The Multi-Stage Experience: the Simulated Work Task Approach to Studying Information Seeking Stages

Hugo C. Huurdeman
h.c.huurdeman@uva.nl
University of Amsterdam

Jaap Kamps
kamps@uva.nl
University of Amsterdam

Max L. Wilson
Max.Wilson@nottingham.ac.uk
University of Nottingham

ABSTRACT

This experience paper shines more light on a simulated work task approach to studying information seeking stages. This explicit multistage approach was first utilized in Huurdeman, Wilson, and Kamps [14] to investigate the utility of search user interface (SUI) features at different macro-level stages of complex tasks. We focus on the paper's terminology, research design, methodology and use of previous resources. Finally, based on our experience, we reflect on the potential for re-using our multistage approach and on general barriers to re-use in an Interactive Information Retrieval research context.

KEYWORDS

experience paper, information seeking, search stages

1 INTRODUCTION

In the Interactive Information Retrieval (IIR) community, there is a varied range of terminology, approaches and methods. Bogers et al. [2] assert that it is not straightforward to re-use aspects and materials from previous user studies in IIR research. They list various barriers to reproducibility and re-use, which include the "fragmentary nature" of the organization of resources, the lack of awareness of their existence, insufficient documentation, the research publication cycle, and the inherent effort required for making resources available.

This experience paper shines more light on the simulated work task approach to studying information seeking stages, which we implemented in Huurdeman, Wilson, and Kamps [14]. To uncover various aspects related to re-use and reproducibility, we specifically focus on the paper's terminology, the experience of designing our user study, the adaptation of previous work and the opportunities for the re-use of our approach.

First, we summarize the original paper in Section 2. Then, we discuss the used terminology (Section 3), followed by the methodology and research design (Section 4). Section 5 discusses in which ways previous work was adapted for use in our paper. Next, we discuss the potential re-use of our approach (Section 6). Section 7 concludes this experience paper with a short reflection.

2 SUMMARY OF MULTI-STAGE STUDY

Research into information seeking behavior has shown substantial changes in user behavior during complex tasks involving learning and construction. Models of information seeking, including Kuhlthau [20]'s Information Search Process model and Vakkari [28]'s adaptation, describe fundamentally different macro-level stages. Current search systems usually do not provide support for these stages, but provide a static set of features predominantly focused on supporting micro-level search interactions. Huurdeman et al. [14] delved deeper into this paradox, and described an experimental user study employing (cognitively complex) multistage simulated work tasks, studying interaction patterns with interface and content during different search stages. In this study, a custom search system named SearchAssist was used, and tasks were designed to take users through pre-focus, focus, and post-focus task stages to gather active, passive, and subjective measures of when SUI features provide most value and support.

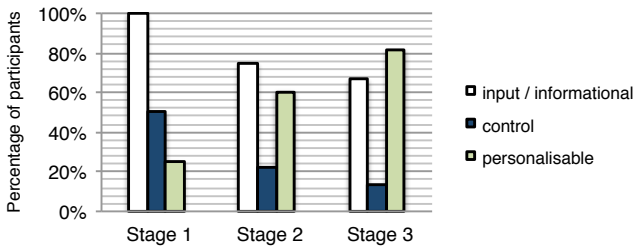

Figure 1: SUI feature categories perceived most useful by stage (from [14])

To our knowledge, this mixed methods study was the first to use an explicit multistage simulated task design using Vakkari [28]'s pre-focus, focus formulation and post-focus stages. The independent variable was task stage, the dependent variables *active utility* (via clicks and queries), *passive utility* (via mouse and eye tracking fixation counts) and *perceived utility* (via questionnaires and interviews) of search user interface features.

First, we looked at *active behaviour*, the behaviour which can be directly and indirectly determined from logged interaction, such as clicks and submitted queries. Our main finding was that some features such as informational features (providing information about results) are used frequently throughout, while input and control features (for refinement of results) are used less frequently after the first stage. Second, we looked at *passive behaviour*, i.e. behaviour not typically caught in interaction logs, such as eye fixations and mouse movements. Our main finding was the difference with the active results: evidently, users look often at actively used features, but other features that are less actively used (such as the recent queries feature) are more used in a passive way, suggesting a different type of support offered by these features. Third, we were interested in the *subjective opinions* of users about the usefulness of features; this data also formed a reference point for interpreting other observed data from the previous research questions.

BIIRRR 2019, 14 March 2019, Glasgow, UK

The paper concluded that the perceived usefulness of features differs radically per search stage, as summarised in Figure 1. First, the most familiar input and informational features (the search box and results list) were perceived as very relevant overall, but declined after the initial stage. Similarly, a set of assistive control features (search filters, tags and query suggestions), less commonly included in SUIs were also perceived as most useful in the beginning, but less useful in consecutive stages. Third, personalisable features (query history and a feature to save results) were considered as less useful in the beginning, but their usefulness significantly increases over time, even surpassing the value of common SUI features. Hence, the results of our paper suggest that the macro-level process has a large influence on the usefulness of SUI features.

