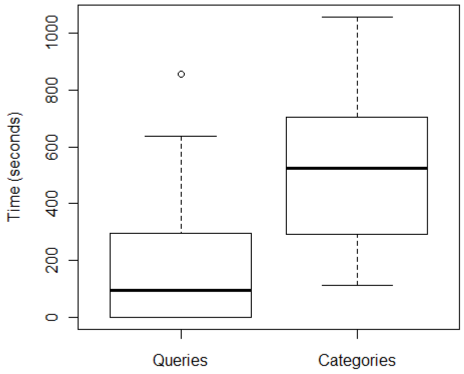


Fig. 7. Estimated time querying vs. browsing by category

The responses to the first question "how did you start?" mentioned the category browser explicitly in 8 of the 12 answers. In most of these cases this was linked to exploring the interface. For example, participant P3 stated:

"I was drawn to the middle then decided to look around at the interface. I decided to look at categories first, picked politics"

Similarly, participant P10 stated:

"I just looked round to see what I could use to explore things. The category browser looked like the most likely candidates because it had descriptions of stuff."

As well as being influenced by the interface, responses from some users suggest that prior interests also played a part. For example, participant P8 stated:

"I just look at the layout of the website and then found that I had a category browser so I went to what I study actually, and I study languages and I try to find something interesting."

The design of the interface, with a relatively small search box, appears to also have had an effect on the choices of at least two of the user, indicated by responses to the second question. Participants P2 and P4 stated:

"Because I only saw that [category]. I didn't see the search until a bit later on." [P2]

"I didn't really see this one at first [the search box] it was a bit obscure." [P4]

For many users, however, the fact that the category browser allowed easy exploration appeared to be the key, with some users making connections to physical museums. For example:

"There is no particular task and so I started from browse to see which information is more interesting to me." [P1]

"If I was going to a museum I would look at the categories [museum sections] that are of most interest to me: arts, old stuff and so this is why I was looking for Mona Lisa." [P5]

The lack of an explicit task was mentioned by some users, and search was explicitly commented on by two users. For example, participant P7 stated *"When I wanted to find something specific I went to the search box."*

3.4 Discussion

RQ1 asks how participants initiate their exploration of the collection. From Table 2 it can be seen that all 20 participants started their exploration using the category browser, rather than a text search. Indeed, the first action for the majority of users (75%) was to select a category. Quantitative data from Section 3.3 backs this up, with 8 out of 12 of the participants for which text transcripts are available explicitly mentioning the category browser as a way of starting their exploration.

Looking at Table 3, it can be seen that there is typically a short delay until participants started their browsing (median 38 seconds, interquartile range of 59). This delay is consistent with participant's comments which suggested that many first spent some time orienting themselves to the interface before starting (e.g. P10 from Section 3.3).

Moving to RQ2 and RQ3, which asked whether participants have a preference for browse or search and why, it is clear from Figure 6 and Figure 7 that there is a general preference for browsing, e.g. from Figure 7 the median estimated time spent browsing using the categories was 524 seconds (IQR 399), compared to 77 seconds (IQR 394) for text queries. Looking at the participant comments, the lack of any explicit task would appear to have played a part in this preference (e.g. P1 and P5 quotes from Section 3.3). In addition to this the design of the interface, with a relatively small text search box at the top, appeared to also play a part, with some users pointing out that they did not see the search box until later in their session (e.g. P2 and P4).

4 Acknowledgements

The research leading to these results was supported by the Network of Excellence co-funded by the 7th Framework Program of the European Commission, grant agreement no. 258191.

References

1. Brendan J. Frey and Delbert Dueck. Clustering by passing messages between data points. *Science*, 315(5814):972–976, 2007.
2. Mark M. Hall, Clough Paul D., Samuel Fernando, Mark Stevenson, Aitor Soroa, and Eneko Aguirre. Automatic generation of hierarchies for exploring digital library collections. forthcoming.