Supporting the complex dynamics of the information seeking process

Huurdeman, H.C.

Publication date
2018

Document Version
Final published version

License
Other

Link to publication

Citation for published version (APA):
Huurdeman, H. C. (2018). *Supporting the complex dynamics of the information seeking process*.

Supporting the Complex Dynamics of the Information Seeking Process

Hugo C. Huurdeman

Supporting the Complex Dynamics of the Information Seeking Process

ILLC Dissertation Series DS-2018-02

INSTITUTE FOR LOGIC, LANGUAGE AND COMPUTATION

For further information about ILLC-publications, please contact

Institute for Logic, Language and Computation
Universiteit van Amsterdam
Science Park 107
1098 XG Amsterdam
phone: +31-20-525 6051
e-mail: illc@uva.nl
homepage: http://www.illc.uva.nl/

SIKS Dissertation Series 2018-05.
The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

Design and lay-out: timelessfuture.com
Cover: Der Wanderer über dem Nebelmeer, Caspar David Friedrich (1818)
Publisher: IR Publications, Amsterdam
ISBN: 978-90-821695-0-8

Supporting the Complex Dynamics of the Information Seeking Process

Academisch Proefschrift

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. K.I.J. Maex
ten overstaan van een door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen in de Agnietenkapel
op woensdag 18 april 2018, te 10.00 uur

door

Hugo Christian Huurdeman

geboren te Amersfoort

Promotiecommissie

Promotores:	Prof. dr. R.A. Rogers	Universiteit van Amsterdam
	Prof. dr. ir. A.P. de Vries	Radboud Universiteit Nijmegen
Copromotor:	Dr. ir. J. Kamps	Universiteit van Amsterdam
Overige leden:	Prof. dr. N.J. Belkin	Rutgers University
	Prof. dr. H.L. Hardman	Universiteit Utrecht
	Dr. A. Helmond	Universiteit van Amsterdam
	Prof. dr. C.M.J.M. van den Heuvel	Universiteit van Amsterdam
	Prof. dr. K.J.P.F.M. Jeurgens	Universiteit van Amsterdam
	Prof. dr. J.S. Mackenzie Owen	Universiteit van Amsterdam
	Prof. dr. P. Vakkari	University of Tampere

Faculteit der Geesteswetenschappen

NWO Netherlands Organisation for Scientific Research

This research was supported by the Netherlands Organization for Scientific Research (NWO), Continuous Access To Cultural Heritage (CATCH) program, project Web Archive Retrieval Tools (WebART), grant number 640.005.001.

Acknowledgments

The *Wanderer* on the cover of this thesis has reached the apex of a mountain, only to reveal the vast landscape still ahead of him. Similarly, my four-year path was an adventure full of twists and turns, ultimately leading to the realization that this PhD thesis is not the end, but merely the beginning of the journey.

My gratitude goes out to Jaap Kamps, for supervising my PhD, explicating the intricacies of research, for countless insights and for keeping me on track. Furthermore, I would like to thank my promotors, Richard Rogers and Arjen de Vries for their vision and feedback. Also, many thanks to my committee members, Anne Helmond, Charles van den Heuvel, Charles Jeurgens, John Mackenzie Owen, Lynda Hardman, Nick Belkin and Pertti Vakkari.

I would also like to thank everyone involved in the WebART project. Thaer Samar, for his insights, friendliness and our collaborations. Anat Ben-David, for her ingenuity, enthusiasm, creativity and support. Sanna Kumpulainen, for wonderful discussions on information seeking and beyond. Many thanks to the people at the Information Access group of the CWI, including Myriam and Gebre, for cool visits, lunches, conversations and table football matches.

My acknowledgements go out to the Koninklijke Bibliotheek. In particular, I would like to thank René Voorburg, for his constant support during the WebART project, as well as our inspiring lunch walks. Furthermore, the various stages of the project involved nice collaborations and discussions with Paul, Hildelies, Barbara, Peter, Lotte, Willem-Jan, Steven, Kees en Martijn.

Many thanks to Max Wilson for the wonderful internship at the Mixed Reality Lab of the University of Nottingham, and to Chris, Horia, Chaoyu and Matthew for being very welcoming, helpful and fun to hang out with.

The process of my PhD was shared with my office companions at the UvA. Marijn, many thanks for our countless interesting discussions and collaborations. Thanks as well to my other office companions and friends, in particular Diana, Hadi, Mostafa, Samira and Alex. Furthermore, I am grateful to the wonderful people at the Digital Methods Initiative, including Anne, Natalia, Bernhard, Erik, as well as the participants in the DMI Summer and Winter Schools, for their enthusiasm, research and many inspiring conversations.

Finally, I would not have reached this point without the continuous and omnipresent support of my family, friends and Lili.

There is no difference between time and any of the three dimensions of space except that our consciousness moves along it.

H.G. Wells, The Time Machine.

Contents

1

Introduction

1.1 Background and Context

In our current times, the World Wide Web is omnipresent, continuously surrounding and supporting our daily life activities. The web's digital information universe has been growing fiercely, and online search engines provide us with handles to find small needles in this giant haystack. Current estimates of the size of search engine indexes suggest they contain tens of billions of pages (van den Bosch et al., 2016). However, the virtually endless amount of content on the web is severely at risk: at any point in time, webpages may appear, change or disappear.

The ephemerality of web content endangers the future understanding of our current times, and has urged institutions, researchers and individuals across the globe to start archiving the dynamic content available on the web. The continuous archiving activities have resulted in Petabytes of valuable web content. However, in contrast to the 'live' web ingrained in all aspects of our daily lives, web archives are desperately waiting for visitors: there is "content now awaiting users, like books in libraries awaiting borrowers" (Rogers, 2013, p.73). The lack of web archive use highlights the importance of understanding the myriad of issues in web archiving, related to harvesting, storage and access (Brown, 2006). Using this understanding we can potentially unlock web archives as a data source for current and future researchers in various fields.

In this thesis, we first focus on the crucial issue of web archive access. In 2001, the Internet Archive introduced the *Wayback Machine* to provide access to the historic websites in their archive via their original URLs. At this point, the majority of institutional web archives across the globe offer URL-based access via the Wayback Machine (Hockx-Yu, 2014). Current web search engines such as Google and Bing, on the other hand, provide a different type of information access, focusing on queries and results lists. While in the 1970s and 1980s, experimental information retrieval systems, often based on document surrogates, were aiming to support "all stages of search performance", subsequent systems have converged to a streamlined feature set. These systems mainly support "query formulation and result list examination, leaving it to the browser to access and display the linked documents" (Ingwersen and Järvelin, 2005, p.137). This streamlined search approach has proved to be extremely effective for basic lookup tasks, and search has become the *de-facto* way to access vastly different types of digital content, for a wide variety of purposes. However, current search approaches are not necessarily suitable for complex research-based tasks, an issue which we will elaborate on next.

1.2 Problem Description and Research Questions

This thesis is inspired by a paradox: on the one hand, search engines on the web provide a world of information at our fingertips, and the answers to many of our common questions are just a simple click away. On the other hand, many of our tasks are complex and multifaceted, and involve a process of knowledge construction: various information seeking models describe a complex set of cognitive stages, influencing the interplay of users' feelings, thoughts and actions (Kuhlthau, 2004; Vakkari, 2001). Despite the evidence of the models, the functionality of search engines, nowadays the prime intermediaries between information and user, has converged to a streamlined set. Even though the past years have embodied rapid advances in contextualization and personalization, our complex information environment is still reduced to a set of ten 'relevant' blue links. This may not be beneficial for supporting complex tasks involving ill-formulated or exploratory needs (White and Roth, 2009), for tasks requiring sustained interaction with information, and for ventures involving the formulation of a deep understanding on a topic (Kelly et al., 2013). This suggests that the currently dominating lookup search approach falls short of the rich interaction needed for task-sharing between user and system (Beaulieu, 2000).

We aim to shed new light upon the apparent contradiction of models describing drastic changes in users' feelings, thoughts and actions, and the limited task support offered in current search systems. Therefore, we need to improve our understanding of the complex tasks involving knowledge construction, but also assess to what extent current search applications constitute helpful frameworks for these tasks. This understanding may facilitate the design of new solutions to support the complex information seeking process, beyond current search interface and system approaches. This general research problem leads to the following main research question of this PhD thesis (**RQ-main**) :

- To what extent do current search approaches support complex information-intensive tasks which involve web content, and how can we support the complex dynamics of the information seeking process?

The first part of this thesis looks at the complex research tasks performed by graduate media and communication researchers, taking the web as their object of study and performing longitudinal analyses of archived web data. To this end, we evaluate search systems and interfaces supporting complex web archive search, but also look at how the underlying *data* may be enriched. The second part of the thesis studies the stages occurring in complex, information-intensive tasks, performed by undergraduate students using web information retrieval systems. Next, we outline the specific research problems and research questions guiding both parts of this thesis.

Research Questions Part I - Search Access to Web Archives

Previous literature, for instance Rogers (2013), has acknowledged the potential of web archives, but also the inherent restrictions that selection policies and access tools impose on the possibilities for research. This inspired us to analyze current access approaches to web archives, but also to assess the possibilities for improving access. Therefore, Part I of this thesis addresses the following research problem (**RP1**):

- to analyze and evaluate search access to web archives in the context of research, and to propose new approaches for search support in a research context.

This first research problem is investigated by means of two main research questions (**RQ1** and **RQ2**). The first research question (**RQ1**), approached in chapter 2, is the following:

RQ1 To what extent do search-based web archive access tools facilitate research in a new media setting?

To address this question, we first perform a study in a new media research context, in which existing access tools were tested, and new search-based access tools were implemented and evaluated. This study, utilizing an action research methodology, looks at the opportunities of the full-text search approach for web research, but also its limitations. A second study expands our perspective to the broader context of media and communication studies. It consists of an analytical literature review of previously published journal papers in media and communication studies which use web content as their data source. This study aims at better understanding which research methods scholars use in their research process.

Regardless of the access interface, there are issues in quality and quantity of the underlying archived web data. In particular, the *completeness* of the archive is an issue: due to legal, organizational and technical limitations, web archives only partially reflect the content of the 'live' web. This incompleteness may influence scholarly research performed using web archives. Therefore, the second research question (**RQ2**) of this thesis was raised:

RQ2 To what extent can representations of unarchived webpages and websites enhance search-based access to web archives?

While the first research question emerged from the need to understand *access systems and interfaces* to web archive data, the second question focuses on the archived *data* itself, aiming to uncover and recover unarchived web content via its underlying link structure. In particular, this research explores how link structure and anchor text in web archives may be used to generate representations of unarchived content. These representations, if rich enough, can potentially be useful to contextualize web archive search systems.

The outcomes of the first research question (**RQ1**) revealed limitations of the full-text search approach to web archive access, in particular transparency issues and a lack of support for the research process. The first issue was tackled in the second research question (**RQ2**). However, it remains an open question what is the best way to integrate *process support* for complex and research-based tasks into actual full-text search systems. To address this question, a better understanding of the role of search system and interface features in complex, information-intensive tasks is needed. As information seeking forms an integral part of the research process, we may be informed by temporally-based information seeking models, which distinguish different **stages** of search (Kuhlthau, 2004; Vakkari, 2001). These models have predominantly been conceived and evaluated in the context of tasks performed by students, as opposed to more experienced researchers. The tasks studied in the context of these models, however, have similarities with the research tasks investigated in the first part of the thesis: they are complex and research-based, and performed in an academic setting. Therefore, we switch the context of our investigation to the performance of research-based tasks by undergraduate students in part II of the thesis, described next.

Research Questions Part II - Supporting Search Stages

Part II covers the second specific research problem (**RP2**) of this thesis:

- to analyze and evaluate the influence of information seeking stages on the usefulness of search system functionality in an online web search context, and to propose new approaches for supporting these stages.

This second research problem is tackled through two research questions (**RQ3** and **RQ4**). In chapter 4, we focus on multistage information seeking models and the support for complex tasks by search systems (**RQ3**):

RQ3 What are the conceptual implications of multistage information seeking models for the design of search systems?

To address this research question, we take a broad approach and aim to bridge the conceptual gap between *macro-level* information seeking models and *micro-level* search systems, by means of a theoretical and practical analysis. On the one hand, we introduce relevant theory from the area of information seeking behavior, including various information seeking models. On the other hand, search user interface (SUI) 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. Then, the theoretical and system perspectives are connected by means of a small-scale user study, followed by a conceptual analysis.

Additional insights beyond the conceptual explorations and the small-scale study of this chapter are needed. Therefore, aiming to understand the value of different categories of SUI features better, we investigate our final research question (**RQ4**) in chapter 5:

RQ4 How can different types of search user interface features support distinct macro-level information seeking task stages?

We look at the actual utility of search user interface (SUI) features at different macro-level *stages* of complex tasks. To this end, a user study is conducted, using simulated work tasks, to explicitly place users within different stages of a complex task. This is combined with a set of measures to detect active and passive user behavior, as well as subjective experiences. We look at the effects of search stages on information seeking, to derive *when* search features are most useful.

Each of the four research questions is divided into various subquestions, which are summarized in Table 1.1. To address the research problems and research questions of this thesis, a varied set of methods is used, described in the next section.

1.3 Methodology Outline

Table 1.2 shows the framing of the thesis per research question, using elements adapted from Ingwersen and Järvelin (2005)'s general model of cognitive information seeking and retrieval.

This thesis uses a mixed approach of qualitative and quantitative methodologies. The research methods and data collection techniques are as follows:

RQ1: Mixed Methods. The first study incorporates an iterative action research design methodology in which needs for web archive search are assessed, and

Table 1.1: Research problems and questions in each part of the thesis.

Part I: Supporting Research Access to Web Archives

RP 1	to analyze and evaluate search access to web archives in the context of research, and to propose new approaches for search support in a research context.
RQ 1	**To what extent do search-based web archive access tools facilitate research in a new media setting?** [Ch. 2]
RQ 1.1	Which limitations of current web archive access tools can be identified in a new media research context? To what extent can search-based tools improve research access to web archives?
RQ 1.2	Which corpus creation, analysis and dissemination methods do media and communication scholars use in the context of web data? What are the implications for search-based web archive access tools?
RQ 2	**To what extent can representations of unarchived webpages and websites enhance search-based access to web archives?** [Ch. 3]
RQ 2.1	What fraction of unarchived web pages and websites can be uncovered based on references to them in the web archive?
RQ 2.2	How can the richness of the representations created for unarchived pages be characterized?
RQ 2.3	To what extent can representations of websites be enriched by aggregating page-level evidence from pages sharing the same hostname?
RQ 2.4	How effective are the derived page-level and site-level representations in a known-item search setting?

Part II: Supporting Search Stages

RP 2	to analyze and evaluate the influence of information seeking stages on the usefulness of search system functionality in an online web search context, and to propose new approaches for supporting these stages.
RQ 3	**What are the conceptual implications of multistage information seeking models for the design of search systems?** [Ch. 4]
RQ 3.1	What are the conceptual implications of multistage information seeking models for the design of search systems?
RQ 3.2	How do current search user interfaces support the information seeking process in the context of complex tasks?
RQ 3.3	To what extent does the search stage influence the flow of interaction at the interface level?
RQ 3.4	How can we reconcile multistage information seeking models and multistage search systems?
RQ 4	**How can different types of search user interface features support distinct macro-level information seeking task stages?** [Ch. 5]
RQ 4.1	How does the user's search stage influence active behavior at the interface level?
RQ 4.2	How does the user's search stage influence passive behavior at the interface level?
RQ 4.3	How is active and passive behavior reflected in the perceived usefulness of features?

Table 1.2: Main framing per part of the thesis

	part I	*part II*
actors	graduate researchers	undergraduate students
context	academic research	
inf. objects	archived webpages	webpages
interface	search interface	
system	full-text search system	

search features are developed, prototyped and evaluated in close collaboration with researchers. This research method involves "the researcher examining current processes, taking action to improve those processes, then analysing the results of the action" (Pickard, 2007, p.134). In the context of this method, the researcher is "an active participant in the process before, during and after the research activity" (p.134). The second study to investigate RQ2 involves a analytical literature review of journal papers in media and communication studies which utilize web data.

RQ2: Quantitative approach. This question involves an analysis of the contents of the Dutch web archive in 2012, and of the richness of representations which can be generated for unarchived content via the archive's link structure and anchor text. The utility of these representations is evaluated via a structured known-item search evaluation. This experimental research, defined by Pickard (2007, p.105) is "an attempt to empirically verify or corroborate the hypothesis of a causal relationship between variables." Evaluation via known-item search, i.e. "searching for a 'known' object or an object 'known to exist', " (Lee et al., 2006), is a common evaluation method in information retrieval. In this case, it involves the indexing of different collections of document representations, creating a set of known-item queries (i.e. queries for a known document in the collection), and calculating the probability of retrieving correct results for each query.

RQ3: Mixed Methods. An extensive literature review on information seeking, information search, information literacy and user interfaces. Taken critical steps in this review are *seeking appropriate sources, evaluation of sources, critical analysis and research synthesis* (as summarized by Pickard (2007, p.26)). This is combined with a small-scale data analysis of experimental eye tracking and log data. Here, the independent variables were task phase (defined as a temporal segment of a task session), while dependent variables were eye tracking fixation counts and *basket* modifications.

RQ4: Mixed methods. An experimental user study employing (cognitively complex) multistage simulated work tasks, studying interaction patterns with interface and content during different search stages. The independent variable is task stage, the dependent variables are 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.

1.4 Conceptual Framework and Research Scope

After having introduced the combination of methodologies to address this thesis' research questions, we focus on our conceptual framework and research scope.

This work is grounded in the broad area of **information behavior**, defined by Wilson (1999) as "the totality of human behavior in relation to sources and channels of information, including both active and passive information seeking, and information use." In this thesis, we predominantly focus on information seeking and searching, subsets of information behavior in Wilson's nested model of research areas (Wilson, 1999). **Information seeking** has been defined by Ingwersen and Järvelin (2005, p.21) as "human information behavior dealing with searching or seeking information by means of information sources and (interactive) information retrieval systems." **Information searching**, in its turn, is a subfield of information seeking in Wilson's nested model, and specifically focuses on the interaction between information user and information system (Wilson, 1999). This thesis also looks at aspects of **information use**, also included in Wilson (1999)'s definition of information behavior. Following various categories of information use defined by Kari (2010), we look at information use as *information processing* (interpreting, analyzing and modifying information), as *knowledge construction* (shaping mental construct as a basis for thinking (Savolainen, 2009)), and as *information production*, creating expressions of knowledge which others can observe.

The concept of **document** has been characterized by Taylor and Joudrey (2009, p.453) as an 'information resource'. As indicated by Buckland (1997), a document may involve "whatever functioned as a document rather than traditional physical forms of documents", hence also encompasses the web-based documents addressed in this thesis. These include webpages, which Koehler (1999) has defined as "collections of Internet objects" (including text, images, videos and scripts), "that can be navigated without recourse to hypertext linkages". He defined websites as collections of one or more webpages that "share some common theme or organizing principle", and which are connected by hypertext links. In a research setting, web documents may be approached at various interrelated analytical levels (Brügger, 2009), including the *web sphere*, *website*, *webpage* and *elements of webpages*. Traditionally, document **representations** (or bibliographic records) entail the "full descriptive and access information for an information resource" (Taylor and Joudrey, 2009, p.446). We also investigate other types of document representations, by utilizing implicit evidence from the content of the web and the archived web. This evidence includes the web's interlinked structure and associated properties such as anchor text. **Metadata** may be assigned to documents, "structured information that describes the attributes of information resources for the purposes of identification, discovery, selection, use, access, and management" (Taylor and Joudrey, 2009, p.89).

Relevance, in the context of information science, is a multifaceted concept. It has been defined by Ingwersen and Järvelin (2005, p.21) as: "the assessment of the perceived topicality, pertinence, usefulness or utility, etc., of information sources, made by cognitive actor(s) or algorithmic devices, with reference to an information situation at a given point in time." Saracevic (1996) has provided the following characterizations of relevance manifestations: *system* (algorithmic) relevance, relations between a query and documents in a system; *topical* (subject) relevance, relations between the topic in a query and the topics reflected in retrieved texts; *cognitive* relevance (pertinence), the relation between the state of knowledge and cognitive information need of a user, and the texts retrieved; *situational* relevance (utility), the relation between the situation and retrieved items; and *motivational* (affective) relevance, the relations between a user's motivation and texts retrieved by a system. Part I of the thesis mainly looks at system, topical and situational relevance, while part II particularly focuses on topical, cognitive and situational relevance.

In this thesis, we distinguish between the full-text search **system**, the back-end of a search engine, and the search user **interface** (SUI), the front-end of a search engine. The search system indexes **document collections**, in the thesis also denoted as **datasets**. These document collections can consist of 'live' web documents, which are defined as documents which are currently available via the World Wide Web. A singular version of such a document is indexed by a search system. Collections of archived web documents, on the other hand, may involve multiple versions of the same resource, which evolve along the temporal dimension, but may also include document duplicates. Both types of documents are usually harvested via a **web crawler**, which iteratively follows links in web content and thus crawls contents of the web.

In the various chapters of this thesis, we investigate work tasks of different actors. **Work task** has been defined as a "job-related task or non-job associated daily-life task or interest to be fulfilled by cognitive actor(s)" (Ingwersen and Järvelin, 2005, p.20). These tasks may be real-life tasks, or assigned simulated work tasks (Borlund, 2003). We look at *complex work tasks*, which can be defined as work tasks which require "understanding, sense-making, and problem formulation" (Byström and Järvelin, 1995). Complex tasks go beyond simple lookup tasks, and involve learning and construction. Work tasks may lead to one or more **search tasks**, defined as "the task to be carried out by a cognitive seeking actor(s) as a means to obtain information associated with fulfilling a work task" (Ingwersen and Järvelin, 2005, p.20).

In terms of these **cognitive actors**, part I of this thesis focuses on *media and* *communication scholars*, taking the web as their object of study. In particular, we look at how (graduate) researchers use archived, temporal web data, and how current URL-based and custom search-based access interfaces influence the use (and non-use) of web archives. The data utilized by researchers may be contained in larger web archives, or may be harvested by the researchers themselves. The subsequent analytical literature review of the broader fields of media and communication may triangulate our findings and extend them to other user groups. Part II of this thesis specifically investigates information seeking by undergraduate students in the context of complex, research-based "essay writing" tasks. These tasks are generally performed using scholarly and non-scholarly sources accessible via the web.

1.5 Contributions and Limitations

The main contribution of this thesis is that it charts the emerging issues of using current search approaches in the context of complex, information-intensive tasks. By understanding these issues, we may ultimately devise solutions in terms of improved search systems and interfaces for complex information tasks. This thesis includes the following contributions, related to three main topics:

- Supporting research use of web archives:
  - An improved understanding of the emergent search access needs of scholars which use the web as a data source [Ch. 2]
  - A better understanding of how search system features support the research process of scholars, and of the limitations of these features [Ch. 2]
  - Methods to reveal and contextualize what material is missing based on the web archive's link structure and anchor text [Ch. 3]
  - An assessment of the utility for retrieval of reconstructed document representations of unarchived contents generated [Ch. 3]
- Supporting complex search processes in search-based access systems:
  - A better understanding of the limitations of contemporary online web search approaches in the context of information-intensive tasks [Ch. 4]
  - Insights into the connections between stages in macro-level information seeking models and micro-level search system features [Ch. 4]
  - A better understanding of passive and active use of search user interface features [Ch. 5]
  - Insights into *when* (categories of) search user interface features are most useful in complex search processes [Ch. 5]
- Helpful frameworks for complex task performance using search systems:
  - A mapping between research process models and information seeking models, thus modeling the contemporary research process as a process of construction [Ch. 6]

  - A better understanding of how low-level search system support for moves and tactics may gradually give way to higher level support for stratagems and strategies [Ch. 6]

Naturally, any approach to this complex problem, including our approach, is not without limitations. We chose two particular target groups for our research, namely (graduate) media and communication scholars, as well as undergraduate students. This may result in limitations with respect to generalizability of our findings. Second, the chosen research methods, including action research methodologies, analytical literature reviews as well as user studies each have inherent strengths and weaknesses, which are described in the chapters in which they are used.

1.6 Thesis Origins

This PhD thesis consists of the introductory matter, followed by four main chapters, each based on one of the four research questions described above. Finally, it closes off with the Conclusions, Appendices and Bibliography. Next, we present the origins of each chapter in terms of the scholarly papers they are based on.

Chapter 2 – Browse, Search & Research: Evaluating & Extending the Functionality of Web Archive Access Tools

Chapter 2 aims to answer the first research question (**RQ1**), about the evaluation and potential extension of search access to web archives. This chapter is directly based on a book chapter and conference paper:

- H. C. Huurdeman and J. Kamps. A Collaborative Approach to Research Data Management in a Web Archive Context. In *Research Data Management - A European Perspective*. Walter de Gruyter GmbH, 2018 [**RQ1.1**]

  Author roles: H.C. Huurdeman initiated the writing, with contributions by J. Kamps.

- H. C. Huurdeman. Towards Research Engines: Supporting Search Stages in Web archives. In *Web Archives as Scholarly Sources conference 2015*, Apr. 2015. `http://events.netlab.dk/conference/index.php/resaw/june2015/paper/view/85` [**RQ1.2**]

Related work, cited in this chapter, has been published as:

- H. C. Huurdeman, A. Ben-David, and T. Samar. Sprint Methods for Web Archive Research. In *Proceedings of the 5th Annual ACM Web Science Conference*, WebSci '13, pages 182–190, 2013. ACM. `http://dx.doi.org/10.1145/2464464.2464513`

- A. Ben-David and H. C. Huurdeman. Web Archive Search as Research: Methodological and Theoretical Implications. *Alexandria*, 25(1-2):93–111, Aug. 2014. http://dx.doi.org/10.7227/ALX.0022

- L. Melgar, M. Koolen, H. C. Huurdeman, and J. Blom. A process model of scholarly media annotation. In *Proceedings of the 2017 Conference on Human Information Interaction and Retrieval*, CHIIR '17, pages 305–308. ACM, 2017. http://dx.doi.org/10.1145/3020165.3022139

Chapter 3 – Lost but Not Forgotten: Finding Pages on the Unarchived Web

Chapter 3 investigates the second research question (**RQ2**), and is directly based on a journal article about uncovering and recovering parts of the unarchived web:

- H. C. Huurdeman, J. Kamps, T. Samar, A. P. Vries, A. Ben-David, and R. A. Rogers. Lost but Not Forgotten: Finding Pages on the Unarchived Web. *Int J on Digital Libraries*, 16(3):247–265, 2015b. http://dx.doi.org/10.1007/s00799-015-0153-3 [**RQ2.1-2.4**]

  Author roles: H.C. Huurdeman initiated the writing and **all authors** contributed. T. Samar carried out data extraction and processing. Further processing and analysis by H.C. Huurdeman, with contributions by T. Samar, J. Kamps and A. Ben-David. Known-item search evaluation by H.C. Huurdeman, with contributions by J. Kamps and A. Ben-David.

Previous versions of this journal article were published as:

- H. C. Huurdeman, A. Ben-David, J. Kamps, T. Samar, and A. P. de Vries. Finding Pages on the Unarchived Web. In *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '14, pages 331–340, 2014. IEEE Press. http://dx.doi.org/10.1109/JCDL.2014.6970188

- T. Samar, H. C. Huurdeman, A. Ben-David, J. Kamps, and A. de Vries. Uncovering the unarchived web. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, pages 1199–1202, 2014. ACM. http://dx.doi.org/10.1145/2600428.2609544

Chapter 4 – From Multistage Information-Seeking Models to Multistage Search Systems

Chapter 4 aims to answer our third research question (**RQ3**), and looks at the conceptual implications of multistage information seeking models on the design of search systems. This chapter is based on two conference papers:

- H. C. Huurdeman and J. Kamps. From Multistage Information-seeking Models to Multistage Search Systems. In *Proceedings of the 5th Information Interaction in Context Symposium*, IIiX '14, pages 145–154, 2014. ACM. http://dx.doi.org/10.1145/2637002.2637020 [**RQ3.1-3.3**]

Author roles: H.C. Huurdeman initiated the writing, with contributions by J. Kamps.

- H. C. Huurdeman and J. Kamps. Supporting the Process: Adapting Search Systems to Search Stages. In *Information Literacy: Moving Toward Sustainability*, number 552 in Communications in Computer and Information Science, pages 394–404. Springer International Publishing, Oct. 2015. `http://dx.doi.org/10.1007/978-3-319-28197-1_40` [**RQ3.4**]

  Author roles: H.C. Huurdeman initiated the writing, with contributions by J. Kamps.

Related work, cited in this chapter, has been published as:

- H. C. Huurdeman, J. Kamps, M. Koolen, and S. Kumpulainen. The Value of Multistage Search Systems for Book Search. In *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum*, volume 1391, 2015a. CEUR-WS. `http://ceur-ws.org/Vol-1391/85-CR.pdf`
- S. Kumpulainen and H. C. Huurdeman. Shaken, not Steered: the Value of Shaking Up the Search Process. In *Proceedings of the First International Workshop on Supporting Complex Search Tasks co-located with ECIR*, 2015. CEUR-WS. `http://ceur-ws.org/Vol-1338/paper_7.pdf`

Chapter 5 – Active and Passive Utility of Search Interface Features in Different Information Seeking Stages

Finally, Chapter 5 is guided by the fourth research question of this thesis (**RQ4**), and looks at the support by SUI features for macro-level information seeking task stages.

This chapter is an extended version of a conference paper:

- H. C. Huurdeman, M. L. Wilson, and J. Kamps. 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, pages 3–12, 2016. ACM. `http://dx.doi.org/10.1145/2854946.2854957` [**RQ4.1-4.3**]

  Author roles: H.C. Huurdeman initiated the writing of this paper, with contributions by M.L. Wilson and J. Kamps. Creator of experimental system: H.C. Huurdeman. Performer of experiment: H.C. Huurdeman. Data analysis by H.C. Huurdeman, with contributions by M.L. Wilson and J. Kamps.

Chapter 6 – Conclusion

An extended version of Section 6.2 has been published as the following article:

- H. C. Huurdeman. Dynamic Compositions: Recombining Search User Interface Features for Supporting Complex Work Tasks. In *CHIIR 2017 Second Workshop on Supporting Complex Search Tasks (SCST 2017)*, pages 21–24. CEUR-WS, 2017. `http://ceur-ws.org/Vol-1798/paper5.pdf`

Part I

Supporting Research Access to Web Archives

Part I:
Supporting Research Access to Web Archives

If we lose our past, we will live in an orwellian world of the perpetual present, where anybody who is in control of the information that is out there, will be able to say what is true and what is not.
Kahle (2014), founder Internet Archive

In the future, web archives may be the main testimonies to our current times. Day by day, our lives progressively move more to the web. The evidence of our lives and society embedded in the internet is crucial to current and future researchers. For instance, it would not be possible to write the history of the 1990s without taking the web into account (Milligan, 2016). To save the web for posterity, institutions across the globe have started to archive its contents. However, many difficulties exist in web archiving procedures and access. As a result, thus far, few studies have been performed using the resulting institutional web archives, even though these archives have amassed Petabytes of web data by now. It is important to understand why this is the case, and how we could better facilitate research use of web archives.

Part I of this thesis focuses on the first research problem (**RP1**): *to analyze and evaluate search access to web archives in the context of research, and to propose new approaches for search support in a research context.* First, in chapter 2, we evaluate the limitations of the data available in archives, and of the access tools that serve as a crucial intermediary between archive and user. This is done via participatory action research in the context of new media studies. In this chapter, also further requirements for scholarly access to web archives are determined via an analytical literature review, focusing on the research process of scholars. Subsequently, following a found lack of transparency of web archives, chapter 3 evaluates approaches to reveal and potentially reduce the incompleteness of web archives via their link structure and anchor text. This results in representations of otherwise lost web contents.

2

Browse, Search & Research: Evaluating & Extending the Functionality of Web Archive Access Tools

Organizations and individuals across the globe capture the Web's evolving content and assemble it into web archives. These valuable archives could be used as research datasets in various settings. Despite their potential value, they have scarcely been used for research thus far. In this chapter, we investigate why this is the case, and uncover underlying reasons for the lack of scholarly use of web archives. To this end, we first introduce the concept of web archiving, the actors involved, and the limitations of web archives as prospective research datasets (Section 2.2). To gain more insights into new media scholars' needs, we perform a study in which we develop, extend and evaluate search-based access tools in an action research setting (Section 2.3). In this setting, we encounter limitations of full-text search tools, including a lack of transparency and a lack of support for various research methods. To address the latter limitation, we further investigate research use of web archives: utilizing a set of eighteen journal papers, we review how media and communications scholars define their dataset, which analysis methods they use and how they disseminate their results (Section 2.4). Based on our findings, we provide recommendations for overcoming the limited transparency of search access tools. Furthermore, we discuss concrete ways to address the lack of research process support in current web archive access systems. This way, we may facilitate a move from mere search engines to potential 'research engines'.

This chapter is based on Huurdeman and Kamps (2018), as well as an extended and revised version of Huurdeman (2015). Related work has been published as Huurdeman et al. (2013); Ben-David and Huurdeman (2014); Melgar et al. (2017).

2.1 Introduction

While large-scale digitization projects have made millions of book, newspaper and journal pages available to researchers, these data sources are still dwarfed by the massive scale of web archives containing born-digital web content. As we will see in the next chapter, the Dutch national library captured 25 million webpages for a selection of 5,000 websites in 2012 alone, and the US-based Internet Archive has archived 491 *billion* webpages during the last 20 years[1]. Still, despite the intricately detailed view that web archives offer on humankind's increasingly 'digital' lives and activities, thus far, web archives have scarcely been used for scholarly research.

This chapter examines why this is the case. The aim of this chapter is to better understand the limitations of current web archive access tools in a research context. This leads to the following main research question (**RQ1**): *To what extent do search-based web archive access tools facilitate research in a new media setting?* This research question is tackled using the following subquestions:

RQ1.1 Which limitations of current web archive access tools can be identified in a new media research context? To what extent can search-based tools improve research access to web archives?

In an action research setting with new media scholars, we look at the prospective use of current web archive access tools by new media researchers, and at the limitations of existing tools.

RQ1.2 Which corpus creation, analysis and dissemination methods do media and communication scholars use in the context of web data? What are the implications for search-based web archive access tools?

This research question features a bottom-up analytical literature review of past journal papers by media and communication scholars to derive a better understanding of appropriate research process support.

This chapter consists of three main sections. First of all, in Section 2.2, we introduce the concept of a web archive, characterize what constitutes the archive, and discuss the actors involved in web archiving. We also look at the properties of web archives as research datasets, and their potential deficiencies and limitations. Then, we introduce a concrete case in Section 2.3, in which we performed experiments towards making web archives available as research datasets. Section 2.4 investigates the scholars' research process via an analytical literature review. Ultimately, the chapter's conclusions are discussed in Section 2.5.

[1] `http://web.archive.org/web/20160703022445/https://archive.org/web/` (accessed: 01/08/16)

2.2 Related Work

2.2.1 An Introduction to Web archives

This section introduces the concepts utilized in the remainder of this chapter. It discusses the rationale behind web archiving, various definitions, and the variety of actors which are involved in archiving.

Preserving the Web

The ever-growing World Wide Web takes up a pivotal role in our everyday lives. We use the web to lookup information, to communicate, and for our daily entertainment and leisure. However, the web is of a highly ephemeral nature: if a server disappears, the content is lost (Masanès, 2006, p.7), and if a website is renewed, content may be moved, changed or deleted altogether. Hence, "the content and structure of the web are constantly in flux", and proactive steps have to be taken to ensure that web content will be preserved (Dougherty and Meyer, 2014). Thus, as Kahle has indicated[2], through web archiving, we may enable a 'memory' of the web and avoid to get stuck in a 'perpetual present'. Various individuals and institutions at local, national and international scales have taken up this challenge, together harvesting Petabytes of valuable web material. In the complex and volatile environment of the current web, however, archiving institutions have a hard time keeping up with the technological developments, but also with the web's massive scale. Estimates based on different samples taken in 2012 indicated that about 35-90% of the web was at least archived once (Ainsworth et al., 2012), but this does not even take into account information in the Deep Web, unreachable for web archive harvesting tools. Further hindrances are formed by privacy issues, intellectual property and copyrights (Masanès, 2006). Before delving deeper into these issues, we first discuss definitions of web archiving.

Defining Web Archiving

Web archiving has been defined by the International Internet Preservation Consortium (IIPC)[3] as "the process of collecting portions of the World Wide Web, preserving the collections in an archival format, and then serving the archives for access and use." Another definition by Ball (2010) focuses on more specific procedural aspects, characterizing web archiving as "the selection, collection,

---

2 `http://brewster.kahle.org/2015/08/11/locking-the-web-open-a-call-for-a-distributed-web-2/` (accessed: 01/08/16)

3 This definition is available via: `http://www.netpreserve.org/web-archiving/overview/` (accessed: 01/08/16)

storage, retrieval, and maintenance of the integrity of web resources". Complementary to this more institutional perspective, Niels Brügger, a Web historian, focuses on the intention and rationale behind archiving: "Web archiving means any form of deliberate and purposive preserving of web material" (Brügger, 2009). Brügger elaborates that this definition implies that archiving is a *conscious* act: the act of preserving the material itself, but also the conscious reasoning about why the material is collected and preserved. Different actors may be involved in this process, discussed in the next section.

Classifying Web Archiving Actors

An increasing number of institutions, companies, groups and individuals are collecting web material[4]. As this very diverse group of actors implies, Web archiving may be done for a great variety of purposes (see Brown (2006); Masanès (2006); Brügger (2009)). The way the archived web is formed differs "based on who does the archiving, when, and for what purpose" (Brügger, 2005). These purposes may include *collection building and preservation* (for example by libraries and archives), *research* (for example in the context of a research institution), or to address applicable *legislation*. The latter may be obliging web archiving due to *legal deposit* laws, requiring institutions to document all published documents in a country (e.g. national libraries in the UK or Denmark)[5], or due to archival laws, obliging government entities to archive their own website. In terms of their funding, the initiatives can further be divided in *state-funded*, *nonprofit* and *commercial* web archives (Masanès, 2006, p.41).

Brügger (2009) provides a broad division of web archiving efforts based on their scale: on the one hand, *macro* archiving entails the archiving of web material by professionals, often in the context of national or local institutions. On the other hand, *micro* archiving involves small-scale archiving carried out by researchers and other individuals, based on a "here-and-now" need to preserve an object of study.

The prime example of an institution applying a *macro* perspective to web archiving is the Internet Archive, a non-profit institution which began archiving the web in 1996 on a massive and transnational scale. Other institutional initiatives, often state-funded, may range from local-level (e.g. a Municipal Archive), to regional and national scales (for instance the UK Web Archive or the Netarkivet in Denmark). Also, an increasing number of commercial initiatives

4 An updated list of web archiving initiatives is available at: `https://en.wikipedia.org/wiki/List_of_Web_archiving_initiatives` (accessed: 01/08/16)

5 To be more precise, legal deposit legislation "defines a legal obligation for publishers to deposit copies of all published works with designated libraries, in order to maintain a comprehensive collection of a nation's published output" (Brown, 2006, p.158)

provide Web archiving as a service[6].

The *micro*-level of archiving is for instance reflected by individual researchers, who may gather their own collections of web material in the context of their research. Finally, a more blurred category exists of bottom-up 'crowdsourced' initiatives like ArchiveTeam, which jointly archive web material, using shared methods and highly collaborative approaches[7]. In this thesis, we mainly focus on the *macro* approach to web archiving, i.e. the larger web archives assembled by cultural heritage institutions.

A Transition from Building Infrastructure to Supporting Use

In the early years of web archiving, the archives predominantly focused on preservation and creating the infrastructure to harvest internet pages – in itself not an easy task, due to the voluminous scale of the web. As Thomas et al. (2010) have argued, much of the preliminary archiving efforts have been done "from the point of view of archiving for its own sake", but less work has been carried out towards the actual *use* of these archives by researchers. As Rogers (2013, p.72) has put it, "Web archiving infrastructure receives scholarly and nonscholarly attention; the archived materials –the primary source materials– gain less notice." A related issue, as indicated by Thomas et al. (2010), is that "the traditional practices of the field of Library and Information Science" have dominated web archive development, not necessarily providing the right handles for humanities and social sciences researchers. Dougherty and Meyer (2014) suggest that there is a "wide gap between the researchers who need archival data sets to support their studies of online phenomena, and the archivists and other practitioners who have the expertise to build such collections and the tools to manage and access them". Moreover, they suggest that these efforts have "so far not yet provided reliable methodological solutions for researchers who wish to use archived web materials". In the next section we zoom in on the issues that may arise in the use of web archives as research datasets.

Summarizing, the past two decades have shown an imminent rise of web archiving initiatives, preserving and providing access to our online past. A gradual move from building infrastructure to supporting use has emerged, but at the

[6] For instance, Archive-It (`https://www.archive-it.org/` (accessed: 01/08/16)), a spin-off of the Internet Archive, provides subscription-based web archiving services, for instance used by cultural institutions. The Canadian-based Pagefreezer (`https://www.pagefreezer.com/` (accessed: 01/08/16)) provides web archiving, also for digital evidence purposes, while Archiefweb (`http://www.archiefweb.eu/` (accessed: 01/08/16)) captures many Dutch government websites.

[7] To give a practical example, the Dutch pre-Facebook social network Hyves was going offline on a short notice, and via ArchiveTeam a joint collective of individuals managed to harvest around 9M public profile pages in time (`http://www.archiveteam.org/index.php?title=Hyves` (accessed: 01/08/16))

moment we are still in the beginning of that transition. The myriad of ongoing initiatives, ranging from bottom-up initiatives to large-scale institutional archiving, shows the importance and multidimensional aspects of web archiving. However, it also results in datasets that vary in multiple ways, due to differences in the purpose of the archives, their approaches and selection criteria. How and to what extent these aspects influence the suitability of web archives as research datasets is the topic of the next section.

2.2.2 Web Archives as Research Datasets

Web archives, due to their vast and diverse contents, can provide a valuable resource for scholars in various disciplines, for instance computer science, the humanities and the social sciences. In theory, web archives may allow for novel research questions and methods, but in practice many carefully crafted archives have remained underused (Dougherty and Meyer (2014)). Utilizing web data in general, and archived web data in particular, introduces various challenges when performing research. This section focuses on these challenges, and looks at various limitations of web archives in research context. These limitations may be divided in three categories: limitations in *data quantity*, limitations in *data quality* and limitations in *access*.

Limitations in Data Quantity

Web archiving is predominantly performed by web crawlers, which "harvest content from remote web servers" (Brown, 2006, p.50). Similar to the crawlers used by common search engines, these crawlers iteratively follow hyperlinks within webpages to capture content. This can be done using three main strategies (Brügger, 2011), which influence the range of data which is captured in archives. Broad *domain* crawls are delimited by the boundary of the national top-level domain (e.g. *.uk*), *selective* snapshots focus on a predefined selection of websites, and *event* harvests focus on important ongoing events. These three strategies feature distinct trade-offs in terms of the breadth and depth of captured content.

First, the *broad domain* strategy entails taking a snapshot of web documents at one or more points in time. This approach typically implies a 'breadth-first' strategy, meaning that the web crawlers focus on capturing the breadth of web material as opposed to the depth. Hence, a wide range of content may be captured for posterity, but this method also implies that material located deep in a website may not be captured. Domain crawls also result in very large datasets, making quality assurance hard to manage, and can take a very long time to complete. For instance, the full *.uk* domain crawl of the British Library in 2013 took almost eleven weeks to complete, leading to a sizable set of material totaling in

31TB[8]. A *selective* strategy, on the other hand, may result in a more manageable set of material. This strategy is based on a (finite) selection of websites, based on certain properties, such as subject, creator, genre or domain (Brown, 2006, p.31). The selective nature implies that material outside the selection lists is excluded from archiving, although using this approach, the amount of captured material per site may be higher: crawlers are usually configured to follow links deeper into a domain. As Brown (2006, p.32) has argued, the selective approach "is likely to facilitate a more detailed understanding of the properties and qualities of the individual resources collected" – it may be feasible for archiving entities to perform quality control, or to adjust crawl settings for individual websites. A third common strategy is *event harvesting*, i.e. the harvesting of material related to events on a local, national or international scale. For instance, in the context of the International Internet Preservation Consortium (IIPC), institutions perform collaborative harvests. Covered events may consist of anticipated and planned events (e.g. the the 2014 Winter Olympics in Sochi) or unplanned events (e.g. the 2005 Katrina hurricane).

Regardless of the approach, temporal omissions exist, and, depending on the webpage captured, these may influence the types of research which can be performed with a web archive. In effect, as Masanès (2006, p.17) has argued, archiving "always implies some selectivity, even if it is not always in the sense of manual, site-by-site, selection", and "the archived portion of the web will always only be a slice in space and time of the original web". Moreover, for individual researchers, limitations may exist in the large multipurpose *macro*-level archives, as they often apply a one-size-fits-all approach to web archiving, which may involve generalized selection policies, crawl settings, and crawling schemes. For instance, a researcher's interest may include sources that lie outside the scope of a web archive's selection criteria (for example highly controversial websites), or a researcher may need more frequent harvests for a certain website. In addition, certain popular websites are difficult or impossible to crawl with regular crawling techniques. For instance, capturing content from social media sites such as Facebook and Twitter may necessitate specialized tools, or API[9] access.

Limitations in Data Quality

There are inevitable limitations in archived data quality, which we initially will define as the extent to which captured web content resembles the original con-

---

8 See: `http://britishlibrary.typepad.co.uk/webarchive/2013/09/domaincrawl.html` (accessed: 01/08/16)

9 API stands for 'Application Programming Interface'. Web APIs may facilitate programmatic access to a website's underlying data and features. An extensive list of APIs is available from `http://www.programmableweb.com/apis` (accessed: 01/08/16). The use of APIs for research purposes lies outside the scope of this thesis.

tent on the 'live' web. First of all, there are technical issues related to the archivability of a website: some data formats and certain types of interactive websites cannot be archived, for instance form-based pages, or dynamic web content based on HTML5 techniques. This leads to an incompleteness at several levels: at the level of the website (individual pages may be missing), but also at the level of the page (page elements, such as embedded material may not be included). Rogers (2013, p.64) summarized it as such: "In a sense, the 'new media' elements (cookies, embedded material, recommendations, comments, etc.) are eliminated for posterity, and a traditional content container, looking somewhat broken for its missing pieces, remains as the 'archived website' ". Furthermore, the "interconnectedness", i.e. the unique hyperlink-based nature of the web may get lost (Masanès, 2006, p.17); (Rogers, 2013, p.63); and in effect, the archived website, an assembled object, becomes detached from its larger context (Helmond, 2015, p.118). Thus, we may arrive at something different from the 'live' web in a multitude of ways.

In addition, temporal inconsistencies may occur (Brügger (2005, 2009, 2011)). For instance, capturing a large website such as `www.cnn.com`, may take a long time, during which contents of some pages have already changed. For instance, the homepage may have been crawled first, but while the crawl is running, other news items have been added. At times, deeper crawled pages may reflect the initial state, but in other cases also later temporal states of the website. These issues lead Brügger to argue that an archived page is a "version" and not a "copy" of a website[10]. Moreover, what is the *right* version of content is often unknown: web servers may adapt content to each request, for instance based on the device that a user utilizes for accessing the web. Hence, the web may be seen as "a black box with resources, of which users only get instantiations" (Masanès, 2006, p.13). Thus, the captured material may in many cases be different than the original resources.

Limitations in Access

Web archiving institutions across the globe are spending substantial efforts on *collecting* and *preserving* our valuable web heritage, but another crucial issue is *providing access*. Several factors influence access to archives.

First of all, legal reasons may impede archive access. While some archives are fully accessible online (e.g. the Portuguese web archive), for the majority of archives this is not the case. Some web archives are only accessible from the institution's premises (e.g. the National Library of the Netherlands), other archives are partially accessible online (e.g. the UK Web Archive), and some archives, so-called *dark* archives, may not provide access to end-users at all. To

[10] For these reasons, Brügger has classified the content of web archives as "re-born digital material" instead of "born digital material"

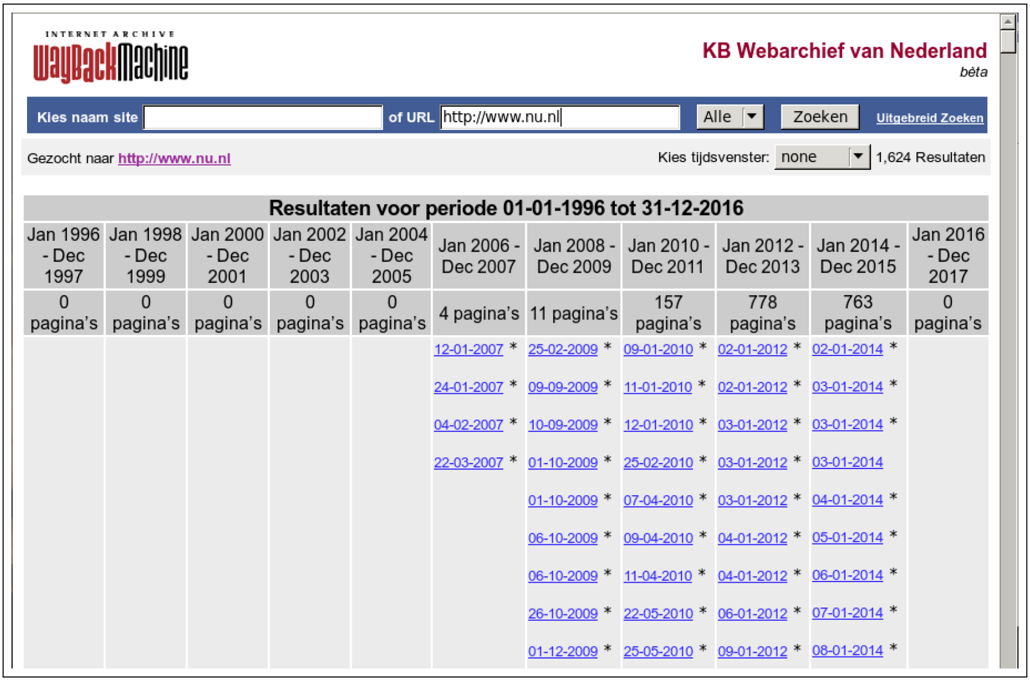

Figure 2.1: Screenshot of Wayback Machine of the Dutch web archive (January 2016), showing the crawl selection screen for the Dutch news website *nu.nl*.

make matters worse, these large institutional web archives usually focus on capturing web content from their own national domain (Rogers, 2013). Combined with the imposed legal barriers, this impedes transnational research using web archives.

The second limitation, the main focus of the remainder of this chapter, lies in the access systems and interfaces, the intermediary between the data in the archive and the potential user of this data. Web archives have taken different approaches to provide access, including *URL-based*, *browse* and *search-based* access options.

The most common way of accessing content is through the Wayback Machine[11], which "allows users to locate archived website snapshots, to differentiate between multiple snapshots of the same site collected on different dates and to navigate across all content collected at a certain point in time, effectively recreating the original context of that content" (Brown, 2006, p.135). The Wayback Machine essentially provides URL-based access to webpages in web

[11] The Internet Archive started harvesting web content in 1996. During the first years no public access interfaces were available. This changed in 2001, however, with the introduction of the Wayback Machine, as evidenced on the following webpage: `https://web.archive.org/web/20011026003810/http://www.archive.org/wayback/press_kit/index.html` (accessed: 01/08/16).

archives, but evidently, this necessitates the knowledge of the URL and date of a certain page or site (potentially unavailable). Moreover, whilst the Wayback Machine interface preserves the flow of surfing, it may "jump cut through time" (Rogers, 2013, p.66). Followed links may lead to an archived page captured at the same time as the source page, to the page "closest" to that date (which may vary considerably), or even to the current page on the 'live' web. Finally, in a research context, the Wayback Machine predominantly facilitates 'single-site' histories (Rogers, 2013, p.66) (Ben-David and Huurdeman, 2014): it is possible to track the evolution of singular webpages, but there is no integrated way to study multiple websites, or perhaps the whole archive.

Additionally, some archives, such as the UK Web Archive, provide ways to browse the contents of the archive through subjects and collections[12]. However, for archives, these hierarchical classifications may be difficult to create and maintain; and a user's navigation may be "limited by the classification decisions made by the archive" (Brown, 2006, p.129).

In essence, both URL and browse-based access approaches are still 'document-centric' methods, focusing on separate documents (Hockx-Yu, 2014). In effect, the current user interface for web archives may work "well with small, curated collections but does not scale up and provide the users with a functional way to use larger collections" (Hockx-Yu, 2014). To a certain extent, the addition of full-text search access to web archives has allowed for a broader view, substantially enhancing access, and "scaling the analysis from the single URL to the full archive" (Ben-David and Huurdeman, 2014)[13]. Search-based access may have many advantages, and overcomes the necessity to know URLs in advance, but the next section will show that also a number of issues are involved.

In sum, web archives potentially provide numerous opportunities for research, but their potential for reuse as research datasets has not been fully harnessed yet. In part, this is caused by issues in *data* and *access*, only corroborated by the variety of actors involved in web archives discussed in the previous section, which take different approaches to archiving. To better understand these limitations, and to potentially amend them, we examine the actual use of the web archive in a practical setting. Taking a large Dutch research project about web archives as its basis, the next section discusses the pitfalls and opportunities of using web archive data for research, and evaluates search-based access approaches.

[12] `http://www.webarchive.org.uk/ukwa/browse/` (accessed: 01/08/16)

[13] Since the late 2000s, an increasing number of archives is also offering full-text search, allowing for textual queries against their collections. At this moment, more than half of the documented web archive initiatives (36 of 59 listed archives listed at: `http://en.wikipedia.org/wiki/List_of_Web_archiving_initiatives` (retrieved 10 April, 2015)) include this service.

2.3 Experiments to Improve Research Access to Web Archives

After discussing the limitations of web archives as research datasets in the previous section, this section describes a number of experiments towards improving access to web archives in a scholarly context. In this section, we investigate the first research question of this chapter (**RQ1.1**): *Which limitations of current web archive access tools can be identified in a new media research context? To what extent can search-based tools improve research access to web archives?*

This section is structured as follows: first, we provide an introduction to the overarching research project. Subsequently, the action research methodology is discussed, followed by the findings from the various research phases.

2.3.1 Introduction to WebART

A highly influential and long-running research program in the Netherlands was the CATCH (Continuous Access To Cultural Heritage) program[14], which aimed at making "the collections of museums, archives and historical associations more accessible." Between 2005 and 2016, the program has funded eighteen multidisciplinary projects, in which researchers and heritage institutes collaborated to improve access to Dutch cultural heritage collections. The WebART project[15] (2012-2016), part of CATCH, has looked at ways to evaluate the current use of web archives and to design novel access methods, both from theoretical and practical perspectives. In the WebART project, the University of Amsterdam[16] and Centrum Wiskunde & Informatica[17] (CWI) joined forces with the National Library of the Netherlands[18] (KB). The interdisciplinary project involved researchers with backgrounds in computer science, information science and new media & digital culture. The main collection studied in the project was the KB's web archive[19]. The KB initiated their web archiving program in 2007, employing a selective policy. As of January 2016, over 10,000 Dutch websites are harvested on a regular basis, the full archive amounting to over 18 Terabytes. This archive is accessible from the National Library's premises, via the URL-based *Wayback Machine*[20].

Exploring Researchers' Needs The WebART project organized and participated in a series of events in 2012 and 2013 (see Table 2.1). These events

[14] `http://www.nwo.nl/catch` (accessed: 01/08/16)
[15] 'Web Archive Retrieval Tools', `http://www.webarchiving.nl/` (accessed: 01/08/16)
[16] `http://www.uva.nl` (accessed: 01/08/16)
[17] `http://www.cwi.nl/` (accessed: 01/08/16)
[18] `https://www.kb.nl/` (accessed: 01/08/16)
[19] `https://www.kb.nl/webarchief/` (accessed: 01/08/16)
[20] `http://archive-access.sourceforge.net/projects/wayback/` (accessed: 01/08/16)

Table 2.1: WebART events and event participation.

Event	*Date*	
(1) DMI Summer School	08/12	Participation in the Digital Methods (DMI) Summer School, developing research scenarios for the Dutch Web archive
(2) Web Archiving: Theorized Practices	12/12	Organization of an ACHI seminar involving renowned scholars using the Web as a corpus
(3) DMI Winter School	01/13	Participation in DMI Winter School, developing research scenarios for the Dutch Web archive
(4) WebART CATCH Event	04/13	Symposium on Web archives with speakers from the British Library, Library of Congress and the University of Amsterdam
(5) Exploring Israeli Politics Online	05/13	Workshop at Bar-Ilan University, Israel, aimed at analyzing political Web archive data
(6) DMI Web Archiving Day	09/13	Workshop and focus group, evaluating all WebARTist tools up to that point
(7) New Media Research Masters	11/13	Seminar for new media Master students, creating proposals for research using the Dutch Web archive

shed more light on the needs of researchers that use web data to perform their research, and that take the web as their object of study. The participants in the events ranged from Master students to PhD-level researchers and renowned senior scholars, reflecting a wide range of potential web archive users and use cases. Many of the participating new media scholars were affiliated with the Digital Methods Initiative (DMI), which is "one of Europe's leading Internet Studies research groups"[21]. It "designs methods and tools for repurposing online devices and platforms (such as Twitter, Facebook and Google) for research into social and political issues". As indicated by Rogers (2013), the term 'Digital Methods' is a counterpoint to 'Virtual Methods'. Virtual methods may translate traditional social science methods, such as questionnaires, and "migrate them onto the web". Digital methods, on the other hand, study the unique characteristics of the web, in particular 'natively digital objects'. The internet, actually, "may be rethought as a source of data about society and culture", which requires new methods (Rogers, 2013, p.38). Research foci of digital methods (Rogers, 2013, p.27) include the *hyperlink*, the *website*, *the search engine*, the *spheres*, the *web(s)* as well as web *platforms*. To facilitate these types of research, the Digital Methods Initiative provides a comprehensive set of tools to study the Web (seventy at the time of writing), including tools for data extraction, scraping, processing, analysis and visualization[22]. Some of these tools can also be used in conjunction with Web archives.

In the WebART project, a set of web archive retrieval tools was designed to accommodate for web researchers' needs. This toolset, named *WebARTist*[23],

[21] `https://digitalmethods.net/` (accessed: 01/08/16)
[22] `https://tools.digitalmethods.net/` (accessed: 01/08/16)
[23] WebARTist standing for 'Web ARchive Temporal Information Search Tools'

was designed during four phases[24] utilizing an action research methodology.