3 TERMINOLOGY

As a first step in analyzing Huurdeman et al. [14], we focus on the terminology, why it was used and how it was developed.

Information behavior, seeking and searching

The paper used commonly accepted definitions in the areas of Library and Information Science (LIS), and (Interactive) Information Retrieval (IIR) to refer to information seeking and searching, concepts which were of key importance to the paper. It was framed using Wilson [33]'s definition of *information behavior*: "the totality of human behavior in relation to sources and channels of information, including both active and passive information seeking, and information use." The paper's main focus was on *information seeking* and *searching*, subsets of information behavior in Wilson's nested model of research areas [33]. We used Ingwersen and Järvelin [15, p.21]'s definition of *information seeking*: "human information behavior dealing with searching or seeking information by means of information sources and (interactive) information retrieval systems." *Information searching*, in its turn, was defined as a subfield of information seeking in Wilson's nested model, and specifically focuses on the interaction between information user and information system [33].

Following Huurdeman and Kamps [13], Wilson [33], we also distinguished between the *macro-level* described by information seeking models, and the *micro-level* of specific system and interface features, and looked at ways to bridge the gap between these levels.

Work tasks, search tasks and their complexity

In the paper, we made the distinction between work tasks and search tasks, and also based this on previous literature in the domain of LIS and IIR. We used Ingwersen and Järvelin [15, p.20]'s definition of *work task*: a "job-related task or non-job associated daily-life task or interest to be fulfilled by cognitive actor(s)". These tasks may be real-life tasks, or in our case, assigned simulated work tasks, for which we used Borlund [3]'s definition and guidance. Work tasks may lead to one or more search tasks, and we used Ingwersen and Järvelin [15, p.20]'s definition: "the task to be carried out by a cognitive seeking actor(s) as a means to obtain information associated with fulfilling a work task".

An important distinction made in the paper is between *simple* work tasks, which can for instance be solved with a single search query, and *complex* work tasks. We utilized the definition of Byström and Järvelin [5]: tasks which require "understanding, sense-making, and problem formulation" . Complex tasks go beyond simple lookup tasks, and might involve learning and construction, as well as different *stages*.

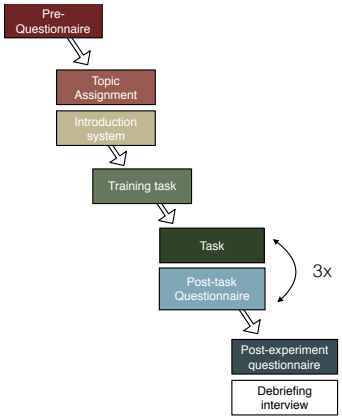

Figure 2: Simplified protocol for the SearchAssist study

Information seeking stages

As a framework we used *temporally-based information seeking models* – as defined by Beheshti et al. [1]. In particular, we were interested in *stages* occurring in information seeking, and utilized previous literature related to tasks involving learning and construction. Kuhlthau [20] has described a succession of stages, during which the feelings, thoughts and actions evolve: *Initiation*, *Topic Selection*, *Exploration*, *Focus Formulation*, *Collection* and *Presentation*. We chose this model, as it was highly cited and one of the most empirically tested information seeking models [1]. The model has been further refined and tested in an information retrieval context by Vakkari [28]. He grouped the stages into three wider stages: *Pre-focus* (Initiation, Topic selection, Exploration), *Focus formulation* (Focus formulation) and *Post-focus* (Collection, Presentation). For the design of our study, we chose to use Vakkari's model, since the grouped stages were more feasible to incorporate in our study than the fine-grained stages defined by Kuhlthau.

Search user interfaces

In our paper, our interest was in the *utility* of potential SUI features. Hence, we needed a way for describing the search user interface and for distinguishing the different types of features. As we did previously in Huurdeman and Kamps [13], we made use of a taxonomy proposed by Wilson [31]. This taxonomy distinguishes *input features* (helping users to express needs), *control features* (allowing users to restrict or modify input), *informational features* (providing results or information about them) and *personalizable features* (which are tailored to a user's experience). We chose this taxonomy because it was focused on Search User Interfaces. Its terminology could help us in framing the study, designing the user interface and in discussing the study's outcomes.

Study setup

The study setup was described using common terminology in previous literature (such as [17, 25]), and via terminology from Borlund [3]. We intended to describe as much of the study's setup as possible within the given 10-page space. This included information on the task design and participants, the full task descriptions, the data and the interface. Finally, we briefly described a validation of topic differences and invoked stages. The latter was important to validate the new multistage simulated task approach used in the paper (see Section 4 for more details). An important element was defining the study's *protocol*, the importance of which also has been underlined by Borlund. Figure 2 depicts a simplified example of the study's protocol.

