System Evaluation of Archival Description and Access

Junte Zhang

ILLC Dissertation Series DS-2011-04

INSTITUTE FOR LOGIC, LANGUAGE AND COMPUTATION

For further information about ILLC-publications, please contact

Institute for Logic, Language and Computation
Universiteit van Amsterdam
Plantage Muidergracht 24
1018 TV Amsterdam
phone: +31-20-525 6051
fax: +31-20-525 5206
e-mail: illc@science.uva.nl
homepage: http://www.illc.uva.nl/

System Evaluation of Archival Description and Access

Academisch Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. D.C. van den Boom
ten overstaan van een door het college voor
promoties ingestelde commissie, in het openbaar
te verdedigen in de Agnietenkapel
op vrijdag 30 september 2011, te 10.00 uur

door

Junte Zhang

geboren te Ruian, China.

Promotiecommissie

Promotor:	Prof. dr. T.H.P.M. Thomassen
Co-promotor:	Dr. ir. J. Kamps
Overige Leden:	Dr. W.M. Duff Prof. dr. R.R. Larson Prof. dr. J.S. Mackenzie Owen Prof. dr. H.R. Tibbo

Faculteit der Geesteswetenschappen

INSTITUTE FOR LOGIC, LANGUAGE AND COMPUTATION

ILLC Dissertation Series DS-2011-04
For further information about ILLC-publications, please contact Institute for Logic, Language and Computation.

SIKS Dissertation Series No. 2011-22
The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

NWO
Netherlands Organisation for Scientific Research

This research was supported by the Netherlands Organization for Scientific Research (NWO) Innovative Research Grants Program (VIDI Scheme) under project number 639.072.601.

Printed and bound by Off Page.
Published by IR Publications, Amsterdam
ISBN: 978-90-814485-6-7

Cover: *Empty Archives* by Hanny Breunese, 2006, Leiden.

Contents

Preface

The cover of this book is titled *Empty Archives*, and is a photo shot by Hanny Breunese in Leiden. We see a picture colored in orange with empty cupboards. Orange is the national color of the Netherlands. The research in this dissertation centers around the evaluation of online archival access to, with as case study, the National Archives of the Netherlands. Empty cupboards imply in a best case scenario that all the materials are in use. Or from a pessimistic perspective, that absolutely nothing can be found. Or in a (distant) future scenario, that everything has been digitized. It is all in the eye of the beholder. Besides the symbolism, it is in my humble opinion just a beautiful picture.

I am grateful and lucky that I have been part of the *Retrieving Encoded Archival Descriptions More Effectively* (README) project, funded by the Netherlands Organization for Scientific Research, for four years, and could contribute to it. This dissertation is based on substantially revised and extended peer-reviewed publications, and I appreciate the comments of the anonymous reviewers. I also thank my doctoral committee for their exhaustive feedback on the first draft of this dissertation.

The work has been supervised by Jaap Kamps. I am very fortunate to have worked closely with him, and he has not only been very generous, but also very supportive in every aspect. I am proud to be a direct academic descendant now.

I owe my deepest gratitude to Henny van Schie of the National Archives of the Netherlands for providing the archival finding aids in EAD, the e-mail correspondence data, and the search log files. Without this valuable data, this book would have consisted of mostly blank pages.

I also received a lot of (mental) support from Nisa Fachry, we had many open and honest talks about life, and she is absolutely one of the nicest persons I know. And Rianne Kaptein, my companion at numerous courses, and for giving me the last-minute tip on the printing company which frees up more money for more fun! If I forgot anyone else, I thank you too.

Finally, I thank my brother, my dad, and my mom.

Chapter 1

Introduction

Imagine someone conducting research on his or her ancestors who lived in Jewish refugee camps in the Netherlands, prior to, and during the country's occupation by Nazi Germany in the Second World War. This person is interested in the internment camp "Westerbork" between 1938-1944. This person wants to learn more about these camps, and about Jewish refugees in the Netherlands.

This is a real scenario sent to the National Archives of the Netherlands (*Nationaal Archief*) by a person who seeks advice on how to proceed with this research. Not coincidentally, the *Nationaal Archief* could provide answers to the questions that this person has. The ancestors of this person, or other people with a similar background, may have been mentioned in archival records that have been stored there. The records are archival materials, which are information objects that serve as evidence of past events, like governmental documents, but for example also personal letters and photos. The information contained in these records may answer the questions that this person has. People's stories can connect with archives and transform archives into social spaces of memory (Ketelaar, 2008). How does this person, with the current state-of-the-art, find these records at the *Nationaal Archief* and request access to them?

The answer is *archival description*, which Pearce-Moses (2005) defines as

> 1. The process of analyzing, organizing, and recording details about the formal elements of a record or collection of records, such as creator, title, dates, extent, and contents, to facilitate the work's identification, management, and understanding. — 2. The product of such a process.[1]

Haworth (2001, p.11) notes that archivists have developed several definitions of archival description, but the shared denominator relates to the primary mission of the archivist to describe archival materials and make them available for use. Archival description results in the creation by archivists of a surrogate, for example an *archival finding aid*. There are surrogates as archival finding aids,

[1] Retrieved 2011/06/24 from `http://www.archivists.org/glossary/term_details.asp?DefinitionKey=2765`.

which are value-added descriptive tools, and others are condensations such as abstracts. Finding aids are traditional inventories, registers, indexes, or guides (i.e. compiled in Microsoft Word) created by archivists to provide detailed information about specific archives. Bearman (1992, p.34) writes,

> "Description is focused on records both as the object being described and as the primary source of information. It seeks to characterize archival materials by constructing a document or unit surrogate. These surrogates, called cataloguing records, finding aids or archival inventories represent a 'unit of material,' or physical records. In archival description systems, these surrogates will be the fundamental record type or central file to which all indexes refer."

So archival finding aids as surrogates are descriptive text documents that identify the scope, content and arrangement of specific archive and manuscript collections. They assist users to gain access to the materials (Pearce-Moses, 2005).[2] The archival finding aids were the backbone of IR systems in 1979 (Bearman, 1979, p.180), and are still driving archival IR systems at their core (Gilliland-Swetland, 2001). People, who need archival records for their research, use archival finding aids to eventually gain access to these records.

The archival finding aids are nowadays increasingly and primarily in electronic form, published on the World Wide Web, and sometimes found by search engines. At the same time, search engines present an overview of archives, and allow for the 'activation' of archival records. To allow for the online 'activation' of archival records, archival finding aids formatted in the technical standard *Encoded Archival Description* (EAD) with *Extensible Markup Language* (XML) have become the *de facto* standard access tool (Pitti, 1997).[3] The archival finding aid in EAD is the study object in this dissertation, and it stands in our context for traditional inventories, but in 'electronic' form.

This dissertation titled *System Evaluation of Archival Description and Access* aims to increase our understanding of the use of archival finding aids in EAD for modern information retrieval from a system evaluation perspective with a queue of studies. This chapter introduces the problem statement and the main research contributions. This dissertation investigates a part of the problem statement through different angles in each chapter. This chapter describes the narrative of the thesis with the outline of the remaining chapters.

1.1 The Research Work

1.1.1 Problem Statement

There are advantages in using EAD, because we can support it with information retrieval (IR) techniques (Baeza-Yates and Ribeiro-Neto, 1999). The archival ac-

[2] For example, the findings of the study of Duff et al. (2004) show that historians highly value finding aids as sources for locating the archival materials that they seek, and as such archival finding aids assist the process of doing historical research. [3] The term *archival finding aid in EAD* is also shortened to *EAD finding aid*.

cess may become more effective, as users may find relevant records that previously were difficult to find. Or the access may become more efficient, as search engines retrieve archival records faster. For example, it enables the application of technologies of XML information retrieval (XML IR; Lalmas (2010); Sigurbjörnsson (2006)), spinning off approaches to gain focused access to archival finding aids in EAD. This means also retaining for the archival domain the findings of previous experimental user studies on XML IR that point to the positive effects of this technology on users by observing the search interaction (Balatsoukas and Demian, 2010; Hammer-Aebi et al., 2006; Kamps and Sigurbjörnsson, 2005; Larsen et al., 2006; Pharo, 2008). Still, how effective is the archival finding aid in EAD as access tool in this digital age?

Archival access depends on archival description. This access is increasingly occuring online with archival finding aids in EAD. The conjecture is that these finding aids in EAD improve archival access. These observations lead to the formal description of the *main problem statement*.

- **With large numbers of archival finding aids published online in EAD, how do searchers interact with these finding aids, and what type of retrieval system is needed to support them?**

The solution consists of establishing an IR evaluation framework with a test collection, studying effective retrieval techniques tailored to EAD finding aids, and investigating archival search behavior from a system perspective.

1.1.2 Scope and Terminology

The following mix of terms may appear as a confusing thesaurus: archive, archival record, fonds, archival collection, archival material, structured document, or archival finding aid in EAD. In this dissertation, the studies *only* focus on archival finding aids in EAD. These are text documents that describe paper archives. The finding aids in EAD are in this dissertation traditional inventories, but in 'electronic' form, for paper archives.

In practice, an archival finding aid in EAD represents archival records, contains descriptions of archival materials, and could be considered a digital archive itself. However, we note the difference between the archival materials and its descriptions. This dissertation focuses on the latter only. Furthermore, for computer scientists, an EAD finding aid is a particular genre of a structured document, may be seen as a type of database, or as a data space.

The archival finding aids in EAD capture the special structure that archivists use to manage and provide archival access. This special structure originates from archival practises and principles. This dissertation primarily studies through system-centered IR measurements this special structure for information access in the archival domain. In other words, we study access to archival finding aids in EAD. It has an information science perspective on archival finding aids in EAD. It does not aim to support or reject any theory in archival science.

1.1.3 Research Plan and Contributions

The research plan for addressing this problem statement consists of six objectives:

(a) an overview of the current state of the art in automated information access to archival materials using archival finding aids in EAD.
(b) an analysis on the evaluation of the different approaches in (a).
(c) a formalization based on the findings of (a) and (b) that results into a design and implementation of a vertical information retrieval (IR) system tailored to archival finding aids in EAD.
(d) the engineering and testing of a re-usable IR test collection for (c) to establish an evaluation framework, based on search log files and other external sources.
(e) the deepening of the understanding of archival search within finding aids by users, and use this to evaluate access to descriptions within EAD finding aids for (c).
(f) investigating the use of search profiles to enhance the accuracy of answering user questions in an EAD search system.

The research is setup modularly with a few dependencies, notably the system building in (c) and the test collection building in (d). The research method is mainly empirical and experimental in nature, based on re-using established theoretical work, and involves system building. Empirical system evaluation as established in the information retrieval research domain will investigate the effectiveness of archival finding aids in EAD as access tool.

In the archival domain, Prom (2011) points out that Web analytics can be used as method to measure user actions, to understand some aspects of user behavior, and to subsequently improve online services. In this dissertation, we also follow-up on this method, and go further by using the combination of available archival finding aids in EAD from the *Nationaal Archief* and its search logs that capture the interaction with these finding aids.

1.1.4 Research Questions

There are five general research questions to address the main problem statement.

- *Q1: How do XML retrieval techniques support a search system driven by archival finding aids in EAD?* (addresses objective (c), see Chapter 2)
- *Q2: Can we use an online archive's search log to derive a domain-specific test collection?* (addresses objective (d), see Chapter 3)
- *Q3: How effective are archival principles—inherited by traditional inventories and subsequently cast on EAD finding aids—for IR?* (addresses objectives (d) and (e), see Chapter 4)

- *Q4: How do we formally identify people's search behaviors with archival descriptions?* (addresses objective (e), see Chapter 5)
- *Q5: Can we use a search log to study different types of users and contextualize the evaluation of their specific needs?* (addresses objective (f), see Chapter 6)

The outline of the thesis lists the other main research contributions, where each chapter investigates one of these research questions. Each general research question has sub-questions that specifically addresses a general question in detail, also see the chapters. Next, we present the structure of this dissertation.

1.2 Structure of the Thesis

The thesis has the following remaining chapters:

Chapter 2 *Usage of XML Retrieval for Archival Access* primarily addresses the presentation of the README system architecture and implementation, and employs XML information retrieval technology on EAD finding aids for archival access. The system provides two-tier access, by first retrieving the whole EAD finding aid, and then the descriptions within it.

Chapter 3 We explore in *Construction of an Archival Test Collection for Evaluation* how an IR test collection can be automatically built, which is a vital component in the evaluation of IR systems. This chapter looks at a test collection for evaluating the retrieval of whole EAD finding aids only by using the README system as developed in Chapter 2.

Chapter 4 *On Archival Description Principles for Retrieval* examines how effective archival description principles projected on EAD finding aids are for effective IR of descriptions within the finding aids. We use the README system as developed in Chapter 2, and go a step further than Chapter 3 by evaluating the retrieval of descriptions within an EAD finding aid.

Chapter 5 After the system evaluation of retrieval of descriptions within EAD finding aids, *Searching within EAD Finding Aids* probes how different people search in an archival finding aid in EAD and interact with its descriptions. We introduce the EAD Search Behavior model, and put it into practice with formal experimentation, so as to capture the search behaviors.

Chapter 6 *User Stereotypes and Evaluation* presents a study on tailoring EAD retrieval systems to different user stereotypes based on interaction data extracted from search logs. Do different user groups also require different retrieval systems? It follows up on the whole EAD finding aids evaluation approach in Chapter 3, and the search behavior analysis in Chapter 5.

Chapter 7 This is the final chapter called the *Conclusion*. Here, the findings of the research will be summarized in order to offer a solution to the main problem statement. Moreover, the contributions of the work will be listed, and the scope and pointers to future work.

Chapter 2

Usage of XML Retrieval for Archival Access

This chapter investigates how EAD can be used for search systems driven by EAD finding aids, and how XML retrieval techniques can support such a system. As proof of concept and research baseline, we build a search system driven by EAD finding aids and supported by XML retrieval techniques.

2.1 Introduction

Archival access is increasingly occurring with electronic resources and shifting to the World Wide Web. The technical standard EAD contributes to this shift. We see a rendezvous between traditional information representation by archivists and information access technologies on the World Wide Web. Archival finding aids in EAD attempt to guide people to archival records. We see more usages of such technologies for archival access.[1] With the widespread use of XML in archives, there are promises, but also pitfalls. The application of XML retrieval on EAD finding aids seems natural, but is it? This leads to the main research question from a system-centered point of view, which is:

- *Q1: How do XML retrieval techniques support a search system driven by archival finding aids in EAD?*

[1] Increasingly, Web 2.0 (O'Reilly, 2007) is becoming a denominator for interactive archival access. The Polar Bear Expedition project of Yakel et al. (2007) is a realization of these ideas, in which an interactive personalized archival access system—driven by finding aids that users can comment on—has been designed. Samouelian (2009) investigates the extend of to which Web 2.0 features have been integrated into archival information systems, and suggests that archival professionals are embracing Web 2.0 to promote their digital content and redefine relationships with users. On the other side of the coin, archival access may be improved by attracting users through social media as Twitter or Facebook, but Crymble (2010) shows that participation in social media does not necessarily lead to more users. We note that in this dissertation, we only focus on IR aspects of archival finding aids in EAD.

To answer the first sub-question, we develop a system to move towards a tangible construct. Evans and Rouche (2004, p.315) point to a methodological issue by discussing the use of systems development research methods, and already suggest that adopting a user-centric prototyping approach in a research context allows for exploration of the interplay between theory and practice, advancing the practice, while also offering new insights into theoretical concepts. Therefore, we add a component in archival research methods (Gilliland-Swetland and Mckemmish, 2004). Duranti (2001) refers to archival science as a system itself, where the properties of the system can be investigated, supporting the development of new knowledge and as a demonstration of the stability of archival theory.

1.1 How can archival description principles that underlie archival finding aids be translated to an archival information system design?

Archival finding aids in EAD have multi-level descriptions in accordance with *General International Standard Archival Description* (ISAD(G), International Council on Archives (1999)).[2] This means that these finding aids can be approached from different levels of details. The second sub-question explores meaningful usages of XML retrieval techniques on EAD finding aids.

1.2 What levels of detail are possible when providing focused intellectual access to archives?

Section 2.2 explains the technical background of XML, EAD, and XML information retrieval. Section 2.3 discusses archival access with EAD finding aids. We describe our system called *README* in Section 2.4. The conclusion in Section 2.5 offers an interpretation of archival access in this digital age, after looking at the theory, real-world instances of archival access applications, and the development of such an application—which is our research baseline.

2.2 Related Work

This section introduces and explains XML, EAD, and XML retrieval with their key concepts relating to the chapter's problem statement.

[2] EAD is an encoding standard, and it enables sharing archival records among archives. This is reflected in the close link between EAD and ISAD(G), which has been approved by the International Council of Archives (ICA) as a standard that defines the elements that should be included in an archival finding aid. It is a hierarchical model. The advent of archival finding aids in EAD has made ISAD(G) particularly useful, because it enables the sharing of records between archives by upholding four principles that all archives in theory should share in their multi-level description of each record, namely (i) the description of the general to the specific; (ii) information should be relevant to the level of description; (iii) descriptions should be linked between levels by making clear the level of the unit of description; (iv) non-repetition of information by using the hierarchy of description.

2.2.1 Extensible Markup Language

Goldberg (2008) presents an illustrative guide on Extensible Markup Language (XML, Bray et al. (2008)). It was designed to manage information. Moreover, it is also a specification for describing the structure of that information. XML evolved from Standard Generalized Markup Language (SGML, Goldfarb (1984)). SGML is a metalanguage—a language that describes languages—that has proved useful in many large publishing applications. However, a problem to adoption is that SGML is too general and is more complex than Web browsers can cope with (Bosak and Bray, 1999). Therefore, Bosak and Bray (1999) created XML consisting of rules that anyone can follow to create a markup language from scratch. EAD is such a markup language.

The basic XML concepts relevant to our research on content-orientated information retrieval—and which we discuss here—are the notion of an *element*, the distinction between *well-formed* and *valid* XML with for example a Document Type Definition (DTD). We show how XML documents, such as an EAD finding aid, can be viewed as *trees*.

Element Angle-bracketed labels, such as `<titleproper>` or `<c01>` in Figure 2.2, are called tags. In XML, there is always a begin tag as `<titleproper>` and its corresponding end tag `</titleproper>`. Anything between this pair of tags is called an XML element. This can be another element or more, which is then called a sub-element, or text (content) only, or instances of both which is called mixed content.

Well-formed The Web's main language is Hypertext Markup Language (HTML), which is a electronic-publishing language, mainly for easily creating the layout—or appearance—of Web pages using a W3C pre-defined set of tags (Berners-Lee et al., 1994; Bosak and Bray, 1999). Unlike HTML, XML is a markup language that is extensible, because it allows for the definition of custom tags that are meaningful to people. However, any XML document must satisfy certain rules in order to be XML. When an XML document satisfies these rules, it is well-formed, and only then it can be computationally processed. Goldberg (2008) lists these rules, such as the case-sensitivity of tag names, and the non-overlap of XML elements, where there is only one root element and each element has only one parent, so as to enforce a strict hierarchy of elements.

Valid An XML document is only XML, when it is well-formed, i.e. conforms to a syntax. However, additional rules can be enforced in terms of the vocabulary (which elements and attributes can be used) and grammar (frequency of elements and nesting), therefore effectively allowing the definition of a custom markup language. A document type definition (DTD) is often used to make a set of XML documents also valid. DTDs are essentially extended context-free grammars expressed in a notation that is similar to the extended Backus–Naur form (EBNF) notation (Knuth, 1964;

Wirth, 1977).

Tree In computer science, a tree is a widely-used data structure that equals a hierarchical tree structure with a set of linked nodes. Given the hierarchy of elements, an XML document also consists of a hierarchical set of linked nodes. Let us take Figure 2.2 as an example of an XML document. We see that `<ead>` is the root node, which is a member of the node set that has no parent (or 'superior'). The lines connecting the nodes (which equals elements) are named 'branches.' Element `<unitid>` is a node without children, and is called a 'leaf' node. The nodes in the tree have several depths, where the root node has a depth of one, element `<archdesc>` has a depth of three, or `<c01>` has a depth of four. Any tree can be mapped to a DTD, and a DTD can be used to construct a tree.

Bosak and Bray (1999) sketch what can be expected from XML. An expectation is that XML will solve some of the Web's biggest problems, which is the difficulty to find the one piece on the Web that you need. Common uses of the tree data structure, and thus XML, is to manipulate hierarchical data and make information easy to search. XML is primarily used to exchange information between either (i) people and systems, (ii) among people, or (iii) among systems. In the case of the former, the logical structure in an XML document is called document-centric, and in the case of the later, it is called data-centric. In XML there is the fundamental concept of separating the logical structure of a document from its layout, where the latter is usually created in Cascading Style Sheets (CSS) or with HTML, possibly in conjunction with XSLT (Goldberg, 2008). However, it is the exploitation of the logical structure for information access that is of primary interest in XML retrieval. What is the logical structure of EAD?

2.2.2 Encoded Archival Description

Forde (2005) discusses changes that have occurred in access and preservation, and mentions surrogacy programmes. The initiative that started in 1997 to materialize by creating 'electronic' finding aids in a technical standard called *Encoded Archival Description* (EAD, (Pitti, 1997)) is also a surrogacy program. EAD is jointly maintained by the Library of Congres (LoC) and the Society of American Archivists (SAA). Surrogacy programmes enable long-term use and remote continuous access to archives. Microform technology was a step towards this access (Forde, 2005, p.194), but with the emergence of the World Wide Web (WWW, Berners-Lee et al. (1994)), remote and continuous access is possible because of online access.

The sharing of geographically 'scattered' records is a motivation for the development of EAD (Pitti, 1997, p.269). An important promise of automation is the desire of archivists to have the ability to cooperate with each other, and to exchange records with other systems (Bearman, 1979, p.183). Haworth (2001, p.21) notes that the emergence of EAD as standard could not be avoided given the plethora of 'electronic' finding aids independently created and contained in a variety of software-independent systems which limited their exploitation and sharing. Kiesling (1997) remarks that strict and unambiguous adherence to the standard is needed to achieve this goal, and EAD may be the catalyst.

An increasing number of archives and manuscript libraries, and also museums (e.g. Chandler (2002)), use EAD to encode archival descriptions that describe unique primary resources in the form of archival materials, such as corporate records and personal (hand-written) papers, and for access (Pitti, 1999). For example, Hill et al. (2005) discuss three different online services providing access to finding aids relating to three different archives, but the similarity is the key role of EAD and its application for creation, storage, indexing, searching and presentation of finding aids. These archival collections may have millions of unique items, which can be in any form or medium. For example, plans, drawings, charts, maps, photographs, audio, and video (Pitti, 1999).

The archives are organized and described hierarchically. EAD consists technically of a set of descriptive elements to describe the archives. Figure 2.1 depicts a part of the official EAD Document Type Definition (DTD). The version of EAD researched in this dissertation is the official 2002 version.[3] The three highest level elements are `<eadheader>`, the optional `<frontmatter>`, and the heart of an archival finding aid consisting of archival descriptions in `<archdesc>`— the body of archival materials (Ruth, 1997, p.320). These descriptions usually consist of a Descriptive Identification Element `<did>` and a Description of Subordinate Components `<dsc>`. The components `<C`$_n$`>` of the whole are recursively nested in `<dsc>` within `<archdesc>`, where $n \in \{01, ..., 12\}$, see Figure 2.1 and 2.4. In other words, `<c02>` is the sub-component of `<c01>`, and so on. A component can also be unnumbered. The EAD files can be deeply nested and lengthy in content with thousands of pages (or more) (Pitti, 1999). Figure 2.2 shows the XML source code of the *Nationaal Archief* EAD finding aid '2.09.70' on the topic 'Bijzondere Rechtspleging' (in English: *Special Justice*). This topic deals with justice in the Netherlands after the Second World War for people who committed crimes during the War. Figure 2.3 sets out the HTML presentation of the *Nationaal Archief* system (of 2009) of this finding aid.[4]

2.2.3 XML Information Retrieval

Traditionally, IR has dealt with the retrieval of complete documents or full-texts. Usually, a user is only interested in a part of a document. As Hjørland (2000, p.33) argues, there is a distinction between document retrieval and fact retrieval, and puts forward the idea to store and only retrieve the facts or 'in-

---

[3] Retrieved 2011/01/11 from `http://www.loc.gov/ead/ead2002a.html`

[4] The *Nationaal Archief* revised its system in the middle of 2011. The examples used in this dissertation, and the studies conducted, are before this change.

```
<!ENTITY % m.desc.full
  '%m.desc.base; | dsc | dao | daogrp | note'>
<!ENTITY % m.did
  'abstract | container | dao | daogrp | langmaterial | materialspec |
   note | origination | physdesc | physloc | repository | unitdate |
   unitid | unittitle'>
<!ELEMENT ead
   (eadheader, frontmatter?, archdesc)>
<!ELEMENT eadheader
   (eadid, filedesc, profiledesc?, revisiondesc?)>
<!ELEMENT archdesc
   (runner*, did, (%m.desc.full;)*)>
<!ELEMENT did
   (head?, (%m.did;)+)>
<!ELEMENT dsc
   ((head?, tspec?, (%m.blocks;)*),
    (((thead?, ((c, thead?)+ | (c01, thead?)+)) | dsc*)))>
<!ELEMENT c01
   ((head?, did, (%m.desc.full;)*, (thead?, c02+)*) |
    (drow+, c02*))>
```

Figure 2.1: A part of the EAD Document Type Definition.

formation' contained in the documents.

The indexing and retrieving of elements, which may contain facts or 'information,' in XML documents is done using XML information retrieval (XML IR, Lalmas (2010)), which is a branch of information retrieval that deals with the retrieval of arbitrary parts of XML files given the document-centric XML structure, and attempts to use the XML markup of documents to the fullest for 'focused' information access by not only providing direct access to a whole document, but also to a part of it. This also conforms to the principle of least effort (Zipf, 1949), as it deals with the burdon of reading more information (i.e. full-text) than is actually necessary. We then also speak of *focused retrieval* (Joty and Sadid-Al-Hasan, 2007; Lalmas, 2010; Trotman et al., 2007). This markup represents the different granularities and complexities of these files. The structure is exploited to expose information. For example, in response to a user query on a collection of digital textbooks marked-up in XML, an XML retrieval system may return the content of a paragraph, section or chapter estimated to best answer that query.

As Lalmas (2010) notes, the concept of XML IR existed before it became known as such. She points out that with the introduction of XML in 1997, there were proposals to gain access to logically structured documents—with DeRose (1997) making the case for EAD—before the existence of the term XML IR. For example, Wilkinson (1994) shows in 1994 that knowledge of the structure of

```
<?xml version="1.0" encoding="utf-8"?>
<ead>
 <eadheader countryencoding="iso3166-1" dateencoding="iso8601"
langencoding="iso639-2b" scriptencoding="iso15924"
repositoryencoding="iso15511" findaidstatus="unverified-full-draft">
   <eadid countrycode="NL" mainagencycode="NL-HaNA"
publicid="-//Nationaal Archief//TEXT (NL-HaNA::2.09.70::Ministerie van
Justitie: Directoraat-Generaal Bijzondere Rechtspleging, Personeelsdossiers)//NL"
urn="2.09.70.ead.xml">2.09.70</eadid>
   <filedesc>
     <titlestmt>
       <titleproper>Inventaris van de archieven van het
       Directoraat-Generaal voor de Bijzondere Rechtspleging
       (1945-1958), met taakopvolgers en uitvoerende instanties
       (1945-1986); Deel III Personeelsdossiers</titleproper>
       <author>Centrale Archiefselectiedienst, Winschoten</author>
     </titlestmt>
     <publicationstmt>
       <publisher>Nationaal Archief, Den Haag</publisher>
       <date normal="2008" era="ce" calendar="gregorian">(c)
       2008</date>
     </publicationstmt>
   </filedesc>
   <profiledesc>
     <creation audience="internal">Deze digitale toegang is in
     <DATE normal="2009" era="ce" calendar="gregorian">
     2009</date>vervaardigd door het Nationaal Archief op basis
     van de richtlijn
     <title linktype="simple">Het gebruik van Encoded Archival
     Description op het Nationaal Archief versie 1.7.2</title>.
     Eindredactie: Henny van Schie,
     <date normal="20090429" era="ce" calendar="gregorian">6 mei
     2009</date>.</creation>
     <langusage>This finding aid is written in
     <language langcode="dut" scriptcode="Latn">
     Dutch</language>.</langusage>
     <descrules audience="internal"/>
   </profiledesc>
   <revisiondesc audience="internal">
     <change>
       <date normal="" era="ce" calendar="gregorian"/>
       <item/>
     </change>
   </revisiondesc>
 </eadheader>
 <archdesc level="fonds" type="inventory">
   <did>
     <head>Beschrijving van het archief</head>
     <unittitle label="Naam archiefblok:">Ministerie van Justitie:
     Directoraat-Generaal Bijzondere Rechtspleging,
     Personeelsdossiers</unittitle>
     <unittitle type="short">DGBR, Personeelsdossiers</unittitle>
     <unitdate label="Periode:" type="inclusive" normal="1945/1986"
      era="ce" calendar="gregorian">
     1945-1986</unitdate>
```


FIGURE 2.2: A part of the *Nationaal Archief* 's XML encoding of its largest EAD finding aid '2.09.70' (*'Inventaris van de archieven van het Directoraat-Generaal voor de Bijzondere Rechtspleging (1945-1958), met taakopvolgers en uitvoerende instanties (1945-1986); Deel III Personeelsdossiers'*) in terms of file size.

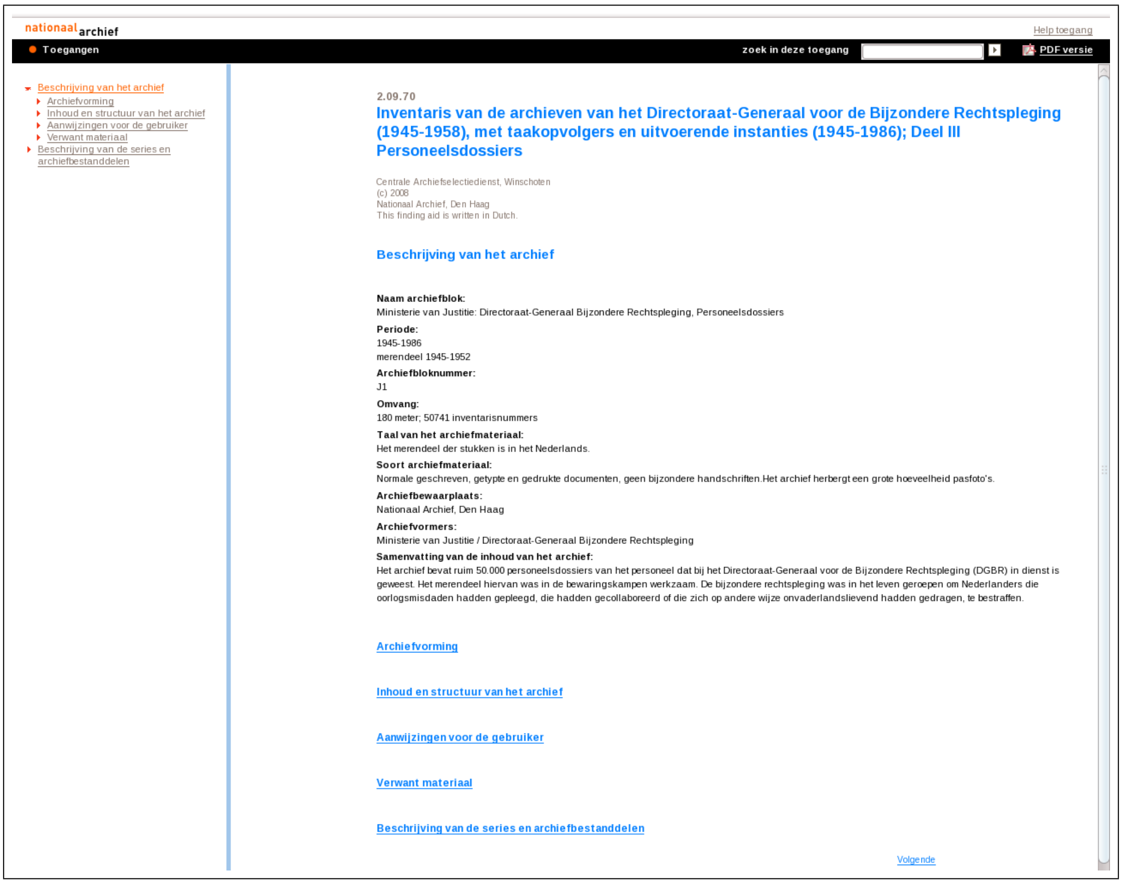


Figure 2.3: *Nationaal Archief* 's HTML presentation of its largest EAD finding aid '2.09.70' in terms of file size. Image courtesy of the *Nationaal Archief*.

documents can lead to improved retrieval performance. Numerous user studies with different experimental setups (e.g. Hammer-Aebi et al. (2006); Larsen et al. (2006); Pharo (2008)) also show that users prefer to interact with smaller nuggets of information. Furthermore, in the archival realm, Bearman (1979, p.188) implicitly referred in 1979 to the pivotal concepts in XML IR of granularity and direct access while the state of the art of that time was subject retrieval based on provenance and content indexing (Lytle, 1980).

The INitiative for the Evaluation of XML retrieval (INEX) started in 2002 to provide an infrastructure—common evaluation benchmark—for evaluating the effectiveness of content-orientated XML retrieval systems (Gövert and Kazai, 2002a). The predominant path to evaluate the system retrieval effectiveness is with the use of test collections, which usually consists of a set of documents, user requests referred to as topics, and relevance assessments specifying the set of 'right answers' for the user requests. The INEX infrastructure is defined by Gövert and Kazai (2002a, p.1) as a test collection of real world XML documents along with standard topics and respective relevance assessments.

We briefly recall the history and methology of INEX. The topic development process consisted of submission of candidate topics by the research participants themselves, then the participants had to judge a topic's top hundred retrieved results. This is manual and labor-intensive. INEX started with a test collection consisting of fifty-five topics inspired on articles from the IEEE Computer Society (Gövert and Kazai, 2002a). There were two types of topics: (i) the *content-only* (CO) topics that request only content related conditions with no knowledge of the document structure, and (ii) *content-and-structure* (CAS) topics which refer explicitly to the XML structure by specifying target elements (e.g. search only in the Summary). The research participants had so submit a ranked list of results that their system produced (i.e. 'runs') so as to compare their results. This type of experimentation continued in 2003 (Fuhr et al., 2004), 2004 (Malik et al., 2004), and 2005 (Malik et al., 2005).

In 2006, INEX used another XML document collection consisting of English articles from the Wikipedia project (Denoyer and Gallinari, 2006; Malik et al., 2006). A reason for this switch is that Wikipedia's content can also appeal to users with backgrounds other than computer science, as it was the case with the IEEE articles, making 'realistic' user studies—observation and measurement of user interaction with XML documents—possible. User studies in IR are driven by tasks, a user is asked to do something.[5] In IR, the domain-knowledge has a positive effect on search interaction (e.g. Lazonder et al. (2000); White et al. (2008)). To ask a random user to search for information related to their interests seems more suitable than a random topic.

INEX consists of different tracks, which are different research paths that investigate a part of the challenges of focused retrieval. Fuhr et al. (2007) discuss the Ad Hoc Track, which is the predominant track of INEX. Ad hoc search is described as a simulation of how a library might be used and involves searching in a static set of documents given a new set of topics. It follows the same methodology as set out in previous years, but also allows passage retrieval (Trotman et al., 2007). Simulating real use cases, this track also features tasks. Prior to 2007, there was the Thorough Task that asks systems to return any elements ranked by their relevance to the topic (Malik et al., 2006). Since 2007, INEX features three tasks: (i) the *Focused Task* that asks to return a ranked-list of non-overlapping results, (ii) the *Relevant in Context Task* that requests the return of a non-overlapping ranked-list of results, but grouped by document, and (iii)

[5] However, user studies are not necessarily task-based. Archival user studies have involved interviews to gather data unrelated to specific tasks. Understanding users is pivotal to evaluate and improve systems, and to understand the interaction—a dialog of a user elaborating and changing the needs and a system responding to this behavior—between systems and users (Lesk, 2004). However a difficulty is, as Lesk (2004, p.227) writes, that users are not always able to make clear what they actually search for. Studies in the archival domain have set the first steps towards understanding archival interaction by investigating the users' information seeking needs and behaviors. The study of Yakel (2004) researches what genealogical researchers want and do by conducting interviews with them. Ultimately, this type of studies helps us not only to understand the user needs, but to improve information systems for people.

the *Best in Context Task* that asks to return a single starting point of a document. Kazai et al. (2004) explain the issue of overlap as the retrieval of multiple overlapping (i.e. nested) result elements, e.g. a system may retrieve the same nugget of information from multiple nested elements. As the results of the user study of Hammer-Aebi et al. (2006) show, users do not appreciate overlap when they search for information, and as such, it must be removed. What does this all mean for EAD retrieval?

2.3 Archival Access: Information Retrieval with EAD

This section aims to investigate this chapter's first research question. We look at the utilization of EAD for information retrieval in terms of its particular structure and its applications by archives.

2.3.1 Retrieval with EAD

Ruth (1997) presents a structural overview that explains the use of EAD markup with illustrative code examples, and also the link with the ISAD(G) principles. DeRose (1997) also points to the difference in content and structure, and the need of structure to search in content.[6]

As illustrated in Tsikrika (2009), structured text retrieval supports the representation and retrieval of the individual document components defined by the logical structure as represented in a hierarchical document, such as an EAD file. This structure can be distinguished in two types of units Tsikrika (2009): (a) atomic units (or 'text content elements') that only contain text and no XML elements, and (b) composite units (or 'nested elements') that contain other units and can be further 'decomposed'. The same is true for EAD, see Figure 2.4, where atomic units such as `<unitid>` or `<unitdate>` are represented as leafs and composite units like `<did>` are non-leaf nodes. However, we extend this representation with *mixed content* nodes, i.e. elements that contain both text and other elements, where there are usually multiple text nodes, and each text node in an element can be retrieved separately in XML Path Language (XPath, Clark and DeRose (1999)). An instance of a mixed node could be the composite unit `<unittitle>` that may have been annotated with a semantic tag like `<persname>`.

There is no shortage of metadata in archival finding aids, but it is "just a matter of finding the right hook to make them more accessible" (Kiesling, 2001,

[6] The move from simple flat-form databases—such as catalogs and abstracts—to highly structured documents— such as EAD finding aids—makes the question of what kind of structure to represent increasingly more important (DeRose, 1997, p.304). An answer is to use structure that increases the ease of use (DeRose, 1997, p.308). The structure of finding aids is characteristically hierarchical with layers upon layers of substructure. Archival finding aids in EAD are also navigation tools with hypertext links, mimicking paper documents with indexes and table of contents (DeRose, 1997, p.307).

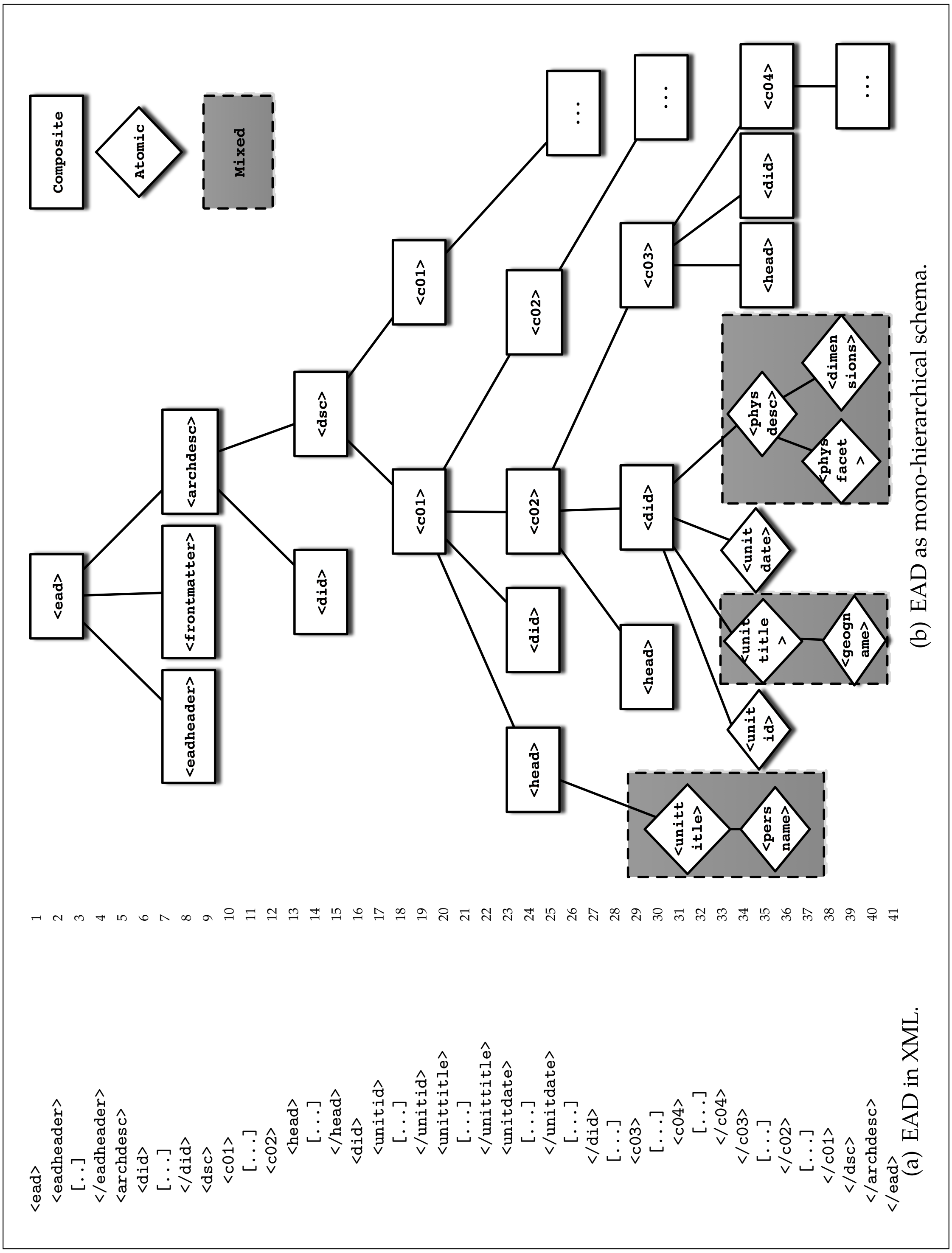


Figure 2.4: Representation of Encoded Archival Description as mono-hierarchical schema using XML elements as nodes for information retrieval (rotated clockwise).

p.87). XML retrieval techniques can be employed to deal with this problem, to maximally and most effectively exploit these 'hooks.' Using this markup, we could zoom into any of them—at the same time index and retrieve them.