2.3.2 Setup and Methodology

As indicated by Pickard (2007), the action research method involves placing the researcher within the research process, and it "envisages a collaborative approach to investigation, that seeks to engage 'subjects' as equal and full participants in the research process (Stringer (1996), as cited by Pickard (2007, p.134)). Using the action research method, the researcher engages in "examining current processes, taking action to improve those processes, then analysing the results of the action" (Pickard, 2007, p.134). In the context of this method, the researcher is "an active participant in the process before, during and after the research activity" (p.134). The implementation within a research context also means that "the form of the solution is completely dependent on the context and the nature of the problem" (Pickard, 2007, p.136).

Action research is a cyclical process, which includes the following stages: *problem identification*, *action planning*, *implementation*, *evaluation* and *reflection* (Pickard, 2007, p.134). In the case of WebART, the full action research cycle occurred between April, 2012 and September, 2013. In action research, the problem identification phase starts by the identification of an issue and the "establishment of the current context from the perspective of those who will be the target of the intervention" (Pickard, 2007, p.135). The next section describes this, together with the action planning of the proposed intervention (2.3.3). Second (2.3.4), we discuss the implementation of the intervention within the research context (the creation of web archive search tools). Then, the success of the intervention is evaluated (2.3.5), done via a workshop with researchers followed by a focus group and survey. Finally, an essential part of action research is reflection (described in Section 2.3.6), here involving a higher-level analysis of the utility of the introduced search-based access tools.

2.3.3 Identifying Problems and Action Planning

In the initial phase in Summer 2012, *problem identification and action planning*, the main issues in web archive research and web archive access were explored via a literature review and by active participation in a summer school. The WebART team participated in the Digital Methods Initiative's (DMI) Summer School (Table 2.1 [1]), a yearly summer school in which motivated scholars "learn and develop research techniques for studying societal conditions and cultural change with the Internet"[25]. In particular, the selection policies and content

[24] Part of this process has been described in: Huurdeman et al. (2013); Ben-David and Huurdeman (2014); Huurdeman (2015);

[25] `https://wiki.digitalmethods.net/Dmi/DmiSummerSchool/` (accessed: 01/08/16)

of the web archive of the Dutch KB were explored, as well as the possibilities of doing research using existing web archive access tools, such as the Internet Archive's Wayback Machine. Gained insights could be used in later development of solutions for improving web archive access.

Thus, the WebART team members collaborated with participants of the summer school in a project-based setting to develop research scenarios using web archives in general, and the Dutch web archive in particular. In the first week of the summer school, a project team with five team members analyzed trackers in the Internet Archive, i.e. objects embedded in the source code of webpages which can track user behavior, often for advertising purposes. The aim of this project was to assess if it is possible to "map website ecologies around websites through invisible back-end linking"[26] [27]. By analyzing the use of trackers of the front page of the New York Times over time, the summer school project showed the feasibility of this research in combination with the Internet Archive. This project also illustrates the wide range of research questions which can be asked to web archives, as well as the importance of looking beyond solely content-based access approaches, towards analysis of the underlying web structure.

In the second week of the summer school, project activities consisted of analyzing the KB's selection lists. A comparison of the websites included in the Dutch web archive and the Internet Archive's showed that the majority of content in the Dutch archive could also be found in the Internet Archive[28]. Also, the popularity of websites (hosts) in the Dutch web archive was explored, by retrieving the number of Facebook likes and Twitter followers for the websites in the archive, as well as for the top 200 sites in the Netherlands according to Alexa[29]. In addition, network analysis of different categories of websites in the Dutch web archives selection list was performed using the 'Issuecrawler'[30] and the DMI tools for analyzing the Internet Archive. Hence, similar to the previous summer school project, there was a focus on the underlying structure of webpages (the hyperlinks). In addition, cross-comparisons were performed, showing the value of additional sources to contextualize the selection lists of the Dutch web archive (such as social media data and the Internet Archive). Ultimately, the projects conducted during the summer school resulted in an initial understanding of the KB's web archive, its selection policies and the

[26] Documented at: `https://wiki.digitalmethods.net/Dmi/TracingTheTrackers` (accessed: 01/08/16)

[27] This built on previous work in the 2012 DMI Winter School looking at tracking ecologies using tools based on 'tracker' fingerprints from Ghostery in the 'live' web. In the summer school project, these methods were applied to the Internet Archive, enabled by combining DMI tools with the Internet Archive's Wayback Machine. The research has later been revisited and extended, see Helmond (2015, p.124).

[28] Albeit this analysis was only performed at the homepage level (as opposed to deeper pages)

[29] `http://www.alexa.com/topsites/countries/NL` (accessed: 01/08/16)

[30] `http://issuecrawler.net` (accessed: 01/08/16)

potential research opportunities of web archives in general.

During the explorations in this phase, also a number of limitations of available access tools were confirmed. In particular, the document-centric and 'single site' approach to web archive research of the Wayback Machine interface posed problems, impeding analysis beyond the page level (Hockx-Yu, 2014; Rogers, 2013; Ben-David and Huurdeman, 2014). Without resorting to external tools, the interface of the Wayback Machine predominantly facilitates qualitative inspections of web archive content, as opposed to analyzing broader patterns and underlying structure. Even though specific tools allowed researchers to analyze multiple URLs[31], and the source code of pages[32], previous knowledge of URLs of online resources was still required. As web archives contain pages from the web of the past, the consequence is that a substantial amount of resources cannot be located. In order to support scholarly use of web archives within WebART, the natural next step was plan the development of search-based access tools, allowing researchers to dynamically search content in the web archive.

2.3.4 Implementation

In the next *implementation* phase, a full-text search system for the Dutch web archive was introduced, potentially offering additional support to new media scholars in their research process. To design the system, the thesis author collaborated with a new media researcher and a computer science researcher on a day-to-day basis. Additionally, further insights were gained via three workshops performed with other new media researchers (Table 2.1 [3,4,5]), the first of which we describe in detail below. Hence, the actual functionality of the search tools (both back-end and front-end) was developed in a bottom-up way (Huurdeman et al., 2013), meaning that researchers in- and outside the WebART project were consulted for building the system's functionality.

Development of Search-based Access The data of the Dutch web archive is stored in the ARC-format, which aggregates web resources and their metadata, as well as crawl-related metadata[33]. After experimentation with different information retrieval solutions, the Terrier Information Retrieval platform (Ounis et al., 2006) was used to create a search environment. Terrier is a highly scalable and customizable open-source search engine written in the programming language Java. Due to the sheer size of the KB's dataset, amounting to over 7 Terabytes at the time, the extraction and indexing had to be carried out via

[31] The DMI 'Internet Archive Wayback Machine Link Ripper' and 'Network Per Year' tools, at `https://tools.digitalmethods.net/` (accessed: 01/08/16)

[32] E.g. tracker fingerprints: `https://tools.digitalmethods.net/beta/trackerTracker/` (accessed: 01/08/16)

[33] `http://archive.org/web/researcher/ArcFileFormat.php` (accessed: 01/08/16)

a Hadoop computer cluster at SURF's Dutch national e-infrastructure.[34] One of the advantages of the Terrier IR platform was that it functioned with this computer cluster.

To ease pilot testing of full-text search approaches, subsets of the Dutch web archive were extracted from the Dutch web archive and CommonCrawl collection of web crawls in the Fall of 2012. CommonCrawl offers "web crawl data that can be accessed and analyzed by anyone"[35]. Initially, based on researchers' requests, 80 crawls from the Dutch news website *nu.nl*[36] were extracted and indexed, later extended to all crawls of this website in 2011 and 2012. This news website was chosen since it was archived most frequently of all sites in the KB archive (daily), and because it provided ample opportunities for news-related analysis, as it was the most popular Dutch news aggregator[37].

The initial featureset of *WebARTist* (see Figure 2.2) included basic capabilities offered by the Terrier retrieval platform, including free-text search and possibilities for query term suggestions (using 'query expansion'). To index the ARC files of the Dutch KB, an extra input parser for Terrier's indexing module was created. Furthermore, Terrier's existing functionality was extended with temporal filters, including filters for the timestamps of pages (the last-modified date) and crawl timestamps (the date a page was harvested). Documentation describing the functionality and querying mechanisms was provided via an online Google Document. Finally, as Terrier is aimed at information retrieval experiments, its standard interface has limited functionality, so we created a custom JSP full-text search interface, where each result list item contained a title, snippet and basic metadata[38]. Term suggestions derived from Terrier's query expansion feature were visualized as a word cloud.

Usage and Extension of Search-based Access To initially explore the utility of the implemented functionality, and to determine which additional features would be needed for research use, the WebART team participated in the DMI Winter School in January 2013 (see Table 2.1 [3]). The theme of this four-day winter school was "Data Sprint: The New Logistics of Short-form Method." PhD candidates, advanced MA students and motivated scholars could participate in the winter school. We organized one of the projects in the associated three-day workshop. The project group consisted of five new media researchers, one computer scientist and one information scientist. The participating researchers were given the opportunity to use the developed full-text search tools

---

34 `http://www.surf.nl/` (accessed: 01/08/16)

35 `http://commoncrawl.org/` (accessed: 01/08/16)

36 `http://www.nu.nl/` (accessed: 01/08/16)

37 According to Alexa it was the 6th most popular Dutch site in 2011: `http://web.archive.org/web/20110923151640/http://www.alexa.com/topsites/countries/NL`

38 The item's *url*, *collection* (KB or CommonCrawl), *the domain*, *content size*, *crawl date* and *page* (last-modified) *date*). By clicking this metadata, a filter could be added to the query.

Figure 2.2: Initial WebARTist search interface prototype (exemplifying the use of 'query expansion').

in combination with the extracted dataset containing *nu.nl*.

For the winter school, all nu.nl harvests in the KB archive and in the CommonCrawl open web crawl were combined. This led to a dataset of 13.64GB, of which all textual HTML content was indexed (3.48GB). The combined dataset contained 64.624 archived pages, harvested on 412 dates for the KB set (Sept. 2011-Dec. 2012) and 20 dates for the CommonCrawl set (Jan.-May 2012).

The full-text search tools offered to the participants allowed for various types of analysis. Five specific *research scenarios* in the context of a searchable news archive were explored during the winter school[39]. First of all, 'temporal analyses of item frequencies' were performed. This entailed analyzing the frequency of news items covering controversial country leaders, taking queries such as 'Mubarak' and 'Assad' as a starting point. At the search system level, this necessitated the addition of a feature to export search results in a structured format, including their timestamps. Second, 'temporal co-word analyses' were performed. Taking the query 'Assad' as a basis, participants analyzed which words frequently co-occured with Assad in the full-text pages returned by the search engine. Again, this evidenced a need for structured exports of resultset items. Third, 'outlink extraction analysis' was done, looking at the frequency of links from the news pages to other sources (for instance in the context of the news coverage about hurricane Sandy and the US 2012 presidential elections). At the system level, this necessitated a feature to obtain lists of outlinks (i.e. inter-domain links) occurring on the pages returned by the WebARTist search

[39] For a more extensive description of the research explorations performed in the Winter School, please refer to Huurdeman et al. (2013)

engine. The fourth explored research scenario was 'geomapping of news wire reporting cities' , looking at the places from which wire reporting services reported news about Syria. To support this research, the search system needed to support indexing of specific HTML fields, and to return them in a structured format. The fifth and final research scenario focused on at 'temporal image analysis'. This scenario looked at the analysis of images used to illustrate news articles over time. To address this scenario, an image search facility had to be added to the WebARTist search engine, as well as a structured export facility for image materials. This allowed for creating visual timelines about issues in the news.

As indicated in Huurdeman et al. (2013), the searchable web archive opened up "new ways of exploring, analyzing and visualizing archived material, in ways that are currently not possible with the single-site approach offered by the URL-search of existing web archives". The participation in the winter school resulted in an overview of the *search features* needed for meeting research questions related to web archive (or historiographical) news analysis research. For instance, the temporal nature of the research explorations resulted in a need for additional query filters, including date ranges. As illustrated by the research scenarios above, the most urgent prospective functionality was a feature to export resultsets. Researchers wished to export results into a structured format, which could be imported in their own analysis and visualization tools (e.g. Excel, Google Fusion Tables, or Gephi). Hence, they wanted to perform their analyses outside of the system. To this end, we added 'export resultset' features for the search result items (including the main paragraph text of each found page), for external links occurring on found pages, for associated images, and for query expansion terms.

The requests for enhanced search features had several implications for the indexed *data* and *metadata*. For instance, to allow for hyperlink analysis, considered "an important way to reconstruct views of the past" (Huurdeman et al., 2013), various types of hyperlinks had to be extracted from each page in the set of archived data. These included site-internal links (within a domain), external links (to other domains), script links, and links to images. Subsequently, this data had to be added to the existing metadata of each archived page, and indexed by the WebARTist search engine.

The consultation and collaboration with researchers in the different workshop settings (Table 2.1 [3,4,5]) ultimately resulted in an overhauled search system, which not only allowed for possibilities to *export* data, but also for partial exploratory data analysis *within* the system. Next, we describe the evaluation of that system.

2.3.5 Evaluation

In September 2013, an *evaluation* phase followed, in which the collaboratively developed search tools were assessed. The 'Web Archiving Day', organized at the Department of Media Studies at the University of Amsterdam, included a morning session with presentations by researchers, and an afternoon session with a workshop for researchers (involving all search tools in the WebARTist toolset), a survey and a focus group evaluation meeting. The purpose of the evaluation was both to assess the developed systems in the WebART project, and to "address the theoretical and methodological implications of searching, mining and visualizing the archived web."[40] In total, ten researchers in the field of new media research participated in the morning session (Table 2.1 [6]). The subsequent workshop and focus group was attended by seven researchers.

System and Participants The researchers could use three search systems, each reflecting a different dataset (see Table 2.2). Documentation describing the functionality and querying mechanisms was provided via an online 'help' function. The first available search system was a search system with an underlying index of the full web archive of the KB (43.5M documents, 2009-2012). This system, based on the search system of the previous iteration, additionally provided filters based on classification codes, duplicate detection, visual statistics, site summaries and results reordering. The second available interface was the *host+1* system (253,649 documents, 2009-2012), which included all features of the full index version, as well as aggregated results (summaries of the whole resultset) and location data. The dataset consisted of all homepages (defined as entry pages of hosts), plus all linked pages up to one level deep. Third, participants could use an updated version of the *nu.nl* interface (57,913 documents, 2011-2012)[41]. Hence, there was an interface containing all data, one containing an enriched subset and one containing solely one site.

Seven researchers attended the workshop and the focus group following the workshop, and completed the survey. The researchers' backgrounds involved new media and digital culture, but a focus on distinct areas. For instance, one researcher researched born-digital archives and how they relate to social memory practices. Another researcher studied methods and tools to study web data for the purposes of social and cultural research. Other participating researchers focused on web culture, the history of connection features of the web, and on

---

[40] `http://www.webarchiving.nl/news/webarchive-search-as-research/` (accessed: 01/08/16)

[41] This includes the additional features added in the the full and host+1 interfaces which were applicable to this dataset. As there was only one site in this interface, no site selection features were necessary.

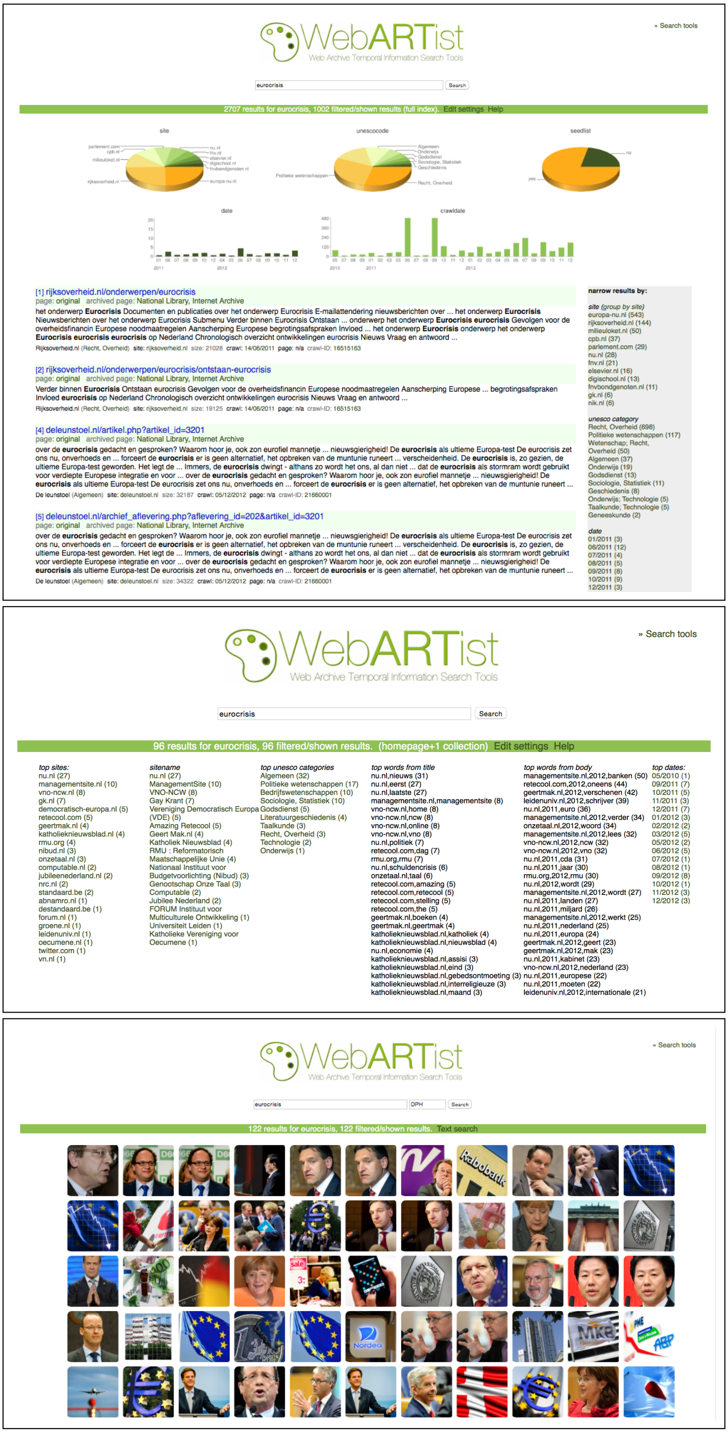

Figure 2.3: Second WebARTist search interface prototype. *Top image:* full index, item view. *Middle:* host+1 index, aggregated view. *Bottom:* nu.nl index, image search view.

Table 2.2: WebARTist features

Feature	*Full index*	*host+1 index*	*nu.nl index*
Unesco classification	X	X	
Results customization	X	X	X
Duplicate detection	X	X	X
Query expansion			X
Statistics	X	X	X
Site Summaries	X	X	
Order results by date	X	X	X
Image search			X
Location data		X	
Aggregated results		X	
Export results	X	X	X
Export outlinks (site/all)	X	X	X

issue mapping[42]. The seven participating researchers included PhD candidates, an associate professor and a full professor. All of these researchers had used web archives for research in the past.

Workshop Findings The purpose of the workshop was to explore the possibilities of WebARTist. During these explorations, the researchers indicated various topics of interest: first of all, in the context of web history, it allowed them to "conjure up past states of the web" using daterange queries, and to "derive periodizations of the web". For instance, a mentioned research topic was the 'rise of social media on the web', which could for instance be studied in WebARTist via the archive's exposed link structure. Other explored possibilities were to create "source hierarchies", i.e. looking at the "most dominant and prominent sources in the archive". For instance, WebARTist would allow a researcher to query for *financial crisis* and find out which are the "top issues of the day" for this topic in the archive. This could be potentially compared with historic or contemporary queries using Google in the Dutch region. Third, researchers could look at the "keyword uptake", the occurrence of competing keywords in the archive over time, also aided by the aggregation and visualization possibilities of the tools. For instance, a researcher may search for *climate change* and assess how various categories of websites (e.g. news or government websites) evolve over time, including the language used on these websites. Another suggested idea was "memory competition", looking at how issues that happened on the same day influence each other, in terms of visible results in the interface; i.e. "does one issue render the other invisible."

A selection-based archive like the Dutch web archive is by nature incomplete.

[42] Issue mapping is "specifically concerned with the application of online techniques, methods and content for the analysis of current affairs," see: `http://www.issuemapping.net/` (accessed: 01/08/16)

However, a feature in WebARTist showed whether a retrieved webpage occurs on the original *seedlist* or not. Using this feature, it was possible to study a phenomenon labeled as "accidental" or "incidental" archiving by the researchers: the occurrence of certain, unselected, sites in the web archive[43].

Hence, the workshop provided insights into a multitude of potential research questions that can be researched using a searchable web archive. These questions focused on the archive itself and its policies ("What is the Dutch approach to digital heritage in this web archive?"), as well as the content of the archive ("How is topic x represented in the archive over time?" or "The uptake of specific terminologies and competing terminologies."). In addition, various remarks of researchers during the workshop provided indications for improvements of web archive access. These remarks embodied different perspectives: the *data* perspective (e.g. limitations in the data due to restrictive selection policies and the fact that the indexed dataset only started in 2009), the *system* perspective (e.g. the maximum number of 10,000 results that could be exported at once), and the *interface* perspective (e.g. not being able to compare two n-grams in one graph). Further investigations into the utility of search-based access followed in a dedicated focus group session, described next.

Focus Group and Survey Findings To evaluate search-based access, a focus group was organized, which "allows a variety of perspectives and explanations [to] be obtained from a single data-gathering session" (Gorman and Clayton (2005), as cited by Pickard (2007, p.219)). In this group of researchers, there was a positive response to the usefulness of the different WebARTist systems. After trying out the possibilities of the toolset and exploring research scenarios, participants of the focus group indicated that "special collections are useful now", by being able to search through them. This reflected "the shift of studying a web archive through queries"; a big step forward as compared to earlier URL-based Wayback Machine interfaces.

In particular, the WebARTist system supported "looking at data rather than single sites", and also to be reflexive about the collection policies. One participant mentioned the facilitation of a move towards studying "data objects" rather than "content". Furthermore, the "aggregate views and bar graphs" were found "extremely useful" by researchers, as they allowed them to view a summary of thousands of results in one page. They make it possible to "reveal underlying structures/patterns within collections", as well as to "explore the archive's statistics and the 'macro' level of analysis." One researcher indicated that this should be the standard view of the interface. Hence, we saw a need for viewing the big picture, i.e. 'distant reading' (Moretti, 2013), reflected in the use of aggregations and visualizations by researchers (e.g. to analyze the

[43] The occurrence of unselected content in the Dutch web archive is further researched in Chapter 3.

metadata and outlinks for the full resultset for certain queries and timeframe). On the other hand, researchers still inspected individual results more closely, for instance to determine if they were relevant. Hence, support for 'close reading'[44], i.e. studying the webpages and websites involved, is still important[45].

A topic of discussion which also came up in previous workshops were the results returned by the web archive search system. Even though search engines for web archives and online search engines may have similarities at the interface level, the listed search results can be quite different[46]. For instance, one researcher searching for pages in the web archive about surveillance remarked that (s)he expected "to find [other] things". The difference in retrieved results coupled with the similarities at the interface level led to a discussion about the differences between a web archive search engine and a live web search engine such as Google. For the majority of participating researchers, the differences between a web archive search engine and a web search engine were clear, even though there is a risk to be "conditioned to thinking about Google style search." To make the difference between the web and the archived web more obvious for prospective users, it was suggested to show the seedlist and assigned categories in the WebARTist interface. Thus, it could be made clear "to the users how the archive was built", and "that it is not Google". According to the researchers, the UNESCO codes, the topical classification categories assigned to each website in the archive, are important as a way of contextualizing the KB's collection. As a researcher remarked, "the codes shows how the institution conceptualizes digital culture heritage. It is important to show them because they make [these conceptualizations] transparent." Another participant remarked that "seeing the categories helps for method rather than for research. It helps understanding why it was curated. That is, if we study the archived process rather than the archive itself." Hence, this type of metadata can serve as search filters (as evidenced in the WebARTist interface), but also as contextualization to the archiving process.

With respect to the included functionality in a web archive access system, the participants indicated in the survey following the workshop that they would like to perform analyses both *within* and *outside* the system. Researchers indicated an interest in having built-in analysis functionality in the search system, for instance facilitating exploratory analysis, but also to have the ability to take the data out of the system and analyse it in their own tools. Hence, some extensions

44 Originally a term from literary criticism

45 Arguably, due to the limitations of web archives as research datasets discussed in the previous section, distant reading methods may not always be successful; see also: `http://www.nytimes.com/2011/06/26/books/review/the-mechanic-muse-what-is-distant-reading.html` (accessed: 01/08/16)

46 The reasons behind this are multifold, and include differences between in sophistication of ranking algorithms of web archives and online search engines, the inclusion of multiple versions of the same results in web archives, and so forth.

of built-in functionality would be helpful, i.e. incorporating "some additional analysis and visualization options". For instance, a researcher indicated in the survey that (s)he "would like statistics for the full dataset: top sites within [the] dataset, top outlinks within [the] dataset." A prospectively useful feature would also entail a "competing keywords" feature, i.e. the ability to directly compare the results of two keywords: "the single line graph of the n-gram viewer is fun, but double or more graphs would be better".

Finally, a need for contextualization emerged from the survey. Most researchers indicated in the post-survey that they would use WebARTist for their research, but that they would combine 'live' web and archived web data. One participant stated in the survey that the search engine "made it possible to build new research questions that go beyond the web site history approach." In addition, it also "offers hope that web archiving is evolving in a more creative field of research."

2.3.6 Reflection

The last step was the *reflection* phase, in which we performed a higher-level analysis of the utility of the introduced search tools. The findings from the evaluation may influence future versions of WebARTist, but also search engines for web archives at large. The focus group pointed out that the experimental prototypes, however useful, also needed more functionality to be usable for research. On a broad level, the researchers requested additional support for *selections*, *analysis*, *collection making*, and more *transparency*.

Our first conclusion is that improved *selection methods* are needed for a searchable web archive[47]. One scholar argued that the limitation of the Wayback Machine is that one always has to start with a URL, while the limitation of WebARTist was that the starting point has to be a query. Suggested expansion included the possibility to start with selecting a site or category of sites, or using lists including "historical web directories, blogrolls and link lists". Also, web archives generally include a massive amount of content. Hence, sampling (using a subset of the large amount of data for initial analysis), was deemed useful for future versions of WebARTist. As a researcher put it, it should be possible to "first view a sample to get a sense of what's in, and if it's worth pursuing, get the full data".

Second, further *analysis* functionality may be needed. In particular, a comparison feature was judged as useful, to directly compare differences in resultsets without manually opening multiple browser windows. In particular, a feature to compute and visualize the differences between result sets was suggested as well as statistics for the whole dataset.

[47] In concordance with researchers participating in an earlier workshop in Israel (Table 2.1 [5])

Third, the possibilities to create custom *collections* and to add annotations are useful. Hence, when a scholar is building a dataset for her research, it should be possible to store it, and to add annotations during the process. These collections may also have appeal for other users: as a researcher remarked, it should be possible to "make the collection you build accessible and annotate it for other users".

Finally, *transparency* is a key issue to be addressed in future systems. Researchers would like to know more about the archive's selection procedures, the (in)completeness of the archive and algorithms. Clearly, an archive's selection policies and the completeness of harvests have a direct influence on the items which are retrieved, leading a researcher to argue that "data is still a crucial factor". In addition to that, ranking and retrieval algorithms in access interfaces may excert a profound influence on the degree of relevant items appearing in results lists[48]. Evidently, these issues could also influence subsequent analysis of datasets derived from the archive. This suggests a need for more contextualization and transparence in future web archive search systems (a topic further investigated in chapter 3). Now, we discuss the implications of our findings so far.

2.3.7 Discussion

Summarizing, we have iteratively refined a toolset for search-based Web archive access in the context of the WebART project. Our first conclusion from the action research setting suggests that more support for various activities in the research process is needed (including selections, analysis and collection making), a topic which we will cover in the next section. Second, there is a need for tools which provide additional transparency indicators of the impact of various variables on the results returned by a search system.

In relation to the issue of **transparency**, Figure 2.4 summarizes the factors affecting the completeness of encountered results. In a "macro" archiving context, various influences exist on the eventual results a researcher retrieves via a search-based web archive access interface. Curators may select web content to include in the archive (such as a list of websites, or a national domain), listed in a *seedlist (a)*. These items are subsequently harvested by a web crawler, using *crawling settings (b)* determined by a crawl engineer. This leads to a set of archived data. However, due to the curatorial decisions, crawler limitations, but also storage limitations, a substantial amount of data may be **unarchived**. The

[48] To take a practical example, in the DMI Winter School (Table 2.1[1]), researchers studied news coverage about the former Egyptian president, Hosni Mubarak. Initial indexing settings resulted in high *recall*, but low *precision* of retrieved items when querying for "Mubarak", impeding temporal analysis of the found articles. Customized indexing settings for the specific news website in question, ignoring the text which surrounds articles on a webpage (e.g. links to other articles and site navigation links) resolved this issue.

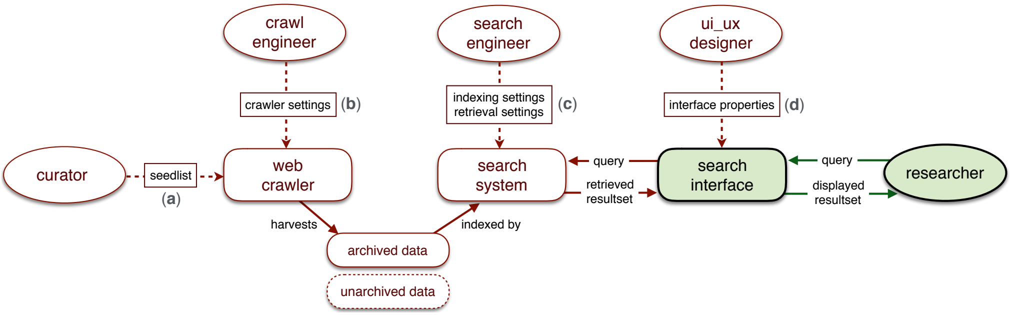

Figure 2.4: Actors and interactions in a "macro" web archiving context. The directly visible elements are indicated in green, while the 'hidden' influences are indicated in red.

data which was archived is subsequently indexed by a search system, for which a search engineer may specify specific *indexing and retrieval settings (c)*. These settings also excert a profound influence. For instance, there is a discrepancy between system (algorithmic) relevance and topical (subject) relevance. Researchers may wish to retrieve documents related to a topic, while the syntactic matching of query words with words occurring in documents by IR algorithms may result in items being erroneously retrieved, or not retrieved at all. System designers may have to find a tradeoff between precision, the degree of retrieved items which are relevant, and recall, the fraction of relevant items retrieved. At the final end of the depicted pipeline, a researcher needs access to this archived data. In most cases, this access is mediated through an access interface, which *interface properties (d)* may have been influenced by a UI designer. This may be a list of 'ten blue links'. However, this condensed list does not provide contextualization about the variables embedded in a "macro" archiving context unbeknownst to a researcher. More advanced interface features, for instance summaries and aggregations, may further aggravate this issue. Distant reading of summaries of resultsets with low precision, without manual checks of individual result items, may result in inaccurate impressions[49].

Hence, a researcher may actually miss a substantial amount of data due to crawling scope and settings, indexing and retrieval parameters and interface issues, potentially be crucial for analysis. This connects to a point previously raised by Bates (1990): "the need for a user to know what has been done for him or her, even when the results are satisfactory, has often been ignored or underestimated by human intermediaries and IR system designers alike." While a simple lookup search in a search engine may need little contextualization, we

[49] An anecdotal example was a query for (then) princess 'Maxima', which also returned a lot of weather forecast pages (with 'maxima' referring to maximum temperatures). When comparing the occurrences of the term in a temporal collection of web pages, e.g. via a timeline chart, one may not immediately be aware of the included false positives. Thus, it is important to make it possible to inspect the underlying result items.

strongly argue for the provision of deeper insights into the variety of underlying variables in research settings.

To improve upon this issue, institutions could make their selection policies, criteria and selected websites available at an institutional level. For instance, the Dutch KB provides this information via its website[50]. Besides this issue, web crawling involves various technical issues. These impede some websites from being harvested correctly. Usually, though, documentation about crawler settings and their influence on obtained data is not provided. Hence, for enabling a better understanding of archived content, this documentation should be made available to researchers as well, preferably in a form understandable for non-technical experts. Finally, standard full-text search systems and interfaces could show the particularities of underlying data, as well as ranking and retrieval properties to their users. A particular design challenge which may be addressed in future work is *how* the various variables could be integrated and visualized in the search interface, but also in which way the influence of these variables can be explained in an understandable and meaningful way.

Hence, at the moment, search systems typically hide which data is *missing* from the archive. By using the information "hidden" in web archives (Rauber et al., 2002), we may contextualize Web archive search. This could for instance be done by providing information about the material which is included and excluded from the archive at retrieval time. Search results in the archive can be combined with found representations of unarchived search results, thereby increasing *transparence* of web archive search tools. In chapter 3, we investigate approaches to potentially achieve this, utilizing the link structure and anchor text in the web archive.

The participants in the focus group evaluation of WebARTist argued that the tool could also be useful for researchers in other disciplines which involve primary source research[51]. Up to this point, however, our evaluation mainly took place in a new media context; in particular within the Digital Methods Initiative at the University of Amsterdam. However, a point of attention, also raised by one researcher participating in the focus group, is whether or not current search tools are "sophisticated enough" for researchers in other areas, and researchers who are not used to working with digital tools. Furthermore, in relation to the need for more research support mentioned in the beginning of this discussion, there is a need for a better understanding of the methods scholars use in their research process. Therefore, we extend the scope of our research in the next section to media and communication studies, and perform an analytical literature

[50] In Dutch, via `https://www.kb.nl/webarchief/` (accessed: 01/08/16)

[51] In the post-survey, the majority of the participants indicated that WebARTist would be useful for Digital Humanities, Internet Studies, Journalism, and the Social Sciences. Half of the participants perceived it to be useful for the Humanities and Information Science

review of previous (and completed) web research. This way, we may arrive at better insights into the limitations of current tools in the context of the research process.

2.4 Scholars' Use of Web Data - Corpus Definition, Analysis & Dissemination

2.4.1 Introduction to Web Data Use

The abundance of content available in web archives has seen limited research use thus far. As follows from the review of previous literature (Section 2.2), several factors influence this *non-use* of web archives, including limitations in data quality and quantity, legal issues and the limits of access interfaces. The previous section (2.3) has evidenced the potential value of moving from URL-based to search-based access. While facilitating a whole new range of research opportunities, search-based web archive access also introduces various limitations, including an increased lack of transparency. We have outlined some of these challenges in a new media research context. In this section, we move to a broader context of media and communication studies.

The setup of the previous section inherently emphasized insights into exploratory research using the Dutch web archive, as most researchers worked with the search tools in a workshop setting. However, there is a need to assess the steps that are taken by researchers between exploration and finished research, and how archive access potentially fits in. This led to the following research question: (**RQ1.2**): *Which corpus creation, analysis and dissemination methods do media and communication scholars use in the context of web data? What are the implications for search-based web archive access tools?*

In this section, we employ a bottom-up approach and review how media and communication scholars use web data in their research process. We chose this approach due to the lack of current use of web archives in a research setting, evidenced in section 2.2. A consequence of this lack of use is that traditional information behavior approaches, closely studying actors which regularly use web archives for their research, would be infeasible in a media and communication context. Based on our literature analysis, we reveal the implications for current search systems for web archives. Subsequently, we offer a tentative outlook on more fluid approaches to overcome these limitations, moving from relatively shallow search engines to a broader 'research engines', potentially improving support for the research process.

2.4.2 Setup and Methodology

This section contains an analytical literature review, which we define in this thesis as a meta-analysis of self-reported research patterns by scholars in their publications. To shed more light on some defining elements of scholars' research tasks, we look at a set of journal papers, in particular in media and communication studies. Due to the previously referenced lack of research use of web archives so far, few papers report the use of web archives as a data source. Therefore, we do not only look at papers using web archive data, but also at papers making use of (temporal) *web collections*, which we define as collections containing longitudinal web data. These may have been personally harvested by the researcher, or may be part of an existing collection of web data gathered for research purposes. We systematically gathered a set of eighteen journal papers which use temporal web data. Specifically, we queried EBSCOhost in April 2015[52] using the databases Communication and Mass Media Complete (CMMC)[53], and Library, Information Science & Technology Abstracts (LISTA)[54]. We generated a list of the most frequently occurring words occurring in the abstracts of a literature review for a previous paper (Ben-David and Huurdeman, 2014). This resulted in the following structured query, issued in EBSCOhost: *AB (((web) OR (research) OR (links) OR (information) OR (sites) OR (evolution) OR (content) OR (changes) OR (events) OR ("over time") OR (longitudinal)))*. From the first 150 results of these query in each database, we selected each journal paper which uses websites, webpages or elements of webpages in their research questions and analysis. Furthermore, we excluded computer science papers, since our aim was to review papers related to the fields of humanities and social sciences. This resulted in 17 journal papers from the CMMC database, and 17 papers from the LISTA database. Our final step consisted of removing the duplicates between databases, and retaining only the papers which perform longitudinal analyses using the web data (i.e. which analyzed multiple points in time), in the period 2007-2015. The final literature set consisted of 18 papers by 17 distinct authors.

The selection of literature was published in 16 distinct journals: *Communication & Society; Information Communication & Society (2 papers); Information and Organization; Internet Research; Journal of Academic Librarianship; Journal of the American Society for Information Science and Technology; Journal of Computer-Mediated Communication; Journal of Documentation; Journal of Information Technology and Politics; Journal of Public Policy and Marketing;*

---

[52] EBSCOhost provides a range of research databases, which index academic papers; `http://www.ebscohost.com` (retrieved 10 april, 2015)

[53] `https://www.ebscohost.com/academic/communication-mass-media-complete` (accessed: 01/08/16)

[54] `https://www.ebscohost.com/academic/library-information-science-and-technology-abstracts` (accessed: 01/08/16)

Journal of Science Communication; Journalism Studies; Library Hi-Tech; New Media & Society; Online Information Review (2 papers); and Scientometrics. Hence, the main areas covered in this literature set were Information Science, Communication, Media Studies and Political Science.

For our set of literature, we chose to select indexed journal papers, leaving out published book chapters and 'grey literature'. An advantage of assessing these journal papers is that their journals constitute a thoroughly validated venue for research. Moreover, the author guidelines of journals often necessitate clear descriptions of used data sources, selection methods and research methodology, which are the main focus of this chapter. The inclusion of published journal papers may result in a focus on existing (as opposed to more novel and digital) research methods. This could actually complement the research explorations of our previous study.

2.4.3 Background Research: Models of the Research Process

Various models of the research process distinguish different stages which occur in research tasks. Case (2012, p.205) lists a 'classic' view on these stages, which typically start with *imagining a research question (1)*, followed by *determining what data are needed and designing a specific study to collect it (2)*, *choosing and implementing research methods (3)*, *analyzing and interpreting observations (4)*, and *considering the overall results (5)*. The exact nature of these stages, however, is usually not as straightforward, and may vary across disciplines. Researchers may take inductive approaches, examining instances and "reason toward generalization", or take deductive approaches, thus reasoning "from the general to the particular" (Case, 2012, p.206). The research process in different research areas has been documented in a large number of models. First of all, 'prescriptive' models of the research process may guide the process, and are for instance included in textbooks, such as Kendall (2012). These models provide an essentially 'idealized' view of the process, which may deviate from practice to a certain extent. Second, a number of empirically-based models exist, derived from direct studies of scholars' research process; for instance, Chu (1999). These models may provide a more realistic (and less linear view) of scholars' research process.

A model by Kendall (2012, p.27) embodies an example of a prescriptive approach to modeling research phases in the social sciences. She explicates the classic deductive method, which involves *defining the research problem (1)*, *reviewing literature (2)*, *hypothesis formulation (3)*, *research design (4)*, *collecting and analyzing data (5)* and *drawing conclusions and reporting findings (6)*. The inductive method, on the other hand, involves a different order of steps, and rather than directly testing a hypothesis, it may *result* in a hypothesis for theory construction (Kendall, 2012, p.30).

In the humanities, on the other hand, the research process cannot necessarily be captured in a linear model. As indicated by Stone (1982), the "importance in humanities of criticism and analysis – including personal observation and opinion – marks a fundamental difference with the literature of science, and the subjective interaction between the humanist and his material is a unique feature." Trace and Karadkar (2017) suggest based on previous literature and an empiral study, that "the typical work process of humanities scholars encompasses a number of interconnected phases or activities that includes exploratory research (using secondary sources), identifying primary (archival) sources, conducting research in the archive, data collection and management, analysis, and writing." In a generalized humanities setting, Stone (1980) (as cited by Chu (1999)) differentiates five general stages, which can be summarized as *thinking and talking (1)*, *reading previous work (2)*, *study sources & notetaking (3)*, *drafting the write-up (4)* and *revising (5)*. She also states in Stone (1982) that differences between groups of disciplines within the humanities may be profound. Chu (1999) distinguishes between six research phases in the context of literary criticism: *idea (1)*, *preparation (2)*, *elaboration (3)*, *analysis & writing (4)*, *dissemination (5)*, and *further dissemination & writing (6)*[55]. A further discussion of these models is outside the scope of this section, but comparisons of models distinguishing stages in the humanities and social sciences can be found in Chu (1999); Bron et al. (2016). In our study, we predominantly focus on media and communication studies, which may use methods from the humanities and social sciences.

This chapter focuses on methods performed in different research phases of scholars which use longitudinal datasets, such as web archives or customly gathered collections of web content. In the context of web archive research, Brügger[56] has suggested four specific research phases relevant to web archives: corpus creation, analysis, dissemination and storage. First of all, during **corpus creation** a researcher identifies and isolates a corpus. In this thesis, we define a *corpus* as a collection of data assembled by a researcher for the purposes of studying her research questions. Corpus creation is followed by **analysis** of the created corpus, using analytical tools and visualizations. After this phase, **dissemination** follows, which involves dissemination of the analysis, for instance in scholarly papers. We can observe that these phases have similarities to phases of various models discussed above. Finally, in a stage not commonly included in research process models, Brügger suggests a need for long-term preservation of corpora and tools in the final **storage** phase. In our literature review, we will look at

[55] In contrast to Kendall's 'prescriptive' model, we can distinguish exploratory 'idea' phases in these evidence-based models, even though the later phases of these models do have similarities with Kendall's model.

[56] Summarized in a presentation from 2015 available at: `http://alexandria-project.eu/wp-content/uploads/2015/11/2nd_alex_ws_niels_bruegger.pdf` (accessed: 01/08/16)

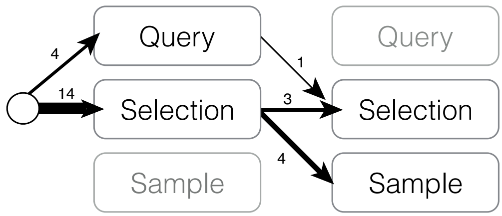

Figure 2.5: Consecutive corpus selection methods in the analyzed literature set (n=18).

the first three phases.

2.4.4 Findings

In this section, we describe our findings with respect to the methods researchers use during the first three phases distinguished by Brügger: corpus creation, analysis and dissemination.

Corpus Creation

For each surveyed journal paper, we assess how scholars collected their data (i.e. defined a corpus), using their descriptions in the papers. In particular, we look at the selection methods scholars used to arrive at their dataset.

The papers in the literature set use three main methods to define a corpus: they utilize a (textual) **query**, they select **webpages or websites**, or they take a **sample**. Four papers start with one or more queries issued at different moments in time, using regular search engines. This includes queries for the term "informetrics" (Bar-Ilan and Peritz, 2008), or descriptors of youth movements (Xenos and Bennett, 2007). The majority of papers (fourteen instances), however, starts with a specific selection of pages or sites, often based on an authoritative list. For example, one article starts with an official list of academic libraries in the USA (Comeaux and Schmetzke, 2013), while another paper utilizes a list of insurance companies from a national UK report (Waite and Harrison, 2007).

Most strikingly, half of the papers combine **multiple methods** to generate their corpus; illustrated in Figure 2.5. For instance, after selecting a list of insurance companies, Waite and Harrison (2007) took a sample of all companies on the list; and after selecting academic webpages, Payne and Thelwall (2008) used a random sample of links. Others select a daily (Karlsson et al., 2015) or weekly (Li et al., 2014) sample for a longer time period. Researchers may take a sample to reduce the expansive size of the initial selection, or to reduce data bias (for instance, using a list of social network sites, John (2013) inspected "each site on the first day of every month.")

Since we focus on longitudinal datasets, in which the time aspect is essential, we evaluate next which temporal ranges the scholars selected for their data. Five papers select web content at **multiple points in time** (for example a singular moment each year, during eight consecutive years (Bar-Ilan and Peritz, 2008)). Eight papers, however, select a **singular temporal range**, for instance one week (Mahrt and Puschmann, 2014). Finally, five journal publications in the set select **multiple timeranges**; for example one week during three consecutive years (Vara-Miguel et al., 2014).

Summarizing, in the analyzed literature set we found three main methods used by scholars to define their corpus, often used in combination: queries, selections and samples. The scope of the selection methods confirms the findings of our previous study in a new media context (Section 2.3), in which researchers explicitly expressed selection needs beyond solely queries. Here, we also found evidence for the value of combinations of selection methods, as well as for the varying granularity of selected time periods. The next question is *how* the scholars analyzed their gathered corpus.

Analysis Methods

The self-reported analysis methods available in the literature set may provide insights into the types of analysis done by researchers. These methods are categorized by the scholars themselves as **quantitative methods** (nine papers), **qualitative methods** (two papers) or a **combination** thereof (seven instances).

Zooming in to the specific methods themselves, the prime analysis method is **content analysis** (sixteen instances). This has been defined as: an "observational research method that is used systematically, objectively and quantitatively to analyse message characteristics" (Kolbe and Burnett (1991) as cited by Waite and Harrison (2007)). Here, we differentiate between *automatic content analysis* and *manual content analysis*. In the first case, existing tools (e.g., Bobby 3.1.1[57]), or custom-developed tools are utilized to obtain (mostly) quantitative statistics about phenomena under study. This is done in four of the surveyed papers. On the other hand, the authors of the majority of papers (twelve instances) manually develop coding schemes of web elements, pages and sites, sometimes based on existing frameworks and checklists. In part, this explains the necessity to take samples described in the previous section (to reduce the number of elements to code). Two papers feature **network analysis**, entailing "a distinct focus on the relationships between entities (individuals, groups, nations or – in the context of the internet– websites, elements on given websites

[57] Comeaux and Schmetzke (2013) use this tool: `http://web.archive.org/web/20000311102551/http://www.cast.org/bobby` (accessed: 01/08/16)

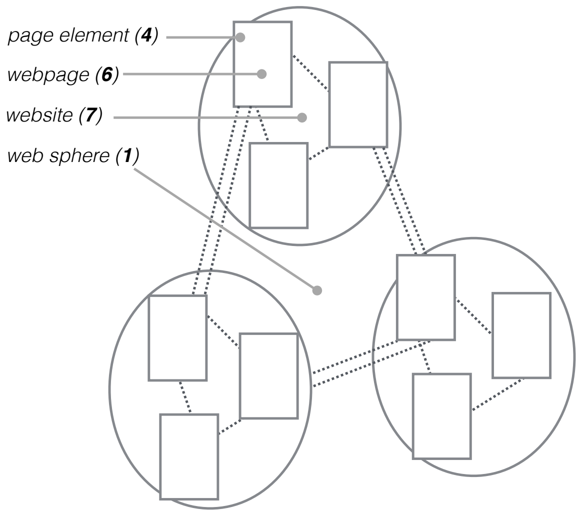

Figure 2.6: Analytical levels of the web, adapted from Brügger (2009). Number of papers indicated per analytical level (n=18).

[e.g. images and textboxes], users). That is, relations are at the centre of network analysis, and these are conceptualised in terms of nodes (entities) and links (connections and patterns of activity among nodes)." (Lomborg, 2012, p.6). For instance, Xenos and Bennett (2007) look at the development of links between electoral websites. Finally, four papers contextualize results derived from web data with other data collection instruments: interviews or surveys.

Brügger (2009) identifies various interrelated **analytical levels** or analytical strata of the web: the individual *elements* of a webpage, the individual *webpage*, the individual *website*, the *web sphere* and the *web as a whole*. The papers in the surveyed literature set use various analytical levels, illustrated in Figure 2.6.

Four of the eighteen analyzed journal papers perform their analysis at the level of the webpage *element*. For instance, Salisbury and Griffis (2014) look at textual mission statements included on academic library websites, and Chua and Banerjee (2013) focus on questions and answers on community question-answering sites. The elements analyzed in the surveyed papers are not just limited to textual elements: Ghobadi and Clegg (2015) analyze videos on YouTube as well as written comments.

The individual *webpage* is the unit of analysis of six papers in our set of literature, for example blog items (Mahrt and Puschmann, 2014). Li et al. (2014) analyze content variations between lowly and highly ranked pages on Google.

Seven papers perform their analysis at the level of the individual *website*.

For example, Koc-Michalska et al. (2014) look at websites of political actors, and Zhitomirsky-Geffet and Maman (2014) investigate real-estate websites. In the surveyed papers, there is no clear consensus about the concept of a website, as authors in the literature set define 'site' differently. Some authors equate website with homepage or entry page (e.g. Koc-Michalska et al. (2014)). Others, on the other hand, conceptualize website as a website's homepage and all top-level pages linked from the homepage (Comeaux and Schmetzke, 2013; Karlsson et al., 2015), or as homepages and a 'selection' of pages, such as political actors' websites (Koc-Michalska et al., 2014). These papers look at a variety of topics: from online campaigning in France, to Swedish broadcasting sites, to the concept of sharing in Web 2.0.

The final, and broadest level of analysis in the surveyed papers is the *web sphere*, defined by Schneider and Foot (2004) as "a hyperlinked set of dynamically-defined digital resources that span multiple websites and are deemed relevant, or related, to a central theme or 'object'." Xenos and Bennett (2007) compare the youth engagement sphere with the electoral web sphere in France.

Summarizing, we observed a combination of qualitative and quantitative methods in the analyzed literature set. The prime analysis method is content analysis, which was predominantly done via manually developed coding schemes, while only a minority of papers used automatic content analysis methods. The focus of the analyses themselves is mostly on the page and website level, and we could observe varying conceptualizations of what exactly constitutes a website. How the scholars' findings are disseminated is described next.

Results Dissemination

The results of the various analyses done by the scholars are naturally disseminated in the surveyed journal papers, and the textual descriptions of findings are usually combined with tabular and visual illustrations. Here, we intend to improve our understanding of the nature of these illustrations.

The authors use different methods to illustrate the results of their studies. In sixteen papers, the results are summarized in **tables**. First, tables in this literature set may illustrate aspects related to the used *data and method*. For instance, Mahrt and Puschmann (2014) include a table with the characteristics of blog posts selected for content analysis. Second, tables may represent *contextual information*, such as expenses for online campaigning in the paper by Koc-Michalska et al. (2014)). Third, the *findings* themselves are summarized in tables. For instance, Bar-Ilan and Peritz (2008) illustrate their findings with tables depicting the most frequently appearing domains per year, and Mahrt and Puschmann (2014) show descriptive statistics of different question types for community question-answering sites.

Table 2.3: Summary of authors' choices in each research phase, expressed as number (perc.) of papers

Corpus definition†	*Analysis*†		*Dissemination*†
select: 13 (72%)	man. content analysis: 12 (67%)	page elem.: 4 (22%)	tables: 15 (83%)
sample: 5 (28%)	auto. content analysis: 4 (22%)	page: 6 (33%)	graphs: 9 (50%)
query: 4 (22%)	network analysis: 2 (11%)	website: 7 (39%)	netw. diagr.: 1 (6%)
	grounded theory: 1 (6%)	web sphere: 1 (6%)	model: 1 (6%)

† *Since multiple methods may be used in each paper, percentages may add up to more than 100%*

In ten papers, researchers also use various types of **diagrams**, often depicting the temporal development of phenomena under study (in particular, graphs and bar charts). For instance, Karlsson and Clerwall (2012) show the degree of news items over time for different news outlets. Chua and Banerjee (2013) illustrate their findings with interaction plots for answer quality and speed across different types of questions (e.g. factoid or complex).

Comeaux and Schmetzke (2013) use screenshots to provide examples of accessibility of websites, while Salisbury and Griffis (2014) provide pie charts of the availability of missions statements on academic library websites. One paper, Xenos and Bennett (2007) depicts the link network of the youth engagement web sphere at different moments of an election cycle in 2004. Finally, Ghobadi and Clegg (2015) derive a model from the results of their study.

Summarizing, we observed a common use of illustrations for the analyses performed by the scholars. Tables may serve to illustrate scholars' data and methodology, to provide contextual information, as well as to summarize the findings themselves. Visual illustrations, such as graphs, depict various aspects of scholars' findings. Considering the longitudinal nature of the scholars' datasets, it comes as no surprise that many of these illustrations depict changes in phenomena over time.

2.4.5 Implications for Access Systems

In our literature analysis, we encountered various ways of how scholars, in the context of web data, define corpora, perform analysis and disseminate their results (summarized in Table 2.3, further details and citations are listed in Appendix 6.2). Most of the scholars in the surveyed literature, however, did not use institutional web archives as a data source, but they created their own collections. This provides further evidence for Dougherty and Meyer (2014)'s observed gap between the large potential community of researchers addressed by web archives, and the small community actually using them. In this analytical literature review, we found a gap between the research activities of scholars, and the functionalities of access systems. We argue that these needs should be understood in the context of the whole research *process*. Therefore, we now discuss the limitations of web archive access systems, as well as the opportunities to improve them, using Brügger's research phases as a 'lens'.

This literature review was initiated by looking at the **corpus creation** research phase. We observed that researchers often define their corpus using *selections* (e.g. authoritative URL lists), *queries*, and *samples*. This confirms the findings from the action research setting described in the previous section. Few elements of these corpus definition methods are supported by current web archive access systems. As discussed in Section 2.2, the commonly used Wayback Machine in essence offers 'single URL' access (facilitating browsing of one URL at a time). Thus, it is possible to select a single website, but it may be problematic to actually retrieve lists of URLs without resorting to external tools[58]. Moreover, there are no integrated ways to query the Wayback Machine, nor to perform automated sampling. Full-text search tools, on the other hand, provide options to query for the text contained in archived webpages, but generally no ways to perform rich list-based selections, or to take a sample of webpages. For instance, in Section 2.3, we have seen that the WebARTist toolset provided various modalities to explore search results. However, it still required scholars to start with a query instead of other entry points.

Furthermore, the datasets of scholars in our literature review were longitudinal in nature. Hence, it is important to include functionality for selecting singular timepoints, timeranges and multiple timeranges in access systems. Most current full-text search tools for web archives, including WebARTist, allow for filtering results based on their timestamps. This filtering is possible for singular points in time as well as timeranges, but not for multiple timeranges.

Moving to the **analysis** phase, the analytical literature review has shown that scholars used automatic and manual content analysis. Hence, custom tools were used to arrive at statistics about observed phenomena, or manual coding schemes were developed to analyze the data at hand. Furthermore, we observed that the scholars' analysis took place at various levels of granularity, ranging from atomic webpage elements to interrelated web spheres.

Despite the essential role of analysis in the workflow of researchers, both URL-based and full-text search access systems for web archives generally do not provide ways to analyze selections of content. The WebARTist toolset provides basic statistics for resultsets, and thus unlocks exploratory analysis possibilities for scholars; even though a researcher may have to export resultsets for full-fledged analysis in other tools. However, within WebARTist and current institutional web archive access tools, it is not possible to save, annotate and

[58] For example using tools made by the Digital Methods Initiative (DMI), as discussed in Section 2.3: `https://wiki.digitalmethods.net/Dmi/ToolDatabase?cat=DeviceCentric&subcat=Wayback` (retrieved 16 April 2015)

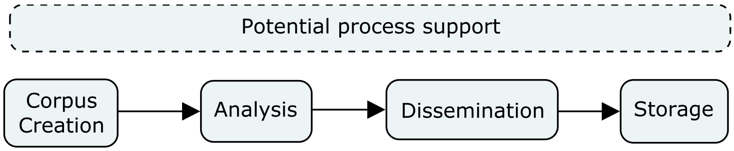

Figure 2.7: Phases of research, adapted from Brügger[24]; and proposed process support.

further analyze customly selected corpora within the system.

Opportunities for integrating analysis tools in future web archive access tools are multifold. Possibilities may range for more options to automatically analyze and generate statistics for page elements, webpages, websites and web spheres within custom corpora, to the support of annotation and the development of coding schemes within the research process.

Finally, for **dissemination** of research results, we observed that the majority of researchers use tables and diagrams to illustrate their findings. As a consequence of the temporal nature of their research problems and datasets, the tables and diagrams often depict evolving aspects of the findings over time.

Few access systems for web archives offer ways to export crawls or resultsets in a structured format (possibly due to the legal issues outlined in section 2.2). Within the research explorations described in Section 2.3, however, we found that there was substantial demand from scholars for these possibilities. Being able to export results may help for doing custom analysis, but also for creating visualizations and dataset summaries.

In addition, some institutional web archives, including the UK Web archive, offer data visualization features, for instance tag clouds or n-grams[59]. The WebARTist toolset also provides opportunities to visualize resultsets as keyword clouds, temporal graphs and geographical maps. These visualizations may be useful in various phases of research. However, these visualization methods currently can only be applied to the full archive, or to resultsets returned by queries. To our knowledge, the ability to visualize user-defined corpora has not been included in a web archive context so far. Hence, further extensions for future web archive access systems would include customizable facilities to summarize and visualize specific data to be used in reports or papers.

2.4.6 Discussion

This analytical literature review based on published research papers using web data suggests various omissions in the current support for corpus creation, analysis and dissemination research phases. Still, the question is how such a varied

[59] The UK Web Archive's visualization tools: `http://www.webarchive.org.uk/ukwa/visualisation` (accessed: 01/08/16)

set of missing features could be combined into an integrated access system for web archives, which still has a high usability[60] and which does not cognitively overload users (Hearst, 2009).

To arrive at this support at a broader level, inspiration may be found in the discussed models of the research process: we propose an approach to more directly support the **research process** of scholars. We may focus on the four stages Brügger has distinguished in the context of web archive research[61]: *corpus creation*, identifying and isolating a corpus; *analysis* of the created corpus, using analytical tools and visualizations; *dissemination* of the analysis, for instance in scholarly papers; leading to final *storage*, involving long-time preservation of corpora and tools. In our view, more support for these research phases should be offered *in-situ*, i.e. within web archive access systems. Research process models may offer inspiration to divide required functionality into different phases at the interface level.

This way, researchers can define their corpora, analyze, disseminate and store them within a workflow in one system. So far, this has been achieved in specialized fields, such as bioinformatics (Zoubarev et al., 2012), and genomics research (Goecks et al., 2010), where systems allowing for collaborative analysis, sharing and reuse have attracted a substantial number of researchers. While it is a major challenge to realize this process approach in a more general research context such as the web archive, these previous examples may provide insights and inspiration. In future work, we aim to extend the WebARTist toolset to provide some of these possibilities.