4 METHODOLOGY

Next, we outline the methodology used in the CHIIR 2016 paper [14], and the decisions made in the process of preparing it.

Methodology, methods and research techniques

For describing aspects related to our methodology here, we use part of the division made by Pickard [25]: *research methodology* (theoretical perspective of the research), *research method* (strategy) and *data collection instruments* (research techniques).

In terms of *research methodology*, the paper used mixed methodology, thus combining qualitative and quantitative methodologies. We decided to use mixed methods to be able to capture the inherently multi-layered ('macro-level') aspects of information seeking and the micro-level behavioral patterns.

With respect to *research method*, we used experimental research, via a lab-based user study. We took this approach (as opposed to e.g. a naturalistic setting) to be able to combine a wide variety of data collection instruments.

The *data collection instruments* directly used in our analysis, and documented in the paper, were chosen based on our research questions, and on examples from previous literature. These were the following:

- Questionnaires (pre-experiment, post-task, post-experiment)
- Interview (post-experiment)
- Transaction logging (clicks, mouse moves, entered text)
- Eye tracking (fixations, saccades)

Furthermore, we made use of other data collection instruments, which were not directly used in our analysis.

- Observations (the investigator observed the participants' behavior and could view their screen contents on a tablet from a distance)
- Screen recordings (a time-stamped screenshot was made every 250 milliseconds)

The rationale underlying the use of the latter instruments is that they were used as a reference during the analysis process (observation notes), and as a backup in case transaction logging instruments would fail (screen recordings).

Further specifics regarding the configuration of data collection instruments, and re-use of previous approaches can be found in Section 5.

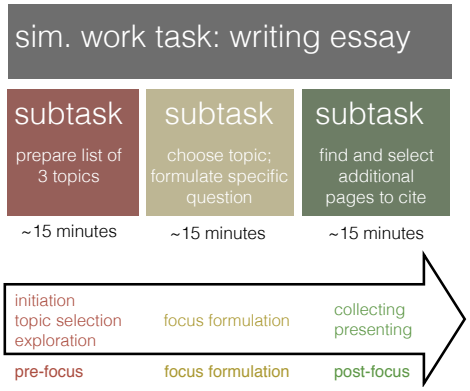

Figure 3: Study design

Research Design

At the moment of writing the paper, Kuhlthau's and Vakkari's models had been studied in longitudinal settings (e.g. [18, 19, 29, 30]), for instance during students' processes of writing a term paper. This means the process was monitored at multiple moments along a broader timeframe (for instance using surveys or by monitoring a search session). At the moment of writing, no longitudinal studies of search user interfaces or their specific features using the model of Kuhlthau or Vakkari existed. Some studies had investigated temporal use of SUI features, but used temporal segmentations of singular search sessions to deduct phases in a session (for instance [7, 13, 24]).

One the one hand, longitudinal settings may not have full possibilities for close monitoring and controlling experimental settings, while on the other hand viewing information seeking stages as temporal search segments might not include the same level of learning as longitudinal studies. Therefore, our aim before the study was to find a middle point - combining aspects of both approaches. As an instantiation of this aim, we set out to study multiple subtasks, representing different stages, within a single simulated work task (see Figure 3).

In our user study, we used a (cognitively complex) multistage simulated work task – the commonly used essay-writing task – which would also be sufficiently familiar to the undergraduate students participating in the study. We studied interaction patterns with interface and content during different search stages, represented by the subtasks. The independent variable was task stage, the dependent variables were active, passive and perceived utility of search user interface features. More specifically, we looked at *active behaviour*, "the behaviour which can be directly and indirectly determined from logged interaction", *passive behaviour*, "behaviour not typically caught in interaction logs, such as eye fixations and mouse movements," and *perceived usefulness*, "the subjective opinions of users about the usefulness of features" [14]. These variables were discussed among the paper authors in advance, and were meant to extend the small-scale data analysis in the paper's predecessor [13].

This approach can be seen as more realistic than the singular search approach, but potentially allow for more experimental control than a longitudinal setting. A challenging aspect, however, was

to formulate simulated work task situations which were representative of Vakkari's stages, and also providing possibilities for learning about a topic. This formulation took place during several months preceding the actual study, and involved the paper authors as well as further information seeking experts. We discuss the re-use of previous materials within the research design in Section 5.

Borlund [3] underlines the importance of counterbalancing tasks. In this case, we focused on work tasks involving learning. Therefore, the three tasks in the study had to be performed in sequence; the stage order could not be counterbalanced without losing cumulative learning and understanding gathered in each subsequent stage. We reckoned that this was a worthwhile tradeoff, since the tasks involved learning, and thus were dependent on each other (e.g., a participant needed to explore topics before making a reasoned decision about which topic to choose).