2.3.2 Archival Metadata Retrieval Systems

Archival data encoded in EAD is structured data. Commonly used relational databases do not provide a perfect solution to store this type of data. XML databases are developed instead to provide a better solution to capture and preserve the richness of the structure in a data structure, though the question is still open whether archival users would seek information using powerful queries with XML structure, and whether it would be more effective for retrieval. A pointer to an answer is the study of Duff and Stoyanova (1998) which presents results of rating elements for access, and not surprisingly, the Title was perceived as the most useful element for retrieval. Another question is whether archivists would accept such technology, with Yaco (2008) pointing to the challenges of the technical implementation of EAD, and with Yakel and Kim (2005) also noting the overall slow adoption due to factors as the small staff size of many repositories, the lack of standardization in archival description, a multiplicity of existing archival access tools, insufficient infrastructure, and difficulty in maintaining expertise.

A multiplicity of existing archival access tools is available, including several open-source solutions, such as eXist (Meier, 2003). Other alternatives more commonly used for archival finding aids in EAD are PLEADE (EAD on the Web) developed in France and used by the French National Archives (Sévigny and Clavaud, 2003), Cheshire3 (Larson and Sanderson, 2006) as used by the Archives Hub in the UK, Archon developed at the University of Illinois (Prom et al., 2007; Schwartz et al., 2007), or the Digital Library eXtension Service (DLXS)[7]. Table 2.1 presents an overview of these and additional systems.

What are the features in terms of archival access do these systems offer? The analysis of search features available for Web sites driven by EAD (instances of digital archives) by Zhou (2007) shows that eighteen of the fifty-eight Web sites (or 22%) did not have any search features. Moreover, nearly half of the tested systems did not have a proper ranking, like ranking by call number order—which is for users often pointless—or the results are presented randomly. However, it should be noted that this study looked at Web sites only, which may be driven by the same systems. Similarly, another survey is presented by Huffman (2008) that shows an overview of search features of thirty-three Web sites driven by EAD that also include a search engine. A finding of that study is that less than half of the investigated archives had options to search in metadata fields, and most performed full-text searches. Although Zhou (2007) and Huffman (2008) look at XML-based EAD search systems, it is still unknown what

[7] Retrieved 2011/05/11 from `http://www.dlxs.org/docs/12a/intro/arch.html`.

systems have to offer to archives in terms of actual XML retrieval.

Therefore, Table 2.1 depicts our overview of the most widely used systems for (archival) metadata retrieval, so we do not investigate the archives, but the technical backbone that drives them. We investigate these systems by auditing their documentation and testing their implementations. The table shows the names of the systems, how the content and structure have been indexed, whether it supports a type of relevance ranking (or whether it is plain batch data retrieval), and institutions which have adopted the system. The systems are either available as open-source software, or as a commercial product. We make a distinction between field-based indexing and full XML-based indexing. The former refers to indexing based on certain metadata fields, like for example Names, Dates, Titles, Places, Subjects, Repository, or Summary, hence generally, indexing happens only with top-level components. The latter means that it is fully XML-based, where all XML structure of the original EAD documents is preserved, including their depth and breadth, and it is possible to optimally query the database with a XML query language like XQuery (Boag et al., 2002; Chamberlin, 2003).

For archives that seek to use open-source software, only Cheshire3 has full XML-based indexing with relevance ranking support, or else they are faced with field-based indexing. For commercial systems, there are the MarkLogic and TEXTML systems which provide both full XML-based indexing and information retrieval based on relevance ranking. All the (archival) metadata systems in Table 2.1 are distinctly different, though XTF partly uses Lucene (Hatcher and Gospodnetic, 2004) for information retrieval.

Besides retrieval using the EAD representation, there is the layout (presentation, user interface). The EAD Cookbook (Prom, 2002) consists of XSLT stylesheets to render a user interface from EADs. There are also content management tools. The Archivists' Toolkit is software particularly developed for this purpose (Westbrook et al., 2007). There are content management systems tailored to archival description that also include an access component, such as Archon—using a home-built module—or ICA-AtoM[8] that uses Lucene.

2.4 README System Description

This section covers the technical presentation of the README system.

2.4.1 Usage Scenarios

We can envision usage scenarios in designing a system. These scenarios are informal stories about user tasks and activities. Similarly, in the archival domain, and with EAD finding aids, we describe the following (original) usage scenarios, which have been randomly picked, that the system should be able to

[8] Retrieved 2011/01/11 from `http://ica-atom.org/`.

Table 2.1: Overview of commonly used (archival) metadata retrieval systems.

	System	Indexing	Relevance Ranking	Examples
Open-source	*Archon*	Field	no	UC Berkeley, Texas A&M University
	Cheshire3	Full XML	yes	British Library ISTC, Archives Hub, SHAMAN Digital Preservation Project
	eXist	Full XML	no	Columbia University Libraries
	Lucene/SOLR	Field	yes	Smithsonian, Rhode Island Archival and Manuscript Collections Online, University of Virginia
	PLEADE	Field	yes	French National Archives
	XTF	Field	yes	Online Archive of California, California Digital Library, Arizona Archives Online, Ohio Archives Online
Commercial	*Adlib Archive*	Field	no	Aston Martin Heritage Trust, National Archives of Ireland
	ARCHI-LOG	Full-text	no	Eastern Townships Research Center, numerous French-Canadian archives
	DigiTool	Field	yes	Boston College, Leiden University, Florida State University, Universidade do Porto, University of Melbourne
	DLXS	Field	yes	National Library of Medicine, University of Michigan Museum of Art
	MAIS-Internet	Full-text	yes	72 Dutch archives on Archieven.nl
	MarkLogic	Full XML	yes	University of Chicago, Elsevier, Greenwood Publishing, McGraw-Hill Education, Oxford University Press, Princeton Theological Seminary, University of Toronto Library, Library of Congress
	TEXTML	Full XML	yes	Northwest Digital Archives, Rocky Mountain Online Archive, A2A

cater to. These typical usage scenarios have been copied and pasted from real world e-mail correspondence between archivists and people asking for assistance from the *Nationaal Archief*. In IR, these are also known as *ad hoc* queries.

- *"Do you know whether the 'Ministerie van Verkeer en Waterstaat' has an own archive which includes PTT documents relating to wireless telegraphy?*
- *"Please can you send me the link or the pdf files that contain the Catalogue of Ancient Dutch Maps."*
- *"I am writing a book about foreigners in Beijing from the Boxer Rising in 1900 to the Communist takeover 1949. Jonkheer Frans Beelaerts van Blokland was the Dutch Minister in Peking during the World War One. I am very interested in seeing any papers that you may hold relating to his years in Peking."*
- *"As per our earlier telephone conversation, I would like to get more information about my family. I have been able to trace back some correspondence concerning my Grandfather (–name removed–) under the dossiers released over the Westerling affair. However, I am unable to trace records covering my Grandfather's activities during WWII."*
- *"I am a trying to find documents related to Mexico's independence. Professor –name removed–, from Leiden University, told me that a Dutch Consul (RHL Heidsieck) who served in Mexico in the mid XIX century, donated some of them to the archive of the foreign ministry. Could you please advise me how to proceed. I am willing to visit the archives soon."*
- *"I am researching the Netherland Cyprus Consulate in XIX. century. Later, I want to extend my research. So my research will include the Netherland Consulates in the Ottoman Empire in XIX. century. Therefore, I want to learn whether there are any berats or fermans in your Archives. If it is, how can I read them? I can read Ottoman Turkish."*
- *"I am conducting research on the Jewish refugee camps and deportation camps in Holland from 1938 to 1944. My family came from Ostfriesland, Germany to Rotterdam in December 1938. They lived in the refugee camp 'Koninginnehoofd' at Wilhelminakade No. 74 in Rotterdam until 1940. I have learned many facts about Camp Westerbork. Is there any way to learn more about Camp Koninginnehoofd or about Jewish refugees in Rotterdam at this time?"*

2.4.2 EAD Finding Aids

We have obtained in total 5,934 EAD finding aids from the *Nationaal Archief* (National Archives of the Netherlands or NA), where the statistics are shown in Table 2.2 with the distribution of the length of content in bytes (without XML markup) and of the structure in terms of XML tags.

When we assume that one type-written page is 2 kilobytes in size, then there are 46 EAD finding aids that have more than 250 pages (512,000 bytes) in terms of the *content only*. On average, the content of an EAD finding aid consists of 19 type-written pages, and the median of the content length is between 3-4 pages.

TABLE 2.2: Statistics of the *Nationaal Archief* 's EAD finding aids.

Feature	Value
Number of Files	5,934
Content Length (bytes)	M=38,605 (Mdn=7,662)
XML Markup (tags)	M=2,322 (Mdn=544)
Correlation Content + Markup	Pearson's $r = .922$, Spearman's $\rho = .855$, Kendall's $\tau = .685$

According to Pitti (1999), detailed archival descriptions average 15-30 pages in length. So on average, the *Nationaal Archief* EAD finding aids can be considered 'detailed.' We show that there is strong positive and significant correlation on a 1% significance level (2-tailed) between content length and XML markup (Pearson's r, Spearman's ρ and Kendall's τ), likely due to the completeness of the markup.

The distribution of the top ten most frequently occurring elements over all finding aids is shown in Table 2.3. We see that `<unittitle>` is the most frequently occurring tag in the *Nationaal Archief* 's set of EAD finding aids, followed up by element `<did>`, and `<unitid>`. These three elements alone make up for more than fifty percent of all EAD structure. These elements are used to describe the lists of files and items in the `<dsc>` nodeset.

TABLE 2.3: Top 10 distribution of elements ($N = 13,775,835$).

Element	Count	Percent
<unittitle>	2,397,086	17.40
<did>	2,347,000	17.04
<unitid>	2,318,869	16.83
<unitdate>	1,595,161	11.58
<physdesc>	647,012	4.70
<p>	556,576	4.04
<c03>	459,269	3.33
<c04>	401,438	2.91
<c05>	361,138	2.62
<c02>	360,561	2.62

2.4.3 Indexing and Storage

Before the indexing and storage, we preprocess the files to make them strictly well-formed XML—which was a prequisite for indexing in an XML database. We do not check for validity, and add minor modifications to the official 2002 EAD standard, of which the most noticeable one is that we convert all XML tags to uppercase. We pre-process the EAD files by adding the physical filename and the title `<titleproper>` as attributes in the root node `<ead>` using XSLT, then process them again in XML Lint, and then cleaning them up (like making all tags uppercase) in Tidy.[9] Since we deal with mostly Dutch language data, we use the ISO/IEC UTF-8 character encoding.

The system is based on MonetDB with the XQuery front-end Pathfinder (Boncz et al., 2006) and the information retrieval module PF/Tijah (Hiemstra et al., 2006). All of the *Nationaal Archief* 's 5,934 EAD finding aids made available to us are indexed into a single main memory XML database that completely preserves the XML structure and allows for powerful XQuery querying. We index the collection without stopword removal, and use the Dutch snowball stemmer.

2.4.4 Retrieval Model

For the retrieval of individual and any arbitrary elements, we employ statistical language models (LM) (Ponte and Croft, 1998), i.e. the probability distribution of all possible term sequences is estimated by applying statistical estimation techniques. The probability of each individual term is calculated using the *maximum likelihood estimate* (mle), which corresponds to the relative frequency of a term t_i in an element e, $P_{mle}(t_i|e) = \frac{\mathrm{tf}_{i,e}}{\sum_t \mathrm{tf}_{t,e}}$ where $\mathrm{tf}_{i,e}$ is the term frequency t_i normalized by the sum of all frequencies in an element e.

We estimate the probability that the element model can generate the given query q. By applying Bayes' theorem, this can be obtained by

$$P(e|q) = \frac{P(q|e) \cdot P(e)}{P(q)} \propto P(q|e) \cdot P(e) \tag{2.1}$$

where $P(q)$ can be ignored for ranking, and the prior $P(e)$ is assumed to be uniform. The query likelihood (or conditional probability) is based on a model that represents an element using a multinomial probability distribution over a vocabulary of terms. For each element, a model on an element is inferred, such that the probability of a term given that model is $p(t|e)$. The model is then used to predict the likelihood that an element could match a particular query q. We make the assumption that each query term can be assumed to be sampled identically and independently from the element model. Applying this assumption, the query likelihood is obtained by multiplying the likelihoods of the individual terms contained in the query:

$$P(q|e) = \prod_{t \in q} P(t|e)^{n(t,q)} \tag{2.2}$$

where $n(t, q)$ is the number of times term t is present in query q.

[9] Retrieved 2011/05/11 from `http://www.xmlsoft.org/xmllint.html` and `http://tidy.sourceforge.net/`.

To deal with zero probabilities because of non-existing terms in case there is sparse data, smoothing techniques are applied. The retrieval model uses Jelinek-Mercer smoothing, which is a mixture model between the element model and the collection as background model, so

$$P(t|e) = (1 - \lambda) \cdot P_{mle}(t|e) + \lambda \cdot P_{mle}(t|C) \tag{2.3}$$

where $P_{mle}(t|C) = \frac{ef_t}{\sum_t ef_t}$, ef_t is the element frequency of query term t in the collection C, and the λ is set to 0.15.

2.4.5 Querying and User Interfaces

We discuss now the three approaches deployed in the README system, which is written in Perl using XHTML, CSS, and JavaScript. The connection with the database server is made using a socket and XML RPC. After processing the query in the system, we post-process the results. An issue in post-processing of XML retrieval results is content overlap due to the hierarchy and nesting of elements. Kazai et al. (2004) explain the overlap problem in content-orientated XML retrieval evaluation. Hammer-Aebi et al. (2006) find that the problem of overlapping elements can be solved in end-user systems at the interface level with a contextual view that groups results by document, instead of using an atomic view of element retrieval with a ranked list. Conversely, for convenience, we remove overlap of content by default in our XML IR systems.

The origin of the finding aid is made clear by showing an icon in front of a result that corresponds to a source. In our case, there is one source namely the *Nationaal Archief*, but the system is also setup to allow for search across institutions. For each retrieval approach, we present user interfaces of the README system with as example the query "bijzondere rechtspleging" (in English: *special justice*) and its retrieved results.

Approach 1: Document Ranking

The XML database is queried using XQuery extended with Narrowed Extended XPath I (NEXI) (Trotman and Sigurbjörnsson, 2004). For document (collection) ranking, we provide the root element `<ead>` (the whole EAD finding aid) as target element, i.e. all descriptions in an archival finding aid in EAD are treated as one element. The corresponding *Whole Fonds* (WF) interface is depicted in Figure 2.5.[10]

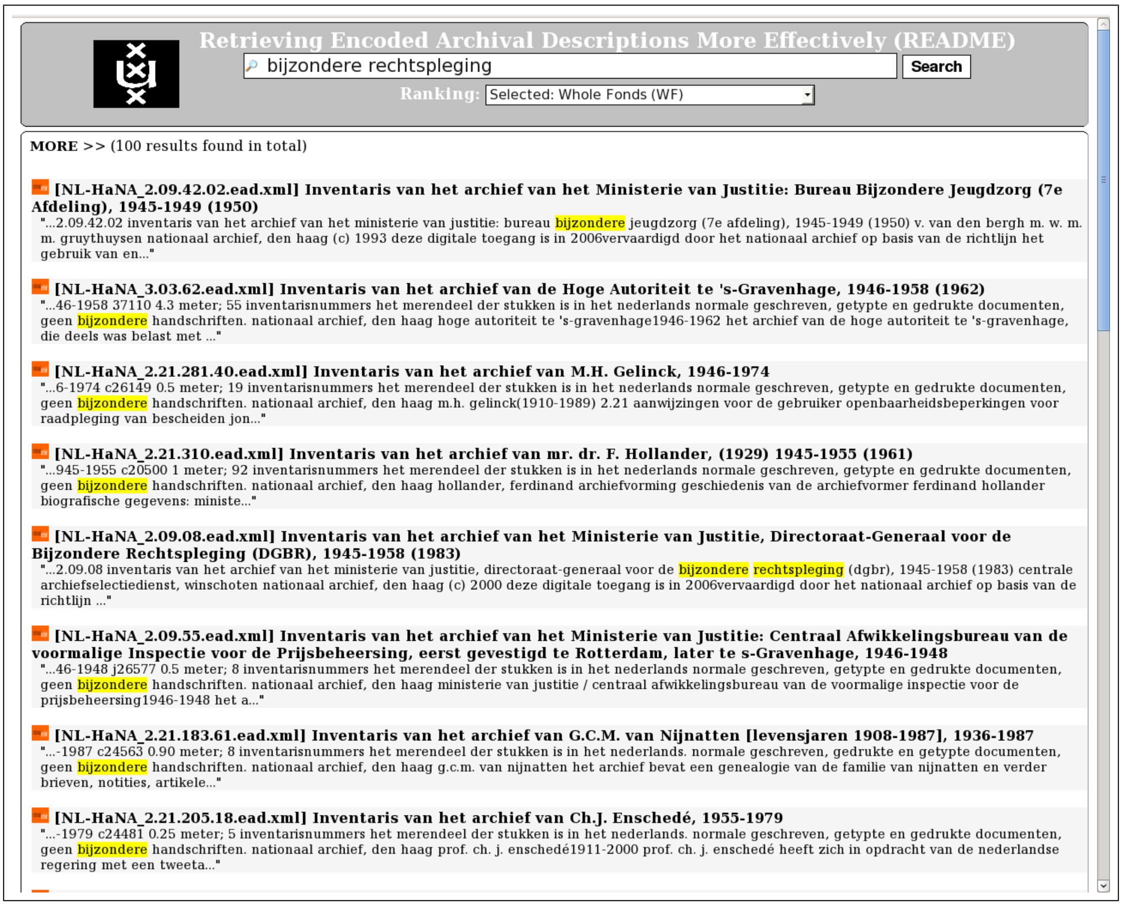


Figure 2.5: Document retrieval with Whole Fonds (WF).

Approach 2: Element Relevance Ranking

For element relevance ranking, see the *Individual Archival Material* (IAM) interface in Figure 2.6, we do not provide a structural hint in the form of a target element, hence any EAD element can be retrieved, including the absolute XPath of an element. Such a path could look like

```
/EAD[1]/ARCHDESC[1]/DSC[2]/C01[4]/C02[8]/DID[1]/HEAD[1]
```

It describes the position of an element in the XML tree hierarchy, starting from the root to the leaf node. The snippet of XQuery code in Figure 2.7 illustrates the procedure element ranking that retrieves N elements stored in `$nodes`.

[10] The term *fonds* is defined by Pearce-Moses (2005) as: "The entire body of records of an organization, family, or individual that have been created and accumulated as the result of an organic process reflecting the functions of the creator." Retrieved 2011/06/24 from `http://www.archivists.org/glossary/term_details.asp?DefinitionKey=756`.

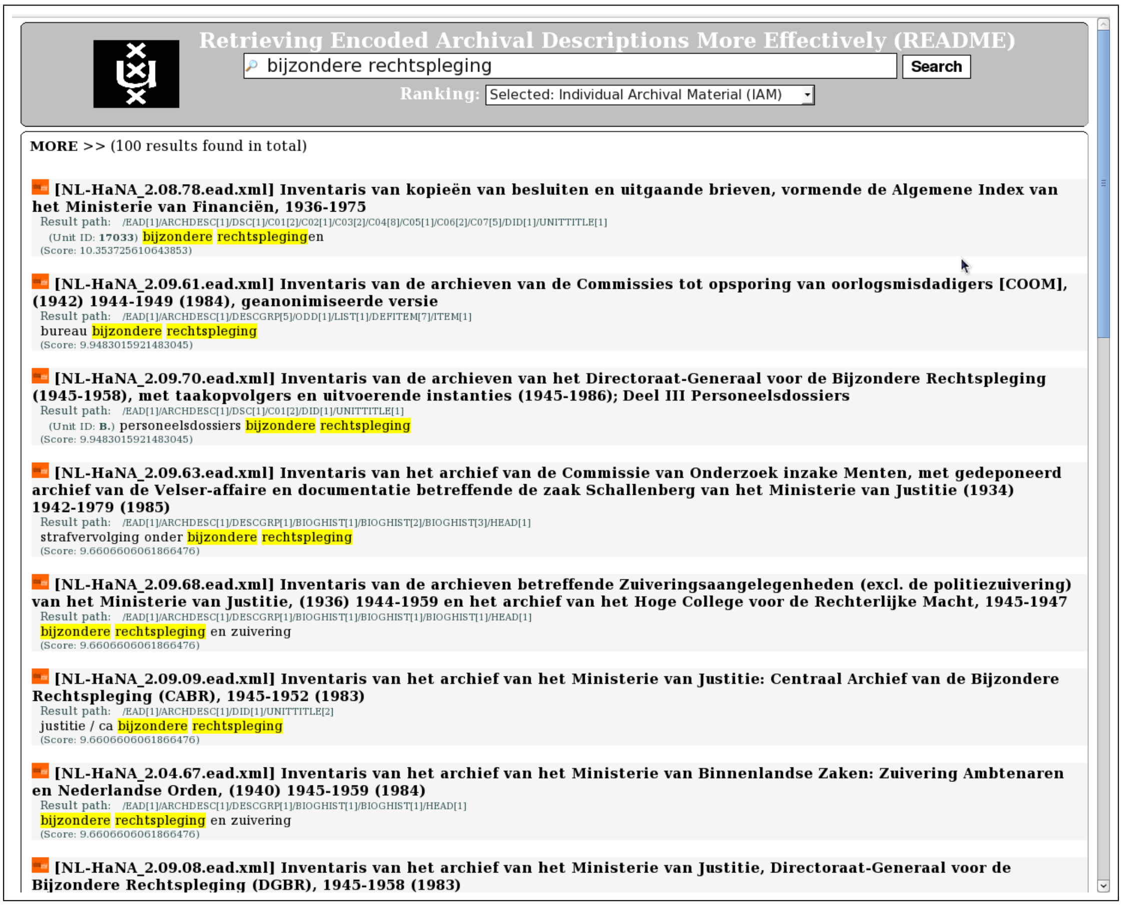


Figure 2.6: Element retrieval with Individual Archival Material (IAM).

```
declare function readme:retrieveAllAMC($query as xs:string, $model as
xs:string, $start as xs:integer, $end as xs:integer) as node()*
  {
     let $options := <TijahOptions ir-model="$model"
            collection-lambda="0.15" rmoverlap="true" prior="no_prior"
            returnNumber="N" />

     let $query_text := tijah:tokenize($query)
     let $query_nexi := concat("//*[about(., ", $query_text, ")]")

     let $qid := tijah:queryall-id($query_nexi, $options)
     let $nodes := tijah:nodes($qid)
```

Figure 2.7: A part of the function readme:retrieveAllAMC. XQuery code that illustrates the initialization of system parameters and the use of NEXI for querying. Here, we search in any nodes using the XPath expression //*, which corresponds to any elements in the document. This piece of code is used for element retrieval.

Approach 3: Aggregation-based Ranking

As Haworth (2001, p.14) notes, archival finding aids have the structure of a tree, and the multi-level description refers to a part-to-whole relationship where de-

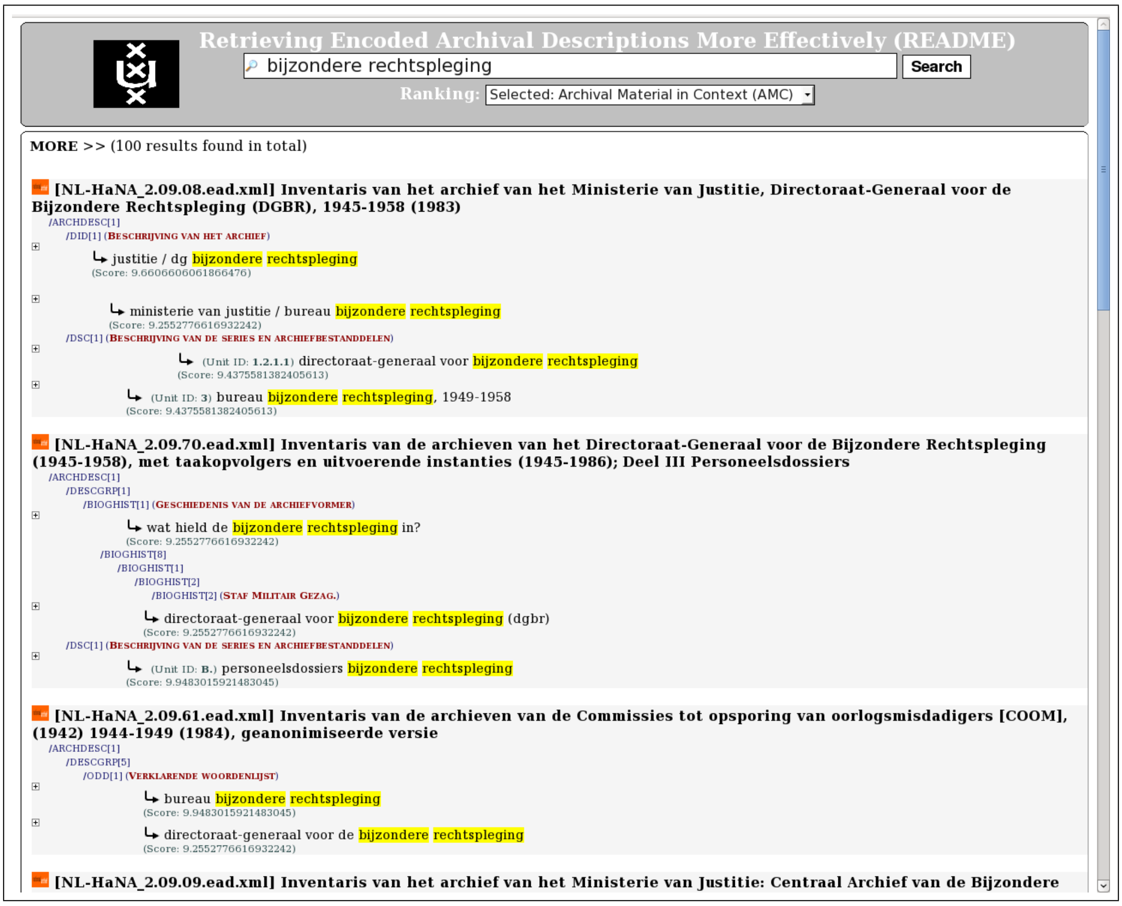


FIGURE 2.8: Aggregation-based retrieval with Archival Materials in Context (AMC).

scriptions are related to each other. The deeper one gets in a finding aid, the shorter the descriptions become. The reason is that according to ISAD(G), information on higher levels cannot be repeated at lower levels, thus descriptions at lower levels are dependent on higher level descriptions. Therefore, descriptions must be displayed together to be comprehensible (Haworth, 2001, p.15).

However, XML retrieval systems, like README, return elements. How do these systems deal with descriptions that inherit the concept of higher level descriptions? How deep should a system go? There is a dilemma for these systems. If the lowest level possible element gets returned, the user may be clueless as there is no information contained in the descriptions of a higher component. If a description of a higher component gets returned, the promise of direct access to relevant information cannot be fulfilled as the user still has the burden to browse the finding aid. Kazai et al. (2004) experiment with the 'overlap' problem, and note that an element may be more exhaustive than any of its descendants alone given that it also contains non-relevant information, while these lower level elements may not contain enough relevant information alone (Kazai et al., 2004, p.73).

```
let $result := for $node at $relevance in $nodes
    return
    <result>
       <rel> { $relevance } </rel>
       <num> { (count($node/preceding::*) + 1) } </num>
       <file> { data($node/ancestor-or-self::EAD/@FILE) } </file>
       <title> { data($node/ancestor-or-self::EAD/@TITLE) } </title>
       <headers> {
          for $node_ancestor in $node/preceding::HEAD[1]
          return
             <heading_before>
                {
                $node_ancestor,
                <headingpath>{readme:getPath($node_ancestor)}</headingpath>
                }
             </heading_before>
       } </headers>
       <path> { readme:getPath($node) } </path>
       <id> {
          for $node_ancestor in $node/preceding-sibling::UNITID[1]
          return <id2> { data($node_ancestor) } </id2>
       } </id>
       <id> {
          for $node_ancestor in $node/preceding-sibling::CONTAINER[1]
          return <id2> { data($node_ancestor) } </id2>
       } </id>
       <text> { $node/normalize-space() } </text>
       <score> { tijah:score($qid, $node) } </score>
    </result>

   let $total := count($result)
   let $tmpRes := <results total="{$total}">  {
      for $result2 in subsequence($result, $start, $end)
          for $res in distinct-values($result2/file)
          let $cs-group := $result2[file = $res]
          let $file := string($cs-group/file)
          let $num := number($cs-group/num)
          order by $file, $num
          return <out id="{$res}">{ $cs-group }</out>
      } </results>
    return
       for $a at $b in $tmpRes
          for $r in distinct-values($a//file)
          let $cs-group2 := $a[//file = $r]
             for $xs at $relevance2 in $cs-group2//out[@id = $r]
             let $score := sum ( for $x in $cs-group2//out[@id = $r]
             return $x//score)
             where $relevance2 <= 8
             order by $score descending
             return <results total="{$total}">  { $xs } <br /> </results>
 };
```

FIGURE 2.9: Part 2 of function readme:retrieveAllAMC. XQuery code that illustrates the retrieval of elements according to relevance, grouping of results by file name, and re-ordering of the results given the original document hierarchy.

To deal with the problem, we introduce an aggregation-based approach. This approach goes a step further than the standard element relevance ranking, so we introduce the *Archival Material in Context* (AMC) interface in Figure 2.8, and see the XQuery code in Figure 2.9. It takes relevance `<rel>` into account.

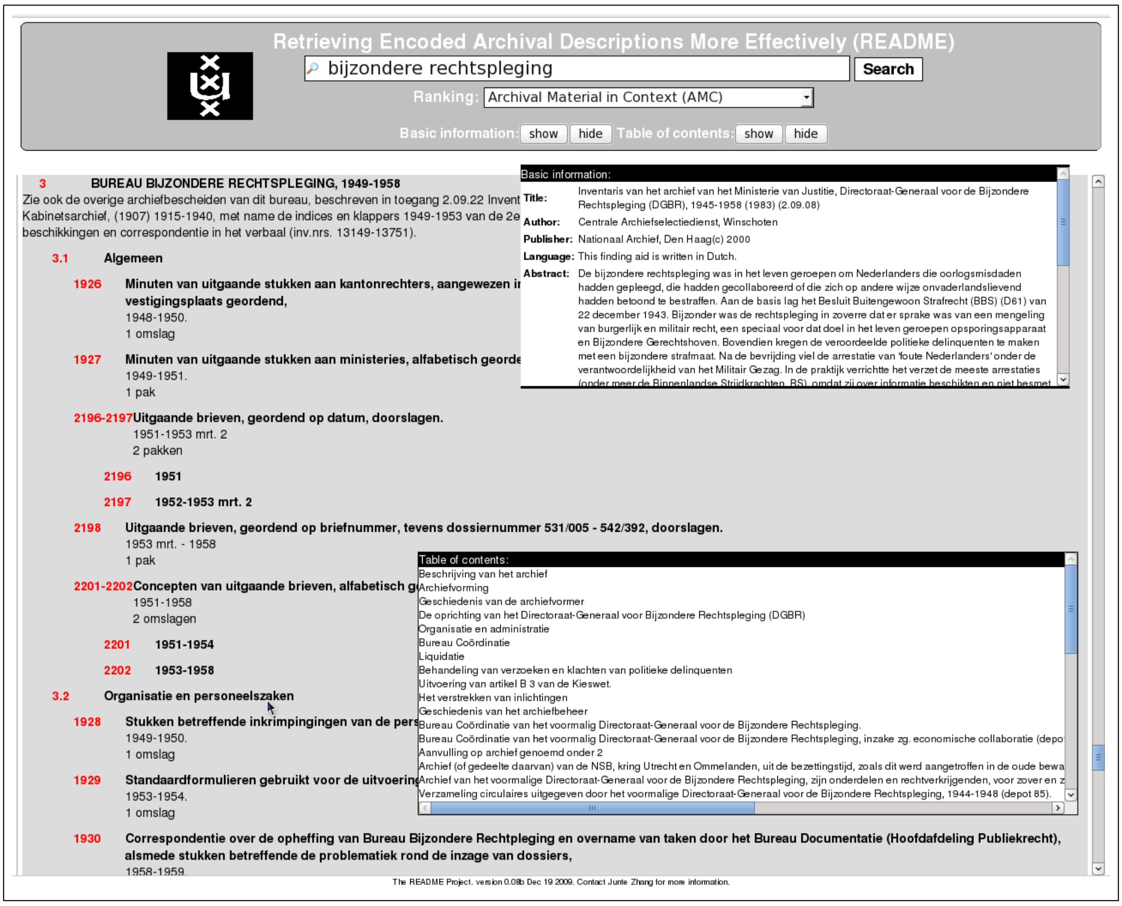


FIGURE 2.10: Deeplinking to the Result Display from the result list with a Basic Information (BI) pane and a Table of Contents. Both the BI and ToC can be dragged, and turned on or off. When a user clicks on an element in the hit list of IAM or AMC, the user gets linked directly to the element in the result display, which the selected element on top in focus.

Any and arbitrary elements can be retrieved. We rank the elements by taking the sum of the top N retrieved element scores e of a file f, so $\text{score}_f = \sum_{e \in f}^{N}$, and rank the files (and its elements) with score_f accordingly, where in our system we set N to 8, but it can be made dynamic by allowing users to move beyond that threshold. The retrieved elements are returned in original order as in the XML file, by computing the distance of the retrieved element to the root node by counting all preceding elements in `<num>`. We group the retrieved elements by its EAD finding aid `<file>`. The aggregation-based approach optimally utilizes the context of the finding aids by aligning the archival principles of provenance and original order of the descriptions within a finding aid (see Chapter 4) to an archival search system using XML IR techniques. It preserves the context to support access.[11]

[11] The aim of respecting the *archival context* is to maintain the quality of information that can

2.4.6 Result Delivery

The hitlist is connected to the Result Display with HTTP parameters using CGI: the query, XPath, source, and file name are always stored in the URL for persistency and to facilitate the analysis of the search logs. The system can deep-link (with the element and aggregation approaches) by rendering HTML anchors for each element using its (unique) XPath as anchor identifier. We deliver a result by physically linking a result to its file, and render its result display with the 'Basic Information' (BI)—e.g. core information of the finding aid, such as the title, creator and summary—and the 'Table of Contents' (ToC) using the SAXON XSLT processor[12]—this is faster than retrieving everything again from the index. There is minimal transformation from the original XML file, because EAD is as much document-centric (directly view-able by users in a browser) as it is data-centric. We use the Yahoo! User Interface Library (YUI)[13] to make the BI and ToC dynamic and enable enhanced interaction, see Figure 2.10. The BI and ToC can be dragged and hidden—making them extra non-obtrusive tools to locate information within the retrieved file.

2.5 Conclusion

After a technical embedding of the background where we explained EAD and XML, we formally introduced our system that employs XML retrieval techniques to gain more focused archival access, effectively exploiting the structure to search and find valuable information in context. A result is the implementation of our vertical (domain-specific) XML IR system driven by archival finding aids: *README*—an online archival information system that is able to retrieve the representation information of the archives by exploiting this representation, specifically the granularity and structure in XML.

Access to archival finding aids in EAD is a two-tier approach. First, the EAD finding aid has to be found. Second, relevant descriptions within the found EAD finding aid have to be returned. We showed that XML IR can be fruitfully applied on archival finding aids in EAD with both tiers. We presented three systems that each provide a different level of detail using the two tiers (from collection to item level) for focused access with archival finding aids in EAD. We also translated archival description principles aimed to preserve the archival

be reused by ensuring its availability, readability, completeness, relevance, representativeness, topicality, authenticity, and reliability (Thomassen, 2001, p.382). Archival research is the study of this context (Thomassen, 2001, p.384). Archival research goes beyond 'information,' and focuses on the interpretation of the 'historical evidence' that consists of these relations (Gilliland-Swetland, 2005, p.236). This rounds off the interrelated argument: when archivists manage their archives by respecting the archival context, they enable their access and use, and make archival research possible, that consists of the study of that context.

[12] Retrieved 2011/01/11 from `http://saxon.sourceforge.net/`.

[13] Retrieved 2011/01/11 from `http://developer.yahoo.com/yui/`.

context, and projected on EAD finding aids, to an information system design with our aggregation-based approach. Fachry et al. (2008) conduct a user study with the systems, and show broad support for this approach.

This dissertation's research focuses on measuring and evaluation access to archival finding aids in EAD, and it does not aim to find out how to design the most usable interface for archival finding aids in EAD, which would be an area too broad to explore. For example, Allen (2005) describes information visualization approaches from the hypertext and visualization research communities for archival access. The retrieved information as expressed in a query should be juxtaposed with proper interfaces. How to make the information and related services *visible* to people with services and functions constitute what access is. The access approach of README is also archival with influence of the archivists who created the EAD finding aids (Hedstrom, 2002), because we aligned archival concepts and principles with the system. We test this influence on access in Chapter 4. More types of system based on README with different ranking algorithms are also possible, but the baseline has been established with the aggregation-based retrieval approach. The question arises: how well does this system work? This question will further be researched in the remaining chapters.

Chapter 3

Construction of an Archival Test Collection for Evaluation

This chapter investigates how to evaluate the IR effectiveness of search systems driven by EAD finding aids like the README system in Chapter 2. We have the system, but we do not know how well it works. We restrict our attention in this chapter to the evaluation of the document ranking approach. In Chapter 4, we also conduct evaluation on the retrieval of descriptions within an EAD finding aid by testing the element relevance ranking and aggregation-based approaches.

3.1 Introduction

Studies can investigate the user needs in advance, such as the study of Yakel (2004), or in conjunction with a system. When the latter is the case, there is also a form of post hoc performance measurement of that system. Fox and Sornil (1999) think that the serious challenge for digital libraries, such as archival information systems, is to have usable systems. Testing the usability of a system is also evaluating a system. An example is the focus group study of Duff and Stoyanova (1998) that investigates the content and user interfaces of archival information systems, and finds suggestions to design more usable displays.

Increasingly more common, the evaluation of systems deals with IR system performance measurements. Robertson (2008) describes early experiments that were formative in the understanding of IR. The Cranfield experiments in the 1950s introduced empirical experimentation to assess the effectiveness of several indexing schemes, and defined the methodology of IR experimentation. The method evolved from known-item search—start from a known document and formulate a question to which that document is a suitable answer—to include judgements of relevance with the aim to find all relevant documents in the collection. The Cranfield experiments measured information retrieval performance with metrics, which later was to become known as *precision* and *recall*. Robertson and Hancock-Beaulieu (1992) also point more in detail to methodological issues, such as laboratory and operational experiments. While the former deals with control over variables, observability, and replicability, the latter deals with realism. Keeping the balance between both methodological approaches is difficult (Robertson and Hancock-Beaulieu, 1992, p.460).[1]

So in IR, the dominant approach to evaluation uses a test collection: a set of documents, a set of search requests (topics), and a set of relevance judgments for each topic (*qrels*). Researchers create such test collections collaboratively on generic (artificial) set of documents, such as newspaper corpora or Wikipedia, which are useful to study generic aspects of the algorithms that drive search systems (i.e. retrieval models). Such test collections provide part of the answers, but fail to address the unique collection and types of search requests of an individual archive. Creating a test collection with this conventional Cranfield approach, for each archive, is simply too expensive in time and effort.

A major challenge for search systems driven by EAD finding aids, is how to evaluate the information retrieval effectiveness given the domain-specifity of their collections, their particular user groups, and how to use this crucial evaluation step to improve a system. We propose a search log-based approach to the IR evaluation of systems driven by EAD finding aids. Nowadays, almost every EAD finding aid system is Web-based, and the interaction between the system and the user is logged in so-called search logs, often hidden deeply away or primarily used to generate descriptive statistics about the general Web traffic of a website. This includes the search requests that have been entered by users, and what and where it was clicked, and so on. Is it reasonable to assume this interaction data can be reused for automated system evaluation? This results in this chapter's research question:

- *Q2: Can we use an online archive's search log to derive a domain-specific test collection?*

We deal with this research question with the following sub-questions:

2.1 How can we derive a test collection from an archival search log?
2.2 How can we validate a log-based test collection?

The clicks of users in the search log may have been greatly influenced by the system that they have used.

2.3 Is there a bias from the original system when using the log-based test collection?

The envisioned test collection is tailored to the archive at hand, representative to both its document collection and its search requests. As a test collection,

[1] Tague-Sutcliffe (1992) also points to methodological issues for carrying out an IR experiment, such as determining the need for testing, choosing the type of test (laboratory or operational), defining the variables, collecting and analyzing the data, and presenting the results.

it can be (re)used for comparative testing under the same experimental conditions. Performance is topic-dependent and this avoids comparing over different topic sets. We apply this approach to a particular domain-specific collection of documents, in a special genre, namely archival finding aids in EAD that describe the archives of persons, families, and corporations.

The remainder of this chapter is structured as follows. Section 3.2 describes the related work. We explain the interaction occurring in the *Nationaal Archief* EAD system and subsequently derive a domain-specific test collection from this search log in Section 3.3, using a range of retrieval models for further testing in Section 3.4. In order to validate the log-based evaluation, we construct a set of topics judged by human experts—based on a set of e-mail reference questions sent to the *Nationaal Archief* and responses from their archivists. The results are analyzed and discussed in Section 3.5. Moreover, we derive a test collection from a research guide compiled by the *Nationaal Archief* and repeat the IR experimentation in Subsection 3.5.4. Finally, we conclude with our main findings and discuss pointers to future research in Section 3.6.

3.2 Background and Related Work

In this section we discuss three strands of related work: search log analysis for IR evaluation, IR test collections, and archival (metadata) retrieval evaluation.

3.2.1 Log Analysis for Information Retrieval Evaluation

Historically, the analysis of log files started and "evolved out of the desire to monitor the performance of computerized IR systems" (Peters, 1993, p.42). The focus has been to analyze how systems are used. Besides system monitoring, it can also be conceptualized as a way to unobtrusively observe human behaviors. Studies in a digital library setting have been reported in (Jones et al., 2000), which focused particularly on the queries that users entered in the system, with the proposition that the analysis can be used to finetune a system for a specific target group of users, but it did not specifically investigate the IR effectiveness.

Research on log analysis in library and information science precedes the research in the World Wide Web, where the latter zooms into IR by analyzing search engines. An overview on search log analysis for Web searching, and a methodology, is presented in (Jansen, 2006), which shows that literature on log analysis for Web-searching is abundant. The logs can be analyzed to better understand how users search on the Web effectively (also see Chapters 5 and 6 of this dissertation for EAD finding aids). An example is the paper of White and Morris (2007), which describes a study about search logs, where the search behavior of advanced and non-advanced users is analyzed by testing the effects of query syntax with query operators on query-click behavior, browsing behavior, and search success.