2.5 Conclusion

The web archives which contain our online past have substantial potential for scholarly use, but this potential has not been harnessed yet. In this chapter, we examined why this is the case, and how we may potentially improve upon this situation. In particular, the aim of this chapter was to better understand the limitations of current web archive access tools in a research context. We performed two studies to evaluate and extend the functionality of search-based web archive access tools. Furthermore, we suggested various steps to improve upon found issues, which may transform web archive access tools from mere search tools into 'research engines'.

First, section 2.2 focused on defining web archiving, classifying web archiving actors and the properties of web archives as research datasets. In particular,

[60] For a definition of usability, see: `http://www.nngroup.com/articles/usability-101-introduction-to-usability` (accessed: 01/08/16)

[61] Summarized in a presentation from 2015 available at: `http://alexandria-project.eu/wp-content/uploads/2015/11/2nd_alex_ws_niels_bruegger.pdf` (accessed: 01/08/16)

some deficiencies came up, such as limitations in data quantity (due to technical and strategic reasons), limitations in data quality (technical issues result in inconsistent data), and limitations in access (due to legal issues and limits of access systems). The remainder of the chapter focused on evaluating and extending the research opportunities of searchable web archives.

In an action research setting, Section 2.3 looked at our initial research question (**RQ1.1**): *Which limitations of current web archive access tools can be identified in a new media research context? To what extent can search-based tools improve research access to web archives?* During a series of workshops and events, the WebARTist toolset, a search system for the Dutch web archive, was created, extended and evaluated. This section concludes that search-based web archive access tools constitute a useful addition to a researchers' toolset. However, also inherent limitations of the search approach came to light, including a lack of rich (corpus) selection methods and analysis functionality, as well as a lack of transparency. We outlined the need to document and reveal the influences of curatorial decisions, crawler limitations and algorithmic influences of search engines on the archived web content retrieved by the researcher. This transparency is of key importance, since the retrieved content may influence scholars' analysis.

Section 2.4 investigated the following question (**RQ1.2**): *Which corpus creation, analysis and dissemination methods do media and communication scholars use in the context of web data? What are the implications for search-based web archive access tools?* It aimed at obtaining a better understanding of scholars' corpus selection, analysis and dissemination methods, based on a analytical literature review of journal papers in media and communication studies. This review and analysis revealed shortcomings of current (browse- and search-based) access tools. In particular, current tools lack proper support of selection methods (e.g. list-based selections and samples, as well as temporal ranges at different levels of granularity). Furthermore, current toolsets do not provide ways to perform content analysis and network analysis directly in the system, nor analysis at different granularities. Visualizing findings, due to a lack of support of corpus generation and analysis methods, is not possible. Hence, we argue for a more inclusive approach, in which potential access systems for web archives provide integrated *process support*, for different activities in different research phases.

Finally, we return to our main research question: *To what extent do search-based web archive access tools facilitate research in a new media setting?* On the one hand, we saw that search-based access are a useful addition to URL-based approaches in the context of scholarly research. URL-based access methods necessitated the knowledge of URLs, and custom tools to analyze more than one URL at a time. On the other hand, we observed that the functionality of the full-text search approach remains limited, thus inhibiting research using web archives. Various selection, analysis and visualization methods useful in the research process are not available. Hence, we concluded that improved *process support* would be desirable, which is no easy challenge to realize. To achieve this aim, more supportive search systems and interfaces are essential, beyond standard full-text search approaches. As information seeking is an important element of the research process, we may be informed by research in the field of information seeking and retrieval (see Part II of this thesis). First, however, the next chapter will focus on increasing the transparency of web archives.

3

Lost but Not Forgotten: Finding Pages on the Unarchived Web

The previous chapter has shown that web archives attempt to preserve the fast changing web, but that they will always be incomplete. Due to restrictive selection policies, restrictions in crawling depth and frequency, as well as technical limitations, large parts of the World Wide Web are unarchived and therefore lost to posterity. This, combined with a lack of insights into indexing and retrieval parameters, leads to a lack of transparency of current web archive search systems: it is unknown which items are *missing* from result lists. Therefore, to improve scholarly access to web archives, it would be worthwhile to devise methods to discover unarchived material, thus providing contextualization about the incompleteness of a web archive.

In this chapter, we propose an approach to uncover unarchived webpages and websites, and to reconstruct different types of descriptions for these pages and sites, based on links and anchor text in the set of crawled pages. We experiment with this approach on the Dutch web archive. After reviewing previous literature on link evidence and anchor text (Section 3.2), we investigate the amount of unarchived material detected in the archive (Section 3.4). We observe that the crawled web contains evidence of a remarkable number of unarchived pages and websites, potentially dramatically increasing the coverage of a web archive. Next, we assess the richness of reconstructed descriptions of unarchived pages (Section 3.5) and sites (3.6), which are succinct in nature. Subsequently, we evaluate the usefulness of page and host-level representations of unarchived contents for retrieval (Section 3.7). In a known-item search setting, we assess whether the succinct descriptions of previously lost webpages and websites are rich enough to uniquely identify pages on the unarchived web. Ultimately, we indicate how our findings may be used to contextualize current search systems for web archives.

This chapter is based on Huurdeman et al. (2015b), previous versions were published as Huurdeman et al. (2014); Samar et al. (2014). The link extraction and analysis was carried out on the Dutch national e-infrastructure with support of SURF Foundation.

3.1 Introduction

Memory and heritage institutions address the ephemerality of the World Wide Web by systematically preserving parts of it. As we have observed in the previous chapter, some instutions intend to archive a selection of websites, while others focus on a country's top-level domain, or even on the grand scale of the entire web. Upon selection, the material is usually harvested by web *crawlers*, iteratively following links occurring in webpages. However, as outlined in Section 2.2, performing web archiving is certainly not without challenges. Legal restrictions and institutional policies may prevent large-scale harvests, but also other trade-offs are involved. For instance, the implemented crawling strategy influence the breadth and depth of material which is stored in the archive, and seemingly insignificant changes in crawl settings may have profound influences on captured material. Moreover, various technical limitations impede harvesting of material, at times causing entire websites, individual pages, and page elements to be absent from the archive. The overall consequence is that our web archives are highly incomplete, and researchers and other users treating the archive to reflect the web as it once was, may draw false conclusions due to unarchived content. This is corrobated by the fact that the various variables causing this incompleteness are usually hidden by access systems. In this chapter, we choose to directly address this lack of **transparency**, and focus on to the following main research question: *To what extent can representations of unarchived webpages and websites enhance search-based access to web archives?*

Hence, we investigate if can we recover parts of the unarchived web. This may seem like a daunting challenge or a mission impossible: how can we go back in time and recover pages that were never preserved? Our approach is to exploit the hyperlinked structure of the web, and collect evidence of uncrawled pages from the pages that were crawled and are part of the archive. Moreover, we move beyond this mere evidence, and aim to generate representations of web pages and websites which are seemingly lost forever. These representations can potentially be integrated in access interfaces.

The chapter investigates the following research questions:

RQ2.1 What fraction of unarchived web pages and websites can be uncovered based on references to them in the web archive?

We exploit the link structure of the crawled content to derive evidence of the existence of unarchived pages. In addition, we characterize and classify the encountered unarchived pages, domains and hostnames.

RQ2.2 How can the richness of the representations created for unarchived pages be characterized?

We build implicit representations of unarchived web pages and domains, based on link evidence, URL words and anchor text, and investigate the richness (or sparseness) of the descriptions in the number of incoming links and the aggregated anchor text. We further break this down over unarchived homepages and other pages.

RQ2.3 To what extent can representations of websites be enriched by aggregating page-level evidence from pages sharing the same hostname?

Here, we look at the additional value of aggregating anchor text and URL words at the host level, in effect creating site-level representations. We investigate the quantity and richness of these additional representations, and compare this with page-level representations of homepages.

RQ2.4 How effective are the derived page-level and site-level representations in a known-item search setting?

As a critical test, we study the effectiveness of the derived representations of unarchived home pages and deep pages in a known-item search setting. Only if the derived page and site-based representations sufficiently characterize the unique page's content, we have a chance to retrieve the page within the top search results.

The remainder of the paper is organized as follows: we first introduce related work (Section 3.2), followed by a description of the experimental setup (Section 3.3). Next, we look at the results of our analysis, characterizing the actual contents of the Dutch web archive and the unarchived pages around the archive (Section 3.4). We study the potential richness of generated representations of unarchived web pages (Section 3.5). Next, we take a look at the comparative richness of aggregated host-level representations of unarchived websites (Section 3.6). The generated representations are evaluated by their utility to retrieve the page or host in a known-item search scenario (Section 3.7). We conclude by discussing the implications of our findings as well as outlining future work (Section 3.8).

3.2 Related Work

In this section, we discuss related work, which falls in two broad areas. First, we discuss related research in web archiving and web preservation. Second, we discuss previous literature on search, based on link evidence and anchor text.

3.2.1 The Content Web Archives Fail To Capture

Experts in the web archiving community discuss the shortcomings of web archiving crawlers in terms of the content they fail to capture (Masanès, 2006; Day, 2003; Hockx-Yu, 2011). Some websites are intentionally excluded, breadth-first crawls might not capture deeper pages of a website, and selective crawlers exclude sites beyond the scope of the selection policy. Moreover, the pace of the development of web crawling functionality does not keep up with the rapid developments of web technologies. Hence, independent from the chosen crawling strategy, websites, web pages and page elements may be missing due to crawling inadequacies (Day, 2003; Hockx-Yu, 2011; Masanès, 2006). Some authors have looked at the impact of missing resources on the display of web pages. For instance, Brunelle et al. (2014) have proposed a damage rating measure to evaluate archive success. Using this measure, they showed that the Internet Archive is missing an increasing number of important embedded resources over the years.

The incompleteness of web archives can impede their value in the context of scholarly research (Brügger, 2012; Hockx-Yu, 2014). However, crawlers do register additional information about web content they encounter. This additional information includes server-side metadata of harvested pages (such as timestamps and HTML response codes), and information embedded in pages (for instance their hyperlinks and associated anchor text). Rauber et al. (2002) have recognized the wealth of additional information contained in web archives which can be used for analytical purposes. Gomes and Silva (2005) used data obtained from the domain crawl of the Portuguese web archive to develop criteria for characterizing the Portuguese web.

The Memento project has expanded the scope of analysis of archived web data beyond the boundaries of a single archive, in order to profile and analyze coverage of archived websites across different web archives. Memento (Van de Sompel et al., 2013) is an HTTP-based framework which makes it possible to locate past versions of a given web resource through an aggregator of resources from multiple web archives. AlSum et al. (2014) queried the Memento aggregator to profile and evaluate the coverage of twelve public web archives. They found that the number of queries can be reduced by 75% by only sending queries to the top three web archives. Here, coverage (i.e. whether a resource is archived and in which archive its past versions are located) was calculated based on the HTTP header of host level URLs.

Instead of just measuring what is missing, we will try to uncover significant parts of the unarchived web. We do so by reconstructing representations of the unarchived web pages and websites using *link and anchor text evidence*, discussed next.

3.2.2 Link Evidence and Anchor Text

One of the defining properties of the Internet is its hyperlink-based structure. The web's graph structure is well studied, and also methods to use this structure have been widely applied, especially in the context of web retrieval (for example PageRank (Page et al., 1999)). The links which weave the structure of the web each consist of a destination URL and an anchor text description. Aggregating anchor text of links makes it for example possible to create representations of target pages. In this paper, we mainly focus on the use of anchor text.

Craswell et al. (2001) explored the effectiveness of anchor text in the context of site finding. Aggregated anchor texts for a link target were used as surrogate documents, instead of the actual content of the target pages. Their experimental results show that anchor texts can be more effective than content words for navigational queries (i.e. site finding). Work in this area led to advanced retrieval models that combine various representations of page content, anchor text, and link evidence (Kamps, 2005). Fujii (2008) presented a method for classifying queries into navigational and informational. Their retrieval system used content-based or anchor-based retrieval methods, depending on the query type. Based on their experimental results, they concluded that content of web pages is useful for informational query types, while anchor text information and links are useful for navigational query types. Contrary to previous work, Koolen and Kamps (2010) concluded that anchor text can also be beneficial for ad hoc informational search, and their findings show that anchor text can lead to significant improvements in retrieval effectiveness. They also analyze the factors influencing this effectiveness, such as link density and collection size.

In the context of web archiving, link evidence and anchor text could be used to locate missing web pages, of which the original URL is not accessible anymore. Martinez-Romo and Araujo (2010) studied the problem of finding replacements of broken links (missing pages) in a web page. They contructed queries based on terms selected from anchor text pointing to the broken link and expanded these queries by adding information from the web page containing the broken link. Then, they submitted the constructed queries to a standard search engine to retrieve candidate replacements for the missing page. Klein and Nelson (2014) computed lexical signatures of lost web pages, using the top n words of link anchors, and used these and other methods to retrieve alternative URLs for lost web pages. The versatility and potential use of anchor text is further exemplified by Kraft and Zien (2004), who show that anchor text can also be used to generate query suggestions and refinements.

Following Kleinberg (1999), Dou et al. (2009) took the relationships between source pages of anchor texts into account. Their proposed models distinguish between links from the same website and links from related sites, to better estimate the importance of anchor text. Similarly, Metzler et al. (2009) smoothed the influence of anchor text which originates from within the same domain, using the 'external' anchor text: the aggregated anchor text from all pages that link to a page in the same domain as the target page. Their proposed approach also facilitates overcoming anchor text sparsity for pages with few inlinks. Along the same line, Broder et al. (2010) used site-level information for improving web search results. They achieved this by creating two indices: a URL index based on the page content, and a site index representing the entire website. They introduced two approaches for creating site-based representations. The first site-level representation was created by concatenating the content text of all pages from a site, or from a sample of pages. The second site-representation was created by aggregating all anchor text of external links pointing to pages in the site. Their experimental evaluation showed that the combination of page and site-level indices is more effective for web retrieval than the common approach of only using a page-level index. Their results also indicate that site-level anchor text represention perform better than site-level representations based on concatenated content text.

Another aspect of anchor text is its development over time: often single snapshots of sites are used to extract links and anchor text, neglecting historical trends. Dai and Davison (2010) determined anchor text importance by differentiating pages' inlink context and creation rates over time. They concluded that ranking performance is improved by differentiating pages with different in-link creation rates, but they also point to the lack of available archived resources (few encountered links were actually available in the Internet Archive).

Our approach is inspired by the previous results on various web centric document representations based on URL and incoming anchor text, typically used in addition to representations of the page's content (Craswell et al., 2001; Kraaij et al., 2002; Ogilvie and Callan, 2003; Kamps, 2005). We focus on the use case of the web archive, which is different from the live web given that we cannot go back and crawl the unarchived page, hence have to rely on these implicit representations exclusively. It is an open question whether the resulting derived representations — based on scant evidence of the pages — is a rich enough characterization to be of practical use. This chapter builds on earlier work on uncovering unarchived pages (Samar et al., 2014) and the recovery and evaluation of unarchived page descriptions (Huurdeman et al., 2014). Parts of the page-level results were presented in Huurdeman et al. (2014), now extended with the analysis of site-level representations to overcome anchor text sparsity at the page level (Broder et al., 2010).

Table 3.1: Crawled web objects per year in the Dutch web archive

year	number of web objects
2009-2011	108,775,839
2012	38,865,673
	147,641,512

Table 3.2: Types of web objects crawled in 2012 (MIME-types)

MIME-Type	count	%
text/html	25,380,955	65.3%
image/jpeg	6,518,954	16.8%
image/gif	1,222,480	3.1%
image/png	1,171,585	3.0%
application/pdf	816,746	2.1%
text/plain	642,282	1.7%
text/xml	488,569	1.3%
rss/xml	483,858	1.2%
other	*2,140,244*	*5.5%*

3.3 Experimental Setup

This section describes our experimental setup: the dataset, the Hadoop-based link extraction methods, and the way the links were aggregated for analysis.

3.3.1 Data

This study uses data from the Dutch web archive (Ras, 2007) at the National Library of the Netherlands (KB). In 2012, the KB was archiving a pre-selected (seed) list of more than 5,000 websites, and this has grown to over 10,000 websites in 2016. Websites for preservation are selected by the library based on categories related to Dutch historical, social and cultural heritage. Each website on the seed list has manually been categorized by the curators of the KB using a UNESCO classification code.

Our snapshot of the Dutch web archive consists of 76,828 ARC files, which contain aggregated web objects. A total number of 148M web objects has been harvested between February 2009 and December 2012, resulting in more than 7 Terabytes of data (see Table 3.1). In our study, we exclusively focus on the content crawled in 2012 (35.7% of the total data harvested between 2009 and 2012). Additional metadata is available in separate documents, including the KB's selection list, dates of selection, assigned UNESCO codes and curators' annotations. Table 3.2 shows a summary of the types of content crawled in

2012, with their counts and percentages. It shows that the majority of content types crawled in 2012 consists of HTML-based textual content (65.3%), and images in various formats (22.9%).

In our extraction, we differentiate between four different types of URLs found in the Dutch web archive:

1. URLs that have been archived intentionally as they are included in the seedlist,
2. URLs that have been unintentionally archived due to the crawler's configuration,
3. unarchived URLs, discovered via the link structure of the archive, of which the parent domain is included in the seedlist (which we will refer to as the *inner aura*)
4. unarchived URLs, discovered via the link structure of the archive, of which the parent domain is not on the seedlist (which we will refer to as the *outer aura*).

Section 3.4 describes these four types of archived and unarchived URLs in more detail.

3.3.2 Link Extraction

We created our dataset by implementing a specific processing pipeline. This pipeline uses Hadoop MapReduce and Apache Pig Scripts for data extraction and processing.

The first MapReduce job traversed all archived web objects contained in the archive's ARC files. Web pages with the following properties were processed:

- crawled in 2012
- having the MIME-type *text/html*, and
- having at least one anchor link including a destination URL and (non-empty) anchor text.

From all pages with these properties, each anchor link found in the page was extracted using JSoup. As we focus on links to textual content, only '`a`' anchors were extracted and other references to embedded content, such as '`script`' links, embedded images (via the '`img`' tag) and embedded '`iframe`' content were ignored. We keep the *source URL* (which is the page URL), *target URL* (the URL of the page that the link is pointing to), and the *anchor text* of the link (the textual description of a link). Links without anchor text were discarded. This extracted link information is combined with basic ARC metadata about the source page (e.g. crawldate). In addition, other relevant information is added, such as the hashcode (MD5) of the source page, the occurrence of the source and target page on the KB's seedlist, and assigned UNESCO classification codes.

A second MapReduce job built an index of all URLs in the Dutch web archive, with their associated crawldate. Using this index, we performed lookups to validate whether or not a target URL found in the link information exists in the archive in the same year. Our final output format for extracted links contains the following properties:

sourceURL, sourceUnesco, sourceInSeedProperty, targetURL, targetUnesco, targetInSeedProperty, anchorText, crawlDate, targetInArchiveProperty, sourceHash

In our study, we look at the content per year. While some sites are harvested yearly, other sites are captured biannually, quarterly or even daily. This could result in a large number of links from duplicate pages. To prevent this from influencing our dataset, we deduplicated the links based on their values for year, anchor text, source, target, and (MD5) hashcode. The hashcode is a unique value representing a page's content, and is used to detect if a source has changed between crawls. We keep only links to the same target URL with identical anchor texts if they originate from unique source URLs.

In our dataset, we include both inter-server links, which are links between different servers (external links), and intra-server links, which occur within a server (site internal links). We also performed basic data cleaning and processing: removing non-alphanumerical characters from the anchor text, converting the source and target URLs to the canonicalized sort-friendly format known as SURT, removing double and trailing slashes, and removing *http(s)* prefixes (see `http://crawler.archive.org/apidocs/org/archive/util/SURT.html`).

3.3.3 Link Aggregation

We combine all incoming links and anchor text to an aggregated page level representation. In this process, we create a representation that includes the target URL, and grouped data elements with source URLs, anchor texts and other associated properties. Using another Apache Pig script, we performed further processing. This processing included tokenization of elements such as anchor text and URL words. This, combined with other processing, allowed us to include counts of different elements in our output files, for example the unique source sites and hosts, unique anchor and URL words, and the number of links from seed and non-seed source URLs. We also split each URL to obtain separate fields for TLD (top-level domain), domain, host and filetype. To retrieve correct values for the TLD field, we matched the TLD extension from the URL with the official IANA list of all TLDs, while we matched extracted filetype extensions of each URL with a list of common web file formats.

We also aggregate all page level evidence within the same hostname to an aggregated site level representation. A site-level representation aggregates evidence for all pages of a given website, or a sample thereof (see for instance Broder et al. (2010)). In this paper, we consider all pages sharing the same parent *host* as a 'site'. We adapted our Pig scripts to aggregate link evidence for all pages under each host, including anchor text and incoming URLs. We carried out similar processing steps as for the page-level representations. For later analysis, we saved up to 100 URLs of pages under each host.

The final aggregated page and site-level representations containing target URLs, source properties and various value counts were subsequently inserted into MySQL databases (13M and 0.5M rows), to provide easier access for analysis and visualization via a web application.

3.4 Expanding the Web Archive

In this section, we study the first research question of this chapter (**RQ2.1**): *What fraction of unarchived web pages and websites can be uncovered based on references to them in the web archive?* We investigate the contents of the Dutch web archive, quantifying and classifying the unarchived material that can be uncovered via the archive.

3.4.1 Archived Content

We begin by introducing the actual archived content of the Dutch web archive in 2012, before characterizing the unarchived contents in the next subsection. As introduced in section 3.3.2, we look at the pages in the archive with a *text/html* MIME-type. Here, we count the unique text-based web pages (based on MD5 hash) in the web archive's crawls from 2012, totaling in 11,041,113 pages. Of these pages, 10,158,586 were crawled in 2012 as part of the KB's seedlist (92%). An additional 882,527 pages are not in the seedlist but included in the archive (see Table 3.3). As discussed in section 3.2.1, each 'deep' crawl of a website included in the seedlist also results in additional ('out of scope') material being harvested, due to crawler settings. For example, to correctly include all embedded elements of a certain page, the crawler might need to harvest pages beyond the predefined seed domains. These unintentionally archived contents amount to 8% of the full web archive in 2012. Close dependencies exists between the chosen crawl settings and the resulting harvested material, the details of which are beyond the scope of this chapter. To avoid influence of artificial effects caused by variations in these crawl settings, we chose only one year for our evaluation instead of multiple years, giving us a relatively stable setting to work with.

Table 3.3: Unique archived pages (2012)

	on seedlist	not on seedlist	total
pages	10,158,586 (92.0%)	882,527 (8.0%)	11,041,113

Table 3.4: Unique archived hosts, domains & TLDs

	on seedlist	not on seedlist	total
hosts	6,157 (14.2%)	37,166 (85.8%)	43,323
domains	3,413 (10.1%)	30,367 (89.9%)	33,780
TLDs†	16 (8.8%)	181 (100%)	181

†*Since the values for the TLDs overlap for both categories, percentages add up to more than 100% (same for Table 3.8).*

Table 3.5: Coverage in archive

mean page count	on seedlist	not on seedlist
per host	1,650	24
per domain	2,976	29
per TLD	634,912	4,876

We can take a closer look at the contents of the archive by calculating the diversity of hosts, domains and TLDs contained in it. Table 3.4 summarizes these numbers, in which the selection-based policy of the Dutch KB is reflected. The number of hosts and domains is indicative of the 3,876 selected websites on the seedlist in the beginning of 2012: there are 6,157 unique hosts (e.g. `papierenman.blogspot.com`) and 3,413 unique domains (e.g. `okkn.nl`).

The unintentionally archived items reflect a much larger variety of hosts and domains than the items from the seedlist, accounting for 37,166 unique hosts (85.8%), and 30,367 unique domains (89.9% of all domains). The higher diversity of the non-seedlist items also results in a lower coverage in terms of number of archived pages per domain and per host (see Table 3.5). The mean number of pages per domain is 2,976 for the domains included in the seedlist, while the average number of pages for the domains outside of the seedlist is only 29.

According to the KB's selection policies, websites that have value for Dutch cultural heritage are included in the archive. A more precise indication of the categories of websites on the seedlist can be obtained by looking at their assigned UNESCO classification codes. In the archive, the main categories are Art and Architecture (1.3M harvested pages), History and Biography (1.2M pages) and Law and Government Administration (0.9M pages) (see Table 3.6). The pages

Table 3.6: Categories of archived pages in KB seedlist (top 10)

inner aura	count	%
1 Art & Architecture	1,328,114	13.0
2 History & Biography	1,174,576	11.5
3 Law & Government Administration	910,530	8.9
4 Education	803,508	7.9
5 Sport, Games & Leisure	712,241	7.0
6 Sociology, Statistics	694,636	6.8
7 Political Science	663,111	6.5
8 Medicine	590,862	5.8
9 Technology & Industry	469,132	4.6
10 Religion	377,680	3.7

Table 3.7: Unarchived *aura* unique pages (2012)

	inner aura	outer aura	total
pages	5,505,975 (51.5%)	5,191,515 (48.5%)	10,697,490

harvested outside of the selection lists do not have assigned UNESCO codes. A manual inspection of the top 10 domains in this category (35% of all unintentionally harvested pages) shows that these are heterogeneous: 3 websites are related to Dutch cultural heritage, 2 are international social networks, 2 websites are related to the European Commission and 3 are various other international sites.

3.4.2 Unarchived Content

To uncover the unarchived material, we used the link evidence and structure derived from the crawled contents of the Dutch web archive in 2012. We refer to these contents as the web archive's *aura*: the pages that are not in the archive, but which existence can be derived from evidence in the archive.

The unarchived *aura* has a substantial size: there are 11M unique pages in the archive, but we have evidence of 10.7M additional link targets that do not exist in the archive's crawls from 2012. In the following sections, we will focus on this aura, and differentiate between the *inner aura* (unarchived pages of which the parent domain is on the seedlist) and the *outer aura* (unarchived pages of which the parent domain is not on the seedlist). The inner aura has 5.5M (51.5%) unique link targets, while the outer aura has 5.2M (48.5%) unique target pages (see Figure 3.1 and Table 3.7).

Like the number of pages, also the number of unique unarchived hosts is

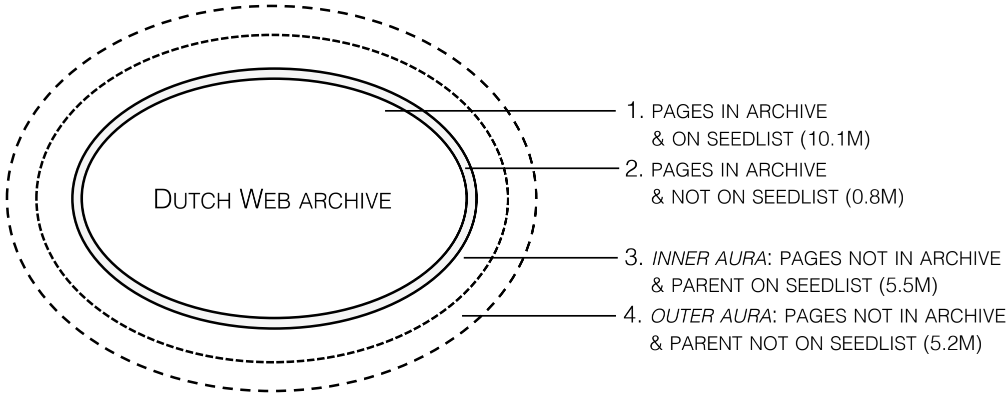

Figure 3.1: 'Layers' of contents of the Dutch web Archive (2012)

Table 3.8: Unarchived unique hosts, domains & TLDs

	inner aura	outer aura	total
hosts	9,039 (1.8%)	481,797 (98.2%)	490,836
domains	3,019 (0.8%)	369,721 (99.2%)	372,740
TLDs	17 (6.6%)	259 (100%)	259

quite substantial: while *in* the archive there are 43,323 unique hosts, we can reveal a total number of 490,836 hosts in the unarchived aura. There is also a considerable number of unique domains and TLDs in the unarchived contents (see Table 3.8).

The tables above also show the difference between the *inner* and *outer* aura. The outer aura has a much larger variety of hosts, domains and TLDs compared to the inner aura (Table 3.8). On the other hand, the coverage in terms of the mean number of pages per host, domain and TLD is much greater in the inner aura than the outer aura (see Table 3.9). This can be explained by the fact that the pages in the inner aura are closely related to the smaller set of domains included in web archive's seedlist, since they have a parent domain which is on the seedlist.

Finally, to get an overview of the nature of the unarchived resources, we have matched the link targets with a list of common web file extensions. Table 3.10 shows the filetype distribution: the majority consists of URLs without an extension (http), html, asp and php pages for both the inner and outer aura. Only a minority of references are other formats, like pdfs and non-textual contents (e.g. jpg files in the outer aura). This suggests that the majority

Table 3.9: Unarchived *aura* coverage (2012)

mean page count	inner aura	outer aura
per host	609	10
per domain	1,823	14
per TLD	323,881	20,044

Table 3.10: Unarchived *aura* filetypes

inner aura	count	%	outer aura	count	%
http	4,281,750	77.8	http	3,721,059	71.7
html	351,940	6.4	php	585,024	11.3
php	321,095	5.8	html	582,043	11.2
asp	38,0964	6.9	asp	181,963	3.5
pdf	70,371	1.3	jpg	30,205	0.6

of references to the unarchived aura which we extracted from the web archive points to textual web content.

As we observe a large fraction of URLs without an extension, we performed an additional analysis to shed more light on the included filetypes. We retrieved HTML status codes and MIME-types from the 'live' web, for a random sample of 1,000 unarchived URLs in the inner aura, and 1,000 URLs in the outer aura. For the inner aura, 596 of all URLs from the sample were available, while the remaining 404 were not accessible (resulting in 30x and 404 status codes). Of the resolvable URLs, 580 (97.3%) were of the MIME-type *text/html*, while only 16 URLs (2.7%) led to other filetypes (i.e. images, videos or PDFs). For the outer aura, 439 URLs were still accessible, and 561 of all URLs could not be retrieved. Of the URLs which resolved, 412 (93.8%) were of the *text/html* type, while only 27 (6.2%) led to other MIME-types. Hence, there is clear evidence that the uncovered URLs in both the inner and outer aura are predominantly text-based pages.

3.4.3 Characterizing the "Aura"

Here, we characterize unarchived contents of the archive based on the top-level domain distribution and the domain coverage.

From the top-level domains (TLDs) we derive the origins of the unarchived pages surrounding the Dutch web archive. Table 3.11 shows that the majority of unarchived pages in the inner aura (95.69%) have Dutch origins. The degree of .nl domains in the outer aura is lower, albeit still considerable, with 31.08% of all 1.8M pages. The distribution of TLDs in the outer aura seems to resemble the TLD distribution of the open web. Even though the regional focus of the

Table 3.11: TLD distribution

inner aura	count	%	outer aura	count	%
1 nl	5,268,772	95.7	1 com	1,803,106	34.7
2 com	130,465	2.4	2 nl	1,613,739	31.1
3 org	52,309	1.0	3 jp	941,045	18.1
4 net	44,348	0.8	4 org	243,947	4.7
5 int	8,127	0.2	5 net	99,378	1.9
6 *other*	1,954	0.1	6 eu	80,417	1.6
			7 uk	58,228	1.1
			8 de	44,564	0.9
			9 be	43,609	0.8
			10 edu	29,958	0.6

selection policy of the Dutch web archive is apparent in the distribution of the top 10, the comparison does provide indications that the outer aura is more comparable to the full web. The prominence of the .jp TLD can be explained by the fact that some Japanese social networks are included in the unintentionally harvested pages of the Dutch archive.

Another way to characterize the unarchived contents of the Dutch web is by studying the distribution of the target domain names. This distribution is quite distinct in the two subsets of the aura: while the inner aura contains many specific Dutch domains, as selected by the KB (e.g. `noord-hollandsarchief.nl` and `archievenwo2.nl`), the outer aura contains a much more varied selection of sites, which include both popular international and Dutch domains (e.g. `facebook.com` and `hyves.nl`), and very specific Dutch sites potentially related to Dutch heritage (e.g `badmintoncentraal.nl`).

To get more insights into the degree of popular websites in the unarchived aura, we compare the domains occurring in the aura against publicly available statistics of websites' popularity. Alexa, a provider of free web metrics, publishes online lists of the top 500 ranking sites per country, on the basis of traffic information. Via the Internet Archive, we retrieved a contemporary Alexa top 500 list for sites in the Netherlands (specifically, `http://web.archive.org/web/20110923151640/alexa.com/topsites/countries/NL`). We counted the number of domains in Alexa's top 100 that occur in the inner and outer aura of the Dutch archive (summarized in Table 3.12). The inner aura covers 7 domains of the Alexa top 100 (including Dutch news aggregator *nu.nl* and *wikipedia.org*), while the outer aura covers as much as 90 of the top 100 Alexa domains, with a considerable number of unique target pages. For these 90 domains, we have in total 1,227,690 URL references, which is 23.65% of all unarchived URLs in the outer aura of the archive. This means that we have potentially many representations of the most popular websites in the Netherlands, even though they have

Table 3.12: Coverage of most popular Dutch domains (*Alexa position*)

inner aura	count	outer aura	count
nu.nl (*6*)	74.2K	twitter.com (*9*)	266.7K
wikipedia.org (*8*)	17.4K	facebook.com (*3*)	227.0K
blogspot.com (*15*)	3.5K	linkedin.com (*7*)	184.9K
kvk.nl (*90*)	2.2K	hyves.nl (*11*)	125.6K
anwb.nl (*83*)	1.7K	google.com (*2*)	106.4K

Table 3.13: Top 10 of inner aura UNESCO categories (*rank in archive*)

inner aura	count	%
1 History & Biography (*2*)	2,444,188	44.4
2 Art & Architecture (*1*)	609,271	11.1
3 Religion (*10*)	567,604	10.3
4 Education (*4*)	235,529	4.3
5 Political Science (*7*)	233,095	4.2
6 General (*12*)	190,727	3.5
7 Law & Government Administration (*3*)	187,719	3.4
8 Sports, Games & Leisure (*5*)	132,576	2.4
9 Technology & Industry (*9*)	114,926	2.1
10 Medicine (*8*)	108,874	2.0

not been captured in the selection-based archive itself.

In addition to the discussed Alexa rankings, the assigned UNESCO classification codes provide indications of the categories of pages in the archive. 98.39% of the pages in the inner aura can be categorized using UNESCO codes, since their parent domain is on the seedlist. The most frequently occurring classifications match the top categories in the archive: History & Biography (e.g. `noord-hollandsarchief.nl`) and Art & Architecture (e.g. `graphicdesignmuseum.nl`), as previously summarized in Table 3.6. A few categories have different positions though: the Religion (e.g. `baptisten.nl`) and General (e.g. `nu.nl`) categories are more frequent in the inner aura than in the archive, and the opposite holds true for Law & Government Administration (e.g. `denhaag.nl`).

For the domains in the outer aura, virtually no UNESCO categorizations are available (only for 0.04% of all pages), since they are outside of the scope of the selection policy in which these classifications codes are hand-assigned. Therefore, we generated a tentative estimate of the categories of target pages by counting the UNESCO categories of source hosts. Consider for example the host `onsverleden.net` ('our history'), of which the pages together receive inlinks

Table 3.14: Outer aura UNESCO categories (top 10), derived from link structure

outer aura	count	%
1 n/a	1,582,543	29.1
2 Art and Architecture	582,122	10.7
3 Education	409,761	7.5
4 General	406,011	7.5
5 History and Biography	405,577	7.5
6 Political Science	362,703	6.7
7 Law & Government Administration	360,878	6.6
8 Sociology & Statistics	292,744	5.4
9 Medicine	160,553	3.0
10 Commerce	117,580	2.2

from 14 different hosts. Eight of these hosts are categorized as 'Education', six have 'History and biography' as their category, and two are part of 'Literature and literature history'. Therefore we have an indication that the topic of the target site is likely related to education and history. This is validated by a manual check of the URL in the Internet Archive (as the URL is not available in the Dutch web archive): the site is meant for high school pupils, and contains translated historical sources. For each host in the aura, we then chose the most frequently occurring category. This resulted in the count-based ranking in Table 3.14. 29.1% of all pages in the outer aura cannot be categorized (due to e.g. inlinks from archived sites which are not on the seedlist). Of the remaining 70.9%, the pages have similar categories as the pages in the *inner aura*, but different tendencies, for instance a lower position for the *History & Biography* category (position 5 with 7.5% of all pages).

Summarizing, in this section we have quantified the size and diversity of the unarchived websites surrounding the selection-based Dutch web archive. We found it to be substantial, with almost as many references to unarchived URLs as pages in the archive. These sites complement the sites collected based on the selection policies, and provide context from the web at large, including the most popular sites in the country. The answer to our first research question is resoundingly positive: the indirect evidence of lost webpages holds the potential to significantly expand the coverage of the web archive. However, the resulting webpage representations are different in nature from the usual representations based on webpage content. We will characterize the webpage representations based on derived descriptions in the next section.

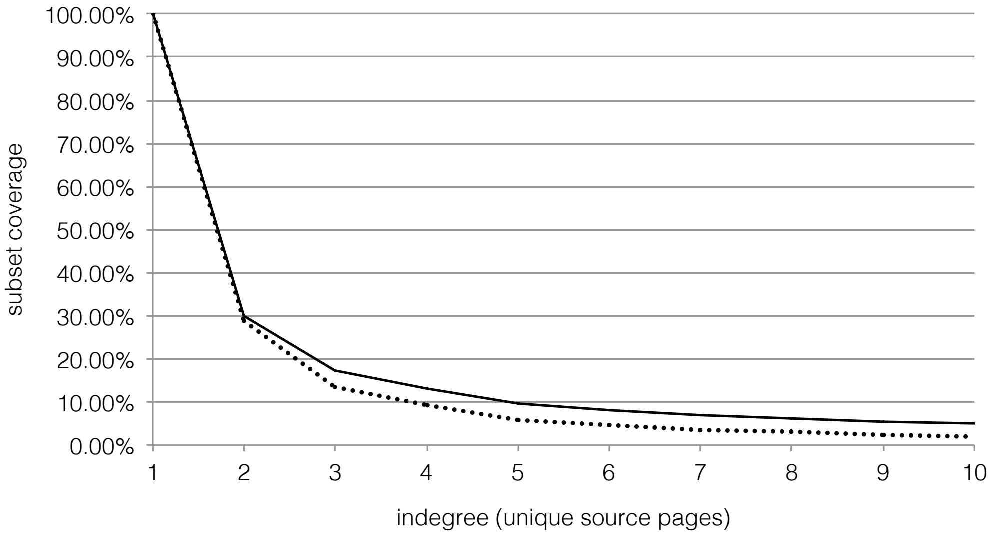

Figure 3.2: Indegree (number of incoming links from unique source pages, based on MD5 hash), compared to subset coverage (*dotted line: inner aura, solid line: outer aura*)

3.5 Representations of Unarchived Pages

In this section, we study **RQ2.2**: *How can the richness of the representations created for unarchived pages be characterized?* We build implicit representations of unarchived web pages and domains, based on link evidence, URL words and anchor text, and investigate the richness (or sparseness) of the resulting descriptions in the number of incoming links, aggregated anchor text and URL words. We break this down over unarchived home pages and other pages.

3.5.1 Indegree

In general, a representation which we can generate for a target page may be richer if it includes anchor text contributed from a wider range of source sites, i.e. has a higher indegree. Therefore, we look at the number of incoming links for each target URL of the uncovered archive in Figure 3.2. It reflects a highly skewed distribution: all target representations in the outer aura have at least 1 incoming link, 18% of the collection of target URLs has at least 3 incoming links, and 10% has 5 links or more. The pages in the inner aura have a lower number of incoming links than the pages in the outer aura. To check whether this is related to a higher number of intra-server (internal site) links, we also assessed the types of incoming links.

We differentiate between two link types that can be extracted from archived web content: *intra-server* links, pointing to the pages in the same domain of

Table 3.15: Link types of unarchived URLs

	inner aura	%	outer aura	%
intra-server	5,198,479	94.4	2,065,186	39.8
inter-server	289,412	5.3	3,098,399	59.7
both	18,084	0.4	27,930	0.5

a site, and *inter-server* links, that point to other websites. Table 3.15 shows the distribution of link types to unarchived URLs. The majority of unarchived URLs in the inner aura originate from the same source domain (i.e. a site on the seedlist), while the degree of intra-server links pointing to unarchived URLs in the outer aura is much smaller. There are very few link targets with both intra-server and inter-server link sources in the inner and outer aura.

3.5.2 Anchor Text Representations

A key influence on the utility of possible representations of unarchived pages is the richness of the contributed anchor text. In the aggregated anchor text representations, we counted the number of unique words in the anchor text. Figure 3.3 shows the number of unique words compared to subset coverage. Like the previous distribution of incoming source links, the distribution of unique anchor text is rather skewed. While 95% of all target URLs in the archive have at least 1 word describing them, 30% have at least 3 words as a combined description, and around 3% have 10 words or more (though still amounting to 322,245 unique pages). The number of unique words per target is similar for both the inner and outer aura.

3.5.3 URL Words

As the unique word count of page representations is skewed, we also looked at other sources of text. One of these potential sources are the words contained in the URL.

For instance, the URL `http://aboriginalartstore.com.au/aboriginal-art-culture/the-last-nomads.php` can be tokenized, and contains several unique words which might help to characterize the page. Specifically, we consider alphanumerical strings between 2 and 20 alphanumerical characters as words. Hence, this URL contains 10 words: ‘aboriginalartstore’, ‘com’, ‘au’, ‘aboriginal’, ‘art’, ‘culture’, ‘the’, ‘last’, ‘nomads’ and ‘php’.

Figure 3.4 indicates the number of words contributed for each subset. It shows a difference between the inner and outer aura: the pages in the inner aura have more URL words than those in the outer aura. One likely reason is that the URLs for the pages in the inner aura are longer, and therefore contribute

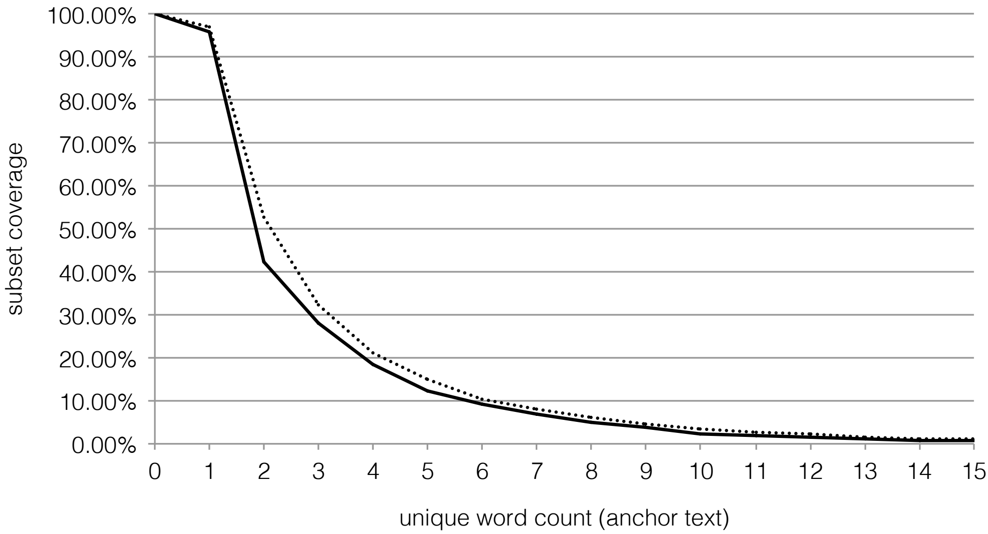

Figure 3.3: Number of unique anchor words in the anchor text representation compared to subset coverage (*dotted line: inner aura, solid line: outer aura*)

Table 3.16: Target structure distribution (*bold: most frequent slash count*)

slash count	inner aura	%	slash count	outer aura	%
0	3,765	0.1	0	324,782	6.3
1	373,070	6.8	1	921,719	17.8
2	587,416	10.7	**2**	**1,543,129**	**29.7**
3	662,573	12.0	3	535,293	10.3
4	**1,098,947**	**20.0**	4	417,361	8.1
5	535,564	9.7	5	284,237	5.5

more words. To obtain a better view of the distribution of pages at different site depths, we also looked at the slashcount of absolute URLs (see Table 3.16). This analysis shows that the pages in the outer aura are mainly located at the first levels of the site (i.e. homepage to third level). The links towards the inner aura, however, are pointing to pages that are deeper in the hierarchy, probably because 94.4% of this subset consists of intra-site link targets (links within a site).

3.5.4 Homepage Representations

As mentioned in section 3.2.2, anchors have been used for homepage finding, since links often refer to homepages. A homepage, as defined by the Merriam-Webster dictionary, is “the page typically encountered first on a Web site that

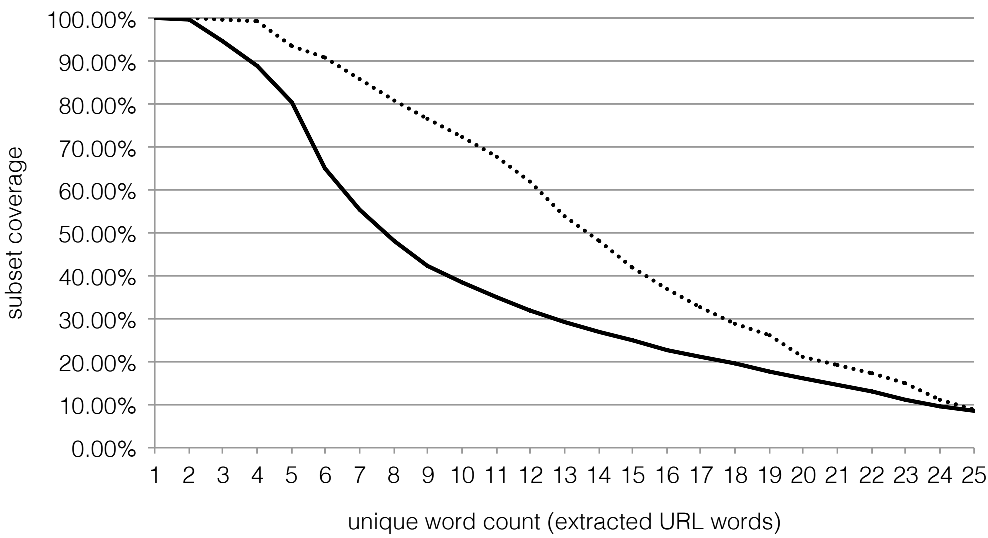

Figure 3.4: Number of unique words in the URL representation compared to subset coverage (*dotted line: inner aura, solid line: outer aura*)

usually contains links to the other pages of the site". To verify to what extent our dataset contains homepages, we looked at whether a homepage is available for each captured host in the outer aura. A basic way to calculate this is to equate homepages with the entry-level pages of hosts in the unarchived aura. Hence, we counted all unarchived target URLs consisting of only a hostname, resulting in 324,807 captured entry-level pages for the outer aura of the archive. In other words, 67.0% of all hosts have their homepage captured in our dataset. Another way is to count pages having a slashcount of 0, but also counting additional pages with higher slashcounts using manual string-based filters (e.g. URLs including '/index.html'), yielding homepages for 336,387 hosts (69.8%).

This can be important from a preservation and research perspective, since homepages are essential elements of websites and are often studied by scholars. The considerable degree of available homepages is also important for the representations that we can generate from the link evidence, because homepages tend to have a higher indegree and more available anchor text. In our dataset, this is for instance reflected in the higher average number of anchor words (2.71) for the homepages as compared to the non-homepages (2.23 unique words). Here, we looked at the specific homepages available in the dataset, but not at the lower pages in the hierarchy. In section 3.6 we look at the potential added value of aggregating link evidence from all pages under a host, creating site-level representations.

Table 3.17: Sample aggregated anchor and URL words

(A) vakcentrum [domain]	(B) nesomexico [non-domain]
vakcentrum.nl (6)	mexico (3)
detailhandel (2)	government (1)
zelfstandige (2)	overheid (1)
ondernemers (2)	mexican (1)
levensmiddelen (2)	mexicaanse (1)
brancheorganisatie (1)	beurzen (1)
httpwwwvakcentrumnl (1)	nesomexico (1)
vgl (1)	scholarship (1)
vereniging (1)	programmes (1)

3.5.5 Qualitative Analysis

Finally, we provide some concrete examples of representations that we can create for target URLs in this dataset. We first look at a homepage from our evaluation sample: `vakcentrum.nl`, a Dutch site for independent professionals in the retail sector. It has 142 inlinks from 6 unique hosts (6 different anchor text strings), resulting in 14 unique words. Table 3.17 (A) displays 9 of the unique words (excluding stopwords). They provide a basic understanding of what the site is about: a branch organization for independent retailers in the food sector.

For other non-homepage URLs it is harder to represent the contents based on the anchor text alone. Take for example `http://knack.be/nieuws/boeken/blogs/benno-barnard`, a page that is not available on the live web anymore. It only has 2 anchor text words: 'Benno' and 'Barnard'. From the URL, however, we can further characterize the page: it is related to news ('nieuws'), books ('boeken') and possibly is a blog. Hence, we have discovered a 'lost' URL, of which we can get a (basic) description by combining evidence. Other non-homepage URLs have a richer description, even if the source links only originate from 1 unique host. For example `http://nesomexico.org/dutch-students/study-in-mexico/study-grants-and-loans` is a page that is not available via the live web anymore (3 incomplete captures are located in the Internet Archive). The anchor text, originating from `utwente.nl` (a Dutch University website), has 10 unique words, contributed from 2 unique anchors. In Table 3.17 (B) the anchor words are shown. The URL words can enrich the representation, providing an indication of the page's content together with the anchor text. Of course, the richness of potential descriptions varies for each recovered target URL. The number of unique words in both anchor text and URL can serve as a basic estimate of the utility of a representation.

Summarizing, the inspection of the richness of representations of unarchived URLs indicates that the incoming links and the number of unique words in the anchor text have a highly skewed distribution: for few pages we have many descriptions which provide a reasonable number of anchors and unique terms, while the opposite holds true for the overwhelming majority of pages. The succinct representations of unarchived webpages are indeed very different in nature. The answer to our second research question is mixed. Although establishing the existence of 'lost' webpages is an important result in itself, this raises doubts whether the representations are rich enough to characterize the page's content. Therefore we investigate in the next section if aggregations of unarchived pages at the host level will improve the richness (and utility) of derived representations.

3.6 Representations of Unarchived Websites

In this section, we study **RQ2.3**: *To what extent can representations of websites be enriched by aggregating page-level evidence from pages sharing the same hostname?* We build aggregated representations of unarchived contents at the host level. We use the combined anchor text and URL words to enrich representations of unarchived websites in the outer *aura*.

3.6.1 Rationale and Method

From a perspective of preservation and research, entry pages are important pages to capture in web archives. For instance, Section 2.4 showed in a surveyed set of journal papers that researchers often study homepages. We observed in the previous section that the indegree and number of unique words for the homepages (defined as entry pages at the host level) are slightly higher than for other pages of a site. However, these succinct representations might not always be rich enough to characterize full unarchived websites. To amend this, we now focus on generating richer representations for hosts in the Dutch web archive.

We focus on the outer aura of the archive, since generating site-level representations for hosts in this subset is potentially more valuable than for the inner aura of the archive (as the main contents for the inner aura are already in the archive). We aggregate pages of unarchived websites at the *host* level (i.e. we create separate representations for `zorg.independer.nl`, `forum.independer.nl` etc.), because this results in more fine-grained representations than aggregation at the *domain* level (i.e. aggregating `*.independer.nl` under one representation).

In the previous section we found homepages for 324,807 out of 481,797 detected hosts in the outer aura of the web archive. Here, we take this a step further and create site-level representations for each host: this representation consists of aggregated evidence such as incoming URLs, incoming anchor words,

Table 3.18: Richness of host representations at the page-level (entry pages), and site-level (aggregated pages). *Table shows mean values for indegree, anchor words, URL words and combined words*

representations	count	in-degree	anchor words	URL words	combined uniq words
page-level	325K	24.14	2.71	2.32	4.56
site-level	481K	56.31	10.06	11.82	17.85

unique sources pages, and so forth. We aggregate this information for all uncovered pages under a given host. For example, a site-level representation of `3fm.nl`, a Dutch radio station, contains information from pages under `3fm.nl/nieuws`, `3fm.nl/dj`, `3fm.nl/megatop50`, plus all other child pages of the uncovered host. Taken together, these might provide a richer description of an unarchived website, reducing anchor text sparsity. In case there is no entry-level page captured in the outer aura, we still aggregate the deeper pages of a site under a certain host (e.g. `fmg.ac/projects/medlands` and other non-entry pages are aggregated under `fmg.ac`). This way, we generated 481,305 site-level representations for uncovered hosts in the outer aura of the archive.

3.6.2 Comparisons

Table 3.18 shows a comparison between page-level homepage representations, and aggregated site-level representations. The mean indegree for the site representations is more than two times higher than for the page-based representations. In addition, the mean number of anchor text words and URL words are substantially higher, as well as the combined unique words from both sources. The mean overlap between anchor text and URL words is 0.48 words for the page-based representations, and 4.03 words for the site-based representations, indicating higher similarity between anchor and URL words for the site-based representations. We now look more in detail at the involved indegree, anchor text and URL word distributions.

3.6.3 Indegree, Anchor Text and URL words

Figure 3.5 shows the indegree distribution of site-level anchor text representations in the outer aura. This indegree is based on the number of inlinks from unique source pages (based on the MD5 hash) to all pages of a given host. The graph also includes the indegree for the page-level representations of hosts in the outer aura. We see that the site-based approach results in more representations of hosts, with a higher number of unique inlinks. For example, there are 123,110 site representations (25.6%) with at least 5 inlinks, compared to only

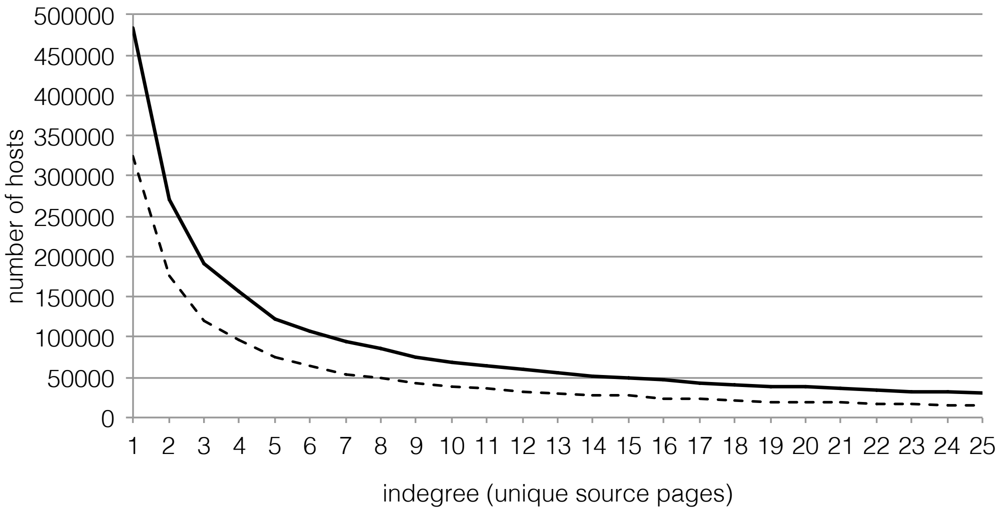

Figure 3.5: Indegree of site-level representations (*solid line*) versus page-level homepage representations (*dotted line*)

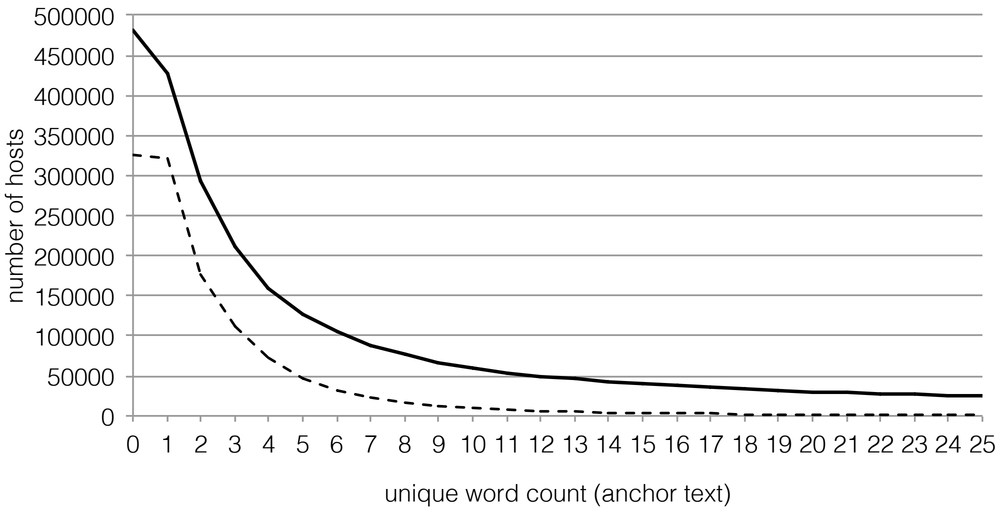

Figure 3.6: Number of unique words in the anchor text for site-level representations (*solid line*) versus page-level homepage representations (*dotted line*)

74,092 page-based representations of entry level pages with at least 5 inlinks (22.8%). The higher indegree could contribute to a richer representation, also in terms of unique anchor text words.

In Figure 3.6, the number of unique anchor text words for each site-based anchor text representation is summarized. The number of unique words is still skewed, but much richer than the page-based representations. There are 126,218 site-level representations with at least 5 anchor words (26.2% of all site-level representations), while at the page-level there is only available evidence for 46,795 entry pages of hosts with 5 anchor words or more (14.4% of all page-level homepage representations).

Figure 3.7 shows the distribution of the number of URL words for homepage and site representations. Naturally, the URL word count for the page representations is quite small: representations of entry pages usually have short URLs. For the site representations, we aggregate the tokenized words from up to 100 URLs per host. We observe the value of aggregating URL words at the host level: substantially more words are available for website representations.

The higher number of available anchor text and URL words means that generated representations are potentially richer, as more words can shed light on more facets of a site. The question is if these words also improve the potential to correctly characterize a site, since they are contributed from different pages and sections of a site.