Task and Stage Validation

We also validated the multistage approach, of both task and stage within the process. We examined the validity of our task descriptions in terms of invoking correct stages.

In post-stage questionnaires users selected the activities they had conducted from a randomized list[1] derived from Kuhlthau's model [20]. From the results of this validation, we concluded that even though changes between stages are sometimes gradual, our experiment correctly invoked the main activities in each stage (for instance, 'exploring' in the first subtask, 'focusing' in the second/third subtask, and 'collecting' in the third subtask). The fact that the first task was seen as explorative was also reflected in the type of information sought, reported in the questionnaire as evolving from 'general' (in the questionnaire after stage 1), to 'specific' (after stage 2 and 3).

Further parts of the stage validation matched the results of the validation, but could not be included in the CHIIR paper, due to a lack of space. However, they were included in an extended version in Huurdeman [12]. This included an assessment of the feelings of participants during the experiment, to monitor the concordance with the stages described in Vakkari [28]. To this end, we used a word list from previous user studies by Kuhlthau [20], Todd [27]. Participants had to choose from a list of ten words (in random order) which could represent their state of mind near the end of each task phase[2]. Some fluctuations in reported feelings could be detected, showing some evidence of Kuhlthau's findings on gradually reduced uncertainty and rising optimism.

Participant recruitment

We aimed at recruiting undergraduate students in the Computer Science department of the University of Nottingham (UK campus), since this is where the lab study took place, and since we could customize the tasks to be relevant to this particular audience. We used a multifaceted approach to cast a wide net: we announced the study via posters in the School of Computer Science, via the institution's Facebook page, via an email list, and via the website callforparticipants.com. Since direct payment was not possible, participants received a 10 GBP Amazon voucher. This amount was based on the available budget, and previous studies in the department.[3] To add additional incentive for our tasks involving learning, we awarded 25 GBP for the best task outcome.

5 RE-USE OF PREVIOUS RESOURCES

This section describes the resources from previous literature which we re-used in our study, as well as resources related to system, data and user interface.

Research design

As mentioned in Section 4, no multistage simulated work task design existed at the time which used Vakkari's stages, but when designing the tasks, we did incorporate elements from previous work, albeit often in adapted form. First of all, Kuhlthau's book [20] and Vakkari's work (e.g. [28]) were an inspiration. Further resources had been found in our previous literature survey on information seeking stages [13], and additional examples were sought during the preparation of this paper. This was both via online literature search systems and via the RepAST repository of assigned search tasks[4]. Another source of information for the subtasks were existing research process and information literacy models. This included Kumar [22, p.51-53] and various online resources[5]. These models include the idea of formulating broad topics, selecting a specific topic and questions related to the topics.

For the textual contents of the tasks, we used elements of work tasks from Kules and Shneiderman [21], Liu and Belkin [23]. For instance, to inspire participants we indicated that ideas for topics should "cover many aspects of the topic" and that "unusual or provocative ideas are good", part of the task description in Kules and Shneiderman [21]. Furthermore, we took inspiration from Liu and Belkin [23], which used an "approach using different subtasks accomplished in different search sessions at different times." In our case, however, subtasks were performed within different search sessions in a single user study – with small breaks in between, in which participants switched from focusing on the screen to filling out a paper-based questionnaire.

Questionnaire design and scales

For the design of our pre-experiment, post-task and post-experiment questionnaires, we combined different sources. A number of questions were based on those described for the related user studies described in Diriye [8]. Some questions were used directly (e.g. referring to task and topic understanding and interest, and to the used interface), and other questions were added or reformulated based on our research questions and particular multi-stage setup. Furthermore, we directly used Kuhlthau [20], Todd [27] for various validation questions within the process, as described in the previous section 4. In terms of scales and possible answers, we used

[1] Specifically: *exploring*, *focusing*, *formulating*, *collecting*, *gathering*, *becoming informed*, *choosing*, and *getting an overview*

[2] In particular: *confident*, *disappointed*, *relieved*, *frustrated*, *sure*, *confused*, *optimistic*, *uncertain*, *satisfied*, *doubtful*.

[3] As the findings of a replication study by Wilson [32] indicate, remuneration might have an effect on motivation of participants. In the CHIIR study, we tried to optimize motivation by ensuring that the participants selected a topic of their liking (using elicitation questions in a pre-questionnaire), and by asking if participants wanted to be eligible for a prize for the best topic.

[4] https://ils.unc.edu/searchtasks/

[5] e.g., http://ischoolapps.sjsu.edu/static/courses/250.loertscher/modelstrip.html

the approaches in the previously mentioned literature as a basis, including open questions and 7-point Likert Scales – for the latter setup, we used guidance from Pickard [25].

