# Adaptive Search Support for Information Seeking Stages

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

## ABSTRACT

We use the Web for work, leisure, and research, assisted by various search systems in the task of satisfying our information needs. We utilize these systems to perform our daily tasks, ranging from simple lookup tasks to complex, exploratory and analytical ventures. The more complex tasks may involve multiple information seeking stages, with evolving inherent needs for each stage. Most search systems, however, only support these complex tasks in an elementary manner, and offer a 'one-size-fits-all' interface optimized for shallow lookup search.

In addition to the wealth of information available on the live Web, historical Web content is currently available in Web archives, containing snapshots of the Web that once was. These Web archives can enable new opportunities for analytical tasks, serving as data sources for researchers in various fields. At the moment, however, few archives offer full-text search, and the search systems that are available fall short of the rich functionality needed for analytical tasks.

This PhD research proposal takes Kuhlthau's ISP model as its framework, and addresses search support for different 'stages' of complex search tasks. It discusses the theoretical implications of multistage information seeking models for the design of search systems. The proposal examines the effects of information seeking stages on the flow of interaction with actual search systems. The understanding on both the theoretical and practical level are used to design and evaluate multistage search systems, firstly in a general Web search setting, and secondly in a Web archive search setting. Finally, this leads to design recommendations for supporting different stages of complex tasks in search systems.

This proposal consists of five parts: first, we introduce the topic and research problem. Then, we state the research questions. Subsequently, background literature is briefly discussed, followed by the employed methodology. Next, current progress is discussed, before summarizing this proposal in the conclusion.


*Digital Libraries 2014 Doctoral Consortium* September 8, 2014, London, UK.

## Keywords

Information seeking, Search process, Search stages, User Interfaces, Web archiving

## 1. INTRODUCTION AND MOTIVATION

The wealth of digital information available in our time has become indispensable for a rich variety of tasks. We use data on the Web for work, leisure, and research, aided by various search systems, allowing us to find small needles in giant haystacks. However, despite recent advances in personalization and contextualization, various types of tasks, ranging from simple lookup tasks to complex, exploratory and analytical ventures, are mainly supported in elementary, 'one-size-fits-all' search interfaces. This unified approach might limit users in performing their more complex tasks.

Web archives, keepers of our *future* cultural heritage, have gathered petabytes of valuable Web data, which characterize our times for future generations. Access to these archives, however, is surprisingly limited: online Web archives usually provide a URL-based Wayback Machine interface, sometimes extended with rudimentary search options. As a result of limited access, Web archives are not widely used yet. For the more complex types of tasks which could be performed using Web archives, such as research tasks, there is a need to move beyond URL-based and simple search access, towards providing support for analytical search and research tasks. The Dutch WebART project[1] takes on this challenge, and aims to improve research access to Web archives in both conceptual and concrete ways. WebART is a multidisciplinary collaboration between the University of Amsterdam, the Centrum Wiskunde & Informatica (CWI) and the National Library of the Netherlands (KB). In this proposal, we look at the support for complex search tasks in both the context of general Web search, and in the context of analytical Web archive search.

**Research problem and framing** The main research problem, posed in the context of *complex work tasks* is to *analyze and evaluate the influence of information seeking stages on the interactive information retrieval process, and how to provide customized search support for these stages.*

*Complex work tasks* can be defined as work tasks which require "understanding, sense-making, and problem formulation" [5]. These tasks go beyond simple lookup tasks, and involve learning and construction. As various information seeking literature has evidenced [6], users often experience different *cognitive stages* in their complex search endeavors.

[1]Web Archive Retrieval Tools (`www.webarchiving.nl`)

In this work, we take Kuhlthau's and Vakkari's stages as our framework [18, 29]. Consider, for example, an often-studied "information-intensive, constrained-based" [28] work task: the preparation of a research paper by a student. Student may experience initial *prefocus* exploratory stages, in which a topic is selected and information is explored. At some point this is followed by a *focus formulation* stage in which a focused perspective is formulated, before finally moving to *postfocus* stages of pinpointed data collection and synthesis. In these stages, types of information sought, relevance criteria and search tactics evolve. In effect, optimal *search support* for users' search activities may vary as well. In this research proposal, we explore the possibility to provide tailored search support for these stages by offering differentiated sets of functionalities in search systems. Here, the focus is not on automatic detection of search stages, but on defining and exploring ways to potentially offer customized support for different stages.