3.6.4 Qualitative Analysis

So far, we have seen indications of the added value of site-level representations to reduce anchor text sparsity. Take for example, `webmath.com`, a website to solve math problems. For the homepage of this site, we have 1 unique incoming anchor, contributing only one unique word ("webmath"). However, as the previous sections have shown, we can also generate representations for websites by aggregating all anchor text for a given host. In the case of `webmath.com`, we then

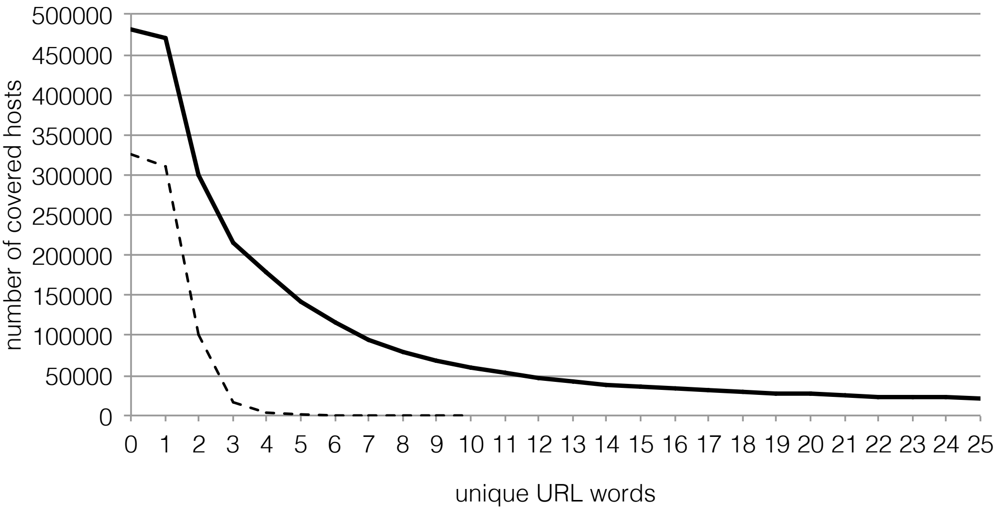

Figure 3.7: Number of unique words in the URL of site-level (*solid line*) versus page-level homepage representations (*dotted line*)

Table 3.19: Aggregated anchor words *webmath.com*

(A) page-level anchors	(B) host-level anchors	
webmath (1)	math (9)	webmath (4)
	algebra (2)	geometry (2)
	calculus (2)	web (2)
	general (2)	everyone (2)
	plots (2)	stepbystep (1)

have 10 unique incoming anchors from 4 unique source hosts. These together contribute 15 unique words, which characterize the site in a better way than the single word for the homepage (see Table 3.19). On the other hand, this does not necessarily apply to all websites: for some hosts, adding aggregated anchor text introduces noise. For example, the Centre for European Reform (`cer.org.uk`) has 5 unique anchor words, including "centre" and "reform". Aggregating anchor text evidence from all pages under this host results in 30 unique words, contributed from 17 unique hosts. These words include words useful for the description of the site, e.g. "european", and "future". However, there are also many words added to the site representation, such as "Nabucco", and "India", that might not be suitable to characterize the whole site. The question that follows from our qualitative analysis is whether the site-based representations are actually specific enough to characterize a given site's content, as noise might be added when aggregating anchor text evidence on the site level.

Summarizing, we looked at the impact of host-level aggregations of link evidence on the richness of generated representations. Our analysis showed a significant increase of unique anchor words available for each site, potentially overcoming anchor text sparsity. In our qualitative analysis we saw examples of improved representations, but also of added noise caused by the aggregation. Hence the answer to our third research question is still mixed. Although establishing the existence of 'lost' webpages and websites is an important result in itself, the resulting representations are sparse, and may not be rich enough to characterize their unique content. We investigate this in the next section.

3.7 Finding Unarchived Pages and Sites

In this section, we study **RQ2.4**: *How effective are the derived page-level and site-level representations in a known-item search setting?* We focus on the retrieval of unarchived webpages based on their derived representations in a known-item search setting and compare page-level with host-level representations.

3.7.1 Evaluation Setup

To evaluate the utility of uncovered evidence of the unarchived web, we indexed representations in the *outer aura* of the archive. Hence, we indexed the unarchived pages, detected via the archive's link structure, of which the parent domain is *not* on the seedlist. These representations consist of unarchived URLs, aggregated anchor text and URL words of unarchived pages and hosts. We indexed these documents using the Terrier 3.5 IR Platform (Ounis et al., 2006), utilizing basic stopword filtering and Porter stemming. We indexed three sets of representations:

- page-level representations for all 5.19M unarchived URLs
- page-level representations for 324,807 homepages (entry pages of hosts)
- aggregated site-level representations for 324,807 unarchived hosts

For each set of representations, we created three indices. The first index of every category uses only the aggregated anchor words (*anchT*). The second index (*urlW*) uses other evidence: the words contained in the URL. Non-alphanumerical characters were removed from the URLs and words of a length between 2 and 20 characters were indexed. Finally, the third index for each set of representations consists of both aggregated anchor text and URL words (*anchTUrlW*).

To create known-item queries, a stratified sample of the dataset was taken, consisting of 500 random non-homepage URLs, and 500 random homepages. Here, we define a non-homepage URL as having a slashcount of 1 or more, and a homepage URL as having a slashcount of 0. These URLs were checked against the Internet Archive (pages archived in 2012). If no snapshot was available in the Internet Archive (for example because of a *robots.txt* exclusion), the URL was checked against the live web. If no page evidence could be consulted, the next URL in the list was chosen, until a total of 150 queries per category was reached. The consulted pages were used by two annotators to create known-item queries. Specifically, after looking at the target page, the tab or window is closed and the topic creator writes down the query that he or she would use for refinding the target page with a standard search engine. Hence the query was based on their recollection of the page's content, and the annotators were completely unaware of the anchor text representation (derived from pages linking to the target). As it turned out, the topic creators used 5-7 words queries for both homepages and non-homepages. The set of queries by the first annotator was used for the evaluation (n=300), the set of queries by the second annotator was used to verify the results (n=100). We found that the difference between the annotators was low: the average difference in resulting MRR scores between the annotators for 100 homepage queries in all indices was 8%, and the average difference in success rate was 3%.

For the first part of our evaluation (section 3.7.2), we ran these 300 homepage and non-homepage queries against the *anchT*, *urlW* and *anchTUrlW* page-level indices created in Terrier using its default InL2 retrieval model based on DFR[1], and saved the rank of our URL in the results list. For the second part of our evaluation (3.7.3), the 150 homepage queries were ran against page-level and site-level indices of hosts in the outer aura of the archive.

To verify the utility of anchor, URL words and combined representations, we use the Mean Reciprocal Rank (MRR) for each set of queries against each respective index.

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{Q} \frac{1}{rank_i} \tag{3.1}$$

The MRR (3.1) is a statistical measure that looks at the probability of retrieving correct results. It is the average over the scores of the first correct result for each query (calculated by $\frac{1}{\text{rank}}$). We also compute the success rate at rank 10, that is, for which fraction of the topics do we actually retrieve the correct URL within the first 10 ranks.

We used unarchived pages uncovered from the Dutch web archive, that are either available in the Internet Archive, or still available on the live web, in order to have the ground truth information about the page's content. This potentially introduces bias—there can be some difference between the pages that still are active, or have been archived, and those that are not—but the URLs did not suggest striking differences. Out of all randomly chosen homepages surveyed, 79.9% were available via either the Internet Archive or the live web. However, this was not the case for the non-homepages (randomly selected pages with a slash count of 1 or more), as only 49.8% could be retrieved via the Internet Archive or the live web. The underlying reasons that many URLs could not be archived include restrictive robots.txt policies (e.g. Facebook pages), contents specifically excluded from the archive (e.g. Twitter accounts and tweets), but also links pointing to automatically generated pages (e.g. LinkedIn 'share' links). The unavailability of URLs strengthens the potential utility of generated page representations, for example via aggregated anchor text, since no page evidence can be retrieved anymore.

3.7.2 Page-based Representations

This section contains the first part of our evaluation, focusing on page-based representations of unarchived content. We use the indices of the 5.19M page

[1] Terrier DFR, `http://terrier.org/docs/v3.5/dfr_description.html`

Table 3.20: Mean Reciprocal Rank (MRR)

MRR	Queries	anchT	UrlW	anchTUrlW
homepages	150	0.327	0.317	0.489
non-homepages	150	0.254	0.384	0.457
combined	300	0.290	0.351	0.473

Table 3.21: Success rates (target page in top 10)

Success@10	Queries	anchT	UrlW	anchTUrlW
homepages	150	46.7%	39.3%	64.0%
non-homepages	150	34.7%	46.0%	55.3%
combined	300	40.7%	42.7%	59.7%

representations in the outer aura of the archive, combined with the 150 homepage and 150 non-homepage known-item queries.

MRR and Success Rate

MRR scores were calculated for the examined homepages and non-homepages to test to what extent the generated representations suffice to retrieve unarchived URLs. The final results of the evaluation based on MRR are summarized in Table 3.20. We found that the MRR scores for the homepages and non-homepages are quite similar, though some differences can be seen. Using the anchor text index, the homepages score higher than the non-homepages, possibly because of the richer representations available for these homepages. The scores for the URL words index are naturally higher for the non-homepages: they have longer URLs and therefore more words that could match the words used in the query. Finally, we can see that the combination of anchor and URL words evidence significantly boosts the retrieval effectiveness: the MRR is close to 0.5, meaning that in the average case the correct result is retrieved at the second rank.

We also examined the success rate, that is, for which degree of the topics do we actually retrieve the correct URL within the first 10 ranks? Table 3.21 shows that again there is some similarity between the homepages and non-homepages. The homepages score better using the anchor text index than the non-homepages: 46.7% can be retrieved. On the other hand, the non-homepages fare better than the homepages using the URL words: 46.0% of the non-homepages is included in the first 10 ranks. Again, we see that combining both representations results in a significant increase of the success rate: we can retrieve 64.0% of the homepages, and 55.3% of the non-homepages in the first 10 ranks.

Table 3.22: Division based on indegree of unique hosts

indegree	pages	word cnt	MRR anchT	homepage
1	251	2.9	0.29	42.6%
2	28	3.8	0.19	82.1%
3	12	4.5	0.29	100%
4+	9	7.3	0.49	88.9%

The MRR scores indicate that anchor text in combination with tokenized URL words can be discriminative enough to do known-item search: the correct results can usually be retrieved within the first ranks. Secondly, the success rates show that by combining anchor text and URL word evidence, 64% of the homepages, and 55.3% of the deeper pages can be retrieved. This provides positive evidence for the utility of these representations.

The performance on the derived representations is comparable to the performance on regular representations of web pages (Hawking and Craswell, 2005). Here we used a standard retrieval model, without including various priors tailored to the task at hand (Kraaij et al., 2002).

Impact of Indegree

We now examine the impact of the number of unique inlinks on the richness of anchor text representations at the page level. For example, the homepage Centre for European Reform (`cer.org.uk`) receives links from 3 unique hosts: `portill.nl`, `europa-nu.nl` and `media.europa-nu.nl`, together contributing 5 unique anchor words, while the page `actionaid.org/kenya` has 1 intra-server link from `actionaid.org`, contributing only 1 anchor word. For the combined 300 topics (domains and non-domains together), we calculated the mean unique word count, the MRR and the degree of homepages in the subset.

The results in Table 3.22 show that an increase in the number of inlinks from unique hosts results in a rise of the mean word count. However, it also illustrates the skewed distribution of our dataset: the majority of pages (251 out of 300) have links from only one source host, while a much smaller set (49 out of 300) have links from 2 or more unique source hosts. The table also provides evidence of the hypothesis that the homepages have more inlinks from unique hosts than non-homepages: at an indegree of 2 or more, the homepages take up more than 80% of the set of pages. We can also observe from the data that the MRR using the anchor text index in our sample is highest when having links from at least 4 unique hosts.

Table 3.23: Site representations: Mean Reciprocal Rank (MRR)

MRR	#Q	AnchT	UrlW	AnchTUrlW
page-level	150	0.435	0.393	0.590
site-level	150	0.452	0.412	0.626

Table 3.24: MRR score comparison for homepage queries in site-level (*srAnchTUrlW*) and page-level (*plAnchTUrlW*) anchor text indices

site-level	better	same	worse	*page-level*
	21	115	14	

3.7.3 Site-based Representations

The encountered importance of a higher indegree and unique word count for representing unarchived pages encouraged us to experiment with other approaches to improve representations. This second part of our evaluation compares the retrieval effectiveness of entry page representations of hosts (at the page-level) with aggregated representations of all pages under a certain host (at the site-level). We use the indices created using these two sets of 324,807 representations in combination with the 150 known-item homepage queries.

MRR and Success Rate

Table 3.23 summarizes the MRR scores for the page and site-level representations. The score for the site-level index based on anchor text is 4% higher than the page-level anchor text index. In our derived site-level representations, we aggregated up to 100 URLs and tokenized the URL words. The value of this approach is seen in the MRR score for the site-level *urlW* index: the MRR score rises with 5%. Finally, the combined representations (*AnchTUrlW*) work best, with a higher MRR rating than both *AnchT* and *UrlW* indices alone. The MRR score is improved almost 6% by aggregating the anchor text and URL word evidence, showing the value of site-level aggregation of link-based evidence.

To get a better insight into which queries perform better and worse, we look more in detail at the differences in performance for all 150 homepage queries across the page-level and site-level AnchTUrlW indices (Table 3.24). For 21 topics, the site-based representations fare better, and for 14 topics, the page-based representations fare better. Hence we see some evidence for both the improvement of site representations and for the potential introduction of noise (influencing the topics that did not perform better). Another striking observation is that the scores for 115 of the topics remain the same. The reason might

Table 3.25: Coverage of URLs by site representations and associated counts, mean number of words and MRR

covered URLs	repr cnt	page-rep words	site-rep words	page-rep MRR	site-rep MRR
1	103	4.3	4.3	0.62	0.64
2	22	4.8	7.0	0.47	0.52
3	6	7.1	10.7	0.44	0.57
4+	19	7.3	36.9	0.62	0.66

Table 3.26: Site representations: Success rates (target page in top 10)

Success@10	#Q	AnchT	UrlW	AnchTUrlW
page-level	150	56.0%	46.0%	74.7%
site-level	150	55.3%	46.7%	74.7%

be related to the skewed distribution of our dataset: for some hosts we might have few captured URLs which could be used in a site-based representation.

Therefore, we now look at the number of URLs available per host, as this might influence the richness of representations. Table 3.25 shows that for the majority of target hosts (103 out of 150 known-item queries), there is only one URL in our dataset. For 47 hosts, 2 or more URLs are available. In the table, we also see that the mean number of unique words increases when the number of covered URLs increases. For the 47 hosts with two or more covered URLs, the MRR values for the site-level representations are clearly higher than for the page-level representations.

Finally, we look at the success rates (the degree of topics with the correct URL in the first 10 results). The success rates in Table 3.26 show a different outcome than the MRR score comparison: page-level and site-level representations score remarkably similar. There is a slightly lower success rate for the site-level anchor text index, and a slight improvement for the URL words index. The similar scores might be caused by the skewed distribution of the dataset. As we have seen in Table 3.25, for 103 out of 150 hosts in our evaluation set we have just 1 captured URL. In those cases, aggregating URLs by host does not increase success rates.

Summarizing, we investigated whether the derived representations characterize the unique content of unarchived web pages in a meaningful way. We conducted a critical test cast as a known-item finding task, requiring to locate unique pages amongst millions of other pages—a true needle-in-a-haystack task. The outcome of the first part of our evaluation is clearly positive: with MRR scores of about 0.5, we find the relevant pages at the second rank on average, and for the majority of pages the relevant page is in the top 10 results. The second part of our evaluation compared page-level representations with site-level representations. We found that using site-level representations improves retrieval effectiveness for homepage queries with 4-6%, while the success rates remain stable. Hence, the answer to our fourth research question is again positive: we can reconstruct representations of unarchived web pages that characterize their content in a meaningful way.

3.8 Discussion and Conclusions

Every web crawl and web archive is highly incomplete, making the reconstruction of the lost web of crucial importance for the use of web archives and other crawled data. Researchers may take the web archive at face value, and equate it to the web as it once was, leading to potentially biased and incorrect conclusions. The main insight of this chapter is that although unarchived web pages are lost forever, they are not forgotten in the sense that the crawled pages may contain various evidence of their existence.

The aim of this chapter was to uncover and recover unarchived web content via its underlying link structure. We proposed a method for deriving representations of unarchived content, by using the evidence of the unarchived and lost web extracted from the collection of archived webpages. We used link evidence to first *uncover* target URLs outside the archive, and second to *reconstruct* basic representations of target URLs outside the archive. This evidence includes aggregated anchor text, source URLs, assigned classification codes, crawl dates, and other extractable properties. Hence, we derived representations of webpages and websites that are not archived, and which otherwise would have been lost. We tested our methods on the data of the selection-based Dutch web archive in 2012.

We investigated our first research question (**RQ2.1**) in Section 3.4: *What fraction of unarchived web pages and websites can be uncovered based on references to them in the web archive?* As a first step, we characterized the contents of the Dutch web archive, from which the representations of unarchived pages were subsequently uncovered, reconstructed and evaluated. Our findings indicate that the archive contains evidence of roughly the same number of unarchived pages as the number of unique pages included in the web archive—a dramatic increase in coverage. In terms of the number of domains and hostnames, the increase of coverage is even more dramatic, but this is partly due to the domain restrictive crawling policy of the Dutch web archive. Whereas this is still only a fraction of the total web, by using the data extracted from archived pages we reconstructed specifically those unarchived pages which once were closely interlinked with the pages in the archive.

The potential value of representations of unarchived pages led to our second research question (**RQ2.2**), described in Section 3.5: *How can the richness of the representations created for unarchived pages be characterized?* Given that the original page is lost and we rely on indirect evidence, the reconstructed pages have a sparse representation. For a small fraction of popular unarchived pages we have evidence from many links, but the richness of description is highly skewed and tapers off very quickly – we have no more than a few words. Although establishing the existence of unarchived web pages is important, it raises doubts whether the representations are rich enough to characterize the pages' content. Therefore, we next investigated a way of enriching these representations.

In Section 3.6, we looked at our third research question (**RQ2.3**): *To what extent can representations of websites be enriched by aggregating page-level evidence from pages sharing the same hostname?* Aggregating link evidence resulted in a substantial increase of unique anchor words available for each site. We saw examples of improved representations, but also of added noise due to the aggregation. This raises doubts on the utility of both page and host-level representations: are these rich enough to distinguish the unique page amongst millions of other pages?

We investigated this issue in Section 3.7, via our fourth research question (**RQ2.4**): *How effective are the derived page-level and site-level representations in a known-item search setting?* We addressed this with a critical test cast as a known item search in a refinding scenario. As it turns out, the evaluation of the unarchived pages showed that the extraction is rather robust, since both unarchived homepages and non-homepages received similar satisfactory MRR average scores: 0.47 over both types, so on average the relevant unarchived page can be found in the first ranks. Combining page-level evidence into host-level representations of websites leads to richer representations and an increase in retrieval effectiveness (an MRR of 0.63). The broad conclusion is that the derived representations are effective, and that we can dramatically increase the coverage of the web archive by our reconstruction approach.

Finally, the chapter's main research question was: *To what extent can representations of unarchived webpages and websites enhance search-based access to web archives?* Our first insight was that the representations generated for unarchived content are relatively sparse. However, the derived representations were rather effective in retrieval, suggesting the value of integrating derived representations into search systems for web archives (see Figure 3.8). By explicitly embedding unarchived content into web archive access interfaces, it is possible to address some of the **transparency** issues discussed in the previous chapter 2. In particular, search systems may be enriched with contextualization about the *coverage* of a web archive. This approach may highlight unarchived pages and sites detected from the archive's contents, and thus provide explicit insights into effects of curatorial decisions and crawler settings to a prospective researcher.

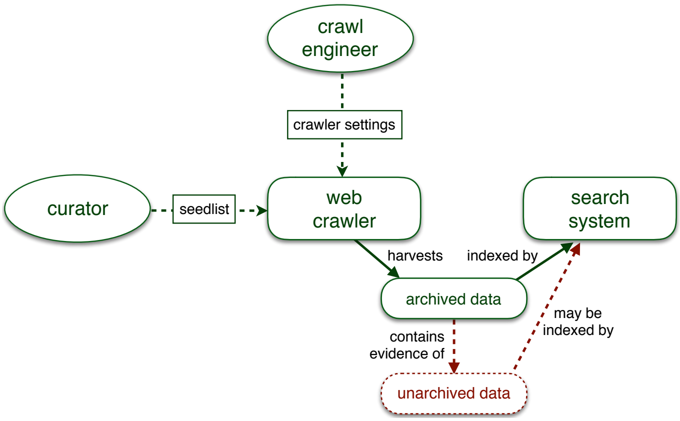

Figure 3.8: The coverage of web archive search engines may be increased by including evidence of unarchived material, derived from link structure and anchor text.

Including the representations of pages in the outer aura, for example, is of special interest as it contains evidence to the existence of top websites that are excluded from archiving, such as Facebook and Twitter. This is supported by the fact that only two years since the data was crawled, 20.1% of the found unarchived homepages and 45.4% of the non-home pages could no longer be found on the live web nor the Internet Archive.

In the context of institutional web archiving practices, the recovered pages may help curators to extend the seedlist of the crawlers of selection-based archives, as the unarchived pages outside the seedlist are potentially relevant to the archive. Additionally, missing pages may be detected for hosts which are on the seedlist. These may provide insights into the effects of crawler settings on crawled (and uncrawled) content, and thus aid crawl engineers in adjusting settings.

This directly leads to the question of how these representations of unarchived content can practically be integrated into access interfaces. A first way is to display archived and unarchived results in a *non-blended* form. A standard results list could show the archived results, while a second result list additionally shows the found unarchived results (for instance in a tab or sidebar widget). A second way is to show *blended results*, and thus combine archived and unarchived results in one ranked result list. However, the representations of unarchived web content are relatively sparse, which may lead to low rankings for unarchived representations in comparison to fully archived content. There are two standard approaches to avoid this. First, a search index for archived results could be combined with a second search index for unarchived results, and archived and unarchived results can be interleaved, either by round robin or a score weighted interleaving, ensuring highly ranked archived and unarchived results will appear together. Since the retrieval scores in the two indices may differ, an option is to apply score transformation. A second approach is the use of an anchor text index for ranking both archived and unarchived results. The use of the same type of evidence for ranking both types of content also facilitates the concurrent display of both archived and unarchived results in web archive search systems[2]. In this context, further testing of ranking and results display is necessary.

There are some limitations to the method as described in this study, that could be addressed in follow up work. The first concerns the aggregation of links by year, which may over-generalize timestamps of the unarchived pages and therefore decrease the accuracy of the representation. Further study is needed on the right window, or weighted representations, taking into account estimates of the volatility or dynamic nature of the websites and web pages at hand. Second, we used data from a selective archive, whose crawler settings privilege selected hostnames and are instructed to ignore other encountered sites. This affects the relative distribution of home pages and non-homepages, both in the archive as well as in the unarchived pages. Hence, the exact impact of the crawling strategy remains an open problem. It would be of interest to determine which crawling strategies provide the best starting point for reconstructing the associated unarchived web. Recent related work has investigated the influence of crawling strategy on topic coverage, using the depth-first KB web archive and the breadth-first CommonCrawl open web crawl (Samar et al., 2016). Third, our initial results in this chapter are based on straightforward descriptions of pure anchor text and URL components and standard ranking models. In follow up research we will examine the effect of including further contextual information, such as the text surrounding the anchors, and advanced retrieval models that optimally weight all different sources of evidence. Fourth, the recovered representations are rather skewed, hence most of the uncovered pages have a relatively sparse representation, while only a small fraction has rich representations. We addressed this by generating site-level representations. However, advanced mixture models at various levels of representation, and advanced weighting schemes treating the observed evidence as a sample from a larger population, can further enrich the representations.

After researching methods to increase the transparency of web archive search systems in this chapter, we now will focus again on *process support* in current search systems. Considering the importance of information seeking in present day's research processes, we will investigate search systems and interfaces and their role in the information seeking process in Part II of the thesis.

[2] We have successfully tested out this approach in a prototype version of WebARTist, which showed unarchived hosts in the same results list as archived results.

Part II

Supporting Information Seeking Stages

Part II: Supporting Information Seeking Stages

The art of information system design (which, I am certain, has a long future) is to find the form and timing of information presentation which will best aid the system user in whatever task he has in hand.

Oddy, Information Retrieval via Man-Machine Dialogue (1977)

Search engines on the web provide a world of information at our fingertips, and the answers to many of our common questions are just one click away. However, for the complex and multifaceted tasks involving a process of knowledge construction, various information seeking models describe an intricate set of cognitive stages. These stages influence the interplay of users feelings, thoughts and actions. Despite the evidence of the models, common search engines, nowadays the prime intermediaries between information and user, still feature a streamlined set of 'ten blue links'. While efficient for lookup tasks, this approach is not necessarily beneficial for supporting sustained information-intensive tasks and knowledge construction (Kelly et al., 2013).

In Part I of this thesis, we arrived at a need for search systems which provide better transparency indicators and process support in the context of web archive research. As information seeking is pivotal in current system-mediated research processes, we need a better understanding of the complex information seeking process, as well as the support search system features provide for it. Here, we further investigate this need in a broader web search context, and focus on the following research problem (**RP2**): *to analyze and evaluate the influence of information seeking stages on the usefulness of search system functionality in an online web search context, and to propose new approaches for supporting these stages.* First, in chapter 4, rich information seeking models are connected with potential multistage interfaces. Using a theoretical and practical analysis, we describe ways to support information seeking stages in search systems. Our findings show the issues, but also the opportunities of designing micro-level search system features supporting macro-level search stages. Building upon this work, chapter 5 zooms in on the utility of specific search user interface features and feature types in different stages, by means of a user study. The findings of this chapter show how the utility of different types of SUI features is dynamically evolving across information seeking stages.

4

From Multistage Information-Seeking Models to Multistage Search Systems

The ever expanding digital information universe makes us rely on search systems to sift through immense amounts of data to satisfy our information needs. Our searches using these systems range from simple lookups to complex and multi-faceted explorations. Part I of the thesis has shown examples of the complex research tasks scholars may be engaged in, which often also involve information seeking. The current part of the thesis looks at research-based tasks performed by students. A multitude of models of the information seeking process, for example Kuhlthau's *Information Search Process* model, divide the information seeking process for complex search tasks into multiple stages. Current search systems, in contrast, still predominantly use a "one-size-fits-all" approach: one interface is used for all stages of a search, even for complex search endeavors. The main aim of this chapter is to bridge the gap between multistage information seeking models, documenting the search process on a general level, and search systems and interfaces, serving as the concrete tools to perform searches. To find ways to reduce the gap, we look at existing models of the information seeking process, at search interfaces supporting complex search tasks, and at the use of interface features over time. In this chapter, we conceptually bring together macro-level information seeking stages and micro-level search system features. We highlight the impact of search stages on the flow of interaction with user interface features, thus providing new handles for the design of multistage search systems.

This chapter is based on both Huurdeman and Kamps (2014) and Huurdeman and Kamps (2015). Acknowledgements to Vu Tran (Information Engineering Group, Univ. of Duisburg-Essen), for allowing us to analyze data collected with the ezDL interface.

4.1 Introduction

In the current, information-abundant age, search plays a pivotal role in our daily lives: we encounter, explore and acquire information via online search systems. The range of activities performed via search systems is steadily increasing, and performed interactions range from simple lookups to multifaceted and complex searches, during sessions spanning seconds, minutes, hours, or even days. The more complex interactions might include different phases, and involve learning and construction.

Models of the information seeking process, such as Kuhlthau's ISP model Kuhlthau (1991), divide the search process for complex search tasks into multiple 'stages,' which occur over a period of time. However, current search systems are mainly using a "one-size-fits-all" approach: one interface is used in all stages of the search process. Therefore, the search process of the user might not be adequately supported in the context of complex search tasks (Brusilovsky and Maybury, 2002; Hearst, 2009; Kamps, 2011; White and Roth, 2009; Ingwersen and Järvelin, 2005; Wilson et al., 2010). However, so far, information seeking literature has not been very specific in discussing concrete support for search stages, and few systems explicitly supporting multiple search stages are in existence (discussed below). Hence, the main aim of this chapter is bridge the conceptual gap between *macro-level* information seeking models and *micro-level* search systems, by means of a theoretical and practical analysis. This is investigated through the following main research question: *What are the conceptual implications of multistage information seeking models for the design of search systems?*

In this chapter, we look conceptually at different ways to divide the information seeking process in the context of complex tasks into multiple stages and at how these stages could be supported by systems. We summarize literature from different fields related to search stages and user interfaces for complex search, and discuss their implications for multistage search systems. To gain further insights on possible stage influence on search features, we look at the temporal use of interface features in search user interfaces (SUIs). Finally, we discuss the potential ways to support stages in a system.

This chapter investigates the following research questions:

RQ3.1 What are the conceptual implications of multistage information seeking models for the design of search systems?

Using available literature, we introduce different types of information seeking and information literacy process models characterizing the search process over time, and discuss the implications of distinguishing information seeking stages for search systems.

RQ3.2 How do current search user interfaces support the information seeking process in the context of complex tasks?

We discuss user interface frameworks, various search user interface paradigms and concrete interfaces in the context of cognitively complex tasks. This will provide insights into currently employed strategies in search interfaces to support users' complex search interactions, and their broader information seeking behavior.

RQ3.3 To what extent does the search stage influence the flow of interaction at the interface level?

To investigate whether concrete search features are used differently over information seeking stages, we study existing literature and analyze the use of SUI features using eye tracking and log data collected in a previous user study, and subsequently discuss the implications for search systems.

RQ3.4 How can we reconcile multistage information seeking models and multistage search systems?

Finally, based on the results of the three previous research questions, we discuss ways to support the multistage search process in potential stage-aware search systems. We discuss stage-based adaptation and stage-based instruction, and the requirements of such approaches.

The remainder of the chapter is organized as follows: first, we look at models of the information search process, which divide the search process in different stages (Section 4.2). In Section 4.3, we look at frameworks of search user interface features and existing SUI paradigms. Section 4.4 discusses previous research that provides indications of the use of features in different stages, and newly analyzed data from a user study using a 'digital bookstore.' Section 4.5 suggests ways to reconcile information seeking and search user interface perspectives. Section 4.6 contains the discussion and conclusions of this chapter.

4.2 Multistage Information Seeking Models

In this section, we study **RQ3.1**: *What are the conceptual implications of multistage information seeking models for the design of search systems?* A very large body of work captures research on users and their search process. Here, we look at models documenting the multistage search process, mainly from the fields of Library and Information Science (LIS) and (Interactive) Information Retrieval (IIR). We discuss the models' consequences for search stages, and potential implications for multistage user interfaces.

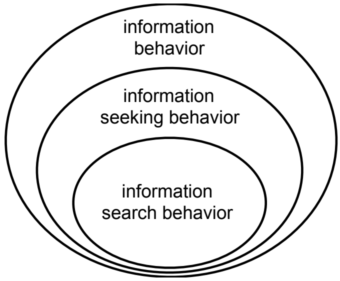

Figure 4.1: Wilson's layered model; figure adapted from Wilson (1999).

4.2.1 Information Seeking Models

Information Behavior, Seeking and Searching

Information behavior was briefly introduced in Section 1.4. It is a broad concept, defined by T. Wilson (1999) as "the totality of human behavior in relation to sources and channels of information, including both active and passive information seeking, and information use." This definition includes information seeking, but also other behavior, like the passive reception of information, or even active avoidance of information (Case, 2012). For the purposes of this chapter, we focus on information seeking, a subset of information behavior in Wilson's nested model of research areas (Wilson, 1999). It is defined by Ingwersen and Järvelin (2005) as "human information behavior dealing with searching or seeking information by means of information sources and (interactive) information retrieval systems." Information searching, in it's turn, is a subfield of information seeking in Wilson's nested model, and specifically focuses on the interaction between information user and information system (Wilson, 1999).

Information Seeking Models

The process of information seeking can be modeled in a large number of ways, depending on the used perspective, and therefore has been described in a multitude of models, "frameworks for thinking about a problem" (Wilson, 1999). In the context of information seeking, examples of models are Kuhlthau (1991)'s Information Search Process model,[1] Ellis (1989)'s behavioral model, Foster (2005)'s nonlinear model of information seeking, and Wilson (1999)'s Problem Solving Model. These models shed light on different aspects of information seeking, using different approaches. For example, Ellis' behavioral model consists of information seeking patterns, that are "not meant to indicate a fixed sequence of events," but which interact in various ways (Ellis, 1989), while Kuhlthau's pro-

[1] Despite the name, this is an information *seeking* model, as is also pointed out by Cole (2011).

cess model uses a more sequential approach, consisting of search stages. In this chapter, we have chosen to focus on the latter, temporally-based model (Beheshti et al., 2014). Hence, we use Kuhlthau's Information Search Process model, and Vakkari (2001)'s adaptation of this model as our framework, discussed below.

In this chapter, our focal point is on cognitively complex tasks, which can be carried out in work settings, but also in educational or daily life settings. In complex tasks, as Byström and Järvelin (1995) indicate, "understanding, sense-making, and problem formulation are essential, and require different types and more complex types of information." Tasks can be categorized in different ways, for example based on complexity, but also based on their specificity or nature, e.g. *exploratory* versus *lookup* tasks (Wildemuth and Freund, 2009). Employed tasks can have a considerable effect on information seeking behavior (Vakkari, 2003), and can be viewed on different levels: *search tasks* are usually contained in larger *work tasks*,[2] which in their turn are contained in a particular *environment* (Toms, 2011).[3] In this chapter, we look at search tasks and overarching work tasks. Both Kuhlthau's and Vakkari's models, discussed next, have mainly been constructed based on longitudinal examinations of particular "information-intensive, constraint-based" work tasks (Toms, 2011): the preparation of papers and research proposals by students.

Kuhlthau's & Vakkari's Models

Information Search Process Based on several longitudinal studies (e.g. Kuhlthau (1988b,a)), Carol Kuhlthau developed a multistage model of the Information Search Process (ISP) (Kuhlthau, 1991), which "depicts information seeking as a process of construction". Kuhlthau's model is descriptive, documenting "common patterns in users' experience in the process of information seeking" for complex tasks requiring construction and learning, with a discrete beginning and ending (Kuhlthau, 2005). So far, as Case (2012) indicates, the model has been predominantly applied in the context of education. In this setting, the students participating in the studies generally have a lower domain knowledge than in work tasks carried out by domain experts, potentially influencing the initial (exploration) stages of the tasks. However, similar stages were observed in studies with a securities analyst and lawyers performing complex work tasks "that require extensive construction of new knowledge" (Kuhlthau and Tama, 2001; Kuhlthau, 2004). Despite the numerous developments in information access since its conception, Kuhlthau's model can still be used to describe information seeking (Kuhlthau et al., 2008).

---

2 Here, we use Ingwersen and Järvelin (2005)'s definition of work task, which includes includes both professional (e.g. job-related) and daily life tasks.

3 Also defined as organizational, social and cultural *context* in Ingwersen and Järvelin (2005, p.261)'s general model of cognitive information seeking and retrieval.

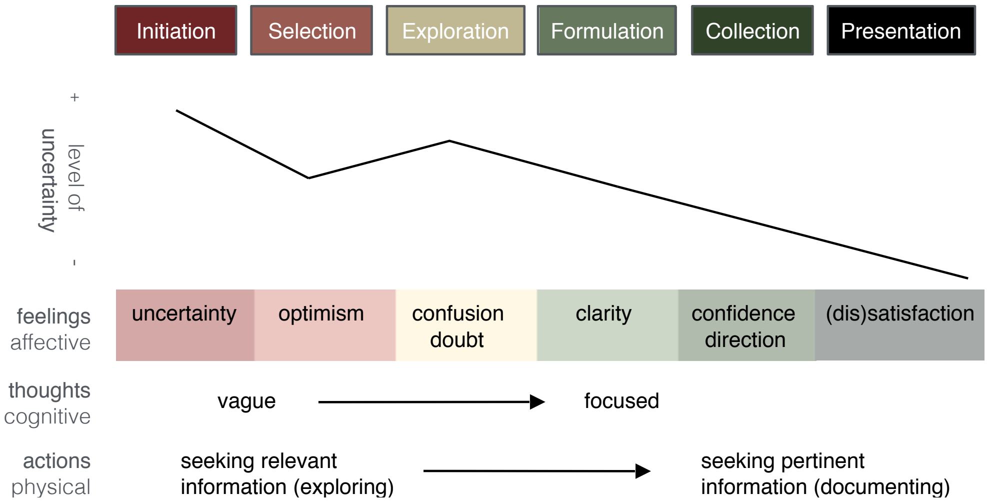

Figure 4.2: ISP Model documenting stages in tasks involving construction; figure adapted from Kuhlthau (2004, p.206).

Table 4.1: Kuhlthau's search stages, adapted from Kuhlthau (2005)

Stage	*Description*
1. Initiation	becoming aware of a lack of knowledge or understanding, often causing uncertainty
2. Selection	identifying & selecting general area, topic or problem, sense of optimism replaces uncertainty
3. Exploration	exploring & seeking information on the general topic, inconsistent info can cause uncertainty
4. Formulation	focused perspective is formed, uncertainty is reducing, while confidence increases
5. Collection	gathering pertinent information to focused topic, less uncertainty, more interest/involvement
6. Presentation	completing the search, reporting and using results

The ISP model consists of six stages: *initiation*, *selection*, *exploration*, *formulation*, *collection* and *presentation* (see Table 4.1 for details). Upon introduction, a novel aspect of this model was the inclusion of affective aspects, together with the cognitive and physical aspects (the interplay of thoughts, feelings and actions) (Kuhlthau, 2005). Uncertainty is one of the key concepts in this process, often initiating the information seeking process, and fluctuating as a person moves through different stages of their search process, encountering different kinds of information. This concept of uncertainty has been further operationalized and tested by Wilson et al. (2002).

Task-based Information Retrieval Process In 2001, Vakkari refined Kuhlthau's model in the context of Information Retrieval (IR) into a tentative theory of the task-based IR process (Vakkari, 2001), based on a longitudinal study with twelve students (Vakkari and Hakala, 2000; Vakkari, 2000b,a). Here, he refined concepts used by Kuhlthau in the context of task performance, and summarized Kuhlthau's six stages into three categories: *pre-focus* (Kuhlthau's stage 1, 2 and 3), *focus formulation* (stage 4), and *post-focus* (stage 5 and 6). Vakkari emphasizes the crucial role of finding a focus in the search process. In the pre-focus phase, thoughts are "general, fragmented and vague", and it is hard for a searcher to express concretely what information is needed. After forming a focus, the search is more directed, leading to more relevant information being sought for. Finally, in the post-focus phase, searches are more specific; this phase might also include rechecking for additional information (Vakkari and Hakala, 2000). There is a high degree of similarity between Kuhlthau's and Vakkari's findings. Subsequent testing of Vakkari's theory confirmed its validity, but also indicated that the experience of searchers should be included in the theory's scope (Vakkari et al., 2003).

Consequences of Search Stages

Given the large amount of empirical support for the models discussed in the previous section, clear indications exist that there are different stages in the search process of users in the context of complex tasks. A key question is whether there are also differences in the interaction with (interactive) information retrieval systems in these stages. While Kuhlthau looked less at the implications of the search stages on IR systems, Vakkari studied some of these effects of task stages. He demonstrated that the *information sought for*, the *relevance* and the *search tactics, terms and operators* varied during different stages.

Information sought In Kuhlthau's model, information sought, in the context of a term paper assignment, converges from *general* (background) information in the first stages, to *specific* (relevant) information in the middle and to *pertinent* information (related to the focused topic) in the final stages (Kuhlthau, 2004). A user encounters high uniqueness (new information) and low redundancy (familiar

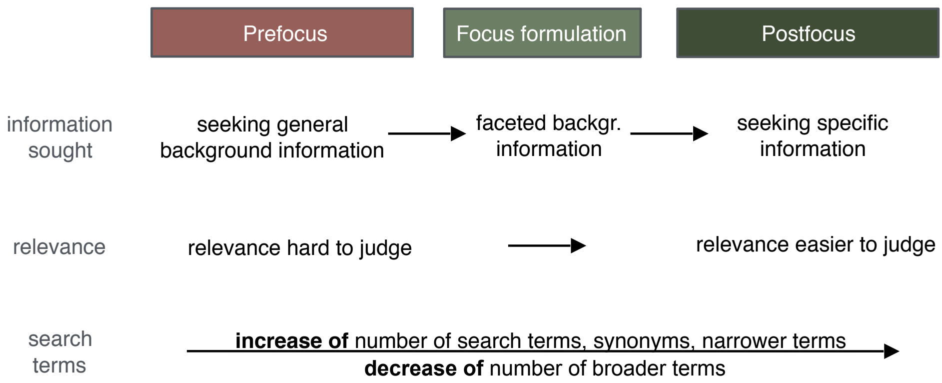

Figure 4.3: Effects of search stages; diagram summarizes findings Vakkari (2001).

information) of found documents in the beginning, while the final stages are characterized by the opposite: low uniqueness and high redundancy.

Vakkari provides a more precise, and slightly altered conception of information looked for in different stages: in the beginning (pre-focus stage) users mostly search for *general background information*, including models and conceptualizations of the topic. In the focus formation stage, the information sought is mostly *faceted background information* (broad sub-fields of the topic), and texts with methodological advice and examples about cases. Finally, in the post-focus stage participants in the study were searching mainly for *specific information* (Vakkari, 2001). Hence, as previously observed by Kuhlthau, the information sought for evolves during different stages.

Relevance Relevance, first defined in Section 1.4, is a key aspect in the context of information seeking and retrieval research. In the context of information seeking, cognitive actors have to assess the "topicality, pertinence, usefulness or utility" of information sources (Ingwersen and Järvelin, 2005, p.21), and this assessment may vary over time. Saracevic (1996) has outlined various categories of relevance manifestations, which include system (algorithmic) relevance, topical (subject) relevance and situational relevance (utility).

One of the factors influencing relevance judgements by users is their stage of search. As Vakkari (2000b) indicates, in the beginning of the search process, the ability of the study's participants to differentiate between relevant and irrelevant material is low, due to unstructured mental representations of the topic. As evidenced by other studies as well (Vakkari and Hakala, 2000; Pharo and Nordlie, 2012; Spink et al., 1998; Vakkari et al., 2003), if users know less of the topic, they are uncertain if a source is relevant or not, and they judge more documents as partially relevant.

Other research has looked at the categories of relevance criteria used in different stages of the search process. Vakkari studied the relevance criteria in different search stages, of which 'topicality' and 'interest' evolved considerably (Vakkari and Hakala, 2000). In a larger study, Arthur Taylor et al. did not find significant differences for 'interest', but 'specificity' and 'source novelty' did vary significantly during different stages (Taylor et al., 2007). Hence, the categories of relevance criteria are dynamic, and evolving through the various stages. The studies also show that the notion of relevance is quite complex, and the different settings of the studies make them hard to compare.

Search tactics, terms and operators In both models, an assessment is made of the search terms, operators and tactics used by searchers. Kuhlthau indicates that the "searcher's ability to express precisely what information is needed grows", while the "degree of efficient and effective interaction between the system and the user increases", without going into specific details. More concretely, Vakkari (2001) observes that the number of search terms used increases, and the number of synonyms, narrower terms and related terms increases, while the number of broader terms decreases. In this study, he concludes that the searchers are using a "larger and more specific vocabulary" in successive searches, coupled with an increased usage of operators.

A search tactic, as defined by Bates (1979), is a "move made to further a search." In Kuhlthau's model, search tactics could be classified as browsing and querying (Vakkari, 2001), while Vakkari, in his theory, makes use of a much larger classification, consisting of 12 tactics; in the study sample of eleven students in Information Studies, evidence was found that the used tactics evolved during the different stages. Another study by Vakkari et al. (2003), in the context of psychology students, however, did not show all of these tendencies (which might also be caused by other experimental factors and the participants' search experience).

4.2.2 Implications for Multistage Interfaces

This section has focused on the conceptual implications of multistage information seeking models for search systems. In terms of impact, Kuhlthau (1999) puts forward that her ISP research and model has had "considerable impact" on library and information services, but "little impact" on IR systems' design. Without providing clear guidelines on how to implement them, she mentions that different concepts could be used in the design of IR systems, like the *process* concept, the *uncertainty* principle, the relation between *uniqueness* and *redundancy*, the *mood* or stance of an individual in the process, user's evolving *interest*, *complexity* and the concept of *enough* in solving a problem. Further guidance in Kuhlthau's work (Kuhlthau, 2004) includes the advice for information systems to not "overwhelm the users" in the beginning: new tools provide access to a large number of sources and therefore intensified users' confusion and uncertainty. A few "well-chosen introductory pieces" might be better in the first (orientation) search stages.

Similarly, Vakkari (2000b) emphasizes that "more support is needed in the initial stages of a task," when students have an unstructured mental model. At this stage, users are building up their knowledge frame of a topic, which is needed before focus formulation can take place (Cole, 2011). Sources containing background information, conceptualizations and frameworks about the topic might be useful at the early stages, in addition to links to sources of general information (e.g. textbooks, encyclopedias and reviews). Several studies, by Vakkari and others have also indicated that relevance changes during different search stages. It is hard for searchers to judge relevance in the first stages, and criteria of relevance evolve, i.e. important relevance criteria in the beginning might be less important at the end of the process. Finally, evidence exist that the search tactics, the search terms and operators used evolve over time, at least for experienced searchers.

In terms of our main aim to explore ways to bridge the gap between multistage information seeking models and search systems, we observe the following. From the perspective of information seeking, Kuhlthau and Vakkari have thoroughly described and validated the multistage nature of the search process, but they have provided less handles to actually implement system support for these stages and their temporal progression. As their models describe the information seeking process more on a *macro* level, it can be hard to implement specific system and interface features guided by the models' implications at the *micro* level (Wilson, 1999). Our main conclusion in this section is that there is a good general understanding of the information seeking stages at the macro level, but that the translation into system and user interface design choices at the micro level remains unsolved. In the next section, we look at search interfaces supporting complex search tasks, and whether they incorporated the multistage information seeking process.

4.3 User Interfaces Supporting Information Seeking

In this section, we study **RQ3.2**: *How do current search user interfaces support the information seeking process in the context of complex tasks?* After studying multistage information seeking from a conceptual angle in the previous section, we now focus on search user interfaces, and interface features that can provide support in various search stages. The following sections focus on SUI frameworks, SUI interface paradigms and features, from the fields of Interactive IR and Human-Computer Interaction (HCI).

4.3.1 User Interfaces

Evidently, user interfaces play a crucial role in the interaction with search systems. Most commonly, by interacting with the interface, users specify their needs, and via the interface, users retrieve the results of their queries. In Section 4.2, we have encountered the large body of theoretical work related to information seeking. For the design of concrete search user interfaces, however crucial they are in the search process, a smaller number of general frameworks and theories exist. In general, it is no straightforward task to design an interface with a high *usability*: as Shneiderman and Pleasant (2005) argue, designing a user interface is “a complex and highly creative process that blends intuition, experience, and careful consideration of numerous technical issues”. In the past, some authors even claimed that interface design is more an “art” than a “science” (Smith and Mosier, 1986). However, the field is evolving, and gradually more structured frameworks, guidelines and design pattern libraries for search user interfaces have emerged.[4]

The constituent elements of SUIs serve as the tools for users to specify their information needs. In this context, M. Wilson (2011) has created a “starting framework for thinking about SUI designs”. It divides search user interface features into the following groups: *input features*, *control features*, *informational features* and *personalizable features* (adapted in Table 4.2). Input features make it possible for users to express what they are looking for, control features allow modifying or restricting input, informational features provide results, or information about results, and personalizable features “tailor the search experience to the searcher, either by their action or by those of other searchers” (Wilson, 2011a). Using these four groups, we can characterize SUIs on a basic level. The categorization of features is not always unambiguous: as Wilson indicates, some features have characteristics that can belong to multiple groups, for example the search box is primarily used as an *input* feature to enter keywords, but also as an *informational* feature, since it additionally informs users which query they have previously entered. Related to this work, Wilson et al. (2010) created a taxonomy of search result visualization techniques, divided by level of search support, evaluation depth and prevalence. The taxonomy showed that some search visualization methods were at the time heavily studied, but rarely used (e.g. facets), and others were heavily used, but rarely studied (e.g. tag clouds).

Another way to differentiate search systems and interfaces supporting complex tasks is the level of system involvement. Bates (1990) has looked at “the degree of user vs. system involvement in the search”. This encompasses a continuum, ranging from fully manual search activities to fully automated searches. Furthermore, she distinguishes various levels of search activities. The lower level

[4] E.g. the Endeca User Interface Design Pattern Library `http://www.oracle.com/webfolder/ux/applications/uxd/endeca/content/library/en/home.html` (accessed: 01/08/16).

Table 4.2: Framework SUI Features (adapted from (Wilson, 2011a))

Group	*Feature example*
Input	Search box, Categories, Clusters, Faceted metadata, Social metadata
Control	Related searches, Corrections, Sorting, Filters, Grouping
Informational	Results display, Text snippets, Deep links, Thumbnails, Immediate feedback, Visualizations
Personalizable	Recent searches, Item tray

activities are *moves* (simple actions) and *tactics* (one or more moves to further a search), while higher level activities include *stratagems* (a complex set of tactics and moves), and *strategies* (a plan for the entire information search).

In the next sections we study how actual search interfaces offer search features in the context of complex search settings.

4.3.2 Traditional Search

Some early command-line information retrieval systems in the 1970s were inspired by the dialogues occurring between (library) intermediary and user (Wilson, 2011b). Such dialogue-based systems, such as THOMAS (Oddy, 1977), would ask a user questions, and based on the users' answers would ideally retrieve a focused set of results, usually in the form of a number of references. As Ingwersen and Järvelin (2005) have argued, also various systems in the 1980s and early 1990s explicitly supported "all stages of task performance" (p.137): these "intelligent intermediary systems" were "to act as an intermediary between an end user and the IR mechanism - and perform similar functions as human expert intermediaries used to perform" (p.162). However, research on these intermediary systems gave way to other approaches. Later IR systems became increasingly streamlined, focusing on query formulation and results list inspection, and leave it to the user to perform the task itself. Hence, as stated by Beaulieu (2000), these IR systems may not provide a mode of interaction which is rich enough for task-sharing between user and system. This tendency can still be observed in current digital libraries (Mi and Weng, 2013), but also in the clean, general-purpose search engines like Google, Yahoo and Bing, even though novel contextualization and personalization features are increasingly utilized.

Some motivations behind the simple design are related to cognitive aspects: search tasks are usually part of larger work tasks, and the interface should distract as little as possible (Hearst, 2009). This may be related to *cognitive load theory*, which describes cognitive load as the load on working memory (Sweller et al., 1998). The working memory has a limited capacity for processing information, as opposed to to the 'effectively unlimited' long term memory, in which knowledge schemas can be stored. The act of processing and incorporating information in knowledge schemas that may be part of information-intensive work tasks is already demanding, i.e. has a high *intrinsic* cognitive load. Overly complex search interfaces may further increase *extraneous* cognitive load, and thus leave less cognitive resources available for the core task. Moreover, as Hearst (2009) indicates naturally, general-purpose search engines need to be accessible and understandable to a large audience with varying levels of system knowledge and search experience, which is not always the case, as various studies indicate (Hoelscher and Strube, 2000).

Still, results in modern online search engines such as Google and Bing are increasingly *personalized*. Personalization, in the context of Web search, has been described as "tailoring search results to an individual's interests" (Hearst, 2009). Personalization can be based on explicit preferences of a user, or based on implicit preferences (i.e. detected by the system). Search results may for example be personalized towards a user's context (for instance location and language), or based on previous interactions with a search engine (for example frequently searched topics). In experimental IR settings, further ways of personalization have been explored, for example based on potentially detectable user characteristics, such as reading proficiency (Collins-Thompson et al., 2011).

While search engines' functionality and retrieved results may be highly relevant to a searcher's query and context, they are not necessarily relevant for the searcher's stage of search. Personalization, for instance, does currently support displaying search results relevant to individual users' characteristics and preferences, but not the learning or construction occurring within a complex task.

Moreover, general search engines and their input, control and informational features are highly optimized for *lookup tasks*: retrieving a focused set of results for a specific query, but less suited to open-ended queries (Marchionini, 2006). Therefore, many authors argue for a move beyond the *lookup* paradigm, as "general-purpose systems will no longer suffice for the complex search tasks in which users engage" (White and Roth, 2009). This has led to initiatives to provide explicit support for exploratory search.

4.3.3 Exploratory Search

Exploratory search is a form of information-seeking which is complex, multifaceted and open-ended, as White and Roth (2009) indicate. They point out that exploratory search is motivated by complex information problems, poor understanding of terminology and information space structure, and often a 'desire to learn'. While traditional search usually consists mainly of *lookup* activities, exploratory search, according to Marchionini, also includes *learning*, and *investigation* activities (Marchionini, 2006). Like in Kuhlthau's model, searchers

Table 4.3: Exploratory search systems' features (adapted from White and Roth (2009)) and categorized using Wilson (2011a)

Exploratory search feature	*Category*	*Example*
Rapid query refinement	input	FilmFinder (Ahlberg and Shneiderman, 1994)
Facets/metadata-based filtering	input, control	Flamenco (Yee et al., 2003)
Leveraging context	informational	WebWatcher (Armstrong et al., 1995)
Visualizations	informational	manyEyes (Viegas et al., 2007)
Histories/workspaces/progress	personalizable	HunterGatherer (schraefel et al., 2002)
Task management	personalizable	SearchBar (Morris et al., 2008)
Learning & understanding	–	SuperBook (Egan et al., 1989)
Collaboration	–	SearchTogether (Morris and Horvitz, 2007)

experience various levels of uncertainty, and the uncertainty might subside when the process moves from exploratory browsing to focused searching (White and Roth, 2009).

Table 4.3 lists a set of features and examples proposed by White and Roth (2009), which should be supported by exploratory search systems, composed in a series of expert discussions and workshops. We may categorize the features using Wilson's SUI framework: *input* and *control* features, like support for (dynamic) queries and facets, *informational* features like visualizations, and *personalizable* features such as histories and task management. Some of the features in Table 4.3, however, are not common in search systems and not included in Wilson's feature set, like explicit features for learning, understanding and collaboration.

Most current systems only support few features of this list. For example, many library systems and online bookstores contain facets that can be used to select and filter results. Some prototypes, however, integrate more of the features mentioned above. An example of those is Golovchinsky et al's Querium (Golovchinsky et al., 2012), which includes queries, relevance feedback, facets and metadata-based result filtering, visualizations and task management. Bozzon et al. (2013) developed an exploratory search framework, *SeCoQL*, that supports Kuhlthau's stages. They interpreted Kuhlthau's stages on a process level, and mapped these to concrete/operative actions represented as a finite-state automaton (FSA). The multi-domain system explicitly supports complex and multifaceted activities, like booking a trip to a foreign city, via interconnected sets of widgets for data exploration. Their evaluation shows that the most relevant Kuhlthau stages in their system were *Initiation*, *Selection* and *Exploration*.

While it is not necessarily an element of a larger overarching search process like in Kuhlthau's model (White and Roth, 2009), there is overlap between Kuhlthau's initial stages and exploratory search: searchers are unfocused and experience various levels of uncertainty. Therefore, we argue that the act of exploratory search is similar to the initial stages of Kuhlthau's model, in particular the *Exploration* phase, and we could thus place it in the early stage of search. In effect, various system features useful for exploratory search (as exemplified in Table 4.3), could be valuable for adaptive systems supporting the full search process as well.

4.3.4 Sensemaking and Analytics

The combined process of information seeking, analysis and synthesis in the context of HCI is often described as *sensemaking*, or "the iterative process of formulating a conceptual representation from a large volume of information" (Hearst, 2009). Hence, besides information search, the analysis and synthesis steps also play an important role (Pirolli and Russell, 2011).

Sensemaking is often associated with complex, information intensive tasks, for example carried out by intelligence analysts (Pirolli, 2009), but also applies to other complex tasks. Pirolli and Card distinguish two major loops in sensemaking based on research conducted among information analysts: an *information foraging* loop, involving "processes aimed at seeking information, searching and filtering it", and a *sensemaking* loop, involving "iterative development of a mental model that best fits the evidence" (Pirolli and Card, 2005).

Like exploratory search, sensemaking can be supported in information search interfaces. Hearst (2009) discusses examples of sensemaking interfaces and their constituent elements, which include flexible arrangement and grouping of information, integrating notetaking and sketching, hypothesis formulation and collaborative search. Ideally, these elements work together and support *flow*, "a fluid and effortless move between operations such as querying, reading, saving, annotating, organizing and labeling" (Hearst and Degler, 2013)

For example, *CoSen* is a system that allows for sensemaking, by organizing retrieved information in a tree structure, showing past queries and by providing clustering tools (Qu and Furnas, 2008). *Sandbox* is a 'thinking environment' which allows for organizing results visually and facilitates hypothesis generation, aimed at information analysts (Wright et al., 2006). Finally, *CoSense* (Paul and Morris, 2009) is a system to facilitate sensemaking for collaborative search tasks on the Web. Note the overlap here with exploratory search interfaces, for which collaboration features also are suggested (see Table 4.3). Other interfaces, not necessarily categorized as sensemaking interfaces, also support analytical tasks. An example is Dunne et al's *Action Science Explorer* (ASE) (Dunne et al., 2012). This tool, intended for researchers and analysts to rapidly understand scientific paper collections, integrates search, statistics, text analytics and visualizations.

The sensemaking and analytical interfaces discussed in this section potentially cover a wider range of search stages, as compared to exploratory search systems. Many of these interfaces are largely aimed at researchers and information analysts. Besides traditional query and results they may offer additional features to 'make sense' of encountered materials, in order to analyze, organize, synthesize and collaborate. This means that some of these systems conceptually support the intermediate and final stages of Kuhlthau's and Vakkari's models.

Consider, for example, Kuhlthau's *Formulation*, *Collection* and *Presentation* stages which, according to Kuhlthau, involve processes similar to hypothesis generation, data collection, information organization and the preparation of a "personalized synthesis of the topic" (Kuhlthau, 2004, p.194). Hence, there is overlap between the progression of Kuhlthau's search stages, and the "flow" of sensemaking systems.

4.3.5 Implications for Multistage Interfaces

After taking the broad perspective of information-seeking models and the implications of search stages at the *macro* level in section 4.2, we here took a different perspective and looked at the support for information seeking stages in actual search user user interfaces. While a large number of interfaces support information search using *micro*-level UI features, we encountered less examples of interfaces explicitly supporting the macro stages of the higher-level information seeking process. Conceptually, however, elements of exploratory search, could fit in the early stages of Kuhlthau's and Vakkari's model, including a move from exploratory browsing (pre-focus) to focused searching (formulation). The concept of sensemaking has a relationship with the intermediate and later stages of the models. Hence, certain search features might be useful for initial exploratory stages of more cognitively demanding tasks, or could help "making sense" of data and information, even though we observed overlap between described sensemaking and exploratory search features.