Difficulties in this phase were related to finding, adapting and formulating questions suitable for the multistage search setting in our project, as well as the used models from Kuhlthau and Vakkari. Moreover, many previous papers did not document their questionnaire contents. An alternative would be to use standard questionnaires instead, though the inspected examples were deemed not specific enough at the time of our study, and the duration to fill out some of the longer standardized questionnaires would limit the possibilities within the originally planned 60 minute timeframe of the study.

Due to space constraints and a lack of time to provide further documentation, the questionnaires were not included in the CHIIR paper itself, but were only described – thus, they could not directly be re-used at the time[6].

Experimental System

The study used custom-built components for the SearchAssist system (depicted in Figure 4). Components were created using Javascript and PHP, and the libraries JQuery and JQuery-UI were utilized. For logging system events, we used a custom MySQL component and Log4Javascript. This way, user actions were both logged to a database as well as stored in raw text files (for redundancy). Furthermore, we exported browser history using the "Export History" Chrome browser extension. For mouse logging, we used a Javascript based method available online[7].

We decided to create the system from scratch, although we reused previous frameworks and components when possible. Existing systems for IIR experiments, such as PyIRE[8], were considered, but not used due to several reasons. Since we made use of paper-based questionnaires, and since we used one interface for the three stages, we did not need a system for handling the experimental flow. Also, we had limited time to setup the needed system, adapting it to our needs (e.g. for using eye-tracking and including the SUI features we wished to evaluate), and to obtain the data that would be necessary to populate a search system representing a general-purpose web search engine, discussed next.

Data

For the data underlying the search system, different options were considered. For instance, creating a search engine using the ClueWeb[9] or Amazon / LibraryThing dataset[10]. However, in the end we decided to use the Bing Web Search API for the search results. This was chosen because it would a) ease the creation of a suitable search interface, b) would provide realistic and recent search results to participants, and c) because participants could open the full web resources listed in the search results. To avoid different users seeing

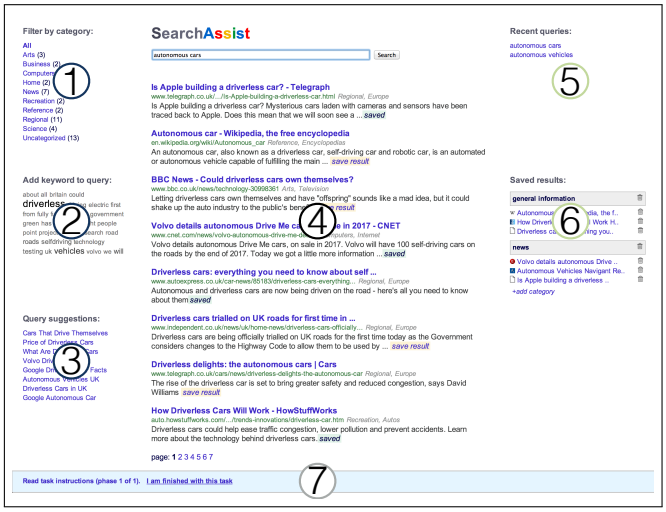

Figure 4: Screenshot SearchAssist. *Left column* (1, 2, 3): control features. *Middle* (4): input and informational features. *Right Column* (5, 6): personalisable features. (7): task bar

different results within the timeframe of the study (approximately two weeks), we cached search results for each query locally, meaning a second user would see the same results if the same query was entered[11]. For spelling corrections to queries we utilized the Bing Spelling Suggestions API. For documentation on the use of these APIs, we used Microsoft's "Bing Search API Quick Start and Code Samples" document.

Interface

The inspiration for the design of this search interface was based on the then-current Google search interface, including the basic layout and color scheme (see Figure 4). This way, we intended to offer a familiar environment to users. All functionality was tested and adapted based on a small pilot study with two participants. For specific features in the interface, the following components were used:

(1) Category Filters: a clickable list, generated by matching hostnames of results with a converted list of URLs and top level category names downloaded from the Open Directory Project (DMOZ)[12].
(2) Word cloud: a basic word cloud created via *jquery.tagcloud.js*. Words could be added to the current query.
(3) Query suggestions: a clickable list generated from the Bing Query Suggestions API.
(4) Search Results: originating from the Bing Web Search API, combined with DMOZ category information.
(5) Recent queries: use of a local MySQL database to display a clickable list with the last 15 queries
(6) Saved results feature: custom built and tested with colleagues (interaction design experts) at the department. Includes possibilities for adding categories and drag 'n drop reordering of saved items / categories, as well as deletion of items / categories.