There has been substantial interest in using click-through data from search logs as a form of implicit feedback (Dumais et al., 2003). A range of implicit feedback techniques have been used for query expansion and user profiling in information retrieval tasks (Kelly and Teevan, 2003; Kim et al., 2001). Joachims et al. (2005, p.160) conclude that "the implicit feedback generated from clicks shows reasonable agreement with the explicit judgments of the pages." There is active research on building formal models of interaction from logs to infer document relevance (Dupret and Liao, 2010).

3.2.2 IR Test Collections

IR evaluation can be traced back to the work-process of a librarian working with card indexes using library classification schemes (Robertson, 2008). The basic methodology for IR experimentation has been developed in the 1950s with the Cranfield experiments, focusing on retrieval effectiveness by the comparative evaluation of different systems (then indexing languages, nowadays retrieval algorithms). Much of the experimentation focuses on building a test collection consisting of a document collection, a set of topics and judgments on which documents are relevant for each topic (Spärck Jones and van Rijsbergen, 1976).

Test collections can be reused by evaluating new or adapted systems or ranking algorithms under the exact same experimental conditions. A notable collaborative ('pooling') approach is the Text REtrieval Conference (Voorhees and Harman, 2005). There are test collections for a variety of domains. Examples included the Cystic Fibrosis database in the medical domain (Shaw et al., 1991), and WT10g for the Web in general (Bailey et al., 2003). In the field of XML retrieval, the Initiative for the Evaluation of XML Retrieval (INEX) constructed test collections based on XML files (Gövert and Kazai, 2002b; Lalmas, 2010). Creating a traditional test collection requires substantial (collaborative) human effort in topic creation and relevance judgments (domain knowledge), and such a test collection does not exist yet in the archival domain.

3.2.3 Evaluating Archival Metadata Retrieval

Published research that empirically or experimentally deals with the evaluation of archival metadata retrieval is scant (Hutchinson, 1997). Experiments that specifically examine the retrieval performance potential of archival finding aids in specifically EAD is almost non-existent, despite the emergence of EAD in 1997 (Pitti, 1999) and its increasing adoption and popular use in archives.

The first study in the archival field that empirically tested different subject retrieval methods was Lytle (1980). Subsequently, there were a few studies that tested the effects of some external context knowledge on retrieval, such as controlled vocabulary terms (Ribeiro, 1996) or document-collection granularity (Hutchinson, 1997). Hutchinson (1997) outlines and tests several strategies for searching in EAD finding aids, namely as a whole unit, only the introductory components of a finding aid before the actual inventory descriptions, introductory components enhanced with controlled access terms extracted from the catalogue records, and catalogue records only. The retrieval of online archival finding aids (not in EAD) have been examined in the study of Feeney (1999) by counting the number of finding aids returned by search engines using different types of query formulations, i.e. keyword, phrase, and Boolean searches using the topical subject and names headings as queries. The retrieval experiments of Tibbo and Meho (2001) on finding aids as full text HTML documents on the World Wide Web pointed to the effectiveness of phrases for the retrieval of finding aids (not in EAD). Moving away from the system-centered view, Daniels and Yakel (2010) present a user-centric experiment, where they observe what users do when they search, how they search, and how often their recall-oriented search was succesful.

3.3 From Search Log to Test Collection

In this section, we study how to derive a test collection from a search log. We look at the search interaction possible with the system of the *Nationaal Archief* and use its search log to create domain-specific test collections for IR evaluation, tailored to their EAD finding aids collection and its users.

3.3.1 *Nationaal Archief* System Interaction

The flow charts in Figure 3.1, 3.2, and 3.3 depict the work-flow of the *Nationaal Archief* system as of 2009, and we have numbered each step possible during that work-flow. This work-flow consists of starting the Web site's frontpage **(1)**, then moving to search in **(2)** by formulating a query (in this example query *Juliana*, the name of a former Dutch queen, 1909-2004) and entering it in the search box. The NA system returns found results in **(3)** from different sources, such as EAD finding aids, the image repository, several thematic databases, and remaining resources. We zoom in on the list of EAD finding aids in **(4)**. From this point, the actual search interaction starts. Figure 3.2 on page 39 and Figure 3.3 on page 40 show divergent search paths stemming from **(4)**.

1. The first search path as illustrated in Figure 3.2 is an example of full-text (or document) search, where the returned results in **(5a)** consist of links to the archive overview of an EAD finding aid in **(6a)**, and then to the beginning of an EAD finding aid in the interface of **(7a)**. The user has to start to use the table of contents on the left side of the screen to browse through the whole EAD finding aid.
2. The second search path in Figure 3.3 is an example of focused (or sub-document) search, where a user can directly go to an archival description of a component in an EAD finding aid by clicking on it in **(5b)**, and then views the

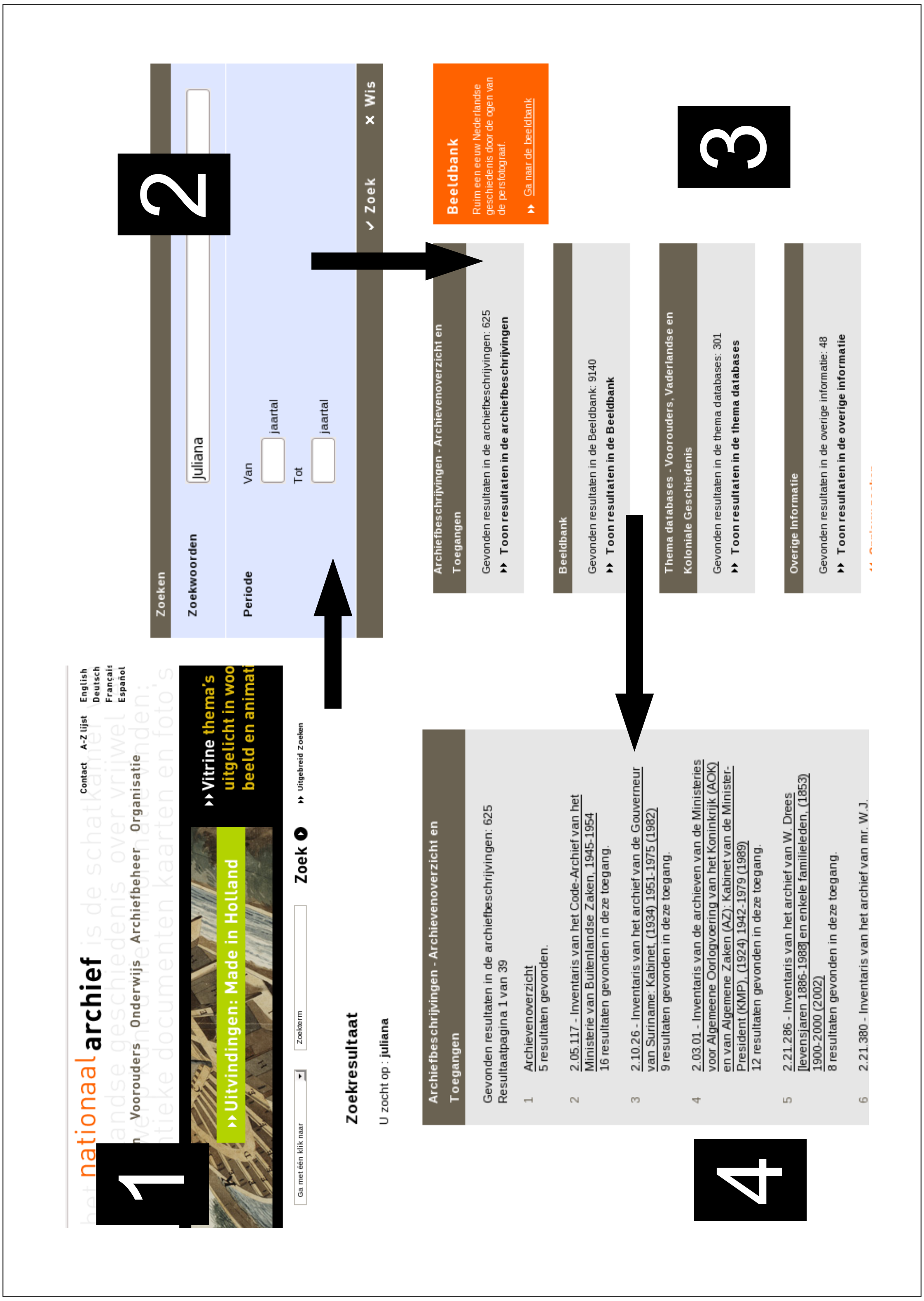


Figure 3.1: *Nationaal Archief* System Interaction (1). Image courtesy of the *Nationaal Archief* (rotated clockwise).

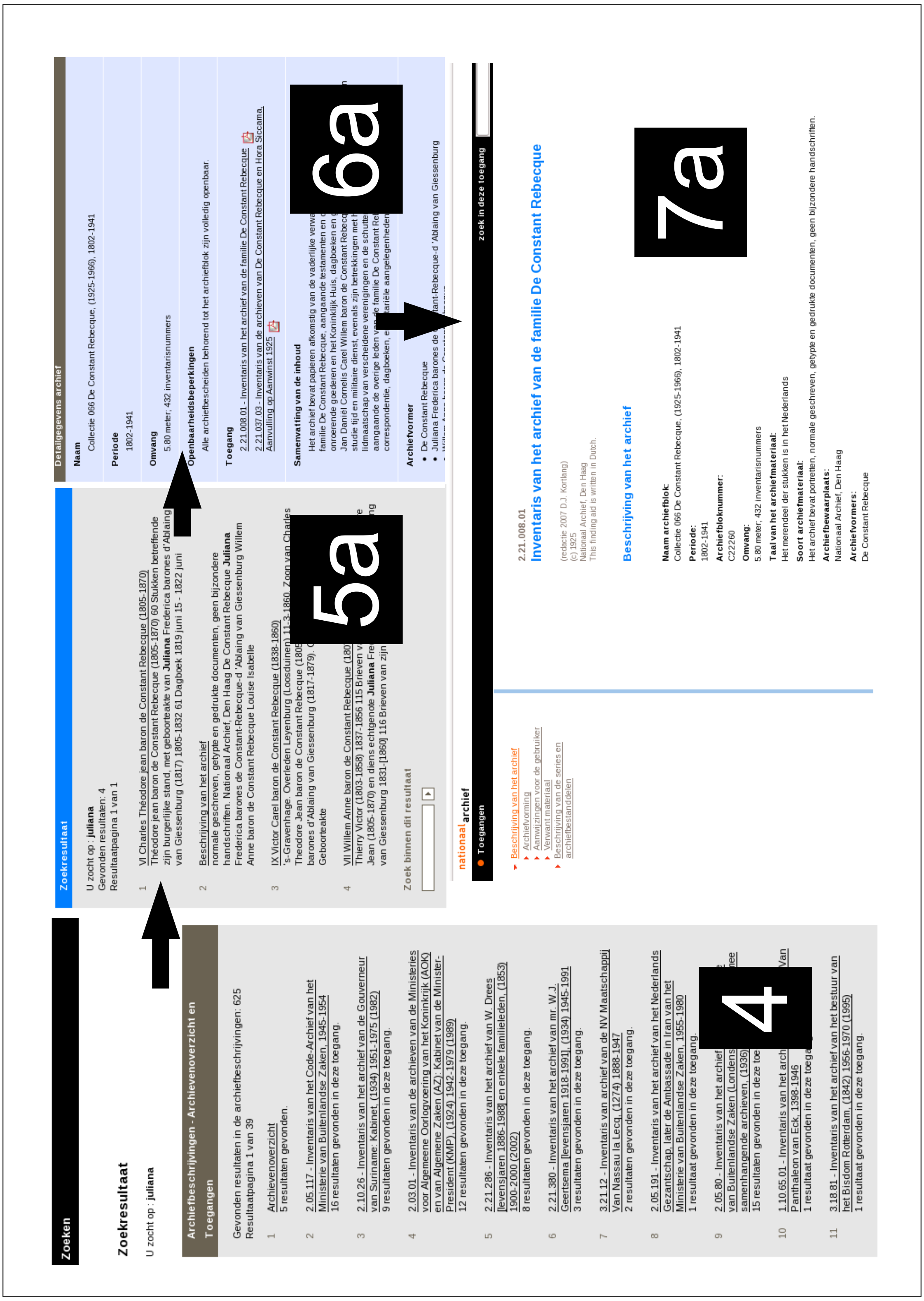


FIGURE 3.2: *Nationaal Archief* System Interaction (2). Image courtesy of the *Nationaal Archief* (rotated clockwise).

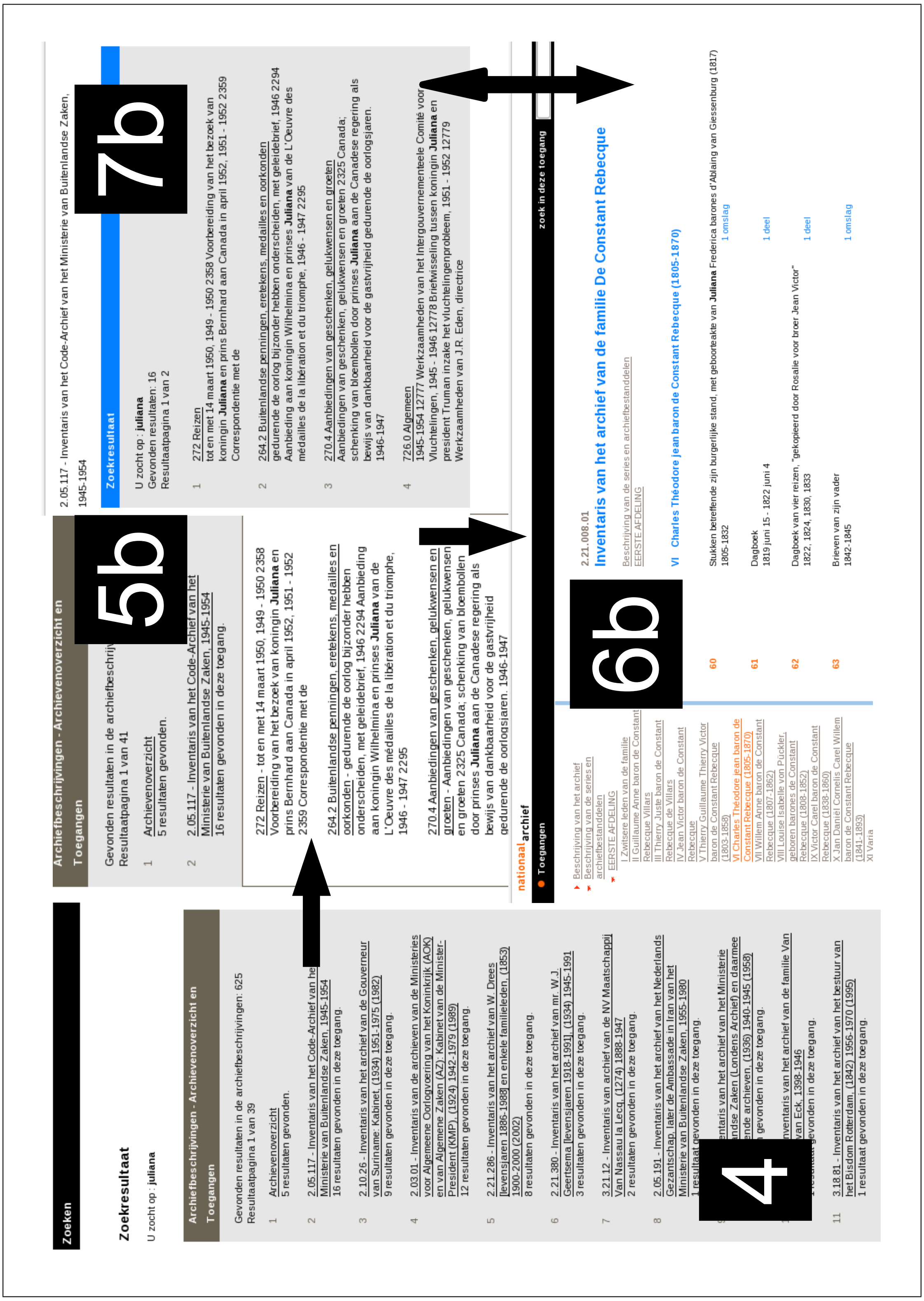


Figure 3.3: *Nationaal Archief* System Interaction (3). Image courtesy of the *Nationaal Archief* (rotated clockwise).

interface that zooms into the clicked archival component in **(6b)**. Further search in the EAD finding aid is possible by entering a query in the search box on the top-right side of the screen, which directs the user to the list of archival descriptions of components in **(7b)**, and eventually back to the interface of archival descriptions in **(6b)**.

This search interaction has been stored in the NA log files. From **(4)**, we can trace each and every click on an EAD finding aid back to a query that has been entered in the search box in **(2)**. This resembles a user request with desired answers to that request.

3.3.2 Search Log Files and Document Collection

We use the search logs of the *Nationaal Archief*. The history preserved at this institution goes back to more than 1,000 of years, with materials preserved in archives which stretch more than 93 kilometers or 58 miles. It also includes maps, drawings, and photos—much of it is published on the *Nationaal Archief* website (`www.nationaalarchief.nl`). The website provides access by offering a search engine, which includes searching in archival finding aids compiled in EAD, image repositories, and separate topic-specific databases.

The NA logs are 91.1 GB in size, with 39,818,981 unique IP-addresses, and collected from 2004 to a part of 2009 on a Microsoft IIS server. This illustrates that the *Nationaal Archief* attracts high traffic. The information contained in the search logs were recorded from 2004-2006 in the *Common Logfile Format* (CLF), and from 2007 to 2009 in the *W3C Extended Logfile Format* (ELF). The information in the CLF format includes a date, a time-stamp of a hit, unique identifier representing a user, the URL of the link that was visited, the query string, and a browser identifier. In the ELF format, it also includes a referral, and transactions have been recorded in detail within each second.

In our experiments, we only focus on click-through data of user interaction with EAD finding aids, where each click contains the filename of an EAD finding aid, which can be traced back to a corresponding query. The reason for this selection is that we have also obtained the corresponding EAD finding aids—also supplied by the *Nationaal Archief* —for analysis and further experimentation. Each EAD finding aid encodes an archival description that describes the content and context of archival material. We use 5,934 EAD finding aids in XML—937 MB of data—which are mostly written in the Dutch language.

3.3.3 Information Extraction from Logs

An archival search log contains both searching and browsing behavior, with complete sessions starting from an initial query. Given the massive size of the log, we pre-processed it by extracting the click-through data that consist of the subset of clicks to EAD URLs. The query string, clicked URL, and the IDs of

Table 3.1: Example of information in sessions extracted from the log.

Query (Topic)	File	Session ID	#
burgerlijke stand suriname	1.05.11.16	504d2bbe246d877bda09856ecc300612.5	28
burgerlijke stand suriname	1.05.11.16	212de7cab1c3709be3a95ac1a37a7873.1	6
burgerlijke stand suriname	3.223.06	22fe3a65b0c9223280f2dd576c57a012.35	1
burgerlijke stand suriname	1.05.11.16	2b844140ef7cfd438300da7ec6278de0.147	1
burgerlijke stand suriname	2.05.65.01	3784a93938e29a6aef8f50baa845a6f3.1	1
burgerlijke stand suriname	1.05.11.16	8b21ec51722f3a52cfaf35d320dfacb0.3	1
burgerlijke stand suriname	1.05.11.16	212de7cab1c3709be3a95ac1a37a7873.2	1
burgerlijke stand suriname	1.05.11.16	9235756a6dbdcffba9179d75108cd220.433	1
burgerlijke stand suriname	3.231.07	3c34072bef0d505467ca9394c392888d.2	1

the user are extracted. It is further processed by aggregating the clicks for each query in a session and keeping track of the count. We define a session as a subset of n clicks from the same IP address, if and only if the difference between i and $i + 1 < 30$ minutes (or 1,800 seconds), where i is a click. This results eventually in 194,138 sessions. Table 3.1 presents the extracted interaction data on an aggregated level. This is used to derive a test collection.

When we focus on Table 3.1, we notice that for query "burgerlijke stand suriname" (in English, "registry of births, deaths and marriages suriname") clicks exist in 9 different sessions, coming from 8 different IPs. There were 28 clicks in one session for EAD finding aid "1.05.11.16," and the same file was clicked in total 6 different sessions. The same file was re-clicked from an IP address in the next session. So "1.05.11.16" could be regarded as "relevant."

Although we regard here "clicked pages" as pseudo-relevant, we make no particular claims on the interpretation of clicks. We make the reasonable assumptions that searchers found these pages of sufficient interest—for whatever reason—to consult them more closely, and that a more effective ranking algorithm will tend to rank such pages higher than those that do not receive clicks. In this chapter, we are interested in the potential of log-based evaluation, and a relatively naive click model is sufficient for that purpose.

3.3.4 Types of Test Collections

A subset of the search log files is used, namely the clicks on archival finding aids in EAD, which is rapidly growing in use for online archival access. We notice that the usage of EAD started to take off in 2006 (19.9MB out of 9.8GB; 0.20%), and this trend in popularity was upward, as it also increased in 2007 (1.5GB out of 31.5GB; 4.8%), and in 2008 (2.8GB out of 41.2GB; 6.8%), and a part of 2009 (304.4MB out of 3.8GB; 7.8%). Hence, the Web traffic of the *Nationaal Archief* is increasingly consisting of the use of EAD, although the amount of EADs published online has increased as well.

We extract from the search log files in total 50,424 unique topics (after string processing, i.e. squeezing white spaces, conversion to lowercase, removal of punctuation), which have been created by 110,805 unique IP-addresses. There are in total 441,575 clicks with 91,009 unique topic-click pairs. Since the collection consists of 5,934 EAD files, numerous topics matched with these files. The queries have a long-tail distribution, where the majority of the topics consist of unique queries that occurred just once. This is also typical in the archival domain, for example genealogists looking for (unique) family names, and this was also the case in the *Nationaal Archief* logs.

We derive different test collections from the logs. We use clicks on the file-level in order to evaluate full-text retrieval. The two types of test collections used:

Complete Log Test Collection The set of 50,424 unique topics, and their corresponding clicks to EAD files, where each and any click is treated as a pseudo-relevance judgment.

Test Collections Based on Agreement Subsets of the search log filtered by the agreement reached among multiple users when they clicked on the same EAD finding aids for a given topic. For example, for agreement 2, we only retain documents clicked by at least two users, which scales down the test collection to 2,455 topics.

We test the two types of test collections separately in the next section.

3.4 Log-Based Evaluation in Action

In this section, we use the log-based test collections to determine the retrieval effectiveness of different ranking methods. We look at both the complete log, as well as on smaller subsets based on agreement. Recall that test collections are used for the comparative evaluation of systems or ranking algorithms, hence we need a number of variant systems in order to show their retrieval effectiveness.

3.4.1 IR Models and Systems

We use the README system framework, which has been described in the previous chapter. To test the effectiveness of the two types of test collections, we regard the sets of queries of the test collections and the retrieval system as independent variables, with the retrieval effectiveness as dependent variable.

We measure the retrieval effectiveness using three IR measures, namely Mean Average Precision (MAP), which is the most frequently used summary measure for a set of ranked results, Mean Recipropal Rank (MRR), and Normalized Discounted Cumulative Gain (nDCG). The MRR is a static measure that looks at the rank of the first relevant result for each topic, and the nDCG measure that uses the number of clicks on each result as a form of graded relevance judgment.

The systems use the default parameter values, the collection λ is set to 0.15—and we set the threshold of returned results for each topic to 100.

BOOL is the Boolean model, where there is no ranking, but a batch retrieval of exact matching results (e.g. Belkin and Croft, 1992). The query is interpreted as AND over all query terms, and the resulting set is ordered by document id.

LM is standard language modeling without smoothing, all terms in the query need to appear in the result (e.g. Ponte and Croft, 1998).

$$P(q|d) = \prod_{t \in q} P(t|d)^{n(t,q)} \tag{3.1}$$

where $n(t,q)$ is the number of times term t is present in query q.

LMS is an extension of the first model by applying smoothing, so that results are also retrieved when at least one of the keywords in the query appears (e.g. Chen and Goodman, 1996).

$$P(t|d) = (1-\lambda) \cdot P_{mle}(t|d) + \lambda \cdot P_{mle}(t|C) \tag{3.2}$$

where $P_{mle}(t|C) = \frac{df_t}{\sum_t df_t}$, df_t is the document frequency of query term t in the collection C.

NLLR is the NLLR or length-normalized logarithmic likelihood ratio, is also based on a language modeling approach. It normalizes the query and produces scores independent of the length of a query (e.g. Kraaij, 2004).

$$NLLR(d,q) = \sum_{t \in q} P(t|q) \cdot \log\left(\frac{(1-\lambda) \cdot P(t|d) + \lambda \cdot P(t|C)}{\lambda \cdot P(t|C)}\right) \tag{3.3}$$

OKAPI is Okapi BM25, which incorporates several more scoring functions to compute a ranking, such as also the document length as evidence (e.g. Robertson et al., 1994).

$$BM25(d,q) = \sum_{t \in q} IDF(t) \cdot \left(\frac{f(t,d) \cdot (k_1+1)}{f(t,d) + k_1 \cdot (1 - b + b \cdot \frac{|d|}{avgdl})}\right), \tag{3.4}$$

where we set $k_1 = 2.0$ and $b = 0.25$. We use $IDF(t) = \log \frac{N - n(t) + 0.5}{n(t) + 0.5}$, where N is the total number of documents in the collection, and $n(t)$ is the function that counts the number of documents that contains query term t.

3.4.2 Study 1: Complete Log Test Collection

When we use all topics for evaluation, and look at all measures, we see in Table 3.2 that *BOOL* is obviously the worst performing system. We note that the

Table 3.2: System-ranking of runs over all topics.

	BOOL	LM	LMS	NLLR	OKAPI
MAP	0.1808 (5)	0.2493 (4)	0.2548 (3)	0.2591 (2)	0.2631 (1)
MRR	0.2015 (5)	0.2940 (4)	0.2980 (3)	0.3024 (2)	0.3077 (1)
nDCG	0.2659 (5)	0.3289 (4)	0.3547 (3)	0.3605 (2)	0.3652 (1)

differences among the other systems are modest, but these differences are all significant (1-tailed) using the paired-samples t-test on a 1% significance level.

When looking at the recall over all topics, we see that the BOOL and LM systems retrieve 48,096 relevant results out of 87,057 (55.25%), the LMS system returns 57,935 of 89,906 (64.44%), the NLLR system has a recall of 65.18%, and the OKAPI system returns most relevant results with 65.69%. It shows that the system using the OKAPI model performs best with our document collection, and that exact matching using the BOOL and LM systems both do not pay off for the early rank (MRR), and as expected hurts the recall. The recall values can be clarified by the long-tail distribution of query terms, which contains many non-occurring names.

In summary, the BOOL system is the worst-performing system, then the LM system, with the LMS system improving over the LM system, and that the differences among the LMS, NLLR, and OKAPI systems are modest (but significant). We will validate the relative system ranking against a set of humanly judged topics in the next section, but first we will look at the system ranking induced by smaller subsets of topics based on agreement.

3.4.3 Study 2: Test Collection Based on Agreements

The search log-based test collection has considerably more topics than a traditional test collection with 25-200 topics. While having thousands of topics opens up new uses, such as focusing on various breakdowns of the topic set even on relatively rare phenomena, it also presents an efficiency challenge: many online archival search systems may crumble under thousands of queries, preventing fast and efficient evaluation. Therefore, we take into account the agreement that exists among different searchers. For example, when we pay attention to Table 3.1, this means that only EAD file "1.05.11.16" is included as a relevance judgement in the test collection.

We see in Table 3.3 that as we increase the threshold of agreement, the number of topics decreases significantly. Take for example notice that in the case that if the agreement is set to 2, the topic set size decreases to 2,455 from 50,424, and when we set the threshold to 4, only 533 topics are left over. What does this mean for evaluating a system with such a set size? The results of this experiments are presented in Table 3.4, where we focus on the MAP scores.

Table 3.3: Distribution (in percentages) of N topics over query length for all topics compared to when filtering on agreements, and e-mail references.

# Tokens	All	Agree 2	Agree 3	Agree 4	E-mail
1	37.66	76.82	82.38	81.99	17.81
2	33.43	15.76	12.02	12.76	19.18
3	16.97	5.87	4.87	4.69	30.14
4	6.68	1.14	0.73	0.56	19.18
N	50,424	2,455	965	533	73

Table 3.4: System-ranking of runs over topics with agreement.

	Agree	BOOL	LM	LMS	NLLR	OKAPI
MAP	2	0.1522 (5)	0.3605 (4)	0.3620 (3)	0.3629 (2)	0.3751 (1)
	3	0.1120 (5)	0.3891 (3)	0.3888 (4)	0.3894 (2)	0.3991 (1)
	4	0.1071 (5)	0.3637 (4)	0.3639 (3)	0.3641 (2)	0.3793 (1)
MRR	2	0.1629 (5)	0.4020 (4)	0.4030 (3)	0.4039 (2)	0.4157 (1)
	3	0.1188 (5)	0.4253 (2)	0.4247 (4)	0.4253 (2)	0.4356 (1)
	4	0.1132 (5)	0.3943 (3)	0.3942 (4)	0.3945 (2)	0.4110 (1)
nDCG	2	0.2734 (5)	0.4521 (4)	0.4564 (3)	0.4578 (2)	0.4726 (1)
	3	0.2384 (5)	0.4750 (4)	0.4767 (3)	0.4778 (2)	0.4913 (1)
	4	0.2315 (5)	0.4520 (4)	0.4552 (3)	0.4560 (2)	0.4735 (1)

We focus on the differences of the MAP scores when we take an agreement between two different IPs. The BOOL system is significantly performing worst, and the OKAPI system is performing the best compared to either the LMS system with a significant improvement of 3.62% (t(4835) = 5.50, $p < 0.01$, one-tailed), or similarly an 3.36% significant improvement over the NLLR. Although the improvement between the LM and LMS systems was only 0.42%, it was also significant (t(4835) = 3.76, $p < 0.01$, one-tailed). This is completely in line with our findings when using the full set of topics.

What happens when we take an agreement of a click among 3 different IPs? Again we focus on the MAP scores. We see that the BOOL system is again significantly the worst performing system, and the OKAPI is significantly performing better on a 1% significance level. The LM and LMS systems also swap from position. However, we did not see significant differences among the other three models. This is partly clarified due to the distribution of the number of keywords in a query as presented in Table 3.3, because the effect of smoothing is leveled out when the topics become in majority singleton queries. As Table 3.3 shows, when we increase the agreement threshold, there are only 965 queries are left, which are predominantly very short (limiting the impact of smoothing) and many of them having having only a single relevant document.

Finally, what happens when we take an agreement among 4 different IPs?

We still see the same pattern as the previous runs, with the BOOL being the worst, and the OKAPI system the best (+4.17%, $t(1300) = 3.21$, $p < 0.01$, one-tailed) compared to the runner up system NLLR. The findings are also consistent with the nDCG scores, though we see that with MRR, the LM and LMS systems swap position with a minor non-significant difference.

In summary, there are two implications. First, deriving a test collection using agreement of 2 seems a viable alternative for using the whole log file. Second, the system rankings are similar when treating the clicks as binary pseudo-relevance judgements (MAP, MRR) and as graded judgements (nDCG).

3.5 External Validation

We investigate the validity of the log-based test collection in terms of the resulting system ranking. As ground-truth, we use a test collection constructed by human experts: responses of archivists to e-mail reference questions. We assume that the 'manual approach' of archivists providing answers to the questions of users are of high quality, but these are difficult to obtain, hence the comparison with the log-based approach. The system rankings of the search log-based test collection are compared to this ground-truth. Additionally, we compare the search log-based test collections with the original ranking of the *Nationaal Archief* system. This is the original system used by users in the log-based approach, and archivists in the e-mail references.

3.5.1 Test Collection Based on E-Mail Reference Requests

We analyze the e-mails that the *Nationaal Archief* received from users, and with replies from archivists that referred explicitly to EAD files. We look at all correspondence (4.1GB of data). The e-mails are converted from *PST* file format to *mbox* format, which comes in readable textual form. Eventually, we manually select 73 different topics (and recommended EAD links) from the e-mail files that have a clear information need expressed by the user and a clear recommendation for at least one EAD finding aid given by an archivist.

A typical example is the information request in Figure 3.4. The explanation of the information request is included in `<narrative>`, the topic in `<title>`, and the relevant files for that topic in `<file>`. We select typical replies from an archivist who linked to EAD files using the query, or recommended the EAD finding aids which are relevant.

3.5.2 System Rank Correlations

First, the recall values of the e-mail runs. We find for the BOOL and LM systems that there are 61 topics with relevant results, where 46 out of 90 (51.1%) EADs have been found. For the LMS and NLLR systems, there are 63 valid topics that return 66 relevant EADs out of 92 (71.7%). For the OKAPI system, there are 63

```
<topics>
  <topic nr="1">
    <title> wateringen </title>
    <narrative> I. Purpose: Query for Commissioned Research Support II. Topic: Historical Water and Flood Management in Holland III. Geographic Location: My research focuses on two locations: 1. Along the river channels, uiterwaarden, and polders of the Nederrijn and Lek Rivers; and 2. along the Oude Rhine (region), between about Woerden and Katwijk. This includes the “waterschaapen" of Rhineland and Delftland. IV. Type of materials: maps, surveys, and tables of data that pertain to land use (agriculture types), water/flood management (dijk, sluice, drainage ditches, pumping, etc...) I am especially interested in archival materials that relate to the early period (before 1300), but including up to about 1925.</narrative>
    <files>3.01.04.01; 4.VTH; 4.KIVI; 4.ZHPB4; 2.16.91; 2.16.06</files>
  </topic>
  <topic nr="2">
    <title>ministerie van verkeer en waterstaat ptt draadloos telegrafie</title>
    <narrative>Do you know whether the “Ministerie van Verkeer en Waterstaat" has an own archive which includes PTT documents relating to wireless telegraphy?</narrative>
    <files>2.16.09; 2.16.93</files>
  </topic>
  <topic nr="34">
    <title>Frans Beelaerts Blokland Peking Beijing</title>
    <narrative>I am writing a book about foreigners in Beijing from the Boxer Rising in 1900 to the Communist takeover 1949. Jonkheer Frans Beelaerts van Blokland was the Dutch Minister in Peking during the World War One. I am very interested in seeing any papers that you may hold relating to his years in Peking.</narrative>
    <files>2.05.90 ; 2.05.19 ; 2.21.253</files>
  </topic>
</topics>
```

FIGURE 3.4: Examples of three topics based on e-mail reference requests, where the <title> is used as query, and if necessary, translated to Dutch.

topics with relevant results, i.e. 67 out of 92 EADs (72.8%). We note that there are 73 topics, but the systems could not find results for all topics.

The results of Table 3.5 show that the BOOL system performs worst, and the OKAPI system is performing best. We again use the paired-samples t-test to check for significance. We look at the MAP scores. When we rank with the LM system, there is also a significant improvement of 65.49% over the BOOL system ($t(60) = 3.03$, $p < 0.05$, one-tailed). When we use the LMS system, we see a 7.45% significant improvement over LM without smoothing ($t(62) = 2.20$, $p < 0.05$, one-tailed). However, the difference between the LMS and NLLR models was only 0.47%, and is not significant. Moreover, the OKAPI system performed 8.77% better than *NLLR*, but it is not significantly better. The findings are similar using the MRR and nDCG measures, though we do not see a difference between the LMS and NLLR systems.

How reliable are our log-based test collections when compared to the test collection semi-automatically derived from the e-mails and their experts' replies? When we compare the system rankings of the test collection from the whole

Table 3.5: System-ranking of runs over e-mail topics.

	BOOL	LM	LMS	NLLR	OKAPI
MAP	0.1678 (5)	0.2777 (4)	0.2984 (3)	0.2998 (2)	0.3261 (1)
MRR	0.1896 (5)	0.3197 (4)	0.3313 (2)	0.3313 (2)	0.3589 (1)
nDCG	0.2646 (5)	0.3624 (4)	0.4167 (2)	0.4176 (2)	0.4361 (1)

log (Table 3.2) with the e-mail topics (Table 3.5) using the MAP scores, we see a complete agreement with a Kendall's Tau value of 1. Overall, we see full agreement between the log-based evaluation and the reference requests, and among the test collections of the log-based evaluation approach.

The results show that the *OKAPI* performs best for EAD finding aids written in the Dutch language. It confirms earlier findings by Savoy (2003) who showed that *OKAPI* also works best for information retrieval with Dutch texts.

These results are promising, but we further research the rank correlations between the system rankings of both approaches by adding the *Nationaal Archief* system itself in the next experiments. For the e-mail topics, archivists have been using the *Nationaal Archief* system in their answers. Is there any effect of this? Are there any limitations in the log-based approach? A limitation could be the position-bias due to the *Nationaal Archief* system, where users tend to click on the first results they see.

3.5.3 Comparison with the *Nationaal Archief* System

Is there a bias from the original system when using the log-based test collection? The clicks of users in the search log may have been greatly influenced by the system that they have used. We compare the results with the ranking of the *Nationaal Archief* system (NA) itself. We use each and every topic of a part of the test collection, namely a month (January) of 2009, to query the *Nationaal Archief* system. The hit lists that this system return for every topic are downloaded. There are 4,110 topics in this month. The returned lists are another type of run. We use the same set of topics to query our systems. See Table 3.6.

Table 3.6: System-ranking of log-based approach compared to the *Nationaal Archief* system over January 2009.

	NA	BOOL	LM	LMS	NLLR	OKAPI
MAP	0.4283 (1)	0.1501 (6)	0.2703 (5)	0.2752 (4)	0.2787 (3)	0.2896 (2)
MRR	0.5112 (1)	0.1627 (6)	0.3100 (5)	0.3142 (4)	0.3177 (3)	0.3319 (2)
nDCG	0.5020 (1)	0.2473 (6)	0.3542 (5)	0.3728 (4)	0.3773 (3)	0.3896 (2)

There is a change. We find here that the *Nationaal Archief* system heavily outperforms the other systems and has a system rank of 1. This illustrates the click-bias existing in the log-based approach. When looking at the MAP score, the difference between the *Nationaal Archief* system and the OKAPI system, we find an improvement of the *Nationaal Archief* system of 47.89% ($t(4162) = 8.17$, $p < 0.01$, one-tailed). The differences among the other systems are also significantly on a 1% significance level.

We also show the contrast between the topics of the e-mail reference requests with our IR models and the *Nationaal Archief* system. The results for the performance of these topics with our systems have already been shown in Table 3.5, but now we append the results of the *Nationaal Archief* system. For the *Nationaal Archief* system, we find a MAP score of 0.2738, a MRR score of 0.3039, and a nDCG score of 0.3432. Here, we see that the system ranking of the *Nationaal Archief* system is positioned between the BOOL and LM systems, but the actual difference between the *Nationaal Archief* and LM systems is 1.42% and not significant. So the e-mail test collection does not have the same click-bias as the log-based approach has. Now, when we look at the Kendall's tau, we see a value of 0.467 over the six systems between the log-based and e-mail reference request approaches, so adding just one extra system shakes up the rank correlation of the system rankings between the two types of test collections and approaches.

In the next section, we further research the scope of the search log-based approach by exploring another third source, namely a research guide of the *Nationaal Archief* consisting of twenty-five questions and answers, which is recall-oriented.

3.5.4 The War Research Guide

In this section, we present a study on using the list of twenty-five questions of the *Nationaal Archief* (see Figure 3.5) called *De Oorlogsgids* (In English, *The War Guide*)[2] to construct an archival IR test collection. We present the experiment, compare and validate its results with the previous experiments.

Experimental Setup

There are twenty-five questions (topics or requests). These questions are example case-studies that are aimed to teach archival users to do research in archives, and refer to archives related to the Second World War (WWII, 1939-1945). These includes archives of Dutch governmental institutions or of persons/families, which were created or changed as a result of the German occupation of the Netherlands during WWII or the Japanese occupation of the former Dutch East Indies (nowadays Indonesia). The majority of the archival materials are publicly accessible.

[2] Retrieved 2011/03/15 from http://www.nationaalarchief.nl/collectie/ondersteuning/onderzoeksgids/default.asp?ComponentID=11865&SourcePageID=16483

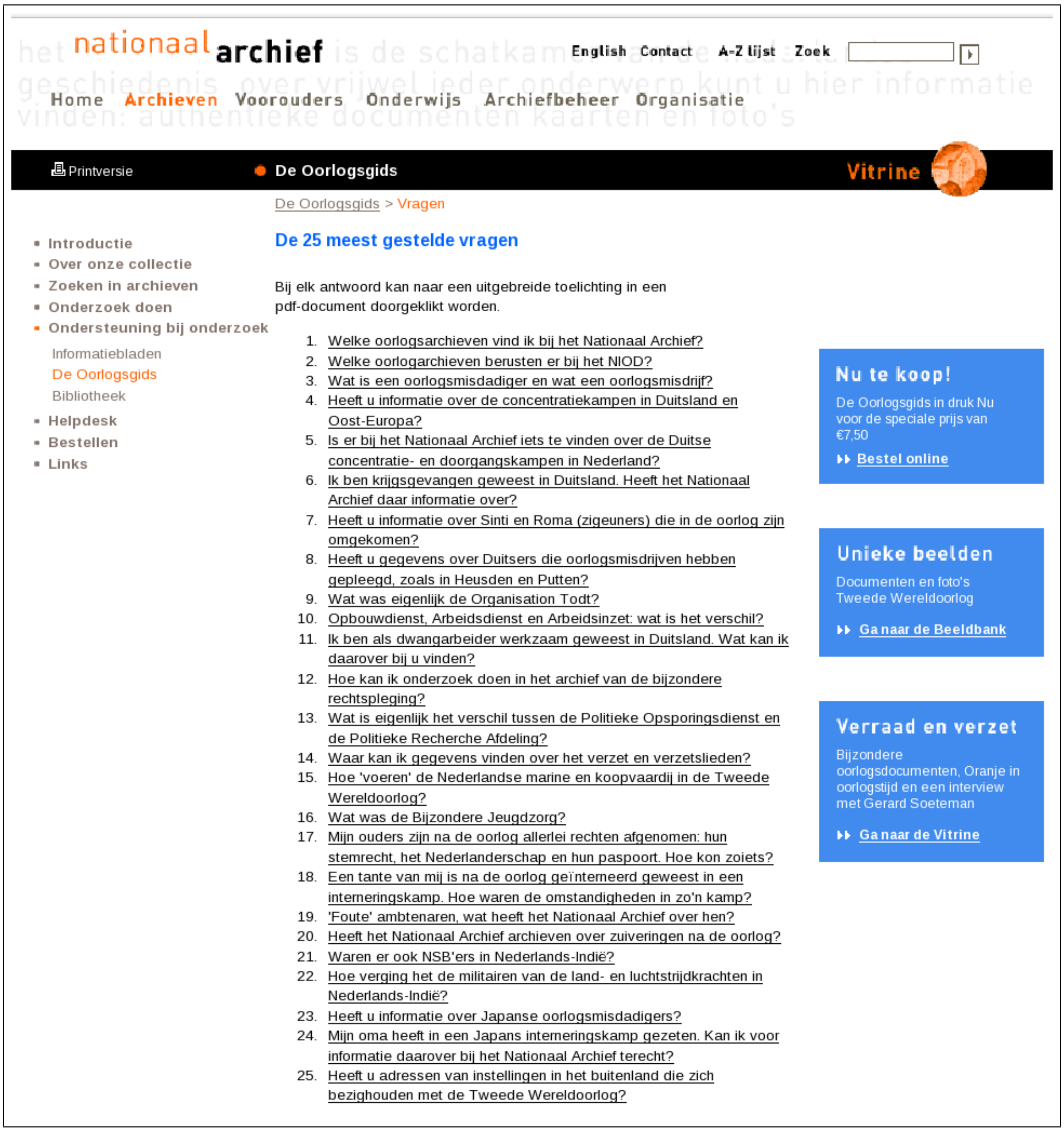


FIGURE 3.5: Research guide with twenty-five questions on the archives about the Second World War. Image courtesy of the *Nationaal Archief*.