Despite the large number of prospective features to aid users in their complex searches, popular search engines usually present a streamlined experience to users, providing only the most essential interface features. Other initiatives take a different approach by combining many features into one search interface (e.g. Querium (Golovchinsky et al., 2012) and ASE (Dunne et al., 2012)), often focused on experts and researchers. They provide users with advanced functionality, which can be integrated in their workflow and used as a *tool* (Bederson, 2004), but might involve a steeper learning curve: a result of the multitude of on-screen features is that user interfaces become more complex. Possible drawbacks of this approach include the large screen space needed, and increased perceptual and cognitive load (Dunne et al., 2012). As Diriye et al. (2010) indicate, excessive search features might even impede information seeking.

In terms of our main aim to bridge the gap between multistage information seeking models and search systems, we observe the following. From the perspective of user interface design, no matter how helpful features can be on an atomic level, it is no straightforward task to integrate advanced exploratory and sensemaking features into one interface, of which the design in itself is already complicated. This is related to the somewhat evasive concept of "flow" (Bederson, 2004; Shneiderman and Pleasant, 2005; Hearst, 2009). As Shneiderman and Pleasant (2005) point out, "creating an environment in which tasks are carried out almost effortlessly and users are 'in the flow' requires a great deal of hard work by the designer." Our main conclusion in this section is that there is a good understanding of search user interface features at the *micro* level, but that our general understanding of behavior at the *macro* level is fragmented at best—a completely opposite conclusion from the previous section.

This immediately suggests ways of connecting and reconciling these two views: what if we use the understanding of information seeking models at the macro level as a guide for understanding the flow of interaction at the micro level? In the previous section, we saw that search stages in complex search tasks have effects on factors such as types of information sought, relevance and search tactics. Based on the occurrence of these effects, we hypothesize that also the flow of users' atomic actions in search user interfaces at the micro level is influenced by search stages at the macro level. To shed more light on this hypothesis, we now will investigate whether search features are actually used differently in distinct search stages.

4.4 Interface Features and Search Stage

In this section, we combine the perspectives of information seeking models and search user interfaces explored in the previous two sections. We study **RQ3.3**: *To what extent does the search stage influence the flow of interaction at the interface level?* To do so, we explore SUI feature use over time. We look if we can find indications of a connection between the information seeking stages at the macro level, and at the interaction flow of feature use at the micro level. Different usage patterns of search features might occur at different moments of a complex search task. We first look at existing literature that has tracked the use of search features over time, and secondly perform a small-scale analysis of data from a previous user study. Finally, we discuss the implications for multistage interfaces.

4.4.1 Interface Features & Search Stage

Some previous studies have focused on the use of search system features over time, usually based on analysis of system log and questionnaire data. An example of previously studied functionality is Relevance Feedback (RF). RF "modifies an existing query based on available relevance judgments for previously retrieved documents" (Koenemann and Belkin, 1996). This can be done explicitly, for instance by users of a system judging the relevance of retrieved items, but also implicitly, based on implicit indicators of relevance (e.g. the time a user spends in viewing different documents). White et al. (2005) conducted a study among "a mixture of students, researchers, academic staff and others". The results of the study indicate that implicit Relevance Feedback was used in the middle of search tasks, while explicit RF was used more towards the end of search tasks. Other potentially useful features are query suggestions. Kelly (2009); Niu and Kelly (2014) looked at the use of query suggestions by undergraduate university students. Their experimental results indicate that query suggestions were used more for difficult topics, and in later search stages, potentially working in a similar vein as Bates (1979) "idea tactics", i.e. helping users to generate new ideas or solutions. They suggest that participants, in the latter parts of a task, "may be exploring the various facets of the topic and/or looking for specific information". In addition, they may have exhausted their original ideas and "need alternative queries." Niu and Kelly (2014) conclude that query suggestions "can provide support in situations where people have less search expertise, greater difficulty searching and at specific times during the search."

Another potentially valuable approach is the use of eye tracking to detect passive use of search features and other implicit indicators (see e.g. Liu et al. (2010)). In the remainder of this section, we focus on eye tracking studies in the context of search stages. Kules et al. (2009) examined searchers' interaction with faceted library catalogs in a study among undergraduate and graduate university students. A significant difference was found in the searchers' average gaze durations of the facets, query and results *Area of Interests* (AOIs) over time: in the first results page viewed, the users looked at facets, query and results about equally, while in the second and third page viewed, users significantly looked more at the results; potentially related to users extracting information from search results. In a subsequent larger study in a setting with undergraduate students, Kules and Capra (2012) looked at searchers' interactions with a faceted library catalog. Using a number of assigned exploratory search tasks, they examined differences in gaze behavior on four SUI features (query box, results, facets, and breadcrumbs). They distinguished between *query terms*, *overview*, *extracting*, *deciding next* and *deciding topic* stages, based on elements of different information seeking and searching models. In this research, again evidence was found that searchers utilize different elements of the interface at different stages of their searches. Here, the results indicate that facets play not only an important role in the initial search stages, but also in the decision making stages of the search process. According to Kules and Capra, this points to the usefulness of facets in cognitively demanding stages, similar to the use of query suggestions (Kelly, 2009). While not focusing explicitly on stages of search, a user study among undergraduate students by Capra et al. (2015) found that task-based search assistance using a feature summarizing previous search trails was not used to "get started", but in later phases of the session. They also found that users mainly utilized this feature during the more complex tasks of the study.

Diriye et al. (2013) performed an eye tracking study using an experimental interface with a rich feature set, the participants being university students. This study confirmed the finding that certain search interface features are *search stage specific* and thus useful at certain points in the information seeking process. Examples of these features are the query box and 'starter pages' (pages containing basic information about the topic), which are mainly useful in the beginning of the process. They also indicate that other features are *search stage agnostic*, i.e. useful at any stage of the information seeking process, in this case search facets and search filters. Also the tasks had an influence on feature use: in more complex tasks the number of used search support features was higher.

4.4.2 Experimental Setup

While the studies discussed in the previous section shed light on the use and usefulness of different *input*, *control* and *informational* features, we also would like to take a look at *personalizable* features used over time, and obtain a more in-depth overview of the usage patterns of SUI features. Therefore we take a tentative look at data from a user study featuring complex tasks carried out using a feature-rich search interface in the following section.

We use the dataset of a previous user study by Tran and Fuhr (2012b) which used the ezDL system, an advanced open-source IR frontend system supporting search and retrieval activities (Beckers et al., 2012), developed at the University of Duisburg-Essen. The data indexed for the experiment consisted of a collection of 2.7 million book records from Amazon, in combination with LibraryThing data (see Tran and Fuhr (2012b)). Twelve Computer Science students completed 3 tasks (2011), with a time limit of 15 minutes per task. The tasks consisted of narrow tasks, complex tasks, and a user-defined task. In this analysis, we focus on the complex tasks (specified in Table 4.4), one of which was self-selected by each participant. As can be seen in table 4.4, the simulated tasks in this experiment are information-intensive and constraint-based (i.e. the user is free to decide how to carry out the assigned task), like the paper and proposal writing tasks examined by Kuhlthau and Vakkari (see Section 4.2). In this analysis,

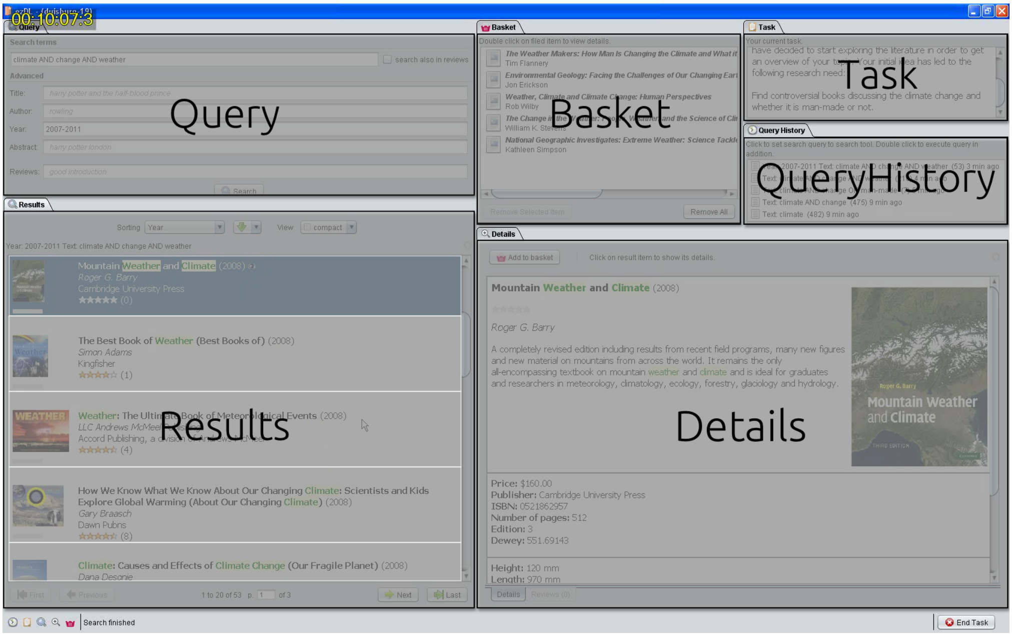

Figure 4.4: Screenshot ezDL interface with Areas of Interest

Table 4.4: ezDL tasks

You are at the early stage of working on an assignment, and have decided to start exploring the literature in order to get an overview of your topic. Your initial idea has led to the following research need:
1. Find trustworthy books discussing the conspiracy theories which developed after the 9/11 terrorist attacks in New York. 2. Find controversial books discussing the climate change and whether it is man-made or not. 3. Find highly acclaimed novels that treats issues related to racial discrimination.

we assume that searchers experience various 'mini' stages during completion of the complex search task (similar to Kelly (2009); Niu and Kelly (2014)), even though we did not perform a longitudinal study like Kuhlthau and Vakkari (which focused on a higher-level work task).

The ezDL search system has an interface with a considerable number of features, which are described in Table 4.5). The use of these features can be tracked via the system log, and the corresponding Areas of Interests (AOIs) via eye tracking. The *AOILog* software (Tran and Fuhr, 2012b,a) allows for monitoring not only static AOIs, but also dynamic AOIs, for example each specific result list item, by keeping track of position, visibility and size of all

Table 4.5: ezDL system features (using Wilson (2011a))

Category	*Feature*
Input	Search box, Social metadata
Control	Sorting, Filters
Informational	Results display, Text snippets, Images/thumbnails
Personizable	Recent searches, Item tray

Table 4.6: Mean fixation counts and percentage per phase

	Beginning	*Middle*	*End*
QueryView	68.8 (16.2%)	26.1 (5.0%)	18.9 (3.2%)
ResultView	167.3 (39.5%)	216.8 (41.4%)	253.1 (42.2%)
DetailView	178.3 (42.1%)	258.9 (49.4%)	302.1 (50.3%)
BasketView	9.2 (2.2%)	22.41 (4.3%)	26.3 (4.4%)

user interface objects.

4.4.3 Findings

To detect changes in interface use in different stages of a search session, we divide the search session for the combined tasks 1, 2 and 3 (n=12) into three parts, based on a linear approximation of search stages: *beginning* (the first 33.3% of task time), *middle* (the second 33.3%) and *end* (the last third). For example, a 15-minute search session of a certain user is divided in three parts of 5 minutes. The results in Table 4.6 include the mean fixation counts[5], and fixation count percentages per stage. The results indicate a strong decline in the views of the query AOI after the initial stage, and a gradual increase in views of results and details (book details of a selected result). In the middle and end stage, the mean fixation counts for the basket rise. The basket is a space in the ezDL interface where encountered books can be stored (analogous to a *shopping basket* in e-commerce sites).

An ANOVA analysis shows that the changes in the fixation counts for the query AOI over time are significant at $p<0.01$. A pairwise comparison for the query AOI shows a significant difference for the *beginning* stage (compared to the *middle* and *end* stage). The other AOIs do not show significant differences over time.

To get a more detailed overview of the changes in SUI use over time, figure 4.5 shows a more detailed distribution of the use of interface features during task 1

[5] In our study, we used a minimum fixation length of 80ms.

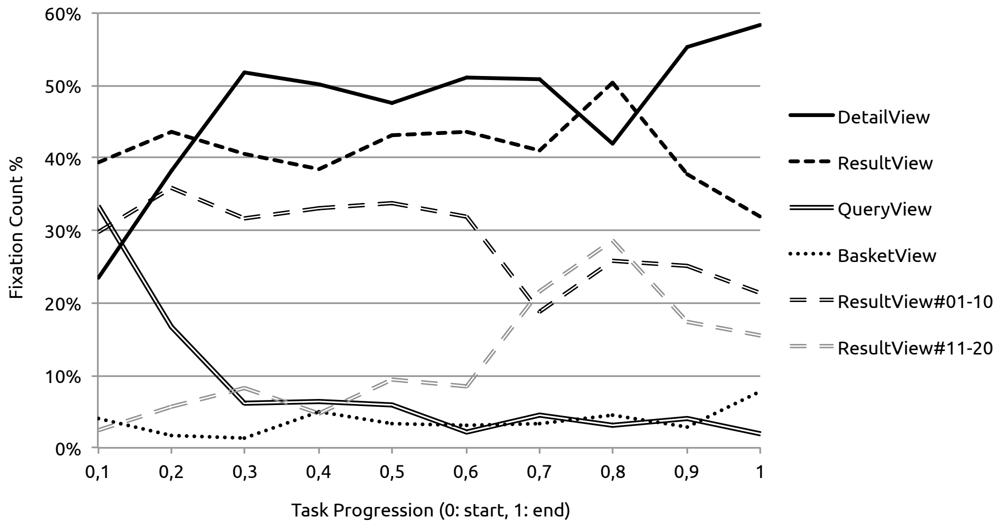

Figure 4.5: Eye tracking log complex Task 1-3 (n=12)

to 3 over time[6], including an indication of the depth of result list items inspected. The horizontal axis represents task progression, 0 being the start of the task, and 1 the end of the task (mean task time: 11.4 minutes). The initial stages of the task are characterized by a predominant focus on the query (*QueryView*), while this quickly transforms into increased results list (*ResultView*) and subsequent item detail inspections (*DetailView*). We can also observe a change in the inspection of results items: while in the beginning, predominantly the top items are inspected, in the intermediate and later phases of the task also the lower items in the results lists are viewed. Finally, in the last 20% of the task, the main focus lies on the item details again, with a rise in fixations on the basket (*BasketView*) as well.

The aforementioned basket in the ezDL interface plays an essential role to gather materials relevant to the task goals. Figure 4.6 visualizes the role of the basket during task completion (based on the systems logs). Here, we focus on complex task 1, performed by the highest number of participants (n=5). If we focus on basket modifications and item count as an indicator of task progress, we first observe few additions to the basket in the initial phase of the task. Second, participants seem to find relatively many items, and modify the basket contents. This is followed by a third phase in which partipants appear to find few things. The fourth and final phase sees participants ‘managing’ the items in the basket, which matches with the increase in eye fixations on the basket previously observed for the three combined tasks. Hence, there are indications

[6] An analysis using mean fixation time yielded similar results, but is not included here for brevity.

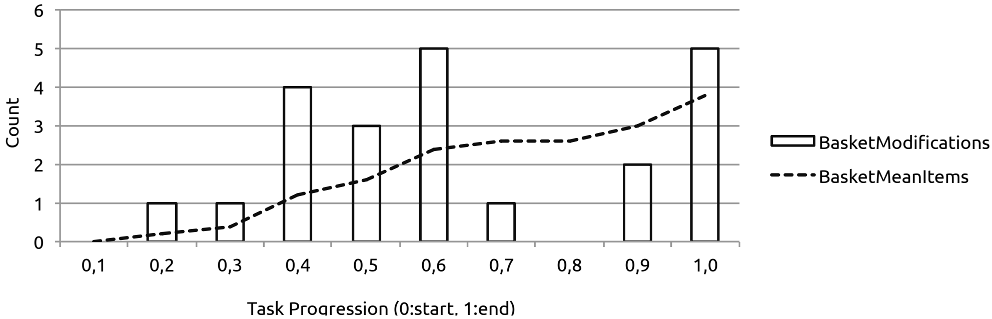

Figure 4.6: Basket modifications (bars) and mean number of items (dotted line) - Complex task 1 (n=5)

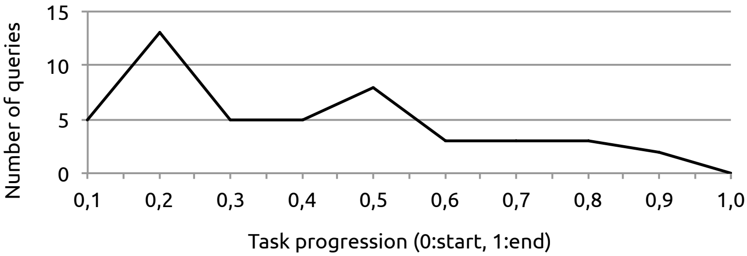

Figure 4.7: Total number of issued queries during task progression - Complex task 1 (n=5)

for variations in various phases of the surveyed task.

Finally, Figure 4.7 shows the total number of queries over time for task 1, and shows akin to Figure 4.5 that users issue a decreasing number of queries over time, and thus seem to be less focused on the query in later task phases.

4.4.4 Implications for Multistage Interfaces

This section investigated the influence of search stage on the flow of interaction with SUI features. While there are limits to the number of previous studies, and to the size of the dataset analyzed here, patterns in the use of search user interface features for the three complex tasks involved could be observed. In our analysis, we found differences in the interaction flow with SUI features at an early and late stage of search. The initial stage is characterized by a significant focus on the query (an *input* feature), followed by increased results and detail inspections (*informational* features). The final stage features slight changes in the focus on results and details features, also in terms of the depth of inspected result list items. While most of these findings are in line with previous literature (Kules et al., 2009; Kules and Capra, 2012; Diriye et al., 2013), we also found variations in the use of *personalizable* features (features relating to previous interactions). The basket is not used immediately, but starts to be used after the initial phase of the task. While this tendency might be a straightforward observation (a user first has to formulate a query and obtain a decent result set before gathering elements relevant to the task), glimpses of variations of use in the intermediate stage of the task, and a spike of increased usage near the end of the task could be observed: possible evidence of users inspecting and reviewing the collected basket items in the final stage of their search. Focusing on the basket modifications as indicator of task progress, we see support in our data for a final stage with characteristics of a post-focus stage.

In terms of our main aim to investigate the gap between multistage information seeking models and search systems, we observe the following. There is no clear dichotomy between the stages: feature use changes only gradually over time, especially the use of essential *informational* features like results lists. Hence, features might be useful at different stages, meaning that some features cannot easily be left out in a multistage interface. However, the results also indicate that some *input* and *personalizable* features are, indeed, search stage sensitive and could be offered at the moment they are needed, or gradually adapt themselves to different search stages (e.g. show different amounts of details), thus assisting the user and potentially reducing cognitive load. Our main conclusion in this section is that we see differences in the flow of interaction between the information seeking stages for some of the user interface features, supporting our hypothesis that the flow of users' atomic actions in search user interfaces at the micro level is influenced by search stages at the macro level.

While we did not focus on longitudinal tasks as in the case of Kuhlthau and Vakkari, our findings showed evidence for behavior changes in different phases of relatively short search episodes. These insights may be useful for constructing future support for complex search tasks: acknowledging the dynamic nature of complex search tasks suggests that interfaces and systems could be more dynamic as well. We further look at this possibility in the next section.

4.5 Reconciling Perspectives – Towards Stage-Aware Systems

This section looks at the following question (**RQ3.4**): *How can we reconcile multistage information seeking models and multistage search systems?* Based on the findings of previous sections, we suggest ways to increase task-sharing between searcher and system. We introduce the concept of adaptive, stage-aware systems, incorporating elements of *macro-level* information seeking and problem-solving stages. Thus, we explore the idea posed by T. Wilson (1999) to use aspects of models of information-seeking behavior "to inform the general design principles of such systems".

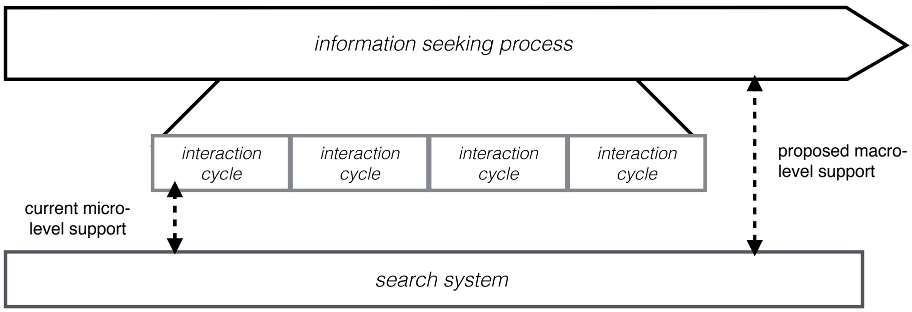

Figure 4.8: Micro and macro-level support

4.5.1 Designing Stage-Aware Search Systems

We define a stage-aware system as a potential tool supporting not just *micro*-level cycles of interactions with search systems, but also providing support for *macro*-level information seeking stages (see figure 4.8). According to Kuhlthau (2004), there are different timepoints in which instructors could intervene, for example at moments of increased uncertainty. In these "Zones of Intervention", guidance and assistance may help users to accomplish what they cannot do on their own (p.129). We could extend this view to the search system, and potentially offer different levels of support and assistance in different stages, by means of adapted *functionality and content*, and by introducing explicit *guidance and reflection*.

Stage-based Adaptation of Functionality and Content

Stage-aware search interfaces could support a searcher's process by adaptively introducing *functionality* in a certain information seeking stage. In early stages of complex information-intensive tasks, searchers may have a limited domain knowledge. At this point, they may engage in information seeking "hoping to resolve some problem, or achieve some goal, for which their current state of knowledge is inadequate" (Belkin, 2000). Searchers, as Vakkari (2001) has stated, "need support to expand and differentiate their conceptual model of the topic". However, at this point, searchers "may not be able to specify the salient characteristics of potentially useful information objects" (Belkin, 2000).[7] Furthermore, as described in Section 4.2, users may have trouble judging the relevance of information items (Vakkari, 2001). Hence, in early stages, with higher

[7] This is related to the ASK hypothesis, which Belkin et al. (1982) defined as follows: "The ASK hypothesis is that an information need arises from a recognized anomaly in the user's state of knowledge concerning some topic or situation and that, in general, the user is unable to specify precisely what is needed to resolve that anomaly."

levels of uncertainty, more potentially assistive features could be included, inspired by features useful for exploratory search (as discussed in Section 4.3.3). These features may help searchers to formulate queries, to generate initial ideas (Kelly, 2009), and to extend their initial knowledge. Concrete features potentially offering these types of support may are for instance query suggestions, keyword clouds and result item recommendations. The type and quality of assistance requires thoughtful consideration. For instance, the provision of unexpected suggested terms may spark serendipitous ideas (Foster and Ford, 2003), but may just as well cause tension and distract searchers from their core task (Belkin, 2000).

Returning to Kuhlthau's ISP model, a careful balance of support is needed. First of all, interventions may be intrusive and have adverse effects if an individual is "self-sufficient" (Kuhlthau, 2004, p.128). During later stages, as Kuhlthau indicates, users are increasingly able to specify their needs, and to perform comprehensive searches. Hence, in those stages, less assistance may be needed, but search interfaces could for instance provide functionality for categorizing and organizing encountered results. The design of these features could be inspired by common approaches to support sensemaking, as discussed in Section 4.3.4, such as notetaking tools. Hence, interventions, such as SUI features which appear at the time a user needs them, should aid users in the "accomplishment of their task" (Kuhlthau, 2004, p.129), but careful balancing is required to support and not inhibit a user's performance, especially in the crucial early stages of a user's complex task.

A second way to adapt search systems to the various information seeking stages is at the *content* level. This could be achieved by differentiating the ranking of retrieved search results (or by selectively showing results). From Kuhlthau's model we can derive the importance of showing introductory sources in the early stages, and the idea to not 'overwhelm' the users. This could be performed by ranking introductory sources highly in the beginning of the process, while systems could show specific and in-depth sources (pertinent to the focused topic) more prominently in later stages of the process. Hence, such a system may rank sources highly which are relevant to a user's current stage in the *process*, not just relevant to a user's query. Thus it may provide information items which have a high cognitive relevance, supporting a users current "cognitive state and processes" (Saracevic, 1996), as well as items with a high situational relevance, related to the task stage at hand. Examining adaptive ranking is beyond the scope of this thesis, but would be worthwhile to consider in further research.

Stage-based Guidance

Reflection is a core element of tasks involving construction and learning. Kuhlthau's ISP model, but also various information literacy models (e.g. Eisenberg and

Berkowitz (1990)'s BigSix model and Stripling and Pitts (1988)'s Research Process model), suggest the positive influence of 'being aware' of one's own information seeking process; for instance by encouraging reflection on the material. A system which aids a user in distinguishing their stage, but also provides search-stage specific guidance, thus may be helpful. This can be backed up by literature related to information seeking and retrieval: experimental results Moraveji et al. (2011) have shown that including search tips can have beneficial effects on search skills, even after their experiment finished. Another example is embodied by Bateman et al. (2012)'s 'search dashboard', which could be used for "reflection on personal behavior" (e.g. summarizing search techniques and topics sought for). Moreover, we may find connection points to the concept of *scaffolding* from educational psychology (Rosenshine and Meister, 1992), in which instructional support during the learning process is gradually removed. Likewise, certain assistive features could be only offered in the early stages of a complex task, but gradually fade away to avoid potentially countereffective effects in later stages.

Substantial search effort may be needed for positive outcomes of complex tasks involving learning. Vakkari and Huuskonen (2012) have shown that systems requiring more effort may lead to improved learning outcomes[8]. This leads to the observation that 'perfect' retrieval systems may not necessarily aid a user in performing a complex task, since personal interaction and reflection on material is needed. Especially in educational settings, inexperienced searchers may perform 'shallow' and short-lived searches, not leading to the formulation of focus within complex tasks (such as essay-writing). An idea tentatively introduced in Kumpulainen and Huurdeman (2015) is to 'shake up' a researcher's shallow search process, in order to enable more "deep" interactions with a user's information environment.

Search systems which encourage reflection and provide guidance may also have drawbacks. For instance, they may lead to "lockstep strategies", which could force "one specific method for problem-solving and decision-making" upon a user (Eisenberg, 2008). Therefore, stage-aware tools should allow users to flexibly switch between 'stages' and interface panels, and a user should be able to remain in control. Also, we have to bear in mind the risks of a 'tick the box' approach in the context of information literacy posed by Johnston and Webber: the idea of "reducing a complex set of skills and knowledge to small,

[8] This has similarities with past literature on adaptive hypermedia systems: students were forced to be more metacognitive with a "less structured" system. Using a highly structured system, students were less thoughtful about their choices and performed worse on an essay posttest (Shapiro, 1998). In a similar context, Shapiro and Niederhauser (2004) further indicate the following: "Rote learning is often aided by easily accessed structures that make fact retrieval simple. Deeper learning is aided by systems that promote a bit of 'intellectual wrestling.' "

discrete units" (Johnston and Webber, 2003). This implies a careful balancing of potential system guidance towards learners and searchers.

4.5.2 Requirements

An essential requirement of a potential stage-aware system is the detection of stages occurring in a user's information seeking process. We distinguish between *manual* and *automatic* approaches.

First of all, a system could rely on the input of a user to select which 'stage' of an interface to show. This approach has been followed in the *Interactive Social Book Search Track* of the CLEF conference (Gäde et al., 2015; Koolen et al., 2015). Searchers for books could manually select panels of an experimental multistage interface, representing exploration, search and review stages in the book search process. The initial outcomes of this large-scale collaborative study (192 participants) suggest that the users of the multistage interface explore more different kinds of books, and have higher levels of engagement as compared to the baseline interface (Huurdeman et al., 2015a; Gäde et al., 2015).

Second, a multistage system could rely on automatic approaches to detect a stage a user is in, instead of manual input. Considering the complex nature of learning tasks (Freund et al., 2014), this is not straightforward. To derive a user's current stage, extensive logging of a user's interaction with a system is needed. Some of these changes in behavior, as well as other learning aspects (Eickhoff et al., 2014) may be detected via the interaction with a search system.

Summarizing, this section has looked at ways to reconcile information literacy and system perspectives. While some 'intelligent' information retrieval systems from a distant past initiated a dialog with their users to perform task-sharing between user and system, current systems are predominantly focused on queries and results list inspection. To increase task-sharing between user and system, we have introduced the concept of stage-aware tools, which support stages occurring in the information seeking process. We discussed stage-based *adaptation* and stage-based *guidance*, and pinpointed some of the requirements for stage-aware systems.

4.6 Discussion and Conclusions

In this chapter, we focused on moving beyond the "one-size-fits-all" approach in search systems for complex tasks involving learning and construction of new knowledge. The main aim of the chapter was to bridge the conceptual gap between *macro-level* information seeking models and *micro-level* search systems, by means of a theoretical and practical analysis. Via a literature study, we theoretically bridged information seeking models and the design of search systems. In addition, using a practical approach, we highlighted differences in the use of search user interface features over time, detecting variations in the the use of these features, especially in in the initial and final stages of a search episode.

Section 4.2 focused on the first research question (**RQ3.1**), *What are the conceptual implications of multistage information seeking models for the design of search systems?* It looked at temporally-based information seeking models, which differentiate various search stages over time, based on empirical evidence. During these stages, the information sought, the relevance, and the search tactics and strategies evolve. Authors like Kuhlthau and Vakkari have accurately pinpointed the issue of stage-specific search support, but provide less concrete pointers to implementation in search systems and interfaces. As T. Wilson (1999) has indicated, many information seeking models focus on the *macro* level of the search process, while information system designers focus more on the *micro* level of search. However, indications for the provision of search stage support in search systems can be determined from the discussed theory, not only at the interface level (providing specific features supporting stages), but also at the content level (for example providing search stage adaptive ranking). Our main conclusion was that there is a good general understanding of the information seeking stages at the macro level, but that the translation into system and user interface design choices at the micro level remains unsolved.

To get more insights into the SUI features that could support complex, information-intensive search tasks, we investigated our second research question (**RQ3.2**): *How do current search user interfaces support the information seeking process in the context of complex tasks?* Hence, Section 4.3 focused on concrete SUI features in the context of M. Wilson (2011)'s framework for interface features. We argued that there is an abundance of interfaces which support information *search*, but few systems provide explicit support for the higher-level information *seeking* process in the context of complex tasks. However, overarching interface paradigms have similarities with temporal search stages. We showed that exploratory search, though slightly different in nature due to the open-endedness of the tasks, could fit in Kuhlthau's and Vakkari's models, in particular in the early pre-focus stages. In addition to that, elements of search paradigms like sensemaking, could fit in the more advanced stages of search of Kuhlthau and Vakkari. There is, however, no integrated system, and many authors point at the complexity to understand the impact of design choices on the overall usability, and the complexity of creating a seamless and effortless flow of interaction (Bederson, 2004; Shneiderman and Pleasant, 2005; Hearst, 2009). Our main conclusion was that there is a good understanding of search user interface features at the micro level, but that our general understanding of behavior at the macro level is fragmented at best. This immediately suggested ways of connecting and reconciling these two views: what if we use the understanding of information seeking models at the macro level as a model to understand the flow of interaction at the micro level?

Section 4.4 focused on the third research question (**RQ3.3**): *To what extent does the search stage influence the flow of interaction at the interface level?* We looked at the influence of search stage on the flow of interaction, and we observed different use of features over time, based on previous literature and an analysis of eye tracking and system data from a small-scale user study. Here, we defined 'stage' as a temporal segment of a search session. Some *informational* features (results lists and details) are generally used in all stages of the search, albeit in different depths, and therefore could be considered stage insensitive. However, the use of a subset of search features varied over time, like the gaze towards the query box (an *input* feature), and the use of the basket (a *personalizable* feature). Especially, we saw variations in the use of interface features in the beginning and end of a complex search task. This provides initial indications of different usage patterns of search user interface features in different search stages, which could be informative for the design of search systems.

Our main conclusion was that we see differences in the flow of interaction between the information seeking stages for some of the user interface features, providing support for our hypothesis that the flow of users' atomic actions in search user interfaces at the micro level is influenced by search stages at the macro level. Hence, even though we performed a small-scale study with relatively short tasks, we observe some elements of the changes documented in the longitudinal models devised by Kuhlthau and Vakkari. The observation that elements of stages may also occur in shorter (albeit complex) tasks, may be researched further in future work.

Finally, in Section 4.5, we discussed **RQ3.4**: *How can we reconcile multistage information seeking models and multistage search systems?* We suggested ways to integrate the knowledge gained from macro-level information seeking models into micro-level search user interfaces. We introduced the concept of a 'stage-aware' search systems, potential tools supporting not just micro-level cycles of interactions with search systems, but providing support for stages occurring in the macro-level process. This may be achieved by means of stage-based adaptation at the *interface* (adaptive SUI features), or *content* level (adaptive ranking and filtering). This approach is illustrated by Figure 4.9. A second suggested way to create more supportive systems is by providing stage-based guidance to non-expert searchers, thereby encouraging reflection on encountered materials.

Concluding, the aim of this chapter was to answer the following main research question (**RQ3**): *What are the conceptual implications of multistage information seeking models for the design of search systems?* Our conceptual analysis clearly revealed differences in the levels of understanding of information seeking

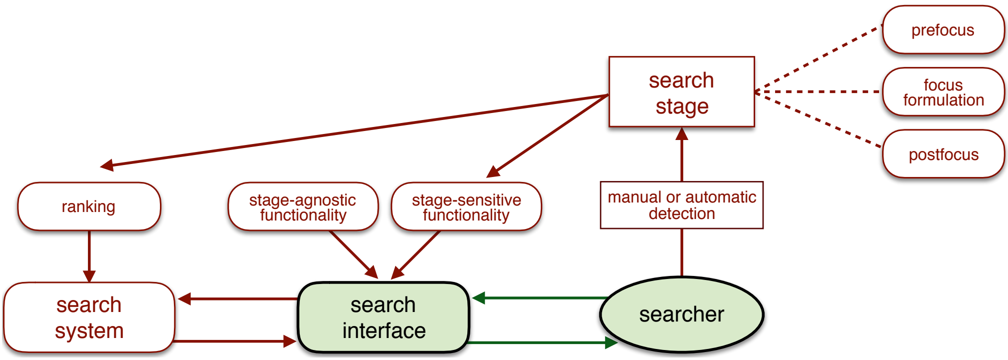

Figure 4.9: Functionality of search systems may be adapted at the system and interface level based on search stages, leading to more *dynamic* support for stages in the information seeking process.

behavior at the macro level, and of systems and interface design at the micro level. Temporally-based information seeking models document complex tasks and search stages over time. These models describe information seeking behavior on a general, or macro level, and the high-level nature of these models makes it hard to directly implement their implications, even though they might be useful in supporting users' complex tasks. User interfaces supporting search, on the other hand, are often "one-size-fits-all" interfaces, containing a streamlined set of micro level features in traditional search systems, or a larger array of features in more analytical search systems. The SUI features of these interfaces can be used at different stages of a search, but the former approach might not be ideal for complex search tasks, while the latter approach might involve a steeper learning curve.

Based on our analysis of information seeking models, search user interfaces and search feature use over time, we hypothesize that there are differences in the interaction flow of SUI feature use at the micro level, depending on the current stage of search at the macro level. Taking Vakkari's stages as an example, when a user is in the pre-focus stage, patterns of interface use and system interaction are different than in the focus formulation, or post-focus stage. This suggests interface elements which are search stage sensitive and we could customize the way search system features are shown during task progression.

While past research on Kuhlthau's and Vakkari's stages has focused on longitudinal tasks, taking weeks or even months to perform, future information seeking research should further investigate to what extent these stages also manifest themselves more short-lived (albeit appropriately complex) tasks. Moreover, at the interface level, future work still has to show whether the multistage approach can be naturally integrated in the user's flow, for different complex tasks and user contexts, without being confusing or intrusive. Multistage systems may provide new ways to reduce unnecessary *extraneous* cognitive load (as defined by Sweller et al. (1998)) by hiding superfluous interface elements, and increase *germane* cognitive load, focused on the stage of the learning task at hand. An essential aspect of multistage systems is that the user should remain in control and have the freedom to switch between interface units. Initial evidence for positive effects has been found in an evaluation of a multistage book search system (Gäde et al., 2015; Hall et al., 2014; Gäde et al., 2016; Huurdeman et al., 2015a).

Taking the information literacy perspective, one could take this even further and build prescriptive systems that actively guide searchers in their search process, in particular targeting those with poor search literacy and stimulate their critical use of information, up to the point that it changes their information behavior (Allan et al., 2012). The positive effects of information literacy interventions (e.g. Walton and Hepworth (2011)), but also of experimental search systems encouraging reflection on encountered materials and search behavior Bateman et al. (2012); Moraveji et al. (2011), suggest that increased support related to a user's *process* may have positive effects on the outcomes of learning tasks. Furthermore, including information literacy instruction into search tools used in the context of research-based tasks may encourage learners to learn by doing, and apply these skills in their later information ventures.

This chapter has provided manifold insights into the potential utility of search functionality in different information seeking stages. However, some limitations have to be addressed in future work: first of all, the concrete study in this chapter used a linear approximation of search stages, based on temporal segments of a search task. A true multistage task design may provide more insights into the use of different SUI features in different stages. Furthermore, the number of studied search features was relatively limited in this study. Hence, the exact usefulness of a broader set SUI features during complex seeking tasks, in terms of active, passive and perceived utility across stages, needs additional investigation. We perform a wider investigation in the subsequent chapter 5.

5

Active & Passive Utility of Search Interface Features in Different Information Seeking Task Stages

The previous chapter focused on bringing together macro-level information seeking models, and micro-level search systems. This chapter, on the other hand, specifically focuses on investigating the utility of search user interface (SUI) features at different macro-level stages of complex tasks. Models of information seeking, including Kuhlthau's *Information Search Process* model, describe fundamentally different macro-level *stages* during these complex tasks. 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. A user study was designed, using simulated work tasks, to explicitly place users within different stages of a complex task: pre-focus, focus, and post-focus. Active use, passive use and perceived usefulness of features were analysed in order to derive *when* search features are most useful. Our results identify significant differences in the utility of SUI features between each stage. From these findings, we conclude that features less commonly found in web search interfaces can provide value for users, without cluttering simple searches, when provided at the right times.

This chapter is an extended and revised version of Huurdeman et al. (2016). In addition to the NWO grant supporting this thesis, this research was supported by EPSRC Platform Grant EP/M000877/1.

5.1 Introduction

Research into Search User Interfaces (SUIs) (Hearst, 2009; Wilson, 2011b; Russell-Rose and Tate, 2012) has proposed many different interactive features, from search suggestions (Niu and Kelly, 2014) to facets (Tunkelang, 2009) to personal spaces to collect useful results (Donato et al., 2010). Although their usefulness has been proven in micro-level studies of complex and exploratory tasks, many of these features have not been adopted by search engines, perhaps because they can impede search during simple lookup tasks (Diriye et al., 2010). In contrast, information seeking theory (Kuhlthau, 2004; Vakkari, 2001) often highlights the existence of *stages* of search *within tasks* involving learning and construction, suggesting that we should consider *when* SUI features might be useful within tasks, rather than whether they are useful for tasks. Different categories of features, such as *input*, *control*, *informational* and *personalizable* features (Wilson, 2011b), might support users in different ways, both actively and passively. An understanding of the utility of features at different stages may help to overcome the apparent divide between the dynamic stages documented in macro-level information seeking models and the more static SUIs currently available online.

In chapter 4, we identified a need for more insights into the role of SUI functionality during complex search tasks. Thus, this work aims to understand the value of different SUI features better, through a user study using a custom search system called SearchAssist (see Section 5.3.2). This leads to the following main research question: *How can different types of search user interface features support distinct macro-level information seeking task stages?* Tasks were designed to take users through pre-focus, focus, and post-focus task stages (Vakkari, 2001) in order to gather active, passive, and subjective measures of when SUI features provide most value and support. More specifically, we have three research questions.

RQ4.1 How does the user's search stage influence active behavior at the interface level?

For the first research question, we looked at *active* behavior, the behavior which can be directly and indirectly determined from logged interaction, such as clicks and submitted queries.

RQ4.2 How does the user's search stage influence passive behavior at the interface level?

For our second research question, we looked at *passive* behavior, i.e. behavior not typically caught in interaction logs, such as eye fixations and mouse movements.

RQ4.3 How is active and passive behavior reflected in the perceived usefulness of features?

For our third question, 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.

The remainder of this chapter is structured as follows: Section 5.2 discusses related work, 5.3 details the experimental setup, 5.4 discusses the results in light of active behavior, 5.5 for the passive behavior, and 5.6 covers the perceived usefulness of features. Finally, Section 5.7 discusses the conclusions and implications of our findings.

5.2 Related Work

The previous chapter featured an in-depth analysis of multistage information seeking models (Section 4.2), and user interface support for complex tasks (Section 4.3). This section briefly discusses related work relevant to the user study in the current chapter, in the context of task-based information seeking and searching, search user interfaces, and the utility of SUI features over time.

5.2.1 Task-based Information Seeking and Searching

As Toms (2011) has indicated, the "*raison-d'être* of information retrieval systems is to deliver task-specific information that leads to problem resolution." Tasks may have different levels: a *work task* may be composed of several *search tasks*, set in a particular *environment* (Toms, 2011). Categorizations of tasks may include complexity and specificity (Wildemuth and Freund, 2009; Wildemuth et al., 2014). For instance, tasks can range from simple lookup tasks, to exploratory and open-ended tasks (Marchionini, 2006). Past research has shown that search behavior varies significantly by task type (Liu et al., 2010). Complex tasks may involve learning, and "understanding, sense-making and problem formulation are essential" (Byström and Järvelin, 1995). In this chapter, we use the often-used paper writing task, as employed by (Vakkari, 2001) and (Kuhlthau, 2004), to study information seeking and information searching.

As indicated in 4.2.1, there is a difference between information behavior, information seeking behavior and information search behavior (Wilson, 1999). Here, we briefly discuss models classified by Wilson as information *seeking* and information *search* models. At the level of information *seeking*, various models exist which describe the information seeking process from a macro perspective. These models include for instance Wilson's problem-solving model (Wilson, 1999) and Foster's non-linear model (Foster, 2005). Ellis (1989)'s model includes behavioral patterns of information seeking, which are not necessarily linear.

Carol Kuhlthau, in her Information Search Process (ISP) model (Kuhlthau, 2004), describes a more sequential and temporally-based set of stages, as previously detailed in Section 4.2. The thoughts, feelings, uncertainty, and actions of a user rise, fall, and evolve as the users pass through different stages. Vakkari (2001) later refined Kuhlthau's model and summarized its stages into *pre-focus*, *focus formulation* and *post-focus* stages. By studying students at three stages during a semester-long project, Vakkari found changes in relevance judgements, search tactics, terms and operators across stages.

Whilst information seeking models may inform the general design of IR systems, information *search* models (or information retrieval interaction models) often times may provide more specific *means* to improve design, in particular the interaction between users and information systems. According to Wilson (1999)'s categorization, information searching models include Spink (1997)'s model of the IR interaction process, which describes cycles of interaction with IR systems, including user judgements, search strategies, tactics and moves. Saracevic (1997)' Stratified model of Information Retrieval Interaction views IR interaction as a dialogue between user and computer, and includes different levels (strata) of interactions. Belkin et al. (1995) has modeled the behavior "people engage in while searching for information in some knowledge resource" as Information Seeking Strategies (ISS). These may be seen as interactions between user and IR system components, and an 'episode' may consist of a sequence of ISSs. ISSs can be described using four dimensions: method of interaction, goal of interaction, mode of retrieval, and considered resources. Finally, Marchionini's (1995) Information-seeking Process Model consists of various sub-processes and their relationships (e.g. 'define problem', 'select source' and 'formulate query'). These are *micro*-level models, which can help us to *design* novel SUI features.

5.2.2 Search User Interfaces

Search user interfaces (SUIs) serve as an intermediary between the user and the underlying data in an information retrieval system. As Hearst (2009) indicates, SUIs aid "users in the expression of their information needs, in the formulation of their queries, in the understanding of their search results, and in keeping track of the progress of their information seeking efforts." SUIs may be designed in vastly different ways, though designing effective SUIs with a high usability is a complex process, as Shneiderman and Pleasant (2005) argue, and it often involves finding trade-offs in simplicity and functionality. This difficulty in designing effective SUIs has led to a growing number of guidelines and theories (Shneiderman and Pleasant, 2005).

Research into Search User Interfaces (SUIs) (Hearst, 2009; Wilson, 2011b; Russell-Rose and Tate, 2012) has suggested many different interactive features, from search suggestions (Niu and Kelly, 2014) to facets (Tunkelang, 2009) to personal spaces to collect useful results (Donato et al., 2010). Though their usefulness has been proved in various studies, most of these features have not been adapted in common search engines. Hearst suggests some underlying reasons for the lack of adoption of advanced features: searching is used as a means to achieve a broader aim, search is mentally intensive and search systems should be understandable for people with different knowledge and experience (Hearst, 2009). Hence, overly complex search engines may distract from a user's core task.

The usefulness of features may depend on the task type and complexity. Some work ties the need for advanced features to different types of tasks, such as Exploratory Search tasks (White and Roth, 2009). Although Diriye et al. (2010) argued that in the context of known-item search tasks, excessive search features may impede people's information searching, most tasks involve at least some exploratory elements (White and Drucker, 2007).

Given the multitude of features which could potentially be integrated in SUIs, it may be useful to divide the types of features in different groups, based on their functions. In the previous chapter, we introduced the taxonomy proposed by Wilson (2011b), which distinguishes four groups of interface features. *Input* features aid users in expressing their needs, *control* features allow users to restrict or modify their input, *informational* features provide results or information about results, and *personalizable* features are tailored to the search experience of a user. In this chapter, we use this categorization to analyze the usefulness of features in different stages of search.

5.2.3 Utility of SUI Features Over Time

In Section 4.4, we have discussed the results of a number of user studies which have looked at the utility of search system features across stages. In those studies, most authors consider 'stage' a temporal segment of a singular session, and have retrospectively identified stages in people's search. Very few authors, however, have used an explicit multistage task design. Liu and Belkin (2015), for example, used one motivating work task but performed during three distinct sessions. They looked at the influence of task stage, type and topic knowledge on the interpretation of dwell time over multiple task sessions. While not directly looking at the use of SUI features, they found that task stage and topic knowledge could help to interpret time as an indicator of usefulness. Similarly, Wilson and schraefel (2008) conducted a longitudinal study of keyword and faceted search, finding that the latter only occurred after the second visit of an online video archive (likely due to confidence and interface understanding).

Other authors did not perform longitudinal studies, but used various simulated work tasks, performed during one session. In Kules and Capra (2012), users were asked to retrospectively assign stages to segments of their search sessions on the basis of a customly defined set of stages (most similar to the micro-analysis of search discussed before). Similarly, White et al. (2005); Niu and Kelly (2014) divided search sessions into equal parts ('beginning', 'middle' and 'end') to look at stage differences. Diriye et al. (2010) looked at the temporal distribution of the use of four interface features during a task, concluding that certain features (starter pages and search box) were search stage sensitive and other features search stage agnostic (facets and filters).

Finally, in chapter 4 we looked at conceptual ways to bridge macro and micro-level information seeking models. Based on changes in gaze behavior of a small-scale user study involving book search among computer science students, we found evidence for differences in the use of *input* and *personalizable* features over time, such as the query box and book basket, while other features were used throughout the task. While the previous chapter, and most previous literature mainly focused on singular tasks, we use an explicit *multistage* task approach in the current chapter. This way, we look at the passive and active utility of different SUI features across macro-level stages, to provide richer insights into exactly when different types of features become more useful. The next section discusses the setup of our multistage approach.

5.3 Experimental Setup

This section details our experimental setup. To study the active and passive use of interface features in different search stages, we conducted a within-subjects user study with task stage as the independent variable. For dependent variables, active system interactions were logged, passive mouse and eye movements were tracked, and questionnaires were used to collect data on perceived usefulness. Participants made use of *SearchAssist*[1], an experimental search system similar to a regular Web search engine, with different categories of SUI features potentially useful for each stage.

5.3.1 Task Design and Participants

While some prior work has inferred task stage, we constructed 3 task descriptions to explicitly represent three key *stages*, inspired by previous literature on tasks involving learning and construction (Kuhlthau, 2004; Vakkari, 2001). Stage 1 was modeled after the initial stages of Kuhlthau's ISP model (initiation, topic selection, exploration), summarized by Vakkari (2001) as the pre-focus stage. Stage 2 was aimed to make users formulate their specific topic (focus formulation), and a question about this topic. Finally, Stage 3 was based on the final

[1] The source code of SearchAssist and eye tracking software used in this study is available via: `https://github.com/timelessfuture/SearchAssist`

Table 5.1: Assigned multistage tasks

Introduction: For a class called "Computers in Society", the professor has given you the assignment to write a 5-page essay on some aspect of [topic]. Having a good grade for this essay is critical for passing the course. The essay is due in a week, but you have yet to decide on an exact topic. In a deliverable due tomorrow, you have to define your topic, a specific question about the topic and a list of sources.
Stage 1: Prepare a list of at least 3 ideas for a topic to write about in the context of [topic]. They should cover many different aspects of the topic, and unusual or provocative ideas are good. Search the web using the SearchAssist system to find out what information is available. Write down your ideas for topics in the text field below. Save any webpages you encounter via the SearchAssist system which are useful for writing on these topics (utilizing the "save result" feature).
Stage 2: Select one of the topics which you defined in the previous task. Choose the topic which interests you most, about which you are able to find enough information, and which you think you are able to finish in the allotted time. Use the SearchAssist system to find information to help you to decide on the topic, and save sources if needed using "save result". Write down the topic in the text box below. After having selected a topic you ask yourself the question "what is it that I want to find out about this topic?" Search the web using the SearchAssist system and formulate a specific question you would like to ask about this topic. You can save any pages you encounter which are useful for answering this question.
Stage 3: To be able to start writing your essay, take the specific question you have formulated in the previous step, and gather as much useful information as you can by searching the internet using SearchAssist. Find around 20 additional pages. Select the 5-10 pages that you could cite in your essay, and which are most relevant for answering the question you formulated in the previous step. If you have time left, formulate a draft answer to your question based on the information you have encountered (max. 300 words) and write it in the text box below.

stages of Kuhlthau's model (collection and presenting), summarized by Vakkari as the *post-focus* stage. In this stage, users had to collect sources relevant to their focused topic, and to provide a draft answer to the formulated question about their topic. In Section 5.3.5 we perform a validation of the simulated stages.

Written as simulated work tasks (Borlund, 2003) (see Table 5.1), the stage descriptions used elements of exploratory work tasks from previous studies (Kules and Shneiderman, 2008; Liu and Belkin, 2015), focused on the often used 'essay writing' task. Following Borlund (2003)'s guidance, the simulated work tasks were designed so that participants could relate to them, that they were topically interesting, and would add 'enough imaginative context'. After pilot tests and discussions with staff, two topics were selected: 'virtual reality', and 'autonomous vehicles'. The participants were undergraduate students of the School of Computer Science of the University of Nottingham (UK campus). The participants were recruited via posters, the Facebook page of Mixed Reality Lab, e-mails, and via *callforparticipants.com*. Upon completing the experiment, participants received a £10 Amazon voucher, and an additional £25 Amazon voucher was awarded to the participant with the best task outcome. In total, 26 participants joined the experiment, after a formative pilot study with 2 participants. Two participants, however, were excluded from our analysis, where one was unable to complete all three stages, and the eye tracking data was not sufficiently accurate for the other. Of the remaining 24 participants, 18 were male and 6 were female; 22 participants were aged 18-25, and 2 were between 26-35. Measured using a 7-point Likert scale (ranging from “novice” to “expert”), the participants assessed their computer proficiency as 5.5 (s.dev: 1.0), and their search expertise as 5.4 (s.dev: 0.8). Hence, the participants in the experiment were experienced computer users and searchers.

5.3.2 Data and Interface

For this study we designed *SearchAssist*, an experimental search system based on PHP, Javascript and MySQL, depicted in Figure 5.1. Search results, query corrections and query suggestions were retrieved in the JSON format via the Bing Search API and displayed as a familiar Web interface, similar to common Web search engines. The use of the Search API allowed participants to access a variety of sources, including scholarly, encyclopedic and news sources.

The SearchAssist interface consisted of the following elements:

- *1. Category filters.* Using the category filters, searchers could filter the set of results. The categories were derived from the top-level categories of the Open Directory Project (DMOZ). Retrieved results were matched against all DMOZ categories using the hostname of each result, and for each result occurring in DMOZ, the top-level category could be used to filter the result set. The results without a matching DMOZ category were shown under the label ‘Uncategorized’.

- *2. Tag cloud.* Using the tag cloud, it was possible to add one or more keywords to a query. The tag cloud was generated based on the most frequently occurring words in the snippets of the first 50 retrieved results.

- *3. Query suggestions.* Query suggestions were retrieved from the Bing Query Suggestions API, and they could be clicked to perform a new search.

- *4. Search box and results.* The SearchAssist interface featured a standard search box and results were retrieved from the Bing Web Search API; the Bing Spelling Suggestions API was also used. Each resultset item

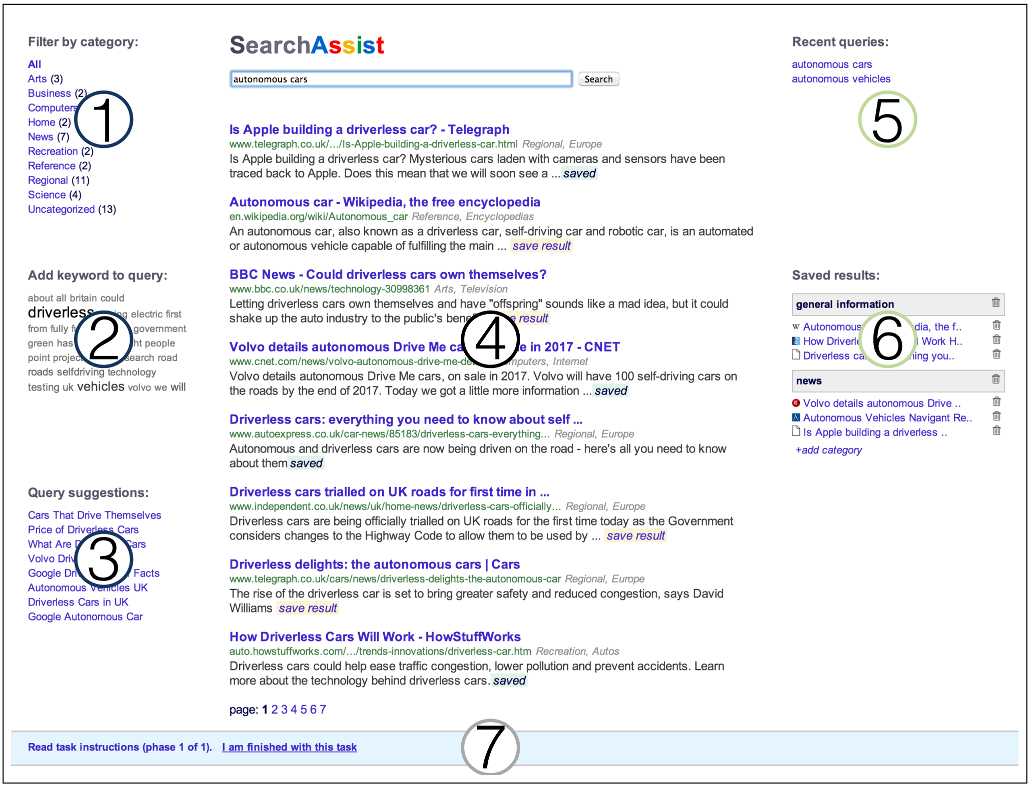

Figure 5.1: Screenshot SearchAssist. *Left column* (1, 2, 3): control features. *Middle* (4): input and informational features. *Right Column* (5, 6): personalizable features. (7): task bar

contained the title of the page, a URL, the DMOZ category, the snippet and a button to save a result. To better facilitate eye tracking, 8 search results were displayed at a time, similar to for example Jiang et al. (2014).

- *5. Recent queries.* The recent queries feature showed the last 15 queries performed across all tasks of the experiment, and allowed them to be resubmitted to the search engine.
- *6. Saved results.* The saved results feature allowed users to view (and remove) saved webpages, to reorder collected webpages by dragging and dropping, and to add (or remove) category labels to the gathered results.
- *7. Task bar.* The task bar contained task-related material, including a link to open the task instructions and a link to finish the current task, after which a user was prompted to fill out the corresponding questionnaire. The task instructions were shown in a Google Doc, which was also used to collect their responses.

5.3.3 Protocol

The experiment started with signing the consent forms and a pre-questionnaire, asking for demographics and ratings for knowledge about the potential task topics. As domain experts would behave differently than domain novices, participants were assigned the topic that they knew least about. Participants were then introduced to the features of the experimental system via a structured Powerpoint presentation[2] and given a training task (approx. 5 min.), which was used to mitigate the familiarity affects in the study, and to check the calibration of the eye tracker. The task stages were performed in sequence; the stage order could not be counter-balanced without losing the cumulative learning required from stage to stage. Participants were given 15 minutes for each stage, including a one minute warning, however participants were allowed to continue after this final minute passed. After each stage, users filled out a questionnaire about the perceived usefulness of features. After the final stage, participants also completed the post-questionnaire and a short debriefing interview (taking 5-10 minutes), focused on their experiences with the system. The total time to participate in the experiment varied between 55 and 90 minutes.

5.3.4 Logging and Eye Tracking

The system logged the active and passive interactions in three ways: via system logging, browser history and eye tracking, and the experiment was carried out

[2] To obtain forthright feedback on the experimental system, participants were informed that an external company designed the system (as opposed to the researchers performing the study).

using the Chromium browser. After each experiment, the browser history was exported in JSON format using the "Export history" browser extension, and the local browser history was deleted. All user actions were saved in a database via MySQL, and as plain text files using Log4Javascript. The logged data included all clicked interface features, all entered text (in the query box), and which page was active in the browser (the search interface, a webpage or the task page). In addition, all results items, query suggestions and query corrections retrieved via the Bing API were saved in their original JSON format.

For passive behavior, the system logged the position of the mouse cursor, and for context, took a screenshot of the user's screen four times per second. Eye tracking was performed using the EyeTribe eye tracker, calibrated using the included software. The Python-based PyGaze framework (Dalmaijer et al., 2013), and the PyTribe toolbox (a Python wrapper for the EyeTribe eye tracker) were customized to our needs and tightly integrated with the experimental interface. For the eye tracking data, the fixation counts and durations were calculated. Fixations were considered as sequences of eye tracking measurements within a 25 pixel radius; within a timeframe of at least 80ms (similar to e.g. Buscher et al. (2008)). We defined bounding boxes for each SUI element of the *SearchAssist* interface to detect the Area of Interest (AoI) of the fixation. In addition, to derive the depth of results list items inspected, we defined a bounding box for each results list item. The same methods were used to calculate the counts and duration of mouse movements in each AoI.