[6] At this point, though, they can be accessed via https://github.com/timelessfuture/searchassist/tree/master/chiir-study-materials
[7] Available at: https://stackoverflow.com/questions/7790725/javascript-track-mouse-position/34348306
[8] https://pyiire.readthedocs.io/en/latest/
[9] http://lemurproject.org/clueweb12/
[10] As in e.g. the INEX/CLEF Interactive Social Book Search Track [9]: http://inex.mmci.uni-saarland.de/data/documentcollection.html
[11] These cached results were later securely stored in conjunction with the experimental data, for future reference and analysis.
[12] Now offline, archived version at: http://web.archive.org/web/20141102025545/http://www.dmoz.org/docs/en/rdf.html

(7) Task bar: clickable links to the task instructions and response form (in an editable Google Document) and an option to end the current task.

A link to the source code of the used experimental system[13] was included in the final CHIIR 2016 paper. This consisted of the search interface, task configurations and all used back-end components (including custom usage logging), along with brief documentation. Although tailored to our experiment, the different elements of this system could be re-used for future studies, keeping in mind the crucial aspect of maintenance (further discussed in Section 7)[14].

Eye tracking and eye tracking analysis

With respect to eye tracking, we made use of the approach employed by Jiang et al. [16]. This pragmatic approach involved showing up to 8 results at a time in the search interface, instead of the more regular 10. This allowed for easier analysis of fixations on certain parts of the search user interface (since there was no scrolling within the search screen itself). The use of a relatively large screen with sufficient screen resolution allowed for the displayal of all features.

For our paper, we looked at common eye tracking metrics (see e.g. Poole and Ball [26]), and chose to analyze fixation counts and fixation durations. To distinguish fixations, we used Buscher et al. [4]'s strategy, which defined a fixation as sequences of eye tracking measures within a 25 pixel radius, within a timeframe of at least 80ms. Since both fixation count and duration measures had similar results, we focused on reporting only fixation counts in the paper, due to stringent space limits for the CHIIR paper[15].

For transparency and flexibility, we decided to use an open-source Python framework to perform the eye tracking and do the subsequent analysis. For the eye tracking, we utilized the PyGaze framework [6], as well as the PyTribe toolbox - a wrapper for the used EyeTribe eye tracker[16]. Using the PyGaze software, it was then possible to generate heatmaps and other eyetracking visualizations, but also to analyze fixations using our own metrics.

6 POTENTIAL FOR RE-USE OF OUR APPROACH

The simulated work task approach to studying information seeking stages, as applied in our paper, has re-use potential. This is reflected by the fact that the approach has been re-used in two papers so far [10, 11].

First, Hoeber et al. [11] explicitly state that they drew inspiration from Huurdeman et al. [14] for the organization of their study, in which they evaluated an interactive search interface entitled "Lensing Wikipedia". The paper utilizes a similar research design as our paper, also using Vakkari [28] stages and Wilson [31]'s taxonomy of interface features. The essay-writing task and its descriptions are adapted to the domain described in the paper (a history course), but are otherwise similar to those in our paper. Instead of three research ideas, users selected three persons in the pre-focus stage, followed by the selection of one person and further investigating this person (focus formulation) and collecting materials (post-focus). Their research questions had different focal points, looking at *feature use* (active utility in our paper), *knowledge gain* (captured in our study, but not used in our paper), *perceived usefulness* (included in our study) and *overall perceived usefulness and satisfaction* (captured in our study, but not reported due to space limitations). Hoeber et al. [11]'s research outcomes in terms of feature use across stages confirm our findings.

Second, Gaikwad and Hoeber [10] used Vakkari [28]'s model as "a design guide, and as a mechanism for controlling the laboratory-based evaluation." This paper uses a multistage task design, but focuses on interactive image retrieval. Therefore, participants explored (pre-focus), selected (focus formulation) and organized images (post-focus), and this is reflected in the task descriptions, which focus on holiday plans, food blogging and self-selected tasks.

7 DISCUSSION AND CONCLUSION

This experience paper has reflected on the various aspects related to creating a simulated work task approach to studying information seeking stages. This approach was first applied in the context of Huurdeman, Wilson, and Kamps [14]. Most terminology from this paper originated from previous literature, and was adapted for use in our paper. Our methodology extended previous approaches in combining a variety of data collection instruments, as well as in taking a new approach to designing multistage studies. We also discussed the re-use of previous resources, including encountered difficulties. Finally, the subsequent use of the multistage approach [10, 11] has shown that re-use of research design and tasks is a feasible prospect.