Each of these topics also refers to a PDF file that answers a question, i.e. describes the context of the question and refers to relevant EAD finding aids for researching the answer to this question (relevance judgements and document collection). Here, there are 22 questions that also refer to EAD finding aids, so we have 22 topics. Therefore, we have the building blocks necessary to construct an IR test collection. Figure 3.6 illustrates these building blocks with two example topics. We refer to this as the "Research Guide" approach. A difference with the other approaches, is that the Research Guide approach is more oriented towards recall, as the lists of EAD finding aids suggested by the *Na-*

```
<topics>
  <topic nr="1">
    <title> war archives </title>
    <description> Which archives on the Second World War can be found at the Nationaal Archief? </description>
    <narrative> The War Guide consists of 25 most frequently asked questions on archives related to the Second World War. This guide tells us which archives have to be researched in order to get an answer to the questions. </narrative>
    <files> refers to 241 EADs (not listed to conserve space) </files>
  </topic>
  <topic nr="4">
    <title> concentration and internment camps </title>
    <description> Do you have information on the concentration and internment camps in Germany and Eastern Europe? </description>
    <narrative> Yes, the Nationaal Archief has archival material on that topic. It mostly touches upon information after the period of the closure of these camps. You will find data on the preparation for the relief of the war victims and—not always complete—lists of names of prisoners, survivors, and of those who have passed away. </narrative>
    <files> refers to 23 EADs (not listed to conserve space) </files>
  </topic>
</topics>
```

FIGURE 3.6: Example of topics based on frequently asked questions posed to archivists of the *Nationaal Archief* (here translated to English from Dutch).

tionaal Archief are exhaustive. For the log-based and e-mail approaches, there is a focus on precision. There are also shallow clicks in the log, for example, where users click once on an EAD finding aid.

We test this test collection with the following experimental setup. We use the ranking (or 'archival overview') that the *Nationaal Archief* system provides as ground truth. We use the systems as described in Subsection 3.4.1 on page 43. Additionally, given the topic specification in Figure 3.6, we have five types of runs: (**R1**) `<title>`, (**R2**) `<description>`, (**R3**) `<title>` + `<description>`, (**R4**) `<narrative>`, and (**R5**) `<title>` + `<narrative>`. So we use the (combination of the) content of these XML elements as queries to retrieve EAD finding aids. These runs stand for the amount of keywords used to express an information need into a query, where R1 is the most concise query possible and R5 is the most elaborate. We measure the results with the same performance measures as in the previous experiments. We return the top 1000 results for each topic.

Results

Now, the results. Table 3.7 shows the numbers. When we look at the MAP scores, we see that for run R1 and run R2, which consist of shorter content, the LM system is working best—this is different compared to the log-based approach. The OKAPI system comes second. We did not find significant improvements among the systems. This system ranking suggests that the keywords in

`<title>` and `<description>` are far from being optimal for recall, and very centered towards early high precision, though the suggested EADs in the qrels are exhaustive.

Table 3.7: System-ranking of five types of runs over the Research Guide topics.

	Run	NA	BOOL	LM	LMS	NLLR	OKAPI
MAP	**R1**	0.1095 (5)	0.0806 (6)	0.1272 (1)	0.1165 (4)	0.1167 (3)	0.1206 (2)
	R2	0.0871 (3)	0.0516 (6)	0.1041 (1)	0.0805 (5)	0.0817 (4)	0.0971 (2)
	R3	0.0970 (3)	0.0516 (5)	0.0886 (4)	0.0024 (6)	0.1037 (2)	0.1099 (1)
	R4	0.0297 (3)	0 (5)	0 (5)	0.0020 (4)	0.0962 (2)	0.1474 (1)
	R5	0.0740 (3)	0 (5)	0 (5)	0.0021 (4)	0.1034 (2)	0.1562 (1)
MRR	**R1**	0.3250 (5)	0.2934 (6)	0.4172 (1)	0.3746 (4)	0.3748 (3)	0.4081 (2)
	R2	0.2514 (6)	0.3274 (2)	0.4107 (1)	0.2813 (5)	0.2987 (4)	0.3217 (3)
	R3	0.3104 (5)	0.3274 (4)	0.3810 (2)	0.0241 (6)	0.3794 (3)	0.3840 (1)
	R4	0.1250 (3)	0 (5)	0 (5)	0.0024 (4)	0.3747 (2)	0.5095 (1)
	R5	0.4114 (2)	0 (5)	0 (5)	0.0024 (4)	0.3969 (3)	0.5418 (1)
nDCG	**R1**	0.2523 (5)	0.2224 (6)	0.2745 (4)	0.3150 (3)	0.3211 (2)	0.3382 (1)
	R2	0.2546 (2)	0.1441 (6)	0.1970 (5)	0.2381 (4)	0.2476 (3)	0.3179 (1)
	R3	0.2721 (3)	0.1441 (5)	0.1823 (4)	0.0492 (6)	0.2965 (2)	0.3449 (1)
	R4	0.2818 (3)	0 (5)	0 (5)	0.0581 (4)	0.3142 (2)	0.4114 (1)
	R5	0.2843 (3)	0 (5)	0 (5)	0.0554 (4)	0.3257 (2)	0.4227 (1)

When we add more text, or evidence, to the query (i.e. for runs R3, R4, and R5), we see that the OKAPI system still gives the best results. The LMS system is not working here as the MAP scores are just 2%, but when we remove the smoothing in the LM system, the MAP scores even plummet to 0, because there is no EAD finding aid that contains all the keywords of the query. In run R4, we note that the difference between the *Nationaal Archief* system and the OKAPI system is the greatest, where the performance almost quadruples with 396.30% ($t(20) = 6, p < 0.01$, one-tailed). In run R5, the improvement of OKAPI over the NA system more than doubles with 111.08% ($t(20) = 5.64$, $p < 0.01$, one-tailed). We note that the OKAPI system performs better overall in all runs in comparison with the standard *Nationaal Archief* system.

When we look at the rank correlations between the system rankings of these runs and the system ranking of the log-based approach over January 2009 in Table 3.6 using Kendall's tau, we see for R1 a tau value of 0.067, for R2 a value of 0.333, for R3 a value of 0.467, and for R4 and R5 a value of 0.733. For the rank correlation between the system rankings of the e-mail reference request approach and the Research Guide approach, we see for R1 a tau value of 0.6, for R2 and R3 a value of 0.467, and for R4 and R5 a value of 0.7333. When we omit the *Nationaal Archief* run for R4 and R5, we still see a full agreement between the log-based and e-mail reference request approaches with the Research Guide approach, as both then have a rank correlation of 1. However, the rank correlations of the R1, R2 and R3 runs do not show such consistency.

Moreover, we note that the test collection derived from the Research Guide does not have the same impact of the click-bias as the test collections derived from the log-based approach seem to have. Another finding is that the OKAPI system is the most effective for (very) long queries across the compared systems, and also for shorter queries. This is a similar finding as with the log-based approach, where it also performed best, given our experimental setup with the tested systems. However, there are also limitations to the Research Guide approach, because we find that the topics here do not seem to be optimally formulated, and focused towards exact matching, while the qrels are oriented towards measuring recall. When we add more evidence of the information need to the query, we see that the system rankings become similar as with the other test collections. As the Research Guide qrels are oriented towards recall, we see that the MAP scores are considerably lower than the results of the other approaches.

3.6 Conclusion

This chapter investigated a search log-based approach to the evaluation of access systems driven by EAD finding aids. Information retrieval has developed standard benchmarks that can be used to evaluate the retrieval effectiveness. However, these generic benchmarks focus on a single document genre, language, media-type, and searcher profile that is radically different from the unique content and user community of a particular archive. One of the main challenges in the evaluation is to develop contextualized evaluation methods that closely capture the unique setting at hand. Therefore, we have proposed the following. By using an archive's own collection and exploiting readily available interaction data in search logs, we can create a domain-specific test collection tailored to the case at hand. That is, having a representative document collection and representative sets of search requests. As a test collection, it can be used and reused for comparative testing under the same experimental conditions.

We conducted a large case study using a large set of EAD finding aids and search logs of the *Nationaal Archief* EAD search system. These logs are massive as these cover several years of interaction with this system. This resulted in a test collection to evaluate the full-text retrieval of archival finding aids in EAD. We presented generic methods to derive a domain-specific test collection, and used a range of retrieval models to determine the effectiveness of the test collections. Our extraction methods are naive—we treat every clicked document as pseudo-relevant—but suffice to illustrate the viability of the approach.

We compared the results by repeating the experiment using a set of traditional topics derived from email requests to the archive and the archivist's responses. We found complete agreement between the log-based evaluation and the traditional topics. However, when we extended the experimentation by adding one extra system, namely the *Nationaal Archief* system itself, we have to nuance our findings, as the rank correlations then changes due to the bias of this system. When we repeated the experimentation by deriving another test collection from a *Nationaal Archief* research guide, we also saw similarities and differences, which may due to the recall-based orientation of this test collection, compared to the log-based approach that focused on precision. Still, our initiative of a log-based approach to IR evaluation is a step forwards towards automated system-centered evaluation of access to EAD finding aids.

So far, we looked at full-text retrieval only, but we can also evaluate on an XML element level by looking at descriptions within an EAD finding aid. This is what we will investigate in Chapter 4.

Chapter 4

On Archival Description Principles for Retrieval

Access to archival finding aids in EAD is a two-tier approach. In Chapter 3, we looked at the evaluation of document retrieval of whole EAD finding aids. Now, we go a step further by investigating the evaluation of XML retrieval on EAD finding aids with the systems developed in Chapter 2, and specifically look at the special archival structure that also exists in EAD finding aids.

4.1 Introduction

The French phrase *respect des fonds* is a pivotal concept in archival science (Ketelaar, 1996, p.34). The term *fonds* is defined by Pearce-Moses (2005) as

> The entire body of records of an organization, family, or individual that have been created and accumulated as the result of an organic process reflecting the functions of the creator.[1]

Horsman (2002, p.2) notes that "the fonds is believed by many archivists to embody the core principles that the profession must use for the arrangement and description of archives." In his paper, he discusses the revisions by archivists of the concept of respect des fonds. Notably, a fonds can be expressed almost as a mathematical formula (Horsman, 2002, p.17):

> A fonds (F) is a any set of relationships (r_1, r_2, r_3, etc.), where a record (a_1, a_2, a_3, etc.) is an element in any of the identified (and non-identified) relationships. Evidently, a record can be part of two or more relationships, and two or more fonds.

This definition puts respect des fonds in a new limelight, compared to the first written formalization of classical archival science in the Dutch manual of 1898 by Muller, Feith and Fruin (Horsman et al., 2005). This manual formalized

[1] Retrieved 2011/06/24 from `http://www.archivists.org/glossary/term_details.asp?DefinitionKey=756`.

two pillars of classical archival theory that can be summarized as the *Principle of Respect for Archival Structure*. This principle states that the fonds is a whole, whose historically determined structure should not be shifted to another arrangement, and if so, the original arrangement should be restored (Ketelaar, 1996, p.34). The foundation of the principle, originally based on paper archives and paper finding aids, are two concepts (Cook, 1997, p.21).

Provenance This concept states that archival records in an archive need to be carefully separated according to their origin. This means that records in a fonds cannot be mixed with other creators' fonds.

Original Order This concept is considered to be the most important rule from which all other rules follow. This concept says that records should not be placed in artificial arrangements based on chronology, geography, or subject, but in the original organization of the creator.

This manual was authoritative, and its principles with explanations became eventually unquestionable dogmas (Ketelaar, 1996, p.35). The rationale is that if we respect the form, structure and context of creation, then we also respect the relations between the content data, the relations between the records, and the relations between the records and their creators with their mission statements (Thomassen, 2001, p.383).

The principles are inherited and also projected on archival finding aids in EAD. The finding aids were the backbone of IR systems in 1979 (Bearman, 1979, p.180), and are still driving archival IR systems at their core (Gilliland-Swetland, 2001). Haworth (2001, p.24) wonders whether EAD has provided us with, what (Duranti, 1992) calls, a 'principal, multipurpose descriptive instrument' that re-integrates preservation of meaning, exercise of control, and provision of access.

A promise of 'electronic' finding aids (e.g. in EAD) is to move beyond the 'fixity' of paper finding aids, and enable the flexibility of archival finding aids as information discovery and retrieval tools for different users who may wish to use these 'electronic' finding aids differently (Gilliland-Swetland, 2005, p.200). However, then we may no longer adhere to the *Principle of Respect for Archival Structure* that have been applied on traditional paper inventories, and inherited by archival finding aids in EAD. Still, Menne-Haritz (2001, p.63) does not see this controversy, and notes that access does not mean that the description and presentation of archives are user driven, as the emphasis put on access is to 'enable,' and does not present data or other information as true representations of reality. In this digital era, the archival finding aid can make different narratives of an archive explicit through 'flexible' retrieval and display, and this is proposed by Gilliland-Swetland (2001) so as to 'popularize' the archival finding aid. This chapter investigates the IR effectiveness of the archival finding aid in EAD when we let *respect des fonds* govern, and when we 'popularize' it.

4.2 Problem Statement

The retrieval experiments conducted in the archival domain is scant. Studies that investigate the retrieval effectiveness of systems that operate on archival finding aids in EAD is almost non-existent. As pointed out by Cox (2008), archivists have been creating finding aids in the last three decades in standardized formats. Archivists have conducted descriptive work with "little knowledge of how researchers find and use archival sources" (Cox, 2008, p.5). Archival finding aids are value-added descriptive tools, such as traditional inventories, registers, indexes, or guides (for example compiled in word processors like Microsoft Word). We focus only on traditional inventories in EAD format.

We can enable the flexibility of archival finding aids (e.g. in EAD) as information discovery and retrieval tools for different users who may wish to use these 'electronic' finding aids differently (Gilliland-Swetland, 2005, p.200). Still, the principle of *respect des fonds* governs the archival finding aid in EAD (e.g. Haworth, 2001; Ketelaar, 1996). This leads to the main research question:

- *Q3: How effective are archival principles—inherited by traditional inventories and subsequently cast on EAD finding aids—for IR?*

We can retrieve any and arbitrary elements in XML retrieval using the *relevance* of returned elements given the keywords of the query that appear in an element. However, archivists have defined principles to arrange, describe and provide access to the materials. These principles are *provenance*, and its corollary *original order* that exists in an EAD finding aid. What do these principles, inherited from traditional paper inventories and projected on EAD finding aids, mean for the effectiveness of retrieval of these finding aids? For a given search request, how many descriptions can be found that a user wants?

We note that the quality of an archival finding aid in EAD depends on the archivist who created it, such as the quality of descriptions. We do not have have control over this. Still, we can investigate the retrieval effects of these principles in a controlled experiment by studying a large and representative sample, which the *Nationaal Archief* finding aids in EAD are.

3.1 What is the effectiveness of retrieval when grouping descriptions by EAD finding aid?

3.2 What are the retrieval effects of returning descriptions in the original order of the archival finding aid in EAD?

The markup is a representation of information. The representation is used to interpret information. The archival representation of information can become very detailed. There is a point that the markup does not longer serve archival access and is no longer meaningful for users, e.g. for adhoc queries. In Cox (2008, p.5) it is argued that the archival representation may diverge with the representation of users (also see Yakel and Torres, 2003). What is the scope of the EAD markup for effectively finding information? Our research focused so far on so-called Content Only (CO) queries as we indexed and queried all elements in the element and aggregation-based approaches of the README system in Chapter 2. However, given the XML structure in EAD files, users can also express their information needs in so-called Content-And-Structure (CAS) queries. This type of query also incorporates some structure in the form of an XML element name in conjunction with keywords. This could be useful when users only want to search in the `archdesc/did` element that contains the description of the collection material itself like the biographical description of the person or organization in `<bioghist>` or a detailed narrative description of the collection material in `<scopecontent>`. The user can opt to search for the second, usually longer part of the `<archdesc>`, namely the full inventory with a nested list of descriptions on different component levels in `<dsc>`.[2]

3.3 Would structural EAD cues expressed in the query be useful?

We adopt a traditional IR system-centered experimental approach. In the next sections, we start with creating test collections on the XML element level, and then move on towards the experimentation and results.

4.3 Test Collection for Within-Fonds Retrieval

In this section, we describe two approaches to create a test collection for IR evaluation. The first approach employs a conventional assessment system where people can select elements in an EAD finding aid they find interesting. The second approach derives a test collection from clicks on descriptions within an EAD finding aid recorded in the search log of the *Nationaal Archief*.

4.3.1 XML IR Test Collection with Manual Assessments

We introduce the README Assessment System for constructing an XML IR test collection with manual assessments tailored to archival finding aids in EAD. Figure 4.1 shows this system. We re-use the README Result Display in Figure 2.10 of Chapter 2 on page 29, and enhance it with features to record what users find relevant in a finding aid in EAD for their search request (topic).

We use Javascript for the system to act on clicks and Perl for post-processing the clicks, to allow for element selection, or de-selection in case a user changes his/her mind. We note that we also make wrapper elements like `<did>` selectable, while for example a component `<c01>` is also a wrapper element and selects the same content—often consisting of a unit ID and unit title—which is verbose but adheres to the official EAD 2002 version. Our rule of thumb is

[2] Stibbe (1992) discussed the idea that the fonds presents the context and the records are described at series level and below.

Figure 4.1: The README Assessment System for selecting useful or relevant elements within a fonds, which can be used to construct an EAD test collection.

that only relevant elements are selected that are closest to the relevant descriptions, so in case both a unit id and title are considered relevant, `<did>` would get selected.

How does it work? The README Assessment System adds a folder icon for every and each element that are present in the EAD finding aid. When a user hovers over this folder icon, a label appears that shows the XML element path or XPath. When a user clicks on this folder icon, it changes to a green 'checked' icon. When a user clicks on this icon again, the selected element gets de-selected, and the icon changes back to the folder icon, and so on. We record for every selection the EAD finding aid, the query, the XPath, and the ranking that have been used. Every selected XPath gets a unique ID by using the XSLT *generate-id* function and its XPath as input. In Figure 4.1, given the search request expressed as the query 'bijzondere rechtpleging' (In English, *special justice*), we see that 3 elements have been selected in the nested components: 2 unit IDs and 1 unit title. These selected elements get stored in a text file, for each search episode that started with a query in an EAD finding aid.

<topics>
<topic nr="1">
<title> ontdekking australie zuidland nieuw holland nieuw zeeland diemensland tasmanie abel tasman willem janszoon </title>
<narrative> I am interested in the discovery of Australasia by European discovers, in particularly the discovery of Australia (South Land, New Holland), Tasmania (Van Diemen's Land) and New Zealand. It is not only about the first contact, but also further discoveries. </narrative>
<files> 1.04.02.ead—Verenigde Oost-Indische Compagnie </files>
</topic>
<topic nr="2">
<title> interland voetbal nederland belgie </title>
<narrative> I am doing research on the relationship between both neighboring countries, and especially which tensions and sentiments result from international football (soccer) matches. I am interested in reports of international matches between the Netherlands and Belgium. </narrative>
<files> 2.19.123.ead—Koninklijke Nederlandse Voetbalbond (KNVB) </files>
</topic>
<topic nr="3">
<title> olympische spelen kandidatuur amsterdam </title>
<narrative> I am searching for applications of the Netherlands to organize the Olympic Games. I am interested in reports or notes of people who were involved in these applications. </narrative>
<files> 2.19.124.ead—Nederlands Olympisch Comite (NOC) </files>
</topic>
<topic nr="4">
<title> korea oorlog wapenstilstand </title>
<narrative> I am interested in the Korean War. I would like to see reports or other documents that give insight in the eventual armistice. </narrative>
<files> 2.03.01.ead—Ministeries voor Algemeene Oorlogvoering van het Koninkrijk (AOK) en van Algemene Zaken (AZ): Kabinet van de Minister-President (KMP), (1924) 1942-1979 (1989) </files>
</topic>
<topic nr="5">
<title> willem drees greet hofmans juliana bernard </title>
<description> </description>
<narrative> I am looking for background information of the Greet Hofmans affair and the role of Willem Drees. I am interested in his notes. I am also interested in literature references. </narrative>
<files> 2.21.286.ead—W. Drees [levensjaren 1886-1988] en enkele familieleden, (1853) 1900-2000 (2002) </files>
</topic>
</topics>

FIGURE 4.2: Topics created for manual assessment of relevant elements within a fonds using the README Assessment System (<narrative> translated to English from Dutch). Relevant is concrete background information in the description of the context of the archive and descriptive subordinate components referring to relevant materials.

The results. We have assessed topics using the README Assessment System, see Figure 4.2 for the search requests. We focus on a specific ad hoc topic in a single most relevant EAD finding aid only, which may range up to a 1,000 pages and contain many thousands of XML elements, and search within a fonds for all relevant descriptions, then judge and select. This results in a relatively modest test collection of in total 73 relevant elements in five EAD finding aids.

4.3.2 From Web Anchor to XML Element

A great value of EAD finding aids is that it can provide focused access to archival descriptions (DeRose, 1997; Kiesling, 2001; Pitti, 1999). On the World Wide Web, references to Web pages called hyperlinks are used. Hyperlinks can link a user to the beginning of Web pages, or to a position in a Web page using a fragment identifier. We study the search interaction with the *Nationaal Archief* system (of 2009), also see Chapter 3, Subsection 3.3.1 on page 37 for the overview of the search interaction possible with this system. The *Nationaal Archief* system also provides access to descriptions with fragment identifiers. This is necessary, because the EAD finding aids can be hundreds of pages in content, therefore the system chunks the archival components in its presentation to users for efficiency reasons, and marks the components by a fragment identifier. The fragment identifiers are specifically known in the *Nationaal Archief* system as a *pageID*. These pageIDs are randomized text strings consisting of six characters and starting with an 'N,' like 'N1015C, 'N10160,' 'N10120,' and etcetera. Furthermore, these links are persistent, ensuring continuous access. There are thousands of these pageIDs present in the search log, each representing a click on a description. How can we make these pageIDs meaningful, so we can understand the clicks, and use them?

An overview first. We extract all the 'pageIDs' of an EAD finding aid from the search log, and its corresponding EAD finding aid. This results in a list, and both are required to link to an archival finding aid in EAD. For example,

```
http://www.nationaalarchief.nl/webviews/page.webview?eadid=NL-HaNA_1.01.01.01&
    pageid=N10168
```

is a link of the NA referring to the EAD finding aid 'NL-HaNA_1.01.01.01' and the description of subordinate component *'A1 delen en banden'* (in English, 'A1 parts and links') identified by the pageID 'N10168.' Also see Figure 4.3.

The *Nationaal Archief* system transforms the EAD finding aids into HTML pages. To discover where a pageID is located in the EAD finding aid, we employ a technique called *data scraping*, where a computer program extracts data from human-readable output, in this case HTML pages, coming from another program. Since the *Nationaal Archief* logs with EAD usage go from 2007 to January of 2009, we note that we could not control for the later (post January 2009) changing of the composition of the NA finding aids (e.g. merging), which may render previously recorded pageIDs unavailable. Moreover, we do not have

Figure 4.3: *Nationaal Archief*'s HTML presentation of a subordinate component in 'NL-HaNA_1.01.01.01,' which gets linked by pageID 'N10168.' Image courtesy of the *Nationaal Archief*.

the complete set of EAD finding aids.[3] So not all clicks that referred to an EAD finding aid could be mapped. However, there is still a substantial amount of pageIDs available for further meaningful representative experimentation and analysis. The procedure for the data scraping is:

1. First, using the list of EADs and their pageIDs, we crawl the *Nationaal Archief* website directly and download the HTML Web pages. In total, we download 10,659 of the original Web pages for one month of log data, and 33,551 Web pages for whole 2008. We note that we only download unique EAD finding aid plus pageID pairs.
2. The second step is to map the pageIDs to an XPath (or XML element path) identifier. An XPath identifier formally notates the absolute position of an element in an XML file. We utilize the 5,934 original EAD finding aids in XML to create a new set of EAD finding aids enhanced with fragment identifiers, where for each element we render an XPath identifier using an XSLT stylesheet compiled with the Saxon processor.
3. The third step is to align each pageID with an XPath reference using both sets of files. Additionally, we extract the anchor text of each pageID. This results in a table that we can use as a dictionary to look up the value of each pageID.
4. Finally, we post-process the Web search logs by enhancing the recorded clicks with this information. We aggregate for each UUID all clicks into a separate file, while also keeping track of the session of each click.

For constructing the test collection, we adopt the log-based approach as out-

[3] For whole 2009, there are 27 EAD finding aids which we do not have. For January of 2009, there are 13 missing EAD finding aids. We have 5,934 EAD finding aids.

lined in Chapter 3. We apply the data scraping procedure on the clicks recorded in the whole of 2008 and in January of 2009. For the whole 2008, there are 31,742 topics with 57,776 types of clicked elements. For January 2009, there are 3,423 topics with 11,520 types of clicked elements.

4.4 The Scope of EAD Markup for Retrieval

In this section, we investigate research question 5.3 first. What elements do people select, and would this knowledge to make retrieval more effective? We posit the background, then present the analysis and the results.

4.4.1 Background

Archival finding aids in EAD could become very long in content and complex in the logical structure due to the archival description (Haworth, 2001; Pitti, 1999). In fact, EAD has been created by studying the structure and functionality of traditional finding aids (Ruth, 1997, p.328).The 2002 version of the EAD DTD consists of 146 elements. The purpose is to optimally describe archival materials using this set of elements. An advantage of EAD is its flexibility, which eases its adoption by different institutions. The trade-off is that differences exist of the same standard in its application in practice.

The different considerations for the structure in EAD finding aids have been explained by (Ruth, 1997). Markup languages, like EAD, should separate content from layout (Bosak and Bray, 1999; Bray et al., 2008). However, for the design of EAD there is not such a strict separation, as 'online and print presentation' was considered one of the functions of an element (Ruth, 1997, p.314). EAD has a broad scope, but what is exactly the scope of EAD for search and retrieval? Duff and Stoyanova (1998) have already studied, before the existence of EAD, the scope of elements in terms of usable displays through a focus group. Can we uncover the scope of EAD by studying the search logs?

4.4.2 Analysis and Results

To answer this question, we count the elements that users have selected in 2008 and January of 2009 in the log-based test collections. Table 4.1 presents the results for the former and Table 4.2 presents the results for the latter. We see that for both distributions, there are 3 main elements that get used. The majority of the within-fonds clicks are on `<head>`, the wrapper element `<did>` gets selected in about a third of the cases, and the `<unittitle>` comes third. We see that the `<unittitle>` gets selected more often in 2009.

Table 4.3 shows the top 10 XML paths for whole 2008, and Table 4.4 depicts this for a month of 2009. The path to the whole fonds (no fragment identifier) `/ead[1]` tops both lists. The remainder of both lists shows eight types of clicks on the `<head>` element. These refer to the headings of descriptions. We see

Table 4.1: Distribution of clicked elements in 2008 ($N = 68,897$).

Element	Count (%)
<head>	42,037 (61.01)
<did>	22,224 (32.26)
<unittitle>	4,431 (6.43)
<corpname>	120 (0.17)
<item>	48 (0.07)
<geogname>	28 (0.04)
<entry>	5 (0.01)
<p>	4 (0.01)

Table 4.2: Distribution of clicked elements in January 2009 ($N = 9,622$).

Element	Count (%)
<head>	5,450 (56.64)
<did>	3,308 (34.38)
<unittitle>	836 (8.69)
<corpname>	16 (0.17)
<item>	5 (0.05)
<persname>	4 (0.04)
<geogname>	3 (0.03)

Table 4.3: Top 10 clicked XML paths in 2008 ($N = 150,408$).

XML Path	Count (%)
/ead[1]	81,511 (54.19)
/ead[1]/archdesc[1]/dsc[1]/head[1]	7,808 (5.19)
/ead[1]/archdesc[1]/did[1]/head[1]	6,726 (4.47)
/ead[1]/archdesc[1]/descgrp[2]/head[1]	3,687 (2.45)
/ead[1]/archdesc[1]/descgrp[3]/head[1]	3,079 (2.05)
/ead[1]/archdesc[1]/dsc[1]/c01[1]/did[1]	2,855 (1.90)
/ead[1]/archdesc[1]/descgrp[1]/head[1]	2,538 (1.69)
/ead[1]/archdesc[1]/descgrp[4]/head[1]	1,975 (1.31)
/ead[1]/archdesc[1]/dsc[1]/c01[2]/did[1]	1,755 (1.17)
/ead[1]/archdesc[1]/descgrp[1]/bioghist[1]/head[1]	1,488 (0.99)

Table 4.4: Top 10 clicked XML paths in January 2009 ($N = 17,267$).

XML Path	Count (%)
/ead[1]	7,645 (44.28)
/ead[1]/archdesc[1]/dsc[1]/head[1]	1,124 (6.51)
/ead[1]/archdesc[1]/did[1]/head[1]	687 (3.98)
/ead[1]/archdesc[1]/descgrp[2]/head[1]	520 (3.01)
/ead[1]/archdesc[1]/dsc[1]/c01[1]/did[1]	447 (2.59)
/ead[1]/archdesc[1]/descgrp[3]/head[1]	411 (2.38)
/ead[1]/archdesc[1]/descgrp[1]/head[1]	337 (1.95)
/ead[1]/archdesc[1]/dsc[1]/c01[2]/did[1]	296 (1.71)
/ead[1]/archdesc[1]/descgrp[4]/head[1]	210 (1.22)
/ead[1]/archdesc[1]/descgrp[1]/bioghist[1]/head[1]	192 (1.11)

that the heading of the descriptive subordinate components `dsc[1]/head[1]` (In Dutch, *Beschrijving van de series en archiefbestanddelen*) gets clicked most often when users search within a fonds.[4] The clicks on the heading of the introductory information `archdesc[1]/did[1]/head[1]` (In Dutch, *Beschrijving van het archief*) come second, which is equivalent to the beginning of an EAD find-

[4] We note <co1> corresponds to the level of series in the *Nationaal Archief* system.

ing aid in the *Nationaal Archief* system. When we look at the <did> elements within a <c> element, we see that both occur in the subordinate component <c01>, namely the first and second <c01> element. The 2002 EAD DTD shows that <unittitle> occurs in these <c> elements. So are structural EAD cues expressed in the query useful? Using XML retrieval techniques, we can narrow down a search by retrieving specific elements only. There is a finite set of elements that gets clicked, but this is due to the *Nationaal Archief* system. It links to these elements only, but users do not search in the content of the headings <head>, since they contain little informative value, but use them for navigation.

The <head> element contains the title of a section, which does not tell us anything. The <did> element is a wrapper. The <unittitle> element occurs in every <c> element, but we do not know at what level. What happens when we remove the <head>, <did> and <unittitle> elements from our analysis and look at their first preceding element (or parent)? Table 4.5 shows the distribution of the preceding elements for all the clicks in 2008, and Table 4.5 for a month of 2009. Compared to the clicks element in Tables 4.1 and 4.2, we see that the number of element types increases by more than a tenfold. The <did> is in the top of both lists. An advantage is that this element covers much of the descriptions in an archival finding aid in EAD. A disadvantage is that it is verbose, and it does not narrow down the information for effective retrieval. The description group <descgrp> is also in the top, which here refers to the context of creation (In Dutch, *Archiefvorming*), and in particular general information such as user instructions and access restrictions. The deepest component is <c09>, though the deeper a component is, the fewer use there is. Again, would structural EAD cues expressed in the query be useful?

This depends on the user's information need, since we do not find a clear favorite element that people use. We could assume that when people are searching for contextual information, the <bioghist> is a 'popular' element because of the content of this element. Moreover, when people search for archival materials, the higher level <c> elements stand out. This suggests that indexing and retrieving all and any elements, as the README system does (see Chapter 2), is the most practical solution for a uniform archival search system.

4.5 Retrieval Effects of Grouping Descriptions

In this section, we investigate research question 5.1 on the the retrieval effects of grouping elements per fonds. We explain the background, then move on to the experiment setup of the studies, and finally present the results.

4.5.1 Background

As Pitti (1999) puts, archival description represents a fonds. First, by the principle of provenance, all material created or received by the same individual,

Table 4.5: Distribution of the first preceding element of the clicked elements in 2008 ($N = 68,897$).

Element	Count (%)
<descgrp>	11,405 (16.55)
<did>	11,157 (16.19)
<dsc>	7,808 (11.33)
<c01>	7,304 (10.60)
<c02>	6,556 (9.52)
<c03>	4,248 (6.17)
<bioghist>	4,162 (6.04)
<c04>	2,471 (3.59)
<scopecontent>	1,499 (2.18)
<prefercite>	1,444 (2.10)
<accessrestrict>	1,442 (2.09)
<custodhist>	1,295 (1.88)
<odd>	1,261 (1.83)
<c05>	1,152 (1.67)
<arrangement>	1,114 (1.62)
<userestrict>	1,068 (1.55)
<processinfo>	845 (1.23)
<relatedmaterial>	727 (1.06)
<appraisal>	443 (0.64)
<c06>	385 (0.56)
<separatedmaterial>	251 (0.36)
<index>	210 (0.30)
<otherfindaid>	158 (0.23)
<bibliography>	84 (0.12)
<origination>	81 (0.12)
<c07>	77 (0.11)
<accruals>	57 (0.08)
<list>	48 (0.07)
<item>	39 (0.06)
<originalsloc>	30 (0.04)
<unittitle>	28 (0.04)
<c08>	16 (0.02)
<c09>	15 (0.02)
<fileplan>	11 (0.02)
<row>	5 (0.01)
<acqinfo>	1 (0.00)

Table 4.6: Distribution of the first preceding element of the clicked elements in January 2009 ($N = 9,622$).

Element	Count (%)
<did>	1,523 (15.83)
<descgrp>	1,503 (15.62)
<c01>	1,265 (13.15)
<dsc>	1,124 (11.68)
<c02>	902 (9.37)
<c03>	629 (6.54)
<bioghist>	508 (5.28)
<c04>	313 (3.25)
<scopecontent>	186 (1.93)
<odd>	179 (1.86)
<accessrestrict>	179 (1.86)
<prefercite>	150 (1.56)
<custodhist>	143 (1.49)
<userestrict>	140 (1.45)
<c05>	138 (1.43)
<processinfo>	121 (1.26)
<altformavail>	116 (1.21)
<arrangement>	114 (1.18)
<phystech>	82 (0.85)
<appraisal>	68 (0.71)
<relatedmaterial>	65 (0.68)
<c06>	51 (0.53)
<separatedmaterial>	29 (0.30)
<otherfindaid>	20 (0.21)
<bibliography>	17 (0.18)
<index>	15 (0.16)
<item>	14 (0.15)
<c07>	8 (0.08)
<origination>	6 (0.06)
<list>	5 (0.05)
<accruals>	3 (0.03)
<unittitle>	3 (0.03)
<fileplan>	1 (0.01)
<c08>	1 (0.01)
<c09>	1 (0.01)

family, or organization is kept together. Second, by the principle of original order, all material of the same creator is stored in its original organization and sequence. In this section, we investigate the retrieval effects of the first principle, provenance. According to Schellenberg (1965, pp.41–42), this is an arrangement of archives "according to their origins in an organic body or an organic activity." In practice, this means keeping (grouping) a fonds together as an organic whole.

Bearman and Lytle (1985, p.14) think archivists have a unique perspective provided by the principle of provenance as it concerns organizational activity, especially how organizations create, use, and discard information. They offer a critique of the principle of provenance. Since this principle is a core concept for archival processing, it also means that archival retrieval is an inferential process Bearman and Lytle (1985, p.16), where archivists have to "infer which organizational units might have undertaken relevant activities and therefore might have produced documentation pertinent to the subject query at hand." The principle of provenance results from the 19-th century view of organizations, which were mono-hierarchical in structure, and Bearman and Lytle (1985) list weaknesses of this model, such as it is a poor model for understanding modern organizations.

Campbell (1967, p.280) points out that an inexpensive way to arrange an archive is by provenance, but like Bearman and Lytle (1985), also notes that to search by inference is not easy because the searcher has to search within the lines of organization created by provenance. EAD finding aids still strictly adhere to the provenance principle (Pitti, 1999). What is its use in terms of finding content in EAD finding aids in the limelight of modern IR?

4.5.2 Study 1: Measuring Concentration Ratios

In this study, we measure how strong and often the principle of provenance is for the topics. We describe our experimental setup and report the results.

Experimental Setup

To investigate the retrieval effects of provenance, we study the log-based test collection. This test collection is based on the clicks of people using the *Nationaal Archief* system (see Section 3.3.1 in Chapter 5 on page 37). This system allows users to search within a fonds and across different fonds.

For our analysis, we use the *concentration ratio* (CR), which is a term from economics (Saving, 1970). It measures the total output that is produced in an industry by a given number of companies in the industry, usually used to show the extent of market control of the largest companies in the industry, and to illustrate the degree to which an industry is oligopolistic.[5] In our case, we can treat an 'industry' as a query (or topic), the 'market control' as the number of clicks, and the 'companies' as EAD finding aids. So the 'extend of the market control' stands for the effectiveness of the principle of provenance. The more extensive the 'market control' is, the stronger the evidence is for the effective-

[5] Oligopolistic is the adjective for the noun 'oligopoly.' 'Oligopoly' is "a market situation in which each of a few producers affects but does not control the market." Retrieved 2011/03/25 from `http://www.merriam-webster.com/dictionary/oligopolistic`.

Table 4.7: Number of clicks per topic over 2008.

# Clicks	#Topics (%)
1	10,398 (43.83)
2	3,642 (15.35)
3	2,123 (8.95)
4	1,352 (5.70)
5	999 (4.21)
6	673 (2.84)
7	538 (2.27)
8	403 (1.70)
9	380 (1.60)
10	331 (1.40)
> 10	2,882 (12.15)

Table 4.8: Number of clicks per topic over 2009.

# Clicks	#Topics (%)
1	1,525 (44.55)
2	490 (14.31)
3	302 (8.82)
4	182 (5.32)
5	154 (4.50)
6	86 (2.51)
7	80 (2.34)
8	83 (2.42)
9	59 (1.72)
10	59 (1.72)
> 10	403 (11.77)

ness of provenance. The concentration ratio is defined as

$$CR_N = \sum_{i=1}^{N} F_i \tag{4.1}$$

where F_i defines here the 'market share' of the i-th EAD finding aid with a total N number of clicks appearing given a query. We focus on topics that have more than a certain click threshold to cope with topics that only have 1 click (means total concentration), or not really used by users, which may give skewed results. We focus our attention on the 'market share' of the top clicked EAD finding aid for every topic, so CR_1. The concentration ratio levels reach from no (0), low (0 – 0.5), medium (0.5 – 0.8), high (0.8 – 1), and total (1) concentration. In case there is a total concentration ratio, we speak of a monopoly.

Results

Now, the results. We count the number of clicks per topic, and group them by that number over whole 2008 in Table 4.7 and January 2009 in Table 4.8. We note that for 2008, 65.3% (98,196) of the clicks come from 2,882 $10+$click topics, and for 2009, 55% (9,497) of the clicks occur in 403 (11.77%) topics with more than 10 clicks. So a majority of the click samples occur in $10+$click topics.

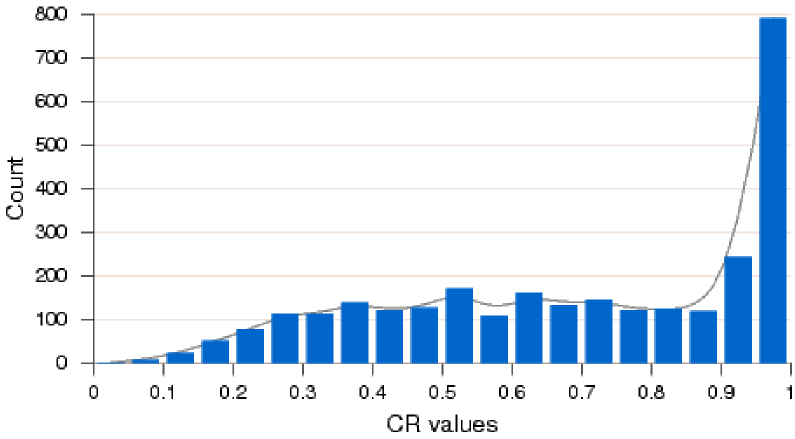
Next, we count the number of different EAD finding aids associated with a topic. We focus only on January 2009 in Table 4.9. The topic *suriname* (in English, *Surinam*), has an association with 39 different EADs, so users have used this range of EAD finding aids for that search request. Other popular topics where people searched across EAD finding aids are *ptt*, which is the acronym for "Staatsbedrijf der Posterijen, Telegrafie en Telefonie" (In English, *Post, Telegraph and Telephone Company*), *"financien"* (In English, *finances*), and *buitenlandse zaken* (In English, *foreign affairs*). However, 2,512 EAD finding aids have one

Table 4.9: Top 10 topics associated with the number of different consulted EADs in January 2009.