5.3.5 Data and Task Validation

First, we sought to confirm that the two topics, 'virtual reality' (VR) and 'autonomous vehicles' (AV), were comparable. No significant differences were found between overall task time, number of queries, results viewed, nor in the majority of usefulness ratings.

Only one significant difference was found, using the Mann-Whitney test, in the post-stage usefulness ratings for the 'saved results' feature for the first ($U=30$, $p=0.01$) and second stage ($U=34$, $p=0.02$), although no significant differences were found for logged usage of this feature. Informal observations indicate that there may have been a higher number of *relevant* results that *could* be found in the AV topic, but that these have not affected the majority of behavior. Overall, however, we conclude that the topics invoked comparable behaviors and continue to analyse the data from both topics as a single set.

A key aspect of Kuhlthau's and Vakkari's models is the learning nature of the tasks. We assessed learning aspects of the tasks via questionnaires of the tasks before, during and after experiment. In the post-questionnaire, participants indicated that they learned considerably about the topic during the experiment. Using a Likert scale of 1 ("none") to 7 ("a great deal"), the average assessment of learning was 5.3, with a standard deviation of 0.9. Participants also had to indicate their level of knowledge about the chosen topic (virtual reality and autonomous vehicles), via a Likert scale ranging from "novice" (1) to "expert" (7). For the chosen topic, this was done before the experiment, and after the first, second and third stage. The self-assessed knowledge about the topic grew from an average value of 2.8, to 4.0 after the first stage, to 4.4 after the second stage, and to 4.9 after the third stage. Hence, participants gradually grew in their expertise about the topic, with the largest increase occurring in the first stage. The self-assessed topic familiarity, ranging from "unfamiliar" (1) to "familiar" (7), showed similar tendencies, and increased from 3.8 before the experiment, to 5.4 after the study finished.

Then, 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 derived from Kuhlthau's model[3]. For the first stage, the most commonly selected activity was 'exploring' (N=17), followed by 'gathering' (16); corresponding to the initiation and exploration activities associated with the initial stages of Kuhlthau's model. After the second stage (focus formulation) users most often chose 'focusing' (16) and 'collecting' (12) as words representing their activities. The common use of focusing corresponds to the focus formulation activity, while collecting may refer to the collected documents in that stage. Finally, for the third task also 'focusing' (17) and 'collecting' (14) were the most common words. The fact that the first task was seen as explorative is also reflected in the type of information sought, reported in the questionnaire, which evolves from 'general' (in the questionnaire after stage 1), to 'specific' (after stage 2 and 3). We conclude that even though the separations between stages are not always dichotomous, our experiment correctly invoked the main activities in each stage.

We also assessed the feelings of participants during the experiment, using a word list adapted from Todd (2006). 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[4]. A large degree of partipants indicates confidence in the questionnaire at the end of all three stages (42% on average), and the same holds true for satisfaction (55% on average). However, also fluctuations can be detected, showing some evidence of Kuhlthau's findings on uncertainty and optimism. Uncertainty changes over time, expressed by 12.5% (n=3) of all participants in the first stage, but by 25% (n=6) in the second stage; and falling again to a degree of 12.5% in the third stage. Optimism, on the other hand, is clearly growing over time: this word is selected by 29% of the participants (n=7)

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

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

in the first stage, and by 46% (n=11) during the third stage. While beyond the scope of the present study, these indications of feelings could compared to the behavioral aspects captured in the active and passive interaction data of the experiment in future work. Next, we discuss the direct results of our study.

5.4 Search Stage & Active Behavior

The following three sections examine the results of the study, and whether the participants showed distinct behavior in the different stages of their overall task. Together, the 24 participants issued 502 queries and clicked on 684 results. Participants spent an average of 32:56 minutes to complete the 3 task stages. Of this time, 36.8% was spent in the SUI, 33.0% on the task screen, and 30.2% on the webpages. Participants spent, on average, 11:32 minutes on the first stage, 8:24 minutes on the second stage, and 12:59 minutes on the third stage.

First, we focus on our first research question (**RQ4.1**): *How does the user's search stage influence active behavior at the interface level?* We define active behavior as the behavior that can be directly and indirectly derived from the logged interactions, such as clicks, queries, and pages visited.

5.4.1 SUI features

Table 5.2 summarizes the main **active interaction** with each available SUI feature. The use of the *Query Box* (counted as the clicks on the 'search' button) is most frequent in the first stage, and decreases in the second and third stage. Using the within-participants, repeated measures ANOVA, we found a significant difference in the use of the search button ($p<0.01$, $F(2)=13.6$). Post hoc tests, using the Bonferroni correction, showed that there is a significant difference between the first and second ($p<0.01$), and the first and third stage ($p<0.01$). Hence, users use the search button more in the first stage, most likely to explore the assigned topic (comparable to Wilson (2009)).

The clicks on retrieved results items, via the *Results List* feature, remain more or less stable without significant differences per stage. The number of times a result is saved using the adjacent 'save result' link, however, is decreasing after the first stage. Users also appear to examine the results beyond the first page more frequently in the third stage (by clicking 'next page') but these differences, like the differences in result clicks and result saves, are not significant.

The *Category Filters* feature is used significantly less frequently after the first stage, and thus seem to be most useful in the initial task stage ($p<0.01$, $F(1.2)=8.6$, Greenhouse-Geisser correction). The differences, with Bonferroni correction, are most prominent between the first and third stage ($p<0.01$), but also occur between the other stages (1->2: $p=0.03$, 2->3: $p=0.03$). Similarly, the clicks on the *Tag Cloud* feature are significantly different ($p<0.01$, $F(1.4)=8.5$, Greenhouse-Geisser correction). Again, the first stage features the highest number of clicks, and using a pairwise comparison, with Bonferroni correction, we found significant changes in clicks between the first and second stage ($p=0.02$), and between the first and third stage ($p=0.01$).

Compared to the other features, the *Query Suggestions* and *Recent Queries* features are not frequently used, and a slight decrease in use of the Query Suggestions and a slight increase in the use of the Recent Queries feature is visible in the data, but are not significant.

Although the differences in the use of the 'Save result' link in the Results List were not significant, the statistics for the *Saved Results* feature indicate that users add categories to these items mostly in the first stage ($p<0.01$, $F(2)=8.1$). Pairwise comparisons, with Bonferroni correction, show that the significant differences occur between the first and second ($p=0.02$), and the first and third ($p<0.01$) stage. Hence, participants save and categorize items in the saved results list most frequently in the first stage. The clicks on the saved results (bookmarks), on the other hand, are clearly most frequent in the last stage ($p<0.01$, $F(1.1)=18.8$, Greenhouse Geisser correction). A pairwise comparison shows significant differences between the first and third stage ($p<0.01$), and the second and third stage ($p<0.01$). Finally, the last stages show a slight increase in the removal of categories and saved items, as opposed to the additions in the first stage, but no significant differences were found.

5.4.2 Queries & Page Visits

As we observed in Table 5.2, participants used the Query Box feature most often in the first stage. Now, we look in more detail at the **queries**, summarized in Table 5.3. The total number of queries submitted, including query suggestions and use of the Recent Queries feature, is significantly decreasing per stage ($p<0.01$, $F(2)=8.9$). A pairwise comparison, with Bonferroni correction, indicates that the differences are significant between the first and second ($p<0.01$), and between the first and third stage ($p<0.01$). Likewise, the unique queries are significantly different ($p<0.01$, $F(2)=7.9$), again with a significant difference between the first and second ($p<0.01$) or third stage ($p<0.01$). These findings overlap with the findings in Section 4.4. Most queries performed were unique, though there is some overlap in the queries between the first, second and third stage, meaning that participants reuse queries in latter parts of the experiment (i.e. by re-entering the same query or using the Recent Queries feature). In the first stage, the majority of queries are initiated from the Query Box. However, subsequent stages show an increase of the relative use of the Recent Queries feature, and a stable share of the Query Suggestions.

While the number of queries decreases after the first stage, the number

Table 5.2: SUI active interaction (clicks), from system logs

mean	stage1	%	stage2	%	stage3	%
Query Box						
search clicks**	8.4	*24.3*	4.5	*19.8*	4.6	*14.9*
Results List						
result clicks	7.3	*20.9*	5.5	*24.2*	7.8	*25.2*
result saves	6.1	*17.5*	4.3	*19*	3.7	*11.8*
next page clicks	0.8	*2.4*	1.2	*5.1*	1.7	*5.5*
Category Filters						
clicks**	2.9	*8.3*	1.1	*4.7*	0.6	*2*
Tag Cloud						
clicks**	1.6	*4.7*	0.7	*3.1*	0.5	*1.7*
Query Suggestions						
clicks	0.8	*2.3*	0.4	*1.7*	0.5	*1.7*
Recent Queries						
clicks	0.3	*1*	0.6	*2.5*	0.8	*2.4*
Saved Results						
clicks**	0.7	*2*	0.9	*3.9*	6.3	*20.4*
add category**	2.2	*6.4*	0.9	*3.9*	0.6	*1.9*
move item	3.5	*10*	2	*8.8*	2.4	*7.7*
remove category	<0.1	*<0.1*	0.3	*1.1*	0.3	*0.9*
remove item	<0.1	*<0.1*	0.5	*2.2*	1.2	*3.8*
Total	34.7	*100*	22.8	*100*	31.1	*100*

*Within-subjects ANOVA: * significant (p<0.05); ** significant (p<0.01)*

Table 5.3: SUI active interaction (queries and page visits)

mean	stage1	stage2	stage3
Queries**	9.5	5.5	5.9
*via Query Box***	*88%*	*81%*	*78%*
via Recent Queries	*3%*	*11%*	*13%*
via Query Suggestions	*8%*	*7%*	*8%*
Unique queries**	8.1	5.1	5.3
Overlap queries prev. stages	0	1.4	1.8
Mean num. query words**	3.2	4.5	4.4
Levenshtein distance (query diversity)	13.2	13.9	17.0
Visited pages**	8.0	6.4	14.2
via Results List	*91%*	*86%*	*56%*
*via Saved Results***	*9%*	*14%*	*44%*
Unique visited pages**	7.3	5.9	10.8
Overlap visited pages prev. stages	0	0.8	2.8
Mean rank visited pages	3.1	5.1	6.4

*Within-subjects ANOVA: * significant (p<0.05); ** significant (p<0.01)*

of **words per query** increases. The highest mean number of query words occurs in the second stage (4.5), and an almost equally high value during the third stage (4.4). The higher number of queries may be related to exploration activities in the first stage, which require various queries to explore various topics. The increasing number of query words, on the other hand, may occur because a person is searching for a more specific topic, and may have built a conceptual representation of a topic (Vakkari, 2001). For example, one user (P.02) started with the query "virtual reality" in the first stage, but queried for "the impact of virtual reality on society art and gaming culture" in the third stage. Or, the queries from another participant (P.06) evolved from short queries such as "autonomous vehicles" to longer queries like "autonomous vehicles costs insurance industry". The differences in the number of query words are significant ($p<0.01$, $F(2)=5.3$), specifically between the first and second stage ($p<0.01$, Bonferroni correction). Finally, we calculated the query diversity, based on the Levenshtein distance between all pairs of unique queries of a user in a certain stage. The query diversity is similar during the first and second stage, but is highest in the third stage, meaning that the edit distance between users' queries is greater; although these differences are not significant.

An analysis of **page visits** reveals that participants in our experiment visited the highest number of pages in the third stage ($p<0.01$, $F(1.3)=11.6$, Greenhouse-Geisser correction), when collecting materials. The differences are significant between the first and third stage ($p=0.02$) and between the second and third stage ($p<0.01$). This variance seems to be explained primarily by the revisiting of pages from the Saved Results feature ($p<0.01$, see previous section), as page visits from the Results List were not significantly different. This finding is reflected in the uniquely visited pages ($p<0.01$, $F(1.5)=8.1$, Greenhouse-Geisser correction), but here the only significant changes occur between the second and third stage ($p<0.01$). Further, the result is also reflected in the mean dwell times on the webpages, which are highest in the first (12.9 sec.) and second stage (14.4 sec.), but lower in the third stage (8.9 sec.). The dwell times are significantly different ($p<0.01$, $F(2)=7.8$), between the first and second ($p<0.01$), and between the second and third stage ($p<0.01$). Participants also explored further down the result set in the later task stages, with the mean visited rank increasing from 3.1 to 5.1 and 6.4 respectively. This was not significant within our current sample, though we found similar behavior in the previous chapter's exploratory analysis (Section 4.4).

Summarizing, this section has focused on the active interaction with the system during the experiment. Utilizing the categorizations of Wilson's framework for SUI features (Wilson, 2011b), the results show various tendencies: *input* features (the Query Box) and *control* features (Category Filters, Tag Cloud and Query Suggestions) are clearly used less often in subsequent stages, while the use of the *informational* (Results list) features remains stable. The results for the *personalizable* features (Recent Queries and Saved Results) differ. The Recent Queries feature is scarcely used, but an increasing tendency can be observed across stages. Similarly, users mostly click on their saved results in the last stage, but save the actual results and add categories most frequently in the first stages. Hence, the Saved Results feature is initially used to store and categorize important results, but later to revisit previous results. Also, users start out with a significantly higher number of queries, as compared to later stages, while the number of page revisits substantially increases in the last stage. Evidence for the learning aspects of the used tasks are found in the increase of the number of query words and query diversity (Kuhlthau, 2004; Vakkari, 2001), as users seem more able to express their needs in queries.

Finally, another contrast can be observed, namely between commonly used features and scarcely used features. Together, the Query Box, Results List and Saved Results features take up over 80% of all clicks, while the remaining set of features takes up less than 20% of all clicks (see Table 5.2). The infrequent *active* use of certain features, in particular the Query Suggestions and Recent Queries features, lead to the question whether some features are perhaps used in a *passive* way, which we will further examine in the next section.

5.5 Search Stage & Passive Behavior

In this section, we focus on the following research question (**RQ4.2**): *How does the user's search stage influence passive behavior at the interface level?* We examine the user's mouse position and eye fixation data to look at the passive behavior in each search stage.

5.5.1 Mouse Hovers

Participants' **mouse movements** can shed more light on the use and utility of SUI features in different stages. Mouse moves in a particular area can be simply movements to reach or click a SUI feature, but may also indicate different types of usage, i.e. mouse moves aiding users in processing the contents of results pages (Rodden et al., 2008). In our analysis, we look at the *passive* mouse movements: the mouse hovers in a SUI feature area that did *not* lead to a click.

Table 5.4 shows the mean count of mouse movements over time. We counted mouse hovers (defined as a change in the coordinates of the mouse pointer) within each SUI feature's Area of Interest. There are significant differences for the following features: the Query Box ($p<0.01$, $F(2)=6.4$), the Results List ($p<0.01$, $F(2)=6.9$), the Category Filters ($p<0.01$, $F(2)=7.0$) and the Tag Cloud feature ($p=0.03$, $F(1.5)=4.5$, Greenhouse-Geisser correction). Mouse hovers in these SUI areas are most common in the first stage, and significantly decrease in the second or third stage. The other features do not show significant changes over time. The results for this measure show overlap with the active interaction measure of the previous section, with the exception of a "dip" in mouse hovers on the Results List in the second stage, and a higher relative amount of hovers over the Query Suggestions, especially in the second and third stage. The higher and more stable degree of mouse hovers around the Query Suggestions may indicate that users use this feature passively in all three stages, as opposed to the decreasing tendency visible in the active use measure. To gain further insights, we next look at passive use, not even involving the mouse, using eye tracking.

5.5.2 Eye Tracking Fixations

To gain an initial overview of eye movements within the SearchAssist interface, we generated **heatmaps** for each stage across all participants. These heatmaps (Figure 5.2) show the spatial distribution of the fixations. A visual inspection reveals a consistent focus on the Query Box and Results List SUI features in each stage (middle pane). The Category Filters, Tag Cloud and Query Suggestions features (left pane) are most intensively used in the first stage, while the Saved Results feature (lower right panel) appears to be most intensely used in the last

Table 5.4: Passive use: mouse hovers *not* leading to a click

mean	stage1	%	stage2	%	stage3	%
Query Box**	344.7	*16.6*	250.2	*19.5*	210.2	*14.6*
Results List**	1226.8	*59.1*	701.9	*54.7*	872.9	*60.7*
Category Filters**	124.6	*6.0*	57	*4.4*	67.7	*4.7*
Tag Cloud*	165.9	*8.0*	73.1	*5.7*	47.2	*3.3*
Query Suggestions	91.3	*4.4*	58.5	*4.6*	56.6	*3.9*
Recent Queries	17.6	*0.8*	18.3	*1.4*	21.3	*1.5*
Saved Results	103.7	*5.0*	123.5	*9.6*	163.3	*11.3*
Total	2074.6	*100*	1282.5	*100*	1439	*100*

*Within-subjects ANOVA: * significant (p<0.05); ** significant (p<0.01)*

Table 5.5: Passive SUI use: mean eye tracking fixation count

mean	stage1	%	stage2	%	stage3	%
Query Box*	58.08	*14*	35	*13.1*	41.42	*11.8*
Results List*	224.88	*54.3*	139.83	*52.5*	187.17	*53.5*
Category Filters*	17.63	*4.3*	10.46	*3.9*	11.42	*3.3*
Tag Cloud**	31.71	*7.7*	14.58	*5.5*	15.5	*4.4*
Query Suggestions*	16.88	*4.1*	9.71	*3.6*	10.83	*3.1*
Recent Queries	10.92	*2.6*	9.79	*3.7*	10.13	*2.9*
Saved Results	54.38	*13.1*	47.17	*17.7*	73.63	*21*
Total	414.48	*100*	266.54	*100*	350.1	*100*

*Within-subjects ANOVA: * significant (p<0.05); ** significant (p<0.01)*

stage.

These differences can be inspected in more detail using the absolute and relative **fixation counts** (with a minimum duration of 80 ms), summarized in Table 5.5. For the most part, the results for the passive use of SUI features confirm the results regarding active use. The number of fixations on the Query Box is significantly decreasing after the first stage (p=0.01, F(2)=4.9), which is comparable with the lower number of unique queries performed in the second and third stage observed in the active interactions. In particular, the difference is significant between the first and second stage (p=0.02). In addition, the less frequent active use of the Tag Cloud and Category Filters is reflected in a decrease in the number of fixations in the second and third stage, and this difference is significant for both Category Filters (p=0.01) and Tag Cloud (p<0.01). A pairwise comparison reveals significant differences between the first and second stage for the Category Filters (p=0.03), and between the first and second (p=0.01) or third stage (p=0.02) for the Tag Cloud.

The fixations on the results list decrease significantly after the first stage ($p=0.02$, $F(2)=4.4$). A significant difference for the fixations on the Results List feature exists between the first and second stage ($p<0.01$, Bonferroni correction), though the relative degree of fixations changes less. Table 5.2 in the previous section, however, did not show a significant difference for the number of clicks on resultset items in any stage. Similarly, the decreasing number of clicks on the Query Suggestions features are coupled with a lower number of fixations on this feature. These differences are significant ($p=0.04$, $F(2)=3.5$) between the first and second stage ($p=0.01$). As in the case of the active interactions, the Recent Queries feature does not show a significant difference, but the relative values for the fixations increase in the second stage. Finally, the fixations on the Saved Results feature rise during the stages, which is similar to the measured increase in the previous section, but the difference for the fixations is not significant ($p=0.09$).

The previous section showed some features which were used frequently, in particular the Query Box, Results List and Saved Results feature, and other features which were used infrequently, such as the Query Suggestions and Recent Queries. We would expect that the often-used SUI features also have a high degree of fixations. The results confirm this: the Results List takes up more than half of the fixations, and also the relative degree of the fixations on the Query Box and Saved Results is high. There is a difference, however, for features that were little used in an active way, such as the Query Suggestions and Recent Queries features. The percentage of fixations on the Query Suggestions over all three stages is 3.6% instead of 1.9% of clicks, and the fixation percentage for the Recent Queries is 3.01% instead of 1.97%. While the difference is relatively small, it does provide evidence that participants may use these features more passively than actively. Another difference can be observed for the Category Filters and Tag Cloud: participants look more at the Tag Cloud (5.8%), than they click on it (3.1%), while a contrary situation exist for the Category Filters (5% of all clicks, but 3.8% of all fixations).

Summarizing, on the one hand, the fixations and mouse movements by users validate the active behavior, showing similar tendencies. The significant differences in the use of the Query Box, Category Filters and Tag Cloud confirm the findings from the previous section, while the eye tracking data also suggests significant changes in the use of Query Suggestions. On the other hand, subtle differences exists in the passive use of less often used features, such as the Query Suggestions and Recent Queries features. This suggests that some features may not be used often in an active way, but that they are still used passively. In Section 5.6, we validate and contextualize these findings with subjective ratings of usefulness and qualitative feedback from participants.

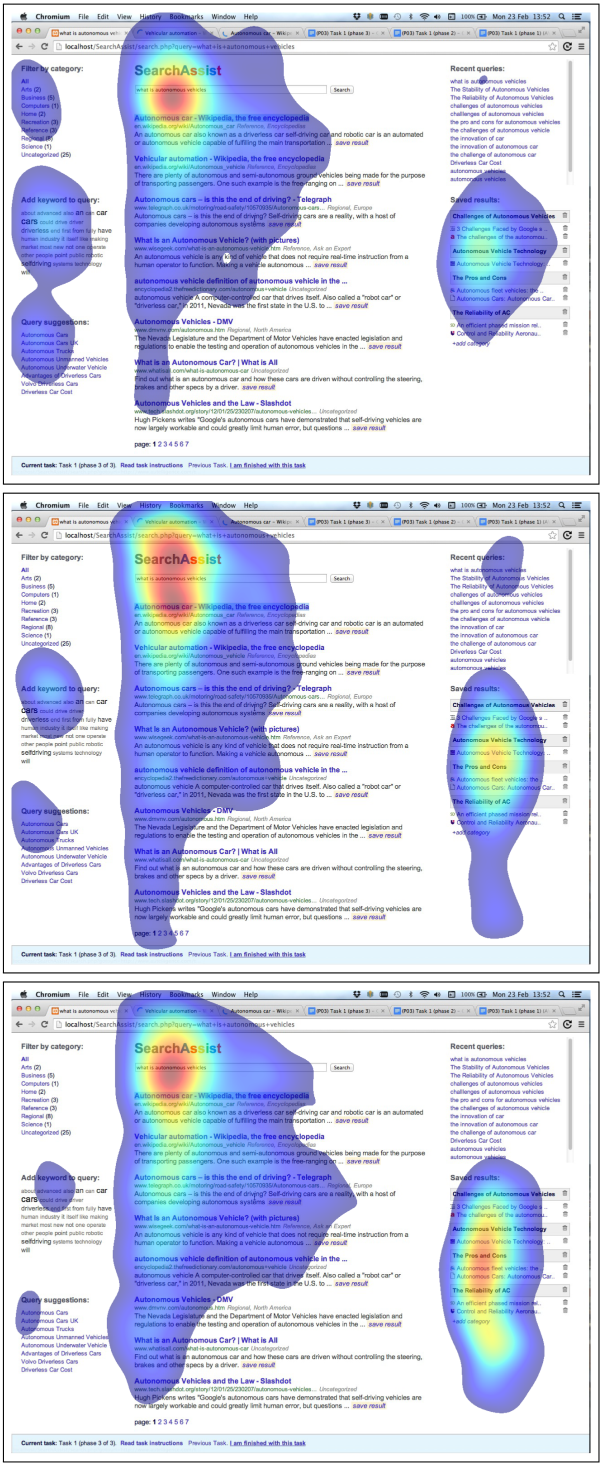

Figure 5.2: Eye tracking heatmaps (Stage 1, 2, 3), based on fixations over 80ms.

5.6 Search Stage & Perceived Usefulness

Our final research question of this chapter looks at the potential influence of search stages on the perceived usefulness of SUI features: (**RQ4.3**): *How is active and passive behavior reflected in the perceived usefulness of features?* Findings from questionnaires after each stage, after finishing the whole experiment, and brief post-experiment interviews are used to contextualize the findings so far.

5.6.1 Usefulness Ratings

After each task stage, participants were asked to rate the usefulness of each SUI feature of SearchAssist on a Likert scale of 1 to 7, as shown in Table 5.6. Somewhat expectedly, the most highly rated features are the Search Box and Results List features. As it turns out, however, this is closely followed by the Saved Results feature, which was also deemed to be very useful by most of the participants. Conversely, the least popular features among the participants were the Tag Cloud and Category Filter features. The most useful features were also rated most consistently among participants: the standard deviation values for the Search Box, Results List and Saved Results are substantially lower than for the other SUI features. Conversely, the most "controversial" feature was the Tag Cloud, with a standard deviation of 1.71, suggesting a relatively high variability of user ratings: some participants found it useful, and others did not perceive it as useful.

Comparing the stages, the Search Box and Results, Category Filters, Tag Cloud and Query Suggestions are all rated most highly in the first stage, which generally corresponds with the results for the active and passive interaction in the previous sections. The inter-stage differences for the Search Box and Results List (non-parametric Friedman test, $p<0.01$, $\chi^2(2)=13.3$) are significant. The Query Suggestions feature has significance ratings close to 0.05 (Friedman, $p=0.07$, $\chi^2(2)=5.4$); and is deemed most useful in the first stage. While the previous features are rated slightly lower in successive stages, the opposite holds true for the Recent Queries and Saved Results features, which both have their highest rating in the third stage. In the case of the Recent Queries feature, the differences are significant (Friedman, $p<0.01$, $\chi^2(2)=15.2$). Here, we note that the Recent Queries feature did not show any significant differences using the previous active and passive interaction measures, though a general increase of use could be observed.

Table 5.7 summarizes the users' ratings after the *whole* experiment, which are also visualised in Figure 5.3. Participants were asked to indicate in which stage or stages a feature was *most* useful, and were allowed zero to multiple answers for each feature. This table shows similar tendencies as Table 5.6, but

Table 5.6: Mean usefulness ratings, gathered after each stage (s.dev.). *Bold: stage with highest rating for feature.*

mean	stage 1	stage 2	stage 3	total
Search Box/Results*	**6.67 (0.7)**	6.33 (0.9)	6.08 (1.1)	6.36 (0.9)
Category Filters	**4.08 (1.5)**	3.79 (1.6)	3.46 (1.7)	3.78 (1.6)
Tag Cloud	**3.92 (1.7)**	3.54 (1.5)	3.63 (1.9)	3.70 (1.7)
Query Suggestions	**4.80 (1.4)**	4.00 (1.7)	4.00 (1.6)	4.26 (1.6)
Recent Queries*	3.46 (1.6)	4.13 (1.7)	**4.71 (1.6)**	4.10 (1.6)
Saved Results	5.83 (1.2)	6.17 (1.1)	**6.30 (0.9)**	6.08 (1.0)

*Non-parametric Friedman test: * significant (p<0.05)*

the differences are more pronounced. Hence, participants judged the usefulness of interface features slightly more explicit after completing the full experiment, perhaps at that moment having an overview of the stages involved in it. A chi-square test indicates that the differences are significant for all SUI features ($p<0.01$, $\chi^2(10)=33.5$). The feature ratings show a clear division: the Search Box/Results List, Category Filters, Query Suggestions and Tag Cloud were most useful in the first and second stage. The opposite is true for the Recent Queries and Saved Results, which were deemed more useful in the later stages.

5.6.2 Questionnaire and Interview Data

The data from the questionnaires and interviews were collected to provide insight into the utility of features at different moments of the task, and to contextualize our measurements. Here, we focus mainly on the *control* and *personalizable* features, which may support a user, but are not commonly included in regular search engines.

The general tendency for the *control* features (the Category Filters, Tag Cloud and Query Suggestions), as visualized in Table 5.7, is that their usefulness decreases over time. In particular, the Tag Cloud feature is deemed less useful in the second and third stage, and is a 'controversial' feature with a considerable variation in user ratings. In the post-stage questionnaires, some participants emphasize the usefulness: *"the tag cloud really aids exploring the topic"* (P.6), and *"the tag cloud came up with words that I hadn't thought of using that were very useful"*. However, especially after the second and third stage, a number of participants (P.05, P.16, P.18, P.21, P.27) indicated that the tag cloud is not so useful, saying that it *"contributed little during this task"* (P.05), that *"the tag system doesn't help to narrow the search much"* (P.18) and that it *"in the end seemed to be too general"* (P.07). P.12 summarizes this in the interview after the experiment: *"The Tag Cloud, I think, was good at the beginning, because when you are not exactly sure what you are looking for, it can give inspiration"*

Table 5.7: Mean post-experiment usefulness ratings – at which moment were the SUI features most useful (% of participants).

perc	stage1	stage2	stage3
Query Box/Results List**	**100.00%**	75.00%	66.67%
Category Filters**	**54.17%**	20.83%	12.50%
Tag Cloud**	**41.67%**	16.67%	8.33%
Query Suggestions*	**54.17%**	29.17%	20.83%
Recent Queries*	12.50%	54.17%	**70.83%**
Saved Results **	37.50%	66.67%	**91.67%**

*Chi-square test: * significant (p<0.05); ** significant (p<0.01)*

(which we can connect to the findings of Kelly (2009)). This can explain the fluctuations in use and perceived usefulness: the Tag Cloud is mainly useful in the beginning of the task, when users are exploring the topic, since provides basic vocabulary to the user (using frequent words in the retrieved snippets), and it may provide inspiration. In another interview, P.15 emphasizes the support of the Tag Cloud feature in generating ideas: *"it was nice to look at what other kinds of ideas [exist] that maybe you didn't think of. Then one word might spark your interest"*. However, once a user had built up a certain level of background knowledge about the topic, the value of the Tag Cloud seemed to diminish, because the user may already be familiar with the words that it displays. A similar situation exists for the Category Filters, as P.11 suggests: *"Category Filters, [those were] good at the start (...) but later I wanted something more specific"*. Hence, the refining of search results using general categories may be useful in the initial stages, but later users have more specific ideas of what they want to search for, and wish for more specific categories. For example, P.16 indicated in the questionnaire after the second stage that *"Category Filters could be more specific in its categories"*, and P.26 ideally wanted to choose a custom set of categories.

The Query Suggestions also have a similar variation in perceived value. While deemed more useful than both the Category Filters and Tag Cloud in the initial stage, the usefulness ratings for the Query Suggestions decrease in the subsequent stages. As in the case of the Category Filters, users ask after the second and third stage for improved precision, and quality (P.2, P.19), and indicate that the suggestions were *"not relevant"* for the current task (P.6, P.8). Again, this can be further contextualized using the interview data: P.11 suggested that the Query Suggestions feature *"was good at the start, but as soon as I got more specific into my topic, that went down"*. P.23 provided a suggestion for design improvement of such a feature, and indicates that over time, the Query Suggestions should take into account previous searches and *"tailor to the kinds results"* he was visiting. Still, some users mentioned, similar to the Tag Cloud, that the Query Suggestions may provide inspiration, but also serendipity: *"I clicked the query suggestions a few times. They gave me sort of serendipitous results, which are useful."* (P.24)

As opposed to the previously discussed *control* features, the *personalizable* features, the Recent Queries and Saved Results features, were increasingly highly rated. Except for some small usability issues (e.g. indicated by P.03 and P.17), users were enthusiastic about the Saved Results feature, and 13 participants wrote down positive comments in the questionnaires (P.04,P.07,P.13-16,P.18,P.21,P.23-27). P.15 remarked: *"I really found the save results feature useful, very easy to use, I wish my search engine had this!"*. The ability to categorize results was also seen as useful: *"The way that I can categorize all the pages I get is useful"* (P.27), and *"I just felt I was organizing my research a little bit."* (P.18). One participant (P.07) also indicated that the Saved Results feature helped to lay out the plans for his search. It also encouraged participants that normally do not use bookmarks to save results. Regarding the usefulness over time, various participants (for example P.12) indicate that the Saved Results *"are most useful in the end"*. One of the participants also provided feedback in the interview that can explain the previous findings that the highest number of links were saved in the first stage (P.20): *"at the start [I was] saving a lot of a general things about different topics. Later on I went back to the saved ones for the topic I chose and then sort of went on from that and see what else I should search."* Hence, users may search and save many items in the beginning, but, if they formulated a focus, will save more specific sources later. Similarly, P.26 said *"I guess in the end I was looking for a more specific search, while my search in the beginning was just simple – [I] just searched virtual reality [and] didn't do anything on top of that"*.

Some participants indicated that the Recent Queries feature, like the Saved Results feature, was more useful later in the experiment: *"Recent queries were more useful in the end because I had more searches from before"* (P.26). The fixation data analyzed in the previous section has shown some evidence that users look more at this feature than that they actually click on it, or hover over it with the mouse. P.23 provides insight into this finding: *"the previous searches became more useful 'as I made' them, because they were there and I could see what I searched before. I was sucking myself in and could work by looking at those."* Thus, the continuous display of recent queries may aid users in their process by providing feedback about the previous paths followed; this may be the case especially in the context of complex tasks.

Summarizing, this section has looked at the perceived usefulness of features. The user ratings of different SUI features largely confirmed the findings from the previous sections, in that certain *input* and *informational* features were deemed highly useful in most stages (Query Box and Results), while *personal-*

izable features become increasingly useful (Recent Queries, Saved Results) and *control* features decreasingly useful (Category Filters, Tag Cloud and Query Suggestions). The changes in ratings after each stage are significant for the Recent Queries and Search Box/Results feature. The increasing use of the Recent Queries feature could be observed in both active and passive interactions, but the significant difference in the usefulness ratings provides more substantial evidence for when this feature provides most value. Finally, the questionnaires and interviews provided contextualization to the active and passive interactions: the variations in the use of certain features, like the Tag Cloud and Query Suggestion feature, are caused by a user's increasing domain knowledge. As the participants indicated in the questionnaires and interviews, the features useful at the start do not provide the specific information needed in later stages, hence do not take into account a user's growing understanding of a topic.

5.7 Discussion and Conclusions

The conceptual analysis of the previous chapter has evidenced the potential influences of information seeking stages on search system use. This chapter has directly examined how different SUI features can support distinct macro-level stages. Specifically, we studied active, passive and perceived utility of SUI features during different stages of a complex task via an explicit multistage task design.

Section 5.4 looked at our first research question (**RQ4.1**): *How does the user's search stage influence active behavior at the interface level?* Hence, we focused on the influence of search stages on active behavior, i.e. the clicks and submitted queries. We concluded that *input* and *control* features were most frequently used in the first stage, while *informational* features were used throughout the experiment. Some features were used infrequently, leading us to validate if those features perhaps were utilized passively instead of actively.

We investigated the passive utility of SUI features in Section 5.5, which focused on the second research question (**RQ4.2**): *How does the user's search stage influence passive behavior at the interface level?* Here, we encountered inherent similarities of passive fixations and mouse movements with the active clicks. 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.

Finally, the third research question, discussed in Section 5.6 (**RQ4.3**) was: *How is active and passive behavior reflected in the perceived usefulness of features?*. This research question investigated the perceived usefulness, based on questionnaire and interview data. Our main finding was that the perceived use-

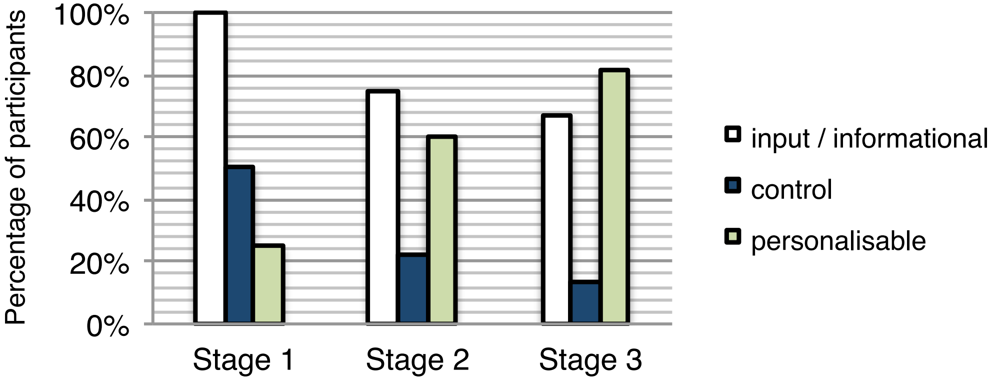

Figure 5.3: SUI feature categories perceived most useful by stage

fulness of features differs radically per search stage, as summarised in Figure 5.3. 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, *personalizable* features (query history and a feature to save results), are considered as less useful in the beginning, but their usefulness significantly increases over time, even surpassing the value of common SUI features. Hence, these results suggest that the macro-level process has a large influence on the usefulness of SUI features.

Finally, we return to our main research question: *How can different types of search user interface features support distinct macro-level information seeking task stages?* To answer this research question, our findings can be connected to the information seeking stages defined by Kuhlthau (2004) and Vakkari (2001).

At a user's initial pre-focus stage, as Vakkari and Hakala (2000) have indicated, thoughts of users are "general, fragmented and vague." Searchers are unable to express "specifically what information is needed," and their "relevance criteria are vague." At this stage, the uniqueness of encountered information is high, while the redundancy of found information is low (Kuhlthau, 2004). The first stage of our experiment represented pre-focus user activities, and at this stage the *input*, *informational* and *control* features are most useful. Naturally, *input* features, are needed at this stage to express users' needs in terms of queries, while users retrieve results via the *informational* features. The user's vague understanding, the trouble in expressing her need, limited domain knowledge, but also the large amount of new information can explain the prominent role of *control* features in this initial stage: users may utilize them to explore different kinds of information, and to control their result set.

As Kuhlthau (2004); Vakkari and Hakala (2000) suggest, the subsequent focus formulation stage is crucial in the process. During this stage, "the search for information becomes more directed", and a better understanding drives persons to seek relevant information, using differentiated criteria. This stage was represented by the second task of our experiment. Our experimental results show that the *control* features become less essential at this point, likely caused by user's improved understanding and emerging focus. This even causes provided categories, suggested tags and searches to be "not specific enough" anymore. The *personalizable* features, on the other hand, become more important during the focus formulation stage and beyond. Contrary to *control* features, *personalizable* features may continously support users in their process, providing feedback on the paths followed in their information journey. These features "grow" with the emerging understanding of a user. For example, users in our experiment repeatedly updated their categorizations and saved results along the way, and one participant even indicated that these features helped him to lay out the plans of his research.

Finally, the third, post-focus stage features specific searches for information, and re-checks for additional information (Vakkari and Hakala, 2000). Searchers may collect information pertinent to their focused topic (Kuhlthau, 2004). At this stage, users are able to "express precisely what information is needed", and encounter low uniqueness, and high redundancy of information (Kuhlthau, 2004). In our experiment, participants performed long, specific queries at this stage, and frequently reopened previous URLs and queries (via the personalizable features). The importance of the *control*, and to a lesser extent, *input* features further declined in the post-focus stage. *Personalizable* features on the other hand, were used relatively often. These features allowed users to keep track of their previous searches and captured material.

Besides insights into *when* SUI features may be useful, our findings have shown that some features were frequently used in an active way, while others were used more passively, but still received a high user rating. Hence, some features, like the Query Box and Result List, directly support users in performing their task, while other features, such as the Recent Queries feature, provide more indirect support, for example by providing context or help them manage their task progress.

Most web search systems have converged over fairly static and familiar designs, where some trialled features, such as the *personalizable* SearchPad (Donato et al., 2010) feature and Google's Wonder Wheel *control* feature, have struggled to provide value for searchers. This is perhaps because, at the wrong times, SUI features can actually impede search (Diriye et al., 2010). Conversely, these more novel control and personalizable features appear consistently in systems like online retail stores, where users are more likely to perform more complex tasks. The results of our work help to provide the insights needed to consider *when* SUI features might be useful during evolving search episodes, such that we could design responsive SUIs that introduce features at the times when they provide value, even on web search. This pleads for UIs that adapt to the needs of the task and task stage at hand.

Our results provide characteristics of behavior observed as students transfer between different stages of a complex essay-writing task, and thus could be used to detect when live users are in pre-focus, focus, and post-focus stages. In future work, this may be extended to other types of complex tasks. Furthermore, this study has focused on one user population (undergraduate students in Computer Science), therefore future work could expand towards other user groups, such as students in other disciplines, as well as more experienced researchers. Future work may also consider turning the analysis around, and try to train a classifier to accurately detect which stage a user is in. Our results, however, indicate that *control* features also need to evolve with the maturity of the the users knowledge level, in effect becoming more *personalizable*. On the other hand, our results also suggest that *personalizable* features provide more support after users move on from initial querying stages. These results support, for example, the premise behind Golovchinsky's work on Querium (Golovchinsky et al., 2012), which personalized control features with metadata about a users search history, to give users filters that develop with their task over time. This naturally leads to further research into task-aware search systems (Kelly et al., 2013) and into additional features which could be useful at different stages (such as co-author visualizations, or user hints and assistance), as well as research into functions which could support interruptions and reinitiating complex search tasks. Thus, future work should directly test how dynamic provision of SUI features does support searchers when exhibiting behavior indicative of different stages, without being impeded when features are not needed. This complex tension between support and impedence, however, is challenging to study.

Concluding, our findings suggest that the active, passive and perceived utility of SUI features across stages, especially in the context of complex and learning tasks, is inherently *dynamic* with different types of features being useful in different task stages. This is in line with macro-level information seeking models, describing broad changes in information behavior across stages, and sheds light on the type of support needed in each stage. This provides new handles to overcome the largely *static* support for information seeking in current search systems, and facilitate a move towards more dynamic and responsive SUIs, providing tailored support to different information seeking stages.

Part III

General Conclusions

6

Conclusions

This section concludes the thesis by synthesizing the main findings of the four research questions (Section 6.1) as well as with a higher-level discussion of our findings (Section 6.2). This chapter summarizes the contributions listed in Section 1.5.

6.1 Research Questions

The first research question of this thesis (**RQ1**), investigated in Chapter 2, was the following: To what extent do search-based web archive access tools facilitate research in a new media setting? An action research methodology was used, which involved the implementation and evaluation of a full-text search system for the Dutch web archive. Our findings indicated that new media scholars were able to address a whole range of novel research questions using a searchable web archive, as opposed to previous URL-based access methods. However, also additional needs for selection and collection making methods, analysis types and transparence emerged. Delving further into this issue in a broader context, a structured literature analysis of 18 journal papers in media and communication studies was performed. Here, we saw that scholars have used varied and combined methods to generate their corpus, to analyze data and to disseminate analysis' results. In a web archive context, this revealed limitations of the traditional search approach, which does not constitute the rich functionality needed for fully supporting the research process.

Conclusion 1.1 The first general conclusion of RQ1 is that search interfaces for web archives lack the **transparency** needed for scholarly research. Standard full-text search approaches predominantly focus on lists of results and their metadata, thereby hiding a whole range of variables. Figure 6.1 (1) summarizes the myriad of influences on the eventual results retrieved in response to a researcher's query. First of all, a curator selects material to be included in the web archive, influenced by his or her context (for instance, legal limitations and institutional selection policies). This selected material, contained in the 'seedlist', is harvested by a *crawler*, which iteratively follows links on the web and captures encountered content. The crawler settings (which influence for instance harvesting strategy, and included file formats), are generally implemented by a crawl engineer. This results in a dataset harvested by the crawler, eventually indexed by a (full-text) *search system*. Typically, a search engineer may customize the indexing settings (e.g. which harvested data types are indexed), and the retrieval settings (e.g. which ranking algorithm to use). These settings influence the final results displayed by a *search interface*, which is the visible intermediary between researcher and data. A user interface designer may determine how results are displayed for a certain query, for instance as a set of 'ten blue links'. All these dependencies together influence exactly which results a researcher retrieves. Insights into the completeness of a resultset are obscured, even though this resultset may become part of a scholars' research corpus. Implications of this conclusion are that more contextual information should be made available by web archiving institutions, and integrated in access interfaces.

Conclusion 1.2 Our second main conclusion derived from RQ1 is that a lack of **process support** exists in traditional full-text search systems, as visualized in Figure 6.1 (2). Based on previous literature, we identified three important phases in web archive research: *corpus creation*, *analysis* and *dissemination*. In many cases, building a corpus by collecting data is a pivotal part of the research process. A first issue of using full-text search systems for this purpose is that these systems predominantly focus on queries. This leaves out other important ways of selecting content, including list-based selections, and sampling methods. These are essential for delineating corpora within the vast datasets contained in web archives. Second, direct analysis methods are generally not supported. Methods include content analysis, both automated and manual, and network analysis. Also, search systems usually only provide access at the page-level, instead of other levels of granularity used in scholars' analysis (in particular, page elements, websites, and web spheres). Third, researchers may visualize data at different moments of the process. Full-text search systems, however, usually do not provide the means to visualize results, nor to visualize custom corpora. Hence, search systems do not constitute the functionality to support key research activities in complex, research-based tasks. Implications are that designers of search-based access systems for web archives should rethink their approaches to support more research methods, in particular in terms of selection options and granularities, corpus creation features, analysis and customized visualization possibilities along the research process.

Subsequently, we focused on ways to improve upon the lack of transparency in web archives (as pinpointed in **Concl. 1.1**). This led to the second research

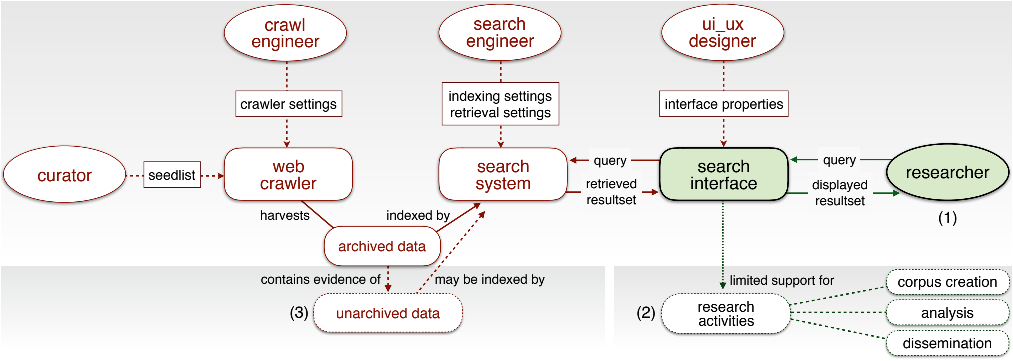

Figure 6.1: The eventual results retrieved by a researcher (**1**) are influenced by a number of 'hidden' actors and interactions, while systems often lack support for research activities (**2**) in different phases of scholars' research process. We may increase transparency by harnessing the link structure and anchor text of an archive (**3**), and increase the coverage of an archive.

question of this thesis (**RQ2**), studied in Chapter 3: To what extent can representations of unarchived webpages and websites enhance search-based access to web archives? Specifically, we performed a quantitative study analyzing the contents of the Dutch web archive. Considering the inherent incompleteness of web archives, we explored how representations of unarchived web material can be generated from evidence in the archive, and how rich (or sparse) these representations are. Our findings show that a web archive contains more than meets the eye: link structure and anchor text evidence can be used to uncover unarchived material, as well as to generate representations of unarchived contents. Thus, they can potentially provide more context and transparency to web archive search tools.

Conclusion 2.1 The first conclusion of RQ2 is that the **coverage** of web archives search systems may be substantially increased by incorporating link and anchor text evidence, especially in the case of selection-based archives. In our study, focusing on the Dutch web archive, we detected roughly the same number of unarchived pages as pages stored in the archive. Further analysis of these uncovered pages showed that they complement the pages collected based on the selection policies, and provide context to the Dutch web at large as they include the most popular websites in the Netherlands. Implications of this conclusion are that (selection-based) archives can use information derived from the archive's link structure for enriching selection criteria and customizing crawler settings.

Conclusion 2.2 After measuring the size and diversity of the unarchived webpages surrounding the Dutch web archive, we devised methods to **reconstruct representations** of the found 'lost' pages. We observed that the representations which can be generated for unarchived pages based on link structure and anchor text are succinct in nature, due the skewed nature of link evidence – the majority of pages has a low number of link and anchor text words, and a minority has rich link evidence. Aggregating representations at the host level partially improved the richness of representations. An evaluation showed that the created representations of unarchived pages and sites, though succinct in nature, may be used for effective retrieval. Hence, integration of these types of representations into web archive search systems may increase coverage of the web archive, and provide more transparency about the completeness, or *incompleteness*, of an archive. This is visualized in Figure 6.1 (3). Implications of the conclusion are that institutional web archives can use representations of unarchived web contents to enrich their metadata and ultimately contextualize access tools.

Conclusion **1.1** and **1.2** suggested the need to increase *transparency* of web archives by incorporating contextual information in search-based access interfaces, and the need to support the *process*. This leads to the question of how this intricate combination of rich information and process support may be integrated into access interfaces, without making them overly complex. As information seeking plays an important role in the research process, Part II aimed at obtaining a better understanding of the information seeking process, as well as insights into how access interfaces may provide support for it. Studies in this part moved the focus from web archives to the 'live' web, and from researchers to students performing research-based tasks in an academic setting.

In this context, the third research question of this thesis (**RQ3**) was investigated in Chapter 4: What are the conceptual implications of multistage information seeking models for the design of search systems? To address this question, we first performed a qualitative analysis of previous literature. The main finding is that there is a gap between macro-level information seeking stages and micro-level search interface features, and that further insights are needed into how concrete search interface features support stages. A subsequent small-scale quantitative analysis of data from a user study involving book search with 12 participants provided indications that some types of search system features are search stage-sensitive, while other features are useful in all stages.

Conclusion 3.1 The first conclusion in relation to RQ3 is that there are issues in the **translation** from the rich stages described in information seeking literature to concrete support in terms of search system features, and vice versa. Information seeking models provide a rich overview of stages occurring during information-intensive tasks involving the construction of knowledge. In particular, their effect on information sought, assessed relevance and search tactics has been generously studied in previous literature. Information sought may evolve from general to specific information, while relevance of encountered information items is dynamic over the course of different stages, being hard to evaluate for a user in the beginning, and easier to judge in later stages. Search terms, tactics and operators evolve as searchers are better able to express their needs over time. However, the translation of these aspects into actual support in terms of search system features remains unclear. Current search applications have converged to streamlined approaches, although specialized search interface features to support complex tasks in the context of exploratory search and sensemaking exist. Nevertheless, the overall understanding of the *utility* of search interface features in different macro-level stages of information seeking was fragmented, and we aimed to improve upon this situation in the remainder of the thesis.

Conclusion 3.2 Our second conclusion derived from RQ3 is that differences in the *micro*-level flow of search user interface feature use may occur during distinct *macro*-level information seeking stages. Contemporary full-text search systems, however, mainly provide support for micro-level cycles of interaction, as opposed to the macro-level information seeking process. Within different phases of search sessions, subtle changes in the flow of interaction with specific user interface features may occur over time, as evidenced in the previous literature and our study. On the one hand, there is no clear dichotomy between the stages: feature use changes only gradually over time, especially *informational* features like results lists. On the other hand, other *input* and *personalizable* features appear to search stage sensitive. This implies that distinguishing **search stage sensitive features** may aid in future search user interface design, as features useful in specific stages may be adaptively offered to a user, reducing cognitive (over)load at the interface level. This is visualized in Figure 6.2. The search stage of a user may be manually indicated or automatically detected, and depending on the search stage, search interface functionality and search system ranking may be adapted. At the interface level, initial (*pre-focus*) stages might involve more assistive features, useful for exploration and the construction of a knowledge frame. Late (*post-focus*) stages may need less assistive features, but more features which provide task support. At the system level, ranking of search results may potentially be adapted, focusing on introductory and general sources in early stages, and on specific sources in late stages.

As we needed a further understanding of exactly how the usefulness of specific types of search functionality evolves, we investigated the fourth research question of this thesis (**RQ4**) in Chapter 5: How can different types of search user interface features support distinct macro-level information seeking task stages? This was researched via a user study with 26 participants, in which the users had to perform three distinct tasks, representing *pre-focus*, *focus* and *postfocus* stages. Using extensive logging and tracking, we gained insights into the active and passive use of features. Questionnaires and interviews, on the other hand, provided indications of how useful the users perceived the features to be over time, allowing for triangulation of findings. The main finding of Chapter

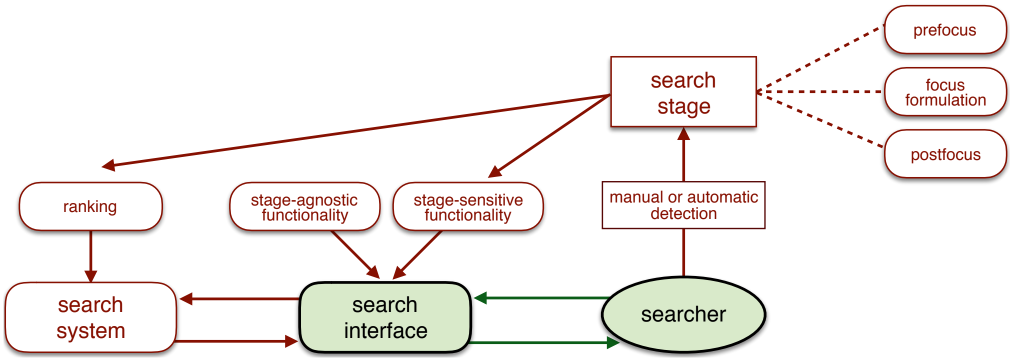

Figure 6.2: Towards *stage-aware* systems: functionality of search systems may be influenced by search stages at the system and interface level.

5 is that within a multistage task involving knowledge construction, the active, passive and perceived usefulness of SUI features differ per information seeking stage.

Conclusion 4.1 The first conclusion of RQ4 is that some SUI features are frequently used in an active way, while other features are used more passively. Certain types of commonly used features, for instance the query box and results list, may directly support users in performing a task. Other features, such as a recent queries feature, were still rated highly by users, despite a lack of active use. The perceived usefulness as expressed by the participants in our study suggests that these features provide indirect support. These types of functionality may provide context to a user, or aid a user in managing the task at hand. This implies that functionality which seems rarely used based on classical interaction measures may still provide value to users. Hence, we arrive at the insight that features can provide **direct and indirect** support to a searcher's overarching work task.

Conclusion 4.2 In line with macro-level information seeking models, describing broad changes in information behavior during complex information-intensive tasks, we conclude that different types of features are useful in different search stages. Hence, we conclude that there is a **dynamic utility of SUI features** across stages occurring in complex and learning tasks. First of all, we conclude that *informational features*, showing search results or information about results, are naturally useful in all information seeking stages, hence are stage agnostic. This confirms findings from the previous chapter; evidently, searchers may assess and utilize search results in all stages of an information seeking venture. Second, *input* and *control features* can be categorized as search stage sensitive features. The value of these features, which include the query box, search filters and search suggestions, is highest in the initial prefocus stage, and decreases over time. This reflects a user's increasing understanding of a topic, during which the value of features to help formulating a query and delimiting a resultset may decrease. Conversely, *personalizable features*, including recent queries and saved results, clearly get more useful in the focus formulation and postfocus stages, as they may "grow" with a user's understanding during the information journey.

6.2 Main Conclusion and Discussion

Here, we return to the main research question raised in the beginning of this thesis: To what extent do current search approaches support complex information-intensive tasks which involve web content, and how can we support the complex dynamics of the information seeking process?

Inspired by the observation that web archives are underused, despite their coverage of our online history since the 1990s, the first part of this thesis looked at the support for research use of web archives. Contributions included a better understanding of the emergent search access needs of scholars using the web as a data source, and of how search system features support the research process. We uncovered limitations of current search approaches: a lack of transparency of current search systems, but also a lack of process support, i.e. support for different inherent phases of complex research-based tasks. Subsequently, we contributed methods to reveal and contextualize what material is missing from the web archive, based on its link structure and anchor text, as well as an assessment of the utility of generated document representations of unarchived contents. We showed that these types of representations can provide additional contextualization for future web archive search systems, thus increasing transparency.

As information seeking is pivotal in current system-mediated research processes, we needed a better understanding of the dynamics of the complex information seeking process, and the support offered by search-based access systems. In the second part of the thesis, our contributions included an assessment of the limitations of contemporary online web search approaches in the context of information-intensive tasks. With respect to search support, we showed a lack of understanding in two directions: the implications of macro-level task stages for micro-level search system design were fuzzy, but also the exact support various micro-level features offer for macro-level stages was unclear. Subsequently, we contributed insights into the connections between the stages and search system features. We showed possibilities to observe evidence of macro-level stages by studying user interactions with search systems at the micro-level. We also contributed to the understanding of passive and active use of search interface features, as well as the moments when (categories of) features are most useful in complex search processes. Using measures capturing active, passive and perceived usefulness of features, we concluded that the utility of search system

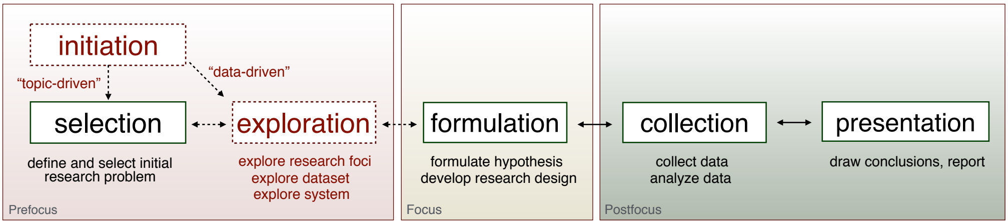

Figure 6.3: Towards a 'constructive' research process approach: integrating Kendall (2012)'s research process model into Kuhlthau (2004)'s ISP model.

and interface features varies during different information seeking stages. This inspires more dynamic approaches to search support, further discussed below.

The encountered value of using the insights from information seeking models for designing search support in Part II of this thesis may also inspire new search approaches for the research context discussed in Part I. While beyond the horizon of this thesis, we may use macro-level research process models derived from scholarly practices to inform the design of search systems supporting research at the micro-level. Formal steps in the research process, such as the formulation of a research problem, corpus creation, analysis and dissemination encompass similarities with stages in Kuhlthau's model and Vakkari's adaptation (see Figure 6.3). However, the models' cognitive and affective aspects, but also the initial exploratory *pre-focus* stages are usually not explicitly included in research process models. Exploration may actually become the starting point of the process in emerging data-driven research facilitated by large-scale computing power. This further underlines the need for transparent search systems, as search tools which produce results, summaries and visualizations of data without revealing their internal data and parameters, may lead to potentially incorrect insights. Future research may look into these aspects, as well as how dynamic approaches may be applied to specific disciplines with different inherent needs and workflows.

The finding that the usefulness of search system features evolves over time provides new inspiration for designing novel search user interfaces. Instead of focusing on singular SUI features, or improving performance of specific ranking algorithms, we propose a more holistic and dynamic approach. This dynamic approach aims at supporting a continuum of macro-level stages in search systems, as opposed to just micro-level interactions. Next, we discuss how this approach may constitute a **helpful framework**[1] for complex tasks.