With respect to the barriers to the re-use of materials, we can observe the *issue of lacking space* to document all aspects of our study – for instance further documentation on decisions within the process. Moreover, there is the typical *lack of time* within the research process and publication cycle, which meant that we could release the source code for the used tools in time for publication, but not the analysis scripts or other resources. A *restrictive consent form* also meant that no actual data from the study (e.g. interaction data) could be released, even anonymously. On a broader scale, we encountered the tension between *flexibility* in terms of research questions, and the possibility to re-use standardized systems and approaches, leading us to create a custom system. There is also the issue of *maintenance*: just four years after our study, the components of the system have changed (e.g. Bing API configurations), as well as the issue of *persistence*: various URLs of used resources are now only available in the Internet Archive[17].

We would fully support the creation of more standardized approaches to documentation and more centralized places to deposit the wide variety of resources related to a user study, as discussed in Bogers et al. [2]. In this light, it is also very positive to observe that conferences such as CHIIR now allow additional space for references and appendices, making it possible to extend publications with pivotal documentation about the process.

[13] Available from: https://github.com/timelessfuture/searchassist

[14] At this point, in 2019, re-use would imply system adaptations to reflect for instance changed search API details and updated hostname lists for the category filters.

[15] For a planned journal extension of the paper, we intend to include both fixation metrics.

[16] PyGaze is available from: https://github.com/esdalmaijer/PyGaze, and PyTribe from: https://github.com/esdalmaijer/PyTribe

[17] We took a proactive approach, however, and archived for instance all webpages opened by participants at the time, using *wget*

REFERENCES

[1] Jamshid Beheshti, Charles Cole, Dhary Abuhimed, and Isabelle Lamoureux. 2014. Tracking middle school students' information behavior via Kuhlthau's ISP Model: Temporality. *J Am Soc Inf Sci Tec* (2014). https://doi.org/10.1002/asi.23230

[2] Toine Bogers, Maria Gade, Mark Michael Hall, Luanne Freund, Marijn Koolen, Vivien Petras, and Mette Skov. 2018. Report on the Workshop on Barriers to Interactive IR Resources Re-use (BIIRRR 2018). *ACM SIGIR Forum* 52, 1 (2018), 10. https://doi.org/10.1145/3274784.3274795

[3] Pia Borlund. 2003. The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. *Inf Res* 8, 3 (2003). http://www.informationr.net/ir/8-3/paper152.html

[4] Georg Buscher, Andreas Dengel, and Ludger van Elst. 2008. Eye Movements As Implicit Relevance Feedback. In *CHI '08 Extended Abstracts on Human Factors in Computing Systems (CHI EA '08)*. ACM, New York, NY, USA, 2991–2996. https://doi.org/10.1145/1358628.1358796

[5] K. Byström and Kalervo Järvelin. 1995. Task Complexity Affects Information Seeking and Use. *Inform Process Manag* 31, 2 (1995), 191–213. https://doi.org/10.1016/0306-4573(95)80035-R

[6] Edwin S. Dalmaijer, Sebastiaan Mathôt, and Stefan Van der Stigchel. 2013. PyGaze: An open-source, cross-platform toolbox for minimal-effort programming of eyetracking experiments. *Behav Res Meth* 46, 4 (2013), 913–921. https://doi.org/10.3758/s13428-013-0422-2

[7] Abdigani Diriye, Ann Blandford, and Anastasios Tombros. 2010. When is System Support Effective?. In *Proceedings of the Third Symposium on Information Interaction in Context (IIiX '10)*. ACM, 55–64. https://doi.org/10.1145/1840784.1840794

[8] A. M. Diriye. 2012. *Search interfaces for known-item and exploratory search tasks*. Doctoral. UCL (University College London). http://discovery.ucl.ac.uk/1343928/

[9] Maria Gäde, Mark Hall, Hugo Huurdeman, Jaap Kamps, Marijn Koolen, Mette Skov, Elaine Toms, and David Walsh. 2016. Overview of the INEX 2016 Interactive Social Book Search Track. In *Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum (CEUR Workshop Proceedings)*, Vol. 1609. http://ceur-ws.org/Vol-1609/16091024.pdf

[10] Manali Gaikwad and Orland Hoeber. 2019. An Interactive Image Retrieval Approach to Searching for Images on Social Media. In *Proceedings of the 2019 ACM Conference on Human Information Interaction and Retrieval (CHIIR '19)*. https://doi.org/10.1145/3295750.3298930

[11] Orland Hoeber, Anoop Sarkar, Andrei Vacariu, Max Whitney, Manali Gaikwad, and Gursimran Kaur. 2017. Evaluating the Value of Lensing Wikipedia During the Information Seeking Process. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval - CHIIR '17*. ACM Press, Oslo, Norway, 77–86. https://doi.org/10.1145/3020165.3020178

[12] Hugo C. Huurdeman. 2018. *Supporting the complex dynamics of the information seeking process*. Ph.D. Dissertation. University of Amsterdam. http://hdl.handle.net/11245.1/1e3bf31a-0833-4ead-a00c-4cb1399d0216