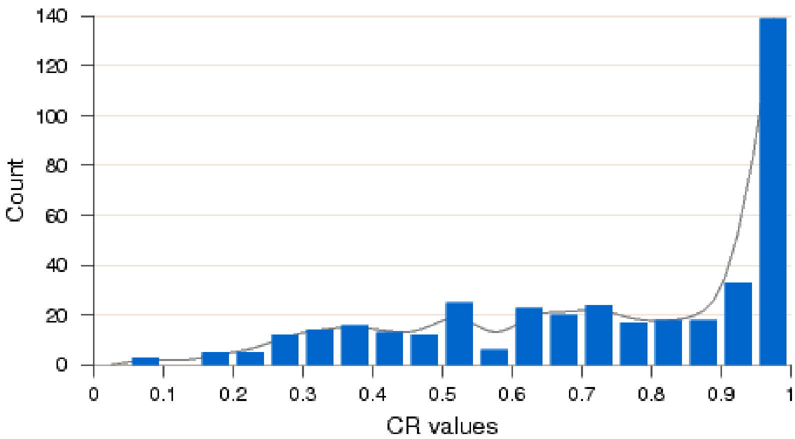
# EADs	Topics (incl. English Translation)	Topic Count
39	*"suriname"* (Surinam)	1
29	*"ptt"* (—)	1
26	*"financien"* (finances)	1
21	*"buitenlandse zaken"* (foreign affairs)	1
19	*"noordwijkerhout"* (—), *"4.miko"* (—)	2
17	*"japanse"* (Japanese), *"genie"* (military engineer), *"wetenschap"* (science), *"marine"*, *"europese"* (European)	5
16	*"rechtbank van eerste aanleg"* (Court of First Instance)	1
13	*"admiraliteit"* (admiralty), *"indonesie"* (Indonesia)	2
12	*"berbice"* (—), *"spoorwegen"* (railways), *"militairen"* (soldiers)	3
11	*"nijmegen"* (—), *"foto"* (photo), *"lisse"* (—)	3
10	*"1.01.04"* (—), *"knil"* (—), *"wic"* (—), *"1.11.01.01"* (—)	4
9	*"hof van holland"* (—), *"leycester"* (—), *"batavia"* (—), *"drees"* (—), *"meijer poorter van vlaardingen in 1789"* (—)	5
8	*"2.12.27"* (—), *"rijkswaterstaat"* (—), *"reisverslag"* (travel report), *"kaarten"* (maps), *"dagboek"* (diary), *"nederlandse kolonie suid america indianen"* (dutch colony south america indians), *"aanwinsten"* (acquisitions), *"overgave"* (surrender), *"paramaribo"* (—), *"van mook"* (—), *"2.12.10"* (—), *"kamer van koophandel"* (Chamber of Commerce), *"staatstoezicht volksgezondheid"* (state control public health)	13
7	*"koedijk"* (—), *"kadaster"* (cadastre), *"deli"*, *"politionele acties"* (police actions), *"1.04.02"* (—), *"geneeskundige dienst"* (medical service), *"voc"* (—)	7
6	not included to preserve space	19
5	not included to preserve space	45
4	not included to preserve space	76
3	not included to preserve space	194
2	not included to preserve space	528
1	not included to preserve space	2,512

associated topic, which indicates that people clicked once and they may have given up to search further with that topic, or that their search request directly matched with one EAD finding aid. The latter also adds support to the principle of provenance, but how many topics are fulfilled by one EAD finding aid?

To answer this question, we look at the mean CR values over all topics, with the click threshold set to > 10 clicks. For 2008, there is mean CR value of 0.70 ($N = 2,882$), and for January 2009, there is a mean CR value of 0.75 ($N = 403$). According to the CR levels, this is a medium concentration. There is a high concentration when a CR value is greater than 0.8. For 2008, there are 1,275 (44.2%) topics that comply with this condition. We see that for 2009, that this is the case for 208 (51.6%) topics. How many of the topics have a total concentration of an



(a) Distribution of CR values over 2008 ($N = 2,882$).



(b) Distribution of CR values over January 2009 ($N = 403$).

Figure 4.4: Distribution plots of the concentration ratios.

EAD finding aid, and thus a monopoly? We plot the distribution of the CR values over whole 2008 (see Figure 4.4(a)) and January 2009 (see Figure 4.4(b)). The distribution charts of Figure 4.4 also show that topics with a monopoly of an EAD finding aid stand out. So we find that provenance is effective for retrieval.

4.5.3 Study 2: Retrieval Effects of Provenance

How effective is it in terms of *retrieval* to group archival descriptions together? Does it improve the retrieval? We measure this and report the results.

Table 4.10: Retrieval performance on the fonds level for standard element relevance ranking and element ranking by provenance.

	Element	Element + Provenance	Fonds
MAP	0.0807	0.1060	0.2816
MRR	0.0887	0.1205	0.3196
nDCG	0.1629	0.1899	0.3765

Experimental Setup

This study uses the log-based test collection based on XML elements. We use the standard README approach as outlined in Chapter 2, but we test whether grouping relevant descriptions by fonds also helps to improve retrieval of a fonds. Concretely, we conduct a retrieval experiment that compares the *Individual Archival Material* (IAM) system, or 'Standard Element,' that retrieves elements according to relevance only with the *Archival Material in Context* (AMC) system, or 'Element + Provenance,' that also takes the principle of provenance into account when ranking the retrieved elements. We use all 3,423 topics in January 2009, and evaluate on the fonds level. The systems return the top 1,000 results. Additonally, we use the *Whole Fonds* system, which is a full-text system and where we retrieve the top 100 results. We index the collection with the Dutch snowball stemmer, and employ, like our previous experimentation, statistical language models (LM) with smoothing in PF/Tijah (Hiemstra et al., 2006). We use the IR evaluation measures MAP, MRR, and nDCG to compare the results.

Results

Next, the results, which are presented in Table 4.10. We see that standard element relevance ranking is the least performing run. Element ranking enhanced with provenance performs significantly better with an improvement of 31.35% ($t(3376)$=11.88, $p < 0.0005$). So we find that the principle of provenance helps to improve the retrieval of a fonds.

However, we note that full-text retrieval based on whole fonds ranking is the best performing system. This is due to a lower recall of the element retrieval-based approaches (2,699 out of 5,302) compared to fonds ranking (3,412 out of 5,280). This result also indicates that in order to improve the retrieval of a whole fonds, we have to use more of the whole fonds in our retrieval models.

4.6 Original Order Versus Relevance

We have probed the principle of provenance in the previous section. In this section, we investigate the retrieval effects of the corollary principle of original order. First, we introduce the background, then the experiment and results.

4.6.1 Background

EAD finding aids are structured in exactly the same way as the archival materials they describe Pitti (1999). First, by the previously investigated principle of provenance. Second, by the principle of original order, all material of the same creator is stored in its original organization and sequence. Archivists consider these principles crucial for archival access, though the concept of original order has been questioned by Boles (1982); MacNeil (2008); Meehan (2010), these have never been tested empirically (Lytle, 1980). Original order corresponds to the preservation of the document hierarchy in an archival description. Physical re-arrangement (such as re-ordering by topic, time or geography), which could enhance user access to the archives, is rejected (e.g. Jenkinson, 1944; Ketelaar, 1996; Lytle, 1980). Specifically, as Lytle (1980) states, even the re-arrangement of archives to suit the needs of historians is disallowed. In recent years, however, archival finding aids in EAD have been put online, giving it a new function as an information retrieval and discovery tool for users. MacNeil (2008, p.24) suggests that original order is "one of many possible orders a body of records will have over time and, therefore, its privileged status needs to be reconsidered." Meehan (2010, p.34) proposes to rethink original order as a conceptual framework "by focusing more on the spirit behind the principle of respect for original order, rather than trying (and failing) to follow it to the letter." Therefore, the actual impact of following the original order of an archive to the letter, when retrieving and presenting information, needs to be re-examined.

Archives may span 100s or 1000s of meters (or yards) of material, and the main purpose of the archival description is to help searchers identify the exact parts of the archive for consultation. There is a direct and natural parallel between locating parts of the archival finding aids in EAD, and the focused access of other XML documents: XML retrieval techniques can be used to exploit the internal structure of an EAD finding aid. This structure could consist of elements that represent lengthy biographies, nested components, all the way down to the single item. However, each and any of these elements can be returned in any order, either by respecting the original order, or returning them according to relevance only, or any other criterion. The retrieval effects of original order are not known, so *given the retrieval of any and arbitrary EAD elements according to the relevance with a query, what are the retrieval effects of returning it in original order?*

On the one hand, original order could improve retrieval as relevant items may appear close to each other due to the intellectual organization by the archival creator. The would correspond with the Cluster Hypothesis, which deposits that closely associated documents tend to be relevant to the same requests (Jardine and van Rijsbergen, 1971, p.219). On the other hand, it may not be a useful feature to improve retrieval because the Cluster Hypothesis may not hold on archival finding aids in EAD.

Table 4.11: Retrieval Performance for first top N results for each topic in terms of relevance (R) and original order (O).

	Top 1,000		Top 500		Top 100		Top 10		Top 5	
	P@10	**MAP**	**P@10**	**MAP**	**P@10**	**MAP**	**P@10**	**MAP**	**P@10**	**MAP**
R	0.1600	0.1454	0.1600	0.1398	0.1600	0.1321	0.1600	0.0917	0.1400	0.0822
O	0.0000	0.0296	0.0000	0.0260	0.0400	0.0517	0.1600	0.0660	0.1400	0.1031

4.6.2 Study 1: Using Manual Assessments Within a Fonds

Experimental Setup

The search requests are reference questions, created with the README Assessment System (see Figure 4.1), resulting in the topics of the 'manual' test collection in Figure 4.2. The used queries consist from three up-to thirteen keywords. Each of these reference questions was judged against a narrative in which the information need is clearly stated, including what is considered relevant. The relevance is determined by locating particular units of archival materials that will likely contain the sought answer. In practice, the descriptions of boxes (folders) and individual files tend to be very succinct—seldom more than a single sentence. Additionally, a finding aid in EAD also contains contextual background descriptions of the archive, which may directly contain relevant information.

The system used in our experimentation has been described in Chapter 2. We indexed the collection without stopword removal, used the Dutch snowball stemmer, and standard parameters. For the retrieval of any arbitrary elements, without element overlap removal (Kazai et al., 2004), we employ statistical language models with smoothing (LM) as explained by Ponte and Croft (1998), i.e. the probability distribution of all possible term sequences is estimated by applying statistical estimation techniques.

Results

Relevance Versus Original Order

We first look at the whole run with 1,000 results in Table 4.11. In the top 1,000 results, both approaches obtain a reasonable recall of 62 out of 73 relevant elements ($R = 0.8493$). Considering that we look for very short descriptions (often a single sentence), the relevance ranking is performing quite well with a MAP of 0.1454 and a P@10 of 0.1600. What happens if we rank these 1,000 results in their original order? The score plummets to almost zero; the relevance ranking is crucial. It should be noted that we are re-ranking the top N of results, and usually there are just a handful of relevant results (also see Table 4.12).

Given that set of results must contain many non-relevant ones, re-ranking the top 1,000 on original order may not fairly reflect the utility of the original

TABLE 4.12: DOM Tree distances in the manual assessed qrels.

EAD	Count	Mean Depth	Mean Distance	Total count <C$_n$>
1.04.02	11	10.000	3.545	17,184
2.19.123	37	10.811	1.492	5,661
2.19.124	7	6.000	1.286	1,491
2.03.01	7	10.143	1.429	14,017
2.21.286	4	9.750	0.750	2,035

order. What will happen if we re-rank a smaller set of results? The remaining columns of Table 4.11 show the results for different sets. The MAP of the relevance ranking drops as expected for the shorter runs. As the cut-off level is decreased, we see that the precision of the original order ranking increases. Interestingly, we see that the MAP for Original Order is higher than the standard element ranking when the cut-off level for each topic is set to 5. This signals that although the relevance ranking is of paramount importance, there is also still potential value in the original order, because relevant results have a tendency to cluster.

Cluster Hypothesis Effects

We want to further investigate the Cluster Hypothesis—how near are the relevant results in the original order of the document? We do this by measuring the distance between the relevant elements in the DOM tree. We restrict our attention to the relevant elements in the descriptive subordinate components <c$_n$> elements. The components <c$_n$> are nested within each other in <archdesc> given the n, where $n \in \{01, ..., 12\}$. A component can also be unnumbered.

The results are shown in Table 4.12. The first topic has 11 results with a mean depth of 10 nodes in the DOM tree. For each pair of results, we look at the distance to a common ancestor, which could be at most the depth itself (i.e., 10). For the first topic the mean distance over all pairs is 3.5 – which signal that the results are somewhat scattered through the archive. However, for the other topics the mean distance is between 0.75 and 1.5, which shows that relevant results occur in close proximity within the archive, especially given the large quantity of <C$_n$> elements per topic (see Table 4.12). For example, in topic *1.04.02* only 11 out of 17,184 (or 0.06%) <C$_n$> elements were seen as relevant.

Sparse Data on the Item Level

A challenge for effective XML retrieval for EAD finding aids is the sparse data, especially in the unit titles, which is a distinct property of the archival descriptions. When we analyze the selected relevant elements, we see that there are very short phrases, sometimes without the occurrence of a keyword, e.g. "*Diverse stukken*, (Unit ID: 824)" (in English: "*Several pieces,*"). The sparse data on the item level can be attributed to concept of *inheritance of description*—each lower level inherits the description of the container (Pitti, 1997), where this context is a crucial cue for assessing relevancy as related relevant items tend to be located in short distance from each other.

The inheritance of description is also an argument for original order, since changing the hierarchy of descriptions may change the context. Or, changing the order might make multilevel description more powerful. In any case, after a search system returns results in the hit list, the user gets a first grasp of what can be found. For the first glimpse, when people assess relevancy or usefulness in the hit list, original order has less value.

4.6.3 Study 2: Experiment with Log-Based Approach

Experimental Setup

We repeat the previous experiment with the XML element test collection based on the log-based approach, where we retrieve all relevant elements, thus without element overlap removal, and we only focus on `<c>` within the inventory (subordinate components). We select topics and a single associated dominant EAD finding aid, where the CR value is near or equal to 1, which means a total concentration. Moreover, we select topics with the largest number of clicks (thus with a broadest set of clicked elements). To align our content-oriented systems (retrieving mostly `<unittitle>`) with the table-of-concent-centric *Nationaal Archief* clicks (retrieving mostly `<head>`), we ignore the following elements (see Tables 4.1 and 4.2) when they are a leaf node (see our explanation in Chapter 2) in both our runs and the log-based test collection: `<unittitle>`, `<did>`, `<head>`, `<note>`, `<corpname>`, `<geogname>`, `<p>`, and `<entry>`. Table 4.13 depicts the top 10 topics, with the original Dutch query used and the translation to English.

Results

Table 4.14 presents the outcome of the retrieval performance for the first top N results for each topic. We see that element relevance ranking outperforms original order in terms of P@10 and MAP, and when looking at the top 1,000 results for each topic, we note that for the MAP score, there is a 5.97% difference between relevance ranking and original order, though this difference is not statistically significant. When we decrease the cut-off level, we see that the MAP scores drop, for both types of runs, because the runs get shorter and the recall decreases. When we set the cut-off level to 5, we see that the P@10 and MAP scores for relevance ranking and order order become equal, with a slight better MAP performance for original order, which also may due to cluster effects.

Table 4.13: Log-based topics with a (near) total concentration of a single EAD finding aid, ordered by the number of (different) clicks.

# Clicks	CR	Topic	EAD
80	0.96	*radiodistributie* (cable radio)	2.16.81.11
74	0.97	*filmkeuring* (motion picture rating)	2.04.60
70	0.92	*kamer van koophandel rotterdam* (Chamber of Commerce Rotterdam)	3.17.17.04
66	1	*centrale vereniging* (central association)	2.19.093
65	1	*orde van de nederlandse leeuw* (Order of the Netherlands Lion)	2.02.32
59	1	*soestdijk en marechaussee* (Soestdijk and military constabulary)	2.04.87
59	0.98	*nbw* (—)	2.09.75
58	1	*snouckaert van schauburg* (—)	1.10.76
50	1	*noordwijkerhout dopen kerken* (Noordwijkerhout baptize churches)	3.18.80
46	1	*volkenkunde* (ethnology)	3.12.16

Table 4.14: Retrieval Performance for first top N results for each topic in terms of relevance (R) and original order (O).

	Top 1,000		Top 500		Top 100		Top 10		Top 5	
	P@10	MAP	P@10	MAP	P@10	MAP	P@10	MAP	P@10	MAP
R	0.0889	0.0817	0.0889	0.0814	0.0889	0.0718	0.0889	0.0212	0.0444	0.0113
O	0.0778	0.0771	0.0778	0.0770	0.0778	0.0668	0.0778	0.0171	0.0444	0.0114

4.7 Discussion

Cox (2008) recalls that in the early 1970s, there was little discussion about descriptive standards, and offers a reflection about the progress made or lessons learned. Indeed, archivists have developed ideas to improve the arrangement and description of, and access to, archives by challenging existing ideas which may be unnecessary complicated or have limitations, while sticking to the archival principles. For example, in 1966, Scott (1966) argues to abandon the record group concept, which is the dominant American view to arrange archives. There is a paradox with this arrangement, as it is based on the principle of provenance, but its application sometimes leads to violations of this principle and of the principle of original order, because the organizations where this arrangement is based on may change. Instead, Scott (1966) proposes to arrange archives on the series level instead, thus removing a layer in the archival hierarchy. As organizations may change, it makes it possible to link records to multiple organizations on this level, whereas on a fonds level this is not possible. For persons and families, Scott (1966, p.502) proposes to use registers to index persons by family or organization, while linking to the archive on the series level.

Evans (1986) goes a step further by calling for a paradigm shift, and argued for the need of authority control, so separating the information describing records and the information describing agencies in the representation by linking the provenance data to the series description data. Evans (1986, p.255) continues by stating that "applying the record group concept to finding aids produces static, out-of-date inventories that provide access to records only through a single, hierarchical path." As Evans (1986) argues, authority control is still based on provenance, as it still groups archival records, but by linking them to an authority record. Evans (1986, pp.255–259) thinks this is superior to one based on the record group concept. Bearman and Lytle (1985) have made a similar argument: an authority system is more flexible in terms of access compared to the hierarchical data model. Therefore, archivists have created *Encoded Archival Context* (EAC, Szary (2005)), and have begun with its realization (Ottosson, 2005; Vitali, 2005). It may facilitate inferential search.

Before the emergence of archival finding aids in EAD, Campbell (1967) has pointed to IR concepts for archives to understand content search. Provenance is an inexpensive way to arrange archives (Campbell, 1967, p.280). Another common and inexpensive arrangement is by date, but for a user who is subject-oriented, arrangement by date without index is frustrating, and arrangements can be combined, such as chronologically arranging archival materials with a sub-arrangement by name (Campbell, 1967, p.280). Archives are also arranged by subject (thesaurus), but this is difficult, because there is a one-to-many relationship between an item and a subject (Campbell, 1967, p.280). Therefore, Campbell (1967) believes that indexing using a controlled vocabulary (and not arrangement) is the key to information retrieval in archives, browsing is not enough as it is an inefficient method to find specific data, and in terms of ranking, Boolean search is mentioned. Nowadays, there are many other options to improve retrieval, which could be explored in the future.

4.8 Conclusion

This chapter has investigated the impact of archival principles—provenance and original order—on the retrieval of archival descriptions. We have started our investigation by looking at what EAD elements people use. We found that this depends on the user's information need, since there is not a clear favorite element that people use. When people are searching for contextual information, the `<bioghist>` is a 'popular' element. When people search for archival materials, the higher level `<c>` elements stand out. This suggests that indexing and retrieving all and any elements, as the README system does, is the most practical solution for a uniform archival search system. This allows for further personalization.

We have investigated the effectiveness of retrieval with provenance in an analysis by looking at the number of clicks that occur within a certain EAD finding aid given a specific search request, and by comparing the grouping of elements by a fonds with standard element relevance ranking in a retrieval experiment. In our analysis, we borrowed an economic term called the concentration ratio, used to measure the dominance of a company in an industry. Similarly, the principle of provenance states that a fonds is an organic whole, and since an EAD finding aid is a representation of a fonds, its consequence is that one EAD finding aid should also have a 'monopoly' for a particular search request. We found that for topics with more than 10 clicks, there is a medium-to-high concentration level. We also found that people search across EAD finding aids, but our analysis and experiment points to evidence that the *provenance* is effective for *retrieval*.

We empirically examined the impact of the archival principle of original order on the ranking of search results by comparing it with a standard archival retrieval system using modern XML retrieval techniques. Our results show that the relevance ranking is of paramount importance, but that the results have a (weak) tendency to cluster. Original order is useful, because physical materials can only be ordered in a single way, and here the traditional archival practices make much sense. With the advent of EAD finding aids, we are no longer bound to the physical and practical limitation of before and we could construct multiple ordering of the same material including those based on a search request or search profile at hand. This opens up a wealth of possibilities to change archival access with archival finding aids in EAD, enabling new and more effective usages of archival description, but will this improve access?

Chapter 5

Searching Within EAD Finding Aids

In Chapter 4, we have conducted experiments with XML retrieval evaluation. We still do not know what users do when they search in EAD finding aids. We use the dataset consisting of clicks within archival finding aids in EAD to observe the information searching behaviors of users within EAD finding aids.

5.1 Introduction

Wilson (1999, p.263) suggests that information behavior may be seen as a series of nested fields, "with information-seeking behavior being seen as a sub-set of the field, particularly concerned with the variety of methods people employ to discover, and gain access to information resources, and information searching behavior being defined as a sub-set of information-seeking, particularly concerned with the interactions between information user (with or without an intermediary) and computer-based information systems, of which information retrieval systems for textual data may be seen as one type." Wilson (1999, p.267) thinks that one might best focus on projects that take a view of information searching as a complex process embedded in the broader perspective of information-seeking behavior, and information behavior in general, rather than on the micro-level of analysis that is typical of IR research.[1]

[1] Belkin et al. (1982, p.62) write that the performance measurement of IR systems seem to be limited to marginal gains in terms of complete precision, recall, or complete user satisfaction. Instead, Belkin et al. (1982) propose a 'radically different' hypothesis that resets traditional IR evaluation (e.g. Spärck Jones (1981)) by focusing on the beginning of search, i.e. the information need but out of an inadequate state of knowledge. The expression of an information need is a statement of what the user does not know (Belkin et al., 1982, p.64). The user's state of knowledge could be dynamic rather than static, changing as he or she proceeds in the search process (Kuhlthau, 1991, p.362). This is a cognitive model that Belkin et al. (1982) call anomalous states of knowledge (or ASK). Therefore, Belkin (1993) states that information seeking behaviors are interactions with texts. The postulations of the ASK model (Belkin et al., 1982) suggest that IR is most properly considered as a form of information seeking behavior, in which the user's interaction with text is the central phenomenon, supported by the rest of the IR sys-

The study of Hodkinson et al. (2000) on consumers' search behavior on the WWW is an example. Through diagrammatic illustrations of different search scenarios, concise insight is given into the search behavior of consumers. First, this allows for examination of individual search styles, and second, it provides an overview for ready comparison between expert searchers and novices. However, Hodkinson et al. (2000) point to limitations of using their diagrammatic approach, like the labor-intensive nature of the observations necessary to generate the state diagrams and attached tables.[2] Their diagrams also do not incorporate the use of within-site search engines.

We can study the search behaviors of users in terms of a state diagram. EAD finding aids are devices that can also be formally represented in a state diagram. Methodologically, by gaining more understanding of search behavior, we can improve the process of searching for information (Belkin, 1993; Ingwersen and Järvelin, 2005). In the archival domain, research to gain more understanding of interaction with archival descriptions—the first step required to assess IR with EAD—is scant. Prom (2004) presents a study that measured user interactions with online archival finding aids in EAD in a controlled setting. That study deals with assessing how people search on the collection-level (fonds) and folder-level (files) using nine different user interfaces. The findings put emphasis on the implications for Web site design for finding aids, but also touches upon search behaviors, and differences among archival users. Although it points to the importance of searching within a fonds, and illustrates search behavior with a few examples, it does not quantitatively define or systematically detail different within-fonds search behaviors in terms of formal state transitions. This leads to the main research question of this chapter.

- *Q4: How do we formally identify people's search behaviors with archival descriptions?*

To investigate this research question, the search logs of the *Nationaal Archief* (NA) are used. Methodologically, we combine the study of a Web search logs with experiments. There are three sub-questions that deal with the search behavior of users when interacting with archival finding aids. The first subquestion delves into comparing disjoint sets of interaction data that represent

tem components. This is dubbed as interactive information retrieval (IIR, also see Kelly (2009)). Likewise, Ingwersen and Järvelin (2005) argue for synthesizing information seeking and retrieval (IS&R) research ranging from systems oriented laboratory IR research to social science oriented information seeking studies.

[2] Formally, a state diagram is a labeled directed graph or a *finite state machine*, which demonstrate the behavior of a device in a formal notation (Kieras and Polson, 1985), who also call it a transition network representation. Such a representation consists of two components. First, a series of nodes, which represent the states that a system can be in. Second, a series of labeled arcs, which represent ways in which a system can make a transition from one state to the other, that interconnect these nodes (Kieras and Polson, 1985, p.380).

different user groups. Specifically, the comparison involves the search behavior of first-time users and users with prior experience in using the NA system.

4.1 Do people who start to use an EAD finding aid system, also click (a) less often, (b) slower, (c) less deep, and (d) less broad in finding aids than people with prior experience?

The hypothesis of the first sub-question is that people who are not yet proficient with archival finding aids use them less exhaustive—in every aspect—compared to frequent users. This is also related to the principle of least effort (Zipf, 1949), which implies that the information seeking behavior stops as soon as minimally acceptable results are found. Would this be true as well for first-time users interacting with EAD finding aids?

The second sub-question looks into the changing search behavior of users who gain more search experience with EAD finding aids.

4.2 What happens when users start to interact more and repeatedly with EAD finding aids?

The expectation for the second sub-question is that users start with little interaction, but as users start to come back and re-use the *Nationaal Archief* system, they will search more, and will have more interaction with the finding aids. And if this is the case, the user search profile will become more similar to a "known user" profile.

The final sub-question deals with researching the common ground that may exist among users who interact with the same EAD finding aid.

4.3 Do different people interact differently with the archival descriptions of the same EAD finding aid?

The conjectures of the third sub-question are that the majority of the people interact little with the finding aids, and a minority use them exhaustively both in terms of the depth and breadth of a finding aid. More popular used finding aids may illustrate the full potential of EAD finding aids as knowledge—or even evidence—discovery tools.

5.2 Related Work

This section presents a comprehensive, but concise as possible overview of search behavior on the World Wide Web, and search behavior in cultural heritage with humanities resources.

5.2.1 Deriving Web Search Behavior from Search Logs

We can study Web search behavior by looking at search logs of Web sites. Usage data in the raw Web server log can disclose the information seeking behavior, for example of virtual scholars (Nicholas et al., 2006), by associating the usage data with search/navigational and/or user demographic data. Jansen and Spink (2006) compare nine studies on Web search logs of five Web search engines by looking at interactions occurring between users and Web search engines from the perspectives of session length, query length, query complexity, and content viewed. A finding is that the wide spread use of Web search engines, employment of simple queries, and decreased viewing of result pages may have resulted from algorithmic enhancements of search engines. A transaction search log analysis can also be domain-specific, and tells us more on how user groups with a particular interest search. Jansen et al. (2010) provide a longitudinal study on people interacting with Web search engines when searching for religious information. The results of the analysis suggest religion is a persistent topic of Web searching as that there is no evidence of a decrease in religious Web searching behaviors. A study in the medical domain also uses a log file to analyze search behavior of people interacting with the PubMed database (Islamaj Dogan et al., 2009).

Wolfram (2008) also studies search characteristics to see whether users engage in different behaviors in different types of Web-based IR environments, which are here a bibliographic databank, an online public access catalog (OPAC), search engine, and a specialized search system. Similar to previous search log-based studies (Jansen, 2006; Jansen and Spink, 2006), the search characteristics of Wolfram (2008) are descriptive statistics and relative frequency distributions related to term usage, query formulation, and session duration. The results reveal that there are differences in search characteristics. Users were more likely to engage in extensive searching using the OPAC and specialized search system. The bibliographic databank search environment resulted in the most parsimonious searching, more similar to a general search engine. So the systems may appear similar here, but users do engage in different search behaviors. White and Dumais (2009) investigate the switching behavior and examine features to predict when people switch to another search engine. White and Drucker (2007) describe a longitudinal log-based study investigating the levels of behavioral variability in users engaged in Web search activities by looking at their search trails. Their findings suggest that there are large differences in variability in the search interaction within and between users, and they identify classes of users, namely navigators and explorers, whose interaction is highly consistent or highly variable.

Aula et al. (2010) examine how search behavior changes as search becomes more difficult by studying the behavioral signals that suggest that a user is having trouble in a search task. These behavior signals have been logged in search log files in order to provide quantitative support for their hypotheses. Findings are that when users face difficult tasks, they start to formulate more diverse queries, they use advanced operators more, and they spend a longer time on the search result page as compared to the successful tasks. The implication of this research is that the behavior signals can be used to predict the user satisfaction in a session, to gain a better understanding of how often users leave search engines unhappy, or how often they are frustrated and in need of help, and perhaps an intervention, at some point during the search session. Vakkari et al. (2003) investigate the subjects' growing understanding of the topic and search experience. By analyzing a search log, Vakkari et al. (2003) find that with more search experience, there is an increased use of specific terms between the sessions, but the use of search tactics and operators remained constant.

5.2.2 Search Behavior in Cultural Heritage

Besides on the Web, there can be a breakdown of the search behaviors of users in a particular domain, such as in cultural heritage (CH). Cultural heritage is defined as the legacy of physical artifacts and intangible attributes of a group or society that are inherited from past generations—widely viewed as a precious and irreplaceable resources—maintained in the present and bestowed for the benefit of future generations because it is seen as essential to the personal and collective identity and necessary for self-respect (Lowenthal, 2005).

Skov and Ingwersen (2008) explore the information seeking behavior in a digital museum context by deriving the information needs from leisure tasks or interests, and the main characteristics of virtual museum visitors' information seeking behavior. This study indicates a broad coverage of different types of needs, and that the information seeking behavior is highly task-dependent. Moreover, there are four different information seeking behaviors of virtual museum visitors. These are (i) highly visual experience, where photos are the most important feature, (ii) meaning making, where implicit connections are made explicit, (iii) known item searching, where more information is sought on known items, and (iv) exploratory behavior, which includes looking for the missing piece in the puzzle and where high recall is important and low precision is accepted. The exploratory research of Zach (2005) investigates how arts administrators find their information, when they stop searching, resulting in a model for their information behavior.

Focusing on a particular user type is the research of Borgman et al. (1995), which investigates children's information searching abilities with OPACs, to understand how IR systems for children should be designed. They set up a longitudinal study conducted over a three-year period at three sites, with four databases, with comparisons to two different keyword online catalogs, and with children participating in each of the four experiments.

In 1980, Lytle (1980) claimed that the reason archivists lagged behind librarians in their study of search behavior may have been the archivists' resistance to social and behavioral science techniques, especially those applied in library and information services (Lytle, 1980, p.70). Lytle also suggested that archival users may be more difficult than library users to observe, since their research needs are difficult to assess, and their needs are diffuse, and the users are unaccustomed to formulate their information needs. In a related user group, Barrett (2005, p.326) also notes that humanities faculty researchers have been described as slow, haphazard, and serendipitous in their search for information. In the humanities in general, the information-seeking behavior of graduate student researchers is described as an idiosyncratic process of constant reading, 'digging,' searching, following leads, hence citation chasing was by far the most frequently described method of finding materials (Barrett, 2005, p.326).

Following up on Lytle's claims, the article of Duff and Johnson (2002) reports on the information-seeking behavior of historians based on semi-structured interviews with ten mid-career historians. The research focuses on how historians locate primary sources, carry out their research, and use archival material. The results show that there are four different types of information-seeking activities, including (i) orienting oneself to archives, finding aids, sources, or a collection; (ii) seeking known material; (iii) building contextual knowledge; and (iv) identifying relevant material. Duff and Johnson (2002, p.492) make a direct link between (iii) and (iv), i.e. acquiring more contextual knowledge means improving the ability to identify relevant material. Although historians often speak about the role of serendipity in their discovery of relevant material, their findings suggest that this process is influenced less by serendipity and more by the deliberate tactics of the expert researcher, so Duff and Johnson (2002, pp.494–495) conclude that what appears to be accidental discovery is actually accidentally found on purpose.

Anderson (2004); Tibbo (2003) study historians' searching behaviors through surveys and interviews. These studies focus on different approaches to search in general, and the main behaviors include e-mailing, telephoning or writing a letter for a copy of a finding aid, primary materials, or assistance of an archivist.

More recently, Prom (2011) shows in the archival domain that Web analytics can be used as method to measure user actions, to understand some aspects of user behavior, and to subsequently improve online services. This dissertation, including this chapter, complements and advances this assertion.

5.3 Analysis of Searching Within an EAD Finding Aid

We can use these clicks to analyse navigational transitions of users who interacted with the NA system (2007 to 2009). A transition (or arc) is an action occurring between two states (or nodes). First, we conduct a first exploration. Second, we identify all the different possible states. Third, we present a transition diagram that captures the search behaviors possible in EAD finding aids.

5.3.1 A First Exploration

An example of a search episode recorded in the log is presented in Figure 5.1, which depicts 30 clicks in session '1027' created by a certain UUID on 2009-

01-05. Each line shows the time a click was recorded, the type of view (i.e. summary or page view), the URL with the filename of the EAD finding aid that the user clicked on, the pageID (or entry point), the textual value of this pageID (anchor text), and its XPath reference.

How can we interpret this data? We see on line 1 that someone clicks on the summary view of EAD finding aid '1.11.01.01,' and more than ten minutes later, clicks on the page view of the finding aid '1.01.01.10.' In case there is no pageID available, the user gets linked to the top (beginning) of a finding aid. On line 3, we see that the user returns to the previous finding aid. Between line 3 and 6, we observe a sequence of interaction within that finding aid. Line 4 shows that a user clicks on a pageID, and gets linked to the Description of Subordinate Components `<dsc>`, or in Dutch, the *"Beschrijving van de series en archiefbestanddelen."* Line 5 shows that a user moves deeper by clicking on the first of the subordinate component `<c01>`. Line 6 illustrates that a user navigates to the first component of `<c01>`, which is `<c02>`. Searching in this finding aid is then halted. However, line 21 displays that this finding aid is re-visited. Between line 22 and 24, we see a repeat of the previously described search behavior. However, line 25 shows that the navigation consists of going one step more in breadth by clicking on the second subordinate component of `<c02>`, namely `c02[2]` that represents *"A2 Losse stukken"* (or in English, *"A2 Separate pieces"*). Line 26 unveils that the user navigates one more component deeper, the first subordinate component of `<c02>`, namely `c03[1]`. Interestingly, the last click in this session exhibits a very deep click on a finding aid previously not seen in this session. This click refers the user to the eighth subordinate component `<c08>`, and has a XML tree depth of 12. Given that this clicks occurs more than twenty minutes later than the previous click, it could be that the user clicked directly using a link provided from another source. Finally, the search behavior in this session only consists of top-down click sequences, where a user starts from the very beginning at the top of the file, and gradually moves deeper and also more in breadth.

In our analysis, we focus on one month of log data recorded in January of 2009. There are 9,389 sessions consisting of at least one click, created by 7,141 UUIDs in a month time. There are in total 31,506 clicks in 3,139 different types of EAD finding aids. Given this amount of data, it is not feasable to manually interpret these clicks. How can we automate this analysis?

5.3.2 The States of Within-EAD Finding Aid Search

A state diagram is used to describe systems or devices, but it can also be used to describe the search behavior of users interacting with a system (Hodkinson et al., 2000). We introduce our *EAD Search Behavior model*, which is based on a state diagram. First, we identify the states of searching in EAD finding aids.

Our ground truth is the official EAD DTD of 2002, also see our represen-

```
08:24:26 summary.webview?eadid=1.11.01.01 - "-" /EAD[1]
08:34:52 page.webview?eadid=1.01.01.10 - "-" /EAD[1]
08:35:03 page.webview?eadid=1.01.01.01 - "-" /EAD[1]
08:35:06 page.webview?eadid=1.01.01.01 N1015C "Beschrijving van de series en
    archiefbestanddelen" /EAD[1]/ARCHDESC[1]/DSC[1]/HEAD[1]
08:35:07 page.webview?eadid=1.01.01.01 N10160 "A Archief van de griffie"
    /EAD[1]/ARCHDESC[1]/DSC[1]/C01[1]/DID[1]
08:35:16 page.webview?eadid=1.01.01.01 N10168 "A1 delen en banden"
    /EAD[1]/ARCHDESC[1]/DSC[1]/C01[1]/C02[1]/DID[1]
08:38:32 page.webview?eadid=2.11.63 - "-" /EAD[1]
08:38:34 page.webview?eadid=4.VEL - "-" /EAD[1]
08:38:36 page.webview?eadid=2.13.25 - "-" /EAD[1]
08:38:42 page.webview?eadid=2.11.63 - "-" /EAD[1]
08:41:24 page.webview?eadid=2.13.25 - "-" /EAD[1]
08:47:29 page.webview?eadid=1.01.01.10 - "-" /EAD[1]
08:47:32 page.webview?eadid=2.04.42 - "-" /EAD[1]
08:48:42 page.webview?eadid=1.01.01.09 - "-" /EAD[1]
08:49:09 page.webview?eadid=1.01.01.02 - "-" /EAD[1]
08:49:12 page.webview?eadid=1.01.01.02 N10120 "Beschrijving van de series en
    archiefbestanddelen" /EAD[1]/ARCHDESC[1]/DSC[1]/HEAD[1]
08:49:22 page.webview?eadid=1.01.01.02 N10124 "A Archief van de audience en de
     secretarie" /EAD[1]/ARCHDESC[1]/DSC[1]/C01[1]/DID[1]
08:50:59 page.webview?eadid=1.01.01.02 N10156 "B Archief van de Hoogduitse
    secretarie" /EAD[1]/ARCHDESC[1]/DSC[1]/C01[2]/DID[1]
08:51:14 page.webview?eadid=1.01.01.02 - "-" /EAD[1]
08:51:17 page.webview?eadid=1.01.01.02 N10120 "Beschrijving van de series en
    archiefbestanddelen" /EAD[1]/ARCHDESC[1]/DSC[1]/HEAD[1]
08:51:32 page.webview?eadid=1.01.01.01 - "-" /EAD[1]
08:51:34 page.webview?eadid=1.01.01.01 N1015C "Beschrijving van de series en
    archiefbestanddelen" /EAD[1]/ARCHDESC[1]/DSC[1]/HEAD[1]
08:51:35 page.webview?eadid=1.01.01.01 N10160 "A Archief van de griffie"
    /EAD[1]/ARCHDESC[1]/DSC[1]/C01[1]/DID[1]
08:51:39 page.webview?eadid=1.01.01.01 N10168 "A1 delen en banden"
    /EAD[1]/ARCHDESC[1]/DSC[1]/C01[1]/C02[1]/DID[1]
08:51:43 page.webview?eadid=1.01.01.01 N101AE "A2 Losse stukken"
    /EAD[1]/ARCHDESC[1]/DSC[1]/C01[1]/C02[2]/DID[1]
08:52:09 page.webview?eadid=1.01.01.01 N101B6 "A.2.1 1576 september - 1581
    augustus" /EAD[1]/ARCHDESC[1]/DSC[1]/C01[1]/C02[2]/C03[1]/DID[1]
08:57:39 summary.webview?eadid=2.10.50 - "-" /EAD[1]
09:13:02 summary.webview?eadid=2.21.005.26 - "-" /EAD[1]
09:32:15 summary.webview?eadid=2.15.43 - "-" /EAD[1]
09:54:28 page.webview?eadid=3.17.17.04 N16B9C "2.1.1.15.4.1 Algemeen"
    /EAD[1]/ARCHDESC[1]/DSC[1]/C01[1]/C02[2]/C03[2]/C04[1]/C05[7]/C06[15]/C07[4]/C08[1]/DID[1]
```

Figure 5.1: Post-processed source code of the search log session ID '1027' by UUID '2b844140ef7cfd438300da7ec6278de0' on 2009-01-05. The time-stamp of each click is depicted in boldface black, the EAD finding aid file in red, the ID of the HTML anchor value in blue, and the mapped XML element path in green.

tation of EAD as a mono-hierarchical tree in Figure 2.4 of Chapter 2 on page 17. The XML tree data-structure enables efficient search in XML documents. The tree structure is strictly hierarchical, where each node has only one parent (except the root), but a node can have an infinite amount of children. Our model attributes the search behaviors within EAD finding aids to the tree data-structure and the EAD model. The searching starts usually at the beginning of the file, or the root node `<ead>`, then the user has one option: go deeper in the hierarchy. It is also possible that a user starts searching within the file, or has arrived at this point, then the user can also go back in the hierarchy. The navigation lines up with going forward or backtrack through the node sets (group of elements in the same hierarchy)—which we call the *depth*. Or moving horizontally (up or down) to sibling nodes—which we call the *breadth*. How can we formalize and capture the search behavior in archival finding aids?

A state diagram is a directed graph that consists of different *states* and transitions. The states S are a finite set of vertices and labeled with unique designator symbols or words written inside them. In theory, any XML element can be a state. However, from a high-level view, we can partition a finding aid into different components. EADs in XML are hierarchical graphs. Consequently, we align EAD to a directed graph by labeling certain elements in the hierarchy as a state. The EAD DTD makes clear that there are three main components in root `<ead>` (label: $(S1)$): `<frontmatter>`, `<eadheader>` (label: $(S2)$), and `<archdesc>` (label: $(S3)$). The `<frontmatter>` is not a mandatory element, and in practice not used by archives, so we discard it. Moreover, the `<archdesc>` consists of two parts: the descriptive elements of the *introductory information* in `archdesc/did` (label: $(S4)$), and the *inventory* consisting of descriptions of subordinate components in `archdesc/dsc` (label: $(S5a)$). Let us zoom on each separately.

Introductory information When someone moves to the `archdesc/did` hierarchy, the so-called 'descriptive identification element declarations' encountered give contextual background information about the archive. These elements include the `<bioghist>` that presents contextual information on the creator of the archive, `<scopecontent>` that contains the scope and content of the archive, `<abstract>` that describes the summary, etc. In our model, we regard all elements in `archdesc/did` as an entity, and one type of navigation, thus as one state.

Inventory When a person moves to the `<dsc>` element, he or she is faced in particular with a complexly nested list of descriptions, illustrating a strong case of searching with depth and breadth. Our model captures this search behavior by analyzing the <c_n> elements, where $n = \{1, ..., 12\}$. In State $(S5a)$, a person can start at the beginning of the list in `dsc/head`, or move directly to a subordinate component <c_n> in State $(S6a)$. A user can go $n - k$ component higher (label: $(S7a)$), or $n + k$ component lower (label: $(S8a)$), where k stands for the number of levels. Additionally, people can navigate in the finding aid in terms of breadth b, so one can move up $b-p$ 'left' (e.g. label: $(S6b)$), or one can move down in the same node set $b+p$, or 'right' (e.g. label: $(S6c)$), where p is the element position.