In this PhD research proposal, we take the perspective of Interactive Information Retrieval, which "focuses on users' behaviors and experiences - including physical, cognitive and affective - and the interactions that occur between users and systems, and users and information" [17].

## 2. RESEARCH QUESTIONS

The main research questions of this proposal are aimed at uncovering the conceptual implications of information seeking stages (RQ1), at evaluating the actual effect of stages on interaction with search interfaces and information (RQ2), and at providing customized search support for stages in Web search and Web archive research (RQ3 and RQ4). The final research question looks at ways to contextualize Web archive search at the content level (RQ5).

**RQ1** What are the conceptual implications of multistage information seeking models for the design of search systems?

This research question takes a broad perspective and looks at ways to bridge the conceptual gap between *macro-level* information seeking models and *micro-level* search systems, by means of a theoretical analysis. On the one hand, we introduce relevant theory from the area of information seeking behavior [32], including various information seeking models and information literacy models. On the other hand, search user interface paradigms and concrete interfaces in the context of cognitively complex work tasks are researched, to gain insights into the support of current search systems for complex tasks.

**RQ2** How do information seeking stages affect the flow of interaction with search systems in a Web search setting?

After taking a theoretical perspective in the previous research question, we look in the second research question at the influence of information seeking stages on the interaction patterns with search systems. By means of a user study we gain insights into concrete interaction patterns at the *interface* and *content* level, in the context of general Web search. The interaction patterns can provide evidence for the importance of certain types of interface features and content at specific moments in the complex task, and be used to derive recommendations for the design of adaptive and multistage systems.

**RQ3** How can we effectively provide search support for information seeking stages at the interface and content level in a Web search setting?

This question looks at ways provide differentiated support for the previously defined information seeking stages in a Web search setting. The findings of the previous research questions are tested by designing an adaptive search system, which provides tailored search support to three main stages of complex tasks. We will experiment with supporting users in their *prefocus*, *focus* and *postfocus* stages. In these three stages, the system could offer differentiated interface features (adaptive features and search tools), and differentiated ranking (adaptive filtering and ranking of results). Further experiments may be done with respect to *interventions*, as systems could aid users at specific moments in the search process.

**RQ4** How can we effectively provide search support for research stages in the context of Web archives?

While the previous research questions looked at information seeking stages in the context of Web data, this question focuses on archived Web data. Web archives contain the Web of the past, and various aspects of Web archives (e.g. duplicates) influence search support. First, limitations of current Web archives in the context of Web research are discussed, followed by a description and evaluation of a search system for the Dutch Web archive.

**RQ5** How can we contextualize search systems for Web archives at the content level?

The final research question is also focused on archived Web data, and discusses ways to ameliorate the incompleteness of Web archives. Part of this research explores the link structure and anchor text of Web archives to generate representations of unarchived contents. We hypothesize that this information can be used to contextualize search systems by showing *un*archived content.

## 3. BACKGROUND

To conceptually support complex search tasks in searchable Web archive systems, we can make use of a rich set of background literature on information seeking [6, 32], which is "human information behavior dealing with searching or seeking information by means of information sources and (interactive) information systems" [16]. Various models indicate that users experience stages in information seeking for complex tasks, and represent these stages in different ways: sequentially [18], [29], as activities that can be recombined in different ways [9], or in nonlinear ways [10]. While the mentioned models of search and research stages generally focus on the *macro* level of information seeking, i.e. higher-level aspects of search, search user interfaces deal with concrete features on the *micro* level [32]. For different reasons, including cognitive load issues [12], most search user interfaces take a basic approach and include only the most essential features. It is possible, however, to employ other approaches in the support of complex tasks: interfaces could offer a wider array of features, e.g. supporting exploratory search [20]

Table 1: Search stages (adapted from Vakkari [29])

| Stage | *Description* |
|---|---|
| *Prefocus* | Topic selection, Exploration |
| *Focus* | Formulating a focused perpective |
| *Postfocus* | Collecting focused information, prepare findings |

and sensemaking [12], but also offer varying functionality to users, depending on their search stage.