---

1 A term introduced by Oddy (1977): "The job of the mechanical part of the system is to create a *helpful framework* within which the user can make problem-solving decisions."

Towards a Helpful Framework for Complex Tasks

Tremendous advances in information retrieval technology have occurred during the past decades. We now have arrived at the point where systems may actually *solve* problems for users. For instance, via common search engines on the web we get 'instant answers' for factual questions ranging from the weather in the next week-end to the birthdate of the current prime minister. Information seeking in the context of more complex tasks, however, is not as straightforward: broader inquiries cannot be directly answered in a succint snippet of information. For instance, gaining novel ideas for research, or finding the appropriate sources for writing an essay requires intensive interaction with information sources.

During the process of information seeking and use, as occurring in complex research-based tasks, the needs and understanding of a user may evolve, moving from broad conceptualizations to a focused perspective. Evidently, to arrive at more *helpful frameworks* for complex tasks featuring sustained information interaction, current *ad-hoc* approaches to search-based information interaction should be rethought. The previous focus on optimizing search engine results displayal for responding to singular queries may not be sufficient, since it does not take into account the overarching process. In this thesis, we propose a fundamentally different approach to information interaction in terms of **process support**.

A searcher's conceptual framework about a topic may evolve over time: during a novice user's information journey, knowledge structures evolve, just as during a scholars' research process, conceptualizations of a topic may undergo changes. For instance, a student may start with a topic she knows little about, but this knowledge advances over time, or a researcher may start with a loose research question, which becomes more focused after interaction with the dataset. However, search interfaces do typically not evolve with a user's knowledge – to become a truly 'helpful framework', a system should support the information seeking *process* of a user, moving from exploratory *pre-focus*, to *focus formulation* and final *post-focus* stages. Our proposed framework is visualized by Figure 6.4.

The first dimension of a helpful framework consists of features offering automatically generated suggestions to users. This support typically takes place at the search activity level (Bates, 1990) of the 'move' (e.g. entering search terms), and 'tactic' (e.g. choosing a broader term). For instance, a word cloud feature may suggest keywords for a query, or a query suggestion feature may propose a broader formulation of a query. The need for this *low-level* support, embodied in various *input* and *control* features, generally decreases over time. When a user's conceptualization of a topic grows, she becomes increasingly able to express herself precisely in the context of that topic, and support at the level of moves and tactics becomes more superfluous.

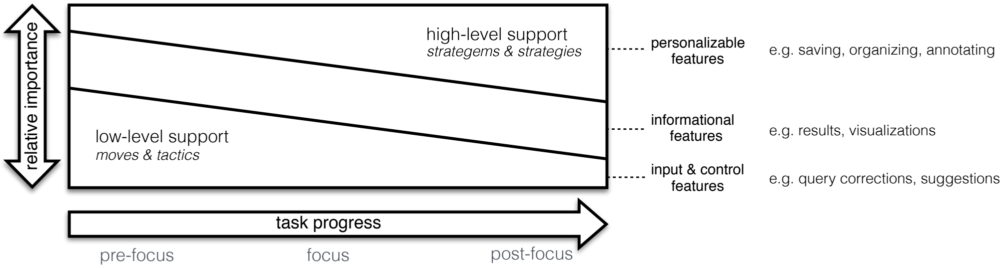

Figure 6.4: Schematic overview of a helpful framework for complex task performance using search systems: low-level support for moves and tactics gradually gives way to higher level support for stratagems and strategies.

The second dimension of the helpful framework is formed by *informational* features. These features provide the actual results, or information about encountered result items. For instance, a search system may show the title of a document, a short snippet and basic metadata. As evidenced in our experiments, these features may be useful throughout the process. They provide low-level support at the move and tactic level, for instance selecting and opening information sources, but also higher level support (for instance by providing visualizations). Extensions are possible: for instance, an author subject search 'stratagem' (a combination of moves and/or tactics) would be facilitated by a feature to group information sources by their author.

The third dimension of a helpful framework consists of features which can support seeking at a higher level. While these types of features may include automated functionality, the main aim is to provide insights into a user's process *through her actions*. As Kuhlthau's model has indicated, processes of hypothesis generation, data collection, information organization and the preparation of a personalized synthesis of a topic take place during processes of knowledge construction. This reflects the highly personalized nature of such complex activities, meaning that automated support may not suffice. Instead, the aim of *personalizable* features should be to aid user's in performing their task. The two parts of this thesis have shown the demand for and use of annotation, saving and organization features by both students and graduate researchers. As opposed to low-level features, these higher-level features may support Bates' 'stratagems' and 'strategies' (planning in the context of an entire search). On the one hand, through logging user's actions and potentially gathering data about the actors' domain knowledge or task at hand, they provide a trail of activities, which may (passively) aid users in locating where they are in the process. On the other hand, they also allow a user to 'work with results', and thus encourage reflection on encountered results. As such, they become increasingly useful throughout a task.

Summarizing, more dynamic support for complex research-based tasks may be achieved by differentiating SUI feature categories and their levels of support. In particular, functionality providing low-level support (in particular *input* and *control features*), are useful in the initial stages of a complex research-based task. Searchers with low domain knowledge, but also researchers exploring a new topic and collection may utilize this functionality to bootstrap their searches. Features providing high-level support (in particular *personalizable* features), may invite searchers to explicitly reflect and interact with results, as well as seeing how these results fit in their process and strategy. In effect, this conclusion trancends the two individual parts of the thesis, as both actors with low domain knowledge (e.g. students) and actors with high domain knowledge (e.g. researchers), may use the different types of features at different stages of their complex tasks.

Current information retrieval approaches utilize extremely advanced algorithms at the back-end, but limited explicit interaction at the front-end: a user enters a short query to receive a list of 'ten blue links', hiding underlying variables and usually independent of the complex process a user is engaged in. To better facilitate complex information-intensive tasks using web archives and the live web, we argue that future research should rethink this approach, thus moving towards more **dynamic support for the complex dynamics of the information seeking process**.

Bibliography

C. Ahlberg and B. Shneiderman. Visual information seeking: Tight coupling of dynamic query filters with starfield displays. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '94, pages 313–317, 1994. ACM. `http://dx.doi.org/10.1145/191666.191775`. (Cited on page 116.)

S. Ainsworth, A. Alsum, H. SalahEldeen, M. C. Weigle, and M. L. Nelson. How much of the web is archived? *ArXiv e-prints*, abs/1212.6177, 2012. `http://arxiv.org/abs/1212.6177`. (Cited on page 21.)

J. Allan, B. Croft, A. Moffat, and M. Sanderson. Frontiers, challenges, and opportunities for information retrieval: Report from SWIRL 2012 the second strategic workshop on information retrieval in lorne. *SIGIR Forum*, 46:2–32, June 2012. `http://dx.doi.org/10.1145/2215676.2215678`. (Cited on page 134.)

A. AlSum, M. Weigle, M. Nelson, and H. Van de Sompel. Profiling web archive coverage for top-level domain and content language. *Int J on Digital Libraries*, 14:149–166, 2014. `http://dx.doi.org/10.1007/s00799-014-0118-y`. (Cited on page 64.)

R. Armstrong, D. Freitag, T. Joachims, and T. Mitchell. Webwatcher: A learning apprentice for the world wide web. In *AAAI Spring symposium on Information gathering from Heterogeneous, distributed environments*, pages 6–12. AAAI, 1995. `https://www.aaai.org/Papers/Symposia/Spring/1995/SS-95-08/SS95-08-002.pdf`. (Cited on page 116.)

A. Ball. DCC State of the Art Report: Web Archiving. Technical Report, University of Edinburgh; UKOLN, University of Bath; HATII, University of Glasgow; Science and Technology Facilities Council, Jan. 2010. `https://www.era.lib.ed.ac.uk/handle/1842/3327`. (Cited on page 21.)

J. Bar-Ilan and B. C. Peritz. The lifespan of "informetrics" on the Web: An eight year study (1998-2006). *Scientometrics*, 79(1):7–25, Nov. 2008. `http://dx.doi.org/10.1007/s11192-009-0401-7`. (Cited on page 50, 51, 53.)

S. Bateman, J. Teevan, and R. W. White. The Search Dashboard: How Reflection and Comparison Impact Search Behavior. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 1785–1794, 2012. ACM. `http://dx.doi.org/10.1145/2207676.2208311`. (Cited on page 129, 134.)

M. J. Bates. Information search tactics. *J Am Soc Inform Sci*, 30(4):205–214, 1979. `http://dx.doi.org/10.1002/asi.4630300406`. (Cited on page 111, 120.)

M. J. Bates. Where should the person stop and the information search interface start? *Inform Process Manag*, 26(5):575–591, Jan. 1990. `http://dx.doi.org/10.1016/0306-4573(90)90103-9`. (Cited on page 44, 113, 175, 176.)

M. Beaulieu. Interaction in information searching and retrieval. *J Doc*, 56(4):431–439, Aug. 2000. http://dx.doi.org/10.1108/EUM0000000007122. (Cited on page 2, 114.)

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 (OSIR 2012)*, pages 9–16, 2012. http://www.is.inf.uni-due.de/bib/pdf/ir/Beckers_etal_12.pdf. (Cited on page 121.)

B. B. Bederson. Interfaces for staying in the flow. *Ubiquity*, 2004. http://dx.doi.org/10.1145/1074068.1074069. (Cited on page 118, 119, 131.)

J. Beheshti, C. Cole, D. Abuhimed, and I. Lamoureux. Tracking middle school students' information behavior via Kuhlthau's ISP Model: Temporality. *J Am Soc Inf Sci Tec*, 2014. http://dx.doi.org/10.1002/asi.23230. (Cited on page 107.)

N. Belkin, R. Oddy, and H. Brooks. ASK for information retrieval: Part I. Background and theory. *J Doc*, 38(2):61–71, Feb. 1982. http://dx.doi.org/10.1108/eb026722. (Cited on page 127.)

N. J. Belkin. Helping People Find What They Don't Know. *Commun ACM*, 43(8):58–61, Aug. 2000. http://dx.doi.org/10.1145/345124.345143. (Cited on page 127, 128.)

N. J. Belkin, C. Cool, A. Stein, and U. Thiel. Cases, scripts, and information-seeking strategies: On the design of interactive information retrieval systems. *Expert Systems with Applications*, 9(3):379–395, Jan. 1995. http://dx.doi.org/10.1016/0957-4174(95)00011-W. (Cited on page 138.)

A. Ben-David and H. C. Huurdeman. Web Archive Search as Research: Methodological and Theoretical Implications. *Alexandria*, 25(1-2):93–111, Aug. 2014. http://dx.doi.org/10.7227/ALX.0022. (Cited on page 19, 28, 31, 33, 47.)

P. Borlund. 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. (Cited on page 9, 141.)

A. Bozzon, M. Brambilla, S. Ceri, and D. Mazza. Exploratory search framework for web data sources. *VLDB J*, 22(5):641–663, Oct. 2013. http://dx.doi.org/10.1007/s00778-013-0326-x. (Cited on page 116.)

A. Z. Broder, E. Gabrilovich, V. Josifovski, G. Mavromatis, D. Metzler, and J. Wang. Exploiting site-level information to improve web search. In *Proceedings of the 19th ACM Conference on Information and Knowledge Management*, CIKM 2010, pages 1393–1396. ACM, 2010. http://dx.doi.org/10.1145/1871437.1871630. (Cited on page 66, 70.)

M. Bron, J. Van Gorp, and M. de Rijke. Media studies research in the data-driven age: How research questions evolve. *J Am Soc Inf Sci Tec*, 67(7):1535–1554, July 2016. http://dx.doi.org/10.1002/asi.23458. (Cited on page 49.)

A. Brown. *Archiving websites: a practical guide for information management professionals*. Facet, 2006. (Cited on page 1, 22, 24, 25, 27, 28.)

N. Brügger. *Archiving Websites - General Considerations and Strategies*. The Centre for Internet Research, 2005. (Cited on page 22, 26.)

N. Brügger. Website history and the website as an object of study. *New Media Soc*, 11(115), Mar. 2009. http://dx.doi.org/10.1177/1461444808099574. (Cited on page 8, 22, 26, 52, 203.)

N. Brügger. Web Archiving Between Past, Present, and Future. *The Handbook of Internet Studies*, pages 24–42, 2011. (Cited on page 24, 26.)

N. Brügger. Web history and the web as a historical source. *Zeithistorische Forschungen*, 9:316–325, 2012. http://www.zeithistorische-forschungen.de/file/2880. (Cited on page 64.)

J. F. Brunelle, M. Kelly, H. SalahEldeen, M. C. Weigle, and M. L. Nelson. Not all mementos are created equal: Measuring the impact of missing resources. In *IEEE/ACM Joint Conference on Digital Libraries*, pages 321–330. IEEE, 2014. http://dx.doi.org/10.1109/JCDL.2014.6970187. (Cited on page 64.)

P. Brusilovsky and M. T. Maybury. From adaptive hypermedia to the adaptive web. *Commun ACM*, 45(5):30–33, 2002. http://dx.doi.org/10.1145/506218.506239. (Cited on page 104.)

M. K. Buckland. What is a document? *J Am Soc Inform Sci*, 48(9):804–809, Sept. 1997. http://dx.doi.org/10.1002/(SICI)1097-4571(199709)48:9<804::AID-ASI5>3.0.CO;2-V. (Cited on page 8.)

G. Buscher, A. Dengel, and L. van Elst. Eye movements as implicit relevance feedback. In *CHI '08 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '08, pages 2991–2996, 2008. ACM. http://dx.doi.org/10.1145/1358628.1358796. (Cited on page 145.)

K. Byström and K. Järvelin. Task complexity affects information seeking and use. *Inform Process Manag*, 31(2):191–213, 1995. http://dx.doi.org/10.1016/0306-4573(95)80035-R. (Cited on page 9, 107, 137.)

R. Capra, J. Arguello, A. Crescenzi, and E. Vardell. Differences in the Use of Search Assistance for Tasks of Varying Complexity. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 23–32, 2015. ACM. http://dx.doi.org/10.1145/2766462.2767741. (Cited on page 121.)

D. O. Case. *Looking for Information : a Survey of Research on Information Seeking, Needs, and Behavior*. Emerald, 2012. (Cited on page 48, 106, 107.)

C. M. Chu. Literary critics at work and their information needs: A research-phases model. *Libr Inform Sci Res*, 21(2):247–273, 1999. http://dx.doi.org/10.1016/S0740-8188(99)00002-X. (Cited on page 48, 49.)

A. Y. K. Chua and S. Banerjee. So fast so good: An analysis of answer quality and answer speed in community Question-answering sites. *J Am Soc Inf Sci Tec*, 64(10):2058–2068, Oct. 2013. http://dx.doi.org/10.1002/asi.22902. (Cited on page 52, 54.)

C. Cole. A theory of information need for information retrieval that connects information to knowledge. *J Am Soc Inf Sci Tec*, 62(7):1216–1231, 2011. http://dx.doi.org/10.1002/asi.21541. (Cited on page 106, 112.)

K. Collins-Thompson, P. N. Bennett, R. W. White, S. de la Chica, and D. Sontag. Personalizing Web Search Results by Reading Level. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 403–412, 2011. ACM. http://dx.doi.org/10.1145/2063576.2063639. (Cited on page 115.)

D. Comeaux and A. Schmetzke. Accessibility of academic library web sites in North AmericaCurrent status and trends (2002-2012). *Libr Hi Tech*, 31(1):8–33, Feb. 2013. http://dx.doi.org/10.1108/07378831311303903. (Cited on page 50, 51, 53, 54.)

N. Craswell, D. Hawking, and S. Robertson. Effective site finding using link anchor information. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, pages 250–257, 2001. ACM. http://dx.doi.org/10.1145/383952.383999. (Cited on page 65, 66.)

N. Dai and B. D. Davison. Mining anchor text trends for retrieval. In *Advances in Information Retrieval, 32nd European Conference on IR Research*, volume 5993 of *LNCS*, pages 127–139. Springer, 2010. http://dx.doi.org/10.1007/978-3-642-12275-0. (Cited on page 66.)

E. S. Dalmaijer, S. Mathôt, and S. Van der Stigchel. PyGaze: An open-source, cross-platform toolbox for minimal-effort programming of eyetracking experiments. *Behav Res Meth*, 46(4): 913–921, 2013. http://dx.doi.org/10.3758/s13428-013-0422-2. (Cited on page 145.)

M. Day. Preserving the fabric of our lives: A survey of web preservation initiatives. In *Research and Advanced Technology for Digital Libraries: 7th European Conference, ECDL '03*, volume 2769 of *LNCS*, pages 461–472. Springer, 2003. http://dx.doi.org/10.1007/978-3-540-45175-4_42. (Cited on page 64.)

A. Diriye, A. Blandford, and A. Tombros. When is system support effective? In *Proceedings of the Third Symposium on Information Interaction in Context*, IIiX '10, pages 55–64. ACM, 2010. http://dx.doi.org/10.1145/1840784.1840794. (Cited on page 119, 136, 139, 140, 162.)

A. Diriye, A. Blandford, A. Tombros, and P. Vakkari. The role of search interface features during information seeking. In *Research and Advanced Technology for Digital Libraries: International Conference on Theory and Practice of Digital Libraries, TPDL '13*, volume 8092 of *LNCS*, pages 235–240. Springer, 2013. http://dx.doi.org/10.1007/978-3-642-40501-3_23. (Cited on page 121, 125.)

D. Donato, F. Bonchi, T. Chi, and Y. Maarek. Do You Want to Take Notes?: Identifying Research Missions in Yahoo! Search Pad. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 321–330, 2010. ACM. http://dx.doi.org/10.1145/1772690.1772724. (Cited on page 136, 139, 162.)

Z. Dou, R. Song, J.-Y. Nie, and J.-R. Wen. Using anchor texts with their hyperlink structure for web search. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 227–234. ACM, 2009. http://dx.doi.org/10.1145/1571941.1571982. (Cited on page 65.)

M. Dougherty and E. T. Meyer. Community, tools, and practices in web archiving: The state-of-the-art in relation to social science and humanities research needs. *J Am Soc Inf Sci Tec*, 65(11):2195–2209, 2014. http://dx.doi.org/10.1002/asi.23099. (Cited on page 21, 23, 24, 54.)

C. Dunne, B. Shneiderman, R. Gove, J. Klavans, and B. Dorr. Rapid understanding of scientific paper collections: Integrating statistics, text analytics, and visualization. *J Am Soc Inf Sci Tec*, 63(12):2351–2369, 2012. http://dx.doi.org/10.1002/asi.22652. (Cited on page 118, 119.)

D. E. Egan, J. R. Remde, L. M. Gomez, T. K. Landauer, J. Eberhardt, and C. C. Lochbaum. Formative design evaluation of superbook. *ACM Trans Inf Syst*, 7(1):30–57, Jan. 1989. http://dx.doi.org/10.1145/64789.64790. (Cited on page 116.)

C. Eickhoff, J. Teevan, R. White, and S. Dumais. Lessons from the Journey: A Query Log Analysis of Within-session Learning. In *Proceedings of the 7th ACM International Conference* *on Web Search and Data Mining*, WSDM '14, pages 223–232, 2014. ACM. http://dx.doi.org/10.1145/2556195.2556217. (Cited on page 130.)

M. B. Eisenberg. Information Literacy: Essential Skills for the Information Age. *DESIDOC Journal of Library & Information Technology*, 28(2):39–47, Mar. 2008. http://dx.doi.org/10.14429/djlit.28.2.166. (Cited on page 129.)

M. B. Eisenberg and R. Berkowitz. *Information problem solving: The Big Six skills approach to library & information skills instruction.* Ablex, 1990. (Cited on page 128.)

D. Ellis. A behavioural approach to information retrieval system design. *J Doc*, 45:171–212, 1989. http://dx.doi.org/10.1108/eb026843. (Cited on page 106, 137.)

K. E. Fisher, S. Erdelez, and L. McKechnie, editors. *Theories of information behavior.* Information Today, 2005. (Cited on page 183, 187.)

A. Foster. Nonlinear information seeking. In Fisher et al. (2005). (Cited on page 106, 137.)

A. Foster and N. Ford. Serendipity and information seeking: an empirical study. *J Doc*, 59(3):321–340, June 2003. http://dx.doi.org/10.1108/00220410310472518. (Cited on page 128.)

L. Freund, H. O'Brien, and R. Kopak. Getting the big picture: supporting comprehension and learning in search. In *Proceedings of the Searching as Learning workshop at the Information Interaction in Context conference (IIiX 2014)*, 2014. http://www.diigubc.ca/IIIXSAL/Papers/FreundObrienKopak.pdf. (Cited on page 130.)

A. Fujii. Modeling anchor text and classifying queries to enhance web document retrieval. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pages 337–346. ACM, 2008. http://dx.doi.org/10.1145/1367497.1367544. (Cited on page 65.)

M. Gäde, M. Hall, H. Huurdeman, J. Kamps, M. Koolen, M. Skov, E. Toms, and D. Walsh. Overview of the INEX 2015 Interactive Social Book Search Track. In *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum*, volume 1391 of *CEUR Workshop Proceedings*, 2015. http://ceur-ws.org/Vol-1391/78-CR.pdf. (Cited on page 130, 134.)

M. Gäde, M. Hall, H. Huurdeman, J. Kamps, M. Koolen, M. Skov, E. Toms, and D. Walsh. Overview of the INEX 2016 Interactive Social Book Search Track. In *Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum*, volume 1609 of *CEUR Workshop Proceedings*, 2016. http://ceur-ws.org/Vol-1609/16091024.pdf. (Cited on page 134.)

S. Ghobadi and S. Clegg. These days will never be forgotten : A critical mass approach to online activism. *Information and Organization*, 25(1):52–71, Jan. 2015. http://dx.doi.org/10.1016/j.infoandorg.2014.12.002. (Cited on page 52, 54.)

J. Goecks, A. Nekrutenko, J. Taylor, and Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol*, 11(8):R86, 2010. http://dx.doi.org/10.1186/gb-2010-11-8-r86. (Cited on page 57.)

G. Golovchinsky, A. Diriye, and T. Dunnigan. The future is in the past: Designing for exploratory search. In *Proceedings of the 4th Information Interaction in Context Symposium*, IIIX '12, pages 52–61, 2012. ACM. http://dx.doi.org/10.1145/2362724.2362738. (Cited on page 116, 118, 163.)

D. Gomes and M. J. Silva. Characterizing a national community web. *ACM Trans Internet Technol*, 5(3):508–531, Aug. 2005. http://dx.doi.org/10.1145/1084772.1084775. (Cited on page 64.)

G. E. Gorman and P. Clayton. *Qualitative Research For The Information Professional: A Practical Handbook*. Facet Publishing, 2nd edition, Feb. 2005. (Cited on page 40.)

M. Hall, H. Huurdeman, M. Koolen, M. Skov, and D. Walsh. Overview of the INEX 2014 interactive social book search track. In *Working Notes of CLEF 2014 - Conference and Labs of the Evaluation forum*, volume 1180 of *CEUR Workshop Proceedings*, 2014. http://ceur-ws.org/Vol-1180/CLEF2014wn-Inex-HallEt2014.pdf. (Cited on page 134.)

D. Hawking and N. Craswell. Very large scale retrieval and web search. In *TREC: Experiment and Evaluation in Information Retrieval*, chapter 9. MIT Press, 2005. (Cited on page 91.)

M. Hearst. *Search User Interfaces*. Cambridge University Press, 2009. http://searchuserinterfaces.com/book/. (Cited on page 57, 104, 114, 115, 117, 119, 131, 136, 138, 139.)

M. A. Hearst and D. Degler. Sewing the seams of sensemaking: A practical interface for tagging and organizing saved search results. In *Proceedings of the Symposium on Human-Computer Interaction and Information Retrieval*, HCIR '13, pages 4:1–4:10, 2013. ACM. http://dx.doi.org/10.1145/2528394.2528398. (Cited on page 117.)

A. Helmond. *The web as platform: Data flows in social media*. PhD thesis, University of Amsterdam, 2015. http://dare.uva.nl/record/1/485895. (Cited on page 26, 32.)

H. Hockx-Yu. The past issue of the web. In *Proceedings of the 3rd International Web Science Conference*, WebSci '11, pages 12:1–12:8, 2011. ACM. http://dx.doi.org/10.1145/2527031.2527050. (Cited on page 64.)

H. Hockx-Yu. Access and Scholarly Use of Web Archives. *Alexandria*, 25(1-2):113–127, 2014. http://dx.doi.org/10.7227/ALX.0023. (Cited on page 1, 28, 33, 64.)

C. Hoelscher and G. Strube. Web search behavior of internet experts and newbies. *Comput Netw*, 33:337–346, 2000. http://dx.doi.org/http://dx.doi.org/10.1016/S1389-1286(00)00031-1. (Cited on page 115.)

H. C. Huurdeman. Towards Research Engines: Supporting Search Stages in Web archives. In *Web Archives as Scholarly Sources conference 2015*, Apr. 2015. http://events.netlab.dk/conference/index.php/resaw/june2015/paper/view/85. (Cited on page 19, 31.)

H. C. Huurdeman. Dynamic Compositions: Recombining Search User Interface Features for Supporting Complex Work Tasks. In *CHIIR 2017 Second Workshop on Supporting Complex Search Tasks (SCST 2017)*, pages 21–24. CEUR-WS, 2017. http://ceur-ws.org/Vol-1798/paper5.pdf. (Cited on page .)

H. C. Huurdeman and J. Kamps. From Multistage Information-seeking Models to Multistage Search Systems. In *Proceedings of the 5th Information Interaction in Context Symposium*, IIiX '14, pages 145–154, 2014. ACM. http://dx.doi.org/10.1145/2637002.2637020. (Cited on page 103.)

H. C. Huurdeman and J. Kamps. Supporting the Process: Adapting Search Systems to Search Stages. In *Information Literacy: Moving Toward Sustainability*, number 552 in Communications in Computer and Information Science, pages 394–404. Springer International Publishing, Oct. 2015. http://dx.doi.org/10.1007/978-3-319-28197-1_40. (Cited on page 103.)

H. C. Huurdeman and J. Kamps. A Collaborative Approach to Research Data Management in a Web Archive Context. In *Research Data Management - A European Perspective*. Walter de Gruyter GmbH, 2018. (Cited on page 19.)

H. C. Huurdeman, A. Ben-David, and T. Samar. Sprint Methods for Web Archive Research. In *Proceedings of the 5th Annual ACM Web Science Conference*, WebSci '13, pages 182–190, 2013. ACM. `http://dx.doi.org/10.1145/2464464.2464513`. (Cited on page 19, 31, 33, 35, 36.)

H. C. Huurdeman, A. Ben-David, J. Kamps, T. Samar, and A. P. de Vries. Finding Pages on the Unarchived Web. In *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '14, pages 331–340, 2014. IEEE Press. `http://dx.doi.org/10.1109/JCDL.2014.6970188`. (Cited on page 61, 66.)

H. C. Huurdeman, J. Kamps, M. Koolen, and S. Kumpulainen. The Value of Multistage Search Systems for Book Search. In *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum*, volume 1391, 2015a. CEUR-WS. `http://ceur-ws.org/Vol-1391/85-CR.pdf`. (Cited on page 130, 134.)

H. C. Huurdeman, J. Kamps, T. Samar, A. P. Vries, A. Ben-David, and R. A. Rogers. Lost but Not Forgotten: Finding Pages on the Unarchived Web. *Int J on Digital Libraries*, 16(3): 247–265, 2015b. `http://dx.doi.org/10.1007/s00799-015-0153-3`. (Cited on page 61.)

H. C. Huurdeman, M. L. Wilson, and J. Kamps. 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, pages 3–12, 2016. ACM. `http://dx.doi.org/10.1145/2854946.2854957`. (Cited on page 135.)

P. Ingwersen and K. Järvelin. *The Turn - Integration of Information Seeking and Retrieval in Context*. Springer, 2005. (Cited on page 2, 5, 8, 9, 104, 106, 107, 110, 114.)

J. Jiang, D. He, and J. Allan. 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, pages 607–616, 2014. ACM. `http://dx.doi.org/10.1145/2600428.2609633`. (Cited on page 144.)

N. A. John. Sharing and Web 2.0: The emergence of a keyword. *New Media Soc*, 15(2): 167–182, Mar. 2013. `http://dx.doi.org/10.1177/1461444812450684`. (Cited on page 50.)

B. Johnston and S. Webber. Information Literacy in Higher Education: A review and case study. *Stud High Educ*, 28(3):335–352, Aug. 2003. `http://dx.doi.org/10.1080/03075070309295`. (Cited on page 130.)

B. Kahle. In: Backlight: Digital amnesia, VPRO Documentary, 2014. `https://www.youtube.com/watch?v=NdZxI3nFVJs`. (Cited on page 17.)

J. Kamps. Web-centric language models. In *Proceedings of the 2005 ACM CIKM International Conference on Information and Knowledge Management*, CIKM '05, pages 307–308. ACM, 2005. `http://dx.doi.org/10.1145/1099554.1099640`. (Cited on page 65, 66.)

J. Kamps. Toward a model of interaction for complex search tasks. In *Proceedings of the Fourth Workshop on Exploiting Semantic Annotations in Information Retrieval*, ESAIR '11, pages 7–8. ACM, 2011. `http://dx.doi.org/10.1145/2064713.2064719`. (Cited on page 104.)

J. Kari. Diversity in the conceptions of information use. *Inf Res*, 15(03):colis709, 2010. `http://www.informationr.net/ir/15-3/colis7/colis709.html`. (Cited on page 8.)

M. Karlsson and C. Clerwall. Patterns and Origins in the Evolution of Multimedia on Broadsheet and Tabloid News Sites. *Journalism Stud*, 13(4):550–565, Aug. 2012. http://dx.doi.org/10.1080/1461670X.2011.639571. (Cited on page 54.)

M. Karlsson, C. Clerwall, and H. Örnebring. Hyperlinking practices in Swedish online news 20072013: the rise, fall, and stagnation of hyperlinking as a journalistic tool. *Information Commun Soc*, 18(7):847–863, July 2015. http://dx.doi.org/10.1080/1369118X.2014.984743. (Cited on page 50, 53.)

D. Kelly. Query suggestions as idea tactics for information search. In *Proceedings of the Third Workshop on Human-Computer Interaction and Information Retrieval*, HCIR '09, pages 9–12, 2009. http://sites.google.com/site/hcirworkshop/hcir-2009. (Cited on page 120, 121, 122, 128, 158.)

D. Kelly, J. Arguello, and R. Capra. NSF workshop on task-based information search systems. *SIGIR Forum*, 47(2):116–127, Jan. 2013. http://dx.doi.org/10.1145/2568388.2568407. (Cited on page 2, 101, 163.)

D. E. Kendall. *Sociology in our times: the essentials*. Wadsworth / Cengage Learning, 8th edition, 2012. (Cited on page 48, 49, 174, 204.)

M. Klein and M. L. Nelson. Moved but not gone: an evaluation of real-time methods for discovering replacement web pages. *Int J on Digital Libraries*, 14(1-2):17–38, 2014. http://dx.doi.org/10.1007/s00799-014-0108-0. (Cited on page 65.)

J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632, 1999. http://dx.doi.org/10.1145/324133.324140. (Cited on page 65.)

K. Koc-Michalska, R. Gibson, and T. Vedel. Online Campaigning in France, 20072012: Political Actors and Citizens in the Aftermath of the Web.2.0 Evolution. *J Inform Tech Polit*, 11(2):220–244, Apr. 2014. http://dx.doi.org/10.1080/19331681.2014.903217. (Cited on page 53.)

W. Koehler. An analysis of web page and web site constancy and permanence. *J Am Soc Inf Sci*, 50(2):162–180, Jan. 1999. http://dx.doi.org/10.1002/(SICI)1097-4571(1999)50:2<162::AID-ASI7>3.0.CO;2-B. (Cited on page 8.)

J. Koenemann and N. J. Belkin. A Case for Interaction: A Study of Interactive Information Retrieval Behavior and Effectiveness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '96, pages 205–212, 1996. ACM. http://dx.doi.org/10.1145/238386.238487. (Cited on page 120.)

R. H. Kolbe and M. S. Burnett. Content-analysis research: An examination of applications with directives for improving research reliability and objectivity. *J Cons Res*, 18(2):243–250, 1991. http://dx.doi.org/10.1086/209256. (Cited on page 51.)

M. Koolen and J. Kamps. The importance of anchor text for ad hoc search revisited. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 122–129. ACM, 2010. http://dx.doi.org/10.1145/1835449.1835472. (Cited on page 65.)

M. Koolen, T. Bogers, M. Gäde, M. Hall, H. Huurdeman, J. Kamps, M. Skov, E. Toms, and D. Walsh. *Overview of the CLEF 2015 Social Book Search Lab*, pages 545–564. Springer, 2015. http://dx.doi.org/10.1007/978-3-319-24027-5_51. (Cited on page 130.)

W. Kraaij, T. Westerveld, and D. Hiemstra. The importance of prior probabilities for entry page search. In *Proceedings of the 33rd International ACM SIGIR Conference on Research *and Development in Information Retrieval*, SIGIR '02, pages 27–34. ACM, 2002. http://dx.doi.org/10.1145/564376.564383. (Cited on page 66, 91.)

R. Kraft and J. Zien. Mining anchor text for query refinement. In *Proceedings of the 13th international conference on World Wide Web*, WWW '04, pages 666–674. ACM, 2004. http://dx.doi.org/10.1145/988672.988763. (Cited on page 65.)

C. Kuhlthau, J. Heinstrom, and R. Todd. The 'information search process' revisited: is the model still useful?, 2008. http://www.informationr.net/ir/13-4/paper355.html. (Cited on page 107.)

C. C. Kuhlthau. Inside the search process: Information seeking from the user's perspective. *J Am Soc Inform Sci*, 42(5):361–371, 1991. http://dx.doi.org/cc9pnq. (Cited on page 104, 106, 107.)

C. C. Kuhlthau. Perceptions of the information search process in libraries: a study of changes from high school through college. *Inform Process Manag*, 24:419–427, 1988a. http://dx.doi.org/10.1016/0306-4573(88)90045-3. (Cited on page 107.)

C. C. Kuhlthau. Longitudinal case studies of the information search process of users in libraries. *Libr Inform Sci Res*, 10:257–304, 1988b. (Cited on page 107.)

C. C. Kuhlthau. Accommodating the user's information search process: challenges for information retrieval system designers. *B Am Soc Inform Inf*, 25(3):12–16, 1999. http://dx.doi.org/10.1002/bult.115. (Cited on page 111.)

C. C. Kuhlthau. *Seeking meaning: a process approach to library and information services.* Libraries Unlimited, 2004. (Cited on page 2, 4, 107, 108, 109, 111, 118, 127, 128, 136, 137, 138, 140, 151, 161, 162, 174, 204.)

C. C. Kuhlthau. Kuhlthau's information search process. In Fisher et al. (2005). (Cited on page 107, 108, 109, 206.)

C. C. Kuhlthau and S. L. Tama. Information search process of lawyers: a call for 'just for me' information services. *J Doc*, 57:25–43, 2001. http://dx.doi.org/10.1108/EUM0000000007076. (Cited on page 107.)

B. Kules and R. Capra. Influence of training and stage of search on gaze behavior in a library catalog faceted search interface. *J Am Soc Inf Sci Tech*, 63:114–138, 2012. http://dx.doi.org/10.1002/asi.21647. (Cited on page 120, 125, 139.)

B. Kules and B. Shneiderman. Users can change their web search tactics: Design guidelines for categorized overviews. *Inform Process Manag*, 44(2):463–484, 2008. http://dx.doi.org/10.1016/j.ipm.2007.07.014. (Cited on page 141.)

B. Kules, R. Capra, M. Banta, and T. Sierra. What do exploratory searchers look at in a faceted search interface? In *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '09, pages 313–322. ACM, 2009. http://dx.doi.org/10.1145/1555400.1555452. (Cited on page 120, 125.)

S. Kumpulainen and H. C. Huurdeman. Shaken, not Steered: the Value of Shaking Up the Search Process. In *Proceedings of the First International Workshop on Supporting Complex Search Tasks co-located with ECIR*, 2015. CEUR-WS. http://ceur-ws.org/Vol-1338/paper_7.pdf. (Cited on page 129.)

J. H. Lee, A. Renear, and L. C. Smith. Known-Item Search: Variations on a Concept. *P*

Am Soc Inform Sci, 43(1):1–17, Jan. 2006. http://dx.doi.org/10.1002/meet.14504301126. (Cited on page 7.)

N. Li, A. A. Anderson, D. Brossard, and D. A. Scheufele. Channeling Science Information Seekers' Attention? A Content Analysis of Top-Ranked vs. Lower-Ranked Sites in Google. *J Comput-Mediat Comm*, 19(3):562–575, Apr. 2014. http://dx.doi.org/10.1111/jcc4.12043. (Cited on page 50, 52.)

J. Liu and N. J. Belkin. Personalizing information retrieval for multi-session tasks. *J Am Soc Inf Sci Tec*, 66(1):58–81, Jan. 2015. http://dx.doi.org/10.1002/asi.23160. (Cited on page 139, 141.)

J. Liu, M. J. Cole, C. Liu, R. Bierig, J. Gwizdka, N. J. Belkin, J. Zhang, and X. Zhang. Search behaviors in different task types. In *Proceedings of the 10th Annual Joint Conference on Digital Libraries*, JCDL '10, pages 69–78. ACM, 2010. http://dx.doi.org/10.1145/1816123.1816134. (Cited on page 120, 137.)

S. Lomborg, editor. *Network analysis: methodological challenges.* The Centre for Internet Research, Aarhus, 2012. http://cfi.au.dk/fileadmin/www.cfi.au.dk/publikationer/cfis_skriftserie/014_lomborg.pdf. (Cited on page 52.)

M. Mahrt and C. Puschmann. Science blogging: an exploratory study of motives, styles, and audience reactions. *JCOM: Journal of Science Communication*, 13(3):1–17, July 2014. https://jcom.sissa.it/archive/13/03/JCOM_1303_2014_A05. (Cited on page 51, 52, 53.)

G. Marchionini. Exploratory search: from finding to understanding. *Commun. ACM*, 49(4): 41–46, 2006. http://dx.doi.org/10.1145/1121949.1121979. (Cited on page 115, 137.)

G. Marchionini. *Information Seeking in Electronic Environments*. Cambridge University Press, 1995. http://dx.doi.org/10.1017/CBO9780511626388. (Cited on page 138.)

J. Martinez-Romo and L. Araujo. Analyzing information retrieval methods to recover broken web links. In *Advances in Information Retrieval: 32nd European Conference on IR Research, ECIR '10*, volume 5993 of *LNCS*, pages 26–37. Springer, 2010. http://dx.doi.org/10.1007/978-3-642-12275-0_6. (Cited on page 65.)

J. Masanès. *Web Archiving*, chapter Web Archiving: Issues and Methods, pages 1–53. Springer, 2006. (Cited on page 21, 22, 25, 26, 64.)

L. Melgar, M. Koolen, H. C. Huurdeman, and J. Blom. A process model of scholarly media annotation. In *Proceedings of the 2017 Conference on Human Information Interaction and Retrieval*, CHIIR '17, pages 305–308. ACM, 2017. http://dx.doi.org/10.1145/3020165.3022139. (Cited on page 19.)

D. Metzler, J. Novak, H. Cui, and S. Reddy. Building enriched document representations using aggregated anchor text. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pages 219–226. ACM, 2009. http://dx.doi.org/10.1145/1571941.1571981. (Cited on page 65.)

J. Mi and C. Weng. Revitalizing the library OPAC: Interface, searching, and display challenges. *Inform Technol Libr*, 27:5–22, 2013. http://dx.doi.org/10.6017/ital.v27i1.3259. (Cited on page 114.)

I. Milligan. Lost in the Infinite Archive: The Promise and Pitfalls of Web Archives. *Intl J Humanities & Arts Computing*, 10(1):78–94, Mar. 2016. http://dx.doi.org/10.3366/ijhac.2016.0161. (Cited on page 17.)

N. Moraveji, D. Russell, J. Bien, and D. Mease. Measuring Improvement in User Search Performance Resulting from Optimal Search Tips. In *Proceedings SIGIR*, SIGIR '11, pages 355–364, 2011. ACM. http://dx.doi.org/10.1145/2009916.2009966. (Cited on page 129, 134.)

F. Moretti. *Distant reading*. Verso Books, 2013. (Cited on page 40.)

D. Morris, M. Ringel Morris, and G. Venolia. Searchbar: A search-centric web history for task resumption and information re-finding. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 1207–1216, 2008. ACM. http://dx.doi.org/10.1145/1357054.1357242. (Cited on page 116.)

M. R. Morris and E. Horvitz. Searchtogether: An interface for collaborative web search. In *Proceedings of the 20th Annual ACM Symposium on User Interface Software and Technology*, pages 3–12, 2007. ACM. http://dx.doi.org/10.1145/1294211.1294215. (Cited on page 116.)

X. Niu and D. Kelly. The use of query suggestions during information search. *Inform Process Manag*, 50:218–234, 2014. http://dx.doi.org/10.1016/j.ipm.2013.09.002. (Cited on page 120, 122, 136, 138, 140.)

R. N. Oddy. Information retrieval through man-machine dialogue. *J Doc*, 33(1):1–14, Jan. 1977. http://dx.doi.org/10.1108/eb026631. (Cited on page 101, 114, 174.)

P. Ogilvie and J. Callan. Combining document representations for known-item search. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, SIGIR '03, pages 143–150, 2003. ACM. http://dx.doi.org/10.1145/860435.860463. (Cited on page 66.)

I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and C. Lioma. Terrier: A high performance and scalable information retrieval platform. In *Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006)*, 2006. http://terrierteam.dcs.gla.ac.uk/publications/ounis06terrier-osir.pdf. (Cited on page 33, 88.)

L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford University, 1999. http://ilpubs.stanford.edu:8090/422/. (Cited on page 65.)

S. A. Paul and M. R. Morris. Cosense: Enhancing sensemaking for collaborative web search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pages 1771–1780, 2009. ACM. http://dx.doi.org/10.1145/1518701.1518974. (Cited on page 117.)

N. Payne and M. Thelwall. Do academic link types change over time? *J Doc*, 64(5):707–720, Oct. 2008. http://dx.doi.org/10.1108/00220410810899727. (Cited on page 50.)

N. Pharo and R. Nordlie. Examining the effect of task stage and topic knowledge on searcher interaction with a "digital bookstore". In *Proceedings of the 4th Information Interaction in Context Symposium*, IIIX '12, pages 4–11, 2012. ACM. http://dx.doi.org/10.1145/2362724.2362730. (Cited on page 110.)

A. Pickard. *Research methods in information*. Facet Publishing, 2007. (Cited on page 7, 31, 40.)

P. Pirolli. Powers of 10: Modeling complex information-seeking systems at multiple scales. *IEEE Computer*, 42:33–40, 2009. http://dx.doi.org/10.1109/MC.2009.94. (Cited on page 117.)

P. Pirolli and S. Card. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of International Conference on Intelligence Analysis*, volume 5, pages 2–4, 2005. `http://web.archive.org/web/20110811113026/https://analysis.mitre.org/proceedings/Final_Papers_Files/206_Camera_Ready_Paper.pdf`. (Cited on page 117.)

P. Pirolli and D. M. Russell. Introduction to this special issue on sensemaking. *Hum-Comput Interact*, 26:1–8, 2011. `http://dx.doi.org/10.1080/07370024.2011.556557`. (Cited on page 117.)

Y. Qu and G. W. Furnas. Model-driven formative evaluation of exploratory search: A study under a sensemaking framework. *Inform Process Manag*, 44(2):534 – 555, 2008. `http://dx.doi.org/10.1016/j.ipm.2007.09.006`. (Cited on page 117.)

M. Ras. Eerste fase webarchivering. Technical report, Koninklijke Bibliotheek, 2007. `http://web.archive.org/web/20120511124528/http://www.kb.nl:80/hrd/dd/dd_projecten/webarchivering/documenten/Beschrijving_webarchivering.pdf`. (Cited on page 67.)

A. Rauber, A. Aschenbrenner, O. Witvoet, R. M. Bruckner, and M. Kaiser. Uncovering information hidden in web archives. *D-Lib*, 8(12):1082–9873, 2002. `http://dx.doi.org/10.1045/december2002-rauber`. (Cited on page 45, 64.)

K. Rodden, X. Fu, A. Aula, and I. Spiro. Eye-mouse coordination patterns on web search results pages. In *CHI '08 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '08, pages 2997–3002, 2008. ACM. `http://dx.doi.org/10.1145/1358628.1358797`. (Cited on page 152.)

R. Rogers. *Digital methods*. MIT press, 2013. (Cited on page 1, 3, 23, 26, 27, 28, 30, 33.)

B. Rosenshine and C. Meister. The use of scaffolds for teaching higher-level cognitive strategies. *Educational Leadership*, 49(7):26, Apr. 1992. `http://www.ascd.org/ASCD/pdf/journals/ed_lead/el_199204_rosenshine.pdf`. (Cited on page 129.)

T. Russell-Rose and T. Tate. *Designing the search experience: The information architecture of discovery*. Newnes, 2012. (Cited on page 136, 138.)

I. Ruthven and D. Kelly, editors. *Interactive Information Seeking, Behaviour and Retrieval*. Facet, 2011. (Cited on page 192, 194.)

P. Salisbury and M. R. Griffis. Academic Library Mission Statements, Web Sites, and Communicating Purpose. *J Acad Libr*, 40(6):592–596, Nov. 2014. `http://dx.doi.org/10.1016/j.acalib.2014.07.012`. (Cited on page 52, 54.)

T. Samar, H. C. Huurdeman, A. Ben-David, J. Kamps, and A. de Vries. Uncovering the unarchived web. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, pages 1199–1202, 2014. ACM. `http://dx.doi.org/10.1145/2600428.2609544`. (Cited on page 61, 66.)

T. Samar, M. C. Traub, J. van Ossenbruggen, and A. P. de Vries. Comparing topic coverage in breadth-first & depth-first crawls using anchor texts. In *Research and Advanced Technology for Digital Libraries: 20th International Conference on Theory and Practice of Digital Libraries, Proceedings*, TPDL '16, pages 133–146. Springer, 2016. `http://dx.doi.org/10.1007/978-3-319-43997-6_11`. (Cited on page 97.)

T. Saracevic. Relevance reconsidered. In *Proceedings of the second conference on conceptions of* *library and information science (CoLIS 2)*, pages 201–218. ACM Press, 1996. http://tefkos.comminfo.rutgers.edu/CoLIS2_1996.doc. (Cited on page 9, 110, 128.)

T. Saracevic. The stratified model of information retrieval interaction: Extension and applications. In *P ASIS Annu Meet*, volume 34, pages 313–327. Learned Information (Europe) Ltd, 1997. http://tefkos.comminfo.rutgers.edu/ProcASIS1997.doc. (Cited on page 138.)

R. Savolainen. Epistemic work and knowing in practice as conceptualizations of information use. *Inf Res*, 14(1):paper 392, 2009. http://www.informationr.net/ir/14-1/paper392.html. (Cited on page 8.)

S. M. Schneider and K. A. Foot. The web as an object of study. *New Media Soc*, 6(1):114–122, 2004. http://dx.doi.org/10.1177/1461444804039912. (Cited on page 53.)

m. c. schraefel, Y. Zhu, D. Modjeska, D. Wigdor, and S. Zhao. Hunter gatherer: Interaction support for the creation and management of within-web-page collections. In *Proceedings of the 11th International Conference on World Wide Web*, WWW '02, pages 172–181, 2002. ACM. http://dx.doi.org/10.1145/511446.511469. (Cited on page 116.)

A. Shapiro and D. Niederhauser. Learning from hypertext: Research issues and findings. *Handbook of research on educational communications and technology*, 2:605–620, 2004. (Cited on page 129.)

A. M. Shapiro. Promoting Active Learning: The Role of System Structure in Learning From Hypertext. *Hum-Comput Interact*, 13(1):1–35, Mar. 1998. http://dx.doi.org/10.1207/s15327051hci1301_1. (Cited on page 129.)

B. Shneiderman and C. Pleasant. *Designing the user interface: strategies for effective human-computer interaction.* Pearson Education, 2005. (Cited on page 113, 119, 131, 138.)

S. L. Smith and J. N. Mosier. Guidelines for designing user interface software. Technical report, MITRE, 1986. http://www.hcibib.org/sam/. (Cited on page 113.)

A. Spink. Study of interactive feedback during mediated information retrieval. *J Am Soc Inform Sci*, 48(5):382–394, 1997. http://dx.doi.org/10.1002/(SICI)1097-4571(199705)48:5<382::AID-ASI2>3.0.CO;2-R. (Cited on page 138.)

A. Spink, H. Greisdorf, and J. Bateman. From highly relevant to not relevant: examining different regions of relevance. *Inform Process Manag*, 34:599–621, 1998. http://dx.doi.org/10.1016/S0306-4573(98)00025-9. (Cited on page 110.)

S. Stone. CRUS humanities research programme. In *Humanities information research: proceedings of a seminar.* Centre for Research on User Studies, University of Sheffield, Sheffield, 1980. (Cited on page 49.)

S. Stone. Humanities scholars: information needs and uses. *J Doc*, 38(4):292–313, Apr. 1982. http://dx.doi.org/10.1108/eb026734. (Cited on page 49.)

E. T. Stringer. *Action research.* Sage Publications, 1996. (Cited on page 31.)

B. K. Stripling and J. Pitts. *Brainstorms and Blueprints - Teaching Library Research as a Thinking Process.* Libraries Unlimited, Englewood, Dec. 1988. (Cited on page 129.)

J. Sweller, J. J. G. v. Merrienboer, and F. G. W. C. Paas. Cognitive Architecture and Instructional Design. *Educational Psychology Review*, 10(3):251–296, Sept. 1998. http://dx.doi.org/10.1023/A:1022193728205. (Cited on page 114, 134.)

A. G. Taylor and D. N. Joudrey. *The organization of information.* Libraries Unlimited Englewood, 2009. (Cited on page 8, 9.)

A. R. Taylor, C. Cool, N. J. Belkin, and W. J. Amadio. Relationships between categories of relevance criteria and stage in task completion. *Inform Process Manag*, 43:1071–1084, 2007. http://dx.doi.org/10.1016/j.ipm.2006.09.008. (Cited on page 111.)

A. Thomas, E. Meyer, M. Dougherty, C. Van den Heuvel, C. Madsen, and S. Wyatt. Researcher engagement with web archives: Challenges and opportunities for investment. Technical report, JISC, London, England, 2010. https://ssrn.com/abstract=1715000. (Cited on page 23.)

R. J. Todd. 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. (Cited on page 146.)

E. G. Toms. Task-based information searching and retrieval. In Ruthven and Kelly (2011). (Cited on page 107, 137.)

C. B. Trace and U. P. Karadkar. Information management in the humanities: Scholarly processes, tools, and the construction of personal collections. *J Am Soc Inf Sci Tec*, 68(2): 491–507, 2017. http://dx.doi.org/10.1002/asi.23678. (Cited on page 49.)

T. Tran and N. Fuhr. Quantitative analysis of search sessions enhanced by gaze tracking with dynamic areas of interest. In *Theory and Practice of Digital Libraries*, volume 7489 of *LNCS*, pages 468–473. Springer Berlin Heidelberg, 2012a. http://dx.doi.org/10.1007/978-3-642-33290-6_53. (Cited on page 122.)

V. T. Tran and N. Fuhr. Using eye-tracking with dynamic areas of interest for analyzing interactive information retrieval. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 1165–1166, 2012b. ACM. http://dx.doi.org/10.1145/2348283.2348521. (Cited on page 121, 122.)

D. Tunkelang. Faceted search. *Synthesis lectures on information concepts, retrieval, and services*, 1(1):1–80, 2009. http://dx.doi.org/10.2200/S00190ED1V01Y200904ICR005. (Cited on page 136, 138.)

P. Vakkari. Cognition and changes of search terms and tactics during task performance: A longitudinal case study. In *Content-Based Multimedia Information Access - Volume 1*, RIAO '00, pages 894–907, 2000a. Centre de Hautes Etudes Internationales D'Informatique Documentaire. http://dl.acm.org/citation.cfm?id=2835955. (Cited on page 109.)

P. Vakkari. Relevance and contributing information types of searched documents in task performance. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '00, pages 2–9, 2000b. ACM. http://dx.doi.org/10.1145/345508.345512. (Cited on page 109, 110, 112.)

P. Vakkari. Task-based information searching. *Annu Rev Inform Sci*, 37(1):413–464, 2003. http://dx.doi.org/10.1002/aris.1440370110. (Cited on page 107.)

P. Vakkari. A theory of the task-based information retrieval process: a summary and generalisation of a longitudinal study. *J Doc*, 57(1):44–60, Feb. 2001. http://dx.doi.org/10.1108/EUM0000000007075. (Cited on page 2, 4, 107, 109, 110, 111, 127, 136, 137, 138, 140, 150, 151, 161, 204.)

P. Vakkari and N. Hakala. Changes in relevance criteria and problem stages in task performance. *J Doc*, 56:540–562, 2000. http://dx.doi.org/10.1108/EUM0000000007127. (Cited on page 109, 110, 111, 161, 162.)

P. Vakkari and S. Huuskonen. Search effort degrades search output but improves task outcome. *J Am Soc Inf Sci Tec*, 63(4):657–670, Apr. 2012. http://dx.doi.org/10.1002/asi.21683. (Cited on page 129.)

P. Vakkari, M. Pennanen, and S. Serola. Changes of search terms and tactics while writing a research proposal: A longitudinal case study. *Inform Process Manag*, 39:445–463, 2003. http://dx.doi.org/10.1016/S0306-4573(02)00031-6. (Cited on page 109, 110, 111.)

H. Van de Sompel, M. Nelson, and R. Sanderson. RFC 7089 - HTTP framework for time-based access to resource states - Memento. RFC, Internet Engineering Task Force (IETF), 2013. http://tools.ietf.org/html/rfc7089. (Cited on page 64.)

A. van den Bosch, T. Bogers, and M. d. Kunder. Estimating search engine index size variability: a 9-year longitudinal study. *Scientometrics*, 107(2):839–856, Feb. 2016. http://dx.doi.org/10.1007/s11192-016-1863-z. (Cited on page 1.)

A. Vara-Miguel, E. S. San Martín, and C. Díaz-Espina. Paid news vs free news: evolution of the WSJ.com business model from a content perspective (2010-2012). *Comunicación y Sociedad*, 27(2):147–167, June 2014. http://hdl.handle.net/10171/36276. (Cited on page 51.)

F. Viegas, M. Wattenberg, F. van Ham, J. Kriss, and M. McKeon. Manyeyes: a site for visualization at internet scale. *IEEE T Vis Comput Gr*, 13(6):1121–1128, Nov 2007. http://dx.doi.org/10.1109/TVCG.2007.70577. (Cited on page 116.)

K. Waite and T. Harrison. Internet archaeology: uncovering pension sector web site evolution. *Internet Research*, 17(2):180–195, May 2007. http://dx.doi.org/10.1108/10662240710737031. (Cited on page 50, 51.)

G. Walton and M. Hepworth. A longitudinal study of changes in learners' cognitive states during and following an information literacy teaching intervention. *J Doc*, 67(3):449–479, Apr. 2011. http://dx.doi.org/10.1108/00220411111124541. (Cited on page 134.)

R. W. White and S. M. Drucker. Investigating Behavioral Variability in Web Search. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 21–30, 2007. ACM. http://dx.doi.org/10.1145/1242572.1242576. (Cited on page 139.)

R. W. White and R. A. Roth. Exploratory search: Beyond the query-response paradigm. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 1(1):1–98, 2009. http://dx.doi.org/10.2200/S00174ED1V01Y200901ICR003. (Cited on page 2, 104, 115, 116, 117, 139, 206.)

R. W. White, I. Ruthven, and J. M. Jose. A study of factors affecting the utility of implicit relevance feedback. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, pages 35–42, 2005. ACM. http://dx.doi.org/10.1145/1076034.1076044. (Cited on page 120, 140.)

B. M. Wildemuth and L. Freund. Search tasks and their role in studies of search behaviors. In *Third Annual Workshop on Human Computer Interaction and Information Retrieval, Washington DC*, 2009. http://ils.unc.edu/searchtasks/publication/publication_1.pdf. (Cited on page 107, 137.)

B. M. Wildemuth, L. Freund, and E. G. Toms. Untangling search task complexity and difficulty in the context of interactive information retrieval studies. *J Doc*, 70(6):1118–1140, 2014. http://dx.doi.org/10.1108/JD-03-2014-0056. (Cited on page 137.)

M. L. Wilson. Keyword search: Quite exploratory actually. In *Proceedings of the 3rd International Workshop on Human-Computer Interaction and Information Retrieval*, HCIR'09, pages 106–108, 2009. `http://eprints.soton.ac.uk/id/eprint/267951`. (Cited on page 147.)

M. L. Wilson. Interfaces for information retrieval. In Ruthven and Kelly (2011). (Cited on page 113, 114, 116, 123, 131, 206.)

M. L. Wilson. Search User Interface Design. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 3(3):1–143, Nov. 2011b. `http://dx.doi.org/10.2200/S00371ED1V01Y201111ICR020`. (Cited on page 114, 136, 138, 139, 151.)

M. L. Wilson and m. c. schraefel. A longitudinal study of exploratory and keyword search. In *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '08, pages 52–56, 2008. ACM. `http://dx.doi.org/10.1145/1378889.1378899`. (Cited on page 139.)

M. L. Wilson, B. Kules, m. c. schraefel, and B. Shneiderman. From keyword search to exploration: Designing future search interfaces for the web. *Foundations and Trends in Web Science*, 2(1):1–97, 2010. `http://dx.doi.org/10.1561/1800000003`. (Cited on page 104, 113.)

T. D. Wilson. Models in information behaviour research. *J Doc*, 55:249–270, 1999. `http://dx.doi.org/10.1108/EUM0000000007145`. (Cited on page 8, 106, 112, 126, 131, 137, 138, 204.)

T. D. Wilson, N. J. Ford, D. Ellis, A. E. Foster, and A. Spink. Information seeking and mediated searching: Part 2. uncertainty and its correlates. *J Am Soc Inf Sci Tec*, 53:704–715, 2002. `http://dx.doi.org/10.1002/asi.10082`. (Cited on page 109.)

W. Wright, D. Schroh, P. Proulx, A. Skaburskis, and B. Cort. The sandbox for analysis: Concepts and methods. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '06, pages 801–810, 2006. ACM. `http://dx.doi.org/10.1145/1124772.1124890`. (Cited on page 117.)

M. Xenos and W. L. Bennett. The Disconnection In Online Politics: the youth political web sphere and US election sites, 2002-2004. *Information, Communication & Society*, 10(4):443–464, Aug. 2007. `http://dx.doi.org/10.1080/13691180701559897`. (Cited on page 50, 52, 53, 54.)