5.3.3 Within-EAD Finding Aid Search Transition Diagram

Finally, connecting the dots. Figure 5.2 depicts our inclusive search behavior model for searching in EAD finding aids. We formally define a transition of n instances as a two-argument function $t_n(argv1, argv2)$, where $argv1$ is the begin state s_{begin}, and $argv2$ is the end state $s2_{end}$, so a transition is $t(s_{begin}, s_{end})$. A transition is also called an *edge* δ, so $\delta : \Sigma \times S \rightarrow S$, where Σ is a finite collection of input symbols or designators. Therefore, we can also express this notation as $argv1 \rightarrow argv2 = t_n$. Given the states, there are twenty-two different types of transitions: $t1((S1),(S1))$ is a user refreshing the view on a whole finding aid; $t2((S1),(S2))$ going from the root to the `<eadheader>`; $t3((S1),(S3))$ going from the root to `<archdesc>`; and etcetera. In practice with the *Nationaal Archief* system, $t2$ is not occurring, because `<eadheader>` is not enabled as entry point to the finding aid. In fact, all clicks in EAD finding aids occur as $t3$. Therefore, we treat State (S3) as State (S1), so $t4((S3),(S4)) = t4((S1),(S4))$. Since not all clicks have a pageID (XPath reference), these are labeled by default as (S1).

In our analysis, we focus on the navigation in the descriptions of subordinate components via $t4((S1),(S4))$, $t5((S1),(S5a))$, and $t6((S4),(S5a))$ in particular. It is possible to move from `archdesc/did` to `archdesc/dsc`, $(S4) \rightarrow (S5a) = \{t6\}$, or directly $(S4) \rightarrow (S6a) = \{t6, t9\}$, and vice versa.

Additionally, we investigate the effects of the variables depth and breadth. Our model expresses a move in the `<dsc>` hierarchy as a combination of both types of transitions. The begin state is always represented (reset) as State (S6a), and the notation for the end state is dependent on this begin state. When a user stays in the same state, there is one state $(S6a) \rightarrow (S6a) = \{t11\}$. In case a user stays on the same depth, and moves in breadth only, $(S6a) \rightarrow (S6b) = \{t14\}$, $(S6a) \rightarrow (S6c) = \{t15\}$. In case a user moves in the same breadth and only in depth: moving back up as $(S6a) \rightarrow (S7a) = \{t12\}$ and moving deeper down as $(S6a) \rightarrow (S8a) = \{t13\}$. We define the other remaining states in `<dsc>` as a two-step transition consisting of a transition in depth and one in breadth: $(S6a) \rightarrow (S7b) = \{t12, t16\}$, $(S6a) \rightarrow (S7c) = \{t12, t17\}$, $(S6a) \rightarrow (S8b) = \{t13, t18\}$, and $(S6a) \rightarrow (S8c) = \{t13, t19\}$. When a user moves deeper and then stays in the same transition, the notation is $(S8a) \rightarrow (S6a) = \{t11\}$, or when a user stays on the same breadth and moves increasingly more in depth, it is expressed as $(S8a) \rightarrow (S8a) = \{t13\}$, conceptualized as $(S8a) \rightarrow (S6a) \rightarrow (S8a)$, but when a user first moves deeper, but then only moves up ('left') in breadth and stays in the same depth, it is notated as $(S8a) \rightarrow (S6b) = \{t14\}$. We note that $(S6a) \rightarrow (S8a)$ equals $(S8a) \rightarrow (S8a)$ as these have the same transition, i.e. moving deeper in the subordinate components. Similarily, $(S6a) \rightarrow (S7a)$

means someone moves back in the hierarchy, and $(S7a) \rightarrow (S7a)$ is conceptualized as $(S7a) \rightarrow (S6a) \rightarrow (S7a)$, which means that someone has moved twice back, where the transition for both is $t12$.

This concludes our formal description of our EAD Search Behavior model, where Figure 5.2 illustrates this model. How can we apply this model on interaction data found in Web search log files?

5.4 Understanding Archival Search Behaviors

This section describes three studies that apply our EAD Search Behavior model on interaction data derived from Web search log files. First, we investigate how first-time users search in the EAD finding aids. Second, we research whether there is a changing search behavior when the same user interacts more and repeatedly with a single EAD finding aid. Third, we probe into the differences among people when they interact with the same finding aids.

5.4.1 User Groups and Variables

There are in total 7,141 UUIDs active in one month of log data. Additionally, we look at the previous recorded search of two years before (so since 2007), so we keep track of the search history of each of the 7,142 UUIDS, e.g. whether they have visited the *Nationaal Archief* system before during these years. We assume that each UUID corresponds to a user. Formally, we have sets of users U, where we know that a set Q consists of people with no previous record of using the *Nationaal Archief* system, and a set R of people who have, so $U = \{Q, R\}$. However, there is an intersection of people between both sets, $\Omega = \{Q \cap R\}$, which visited for the first time, and started to come back and re-use the system. Therefore, we define the following distinct user groups:

Single-visit This group consists of 4,635 people who visit for the first time and only once in January 2009.

Already-visited This group consists of 2,172 people who already visited and used the *Nationaal Archief* system before January 2009, and did again in January 2009.

Follow-up This group consists of 334 people who visited for the first time in January 2009, and later re-used the system in the same month.

Given these groups, the experimental setup has an independent groups design, but we also look at the variance within the groups with repeated measures. The single (dichotomous) independent variable here is the visit count, where there are three groups. We control the results by grouping (or creating 'bins' of) the distributions of the transitions by the same begin (first click) and end state (last click) of a session, resulting in two conditions in each group, namely (i) start from top to bottom, and (ii) start from bottom. A session is here

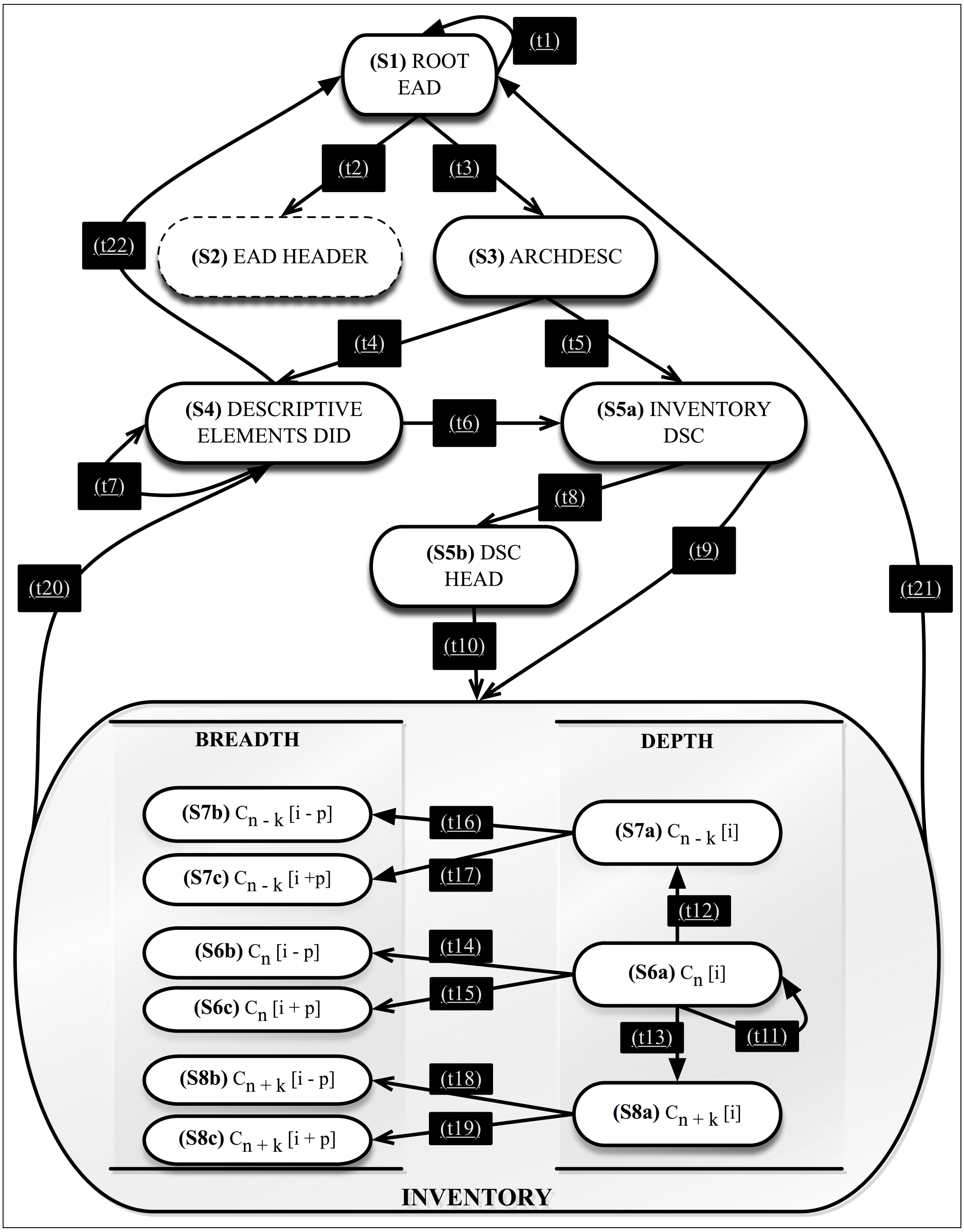


Figure 5.2: EAD Search Behavior model: Inclusive state diagram of user interaction expressed as transitions in archival finding aids in EAD.

defined as a stream of clicks belonging to a user, separated by an interval of 30 minutes of inaction, in the same EAD finding aid. The dependent variables in all conditions are the depth, the breadth, and the search time (seconds).

5.4.2 Study 1: Pioneering into Finding Aids

What is the search behavior of 'new' people pioneering into EAD finding aids? We compare *single-visit* with *already-visited* users.

Table 5.1: Distribution of states ($N = 3,606$) in one-click sessions of single-visit users.

State	Count
($S1$)	2463 (0.6830)
($S5a$)	794 (0.2202)
($S4$)	270 (0.0749)
($S3$)	79 (0.0219)

Table 5.2: Distribution of states ($N = 3,836$) in one-click sessions of already-visited users.

State	Count
($S1$)	2939 (0.7662)
($S5a$)	557 (0.1452)
($S4$)	244 (0.0636)
($S3$)	96 (0.0250)

One-click Sessions

First, we only focus on one-click sessions (no transition). Table 5.1 shows the results of the distributions of the *single-visit* group and Table 5.2 presents the distribution of transitions of the *already-visited* sessions.[3] For both groups, the vast majority of the one-click sessions constitute clicks on the whole file `<ead>` and `<archdesc>` (($S1$) and ($S3$)). Then users click in the descriptions of subordinate components `<dsc>`, which has been 'collapsed' as ($S5a$), and finally in the introductory information in `archdesc/did` ($S4$). In Table 5.2, the already-visited sessions consist of more states. Interestingly, the single-visit users have relatively more ($+7.5\%$) states within the `<dsc>` hierarchy compared to the already-visited group. This suggests that in the already-visited group, more often people stop searching by just having a top-level glance of a finding aid, while the single-visit group consists of more single 'deep' clicks that may have overwhelmed people to stop searching.

Multi-click Sessions

We group and label all transitions by the same begin (first click) and end states (last click) of a session. As laid out by Navarro-Prieto et al. (1999), there are three cognitive strategies in Web search. By looking at the first and last click of a session, we can align a session to one of these strategies. When someone starts in the top and ends down, this can be aligned to the top-down strategy. When

[3] The number of one-click states equals the number of sessions.

Table 5.3: Distribution of transitions belonging to sessions in the single-visit group, where the first click is $(S1)$ and last click is $(S5a)$ as well, where $N = 1,146$. There are 161 of these sessions, created by 150 users.

Transition	Count
$(S5a) \rightarrow (S5a)$	430 (0.3752)
$(S4) \rightarrow (S4)$	292 (0.2548)
$(S1) \rightarrow (S5a)$	100 (0.0873)
$(S1) \rightarrow (S4)$	84 (0.0733)
$(S1) \rightarrow (S1)$	70 (0.0611)
$(S4) \rightarrow (S5a)$	64 (0.0558)
$(S4) \rightarrow (S3)$	33 (0.0288)
$(S3) \rightarrow (S5a)$	20 (0.0175)
$(S3) \rightarrow (S4)$	17 (0.0148)
$(S4) \rightarrow (S1)$	12 (0.0105)
$(S5a) \rightarrow (S1)$	10 (0.0087)
$(S5a) \rightarrow (S4)$	8 (0.0070)
$(S5a) \rightarrow (S3)$	5 (0.0044)
$(S3) \rightarrow (S1)$	1 (0.0009)

Table 5.4: Distribution of transitions belonging to sessions in the already-visit group, where the first click is $(S1)$ and last click is $(S5a)$ as well, where $N = 2,848$. There are 359 of these sessions, created by 158 users.

Transition	Count
$(S5a) \rightarrow (S5a)$	1472 (0.5169)
$(S4) \rightarrow (S4)$	345 (0.1211)
$(S1) \rightarrow (S5a)$	322 (0.1131)
$(S1) \rightarrow (S1)$	258 (0.0906)
$(S4) \rightarrow (S5a)$	109 (0.0383)
$(S1) \rightarrow (S4)$	104 (0.0365)
$(S5a) \rightarrow (S1)$	54 (0.0190)
$(S3) \rightarrow (S4)$	46 (0.0162)
$(S4) \rightarrow (S3)$	43 (0.0151)
$(S5a) \rightarrow (S3)$	34 (0.0119)
$(S3) \rightarrow (S5a)$	29 (0.0102)
$(S5a) \rightarrow (S4)$	13 (0.0046)
$(S4) \rightarrow (S1)$	11 (0.0039)
$(S3) \rightarrow (S1)$	5 (0.0018)
$(S1) \rightarrow (S3)$	3 (0.0011)

someone starts deep and also ends deep, which may be a particular search behavior occurring with archival finding aids, this may be a combination between the bottom-up and top-down strategy.

We focus on two types of sessions. First, sessions that start in $(S1)$ and end in $(S5a)$, also see Tables 5.3 and 5.4, which illustrate how users search top-down from the beginning to eventually deep in the finding aid in the (sub)-series level of the inventory. Second, sessions where the first click belongs to the $(S5a)$ state, the last click also is $(S5a)$, and where users primarily seem to stay in the inventory, see Tables 5.5 and 5.6. Both types of sessions indicate that the users searched particularly deep within an EAD finding aid.

Table 5.3 shows the results for sessions starting in the $(S1)$ state and ending in the $(S5a)$ state in the single-visit group: there are 150 sessions, thus 150 users, with 1,146 transitions. The transition $(S5a) \rightarrow (S5a)$ tops the list with 37.5%, then $(S4) \rightarrow (S4)$ with 25.5%. Table 5.4 presents the distribution of the transitions of the already-visited group, where there is an absolute majority (51.7%) of the transitions of type $(S5a) \rightarrow (S5a)$. Searching in the `archdesc/did` becomes more than half less frequent (-13.4%).

Next, we focus on the second type of sessions starting and ending in the inventory. Table 5.5 lays out the results for the single-visit group. There are 537 transitions created in 143 sessions/users. Transitions within the inventory are overwhelmingly most frequent with 84%. This suggests that for single-visit

Table 5.5: Distribution of transitions belonging to sessions in the single-visit group, where the first click is $(S5a)$ and last click is $(S5a)$ as well, where $N = 537$. There are 146 of these sessions, created by 143 users.

Transition	**Count**
$(S5a) \rightarrow (S5a)$	453 (0.8436)
$(S4) \rightarrow (S4)$	25 (0.0466)
$(S5a) \rightarrow (S3)$	14 (0.0261)
$(S3) \rightarrow (S4)$	11 (0.0205)
$(S4) \rightarrow (S5a)$	9 (0.0168)
$(S3) \rightarrow (S5a)$	8 (0.0149)
$(S4) \rightarrow (S3)$	6 (0.0112)
$(S5a) \rightarrow (S1)$	4 (0.0074)
$(S1) \rightarrow (S4)$	4 (0.0074)
$(S3) \rightarrow (S1)$	1 (0.0019)
$(S1) \rightarrow (S5a)$	1 (0.0019)
$(S1) \rightarrow (S1)$	1 (0.0019)

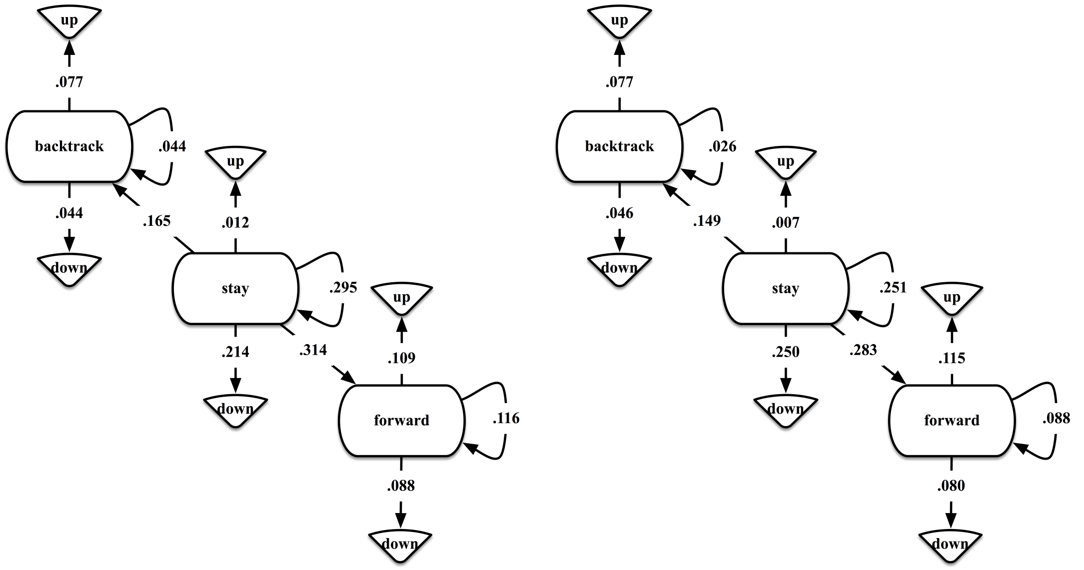
Table 5.6: Distribution of transitions belonging to sessions in the already-visited group, where the first click is $(S5a)$ and last click is $(S5a)$ as well, where $N = 1,215$. There are 176 of these sessions, created by 127 users.

Transition	**Count**
$(S5a) \rightarrow (S5a)$	903 (0.7432)
$(S4) \rightarrow (S4)$	140 (0.1152)
$(S3) \rightarrow (S4)$	30 (0.0247)
$(S4) \rightarrow (S5a)$	30 (0.0247)
$(S5a) \rightarrow (S3)$	29 (0.0239)
$(S5a) \rightarrow (S1)$	15 (0.0123)
$(S4) \rightarrow (S3)$	15 (0.0123)
$(S3) \rightarrow (S5a)$	14 (0.0115)
$(S5a) \rightarrow (S4)$	11 (0.0091)
$(S1) \rightarrow (S5a)$	11 (0.0091)
$(S1) \rightarrow (S4)$	8 (0.0066)
$(S4) \rightarrow (S1)$	4 (0.0033)
$(S1) \rightarrow (S1)$	4 (0.0033)
$(S3) \rightarrow (S3)$	1 (0.0008)

users, once they click deep within the `<dsc>` node set, they also stay in this hierarchy, and thus end here as well. Table 5.6 shows for the already-visited group that people also overwhelmingly click on the descriptions of subordinate components in the inventory with 74.3%. The main difference here with the single-visit group is that people consult the introductory information in `archdesc/did` more than twice as often. The frequency of the $(S5a) \rightarrow (S5a)$ transition warrants further analysis. What type of behavior is exactly exhibited in these transitions?

Inventory Search, Starting Top

Let us center on the sets of sessions that start in the top of the finding aid and finish in the inventory. Figure 5.3(a) depicts the distribution of the transition types in the single-visit group ($N = 430$), and Figure 5.3(b) for the already-visited group ($N = 1,472$). The main difference between both groups is that single-visit users have less interaction (also in absolute numbers), despite numerically being by far the largest group. However, we focus on any measurable interaction that is present in this group for comparison, when we omit one-click sessions, and focus on sessions with at least one transition (or more). We find that the search behavior for both groups is mostly staying in the same node set. In that node set, they go down by clicking on the next siblings of the node they just checked. Going up in the same node set is the least frequent search behavior, and almost non-existent for single-visit users. This could be explained by



(a) Detailing the $(S5a) \rightarrow (S5a)$ transitions of Table 5.3 of the single-visit group, where $N = 430$, and 63 types.

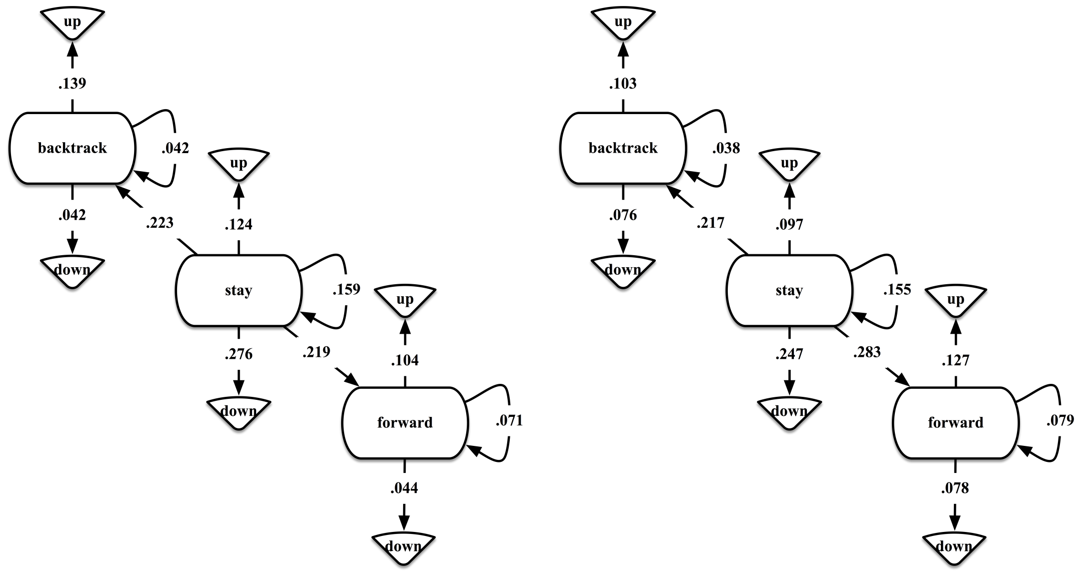
(b) Detailing the $(S5a) \rightarrow (S5a)$ transitions of Table 5.4 of the already-visited group, where $N = 1,472$, and 88 types.

FIGURE 5.3: Detailing the transitions starting from the top and ending in the inventory of the single-visit and already-visited groups in terms of depth and breadth.

users seeking 'new' or more information by naturally going down, and perhaps due to time restraints, do not go back up to previously seen information in the same node set. However, both groups do move back up to a higher component. When already-visited users exhibit this behavior, they primarily go to a higher component, while simultaneously moving up or down in breadth. Single-visit people also have this behavior, but they start less frequently from the start of a higher node set than already-visited users, but tend to stay deep while going a component higher. This may be due to the *Nationaal Archief* system interface, where single-visit users opt to stay the closest to their last visited description, while already-visited users decide to look for more context by going again top-down in the higher node set. When people move down to deeper components, both groups start from the top of the 'new' component. They usually arrived at this point deeper down in a higher component.

Inventory Search, Starting Deep

Now, let us further investigate the search behaviors of $(S5a) \rightarrow (S5a)$ of sessions that start deep and end deep in the <dsc> hierarchy. Figure 5.4(a) shows the details for the single-visit group ($N = 453$). Figure 5.4(b) presents the search behaviors for the already-visited group ($N = 903$).



(a) Detailing the $(S5a) \rightarrow (S5a)$ transitions of Table 5.5 of the single-visit group, where $N = 453$, and 72 types.

(b) Detailing the $(S5a) \rightarrow (S5a)$ transitions of Table 5.6 of the already-visited group, where $N = 903$, and 86 types.

FIGURE 5.4: Detailing the transitions starting and ending in the inventory of the single-visit and already-visited groups in terms of depth and breadth.

First, the comparison between single-visit and already visited users shows similar search behavior as the previous condition: single-visit users have fewer interaction within the inventory compared to people who already have used the *Nationaal Archief* system. Another similarity is that staying in the same node set is the dominant search behavior for both groups, and here, the search behavior is the same for both groups. Their dominant behavior is to stay in the same component, next comes going forward by clicking on a deeper component, and going back up to a higher component is the least found search behavior.

Second, the comparison between single-visit users in the two conditions. Here, we observe a difference in terms of backtracking behavior. When this group starts deep in a finding aid, backtracking is more frequent compared to when they start from the very top. The backtracking search behaviors are distributed equally, although when they start in the inventory, they go fewer times to a higher component and up in that component. In terms of forward-going behavior, there is a difference that single-visit users go relatively fewer times forward in the `<dsc>` hierarchy when they start from the top, compared to when they actually start in these components. They finish their search faster once they have arrived here from the top, and a reason could be that users found the desired description, or due to time restrictions, or it could be a matter of a threshold in the attention span.

Third, the comparison between already visited users in the two conditions. We observe a similar backtracking behavior as with the single-visit group, as more people go back when they started to search in the inventory than when they started at the top. When already-visited people move back, the least occurring behavior is that they only move a component higher, because at the same time they also move in breadth compared to their previous state. In terms of forward-going behaviors, these are constant in both conditions.

Search Time and EAD Use

In the single-visit group, where there are 1,291 multi-click sessions, the mean search time is 214.86 seconds. In the already-visited group, there are 2,267 multi-click sessions, where the mean time is 625.55 seconds. People who already visited and used the *Nationaal Archief* system spend almost thrice more time than single-visit users. We use the independent samples t-test to check for statistical significance of the difference of the search time between these two groups. We find a statistical significant difference in the mean scores between both groups ($t(2539) = 9.19$, $p < 0.0005$, 2-tailed).

Next, we break-down the mean time per session type. For the single-visit group, the $(S1) \rightarrow (S5a)$ session type (N=161) has mean of 262.34 seconds, and for the $(S5a) \rightarrow (S5a)$ session type (N=146) the mean time is 247.96 seconds. For the already-visited group, the mean time of the former session type (N=359) is 589.72 seconds and for the latter type (N=176) the mean time is 646.79 seconds. The search time is dependent on the begin state. For the single-visit group, more time is spent searching when starting in the top of a finding aid. However, for the already-visited group, people spend more time when they start in the inventory. We find significant differences between both conditions in the single-visit group ($t(1682) = -13.69$, $p < 0.0005$, 2-tailed) and significant differences in the already-visited group ($t(4062) = -7.93$).

We measure the mean search times of staying in the same component (node set), going to a deeper component, and moving to a higher component. The single-visit group spends most time searching in the same component, but when already in the inventory, they spend more time going to a higher component, possibly seeking more context, whereas as when they start in the top, they will spend more time going further down and less going back up. We use the one-way repeated measures ANOVA to check for statistical significant difference of the search time across the components in the single-visit group, but we exclude missing values pair wise.[4] We find in the first condition, a significant effect for time (Wilk's $\lambda = .88$, $F(2, 49) = 3.41$, $p < 0.05$, multivariate partial

[4] The search time values in the components do not have the same sample sizes. The size of the samples is dependent on the smallest sample size. When a sample size is bigger, the remaining values of that sample get excluded pair wise in our analysis. We assume that the values are missing randomly.

eta squared = .12). According to Cohen (1988), this result suggests a large effect size. In the second condition, we also report a significant effect for time (Wilk's $\lambda = .89$, $F(2, 54) = 3.31$, $p < 0.05$, multivariate partial eta squared = .11).

In the already-visited group, there is a similar search behavior as in the single-visit group. We use the one-way repeated measures ANOVA. In the first condition, there is a significant effect for time (Wilk's $\lambda = .92$, $F(2, 172) = 7.78$, $p < .001$, partial eta squared = .083), but not in the second condition.

Moreover, we also check how many EAD finding aids each user utilized during their search endeavors, only when there is at least once transition. We find that for the single-visit group, people on average used 1.12 finding aids, where there are 1,151 users who visited in total 1,291 finding aids, of which 833 (64.5%) are unique. For the already-visited group, 815 people visited on average 2.38 finding aids, or 1,938 in total, of which 1,196 (61.7%) are unique. So regardless of the number of clicks, people with more experience search more 'diverse' and will on average use twice as often different EAD finding aids in the same time frame (a month). However, we also see that the already-visited group slightly reuses more often the same finding aids.

Finally, Table 5.7 depicts the top ten most frequently consulted EAD finding aids for the single-visit group, and Table 5.8 for the already-visited one. In both groups, the finding aids '2.09.06' (*"Inventaris van het archief van het Ministerie van Justitie te Londen, (1936) 1940-1945 (1953)"*) and '2.24.05.01' (*"Inventaris van de fotocollectie Elsevier, positieven binnenland"*) top both lists. The remaining of the lists do not show a large similarly, so the topical interest likely diverges between both groups. More people use the top two finding aids in the single-visit group, suggesting these were deemed to be interesting enough for single-visit users to click on, but during search it may have proved otherwise.[5]

5.4.3 Study 2: Incremental Interaction and Search Behavior

What happens when the same user interacts more and repeatedly with an archival finding aid? We study the repeated-visit group, which consists of people who started to interact with the *Nationaal Archief* system for the first time, and kept using it thereafter. Do they exhibit similar search behavior compared to *single-visit* or *already-visited* users?

One-Click Sessions

Table 5.10 presents the distributions of 614 states in one-click sessions. We observe a similar distribution as with the other groups. However, the distribution

[5] There is no information on the demographic background of both groups. One could assume that expert genealogists, most frequent users of archives, would search the finding aids more than once, and would use the most popular finding aids. Historians are more likely to use finding aids that are not used by as many people, because historians have a broader range of interests and might use finding aids that are less frequently used.

Table 5.7: Top 10 EADs used by single-visit users.

EAD ID	Count
2.09.06	280
2.24.05.01	235
2.09.09	61
2.21.315	43
2.21.008.65	40
2.24.03	33
2.21.333.02	26
2.21.270	26
2.06.081	26
1.04.02	23

Table 5.8: Top 10 EADs used by already-visited users.

EAD ID	Count
2.09.06	138
2.24.05.01	85
2.09.08	45
2.09.09	40
2.20.01	35
2.01.15	26
3.09.11.04	26
4.RCMA	26
1.08.11	24
2.15.17	23

Table 5.9: Top 10 most 'popular' EADs used by follow-up users.

EAD ID	Count
2.09.06	16
2.24.05.01	11
2.09.09	9
2.21.286	7
1.04.02	7
2.13.132	7
2.13.09	6
4.VTH	5
2.13.52	5
2.20.23	4

Table 5.10: Distribution of states in one-click sessions of follow-up users ($N = 614$).

State	Count
$(S1)$	458 (0.7459)
$(S5a)$	81 (0.1319)
$(S4)$	48 (0.0782)
$(S3)$	27 (0.0440)

is more similar to the already-visited group, where relatively fewer times people click in the inventory `<dsc>` $(S5a)$ than single-visit users. This suggests that people with more experience do not click deep, unless they are certain (and can be bothered), because one-click sessions do not represent this certainty.

Multi-Click Sessions

Follow-up users hardly use the introductory information anymore, once they start and finish deep in the inventory (see Table 5.12). However, when follow-up users have started their search in the top, they still use the introductory information more frequently (24.3%, see Table 5.11), just as single-visit users (25.5%, see Table 5.3), than presumably more experienced already-visited users (12.1%, see Table 5.5). The usage of introductory information is more constant in the already-visited group with about 12% in both session types, and they make more use of the inventory when they start from the top than single and follow-up users do (see Tables 5.6 and 5.12). This indicates that already-visited users are more persevering in their endeavors as they search more exhaustively in the descriptions of the subordinate components in the inventory.

Table 5.11: Sessions in the follow-up group, where the first click is $(S1)$ and last click is $(S5a)$, where $N = 379$, and 58 sessions created by 41 users.

Transition	Count
$(S5a) \rightarrow (S5a)$	134 (0.3536)
$(S4) \rightarrow (S4)$	92 (0.2427)
$(S1) \rightarrow (S5a)$	46 (0.1214)
$(S1) \rightarrow (S1)$	42 (0.1108)
$(S1) \rightarrow (S4)$	23 (0.0607)
$(S4) \rightarrow (S5a)$	17 (0.0449)
$(S4) \rightarrow (S1)$	6 (0.0158)
$(S5a) \rightarrow (S1)$	5 (0.0132)
$(S3) \rightarrow (S4)$	4 (0.0106)
$(S4) \rightarrow (S3)$	4 (0.0106)
$(S5a) \rightarrow (S3)$	3 (0.0079)
$(S3) \rightarrow (S5a)$	3 (0.0079)

Table 5.12: Sessions in the follow-up group, where the first and last clicks are $(S5a)$, where $N = 260$, and 48 sessions created by 38 users.

Transition	Count
$(S5a) \rightarrow (S5a)$	216 (0.8308)
$(S4) \rightarrow (S4)$	22 (0.0846)
$(S4) \rightarrow (S5a)$	6 (0.0231)
$(S3) \rightarrow (S4)$	3 (0.0115)
$(S5a) \rightarrow (S1)$	3 (0.0115)
$(S5a) \rightarrow (S3)$	3 (0.0115)
$(S1) \rightarrow (S4)$	2 (0.0077)
$(S5a) \rightarrow (S4)$	2 (0.0077)
$(S3) \rightarrow (S5a)$	1 (0.0038)
$(S4) \rightarrow (S3)$	1 (0.0038)
$(S1) \rightarrow (S5a)$	1 (0.0038)

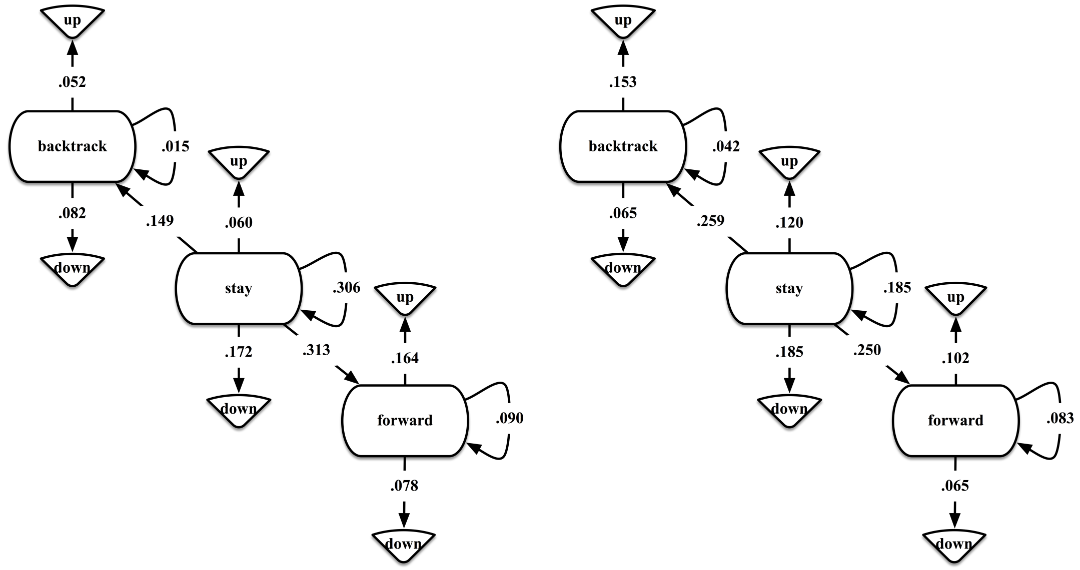
Transitions in Inventory

Figure 5.5(a) shows the distribution of the follow-up group starting in the top and Figure 5.5(b) show the distribution when it starts deep.

First, we expand the within-inventory transitions of Table 5.11 in sessions that start in the top and end in the inventory, and these are similar compared to the other groups. When we count the number of transitions that stay in the same node set, then a majority of 72 (53.7%) are moves within the same component. The total number of times that someone goes a component higher is 20 (14.9%). When we count the number of instances people go a component deeper, then there are 42 (31.3%) transitions. In terms of transition types, there are 14 same node set types, 11 backtracking types, and 18 forward-going types.

Second, we also expand the 216 $(S5a) \rightarrow (S5a)$ transitions of Table 5.12. We count 106 (49.1%) transitions that are actions moving in the same component. There are 56 (25.9%) transitions that are moves back up to a higher component. In terms of forward-going behavior, we count 54 (25%) transitions. There are 17 'staying in the same component' transition types, 25 backtracking types, and 18 forward-going types.

The search behavior of starting deep is different compared to when people start in the top of a finding aid. In case people navigate from the top to the inventory, forward-going transitions are more dominant than when they are already start there. In case they start in the descriptive components, backtracking behavior becomes more dominant. This search behavior is most apparent in the follow-up group.



(a) Detailing the $(S5a) \rightarrow (S5a)$ transitions of Table 5.11 when starting in the top and ending in the inventory, where $N = 134$, and 43 types.

(b) Detailing the $(S5a) \rightarrow (S5a)$ transitions of Table 5.12 when starting and ending in the inventory, where $N = 216$, and 60 types.

FIGURE 5.5: Detailing the within-inventory transitions of the follow-up group in terms of depth and breadth.

Search Time and EAD Use

The mean search time of follow-up users in a session is 289.62 seconds, where there are in total 465 sessions. This is more than the single-visit group where the mean time is 214.86 seconds, but still far less than the already-visit group with a mean time of 625.55 seconds. We use the one-way between-groups ANOVA with planned comparisons to check for statistical significance of the difference of the search time of the follow-up group (contrast coefficient 2) with the other two groups (contrast coefficient -1). We find significant differences ($F(1, 1302) = 14.78$, $p < 0.0005$, 2-tailed). Next, we break-down the search time for both the $(S1) \rightarrow (S5a)$ (N=58) and $(S5a) \rightarrow (S5a)$ (N=48) session types. The mean time for the former session type is 267.14 seconds, which is less than the mean over all sessions. The mean time for the latter session type is 258.33 seconds, which is less than the mean over the former session type. In order to check for statistical signifant difference of the search time for the conditions within each group, we use the paired-samples t-test. We find a significant (2-tailed) difference ($t(638) = -8.09$, $p < 0.0005$). These results show that follow-up people spend less time searching starting in the inventory than when they start from the top. This finding is similar to the single-visit group, and different compared to the already-visited group that spends more time interacting in the EAD finding aid when they start directly in the inventory, rather than when they start in the top. Reasons could be the greater certainty or expectation that something relevant can be found in the 'deep' subordinate components, and less use of the introductory information that precedes these components.

When people start from the top of a finding aid, follow-up users spend most time going to a deeper component, staying in the same component, and moving back up comes last. Here, the top-down behavior is predictable and prevalent in terms of the search time. We find a significant effect for time (Wilk's $\lambda = .46$, $F(2, 14) = 8.14$, $p < 0.005$, partial eta squared = .54). When people actually start in the inventory, going to a higher component requires most time, staying in the same component comes second, and moving further down to a deeper component comes last. Here, the bottom-up search behavior becomes clear, though not significant with the one-way repeated measures ANOVA.

The results also match with the number of transitions in the different components. We use the Pearson product-moment correlation to find out the correlation between search time and number of transitions in a session. For the follow-up group, there is a strong negative correlation ($r = -.94$, $n = 465$, $p < 0.05$, 2-tailed). For the single-visit group, there is a small negative correlation ($r = -.11$, $n = 1291$, $p < .0005$, 2-tailed). For the already-visited group, there is a large negative correlation ($r = -.64$, $n = 2267$, $p < 0.01$, 2-tailed), with the search time associated with the number of transitions.

When we observe the EAD usage in sessions consisting of multiple clicks, we measure that 224 members of the follow-up group use on average 1.77 finding aids, or a total of 396, of which 328 are unique (82.8%). This is more than in single-visit group, where there is a mean of 1.12, but still less than the already-visited group with a mean of 2.38 finding aids. A difference with the single-visit and already-visited group is that the re-use of a finding aid is less for the follow-up group, as 82.8% of the used finding aids are unique here compared to 64.5% and 61.7% respectively (i.e. the lower the percentage, the more there is re-use). This may also partly explain the search time of the sessions by the groups, as more re-use of finding aids means longer search time, and less re-use means that there are more 'newly' consulted finding aids used only once (e.g. only scanned) and used less exhaustively in time. As the follow-up group re-uses the finding aids least often, this may have impeded the search time, though on average it is still longer than the search time in single-visit group. Finally, we also look at the used EAD finding aids, and the same finding aids are topping the list as compared to the other groups (See Table 5.9), but the remainder of the lists make clear the topical divergence among the groups.

To conclude, a finding is that the search behaviors of the follow-up group exhibits characteristics of both the single-visit and follow-up group. Some evident search behaviors have been empirically confirmed. For example, when we focus in detail on the search behavior of follow-up users within the inventory, we have discovered similar search behaviors as the other groups. We find that the begin state of a session has effect on the remaining search behavior in the session, i.e. when users start deep, they tend to go back seeking more context, and when they start in the top, they will go further down in the descriptions. A clearer difference exists with the usage of the introductory information, which is by far more frequently used by people who only used the *Nationaal Archief* system once. The usage of these descriptions drops, as people use the system more often. The search time is also an indication of changing search behavior: as people use the system more often, they on average also spend more time searching, and also in more different EAD finding aids. The results show that repeated (more) usage has indeed effect on the search behavior.

5.4.4 Study 3: On the Uniformity of Archival Search

In this subsection, the research continues by investigating the research question "Do different people interact differently with the same archival finding aid?" We aggregate all clicks by EAD finding aid, and explore the similarities and differences among different people in a group. This study uses the finding aid (and its content, length, and structure) as a control variable. For each finding aid, we list the number of users, the number of sessions, the frequencies of the states and the transitions, and the descriptive statistics on the search time. We study the top ten most 'popular' finding aids. We only focus on the interaction with EAD finding aids used by more than one person.

Table 5.13 presents the results for the single-visit group, Table 5.13 depicts the results for the already-visited group, and Table 5.15 shows the results for the follow-up group. We outline the search behaviors within each single finding aid, per session, of the top 10. These behaviors are represented and measured in terms of the number of users, the type of sessions (one or multi-click), the transition types in the sessions (i.e. the total sum and the number of sessions that have a transition), and finally the search time (i.e. mean, median, maximum and minimum). So far, we have mainly focused on the different *moves* by the different groups. However, as Bates (1990) makes clear, a combination of moves constitutes a *tactic*. Can we uncover different tactics?