Many information seeking models are general models of the information seeking process, potentially applicable to different settings featuring complex search tasks. A number of authors argue that the research process also includes various "stages" [6]. This might imply different information seeking strategies in evolving research phases (as documented by e.g. [3, 7, 21]). However, customized support for stages in the search and research process is currently lacking, suggesting an urgent need for more research and development in this area.

Web archives are a relatively new data source for researchers from various disciplines ranging from the Humanities to Computer Science. Access to Web archives is restriced, both in a practical sense, as many Web archives are only available onsite, and in a technical sense, as most Web archives can only be accessed via the URL-based Wayback Machine. As there is a limitation in search-based access, this limits research which can be performed using Web archives. Various new media scholars have discussed the unique properties of Web archives and their influence on (potentially) performed research using these archives [4, 23, 26], highlighting the opportunities, but also current limitations. To overcome limitations, we may explore new search access methods for Web archives, which can support the flow of search in a research context. This can include user-definable search strategies [8], modifiable as "building blocks" in visual user interfaces. In addition, (meta)data of Web archives could be improved, to uncover information hidden in Web archives [22], and to further contextualize Web archive search systems [19].

# 4. METHODOLOGY AND FRAMING

To answer the mentioned research questions, a mixed approach of qualitative and quantitative methodologies is employed. The used research methods and data collection techniques include:

- an extensive literature review on past research in the area of information seeking behavior, and theory from the subdomain of information search (RQ1).
- user studies employing cognitively complex simulated work tasks, focusing on the use of search system features in different search stages (involving eye tracking, logging and questionnaires) (RQ2, RQ3).
- a user-centric *living lab* research methodology in which needs for Web archive search are assessed, and search features are developed, prototyped and evaluated in close collaboration with researchers (RQ4).
- a structured known-item search evaluation for generated unarchived page and site representations (RQ5).

Table 2: Main framing per research question

| | |
|---|---|
| *context* | academic research (RQ2-5) |
| *actors* | undergraduate students (RQ2-3) |
| | postgraduate researchers (RQ4-5) |
| *inf. objects* | webpages (RQ2-3) |
| | archived webpages (RQ4-5) |
| *interface* | search interface (RQ2-5) |
| *system* | full-text search system (RQ2-5) |

Table 2 shows the framing of the thesis per research question using elements adapted from Ingwersen and Järvelin [16]'s general model of cognitive information seeking and retrieval.

# 5. PROGRESS

## RQ1: Conceptual implications of Search Stages

In previous work [13], we discussed the conceptual implications of *macro* information seeking stages for the design of search systems. We focused on Kuhlthau's and Vakkari's models of the information seeking process, and discussed the impact of search stages on information sought (moving from general to specific), relevance (evolving through the process) and search tactics (with a growth in searchers' ability to precisely express their information needs). While there is an abundance of models describing the information seeking process, there are less examples of search interfaces explicitly supporting stages in the information seeking process. Despite some experimental interfaces support exploratory search and sensemaking, the general tendency has moved towards streamlined, 'one-size-fits-all' search interfaces [16]. We argue that it would be possible to move beyond this paradigm, and the implications from the literature provide handles for creating customized support for different information seeking stages.

## RQ2: Effects of search stages on the flow of interaction

The theoretical perspective in the previous research question has provided strong indications that information seeking stages might influence the interaction patterns with search systems. In various stages of complex tasks, different categories of search interface features may be used differently, both actively (by interacting with the features) and passively (by looking at the features). Also at the content level differences might occur, as information seeking models indicate that there are distinctions between the use of various types of information sources over time (e.g. the use of introductory sources in initial stages, and the use of specific sources in later stages).