K.-P. Yee, K. Swearingen, K. Li, and M. Hearst. Faceted metadata for image search and browsing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '03, pages 401–408, 2003. ACM. `http://dx.doi.org/10.1145/642611.642681`. (Cited on page 116.)

M. Zhitomirsky-Geffet and Y. Maman. "Wisdom of the crowds" and online information reliability: A case study of Israeli real estate websites. *Online Inform Rev*, 38(3):417–435, 2014. `http://dx.doi.org/10.1108/OIR-07-2013-0176`. (Cited on page 53.)

A. Zoubarev, K. M. Hamer, K. D. Keshav, E. L. McCarthy, J. R. C. Santos, T. V. Rossum, C. McDonald, A. Hall, X. Wan, R. Lim, J. Gillis, and P. Pavlidis. Gemma: a resource for the reuse, sharing and meta-analysis of expression profiling data. *Bioinformatics*, 28(17):2272–2273, Sept. 2012. `http://dx.doi.org/10.1093/bioinformatics/bts430`. (Cited on page 57.)

Appendix A: Reviewed Papers and Research Phases

Authors	*Corpus definition*	*Analysis*	*Dissemination*
Bar-Ilan, J., & Peritz, B. C. (2008)	Query ("informetrics")	content analysis (automatic)	tables, graphs
Chua, A. Y. K., & Banerjee, S. (2013)	Selection (Community Question Answering sites)	content analysis (manual)	tables, graphs
Comeaux, D., & Schmetzke, A. (2013)	Selection (list of library websites)	content analysis (automatic)	tables
García-Lacalle, J., Pina, V., Royo, S. (2011)	Selection (list of hospital websites)	content analysis (manual)	tables
Ghobadi, S., & Clegg, S. (2015)	Query ("Iran June 2009") → Selection (highest rated videos)	content analysis (manual), interviews	tables, model
Hoy, M. G., & Park, J. S. (2014)	Selection (Food and Drug Administrative letters) → Selection (addressing internet features)	content analysis (manual)	tables
John, N. A. (2013)	Category (largest social network sites) → Sample (date closest to start of month)	grounded theory	-
Karlsson, M., & Clerwall, C. (2012)	Selection (largest newspapers)	content analysis (manual), interviews	tables, graphs
Karlsson, M., & Clerwall, C. (2015)	Selection (largest newspapers) → Selection (news items) → Sample (daily)	content analysis (manual), interviews	tables, graphs
Koc-Michalska, K., Gibson, R., & Vedel, T. (2014)	Selection (presidential websites)	content analysis (manual), surveys	tables
Li, N., Anderson, A. A., Brossard, D., & Scheufele, D. A. (2014)	Query (related to nanotechnology) → Sample (one week per month)	content analysis (automatic)	tables, graphs
Mahrt, M., & Puschmann, C. (2014)	Selection (blog inspections)	content analysis (manual)	tables
Payne, N., & Thelwall, M. (2008)	Selection (academic webpages) → Sample (random links)	network analysis	tables
Salisbury, P., & Griffis, M. R. (2014)	Selection (list of libraries)	content analysis (manual)	graphs
Vara-Miguel, A., San Martín, E. S., & Díaz-Espina, C. (2014)	Selection (news website) → Selection (news items)	content analysis (manual)	tables, graphs
Waite, K., & Harrison, T. (2007)	Selection (list from annual report) → Sample (number of items from list)	content analysis (manual)	tables, graphs
Xenos, M., & Bennett, W. L. (2007)	Query (related to youth and electoral web spheres)	content analysis (manual), network analysis	tables, graphs, networks
Zhitomirsky-Geffet, M., & Maman, Y. (2014)	Selection (most popular real estate sites) → Selection (10 cities)	content analysis (automatic)	tables, graphs

Figure 6.5: Detailed information about reviewed papers in Section 2.4, per research phase.

Abstract

The World Wide Web, just 25 years after its inception, has become ingrained in nearly every aspect of our daily lives. The traces of our increasingly 'digital' lives and society captured in the web are a highly valuable resource for current and future researchers. However, the content on the web is at risk: at every tick of the clock, websites emerge, change and vanish. The highly ephemeral nature of the web has led to numerous initiatives to perform web archiving. Despite their potential value, web archives have been scarcely used for research thus far. This lack of scholarly use of web archives was this thesis' starting point.

In the first part of this work, we look specifically at search access to web archives in a media and communication research context. Web archives increasingly provide search functionality to access their contents, extending the range of research options for scholars. However, despite the clear value of search-based access, we find that this type of access also obscures a myriad of underlying variables, resulting in a lack of transparency. This leads us to experiment with methods to reconstruct and reveal what is *not* contained in the archive. Furthermore, we uncover a lack of process support: search systems for web archives do not yet provide appropriate support for activities in different phases of a scholar's research process.

The dynamic nature of information seeking in current system-mediated research processes leads us to further investigate the complex information seeking process in part two of this thesis. Inspired by models which document an intricate set of stages, documenting evolving feelings, thoughts and actions of a searcher, we reveal a lack of understanding in two directions. On the one hand, the implications of macro-level stages for micro-level search system design are fuzzy, and on the other hand, the exact support various existing micro-level search system features offer for macro-level stages is unclear. Our studies observe evidence for macro-level stages by studying the flow of user interaction with search systems. This may involve both active interaction (queries and clicks) and passive interaction (eye and mouse pointer movements). Using these measures, we distinguish stage-insensitive and stage-sensitive search user interface features, which could be adapted to a user's information seeking stage. This way, more dynamic support for the information search process may be achieved.

Samenvatting

In de afgelopen 25 jaar is het *World Wide Web* uitgegroeid tot een essentieel onderdeel van ons dagelijks leven. Voor toekomstige onderzoekers is de inhoud van het web van groot belang om onze huidige samenleving te begrijpen. Echter, de *content* op het web is verre van stabiel: webpagina's kunnen op elk moment verschijnen, veranderen en verdwijnen. Het vluchtige karakter van het web heeft wereldwijd geleid tot een groot aantal initiatieven voor webarchivering, die ondertussen vele Petabytes aan informatie hebben vastgelegd. Ondanks deze enorme hoeveelheid aan potentieel onderzoeksmateriaal gebruiken wetenschappers webarchieven nog maar zelden als databron. Dit gebrek aan wetenschappelijk gebruik van het webarchief was het startpunt van dit proefschrift.

Het eerste deel van de dissertatie focust zich op toegang tot webarchieven in een onderzoekscontext. Uit een literatuurstudie blijkt dat webarchieven beperkingen kennen in termen van de kwaliteit en kwantiteit van de data, en de toegang tot de data. De meest algemene manier om gearchiveerde webpagina's te bekijken is via het invoeren van een specifieke URL van een pagina uit het verleden. Daarnaast bieden initiatiefnemers van webarchieven in toenemende mate zoekmogelijkheden aan via trefwoorden, gelijkend op de functionaliteit van online zoekmachines.

Het ontwerp van onderzoeksfunctionaliteit voor webarchieven is echter niet eenvoudig. In het proefschrift analyseren we zoektoegang tot webarchieven en bekijken we hoe deze manier van toegang verbeterd kan worden. Hiertoe werd een zoeksysteem voor het webarchief van de Koninklijke Bibliotheek ontworpen en geëvalueerd via een action research methodologie, in samenwerking met onderzoekers in nieuwe media. We laten enerzijds zien dat deze zoekfunctionaliteit een grote toegevoegde waarde heeft voor onderzoeksdoeleinden. Anderzijds verbergt deze manier van toegang ook een groot aantal onderliggende variabelen. Deze variabelen omvatten de originele selectie van sites, de wijze van archivering, de indexerings- en retrievalinstellingen en de invloed van de grafische zoekomgeving. Toekomstige zoekfunctionaliteit voor webarchieven zal nieuwe manieren moeten vinden om deze complexiteit meer inzichtelijk en *transparant* te maken voor de wetenschappelijke gebruikers van een webarchief.

Een gestructureerde literatuuranalyse laat zien dat zoeksystemen voor webarchieven weinig ondersteuning bieden voor gedocumenteerde activiteiten in verschillende fasen van het onderzoeksproces, waaronder corpus creatie, analyse en disseminatie. De dominante focus op zoekqueries maakt het moeilijk om een dataset voor onderzoek (corpus) samen te stellen op basis van specifieke selecties van websites, of via willekeurige steekproeven. In de meeste gevallen ontbreken analysemogelijkheden in de zoekfunctionaliteit, net als opties om gevonden inhoud van webarchieven te visualiseren. Er is dus sprake van een gebrek aan *procesondersteuning* voor onderzoeksactiviteiten.

De huidige wijze van webarchivering, via zogenoemde 'web crawlers', heeft een grote invloed op de inhoud van het uiteindelijke webarchief. In veel gevallen worden websites, webpagina's of delen van webpagina's niet correct binnengehaald, leidend tot een inherente incompleetheid van webarchieven. Op dit moment is het vrijwel onmogelijk om te zien in hoeverre een webarchief een compleet perspectief op een onderwerp biedt. Via kwantitatieve experimenten met de linkstructuur van het webarchief van de Koninklijke Bibliotheek tonen we aan dat er methoden zijn om zichtbaar te maken wat er *niet* in het archief zit, en niet-gearchiveerde content zelfs deels te reconstrueren. Deze methoden kunnen bijdragen aan het ontwerp van toekomstige zoekmachines voor webarchieven die onderzoekers meer transparantie bieden.

Het eerste deel van het proefschrift liet beperkingen zien in de procesondersteuning door zoekmachines. Om deze beperkingen te kunnen verhelpen is een beter begrip nodig van het zoekproces gedurende complexe taken en de rol die de zoekfunctionaliteit daarin speelt. Dit is de focus van het tweede deel van het proefschrift.

Jaren van razendsnelle ontwikkeling op technisch gebied hebben geleid tot een zeer effectieve ondersteuning voor 'look-up' zoektaken in online zoekmachines. Dit zijn simpele taken waarvoor een duidelijk antwoord beschikbaar is, zoals het opvragen van de openingstijden van een winkel. Meer complexe taken, waarbij onderzoek en het ontwikkelen van kennis een rol speelt, worden in mindere mate ondersteund. Voor dit soort taken beschrijven verschillende 'information seeking' modellen, waaronder Kuhlthau's ISP model, een specifieke set van cognitieve stadia. In de verschillende stadia, die evolueren van exploratief naar gefocust, vinden veranderingen plaats in de gevoelens, gedachten en acties van de gebruikers van een informatiesysteem. Kuhlthau's model heeft een belangrijke invloed gehad op bibliotheek- en informatiewetenschap, maar minder invloed op concrete zoektools: zoekmachines bieden geen expliciete ondersteuning voor de verschillende stadia van het zoekproces.

Een literatuurstudie in het tweede deel van dit proefschrift laat zien dat één van de onderliggende oorzaken een gebrek aan wederzijds inzicht is. Enerzijds beschrijven information seeking modellen het zoeken naar informatie op een conceptueel macro-niveau. Hierdoor is het lastig om de implicaties van zoekstadia voor het ontwerp van concrete zoekfunctionaliteit op micro-niveau te begrijpen. Anderzijds is het nut van specifieke zoekfunctionaliteit in de algehele 'macro' context van het algehele zoekproces onduidelijk. Een eerste aanzet tot een beter wederzijds inzicht bestond uit een kwantitatieve analyse van data uit een eerdere gebruikersstudie. Hierin waren aanwijzingen zichtbaar dat sommige onderdelen van een zoeksysteem zoekstadium-afhankelijk zijn en vooral gebruikt worden in specifieke fases van een zoeksessie. Andere functionaliteit is zoekstadium-onafhankelijk en kan waardevol zijn op verschillende momenten van een zoeksessie.

In een uitgebreidere gebruikersstudie voerden studenten drie taken uit, gemodelleerd naar verschillende zoekstadia uit de literatuur ('pre-focus', 'focus formulation' en 'post-focus'). Hierbij werd er in de evaluatie een onderscheid gemaakt tussen 'actieve' interactie (muiskliks en de invoer van zoekwoorden), 'passieve' interactie (oog- en muisbewegingen) en subjectieve waardering (op basis van vragenlijsten en korte interviews). De resultaten van de studie laten een duidelijk verschil zien in de waarde van zoekfunctionaliteit gedurende verschillende stadia. Zogenoemde 'Informational' features, in dit geval de lijst met zoekresultaten, waren grotendeels zoekstadium-onafhankelijk. Andere functionaliteit bleek zoekstadium-afhankelijk. Dit waren ten eerste de 'Input' en 'Control' features, de functionaliteit waarmee gebruikers zoekvragen invoeren en verfijnen, waaronder het zoekvenster, filters en zoeksuggesties. De waarde van deze functionaliteit neemt af gedurende het proces, deels door de opgedane kennis over een onderwerp. Verder was een tegengestelde tendens zichtbaar voor 'Personalizable' features. Dit is het type functionaliteit dat zich aanpast aan de activiteiten van een gebruiker (in deze studie een lijst met recente zoektermen en bewaarde resultaten). 'Personalizable' features groeien mee met de kennis die een gebruiker gedurende de taak opdoet en worden dus steeds belangrijker. Uit de studie volgt verder dat sommige zoekfunctionaliteit met name actief werd gebruikt, terwijl andere features een meer passieve waarde hadden. Deze nieuwe inzichten kunnen worden gebruikt voor het ontwerpen van betere procesondersteuning in zoekmachines.

Huidige zoekmachines bieden toegang tot een immense hoeveelheid informatie via geavanceerde algoritmes, maar de interactie is beperkt: een korte query leidt tot een korte lijst van tien blauwe links. Dit proefschrift laat zien dat deze toegangswijze een scala aan onderliggende variabelen verbergt, en geen directe verbintenis met het zoekproces van de gebruiker heeft. Voor complexe en informatie-intensieve taken is het daarom nodig om de toegangswijze te herzien, zowel in de context van het webarchief als het 'live' web. Op die manier kunnen toekomstige zoeksystemen een meer dynamische ondersteuning bieden voor de complexe dynamiek van het zoekproces.

List of Figures

List of Tables

Titles in the SIKS Dissertation Series

2011

1 **Botond Cseke** (RUN) *Variational Algorithms for Bayesian Inference in Latent Gaussian Models*

2 **Nick Tinnemeier** (UU) *Organizing Agent Organizations. Syntax and Operational Semantics of an Organization-Oriented Programming Language*

3 **Jan Martijn van der Werf** (TUe) *Compositional Design and Verification of Component-Based Information Systems*

4 **Hado van Hasselt** (UU) *Insights in Reinforcement Learning: Formal analysis and empirical evaluation of temporal-difference*

5 **Base van der Raadt** (VUA) *Enterprise Architecture Coming of Age: Increasing the Performance of an Emerging Discipline*

6 **Yiwen Wang** (TUe) *Semantically-Enhanced Recommendations in Cultural Heritage*

7 **Yujia Cao** (UT) *Multimodal Information Presentation for High Load Human Computer Interaction*

8 **Nieske Vergunst** (UU) *BDI-based Generation of Robust Task-Oriented Dialogues*

9 **Tim de Jong** (OU) *Contextualised Mobile Media for Learning*

10 **Bart Bogaert** (UvT) *Cloud Content Contention*

11 **Dhaval Vyas** (UT) *Designing for Awareness: An Experience-focused HCI Perspective*

12 **Carmen Bratosin** (TUe) *Grid Architecture for Distributed Process Mining*

13 **Xiaoyu Mao** (UvT) *Airport under Control. Multiagent Scheduling for Airport Ground Handling*

14 **Milan Lovric** (EUR) *Behavioral Finance and Agent-Based Artificial Markets*

15 **Marijn Koolen** (UvA) *The Meaning of Structure: the Value of Link Evidence for Information Retrieval*

16 **Maarten Schadd** (UM) *Selective Search in Games of Different Complexity*

17 **Jiyin He** (UvA) *Exploring Topic Structure: Coherence, Diversity and Relatedness*

18 **Mark Ponsen** (UM) *Strategic Decision-Making in complex games*

19 **Ellen Rusman** (OU) *The Mind ' s Eye on Personal Profiles*

20 **Qing Gu** (VUA) *Guiding service-oriented software engineering: A view-based approach*

21 **Linda Terlouw** (TUD) *Modularization and Specification of Service-Oriented Systems*

22 **Junte Zhang** (UvA) *System Evaluation of Archival Description and Access*

23 **Wouter Weerkamp** (UvA) *Finding People and their Utterances in Social Media*

24 **Herwin van Welbergen** (UT) *Behavior Generation for Interpersonal Coordination with Virtual Humans On Specifying, Scheduling and Realizing Multimodal Virtual Human Behavior*

25 **Syed Waqar ul Qounain Jaffry** (VUA) *Analysis and Validation of Models for Trust Dynamics*

26 **Matthijs Aart Pontier** (VUA) *Virtual Agents for Human Communication: Emotion Regulation and Involvement-Distance Trade-Offs in Embodied Conversational Agents and Robots*

27 **Aniel Bhulai** (VUA) *Dynamic website optimization through autonomous management of design patterns*

28 **Rianne Kaptein** (UvA) *Effective Focused Retrieval by Exploiting Query Context and Document Structure*

29 **Faisal Kamiran** (TUe) *Discrimination-aware Classification*

30 **Egon van den Broek** (UT) *Affective Signal Processing (ASP): Unraveling the mystery of emotions*

31 **Ludo Waltman** (EUR) *Computational and Game-Theoretic Approaches for Modeling Bounded Rationality*

32 **Nees-Jan van Eck** (EUR) *Methodological Advances in Bibliometric Mapping of Science*

33 **Tom van der Weide** (UU) *Arguing to Motivate Decisions*

34 **Paolo Turrini** (UU) *Strategic Reasoning in Interdependence: Logical and Game-theoretical Investigations*

35 **Maaike Harbers** (UU) *Explaining Agent Behavior in Virtual Training*

36 **Erik van der Spek** (UU) *Experiments in serious game design: a cognitive approach*

37 **Adriana Burlutiu** (RUN) *Machine Learning for Pairwise Data, Applications for Preference Learning and Supervised Network Inference*

38 **Nyree Lemmens** (UM) *Bee-inspired Distributed Optimization*

39 **Joost Westra** (UU) *Organizing Adaptation using Agents in Serious Games*

40 **Viktor Clerc** (VUA) *Architectural Knowledge Management in Global Software Development*

41 **Luan Ibraimi** (UT) *Cryptographically Enforced Distributed Data Access Control*

42 **Michal Sindlar** (UU) *Explaining Behavior through Mental State Attribution*

43 **Henk van der Schuur** (UU) *Process Improvement through Software Operation Knowledge*

44 **Boris Reuderink** (UT) *Robust Brain-Computer Interfaces*

45 **Herman Stehouwer** (UvT) *Statistical Language Models for Alternative Sequence Selection*

46 **Beibei Hu** (TUD) *Towards Contextualized Information Delivery: A Rule-based Architecture for the Domain of Mobile Police Work*

47 **Azizi Bin Ab Aziz** (VUA) *Exploring Computational Models for Intelligent Support of Persons with Depression*

48 **Mark Ter Maat** (UT) *Response Selection and Turn-taking for a Sensitive Artificial Listening Agent*

49 **Andreea Niculescu** (UT) *Conversational interfaces for task-oriented spoken dialogues: design aspects influencing interaction quality*

2012

1 **Terry Kakeeto** (UvT) *Relationship Marketing for SMEs in Uganda*

2 **Muhammad Umair** (VUA) *Adaptivity, emotion, and Rationality in Human and Ambient Agent Models*

3 **Adam Vanya** (VUA) *Supporting Architecture Evolution by Mining Software Repositories*

4 **Jurriaan Souer** (UU) *Development of Content Management System-based Web Applications*

5 **Marijn Plomp** (UU) *Maturing Interorganisational Information Systems*

6 **Wolfgang Reinhardt** (OU) *Awareness Support for Knowledge Workers in Research Networks*

7 **Rianne van Lambalgen** (VUA) *When the Going Gets Tough: Exploring Agent-based Models of Human Performance under Demanding Conditions*

8 **Gerben de Vries** (UvA) *Kernel Methods for Vessel Trajectories*

9 **Ricardo Neisse** (UT) *Trust and Privacy Management Support for Context-Aware Service Platforms*

10 **David Smits** (TUe) *Towards a Generic Distributed Adaptive Hypermedia Environment*

11 **J. C. B. Rantham Prabhakara** (TUe) *Process Mining in the Large: Preprocessing, Discovery, and Diagnostics*

12 **Kees van der Sluijs** (TUe) *Model Driven Design and Data Integration in Semantic Web Information Systems*

13 **Suleman Shahid** (UvT) *Fun and Face: Exploring non-verbal expressions of emotion during playful interactions*

14 **Evgeny Knutov** (TUe) *Generic Adaptation Framework for Unifying Adaptive Web-based Systems*

15 **Natalie van der Wal** (VUA) *Social Agents. Agent-Based Modelling of Integrated Internal and Social Dynamics of Cognitive and Affective Processes*

16 **Fiemke Both** (VUA) *Helping people by understanding them: Ambient Agents supporting task execution and depression treatment*

17 **Amal Elgammal** (UvT) *Towards a Comprehensive Framework for Business Process Compliance*

18 **Eltjo Poort** (VUA) *Improving Solution Architecting Practices*

19 **Helen Schonenberg** (TUe) *What's Next? Operational Support for Business Process Execution*

20 **Ali Bahramisharif** (RUN) *Covert Visual Spatial Attention, a Robust Paradigm for Brain-Computer Interfacing*

21 **Roberto Cornacchia** (TUD) *Querying Sparse Matrices for Information Retrieval*

22 **Thijs Vis** (UvT) *Intelligence, politie en veiligheidsdienst: verenigbare grootheden?*

23 **Christian Muehl** (UT) *Toward Affective Brain-Computer Interfaces: Exploring the Neurophysiology of Affect during Human Media Interaction*

24 **Laurens van der Werff** (UT) *Evaluation of Noisy Transcripts for Spoken Document Retrieval*

25 **Silja Eckartz** (UT) *Managing the Business Case Development in Inter-Organizational IT Projects: A Methodology and its Application*

26 **Emile de Maat** (UvA) *Making Sense of Legal Text*

27 **Hayrettin Gurkok** (UT) *Mind the Sheep! User Experience Evaluation & Brain-Computer Interface Games*

28 **Nancy Pascall** (UvT) *Engendering Technology Empowering Women*

29 **Almer Tigelaar** (UT) *Peer-to-Peer Information Retrieval*

30 **Alina Pommeranz** (TUD) *Designing Human-Centered Systems for Reflective Decision Making*

31 **Emily Bagarukayo** (RUN) *A Learning by Construction Approach for Higher Order Cognitive Skills Improvement, Building Capacity and Infrastructure*

32 **Wietske Visser** (TUD) *Qualitative multi-criteria preference representation and reasoning*

33 **Rory Sie** (OUN) *Coalitions in Cooperation Networks (COCOON)*

34 **Pavol Jancura** (RUN) *Evolutionary analysis in PPI networks and applications*

35 **Evert Haasdijk** (VUA) *Never Too Old To Learn: On-line Evolution of Controllers in Swarm- and Modular Robotics*

36 **Denis Ssebugwawo** (RUN) *Analysis and Evaluation of Collaborative Modeling Processes*

37 **Agnes Nakakawa** (RUN) *A Collaboration Process for Enterprise Architecture Creation*

38 **Selmar Smit** (VUA) *Parameter Tuning and Scientific Testing in Evolutionary Algorithms*

39 **Hassan Fatemi** (UT) *Risk-aware design of value and coordination networks*

40 **Agus Gunawan** (UvT) *Information Access for SMEs in Indonesia*

41 **Sebastian Kelle** (OU) *Game Design Patterns for Learning*

42 **Dominique Verpoorten** (OU) *Reflection Amplifiers in self-regulated Learning*

43 **Anna Tordai** (VUA) *On Combining Alignment Techniques*

44 **Benedikt Kratz** (UvT) *A Model and Language for Business-aware Transactions*

45 **Simon Carter** (UvA) *Exploration and Exploitation of Multilingual Data for Statistical Machine Translation*

46 **Manos Tsagkias** (UvA) *Mining Social Media: Tracking Content and Predicting Behavior*

47 **Jorn Bakker** (TUe) *Handling Abrupt Changes in Evolving Time-series Data*

48 **Michael Kaisers** (UM) *Learning against Learning: Evolutionary dynamics of reinforcement learning algorithms in strategic interactions*

49 **Steven van Kervel** (TUD) *Ontologogy driven Enterprise Information Systems Engineering*

50 **Jeroen de Jong** (TUD) *Heuristics in Dynamic Sceduling: a practical framework with a case study in elevator dispatching*

2013

1 **Viorel Milea** (EUR) *News Analytics for Financial Decision Support*

2 **Erietta Liarou** (CWI) *MonetDB/DataCell: Leveraging the Column-store Database Technology for Efficient and Scalable Stream Processing*

3 **Szymon Klarman** (VUA) *Reasoning with Contexts in Description Logics*

4 **Chetan Yadati** (TUD) *Coordinating autonomous planning and scheduling*

5 **Dulce Pumareja** (UT) *Groupware Requirements Evolutions Patterns*

6 **Romulo Goncalves** (CWI) *The Data Cyclotron: Juggling Data and Queries for a Data Warehouse Audience*

7 **Giel van Lankveld** (UvT) *Quantifying Individual Player Differences*

8 **Robbert-Jan Merk** (VUA) *Making enemies: cognitive modeling for opponent agents in fighter pilot simulators*

9 **Fabio Gori** (RUN) *Metagenomic Data Analysis: Computational Methods and Applications*

10 **Jeewanie Jayasinghe Arachchige** (UvT) *A Unified Modeling Framework for Service Design*

11 **Evangelos Pournaras** (TUD) *Multi-level Reconfigurable Self-organization in Overlay Services*

12 **Marian Razavian** (VUA) *Knowledge-driven Migration to Services*

13 **Mohammad Safiri** (UT) *Service Tailoring: User-centric creation of integrated IT-based homecare services to support independent living of elderly*

14 **Jafar Tanha** (UvA) *Ensemble Approaches to Semi-Supervised Learning Learning*

15 **Daniel Hennes** (UM) *Multiagent Learning: Dynamic Games and Applications*

16 **Eric Kok** (UU) *Exploring the practical benefits of argumentation in multi-agent deliberation*

17 **Koen Kok** (VUA) *The PowerMatcher: Smart Coordination for the Smart Electricity Grid*

18 **Jeroen Janssens** (UvT) *Outlier Selection and One-Class Classification*

19 **Renze Steenhuizen** (TUD) *Coordinated Multi-Agent Planning and Scheduling*

20 **Katja Hofmann** (UvA) *Fast and Reliable Online Learning to Rank for Information Retrieval*

21 **Sander Wubben** (UvT) *Text-to-text generation by monolingual machine translation*

22 **Tom Claassen** (RUN) *Causal Discovery and Logic*

23 **Patricio de Alencar Silva** (UvT) *Value Activity Monitoring*

24 **Haitham Bou Ammar** (UM) *Automated Transfer in Reinforcement Learning*

25 **Agnieszka Anna Latoszek-Berendsen** (UM) *Intention-based Decision Support. A new way of representing and implementing clinical guidelines in a Decision Support System*

26 **Alireza Zarghami** (UT) *Architectural Support for Dynamic Homecare Service Provisioning*

27 **Mohammad Huq** (UT) *Inference-based Framework Managing Data Provenance*

28 **Frans van der Sluis** (UT) *When Complexity becomes Interesting: An Inquiry into the Information eXperience*

29 **Iwan de Kok** (UT) *Listening Heads*

30 **Joyce Nakatumba** (TUe) *Resource-Aware Business Process Management: Analysis and Support*

31 **Dinh Khoa Nguyen** (UvT) *Blueprint Model and Language for Engineering Cloud Applications*

32 **Kamakshi Rajagopal** (OUN) *Networking For Learning: The role of Networking in a Lifelong Learner's Professional Development*

33 **Qi Gao** (TUD) *User Modeling and Personalization in the Microblogging Sphere*

34 **Kien Tjin-Kam-Jet** (UT) *Distributed Deep Web Search*

35 **Abdallah El Ali** (UvA) *Minimal Mobile Human Computer Interaction*

36 **Than Lam Hoang** (TUe) *Pattern Mining in Data Streams*

37 **Dirk Börner** (OUN) *Ambient Learning Displays*

38 **Eelco den Heijer** (VUA) *Autonomous Evolutionary Art*

39 **Joop de Jong** (TUD) *A Method for Enterprise Ontology based Design of Enterprise Information Systems*

40 **Pim Nijssen** (UM) *Monte-Carlo Tree Search for Multi-Player Games*

41 **Jochem Liem** (UvA) *Supporting the Conceptual Modelling of Dynamic Systems: A Knowledge Engineering Perspective on Qualitative Reasoning*

42 **Léon Planken** (TUD) *Algorithms for Simple Temporal Reasoning*

43 **Marc Bron** (UvA) *Exploration and Contextualization through Interaction and Concepts*

2014

1 **Nicola Barile** (UU) *Studies in Learning Monotone Models from Data*

2 **Fiona Tuliyano** (RUN) *Combining System Dynamics with a Domain Modeling Method*

3 **Sergio Raul Duarte Torres** (UT) *Information Retrieval for Children: Search Behavior and Solutions*

4 **Hanna Jochmann-Mannak** (UT) *Websites for children: search strategies and interface design - Three studies on children's search performance and evaluation*

5 **Jurriaan van Reijsen** (UU) *Knowledge Perspectives on Advancing Dynamic Capability*

6 **Damian Tamburri** (VUA) *Supporting Networked Software Development*

7 **Arya Adriansyah** (TUe) *Aligning Observed and Modeled Behavior*

8 **Samur Araujo** (TUD) *Data Integration over Distributed and Heterogeneous Data Endpoints*

9 **Philip Jackson** (UvT) *Toward Human-Level Artificial Intelligence: Representation and Computation of Meaning in Natural Language*

10 **Ivan Salvador Razo Zapata** (VUA) *Service Value Networks*

11 **Janneke van der Zwaan** (TUD) *An Empathic Virtual Buddy for Social Support*

12 **Willem van Willigen** (VUA) *Look Ma, No Hands: Aspects of Autonomous Vehicle Control*

13 **Arlette van Wissen** (VUA) *Agent-Based Support for Behavior Change: Models and Applications in Health and Safety Domains*

14 **Yangyang Shi** (TUD) *Language Models With Meta-information*

15 **Natalya Mogles** (VUA) *Agent-Based Analysis and Support of Human Functioning in Complex Socio-Technical Systems: Applications in Safety and Healthcare*

16 **Krystyna Milian** (VUA) *Supporting trial recruitment and design by automatically interpreting eligibility criteria*

17 **Kathrin Dentler** (VUA) *Computing healthcare quality indicators automatically: Secondary Use of Patient Data and Semantic Interoperability*

18 **Mattijs Ghijsen** (UvA) *Methods and Models for the Design and Study of Dynamic Agent Organizations*

19 **Vinicius Ramos** (TUe) *Adaptive Hypermedia Courses: Qualitative and Quantitative Evaluation and Tool Support*

20 **Mena Habib** (UT) *Named Entity Extraction and Disambiguation for Informal Text: The Missing Link*

21 **Kassidy Clark** (TUD) *Negotiation and Monitoring in Open Environments*

22 **Marieke Peeters** (UU) *Personalized Educational Games: Developing agent-supported scenario-based training*

23 **Eleftherios Sidirourgos** (UvA/CWI) *Space Efficient Indexes for the Big Data Era*

24 **Davide Ceolin** (VUA) *Trusting Semi-structured Web Data*

25 **Martijn Lappenschaar** (RUN) *New network models for the analysis of disease interaction*

26 **Tim Baarslag** (TUD) *What to Bid and When to Stop*

27 **Rui Jorge Almeida** (EUR) *Conditional Density Models Integrating Fuzzy and Probabilistic Representations of Uncertainty*

28 **Anna Chmielowiec** (VUA) *Decentralized k-Clique Matching*

29 **Jaap Kabbedijk** (UU) *Variability in Multi-Tenant Enterprise Software*

30 **Peter de Cock** (UvT) *Anticipating Criminal Behaviour*

31 **Leo van Moergestel** (UU) *Agent Technology in Agile Multiparallel Manufacturing and Product Support*

32 **Naser Ayat** (UvA) *On Entity Resolution in Probabilistic Data*

33 **Tesfa Tegegne** (RUN) *Service Discovery in eHealth*

34 **Christina Manteli** (VUA) *The Effect of Governance in Global Software Development: Analyzing Transactive Memory Systems*

35 **Joost van Ooijen** (UU) *Cognitive Agents in Virtual Worlds: A Middleware Design Approach*

36 **Joos Buijs** (TUe) *Flexible Evolutionary Algorithms for Mining Structured Process Models*

37 **Maral Dadvar** (UT) *Experts and Machines United Against Cyberbullying*

38 **Danny Plass-Oude Bos** (UT) *Making brain-computer interfaces better: improving usability through post-processing*

39 **Jasmina Maric** (UvT) *Web Communities, Immigration, and Social Capital*

40 **Walter Omona** (RUN) *A Framework for Knowledge Management Using ICT in Higher Education*

41 **Frederic Hogenboom** (EUR) *Automated Detection of Financial Events in News Text*

42 **Carsten Eijckhof** (CWI/TUD) *Contextual Multidimensional Relevance Models*

43 **Kevin Vlaanderen** (UU) *Supporting Process Improvement using Method Increments*

44 **Paulien Meesters** (UvT) *Intelligent Blauw: Intelligence-gestuurde politiezorg in gebiedsgebonden eenheden*

45 **Birgit Schmitz** (OUN) *Mobile Games for Learning: A Pattern-Based Approach*

46 **Ke Tao** (TUD) *Social Web Data Analytics: Relevance, Redundancy, Diversity*

47 **Shangsong Liang** (UvA) *Fusion and Diversification in Information Retrieval*

2015

1 **Niels Netten** (UvA) *Machine Learning for Relevance of Information in Crisis Response*

2 **Faiza Bukhsh** (UvT) *Smart auditing: Innovative Compliance Checking in Customs Controls*

3 **Twan van Laarhoven** (RUN) *Machine learning for network data*

4 **Howard Spoelstra** (OUN) *Collaborations in Open Learning Environments*

5 **Christoph Bösch** (UT) *Cryptographically Enforced Search Pattern Hiding*

6 **Farideh Heidari** (TUD) *Business Process Quality Computation: Computing Non-Functional Requirements to Improve Business Processes*

7 **Maria-Hendrike Peetz** (UvA) *Time-Aware Online Reputation Analysis*

8 **Jie Jiang** (TUD) *Organizational Compliance: An agent-based model for designing and evaluating organizational interactions*

9 **Randy Klaassen** (UT) *HCI Perspectives on Behavior Change Support Systems*

10 **Henry Hermans** (OUN) *OpenU: design of an integrated system to support lifelong learning*

11 **Yongming Luo** (TUe) *Designing algorithms for big graph datasets: A study of computing bisimulation and joins*

12 **Julie M. Birkholz** (VUA) *Modi Operandi of Social Network Dynamics: The Effect of Context on Scientific Collaboration Networks*

13 **Giuseppe Procaccianti** (VUA) *Energy-Efficient Software*

14 **Bart van Straalen** (UT) *A cognitive approach to modeling bad news conversations*

15 **Klaas Andries de Graaf** (VUA) *Ontology-based Software Architecture Documentation*

16 **Changyun Wei** (UT) *Cognitive Coordination for Cooperative Multi-Robot Teamwork*

17 **André van Cleeff** (UT) *Physical and Digital Security Mechanisms: Properties, Combinations and Trade-offs*

18 **Holger Pirk** (CWI) *Waste Not, Want Not!: Managing Relational Data in Asymmetric Memories*

19 **Bernardo Tabuenca** (OUN) *Ubiquitous Technology for Lifelong Learners*

20 **Loïs Vanhée** (UU) *Using Culture and Values to Support Flexible Coordination*

21 **Sibren Fetter** (OUN) *Using Peer-Support to Expand and Stabilize Online Learning*

22 **Zhemin Zhu** (UT) *Co-occurrence Rate Networks*

23 **Luit Gazendam** (VUA) *Cataloguer Support in Cultural Heritage*

24 **Richard Berendsen** (UvA) *Finding People, Papers, and Posts: Vertical Search Algorithms and Evaluation*

25 **Steven Woudenberg** (UU) *Bayesian Tools for Early Disease Detection*

26 **Alexander Hogenboom** (EUR) *Sentiment Analysis of Text Guided by Semantics and Structure*

27 **Sándor Héman** (CWI) *Updating Compressed Column-stores*

28 **Janet Bagorogoza** (TiU) *Knowledge Management and High Performance: The Uganda Financial Institutions Model for HPO*

29 **Hendrik Baier** (UM) *Monte-Carlo Tree Search Enhancements for One-Player and Two-Player Domains*

30 **Kiavash Bahreini** (OU) *Real-time Multimodal Emotion Recognition in E-Learning*

31 **Yakup Koç** (TUD) *On the robustness of Power Grids*

32 **Jerome Gard** (UL) *Corporate Venture Management in SMEs*

33 **Frederik Schadd** (TUD) *Ontology Mapping with Auxiliary Resources*

34 **Victor de Graaf** (UT) *Gesocial Recommender Systems*

35 **Jungxao Xu** (TUD) *Affective Body Language of Humanoid Robots: Perception and Effects in Human Robot Interaction*

2016

1 **Syed Saiden Abbas** (RUN) *Recognition of Shapes by Humans and Machines*

2 **Michiel Christiaan Meulendijk** (UU) *Optimizing medication reviews through decision support: prescribing a better pill to swallow*

3 **Maya Sappelli** (RUN) *Knowledge Work in Context: User Centered Knowledge Worker Support*

4 **Laurens Rietveld** (VUA) *Publishing and Consuming Linked Data*

5 **Evgeny Sherkhonov** (UvA) *Expanded Acyclic Queries: Containment and an Application in Explaining Missing Answers*

6 **Michel Wilson** (TUD) *Robust scheduling in an uncertain environment*

7 **Jeroen de Man** (VUA) *Measuring and modeling negative emotions for virtual training*

8 **Matje van de Camp** (TiU) *A Link to the Past: Constructing Historical Social Networks from Unstructured Data*

9 **Archana Nottamkandath** (VUA) *Trusting Crowdsourced Information on Cultural Artefacts*

10 **George Karafotias** (VUA) *Parameter Control for Evolutionary Algorithms*

11 **Anne Schuth** (UvA) *Search Engines that Learn from Their Users*

12 **Max Knobbout** (UU) *Logics for Modelling and Verifying Normative Multi-Agent Systems*

13 **Nana Baah Gyan** (VUA) *The Web, Speech Technologies and Rural Development in West Africa: An ICT4D Approach*

14 **Ravi Khadka** (UU) *Revisiting Legacy Software System Modernization*

15 **Steffen Michels** (RUN) *Hybrid Probabilistic Logics: Theoretical Aspects, Algorithms and Experiments*

16 **Guangliang Li** (UvA) *Socially Intelligent Autonomous Agents that Learn from Human Reward*

17 **Berend Weel** (VUA) *Towards Embodied Evolution of Robot Organisms*

18 **Albert Meroño Peñuela** (VUA) *Refining Statistical Data on the Web*

19 **Julia Efremova** (Tu/e) *Mining Social Structures from Genealogical Data*

20 **Daan Odijk** (UvA) *Context & Semantics in News & Web Search*

21 **Alejandro Moreno Célleri** (UT) *From Traditional to Interactive Playspaces: Automatic Analysis of Player Behavior in the Interactive Tag Playground*

22 **Grace Lewis** (VUA) *Software Architecture Strategies for Cyber-Foraging Systems*

23 **Fei Cai** (UvA) *Query Auto Completion in Information Retrieval*

24 **Brend Wanders** (UT) *Repurposing and Probabilistic Integration of Data: An Iterative and data model independent approach*

25 **Julia Kiseleva** (TU/e) *Using Contextual Information to Understand Searching and Browsing Behavior*

26 **Dilhan Thilakarathne** (VUA) *In or Out of Control: Exploring Computational Models to Study the Role of Human Awareness and Control in Behavioural Choices, with Applications in Aviation and Energy Management Domains*

27 **Wen Li** (TUD) *Understanding Geospatial Information on Social Media*

28 **Mingxin Zhang** (TUD) *Large-scale Agent-based Social Simulation: A study on epidemic prediction and control*

29 **Nicolas Höning** (TUD) *Peak reduction in decentralised electricity systems - Markets and prices for flexible planning*

30 **Ruud Mattheij** (UvT) *The Eyes Have It*

31 **Mohammad Khelghati** (UT) *Deep web content monitoring*

32 **Eelco Vriezekolk** (UT) *Assessing Telecommunication Service Availability Risks for Crisis Organisations*

33 **Peter Bloem** (UvA) *Single Sample Statistics, exercises in learning from just one example*

34 **Dennis Schunselaar** (TUe) *Configurable Process Trees: Elicitation, Analysis, and Enactment*

35 **Zhaochun Ren** (UvA) *Monitoring Social Media: Summarization, Classification and Recommendation*

36 **Daphne Karreman** (UT) *Beyond R2D2: The design of nonverbal interaction behavior optimized for robot-specific morphologies*

37 **Giovanni Sileno** (UvA) *Aligning Law and Action: a conceptual and computational inquiry*

38 **Andrea Minuto** (UT) *Materials That Matter: Smart Materials meet Art & Interaction Design*

39 **Merijn Bruijnes** (UT) *Believable Suspect Agents: Response and Interpersonal Style Selection for an Artificial Suspect*

40 **Christian Detweiler** (TUD) *Accounting for Values in Design*

41 **Thomas King** (TUD) *Governing Governance: A Formal Framework for Analysing Institutional Design and Enactment Governance*

42 **Spyros Martzoukos** (UvA) *Combinatorial and Compositional Aspects of Bilingual Aligned Corpora*

43 **Saskia Koldijk** (RUN) *Context-Aware Support for Stress Self-Management: From Theory to Practice*

44 **Thibault Sellam** (UvA) *Automatic Assistants for Database Exploration*

45 **Bram van de Laar** (UT) *Experiencing Brain-Computer Interface Control*

46 **Jorge Gallego Perez** (UT) *Robots to Make you Happy*

47 **Christina Weber** (UL) *Real-time foresight: Preparedness for dynamic innovation networks*

48 **Tanja Buttler** (TUD) *Collecting Lessons Learned*

49 **Gleb Polevoy** (TUD) *Participation and Interaction in Projects. A Game-Theoretic Analysis*

50 **Yan Wang** (UVT) *The Bridge of Dreams: Towards a Method for Operational Performance Alignment in IT-enabled Service Supply Chains*

2017

1 **Jan-Jaap Oerlemans** (UL) *Investigating Cybercrime*

2 **Sjoerd Timmer** (UU) *Designing and Understanding Forensic Bayesian Networks using Argumentation*

3 **Daniël Harold Telgen** (UU) *Grid Manufacturing: A Cyber-Physical Approach with Autonomous Products and Reconfigurable Manufacturing Machines*

4 **Mrunal Gawade** (CWI) *Multi-core Parallelism in a Column-store*

5 **Mahdieh Shadi** (UvA) *Collaboration Behavior*

6 **Damir Vandic** (EUR) *Intelligent Information Systems for Web Product Search*

7 **Roel Bertens** (UU) *Insight in Information: from Abstract to Anomaly*

8 **Rob Konijn** (VUA) *Detecting Interesting Differences:Data Mining in Health Insurance Data using Outlier Detection and Subgroup Discovery*

9 **Dong Nguyen** (UT) *Text as Social and Cultural Data: A Computational Perspective on Variation in Text*

10 **Robby van Delden** (UT) *(Steering) Interactive Play Behavior*

11 **Florian Kunneman** (RUN) *Modelling patterns of time and emotion in Twitter #anticipointment*

12 **Sander Leemans** (TUe) *Robust Process Mining with Guarantees*

13 **Gijs Huisman** (UT) *Social Touch Technology: Extending the reach of social touch through haptic technology*

14 **Shoshannah Tekofsky** (UvT) *You Are Who You Play You Are: Modelling Player Traits from Video Game Behavior*

15 **Peter Berck, Radboud University** (RUN) *Memory-Based Text Correction*

16 **Aleksandr Chuklin** (UvA) *Understanding and Modeling Users of Modern Search Engines*

17 **Daniel Dimov** (UL) *Crowdsourced Online Dispute Resolution*

18 **Ridho Reinanda** (UvA) *Entity Associations for Search*

19 **Jeroen Vuurens** (TUD) *Proximity of Terms, Texts and Semantic Vectors in Information Retrieval*

20 **Mohammadbashir Sedighi** (TUD) *Fostering Engagement in Knowledge Sharing: The Role of Perceived Benefits, Costs and Visibility*

21 **Jeroen Linssen** (UT) *Meta Matters in Interactive Storytelling and Serious Gaming (A Play on Worlds)*

22 **Sara Magliacane** (VUA) *Logics for causal inference under uncertainty*

23 **David Graus** (UvA) *Entities of Interest— Discovery in Digital Traces*

24 **Chang Wang** (TUD) *Use of Affordances for Efficient Robot Learning*

25 **Veruska Zamborlini** (VUA) *Knowledge Representation for Clinical Guidelines, with applications to Multimorbidity Analysis and Literature Search*

26 **Merel Jung** (UT) *Socially intelligent robots that understand and respond to human touch*

27 **Michiel Joosse** (UT) *Investigating Positioning and Gaze Behaviors of Social Robots: People's Preferences, Perceptions and Behaviors*

28 **John Klein** (VUA) *Architecture Practices for Complex Contexts*

29 **Adel Alhuraibi** (UVT) *From IT-Business Strategic Alignment to Performance: A Moderated Mediation Model of Social Innovation, and Enterprise Governance of IT*

30 **Wilma Latuny** (UVT) *The Power of Facial Expressions*

31 **Ben Ruijl** (UL) *Advances in computational methods for QFT calculations*

32 **Thaer Samar** (RUN) *Access to and Retrievability of Content in Web Archives*

33 **Brigit van Loggem** (OU) *Towards a Design Rationale for Software Documentation: A Model of Computer-Mediated Activity*

34 **Maren Scheffel** (OUN) *The Evaluation Framework for Learning Analytics*

35 **Martine de Vos** (VUA) *Interpreting natural science spreadsheets*

36 **Yuanhao Guo** (UL) *Shape Analysis for Phenotype Characterisation from High-throughput Imaging*

37 **Alejandro Montes García** (TUe) *WiBAF: A Within Browser Adaptation Framework that Enables Control over Privacy*

38 **Alex Kayal** (TUD) *Normative Social Applications*

39 **Sara Ahmadi** (RUN) *Exploiting properties of the human auditory system and compressive sensing methods to increase noise robustness in ASR*

40 **Altaf Hussain Abro** (VUA) *Steer your Mind: Computational Exploration of Human Control in Relation to Emotions, Desires and Social Support For applications in human-aware support systems*

41 **Adnan Manzoor** (VUA) *Minding a Healthy Lifestyle: An Exploration of Mental Processes and a Smart Environment to Provide Support for a Healthy Lifestyle*

42 **Elena Sokolova** (RUN) *Causal discovery from mixed and missing data with applications on ADHD datasets*

43 **Maaike de Boer** (RUN) *Semantic Mapping in Video Retrieval*

44 **Garm Lucassen** (UU) *Understanding User Stories: Computational Linguistics in Agile Requirements Engineering*

45 **Bas Testerink** (UU) *Decentralized Runtime Norm Enforcement*

46 **Jan Schneider** (OU) *Sensor-based Learning Support*

47 **Yie Yang** (TUD) *Crowd Knowledge Creation Acceleration*

48 **Angel Suarez** (OU) *Colloborative inquiry-based learning*

2018

1 **Han van der Aa** (VUA) *Comparing and Aligning Process Representations*

2 **Felix Mannhardt** (TUe) *Multi-perspective Process Mining*

3 **Steven Bosems** (UT) *Causal Models For Well-Being: Knowledge Modeling, Model-Driven Development of Context-Aware Applications, and Behavior Prediction*

4 **Jordan Janeiro** (TUD) *Flexible Coordination Support for Diagnosis Teams in Data-Centric Engineering Tasks*

5 **Hugo C. Huurdeman** (UvA) *Supporting the Complex Dynamics of the Information Seeking Process*

Titles in the ILLC Dissertation Series

ILLC DS-2009-01: **Jakub Szymanik**
Quantifiers in TIME and SPACE. Computational Complexity of Generalized Quantifiers in Natural Language

ILLC DS-2009-02: **Hartmut Fitz**
Neural Syntax

ILLC DS-2009-03: **Brian Thomas Semmes**
A Game for the Borel Functions

ILLC DS-2009-04: **Sara L. Uckelman**
Modalities in Medieval Logic

ILLC DS-2009-05: **Andreas Witzel**
Knowledge and Games: Theory and Implementation

ILLC DS-2009-06: **Chantal Bax**
Subjectivity after Wittgenstein. Wittgenstein's embodied and embedded subject and the debate about the death of man.

ILLC DS-2009-07: **Kata Balogh**
Theme with Variations. A Context-based Analysis of Focus

ILLC DS-2009-08: **Tomohiro Hoshi**
Epistemic Dynamics and Protocol Information

ILLC DS-2009-09: **Olivia Ladinig**
Temporal expectations and their violations

ILLC DS-2009-10: **Tikitu de Jager**
"Now that you mention it, I wonder...": Awareness, Attention, Assumption

ILLC DS-2009-11: **Michael Franke**
Signal to Act: Game Theory in Pragmatics

ILLC DS-2009-12: **Joel Uckelman**
More Than the Sum of Its Parts: Compact Preference Representation Over Combinatorial Domains

ILLC DS-2009-13: **Stefan Bold**
Cardinals as Ultrapowers. A Canonical Measure Analysis under the Axiom of Determinacy.

ILLC DS-2010-01: **Reut Tsarfaty**
Relational-Realizational Parsing

ILLC DS-2010-02: **Jonathan Zvesper**
Playing with Information

ILLC DS-2010-03: **Cédric Dégremont**
The Temporal Mind. Observations on the logic of belief change in interactive systems

ILLC DS-2010-04: **Daisuke Ikegami**
Games in Set Theory and Logic

ILLC DS-2010-05: **Jarmo Kontinen**
Coherence and Complexity in Fragments of Dependence Logic

ILLC DS-2010-06: **Yanjing Wang**
Epistemic Modelling and Protocol Dynamics

ILLC DS-2010-07: **Marc Staudacher**
Use theories of meaning between conventions and social norms

ILLC DS-2010-08: **Amélie Gheerbrant**
Fixed-Point Logics on Trees

ILLC DS-2010-09: **Gaëlle Fontaine**
Modal Fixpoint Logic: Some Model Theoretic Questions

ILLC DS-2010-10: **Jacob Vosmaer**
Logic, Algebra and Topology. Investigations into canonical extensions, duality theory and point-free topology.

ILLC DS-2010-11: **Nina Gierasimczuk**
Knowing One's Limits. Logical Analysis of Inductive Inference

ILLC DS-2010-12: **Martin Mose Bentzen**
Stit, Iit, and Deontic Logic for Action Types

ILLC DS-2011-01: **Wouter M. Koolen**
Combining Strategies Efficiently: High-Quality Decisions from Conflicting Advice

ILLC DS-2011-02: **Fernando Raymundo Velazquez-Quesada**
Small steps in dynamics of information

ILLC DS-2011-03: **Marijn Koolen**
The Meaning of Structure: the Value of Link Evidence for Information Retrieval

ILLC DS-2011-04: **Junte Zhang**
System Evaluation of Archival Description and Access

ILLC DS-2011-05: **Lauri Keskinen**
Characterizing All Models in Infinite Cardinalities

ILLC DS-2011-06: **Rianne Kaptein**
Effective Focused Retrieval by Exploiting Query Context and Document Structure

ILLC DS-2011-07: **Jop Briët**
Grothendieck Inequalities, Nonlocal Games and Optimization

ILLC DS-2011-08: **Stefan Minica**
Dynamic Logic of Questions

ILLC DS-2011-09: **Raul Andres Leal**
Modalities Through the Looking Glass: A study on coalgebraic modal logic and their applications

ILLC DS-2011-10: **Lena Kurzen**
Complexity in Interaction

ILLC DS-2011-11: **Gideon Borensztajn**
The neural basis of structure in language

ILLC DS-2012-01: **Federico Sangati**
Decomposing and Regenerating Syntactic Trees

ILLC DS-2012-02: **Markos Mylonakis**
Learning the Latent Structure of Translation

ILLC DS-2012-03: **Edgar José Andrade Lotero**
Models of Language: Towards a practice-based account of information in natural language

ILLC DS-2012-04: **Yurii Khomskii**
Regularity Properties and Definability in the Real Number Continuum: idealized forcing, polarized partitions, Hausdorff gaps and mad families in the projective hierarchy.

ILLC DS-2012-05: **David García Soriano**
Query-Efficient Computation in Property Testing and Learning Theory

ILLC DS-2012-06: **Dimitris Gakis**
Contextual Metaphilosophy - The Case of Wittgenstein

ILLC DS-2012-07: **Pietro Galliani**
The Dynamics of Imperfect Information

ILLC DS-2012-08: **Umberto Grandi**
Binary Aggregation with Integrity Constraints

ILLC DS-2012-09: **Wesley Halcrow Holliday**
Knowing What Follows: Epistemic Closure and Epistemic Logic

ILLC DS-2012-10: **Jeremy Meyers**
Locations, Bodies, and Sets: A model theoretic investigation into nominalistic mereologies

ILLC DS-2012-11: **Floor Sietsma**
Logics of Communication and Knowledge

ILLC DS-2012-12: **Joris Dormans**
Engineering emergence: applied theory for game design

ILLC DS-2013-01: **Simon Pauw**
Size Matters: Grounding Quantifiers in Spatial Perception

ILLC DS-2013-02: **Virginie Fiutek**
Playing with Knowledge and Belief

ILLC DS-2013-03: **Giannicola Scarpa**
Quantum entanglement in non-local games, graph parameters and zero-error information theory

ILLC DS-2014-01: **Machiel Keestra**
Sculpting the Space of Actions. Explaining Human Action by Integrating Intentions and Mechanisms

ILLC DS-2014-02: **Thomas Icard**
The Algorithmic Mind: A Study of Inference in Action

ILLC DS-2014-03: **Harald A. Bastiaanse**
Very, Many, Small, Penguins

ILLC DS-2014-04: **Ben Rodenhäuser**
A Matter of Trust: Dynamic Attitudes in Epistemic Logic

ILLC DS-2015-01: **María Inés Crespo**
Affecting Meaning. Subjectivity and evaluativity in gradable adjectives.

ILLC DS-2015-02: **Mathias Winther Madsen**
The Kid, the Clerk, and the Gambler - Critical Studies in Statistics and Cognitive Science

ILLC DS-2015-03: **Shengyang Zhong**
Orthogonality and Quantum Geometry: Towards a Relational Reconstruction of Quantum Theory

ILLC DS-2015-04: **Sumit Sourabh**
Correspondence and Canonicity in Non-Classical Logic

ILLC DS-2015-05: **Facundo Carreiro**
Fragments of Fixpoint Logics: Automata and Expressiveness

ILLC DS-2016-01: **Ivano A. Ciardelli**
Questions in Logic

ILLC DS-2016-02: **Zoé Christoff**
Dynamic Logics of Networks: Information Flow and the Spread of Opinion

ILLC DS-2016-03: **Fleur Leonie Bouwer**
What do we need to hear a beat? The influence of attention, musical abilities, and accents on the perception of metrical rhythm

ILLC DS-2016-04: **Johannes Marti**
Interpreting Linguistic Behavior with Possible World Models

ILLC DS-2016-05: **Phong Lê**
Learning Vector Representations for Sentences - The Recursive Deep Learning Approach

ILLC DS-2016-06: **Gideon Maillette de Buy Wenniger**
Aligning the Foundations of Hierarchical Statistical Machine Translation

ILLC DS-2016-07: **Andreas van Cranenburgh**
Rich Statistical Parsing and Literary Language

ILLC DS-2016-08: **Florian Speelman**
Position-based Quantum Cryptography and Catalytic Computation

ILLC DS-2016-09: **Teresa Piovesan**
Quantum entanglement: insights via graph parameters and conic optimization

ILLC DS-2016-10: **Paula Henk**
Nonstandard Provability for Peano Arithmetic. A Modal Perspective

ILLC DS-2017-01: **Paolo Galeazzi**
Play Without Regret

ILLC DS-2017-02: **Riccardo Pinosio**
The Logic of Kant's Temporal Continuum

ILLC DS-2017-03: **Matthijs Westera**
Exhaustivity and intonation: a unified theory

ILLC DS-2017-04: **Giovanni Cinà**
Categories for the working modal logician

ILLC DS-2017-05: **Shane Noah Steinert-Threlkeld**
Communication and Computation: New Questions About Compositionality

ILLC DS-2017-06: **Peter Hawke**
The Problem of Epistemic Relevance

ILLC DS-2017-07: **Aybüke Özgün**
Evidence in Epistemic Logic: A Topological Perspective

ILLC DS-2017-08: **Raquel Garrido Alhama**
Computational Modelling of Artificial Language Learning: Retention, Recognition & Recurrence

ILLC DS-2017-09: **Miloš Stanojević**
Permutation Forests for Modeling Word Order in Machine Translation

ILLC DS-2018-01: **Berit Janssen**
Retained or Lost in Transmission? Analyzing and Predicting Stability in Dutch Folk Songs

ILLC DS-2018-02: **Hugo C. Huurdeman**
Supporting the Complex Dynamics of the Information Seeking Process

In the context of complex tasks, information seeking has been described as a journey. The correct route, and even the final destination of this journey is often unknown in advance. Searchers may discover new paths, but also encounter ample challenges and dead-ends. In the quest for knowledge, obfuscation and illumination may go hand-in-hand, but ultimately lead to new insights.

The complex interplay of feelings, thoughts and actions during complex tasks involving learning and knowledge construction has been formally documented in various information seeking models. Carol Kuhlthau's Information Search Process model depicts a set of six stages during information-intensive tasks. The feelings, thoughts and actions of searchers evolve throughout these stages, and may include moments of optimism, uncertainty, confusion and satisfaction.

However, despite the evidence of various information seeking models, the functionality of search engines, nowadays the prime intermediaries between information and user, has converged to a streamlined set. Even though the past years have embodied rapid advances in contextualization and personalization, the Web's complex information environment is still reduced to a set of ten 'relevant' blue links. This may not be beneficial for supporting sustained information-intensive tasks and knowledge construction.

This thesis aims to shed new light on the apparent contradiction of models describing drastic changes in searchers' feelings, thoughts and actions, and the limited task support offered by current search systems. Through literature reviews, user studies and information retrieval experiments, this thesis aims to rethink the currently dominating search approach, and ultimately arrive at more dynamic support approaches for complex search tasks.

9 789082 169508