When we focus on the characteristics of the single-visit group in Table 5.13, the interactions of different people in a finding aid are the same, namely sessions consisting of only one click. When we look at the second most 'popular' finding aid, 232 of the 235 sessions consist of just one click. For the 9-th most 'popular' finding aid '2.06.081' (*Inventaris van het archief van de Rijkscommissie Werkverruiming en voorgangers, (1918) 1922-1946*) even all 26 sessions have just one click. This shows that most people do not fully utilize these finding aids as there are not many in-depth transitions. However, there are three finding aids with transitions within or to the inventory.

- In the finding aid used by most people, '2.09.06' (*Inventaris van het archief van het Ministerie van Justitie te Londen, (1936) 1940-1945 (1953)*), there are 46

Table 5.13: Details of searching in the top 10 finding aids by the single-visit group.

EAD	Users	Sessions		Transitions			Time			
		1-Click	*> 1-Click*	*Type*	*Total*	*Sessions*	*M*	*Mdn*	*Max*	*Min*
1. 2.09.06	280	236	44	$(S5a) \rightarrow (S5a)$	46	30	149.73	88.5	1,022	0
				$(S3) \rightarrow (S4)$	6	5	9.66	10	13	5
				$(S5a) \rightarrow (S1)$	13	11	288.84	186	867	17
				$(S5a) \rightarrow (S3)$	9	9	418.66	163	1128	26
				$(S1) \rightarrow (S4)$	10	9	62.7	24.5	347	5
				$(S3) \rightarrow (S5a)$	2	2	78.5	78.5	133	24
				$(S4) \rightarrow (S5a)$	5	5	49.6	14	144	8
				$(S3) \rightarrow (S1)$	1	1	88	88	88	88
				$(S4) \rightarrow (S1)$	1	1	27	27	27	27
				$(S4) \rightarrow (S4)$	23	8	19.30	13	84	2
				$(S4) \rightarrow (S3)$	3	3	17	19	24	8
				$(S1) \rightarrow (S5a)$	3	3	145.33	68	363	5
				$(S1) \rightarrow (S1)$	2	2	89.5	89.5	167	12
2. 2.24.05.01	235	232	3	$(S1) \rightarrow (S3)$	3	3	26.33	29	36	14
3. 2.09.09	61	48	13	$(S3) \rightarrow (S3)$	1	1	2,762	2,762	2,762	2,762
				$(S1) \rightarrow (S3)$	2	2	683	683	1,289	77
				$(S1) \rightarrow (S1)$	14	12	139.14	67	589	1
4. 2.21.315	43	39	4	$(S3) \rightarrow (S4)$	1	1	90	90	90	90
				$(S3) \rightarrow (S3)$	1	1	138	138	138	138
				$(S1) \rightarrow (S3)$	1	1	125	125	125	125
				$(S3) \rightarrow (S1)$	1	1	141	141	141	141
				$(S1) \rightarrow (S1)$	2	2	810	810	1,535	85
5. 2.21.008.65	40	33	7	$(S5a) \rightarrow (S5a)$	3	3	13	7	26	6
				$(S1) \rightarrow (S4)$	3	3	40.33	15	98	8
				$(S1) \rightarrow (S5a)$	3	3	87.33	108	129	25
6. 2.24.03	33	26	7	$(S1) \rightarrow (S4)$	7	7	277.14	34	1345	19
				$(S4) \rightarrow (S4)$	6	1	8.83	8	13	6
7. 2.21.333.02	26	25	1	$(S4) \rightarrow (S4)$	1	1	247	247	247	247
8. 2.21.270	26	19	7	$(S5a) \rightarrow (S5a)$	6	2	32.5	12	115	4
				$(S1) \rightarrow (S4)$	4	4	184.75	99	466	75
				$(S4) \rightarrow (S4)$	3	2	17	9	34	8
				$(S1) \rightarrow (S5a)$	1	1	67	67	67	67
				$(S1) \rightarrow (S1)$	1	1	97	97	97	97
9. 2.06.081	26	26	0	-	-	-	-	-	-	-
10. 1.04.02	23	11	12	$(S1) \rightarrow (S4)$	2	2	64.5	64.5	85	44
				$(S1) \rightarrow (S3)$	2	2	48	48	58	38
				$(S4) \rightarrow (S1)$	1	1	39	39	39	39
				$(S1) \rightarrow (S1)$	17	11	150.23	26	872	5

$(S5a) \rightarrow (S5a)$ transitions, occurring in 30 of the 44 sessions with at least one transition. In 5 cases, people get to the inventory from the top $(S1)$ and $(S3)$. In another 5 sessions people arrive at this point from the introductory information in $(S4)$. In the remaining other 20 sessions, people have started directly in the inventory. We count 14 sessions that have transitions from the top to $(S4)$, and once arrived here, there are in total 23 transitions in 8 different sessions, and in 4 sessions people also moved from $(S4)$ back to the top. 'Backtracking' transitions from the inventory back to the top require most time. Transitions in the inventory involves on average 149.73 seconds (2.5 minute). So in case we look at the subset of users who actually search in a finding aid, the majority of the sessions have similar search behavior, i.e. searching directly in the inventory.

- In the finding aid '2.21.008.65' (*Inventaris van het archief van mr. J.C. Voorduin [levensjaren 1799-1878], 1814-1848*), there are no transitions from the introductory information to the inventory (or vice versa). We identify two search behaviors or tactics. First, people go from the top of a finding aid to the in-

TABLE 5.14: Details of searching in the top 10 finding aids by the already-visited group.

EAD	Users	Sessions		Transitions			Time			
		1-Click	*> 1-Click*	*Type*	*Total*	*Sessions*	*M*	*Mdn*	*Max*	*Min*
1. 2.09.06	138	137	18	$(S5a) \rightarrow (S5a)$	19	9	155.89	25	1,928	3
				$(S3) \rightarrow (S4)$	4	4	13.5	10	30	4
				$(S5a) \rightarrow (S1)$	1	1	135	135	135	135
				$(S1) \rightarrow (S4)$	3	3	36.66	24	81	5
				$(S5a) \rightarrow (S3)$	6	5	246	228	504	2
				$(S3) \rightarrow (S5a)$	2	1	3	3	5	1
				$(S3) \rightarrow (S1)$	1	1	11	11	11	11
				$(S4) \rightarrow (S5a)$	1	1	498	498	498	498
				$(S4) \rightarrow (S4)$	8	3	12.37	11	25	4
				$(S4) \rightarrow (S3)$	3	2	9.66	11	12	6
				$(S1) \rightarrow (S5a)$	2	2	54	54	64	44
				$(S1) \rightarrow (S1)$	3	2	164	99	375	18
2. 2.24.05.01	85	101	5	$(S1) \rightarrow (S3)$	3	3	23	21	34	14
				$(S1) \rightarrow (S1)$	4	2	2,906	2,981	5,432	230
3. 2.09.08	45	43	14	$(S3) \rightarrow (S4)$	3	3	102.33	128	151	28
				$(S1) \rightarrow (S4)$	3	3	629.66	71	1,766	52
				$(S1) \rightarrow (S3)$	3	3	133.66	116	183	102
				$(S4) \rightarrow (S4)$	2	2	128	128.5	228	29
				$(S4) \rightarrow (S3)$	1	1	38	38	38	38
				$(S1) \rightarrow (S1)$	8	8	29.25	22.5	93	8
4. 2.09.09	40	44	15	$(S1) \rightarrow (S4)$	3	3	229.66	46	601	42
				$(S3) \rightarrow (S3)$	1	1	22	22	22	22
				$(S1) \rightarrow (S3)$	2	2	359	359.5	403	316
				$(S4) \rightarrow (S1)$	1	1	136	136	136	136
				$(S3) \rightarrow (S1)$	1	1	212	212	212	212
				$(S1) \rightarrow (S1)$	12	12	882.33	144	7,269	9
5. 2.20.01	35	33	7	$(S3) \rightarrow (S4)$	1	1	4	4	4	4
				$(S1) \rightarrow (S4)$	6	6	181	96.5	691	21
				$(S4) \rightarrow (S1)$	2	2	942	942.5	1,325	560
				$(S4) \rightarrow (S4)$	5	4	145	13	686	6
				$(S4) \rightarrow (S3)$	2	1	74	74	130	18
				$(S1) \rightarrow (S1)$	7	6	166.71	167	344	37
6. 2.01.15	26	22	10	$(S5a) \rightarrow (S5a)$	10	5	77.4	9	656	4
				$(S3) \rightarrow (S4)$	3	1	15.33	11	32	3
				$(S1) \rightarrow (S4)$	5	4	11	9	21	2
				$(S4) \rightarrow (S1)$	2	2	31	31	45	17
				$(S4) \rightarrow (S4)$	18	3	5.27	3	11	2
				$(S4) \rightarrow (S3)$	3	1	8	8	12	4
				$(S1) \rightarrow (S5a)$	2	2	18	18	31	5
				$(S1) \rightarrow (S1)$	6	4	193	70.5	841	0
7. 3.09.11.04	26	26	0	-	-	-	-	-	-	-
8. 4.RCMA	26	24	4	$(S5a) \rightarrow (S5a)$	3	1	635	7	1,897	1
				$(S4) \rightarrow (S5a)$	1	1	4	4	4	4
				$(S1) \rightarrow (S1)$	2	2	19	19	27	11
9. 1.08.11	24	18	11	$(S3) \rightarrow (S4)$	2	2	61.5	61.5	82	41
				$(S5a) \rightarrow (S1)$	1	1	181	181	181	181
				$(S1) \rightarrow (S4)$	9	7	47.77	12	161	4
				$(S3) \rightarrow (S3)$	1	1	57	57	57	57
				$(S1) \rightarrow (S3)$	2	2	1,005	1,005.5	1,526	485
				$(S4) \rightarrow (S5a)$	1	1	358	358	358	358
				$(S4) \rightarrow (S1)$	1	1	190	190	190	190
				$(S4) \rightarrow (S4)$	2	2	104	104	157	51
				$(S4) \rightarrow (S3)$	1	1	187	187	187	187
				$(S1) \rightarrow (S1)$	9	6	91	52	278	13
10. 2.15.17	23	23	0	-	-	-	-	-	-	-

troductory information, and stop after checking the first element. Second, people going from the top to the inventory, after a pause, and then quickly (M=13 seconds) clicking through it. This happens in at least 3 sessions.

- In the finding aid '2.21.270' (*Inventaris van het archief van de familie Laman Trip en aanverwante families, 1710-1986*), we again do not see a transition between the introductory information and inventory. Starting the search in the introductory information is more frequent here compared to the previous finding aid, but people tend to check for more descriptions when they are

TABLE 5.15: Details of searching in the top 10 finding aids by the follow-up group.

EAD	Users	Sessions		Transitions			Time			
		1-Click	*> 1-Click*	*Type*	*Total*	*Sessions*	*M*	*Mdn*	*Max*	*Min*
1. 2.09.06	16	19	8	$(S5a) \rightarrow (S5a)$	9	5	71.11	25	277	4
				$(S3) \rightarrow (S4)$	3	3	10.66	6	22	4
				$(S5a) \rightarrow (S1)$	2	2	75.5	75.5	107	44
				$(S1) \rightarrow (S4)$	1	1	4	4	4	4
				$(S5a) \rightarrow (S3)$	3	3	92.66	26	230	22
				$(S4) \rightarrow (S5a)$	3	3	13.66	14	19	8
				$(S4) \rightarrow (S4)$	10	2	7.3	6.5	14	2
				$(S1) \rightarrow (S5a)$	1	1	74	74	74	74
2. 2.24.05.01	11	14	1	$(S3) \rightarrow (S3)$	1	1	40	40	40	40
				$(S1) \rightarrow (S3)$	1	1	137	137	137	137
3. 2.09.09	9	9	2	$(S1) \rightarrow (S3)$	1	1	39	39	39	39
				$(S1) \rightarrow (S1)$	3	2	447.66	183	1,132	28
4. 2.21.286	7	5	3	$(S5a) \rightarrow (S5a)$	8	3	33.12	9.5	140	2
				$(S4) \rightarrow (S5a)$	1	1	2	2	2	2
				$(S4) \rightarrow (S4)$	3	1	4.66	3	9	2
5. 1.04.02	7	6	5	$(S1) \rightarrow (S4)$	2	2	170	170.5	294	47
				$(S1) \rightarrow (S3)$	1	1	57	57	57	57
				$(S3) \rightarrow (S1)$	1	1	696	696	696	696
				$(S4) \rightarrow (S3)$	1	1	357	357	357	357
				$(S1) \rightarrow (S1)$	6	4	390	84	1702	28
6. 2.13.132	7	4	7	$(S3) \rightarrow (S4)$	2	2	52	52	80	24
				$(S1) \rightarrow (S3)$	3	3	475.33	239	1,042	145
				$(S3) \rightarrow (S1)$	1	1	350	350	350	350
				$(S4) \rightarrow (S3)$	1	1	4,185	4,185	4,185	4,185
				$(S1) \rightarrow (S1)$	1	1	29	29	29	29
7. 2.13.09	6	3	8	$(S5a) \rightarrow (S5a)$	20	4	37.35	13.5	281	2
				$(S3) \rightarrow (S4)$	2	1	9.5	9.5	14	5
				$(S5a) \rightarrow (S1)$	2	2	125	125	166	84
				$(S1) \rightarrow (S4)$	2	2	10.5	10.5	18	3
				$(S5a) \rightarrow (S3)$	2	2	121	121.5	240	3
				$(S3) \rightarrow (S5a)$	2	2	13	13	19	7
				$(S4) \rightarrow (S1)$	1	1	261	261	261	261
				$(S4) \rightarrow (S5a)$	4	3	14.75	14.5	25	5
				$(S4) \rightarrow (S4)$	19	4	12.42	8	71	2
				$(S4) \rightarrow (S3)$	2	2	91.5	91.5	176	7
				$(S1) \rightarrow (S5a)$	2	2	10.5	10.5	15	6
				$(S1) \rightarrow (S1)$	2	2	29	29	45	13
8. 4.VTH	5	4	2	$(S1) \rightarrow (S4)$	1	1	185	185	185	185
				$(S1) \rightarrow (S3)$	1	1	9	9	9	9
				$(S3) \rightarrow (S1)$	1	1	49	49	49	49
				$(S4) \rightarrow (S4)$	1	1	33	33	33	33
				$(S1) \rightarrow (S1)$	3	1	5.33	4	12	0
9. 2.13.52	5	3	3	$(S1) \rightarrow (S1)$	3	3	36.33	44	52	13
10. 2.20.23	4	8	0	-	-	-	-	-	-	-

in the inventory. In 1 session, someone moved from the top to the inventory. In another case, someone started directly in the inventory. We find that the transition from the top to the introductory information requires most time.

Next, we focus on the characteristics of the already-visited group in Table 5.14. We take notice that two EAD finding aids have one-click sessions only: the 7-th most 'popular' finding aid in this group, '3.09.11.04' (*Inventaris van het archief van het Militair Hospitaal te 's-Gravenhage, 1884-1940*) and '2.15.17' (*Inventaris van het archief van het Ministerie van Sociale Zaken: Afdeling Arbeid II; Afdeling Arbeidsverhoudingen, (1930) 1942-1950*). These only have 26 and 23 one-click sessions respectively. So here there is complete agreement among the users for a specific search behavior, i.e. only getting an overview and then stop. In the other EAD finding aids, there are more transition types. There are three EAD finding aids that have transitions within the inventory, i.e. '2.09.06' and '2.01.15' (*Inventaris van de Stamboeken, Naamlijsten, Conduite- en Pensioenstaten van Officieren, onderofficieren en Minderen der Landmacht, ca. 1795-1813*) and '4.RCMA' (*Inventaris van het Kaarten-, Tekeningen- en Fotoarchief van de Rubber Cultuur Maatschappij Amsterdam (RCMA), 1908-1980*).

- In '2.09.06,' transitions in the inventory occur with 9 instances most often. In 7 instances, people moved from the top to the introductory information. Once here, there are 8 transitions in 3 sessions. There is only one transition going from the introductory information to the inventory. This indicates that people only explore this information here, and then stop their search. When we look at the search time, it took someone 498 seconds to go from the introductory information to the inventory, while on average users spent 12.37 seconds to move in the introductory information. Transitions in the inventory require on average 155.89 seconds, which is about 8 seconds less than the 'refresh' clicks, but by far more than the other transitions.
- EAD finding aid '2.01.15' has a similar pattern. There are more sessions where people rather search in the inventory than in the introductory information, but in terms of the total count, searching in the introductory information is more popular. No interaction exists between the introductory information and the inventory. More people go from the top to the introductory information than to the inventory, and once in the introductory information, they all continue at least one transition. When looking at the search time, people refresh their view on the whole finding aid for a substantial number of seconds, and secondly come the clicks in the inventory.
- When we look at EAD finding aid '4.RCMA,' we can deduce different search behaviors. There are in total 26 users who searched in this finding aid, with 28 sessions, where 4 sessions have at least one transition. One search behavior is to stay in the top of an EAD finding aid only, and refresh the page. The other search behavior is to search in the inventory after having explored the introductory information, and to click 3 times here.

Finally, we take a look at results of the follow-up group in Table 5.14. We note that the last EAD finding aid in the table has one-click sessions only. The EAD finding aids in the positions 2, 4, and 9 only have transitions stay in the top. For the finding aids in the positions 5, 6, and 8 there are only transitions from the top to the introductory information. Then there are three EAD finding aids where people actually search in-depth by exploit the hierarchical archival descriptions by going to the inventory. These are '2.09.06,' '2.21.286' (*Inventaris van het archief van W. Drees [levensjaren 1886-1988] en enkele familieleden, (1853) 1900-2000 (2002)*), and '2.13.09' (*Inventaris van het archief van het Ministerie van Oorlog: Stamboeken van Onderofficieren en Minderen van de Landmacht, 1813-1924*).

- For the first EAD finding aid, we notice that that it has been used by 16 users, consisting in total of 27 sessions, of which 8 have multiple clicks. In

4 sessions, people go from the top to the introductory information. There are 2 sessions that combined have more than 10 transitions within the introductory information. In 3 cases, people move from the introductory information to the inventory. In one session, someone moves directly from the top to the inventory, but that step took 74 seconds, which suggests that a user carefully examined the table of contents. In a majority (5) of these sessions, people also search in the inventory with 9 transitions. People spend on average more time going back to the top from the inventory.

- For the second EAD finding aid '2.21.286,' we can deduce three types of behaviors from three sessions: (i) someone starts searching directly in the introductory information, and clicks quickly three times here, (ii) someone directly moves to the inventory, and (iii) once in the inventory, the search is more exhaustive with 8 transitions that take on average 33 seconds each. These sessions may be independent, but could also have followed up on each after, and may constitute a tactic.
- For the third EAD finding aid, the descriptions in the introductory information and in the inventory are equally used, though for the latter transition type more than double of the search time is used. There are 3 sessions with 4 transitions starting from the top to the introductory information. There are 3 sessions with 4 transitions moving from the introductory information to the inventory. Transitions from the introductory information and inventory back to the top require most time, where the latter require more time than the former, as people likely spend more time in the inventory.

We have measured and described the search behaviors within three sets of EAD finding aids that have most frequently been searched by the different user groups. The results illustrate a clear distinction between use and non-use of the EAD finding aids in all groups. The majority of the sessions just have one click. Another related search behavior is to stay in the top of an EAD finding aid, or moving between the top and the introductory information. Here, the search is mainly exploratory, given the number of transitions and the absence of transitions to, within, or from the nested list of descriptions of subordinate components in the inventory, because users do not click further. This could be explained by users expecting high precision, who seek to find relevant EAD finding aids directly in the top of the returned results, although it is hard to know whether some records within the EAD finding aid might be considered relevant until users get to the file or item level. This is the challenge for effective access to EAD finding aids.

5.5 Conclusion

This chapter has outlined research on what users do when they search within an EAD finding aid. This research follows a search log-based approach combined with formal experimentation. The main contribution of this chapter's work is the theory and practice of our framework for measuring archival search behavior with EAD finding aids. We have demonstrated the validity of our EAD Search Behavior model in three studies. Our experiments and analysis based on a search log file show what people do when they search in EAD finding aids. We have derived different user groups based on the visit count. Our results show differences among the groups. People with no previous experience with EAD finding aids have fewer interaction, thus search less deep, have a shorter search time, and search in fewer (different) EAD finding aids, than people with prior experience. This interaction changes as people get more experienced, and our results have shown that the search behavior becomes more like of experienced users.When people started deep, they tend to move back. We have also investigated the search behavior of different users in a single EAD finding aid in detail, and explored the top ten most 'popular' ones of every group, which are likely exemplar of how archival users search in general. We find that 'popular' EAD finding aids are not heavily used when these mostly consist of single-clicks or 'shallow' exploratory behavior in the top of an EAD finding aid, and are used when people search in the inventory. The results illustrate the difficulty of providing focused access to EAD finding aids. Our data and analysis still show that users go to the inventory frequently, and this suggests that they also want to go there.

A large number of studies on Web search behavior employ multiple research methods, and rely on more than one data-gathering method (Hsieh-Yee, 2001). This chapter describes the results of a study that is mainly observational in nature and uses one data source, namely the Web search logs of the *Nationaal Archief*. It adds more to our understanding of archival search behavior, and complements and supports other findings derived from other research methodologies like interviews (Duff and Johnson, 2002, 2003) or models (Gilliland-Swetland, 2001). Moreover, the analysis of the search behavior of interaction with EAD finding aids is not only representative, but also complete given the sheer size of the log files. Hsieh-Yee (2001, pp.180–181) also mentions that other studies have described the behavior of a group of searchers in a particular environment performing tasks of their own or as given by researchers in experimental laboratory studies. The study as reported in this chapter resembles the former type. It exclusively looks at the natural interaction in strongly hierarchical documents consisting of real user needs, and particularly in the archival domain. Our findings do not necessarily reflect behavior of general Web searchers.

Chapter 6

User Stereotypes and Evaluation

Chapter 5 investigated how people use EAD finding aids by observing the information search behavior of users within archival finding aids in EAD. This chapter focuses on active and directed search, and also follows up on the whole EAD finding aids evaluation approach as outlined in Chapter 3.

6.1 Introduction

At the heart of IR systems is the idea to aide the (re)use of information by users. It is crucial that these systems are evaluated so as to find out how well they work for that purpose. The next step is to improve these systems based on IR experiments. The Cranfield tests in the 1950s have been instrumental, "almost entirely for the good" (Spärck Jones, 1981, p.283), in shaping the view and developing the study of IR systems. At its core, these laboratory tests involved searching for source documents on which the questions have been based. Although it was found important to stimulate real life searches, searcher variations were eliminated in order to avoid the sticky issue of relevancy, and also by the need for statistically valid results (Spärck Jones, 1981). As acknowledged by Robertson and Hancock-Beaulieu (1992, p.460), "we do not know how to simulate a real user's reactions" in a laboratory test, and he added, "operational testing is not easy either." This has profound implications for IR evaluation. How can we deal with this issue?

Instead of starting by collecting source documents with queries, it may be possible to automatically collect these data after a certain period of usage. The study of online usage of Web sites is done in Web-search Transaction Log Analysis (TLA), which is as a methodology to "examine the characteristics of searching episodes in order to isolate trends and identify typical interactions between searchers and the system" (Jansen, 2006, p.410). We can explore search patterns with implicit features that exist in the logs for information retrieval and filtering applications (Dumais et al., 2003).

In terms of archival access, it is archival description that facilitates reuse, to allow the activation of an archival record by a user. Therefore, archives seek to disclose their assets online through their websites using archival finding aids (e.g. in EAD), which increasingly often include a search engine. The transactions (or interactions, which includes searches) on websites from archives are automatically logged. Archival access is increasingly shifting online, yet it often not known how well these search systems driven by EAD finding aids perform, and how to improve them for their different users.

In terms of research methodology, there are many user studies in the archival domain that qualitatively examine information access to archival materials using finding aids in EAD (e.g. Daniels and Yakel, 2010; Duff and Johnson, 2003; Duff and Stoyanova, 1998; Duff et al., 2010; Yakel, 2004), though studies that evaluate archival access in a quantitative system-centered manner are scant (e.g. Hutchinson, 1997). Search log files could aid us in quantitatively studying the quality of online archival access, as it previously has been the case for digital libraries (Peters, 1993) and the World Wide Web at large (Jansen, 2006), so as to better understand archival users and improve (archival) information retrieval systems by evaluation. In Chapter 3, it has been shown that log files can be effectively used to construct IR test collections that agree with traditional methods of evaluation. This leads to the following main research question:

- *Q5: Can we use a search log to study different types of users and contextualize the evaluation of their specific needs?*

Although TLA is a very active research area (Jansen, 2006), the archival domain has yet to be explored. We have a massive transaction log covering multiple years of traffic on a high volume site, and the complete collection of archival finding aids that is available.

7.1 Can we use the search logs of an archive to conduct a transaction log analysis and what does it say about archival access?

As discussed by Dumais et al. (2003), implicit measures—this includes for example links, citations, dwell time, scrolling, and viewing—can be used to explore user interests and preferences. An interesting pointer to future work is explore individual or group differences (Dumais et al., 2003). In terms of IR evaluation (Robertson and Hancock-Beaulieu, 1992), we could derive real users' reactions from the log for laboratory evaluation experiments. The study of Yakel and Torres (2003) revealed that there are striking differences between novice and expert users in the archives.

7.2 Can we identify novice and expert archival users in the search log files?

We break down the search episodes into interaction elements (searching, browsing, etc). Do these groups exhibit different information seeking behaviors?

From an archival point of view, it has been stated in (Duff and Johnson, 2003, p.92) that most "archival information systems have been developed to meet the needs of archivists and historians." These are expert users. But how effective are these information systems really for these users? Gilliland-Swetland (2001, p.210) makes a distinction between user-centric and materials-centric approaches. The assertion is that the latter approach has increasingly come under fire from users and within the archival profession itself, because it not sufficiently accomodates the needs of diverse user groups. From an IR point of view, the importance of real life searches is mentioned (Spärck Jones, 1981), yet is is not know how to simulate the actions of real users (Robertson and Hancock-Beaulieu, 1992), and there is a trade-off between searcher variations and statistical valid results (Spärck Jones, 1981). Can we bridge the two? This leads finally to the third sub-question:

7.3 Can we use interaction data in the log files to contextualize the evaluation of the IR effectiveness, specifically for novices and expert users?

Are their different needs best served by different systems? Or is still the same system best for all? The remainder of this chapter is organized with the following contributions. In Section 6.2, we present a literature review regarding the closest related work as outlined in this paper. First, an analysis of our Web search log is presented in Section 6.3. Second, we detail how we derive two user groups from the log file—novices and experts—using implicit measures in Section 6.4. Third, we contextually evaluate their retrieval performance in Section 6.5 by tailoring the evaluation to the two user groups. Finally, we draw conclusions in Section 6.6.

6.2 Related Work

We give a literature overview of related literature. First, we describe literature about transaction log analysis in libraries and on the World Wide Web. We continue by discussing literature on user stereotyping and evaluation.

6.2.1 Transaction Log Analysis

Historically, TLA research started and "evolved out of the desire to monitor the performance of computerized IR systems" (Peters, 1993, p.42). In the late 1970s through the mid-1980s, TLA was applied on online public access catalogs (OPACs), such as the work of Tolle (1983) with a study that focused on the (then) current use of OPACs with early implications for IR. The overview of Kinsella and Bryant (1987) reflected on OPAC research in the UK, and also pointed to computer logging and transaction tape (or log) analysis. As indicated by Kinsella and Bryant (1987), transaction logs enable to quantify the use of an OPAC, and show the (changing) patterns of use in time, but there also limitations such as to delineating user sessions to find individual patterns to reveal real user needs.

A historical overview of TLA research in library and information science is presented by Peters (1993), where it has been pointed out that due to the development of automated IR systems in general, and transaction logging facilities in particular, TLA research gained ground. This overview shows that TLA research is extensive and diverse, with abundant published work on studies applying TLA on OPACs. For example, besides system monitoring, TLA can also be conceptualized as a way to unobtrusively observe human behavior. Fang (2007) experiments with using Google Analytics to analyze library Web sites and subsequently improve the design of these sites.

Other studies in a digital library setting have been reported by Jones et al. (2000), which focused particularly on the queries that users entered in the system, with the proposition that the analysis can be used to finetune a system for a specific target group of users. Research on TLA in library and information science preceded the current active research in the World Wide Web, which zooms into IR by analyzing search engines (e.g. Jansen et al. (1998, 2007)). An overview on search log analysis for Web searching, and a methodology, is presented by Jansen (2006), which shows that literature on TLA for Web-searching is abundant. In the archival domain, Prom (2011) makes the case that Web analytics can be used as method to measure user actions, to understand some aspects of user behavior, and to subsequently improve online services.

6.2.2 User Stereotypes and Evaluation

Stereotypes are proposed by Rich (1979) as a useful mechanism for building models of individual users based on a small amount of information about them, which is also called a *persona* (Chapman et al., 2008). A user stereotype must accurately characterize the users of a system in order to be useful, and are effective in optimizing the utilization of a system. It is also suggested by Ellis (1989), where an IR system for social scientists was developed, that the design of IR systems can be based on user models. For evaluating IR systems, certain human characteristics—like the degree of subject knowledge or professional education—affect relevance judgments and their consistency (Saracevic, 1975). The importance of this user context is stressed when considering relevance and people (Saracevic, 1975). Implicit measures of user interest (e.g. links, citations, etc) can be used for IR applications (Dumais et al., 2003) as these point to information about the users. So user stereotypes can be used for developing and evaluating IR systems.

In the archival domain, the importance of users has increased. The archives paid little attention to their users until the 1990s (Duff and Johnson, 2003). A user model based on genealogists, who are one of the most frequent users of archives, has been presented by Duff and Johnson (2003), and can be used to improve the design of online archival search systems. The study of Yakel and Torres (2003) points to the difference between expert and novice archival users.

Additional insight regarding novice and expert (Web) users has been presented in other user studies (Brand-Gruwel et al., 2005; Holscher and Strube, 2000).

6.2.3 Improving IR Systems and Search Behavior

Wilbur (1998, p.526) notes that it remains to be seen whether there are any simple principles that characterize human performance and which can serve to improve machine performance, or whether we must await the development of machines that can learn by accumulated experience an array of responses, essentially irreducible in its complexity, such as the one that may underlie human behavior (Wilbur, 1998, p.527). The idea to define simple principles that characterize human performance is—in a related way—done by Davis (1989).

Davis (1989) develops and validates new scales for two specific variables, perceived usefulness and perceived ease of use, which are hypothesized to be fundamental determinants of user acceptance of computers (e.g. programs and technologies). Although human performance does not equal to user acceptance of computers, it starts with it. Moreover, the more a user accepts a computer by perceiving it to be useful and of ease of use, the likely the performance will improve. Davis (1989) observes that perceived usefulness and ease of use were significantly correlated with usage behavior, usefulness had a significantly greater correlation with usage behavior than did ease of use, and usefulness could be a causal antecedent of easy of use. In other words, when a system is regarded as useful, the user may perceive it also as easy to use.

Going a step further, and in the IR context, the objective of characterizing human performance into a form of simple principles to improve machine performance is not a farfetched idea either. Ellis (1989) develops a behavioral model—which attempts to move beyond the interface but on the interaction with a database itself—based on empirically investigating the search behavior of social science researchers and uses this model to inform the design of information retrieval systems. Ellis (1989) proposes to replicate this model in other domains. Since then, the behavioral model of Ellis has been applied to compare researchers in the physical sciences with social scientists (Ellis et al., 1993), in the context of an engineering company (Ellis, 1997), or with nursing science, literature/cultural studies, history and ecological environmental science (Talja and Maula, 2003).

Koenemann and Belkin (1996) study interactive information retrieval behavior and effectiveness, and describe an experiment investigating the information seeking behavior of novice searchers who used one of four versions of an IR system to formulate routing queries for two given search topics. The independent variable is here relevance feedback, which is a query formulation support mechanism that modifies an existing query based on available relevance judgements for previously retrieved documents. The results show that relevance feedback appears to be working effectively, especially when it is made transparent ('penetrable') in a system, as users needed fewer iterations to achieve results comparable to, or better than the other, less interactive, feedback conditions.

Wilson et al. (2009) propose a framework to evaluate advanced search interfaces using established information-seeking models through the quantification of the strengths and weaknesses of the interfaces in supporting user tactics and varying user conditions. Rose (2006) considers several characteristics of user search behavior, e.g. the variety of information-seeking goals, the cultural and situational context of search, and the iterative nature of the search task, and suggests ways that interfaces can be redesigned to make searching more effective for users.

Moving beyond the evaluation of interfaces, Yuan and Belkin (2010b) investigate IR support techniques for different information-seeking strategies (ISS). There are four conditions in their within-subjects experiment, where each condition is represented by an experimental system that supports a particular ISS. The findings are that it is possible to support different behaviors with different combinations of IR support techniques, and systems that support these techniques work better. Following up on that study, Yuan and Belkin (2010a) evaluate an integrated system that adapts to support different ISS, which include both scanning and searching behaviors. It is found that combining and supporting different ISS in one system works better than the baseline of a standard IR system. In fact, (Bates, 1990) already stated that users can search more effectively when search activities associated with information searching are incorporated into the system. So Bates (1990, p.590) concludes:

> "A really good information retrieval system that allows us to exercise strategic search choices quickly and easily may, in like manner, lead us to explore knowledge and research our information needs in far more powerful and creatively stimulating ways than we ever imagined in the days of the manual library or the simple online bibliographic database."

On the one hand, a system can be designed that enables search behaviors, assuming that by enabling certain interactions, a system may work better. On the other hand, off-the-shelf data on search behaviors can be (re-)used to improve systems. An instance of the latter case is the study of Beg and Ahmad (2007) that proposes a personalized Web search system by maintaining the search profile of each user, which is the basis of determing the search results of that system. The Web search quality is a component of their framework. This quality is computed by the sequence in which a user picks up the results, the time spent at those documents and whether or not the user prints, saves, bookmarks, e-mails to someone or copies-and-pastes a portion of that document. Interacting with the stored profile of each user, this integrated system is simulated and found to give improved Web search results. Likewise, in the controlled laboratory IR evaluation experiment of Agichtein et al. (2006), it is also shown that incorporating user behavior data can significantly improve the ranking of top results in a real Web search setting compared to the baseline of the original system performance.

For archives, the research consists of studying search behaviors to improve systems. An example is the proposal of Gilliland-Swetland (2001) to enhance archival access by 'popularizing' the archival finding aid in EAD for diverse user groups by re-fitting the model of Ellis (1989). A user group are genealogists, and Duff and Johnson (2003) conceptualize an access system driven by finding aids based on interviews. No research exists in the archival domain that formally and quantifiably analyzes archival search behavior based on off-the-shelf interaction data derived from search logs that comprise of 'real' usage of archival finding aids in EAD, and conducts contextualized system evaluation based on that.

6.3 Archival Search Log Analysis

In this section, we offer an analyzis of the search log files of the *Nationaal Archief* website. This analysis offers insight into the search behavior of archival users.

6.3.1 Preparation

We use Perl to prepare the original logs for analysis. We start by extracting only the click-through data that can be traced to the use of EAD finding aids as we have also obtained the matching EAD finding aids for analysis and experimentation. The URLs refer to the filename, frequently also the query terms, and occasionally include parameters like a sub(category) as subject headings.

The next step is to partition the log files into smaller subsets by identifying user sessions. A *user session* is defined in (Jansen et al., 2007, p.862) from "a contextual viewpoint as a series of interactions by the user toward addressing a single information need." We define a session as a subset of n clicks from the same IP address, if and only if the difference between i and $i + 1 < 30$ minutes, where i is 1 click, hence it is possible that a user has multiple sessions.

6.3.2 Analysis

We analyze the logs to illuminate the search behavior of archival users in general. We mainly look at the queries used, the session length, and session duration as has been done in previous studies with log files in other domains (e.g. Jansen et al., 1998; Jones et al., 2000).

Origin of Users

Table 6.1 gives an overview of the origin of all users who used the *Nationaal Archief* Web site, where we only list the origin of the IPs that could be traced back. It states the official short names in English as given in ISO 3166-1 and the corresponding ISO 3166-1-alpha-2 code elements.[1] Not surprisingly, we see that Dutch users top the list. American users are the second most popular group, perhaps due to the number of Dutch emigrants and their offspring.[2] We observe that users coming from the former Dutch colonies Surinam and Indonesia are in the third and fifth position respectively.

Query Terms

We count the frequencies of all query terms (keywords). Table 6.2 shows the top 10 most frequently used query terms for searching in the archival finding aids in EAD. This count is interpolated over URLs that did not have a query included. That is, we assign the last known query to a hit without query. There are in total 441,575 hits with a query found in the complete log. The distribution of the query terms has a long-tail shape, which means that most users, like genealogists, entered unique keywords—mostly names—when interacting with the *Nationaal Archief* system. This distribution complements previous search log studies (Jansen et al., 1998; Jones et al., 2000), as well as findings in archival studies (Duff and Johnson, 2003). We also see that the users of the system searched for the popular archives of the *Nationaal Archief* —mostly from the Dutch colonial past. At position 1 stands the archive about the *Vereenigde Oostindische Compagnie* (VOC; in English: Dutch East India Company). The users also used other acronyms as queries, such as *Koninklijke Nederlandse Voetbal Bond* (KNVB; in English: Royal Netherlands Football Association). At position 8, we see the query *2.10.01*, which is the UUID belonging to the archive of the "Ministry of Colonial Affairs (1814-1849)." These UUIDs are often used, which implies that there is known-item search, i.e. the user used the search engine as a bookmark tool.

Session Lengths

We explore the session length (Jansen, 2006), i.e. the number of queries used in a session. We fine-tune this by looking at two aspects to explore the session length: (1) the number of unique queries used in a session (i.e. query revision), and (2) the number of clicks in a session. The results are presented in Table 6.3 and 6.4 grouped by frequency.

There are 78,190 sessions with a known query, while there are 194,138 sessions in total. This means that for the majority of the sessions (115,948), no query could be found, or have an 'empty query.' The majority of the sessions with a query have only one type of query (81%), while there are 538 sessions with more than 10 unique queries (see Table 6.3). This implies that users—even

[1] Retrieved 2011/04/14 from `http://www.iso.org/iso/english_country_names_and_code_elements`

[2] According to the US census of 2006 (`http://factfinder.census.gov/servlet/ADPTable?_bm=y&-geo_id=01000US&-ds_name=ACS_2006_EST_G00_`), there are 5,187,864 Americans with Dutch roots, which accounts for 1.6% of the US population.

Table 6.1: Origin of the users of the *Nationaal Archief* Web site.

Origin	Count
NL	7,837,158
US	2,739,877
SR	335,818
BE	181,473
ID	132,745
GB	118,259
AU	105,590
IN	80,445
DE	78,529
RU	75,603
CN	70,673
TR	67,974
JP	63,082
FR	63,044
CA	62,775
SE	52,973
ZA	49,069
CH	41,446
SG	32,504
AN	32,148
IT	31,718
ES	30,361
KR	28,355
AE	25,407
MY	24,912
BR	23,302
AT	19,328
UA	18,842
PT	18,107
IL	17,791
LK	17,702
IE	16,760
NO	16,744
CZ	12,776
TH	12,365
EU	11,446
RO	11,296
PH	10,551
PL	9,389
NZ	8,931
DK	8,292
CL	7,859
AR	7,566
DZ	7,559
SA	7,424
TW	7,176
IR	7,061
HU	6,551
PE	6,507
VN	6,190
GR	6,176
FI	5,614
CO	4,665
HK	4,550
MX	4,452
LV	4,262
CR	4,104
VE	3,807
TN	3,351
KW	3,265
QA	3,145
EG	2,641
LU	2,570
KZ	2,307
PR	2,166
LT	2,166
BH	2,113
MA	2,065
PK	2,060
BY	1,965
BG	1,949
OM	1,846
BB	1,741
SK	1,657
PA	1,553
GH	1,527
HR	1,525
MV	1,520
AG	1,357
NA	1,336
EC	1,328
CY	1,314
NG	1,287
BD	1,188
LB	1,173
MU	1,148
SN	1,029
SI	967
KE	941
EE	935
BN	913
PY	855
PS	843
YE	823
SV	817
ET	800
MZ	789
MT	751
BA	730
KH	691
RS	671
LA	663
JO	654
RW	640
MK	640
CI	607
BS	593
GE	566
IS	538
DO	537
MO	465
TT	444
NC	443
BO	441
ME	439
GT	434
AZ	399
SY	397
FO	393
TZ	377
PF	363
BJ	353
IQ	350
AO	346
ML	340
SD	322
MM	315
JM	306
AP	304
CU	295
AW	292
BW	251
AL	251
CM	248
TG	234
CV	221
UZ	209
AF	202
UG	200
AM	196
UY	186
NP	174
LY	164
MW	161
MG	156
MN	150
GI	149
VI	143
KY	119
FJ	119
BF	100
NI	98
SC	97
GA	85
AD	85
MC	72
MD	67
ZW	66
MP	59
BI	59
HN	57
GU	55
LI	55
GM	54
PG	43
CG	41
DJ	39
KG	39
MQ	37
BM	35
CD	33
VU	32
MR	32
YU	32
AS	31
ZM	30
SL	26
**	25
SO	25
TJ	24
TO	24
AX	24
BZ	21
GQ	16
TC	15
HT	12
CK	10
GN	8
VA	8
GY	8
PW	7
BT	7
TD	6
AI	5
WS	4
LS	4
NF	4
CF	3
SM	3
NE	2
IO	2
FK	1
GP	1

Table 6.2: Top 10 most popular used query strings aggregated and interpolated over each hit from 2004-2009, where the total number of (interpolated) queries is 465,089 with 50,424 unique type of queries.

Position	Query String	Count (%)
1	voc	4,383 (0.94)
2	suriname	4,277 (0.92)
3	knil	2,785 (0.60)
4	knvb	2,506 (0.54)
5	wic	1,891 (0.41)
6	hof	1,633 (0.35)
7	hof van holland	1,567 (0.34)
8	arbeidsdienst	1,541 (0.33)
9	2.10.01	1,510 (0.32)
10	drees	1,334 (0.29)

Table 6.3: Distribution of the aggregated (bins) number of unique known queries used, where the documents have been clicked in total 464,932 times in 78,190 sessions with a known query.

Queries Per Session	Session Count	
	N	%
1	63,549	(81.28)
2	9,524	(12.18)
3	2,516	(3.22)
4	941	(1.20)
5	471	(0.60)
6	229	(0.29)
7	159	(0.20)
8	113	(0.14)
9	96	(0.12)
10	54	(0.07)
> 10	538	(0.69)
	78,190	100

when they visit frequently—mostly search for one query during a session. But how often do users click on different results using these queries?