In previous research, we took a tentative look at the influence of search stages on the flow of interaction [13]. An analysis of eye tracking and system log data of complex tasks performed via *ezDL*, a relatively rich user interface [1], showed differences in the use of interface and search system features occurring at different stages of a search episode. Here, we used Wilson's framework of interface features, which includes *input*, *control*, *informational* and *personalizable* features. Using input features, users can express what they are looking for (e.g. the query box), control features serve to modify or restrict input (e.g. facets), informational features provide (information about) results, and personalizable fea-

tures tailor the search experience to the user (e.g. features to save or bookmark results) [31]. In our analysis, we saw a decrease in the use of input features, and an increase in the use of personalizable features during the search episode. The results of this initial study point towards the potential usefulness of developing adaptive search systems. Since we need more detailed data, we will perform an additional user study, using simulated tasks in combination with an experimental search interface (which uses the *Bing API*), to perform more in-depth analyses of students' use of interface features and content in different stages of a complex information-intensive task.

Further insights into this topic are gained in the INEX Interactive Social Book Search track [11], where explorations into creating and evaluating multistage search systems are done in the context of book search.

## RQ3: Search support for information seeking stages

The variances in the interaction with search features and information in various stages can be used to create stage-sensitive search systems. We are constructing a multistage search system based on the results of the previous two research questions, and evaluate its features in a user study with simulated work tasks. Our experimental system consists of separate subinterfaces for *prefocus*, *focus formulation*, and *postfocus* search stages. Depending on the search stage, this system may adaptively show SUI features, adjust the shown details of features, and change their prominence, position and size. Further adaptation can be done at the content level: content can be showed and ranked differently at various moments of complex tasks. The exact details of this multistage interface will be finalized after analyzing the data gathered in RQ2. An open question, evaluated in the experiment, is what level of search support is supportive in the search process (as opposed to intrusive or confusing).

Additional planned experimentations are related to the idea of *interventions*. Kuhlthau [18] states that intervention is not in all cases helpful or necessary. However, she defines points in time, or *zones* in the information seeking process, where intervention can be most useful. Based on the information from the previous research questions, we will look at appropriate moments for a search system to "intervene" in the search process, explore in which ways this could be done, and which potential ways would be most useful.

## RQ4: An analytical search system for explorative Web archive search

To gain a better understanding of the requirements of search systems for Web archives, we have made use of a *living lab*, or *co-design* setting in the WebART project, in which developers worked in close collaboration with the actual users of the system, in particular New Media researchers at the Media Studies department of the University of Amsterdam.

The created analytical search interface, *WebARTist*, allows for full-text search in the Dutch KB's 7 Terabyte Web archive. As Figure 1 shows (see Appendix A), the WebARTist search system provides options to *explore*, *analyze*, and *synthesize* search results in the Web archive, supporting Web archive researchers in their research process. Current ways to *explore* results currently include regular search results, word clouds, diagrams and maps. The system allows for filtering the results based on various properties (e.g. temporal ranges and outlinks). Furthermore, via the *analyze* tab it is possible to analyze and edit corpora, using statistical tools. Finally, future versions of the interface will allow users to create visualizations and summaries of performed analyses of datasets in the system via the *synthesize* tab. Prototype systems have been evaluated and extended in various *co-design* events, including the Digital Methods Winter School [14], a two day workshop [30] and a focus group with new media researchers [27]. These events showed the large potential of searchable Web archives in the context of research: as a participant noted, search "supports the shift to studying web archives through queries", and the system "made it possible to build new research questions beyond the web site history approach". Beyond looking at the content, new media researchers also looked at the underlying structure, and aggregated (statistical) views were deemed useful for "revealing underlying structures and patterns within collections". Naturally, also new features were requested by the researchers, including creating rich data selections and features to share (annotated) collections. Also, inherent issues with search in the context of research occurred (e.g. influences of indexing and ranking, and researchers' unfamiliarity with temporal Web archive search). We have documented theoretical and methodological implications of Web archive search, showing that searchable Web archives can lend themselves to additional types of research scenarios, but that they also introduce other, unresolved, challenges [2]. One of these challenges includes the problem of unarchived content: Web archives are inherently incomplete, due to harvesting restrictions (for example on a national level), but also due to technical limitations. Therefore, a very large number of pages cannot be archived, but corpus flaws are not immediately visible in retrieval systems. Hence, contextualization is needed, which is researched in RQ5.