[13] Hugo C. Huurdeman and Jaap Kamps. 2014. From Multistage Information-seeking Models to Multistage Search Systems. In *Proceedings of the 5th Information Interaction in Context Symposium (IIiX '14)*. ACM, New York, NY, USA, 145–154. https://doi.org/10.1145/2637002.2637020

[14] Hugo C. Huurdeman, Max L. Wilson, and Jaap Kamps. 2016. Active and Passive Utility of Search Interface Features in Different Information Seeking Task Stages. In *Proceedings of the 2016 ACM Conference on Human Information Interaction and Retrieval (CHIIR '16)*. ACM, New York, NY, USA, 3–12. https://doi.org/10.1145/2854946.2854957

[15] Peter Ingwersen and Kalervo Järvelin. 2005. *The Turn - Integration of Information Seeking and Retrieval in Context*. Springer.

[16] Jiepu Jiang, Daqing He, and James Allan. 2014. Searching, Browsing, and Clicking in a Search Session: Changes in User Behavior by Task and over Time. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '14)*. ACM, New York, NY, USA, 607–616. https://doi.org/10.1145/2600428.2609633

[17] Diane Kelly. 2009. Methods for Evaluating Interactive Information Retrieval Systems with Users. *Foundations and Trends in Information Retrieval* 3, 1âĂŤ2 (2009), 1–224. https://doi.org/10.1561/1500000012

[18] Carol Collier Kuhlthau. 1988. Longitudinal case studies of the information search process of users in libraries. *Libr Inform Sci Res* 10 (1988), 257–304.

[19] Carol Collier Kuhlthau. 1988. Perceptions of the information search process in libraries: a study of changes from high school through college. *Inform Process Manag* 24 (1988), 419–427. https://doi.org/10.1016/0306-4573(88)90045-3

[20] Carol Collier Kuhlthau. 2004. *Seeking meaning: a process approach to library and information services*. Libraries Unlimited.

[21] Bill Kules and Ben Shneiderman. 2008. Users can change their web search tactics: Design guidelines for categorized overviews. *Inform Process Manag* 44, 2 (2008), 463–484. https://doi.org/10.1016/j.ipm.2007.07.014

[22] Ranjit Kumar (Ed.). 2010. *Research Methodology*. SAGE.

[23] Jingjing Liu and Nicholas J. Belkin. 2015. Personalizing information retrieval for multi-session tasks. *J Am Soc Inf Sci Tec* 66, 1 (Jan. 2015), 58–81. https://doi.org/10.1002/asi.23160

[24] Xi Niu and Diane Kelly. 2014. The use of query suggestions during information search. *Inform Process Manag* 50 (2014), 218–234. https://doi.org/10.1016/j.ipm.2013.09.002

[25] Alison Pickard. 2007. *Research methods in information*. Facet Publishing.

[26] Alex Poole and Linden J. Ball. 2005. Eye Tracking in Human-Computer Interaction and Usability Research: Current Status and Future. In *Prospects, Chapter in C. Ghaoui (Ed.): Encyclopedia of Human-Computer Interaction. Pennsylvania: Idea Group, Inc.*

[27] Ross J. Todd. 2006. From information to knowledge: charting and measuring changes in students' knowledge of a curriculum topic. *Inf Res* 11, 4 (2006). http://www.informationr.net/ir/11-4/paper264.html

[28] Pertti Vakkari. 2001. A theory of the task-based information retrieval process: a summary and generalisation of a longitudinal study. *J Doc* 57, 1 (Feb. 2001), 44–60. https://doi.org/10.1108/EUM0000000007075

[29] Pertti Vakkari and Nanna Hakala. 2000. Changes in relevance criteria and problem stages in task performance. *J Doc* 56 (2000), 540–562. https://doi.org/10.1108/EUM0000000007127

[30] Pertti Vakkari, Mikko Pennanen, and Sami Serola. 2003. Changes of search terms and tactics while writing a research proposal: A longitudinal case study. *Inform Process Manag* 39 (2003), 445–463. https://doi.org/10.1016/S0306-4573(02)00031-6

[31] Max L. Wilson. 2011. Search User Interface Design. *Synthesis Lectures on Information Concepts, Retrieval, and Services* 3, 3 (Nov. 2011), 1–143. https://doi.org/10.2200/S00371ED1V01Y201111ICR020

[32] Max L. Wilson. 2013. Teaching HCI Methods: Replicating a Study of Collaborative Search. In *Proceedings of the CHI2013 Workshop on the Replication of HCI Research (RepliCHI 2013) (CEUR Workshop Proceedings)*, Vol. 976. 39–43. http://ceur-ws.org/Vol-976/tpaper5.pdf

[33] Tom D. Wilson. 1999. Models in information behaviour research. *J Doc* 55 (1999), 249–270. https://doi.org/10.1108/EUM0000000007145