## RQ5: Contextualizing Web archive search at the content level

On the content end, we have enriched the Dutch Web archive's data based on researchers' requests, and in generated subcollections we included elements such as link structure and assigned categories. To find ways to alleviate the inherent incompleteness of Web archives, we looked at using link structure and anchor text to uncover and reconstruct unarchived pages [15, 25]. Our analysis showed that a remarkable number of representations for unarchived pages could be generated, and that the retrieval effectiveness in a known item search setting was surprisingly high. Further experiments included the creation of site-based representations, i.e. representations which aggregate unarchived content at the site-level, instead of the page-level. The uncovered content could be useful for contextualizing Web archive search: interfaces could show both archived and unarchived contents to users, to gain a better understanding of the Web that was. A first prototype, *AuraExplorer*, has been created to allow for exploring both archived pages and unarchived representations of pages.

# 6. CONCLUSION

Current search systems predominantly offer 'one-size-fits-all' approaches for exploring the vast reaches of the Web's information landscape. A singular and static search interface is used for simple and complex tasks alike, and for each *stage*

of a complex task. This proposal discussed ways to move beyond this approach, by experimenting with stage-based differentiation of search support at the interface and content level. The implementation of this idea is not necessarily straightforward: while many information seeking models provide in-depth descriptions of seeking at the macro-level, the connections between these macro-level models and concrete, micro-level search features are fuzzy at best.

We conceptually connect information seeking models and concrete search features, providing an understanding of the utility of features at different moments of complex tasks. To gain further insights, we utilize simulated work tasks to derive actual data on the use of search features and content at different stages of a task. Both the theoretical and practical perspectives provide a foundation for the design of stage-based search systems. Using this base of knowledge, we experiment with systems offering tailored support for stages at the interface and content level. Finally, we provide practical insights in designing analytical search systems in a Web archive research setting.

The main contributions of this work are threefold: tightening the connections between model and practice, providing an understanding of users' needs in various stages of complex search tasks, and deriving design recommendations for *deeper* search systems providing customized search support.

## Acknowledgments

We gratefully acknowledge the feedback received at the doctoral consortium of the 2014 ACM/IEEE *Digital Libraries* conference, and are thankful for the received travel support grant. This research in the context of the WebART project is supported by the Netherlands Organization for Scientific Research (NWO, project # 640.005.001).

## REFERENCES

[1] T. Beckers, S. Dungs, N. Fuhr, M. Jordan, S. Kriewel, and V. T. Tran. ezdl: An interactive search and evaluation system. In *SIGIR 2012 workshop on open source information retrieval*, pages 9–16, 2012.

[2] A. Ben-David and H. C. Huurdeman. Web archive search as research: Methodological and theoretical implications. *Alexandria*, 25(1), 2014.

[3] J. Bronstein. The role of the research phase in information seeking behaviour of jewish studies scholars: a modification of ellis's behavioural characteristics. *Information Research*, 12(3), 2007.

[4] N. Brügger. Website history and the website as an object of study. *New Media & Society*, 11(115), Mar. 2009.

[5] K. Byström and K. Järvelin. Task complexity affects information seeking and use. *Inf. Process. Manage.*, 31 (2):191–213, 1995.

[6] D. O. Case. *Looking for Information : a Survey of Research on Information Seeking, Needs, and Behavior*. Emerald Group Publishing, 2012.

[7] C. M. Chu. Literary critics at work and their information needs: A research-phases model. *Library & Information Science Research*, 21(2):247–273, 1999.

[8] A. P. de Vries, W. Alink, and R. Cornacchia. Search by strategy. In *Proc. Exploiting semantic annotations in information retrieval*, pages 27–28. ACM, 2010.

[9] D. Ellis. A behavioural approach to information retrieval system design. *Journal of documentation*, 45(3): 171–212, 1989.

[10] A. Foster. *Theories of information behavior*, chapter Nonlinear Information Seeking. Information Today, Inc., 2005.

[11] M. Hall, H. Huurdeman, M. Koolen, M. Skov, and D. Walsh. Overview of the INEX 2014 interactive social book search track. In *CLEF 2014 Notebook Papers*, 2014.

[12] M. Hearst. *Search user interfaces*. Cambridge University Press, 2009.

[13] H. C. Huurdeman and J. Kamps. From multistage information-seeking models to multistage search systems. In *Proc. IIiX 2014*, 2014.

[14] H. C. Huurdeman, A. Ben-David, and T. Sammar. Sprint methods for web archive research. In *Proc. WebSci '13*, WebSci '13, pages 182–190, 2013. ACM.

[15] H. C. Huurdeman, A. Ben-David, J. Kamps, T. Samar, and A. P. de Vries. Finding pages in the unarchived web. In *Proc. DL 2014*, 2014.

[16] P. Ingwersen and K. Järvelin. *The turn: Integration of information seeking and retrieval in context*. Springer, 2005.

[17] D. Kelly. Methods for evaluating interactive information retrieval systems with users. *Found. Trends Inf. Retr.*, 3(1-2):1–224, Jan. 2009.

[18] C. C. Kuhlthau. *Seeking Meaning: A Process Approach to Library and Information Services*. Libraries Unlimited, 2004.

[19] J. Lin, K. Kraus, and R. Punzalan. Supporting "distant reading" for web archives. In *Proc. Digital Humanities 2014*, pages 239–241.

[20] G. Marchionini. Exploratory search: from finding to understanding. *Commun. ACM*, 49(4):41–46, 2006.

[21] L. I. Meho and H. R. Tibbo. Modeling the information-seeking behavior of social scientists: Ellis's study revisited. *JASIST*, 54(6):570–587, 2003.

[22] A. Rauber, A. Aschenbrenner, O. Witvoet, R. M. Bruckner, and M. Kaiser. Uncovering information hidden in web archives. *D-Lib Magazine*, 8(12):1082–9873, 2002.

[23] R. Rogers. *Digital methods*. MIT Press, 2013.

[24] I. Ruthven and D. Kelly, editors. *Interactive Information Seeking, Behaviour and Retrieval*. Facet, 2011.

[25] T. Samar, H. C. Huurdeman, A. Ben-David, J. Kamps, and A. P. de Vries. Uncovering the unarchived web. In *Proc. SIGIR '14*. ACM Press, New York NY, 2014.

[26] S. M. Schneider and K. A. Foot. The web as an object of study. *New media & society*, 6(1):114–122, 2004.

[27] The DigIn. Scholarly use of web archives - exploring the election web archives. `http://www.thedigin.org/scholarly-use-of-web-archives-studying-israeli-politics-on-the-web`, 2014. Accessed: 2014-08-04.

[28] E. G. Toms. *Interactive Information Seeking, Behaviour and Retrieval*, chapter Task-based information searching and retrieval. In , Ruthven and Kelly [24], 2011.

[29] P. Vakkari. A theory of the task-based information retrieval process: a summary and generalisation of a longitudinal study. *Journal of documentation*, 57(1):44–60, 2001.

[30] WebART. Web archive search as research.

http://www.webarchiving.nl/news/webarchive-search-as-research, 2014. Accessed: 2014-08-04.

[31] M. Wilson. *Interactive Information Seeking, Behaviour and Retrieval*, chapter Interfaces for Information Retrieval. In , Ruthven and Kelly [24], 2011.

[32] T. D. Wilson. Models in information behaviour research. *Journal of documentation*, 55(3):249–270, 1999.

## 7. APPENDIX A: SCREENSHOTS


**Figure 1: WebARTist prototype interface (*Explore* tab)**