As Table 6.4 shows, in almost 45% of all queries, only one result has been clicked. However, for more than 16% of all queries there were more than 10 clicks on different EAD files. The former could mean that a user directly found the desired result, or discovered that further search with a query would not be effective and stopped. The latter could mean that a certain query yielded many relevant results, or the user decided to continue searching regardlessly. What does this mean for the time spent per session?

Table 6.4: Distribution of the aggregated number of documents clicked and viewed per non-empty query, where the documents have been clicked in total 465,089 times.

Clicks Per Query	**Query Count**	
	N	%
1	22,444	(44.51)
2	6,686	(13.26)
3	3,552	(7.04)
4	2,474	(4.91)
5	1,877	(3.72)
6	1,457	(2.89)
7	1,151	(2.28)
8	992	(1.97)
9	898	(1.78)
10	762	(1.51)
> 10	8,131	(16.13)
	50,424	100

Table 6.5: Session duration: distribution of the dwell time grouped per bin.

Time (s)	**Count**	
	N	%
0	118,564	61.07
0-500	54,109	27.87
500-1,000	9,336	4.81
1,000-1,500	4,639	2.39
1,500-2,000	2,924	1.51
2,000-2,500	1,315	0.68
2,500-3,000	888	0.46
$> 3{,}000$	2,363	1.22
	194,138	100

Session Duration and Repeated Visits

We check the time (in seconds) that was spent in a session, which we call the *dwell time*. In case of one-click sessions, this time is set to 0. Table 6.5 shows the distribution of the dwell time grouped per bin. Most of the sessions consisted of a interactions that consisted of just one click. In case there are more clicks, the session lasted no longer than 500 seconds (about 8 minutes). There are 2,363 instances of sessions where a user would search for more than 3,000 seconds (or 50 minutes). This distribution is similar to the ones found in previous studies (e.g. Jansen et al., 2007; Jones et al., 2000). It can be imagined that a user continues searching after a break, so how often do users re-visit and thus re-use the search system?

Table 6.6 depicts the maximum number of repeated visits per user. It shows

Table 6.6: Repeated visits: distribution of the maximum number of sessions and number of users (IP), which shows that the majority of users have interacted in one session.

N Visits	Users		*N* Sessions
	N	%	
1	88,539	79.91	88,539
2	11,660	10.52	23,320
3	41,02	3.70	12,306
4	1,903	1.72	7,612
5	1,119	1.01	5,595
6	717	0.65	4,302
7	478	0.43	3,346
8	389	0.35	3,112
9	295	0.27	2,655
10	212	0.19	2,120
> 10	1,391	1.26	41,231
	110,805	(100)	194,138

that the majority of the users searched in the archival finding aids in EAD only one time, and 1,391 users reused the files more than ten times. It is interesting to note that 41,231 sessions could be traced back to 1,391 users, so there are on average about 30 sessions per user in this group.

6.4 Deriving User Groups

In this section, we will try to uncover specific and interesting groups of users in the log, and analyze their information seeking behavior.

6.4.1 Implicit Features of User Interest

Can we identify different user groups—novices and experts—in the log files? A reasonable assumption is that archival experts—like genealogists or historians–use the archives more frequently than novice users. This supposition is supported by previous user studies (Duff and Johnson, 2003; Yakel and Torres, 2003). Hence our operational definition of "archival experience" is in terms of frequency of visits. We experiment with categorizing the interaction data extracted from the log by visit counts (i.e. number of sessions per user). The search system of the *Nationaal Archief* website presents links to archival finding aids in EAD in a HTML interface that appear to be relevant to a query. The number of visits by a user to these finding aids in EAD suggests the amount of experience that user has with working the search system.

The complete log has been processed and partitioned in sessions. We use these sessions to create 11 groups (or bins)—aggregated over all years—by the maximum session count. We pay special attention to two groups:

First group This group stands for *bin 1*, i.e. the set of sessions that correspond to the single-visit sessions (see Table 6.6).

Last group This group is *bin > 10*, i.e. the set of sessions that can be traced back to users who used the archives more than 10 times in different sessions.

We have identified and extracted the following implicit features that could point to user interest for each bin of sessions. Can we use the following implicit features to identify user groups?

Dwell time The amount of time in seconds that a user spends interacting with a system in a session, where the time-out between 2 interactions is set to 1,800 seconds (30 minutes). A one-click session is a session with a dwell time of 0 seconds.

Query Revision The number of queries used in a session. The Query Revision has a value of 0 when there are no queries found in a session.

Repeated Queries The number of times the first query of a session is repeated later in that session.

Query Length The number of terms in a query.

Deep Linking The number of times the user clicks on an anchor value that links to a part of a document.

Full-text Linking The number of times the user clicks on an anchor value that links to a full-text document.

Additionally, users of the *Nationaal Archief* website searched in the archives using topical (sub)categories. Therefore, we also extract and count the use of (sub)categories. The results are shown in Tables 6.9, 6.10, 6.11, 6.12, and 6.13.

6.4.2 Vocabulary Use

We also look at the vocabulary use of the first group and last group in order to see the differences. When users clicked on an EAD finding aid without having used a query in the session, it is represented as '—'. We see that this the most popular in the last group, which suggests that the last group browses more. Both *suriname* (Surinam) and *voc* (Dutch East India Company) are the most popular used queries in both groups. We see that the last group makes more use of know-item search when they retrieve an EAD finding aid using the number of that finding aid (e.g. *2.10.01*), which is not known unless the user already knows it. So we see similarities and differences in the search requests between both groups.

6.4.3 Results

We see that for the *first group*—set of single-visit sessions—the dwell time is on average the least. We see a clear divide in Table 6.9. It is the highest—on average almost 5 times as long—for sessions belonging to users who have

Table 6.7: Top 10 most popular used query strings of the first group ($N = 14,368$).

Query String	Count (%)
suriname	295(2.05)
voc	293(2.04)
knil	202(1.41)
nsb	85(0.59)
knvb	73(0.51)
arbeidsdienst	69(0.48)
wic	61(0.42)
—	60(0.42)
oorlog	60(0.42)
drees	59(0.41)

Table 6.8: Top 10 most popular used query strings of the last group ($N = 14,866$).

Query String	Count (%)
—	719(4.84)
suriname	506(3.40)
voc	295(1.98)
hof van holland	222(1.49)
curacao	198(1.33)
staten generaal	197(1.33)
2.10.01	181(1.22)
2.10.02	180(1.21)
1.04.02	178(1.20)
drees	153(1.03)

Table 6.9: Statistics about the dwell time and one-click sessions (0 dwell time) found in the log over all bins.

Bin	Dwell Time		One-Click	
	M (SD)	N	Count	%
1	105.07 (347.78)	88,539	60,341	68.15
2	179.83 (481.32)	23,320	13,551	58.11
3	218.15 (570.75)	12,306	6,815	55.38
4	256.66 (646.34)	7,612	4,071	53.48
5	271.46 (704.45)	5,595	2,933	52.42
6	262.18 (660.32)	4,302	2,343	54.46
7	301.32 (743.13)	3,346	1,730	51.70
8	279.76 (688.18)	3,112	1,727	55.50
9	288.15 (736.97)	2,655	1,441	54.28
10	265.71 (682.25)	2,120	1,173	55.33
> 10	520.80 (1,773.87)	41,231	22,436	54.42

visited more than 10 times (*last group*). We test whether there is a significant difference between the mean scores of the dwell time (independent variable) of the *first group* and *last group* using the independent samples t-test. We find a significant difference for the first group (M = 105.07, SD = 347.78) and last group (M = 520.80, SD = 1773.87; t(42713) = -47.17, $p < .000$, two-tailed). We also notice that as the number of visit count is increased, the dwell time also tend to increase. We check whether this is significant using a one-way between-groups ANOVA. We find statistical significant differences at the $p < 0.01$ level for the eleven groups according to the dwell time (F(10, 194) = 591.68).

Table 6.10 show the results related to the query properties: the query revision, repeated queries, and query length. Regarding the query revision and repeated queries, we again see strong differences between the first group and the last group. The former group has on average a query revision value of

Table 6.10: Statistics about the queries found in the log over all bins.

Bin	Query Revision M (SD)	N	Repeated Queries M (SD)	N	Query Length (SD)	N
1	0.4810 (0.9404)	88,539	0.0787 (0.4100)	88,539	1.7295 (1.1629)	42,599
2	0.6907 (1.2810)	23,320	0.1414 (0.7049)	23,320	1.7727 (1.2362)	16,108
3	0.7817 (1.5929)	12,306	0.1491 (0.5799)	12,306	1.7648 (1.2575)	9,619
4	0.8543 (1.5900)	7,612	0.1797 (0.8556)	7,612	1.8630 (1.6133)	6,505
5	0.9040 (1.6048)	5,595	0.1735 (0.6818)	5,595	1.7404 (1.2587)	5,058
6	0.8459 (1.4115)	4,302	0.1690 (0.6408)	4,302	1.8063 (1.2905)	3,639
7	0.9800 (1.7542)	3,346	0.2047 (0.7084)	3,346	1.7557 (1.3177)	3,279
8	0.8969 (1.6782)	3,112	0.1951 (0.8083)	3,112	1.7191 (1.1519)	2,791
9	0.9571 (1.9779)	2,655	0.1992 (0.8294)	2,655	1.6934 (1.1308)	2,541
10	0.9557 (2.4104)	2,120	0.1632 (0.6181)	2,120	1.7493 (1.2046)	2,026
> 10	1.5468 (5.4793)	41,231	0.2113 (1.3042)	41,231	1.5400 (1.2493)	63,778

0.4810, while the latter group revise the query significantly three times more often (t(42364) = -39.23, $p < .000$). Queries are not often repeated, but when they were, the group which used the archives most frequently also reused their queries most often. This is surprising, since we expected that if a query is revised less often, the same query is repeated more often. Overall, fewer interaction is found in the first group than in the last group.

Interestingly, we observe that the last group used on average shorter queries. These query length values are lower than reported in a previous study on digital libraries and the World Wide Web (Jones et al., 2000). A reason could be the particular use of acronyms, which we treated as singleton queries.

The logs also recorded the navigation path among different EAD finding aids. We use the search system of the *Nationaal Archief* to discover the sequence of the different types of links. The interaction flow is as follow (also see page 37). After the user enters a query in the search system, an overview of the results is presented with 2 options.

1. The first option is to click on an overview view which presents potentially relevant links to summary views (*Summary*)—these summary views link to the start of a file (*Page View*) and present contextual information (e.g. title, summary). On a Page View, users can continue the search within an EAD file by deeplinking.
2. The second option is to click directly to a part of a document (*Direct To File*) and skip the Summary.

We focus on the number of times the users clicked on a deep link, or to a full-text EAD file (thus starting from the beginning). We see in Table 6.11 that the users clicked more often on a deep link than a full-text link. This is a feature of the EAD files, which provide access to information to a part of a document. The first group has the fewest number of clicks, whereas the last group has the most. Again, we see that there is more interaction in terms of clicks coming from users who search in the archives more frequently.

TABLE 6.11: Average number of deep links and full-text links found in the log over all bins.

Bin	Deep Linking M (SD)	Full-text Linking M (SD)	N
1	2.03 (5.13)	0.74 (1.23)	88,539
2	3.10 (10.82)	0.94 (1.69)	23,320
3	3.47 (8.55)	1.07 (2.27)	12,306
4	3.88 (10.48)	1.14 (2.62)	7,612
5	4.10 (13.28)	1.17 (2.28)	5,595
6	3.93 (9.63)	1.07 (1.98)	4,302
7	4.36 (10.36)	1.28 (2.47)	3,346
8	4.11 (9.78)	1.08 (2.04)	3,112
9	4.05 (8.85)	1.16 (2.56)	2,655
10	3.87 (10.67)	1.16 (3.29)	2,120
> 10	5.36 (19.37)	2.01 (7.03)	41,231

Table 6.12 shows that the majority of clicks link to the page views, which includes deep links. Then comes the summary views, and finally *Direct To File* clicks. This suggests that users more often start searching at the summary views—and narrow down their search by browsing and clicking within a file—rather than using *Direct To File* links. This is the case for all groups. Again, we see that the number of clicks is least frequent for the first group, and the most for the last group, though the search pattern is the same.

TABLE 6.12: Average number of types of links found in the log over all bins.

Bin	Summary M (SD)	Page View M (SD)	Direct to File M (SD)	N
1	0.3655 (0.8313)	2.2966 (5.2208)	0.2079 (0.6334)	88,539
2	0.5027 (1.1739)	3.4172 (10.5553)	0.3061 (0.8072)	23,320
3	0.5791 (1.6434)	3.8861 (8.9876)	0.3489 (0.8505)	12,306
4	0.6171 (1.8375)	4.3427 (11.0165)	0.3866 (0.9099)	7,612
5	0.6272 (1.5163)	4.5458 (13.5902)	0.4152 (0.9856)	5,595
6	0.5700 (1.2754)	4.3708 (10.0514)	0.4193 (0.9924)	4,302
7	0.7113 (1.8031)	4.8715 (10.8487)	0.4441 (1.0608)	3,346
8	0.5993 (1.4655)	4.5508 (10.1379)	0.4296 (1.1041)	3,112
9	0.6354 (1.7361)	4.5168 (9.4326)	0.4648 (1.2595)	2,655
10	0.6429 (2.2847)	4.3137 (11.2965)	0.4175 (1.0109)	2,120
> 10	1.1896 (4.6813)	6.1017 (21.3702)	0.5252 (1.8338)	41,231

Finally, let us focus on the use of (sub)categories by the users. A distinction between novice and expert users of archives is the search for names (Yakel and Torres, 2003), e.g. users looking for ancestral information by using their names as query—and this happens particularly often in archives. The transition from tracing personal and organization names (novices) to a particular research project (experts) is an essential part of distinguishing both user groups as this enables more effective information retrieval (Yakel and Torres, 2003). Table 6.13 shows the rank order of the use of categories *Families and Persons* and *Persons and Families*. Both are in fact the same, but this division may be due to an artifact of the system. It shows that for the first group, these categories were the most popular, and used less frequently in the last group. This is another indication that the last group exhibits traits of expert users, as experts move away from searching for names, and to a particular research project.

Table 6.13: On the use of Family and Personal names as categories, which shows that these categories were most popular for first group, whereas the last group used other category subjects more frequently.

Category	**Bin 1**		**Bin > 10**	
	Rank	N **(%)**	**Rank**	N **(%)**
Fam. and Pers.	1	805 (18.56)	5	269 (5.30)
Pers. and Fam.	2	799 (18.42)	7	221 (4.35)

6.4.4 Novices and Experts

The results show two clearly different interaction stereotypes. On the one hand, we see a group of users which spends least amount of time time to search of all groups, has most one-click sessions, revises and repeats queries least often, clicks less often on results given their queries, and mostly seem to search for names. On the other hand, we have a group of users which spends more time to search than any other group, revises and repeated queries most often, clicks more than the other groups, and did not primarily search by looking for names. Can we match both types of interaction stereotype with certain user groups?

In a previous study (Holscher and Strube, 2000), a finding was that a user with considerable knowledge in a certain domain spends significantly less time to read documents in that domain. In other words, domain experts have a better performance as they search more efficiently and spend less time. However, we have to note that archival users are different than Web users with different information tasks. Archival finding aids in EAD are also complex document representations, which differ from normal web pages, particularly by the length of content and depth of document structure. Expert archival users are doing research, and have problem-solving tasks, and use the document structure to search for relationships.

Lazonder et al. (2000) examine differences between novice and experienced users in searching information on the World Wide Web. They state that seeking information on the Web is similar to the work of a detective. It involves tracing relevant information, one has to ask the right questions, consult proper sources of information, and creatively combine search outcomes (Lazonder et al., 2000, p.576). Their controlled experiment uses tasks and questionnaires. The participants are divided in two groups, experts and novices, using a set of questionnaires. In terms of finding relevant Web sites, their findings show that experts perform tasks faster and more successfully with fewer interaction than novices. In terms of finding information in Web sites, there is no significant difference.

However, a study on information problem solving processes of novices and experts—e.g. identifying information needs, locate information sources, etc—revealed that experts spend significantly more time to complete a task than novices (Brand-Gruwel et al., 2005). This study showed that experts would spend the maximum available time to try to solve a problem. This is a match with expert archival users, such as genealogists, who continue searching until they have found the information they needed (Duff and Johnson, 2003).

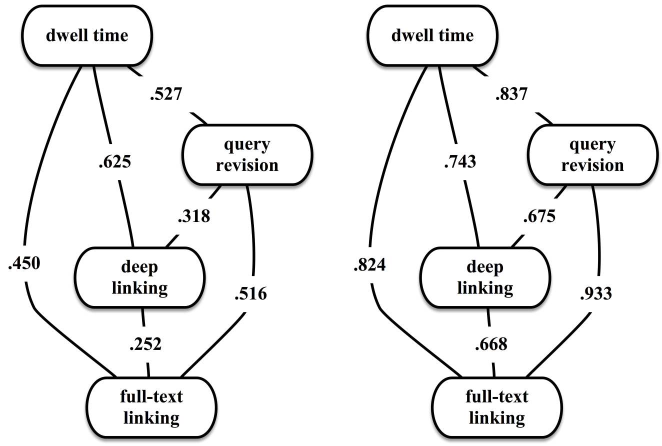
Regarding the query properties, we see a match with a finding of Holscher and Strube (2000), namely that users with little domain knowledge (novices), used longer queries than experts. A reason could be that domain experts know more effective query terms, and needed fewer terms to formulate a query. Another matching finding of Holscher and Strube (2000) with our results is that experts were more inclined to select a target document for assessment than novices (see Table 6.12). This is also in line with the results of Brand-Gruwel et al. (2005), who found out that experts elaborate more often on the content and judge the information more often. Moreover, a similar finding is that experts process information more often than novices (Holscher and Strube, 2000).

In summary, we can assert that the first group shares traits (or stereotypes) that can be matched with novice users. The second group can be matched with expert archival users. Moreover, our analysis in subsection 6.4.3 showed that there are statistical significant differences between the mean values using the implicit features as independent variables.

6.4.5 Correlations

We check the correlations between the dwell time, query revision, deep linking, and full-text linking for both groups using Pearson product-moment correlation coefficient. The correlation values with the novices are presented in Figure 6.1(a), and with the experts in Figure 6.1(b). These values are significant ($p < 0.01$, 2-tailed). Regarding the novices, we observe a strong correlation between dwell time and deep linking. There is medium correlation between dwell time and query revision, dwell time and full-text linking, and between query revision and full-text linking. There is weak correlation between query revision and deep linking, and between deep linking and full-text linking.

Interestingly, we see strong correlations between all variables in the expert user group (see Figure 6.1(b)). This means that the variables are statistical de-



(a) Novice group. (b) Expert group.

FIGURE 6.1: Correlation ($p < 0.01$) between the variables in both the novice and expert groups: dwell time, query revision, deep linking and full-text linking.

pendent on each other in this group. There is very strong correlation (0.933) between query revision and full-text linking, which may be due to the fact that each time a query is revised, the search procedure is in fact re-started as indicated by the full-text links in the summary and page views of the system.

In this section, we identified different groups of searchers corresponding to "novice" and "expert" stereotypes, and saw that these groups exhibit significantly different information seeking behavior: where "novices" follow a hit and run approach, the "experts" actively and interactively explore the information available. In the next section, we try to determine what is the best search system for these different searcher stereotypes.

6.5 IR Evaluation in Context

In this section, we use the interaction data of particular groups of users for contextual evaluation, trying to answer what type of system is best for their types of queries and their choice of results to inspect in detail. In the previous section, we saw significant differences between the interactions of "novices" and "experts" in the archives. Are they also served best by different systems? Or is the same system best for all types of users?

6.5.1 Experimental Setup

Next, we describe retrieval experiments that use the extracted interaction data for a search log-based context-sensitive IR evaluation. We study the transactions between December 31 2008 till January 31 2009—this is a month of data—and focus on the use of EAD files in particular. The information contained in this search log has been recorded in the *W3C Extended Logfile Format* (ELF), which includes a date, a timestamp of a hit, unique identifier for the user, the URL of the link that was visited, the query string, a browser identifier, a referral, and hits were recorded in detail within each second.

The log file—of the month January 2009—is 3.8 GB in size, and after filtering transactions of only EAD hits, the object of study in our experiments is 304.4 MB of data. Moreover, we index 5,934 EAD files—937 MB of data obtained from the *Nationaal Archief* —that could be found in the log. The experimental setup is replicated from the study as described in Chapter 3.

6.5.2 Filtering Assessments from User Groups

In Chapter 3, it is explained how the log files can be used to construct a massive and reliable test collection for IR evaluation. But can we use interaction data in the log files to evaluate the IR effectiveness of specific user groups?

The step is to construct the test collections using the subsets of sessions that have been identified, namely (1) all single-visit sessions, and (2) all sessions that can be traced back to a user with more than 10 visits. We have asserted that the former group can be related to novice users, and the latter group to archival experts. We have large sets of queries and corresponding clicks, both from the selection of search results, and from further browsing within the selected results. We make the reasonable assumption that clicks correspond to results that a user purposefully wants to inspect in full detail, which is related to the relevance of the result (although not necessarily in a strict sense of topically relevant). In short, we treat clicks as pseudo-relevance judgements, and assume that a system that ranks "clicked" results higher is a better system. Using the two lists of sessions, we can derive 2 types of test collections from the log file. Table 6.14 shows that both test collections are large enough (Spärck Jones and van Rijsbergen, 1976).

Table 6.14: Properties of the test collections: number of topics and the number of relevant results.

	#Topics	**#Relevant**
Novice	1,374	1,758
Expert	1,687	3,027

6.5.3 Results

Figure 6.2(a) shows the line graph of the 'novice' group. Figure 6.2(a) shows the line graph of the 'expert' group. Tables 6.15 and 6.16 further summarize the results of our experiments. In our evaluation, we used three IR measures. We first treat every click as a binary relevance judgement. This is captured by

Table 6.15: System-ranking of runs evaluated against judgments from 'novices.'

	MAP	MRR	nDCG
BOOL	0.1716 (5)	0.1791 (5)	0.2686 (5)
LM	0.2753 (4)	0.2935 (4)	0.3607 (4)
LMS	0.2763 (3)	0.2941 (3)	0.3761 (3)
NLLR	0.2817 (2)	0.2995 (2)	0.3827 (2)
OKAPI	0.2912 (1)	0.3102 (1)	0.3920 (1)

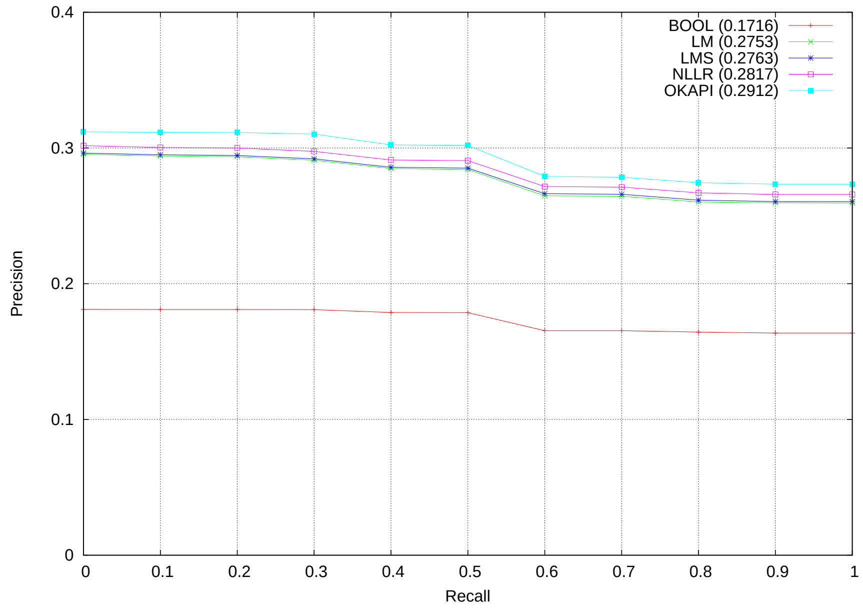
Table 6.16: System-ranking of runs evaluated against judgments from 'experts.'

	MAP	MRR	nDCG
BOOL	0.1070 (5)	0.1217 (5)	0.2041 (5)
LM	0.2436 (4)	0.2982 (4)	0.3272 (4)
LMS	0.2480 (3)	0.3025 (3)	0.3398 (3)
NLLR	0.2492 (2)	0.3037 (2)	0.3419 (2)
OKAPI	0.2666 (1)	0.3272 (1)	0.3616 (1)

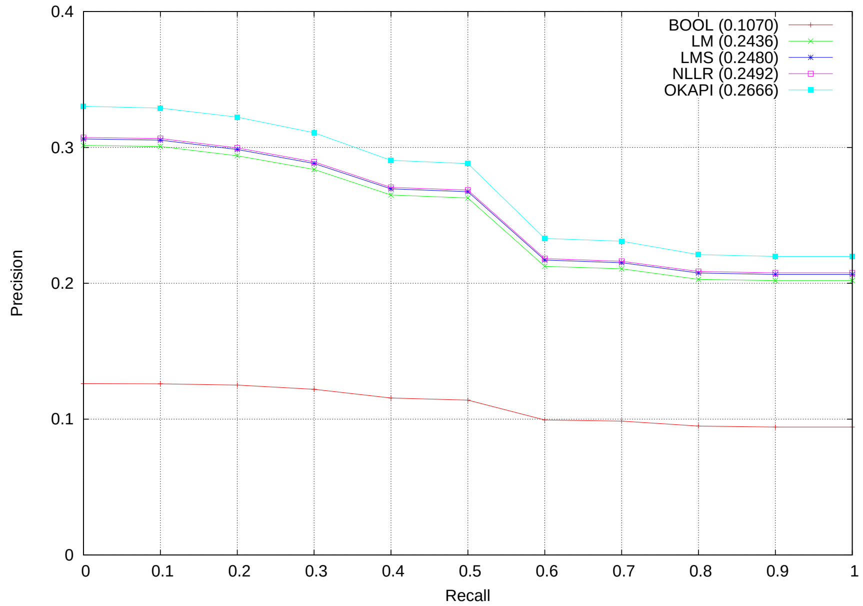
Mean Average Precision (MAP), which is the most frequently used summary measure for a set of ranked results, and Mean Recipropal Rank (MRR). The MRR is a static measure that looks at the rank of the first relevant result for each topic, and focuses on precision. Moreover, we can also use the number of clicks on each result by different results as a form of graded relevance judgment using the Normalized Discounted Cumulative Gain (nDCG).

We observe that the MAP scores are higher when evaluating the derived set of topics from novices than experts. This can be clarified by one of the conclusions of Saracevic (1975, p.163) that the level of subject knowledge varies inversely with the number of relevant judged documents, i.e. "the less the subject knowledge, the more lenient are their judgements."

We take notice that the system rankings are completely in line for the two groups. Okapi is the best performing, BOOL is the worst performing run—which was expected beforehand, and LM smoothing helps the retrieval performance. The MRR scores are also higher than the MAP values, the difference is greater with experts than novices, and the scores are also close between both evaluation sets. We check for statistical significance using the paired samples t-test. We start with Table 6.15. We focus on the MAP scores for the significance tests. BOOL is significantly performing worst. There is a minor but significant improvement of 0.36% for LMS over LM (t(1373) = 3.81, $p < 0.0005$, one-tailed). There is also a significant improvement of 1.95% of NLLR over LMS (t(1373) = 3.00, $p < 0.005$, one-tailed). We also see a significant improvement of 3.37% of OKAPI over NLLR ($p < 0.01$, one-tailed). Next, we focus on the results of Table 6.16. LMS has a significant 1.81% improvement over LM (t(1686) = 3.78, $p < 0.0005$, one-tailed). The difference between LMS and NLLR is only 0.48%,



(a) Precision and recall graphs for the 'novices.'



(b) Precision and recall graphs for the 'experts.'

FIGURE 6.2: Precision and recall line graphs.

but significant ($p < 0.05$). OKAPI has a 6.98% improvement over NLLR, and this was also significant ($p < 0.0005$). These findings are the same as compared to our retrieval experiments in Chapter 3.

In this section, we used the interaction data of particular groups of users for contextual evaluation. In the previous section, we saw significant differences between the interactions of "novices" and "experts" in the archives. There is an open debate in archival science whether the currently used systems, which are tailored to archival experts, are also suitable for novices like incidental visitors to archival Web sites. Our results show that, in terms of retrieval effectiveness, the system that is best for the "experts" is also the best system for the "novices." Even though this result is limited to the options under consideration—we only explored variants of the ranking method—it is a reassuring result. It can also be viewed as a proof of concept of the approach, and further experiments could consider other document representations (such as user tags or queries), recommendation, or even experiments with interface changes in the wild.

6.6 Conclusion

We have investigated the complete Web search logs from the *Nationaal Archief* from 6 years. These logs represented the full searches of archival visitors who sought online archival access with EAD finding aids. The general question is whether we can derive context from the logs. If so, how we can use this for a context-sensitive IR evaluation?

Our results show that we can use the search logs to give more insight into the search behavior of archival users, and we have looked at several generic properties that can be extracted from the logs. These are query terms, session length, and session duration. The main finding is that the search log gives insight in the searches of archival users, which can be used to answer the currently open question on the effectiveness of archival access to currently available information and systems. There is an open debate in archival science whether the currently used systems, which are tailored to archival experts, are also suitable for novices, like incidental visitors, to archival web sites. We experimented with the visit count of a user to group user sessions. Our assumption was that more experienced users consult the EAD finding aids more frequently than novice users. Using implicit features that point to user interest, we have observed two very different interaction stereotypes. Our assertion is that we can match these to novice and expert user stereotypes. Our main finding is that novice and expert searchers exhibit a significantly different information seeking behavior.

The results helped us in constructing a test collection for each group. We can treat each click to a file—one which can (in)directly be traced back to a query—as a pseudo-relevance judgment. The scores for experts were lower than for experts, but the system rankings were the same for the test collections. Our main finding is that, despite significantly different search episodes reflected by their specific information requests and choice of results to inspect in detail, both the experts and the novices are best served by the same system.

To conclude, the results show that the same IR system is working the best for two sets of relevance judgments coming from two different pools of users. The

MAP scores were considerable lower when evaluated using the set of relevance judgments. This could mean that the topics coming from these users are harder, i.e. to deliver the relevant results. This explains the higher dwell time. On the one hand, expert archival users search exhaustively for information, and it would benefit them to improve the IR effectiveness of a system in terms of the MAP scores. However, for the early precision, the difference between both groups is minor using the same systems. This implies, on the other hand, that we could also avoid complicated system personalization issues to fine-tune the IR effectiveness of a system for different users—and use one archival search system for everyone.

Chapter 7

Conclusions

Archival finding aids in EAD are put online. We do not know how users interact with these finding aids, and what type of system is needed to support them in their search. Therefore, we have applied XML retrieval techniques to the EAD finding aids, developed system evaluation of EAD retrieval, and studied the search behavior of archival searchers. The system evaluation consisted of studying effective retrieval techniques tailored to EAD finding aids, taking into account different user stereotypes and contexts, and the textual context of the unit that needs to be returned. Now, we present our findings and contributions.

7.1 Findings and Contributions

First, we present the main contributions, then break-down with an overview of our general contributions and findings emanating from the chapters, then finally the contributions to information retrieval research and archival science.

7.1.1 Main Findings and Contributions

The research scope of this dissertation titled *System Evaluation of Archival Description and Access* is primarily access to archival finding aids in EAD. We asked:

– *With large numbers of archival finding aids published online in EAD, how do searchers interact with these finding aids, and what type of retrieval system is needed to support them?*

The solution is: apply XML retrieval techniques to EAD finding aids, develop system evaluation of EAD retrieval, and study the search behavior of archival users interacting with EAD finding aids. There are three main contributions:

1. the building—for the first time—of a re-usable test collection for EAD finding aids which can be used to automatically measure the performance of search systems driven by EAD finding aids (see Chapters 2, 3 and 4).

2. IR experimentation with EAD retrieval, both across EAD finding aids and within finding aids primarily by treating key archival principles—provenance and original order—as independent variables (see Chapter 4).
3. deepening the understanding of archival users' search behavior—for novices and experts—when they search within EAD finding aids through formal and quantitative analysis and experimentation (see Chapters 5 and 6).

The system evaluation of EAD finding aids is an IR evaluation research methodology to gauge the IR effectiveness. The system evaluation is a different methodological leap forward that complements user studies in the archival domain, and increases our understanding of online information access with archival finding aids in EAD. Specifically, we list the general contributions and findings resulting from our system evaluation of archival description and access next.

7.1.2 General Contributions and Findings

1. **Supporting search in EAD finding aids with XML retrieval.**

   Contributions
   - Showing how XML retrieval may help getting access to archival records via descriptions in EAD finding aids (see Chapter 2).
   - Using the granularity and special archival structure in EAD finding aids to provide a two-tier approach to access, namely whole document ranking (finding the collection) and then element relevance ranking (finding the description) (see Chapter 2).

   Methods
   - We developed a system to move towards a tangible construct. Evans and Rouche (2004, p.315) point to a methodological issue by discussing the use of systems development research methods, and already suggest that adopting a user-centric prototyping approach in a research context allows for exploration of the interplay between theory and practice, advancing the practice, while also offering new insights into theoretical concepts. Therefore, we added a component in archival research methods (Gilliland-Swetland and Mckemmish, 2004). We investigated the access component in archival theory by developing a system.

   Findings
   - We formally introduced our system that employs XML retrieval techniques as a more focused means to gain access to online digital archives, effectively exploiting the structure to search and find valuable information. A result is the implementation of our domain-specific XML IR system driven by archival finding aids in EAD: *README* (Retrieving Encoded Archival Descriptions More Effectively). We showed how XML IR can be fruitfully applied on EAD for archival access.

Reflection

- This dissertation's research focused on evaluating archival access to EAD finding aids, and it did not aim to find out how to design the most usable interface for archival finding aids in EAD, which would be an area too broad to explore. However, the retrieved information as expressed in a query should be juxtaposed with proper interfaces. How to make the information and related services visible to users with services and functions constitute what access is.
- More types of system based on README with different ranking algorithms were also possible, but the research baseline has been established.

2. **System evaluation of whole document (collection) ranking tailored to archival finding aids in EAD.**

   Contributions

   - System evaluation of EAD retrieval based on the first tier, namely whole document ranking (see Chapter 3).
   - Using a search log and external sources for evaluation tailored to the document collection, users, search requests and relevance assessments (see Chapter 3).

   Methods

   - We conducted a study using a large set of EAD finding aids and search logs of the *Nationaal Archief* EAD search system. These logs cover several years of interaction with this system. This resulted in a test collection to evaluate the retrieval of EAD finding aids. We presented generic methods to derive a domain-specific test collection, and used a range of retrieval models to determine the effectiveness of the test collections. Our extraction methods are naive—we treat every clicked document as pseudo-relevant—but suffice to illustrate the viability of the approach.
   - We validated the results against a set of traditional topics derived from email requests to the archive and the archivist's responses. Moreover, we compared the test collections with a test collection derived from the *Nationaal Archief* research guide that consists of questions and recommended EAD finding aids.

   Findings

   - We found complete agreement between the log-based evaluation and the traditional topics derived from email requests. However, when we extended the experimentation by adding one extra system, namely the *Nationaal Archief* system itself, we have to nuance our findings, as the system rank correlations then changes.
   - When we repeated the experimentation by deriving another test collection from the *Nationaal Archief* research guide, we also saw similarities and differences, which may due to the recall-based orientation of

this test collection, compared to the log-based approach that focused on precision.

Reflection

- The results of our 'naive' approach to constructing an IR test collection for evaluation look promising, but still more research and further experimentation is necessary. Pointers to further research include investigating more advanced click models and testing with more types of systems. More complex models of interaction, (e.g., think of Dupret and Liao, 2010) will likely generate a more faithful test collection.
- Alonso et al. (2008) have explored crowdsourcing for relevance evaluation, and this could also be a promising future direction for automated system evaluation of archival access.

3. System evaluation of searching across and within EAD finding aids.

Contributions

- Study on the unit of retrieval shows that a wide range of elements is consulted by searchers, which supports the element relevance ranking and aggregation-based approaches in Chapter 2 (see Chapter 4).
- System evaluation of retrieval within an EAD finding aid with XML retrieval techniques (second tier), after the system evaluation of whole document ranking (first tier) in Chapter 3 (see Chapter 4).

Methods

- First, we have started our investigation by looking at what elements people use when an EAD search system retrieves archival descriptions.
- Second, we have investigated the effectiveness of retrieval with provenance in an analysis by looking at the number of clicks that occur within a certain EAD finding aid given a specific search request, and by comparing the grouping of elements by EAD finding aid with standard element relevance ranking in a retrieval experiment. We borrowed an economic term called the concentration ratio, used to measure the dominance of a company in an industry. Similarly, the principle of provenance states that an archive is an organic whole, and since an EAD finding aid is a representation of an archive, its consequence is that one EAD finding aid should also have a 'monopoly' for a particular search request.
- Third, we empirically examined the impact of the archival principle of original order on the ranking of search results by comparing it with a standard metadata retrieval system using XML retrieval techniques.

Findings

- We found that this depends on the user's information need, since there is not a clear favorite element that people use. When people are searching for contextual information, the `<bioghist>` is a 'popular' element.

When people search for archival materials, the higher level `<c>` elements stand out. This suggests that indexing and retrieving all and any elements, as the README system does, is the most practical solution for a uniform archival search system.
- We found that people search across EAD finding aids, but our analysis and experiment points to evidence that the *provenance* is dominant and is effective for *retrieval*.
- Our results show that the relevance ranking is of paramount importance, but that the results have a (weak) tendency to cluster. Original order is useful, because physical materials can only be ordered in a single way, and here the traditional archival practices make much sense.

Reflection

- With EAD finding aids, we are no longer bound to the physical and practical limitation of before and we could construct multiple ordering of the same material including those based on a search request or search profile at hand. This opens up a wealth of possibilities to improve access, enabling new and more effective usage of archival description.

4. Observing search behaviors of users within EAD finding aids.

Contributions

- Description of a formal model to capture and measure the search behavior of people within an EAD finding aid (see Chapter 5).
- Using the search log to analyze the information searching behavior within an EAD finding aid. (see Chapter 5)

Methods

- We have followed a search log-based approach combined with formal user-centric experimentation. We present a state diagram and use this diagram to capture and measure the search behavior within EAD finding aids.
- We have an independent groups design, and look at the variance within the groups with repeated measures. The single (dichotomous) independent variable is the visit count. There are three groups.
- We control the results by grouping (or creating 'bins' of) the distributions of the transitions by the same begin and end state of a session, resulting in two conditions in each group, namely (i) start from top to bottom, and (ii) start from bottom. A session is here defined as a stream of clicks belonging to a user, separated by an interval of 30 minutes of inaction, in the same EAD finding aid.

Findings

- We introduced and used the EAD Search Behavior model, which is a state diagram that captures the search behaviors of archival users when they interact in an EAD finding aid.

- People with no previous experience with EAD finding aids have fewer interaction, thus search less deep, have a shorter search time, and search in fewer (different) EAD finding aids, than people with prior experience. A similarity across the groups is: starting in the top likely leads to forward-going behavior, and starting deep likely leads to backtracking behavior.
- This interaction changes as people get more experienced, and our results have shown that the search behavior becomes more like of experienced users.
- 'Popular' EAD finding aids are not heavily used when these mostly consist of single-clicks or 'shallow' exploratory behavior in the top of an EAD finding aid, and are used when people search in the inventory. The results illustrate the difficulty of providing focused access to EAD finding aids. Our data and analysis show that users go to the inventory frequently, and this suggests that they also want to go there.

Reflection

- We described the results of a study that is mainly observational and uses one data source, namely the *Nationaal Archief* Web search logs. It adds more to our understanding of archival search behavior, and complements other findings derived from other research methodologies like interviews (Duff and Johnson, 2002, 2003) or models (Gilliland-Swetland, 2001). Hsieh-Yee (2001, pp.180–181) also mentions that other studies have described the behavior of a group of searchers in a particular environment performing tasks of their own or as given by researchers in experimental laboratory studies. This study resembled the former type.
- As we gain more understanding of how archival users search, we can improve access. We can conceptualize search profiles, but the next step could be personalizing a search system so all users are able to use the EAD findings aids as exhaustively as users with more archival search experience are already doing. Our results also add support to adapting a system that entices users to start using an EAD finding aid, because even a majority of the 'popular' EAD finding aids does not seem to get optimally used. The solution is working towards more effective focused access, because the archival finding aid in EAD starts to get really used as people get closer to the descriptions deep down.

5. **System evaluation based on whole document ranking for specific user stereotypes, in particular focusing on prior experience with searching EAD finding aids.**

   Contributions

   - Transaction log analysis of searching with EAD finding aids with a break-down per visit-count of users (see Chapter 6).

  – System evaluation at the whole document level (first tier, see Chapter 2) tailored to specific user groups, in particular 'novice' and 'expert' users (see Chapter 6).

Methods

  – We have investigated the complete search logs from the *Nationaal Archief* from six years with a transaction log analysis. These logs represented the full searches of archival visitors who sought online archival access with archival finding aids in EAD. We experimented with the visit count of a user to group user sessions. Our assumption was that more experienced users consult the EAD finding aids more frequently than novice users. Using implicit features that point to user interest, we have observed two very different interaction stereotypes. We conducted system-centered IR experimentation based on these user stereotypes.

Findings

  – The search log gives insight in the searches of archival users, which can be used to answer the currently open question on the effectiveness of archival access to currently available information and systems.
  – Our assertion is that we can match these to novice and expert user stereotypes. Our main finding is that novice and expert searchers exhibit different information seeking behaviors.
  – The results from this study helped us in constructing two test collections for each group. We can treat each click to a file—one which can (in)directly be traced back to a query—as a pseudo-relevance judgment. The scores for experts were lower than for experts, but the system rankings were the same for test collections. Our main finding is that, despite significantly different search episodes reflected by their specific information requests and choice of results to inspect in detail, both groups are best served by the same type of system.

Reflection

  – The same IR system is working the best for two sets of relevance judgments coming from two different pools of users. The MAP scores were considerable lower when evaluated using the set of relevance judgments of 'experts.' This could mean that the topics coming from these users are harder, i.e. to deliver the relevant results. This explains the higher dwell time. On the one hand, expert archival users search exhaustively for information, and it would benefit them to improve the IR effectiveness of a system in terms of the MAP scores. However, for the early precision, the difference between both groups is minor using the same systems. This implies, on the other hand, that we could also avoid complicated system personalization issues to fine-tune the IR effectiveness of a system for different users—and use one archival search system for everyone.

7.1.3 IR and Archival Science Contributions

The specific contributions are for

- *Information Retrieval:*
  - System evaluation of an important 'real' and domain-specific search task.
  - Usage of transaction logs for deriving domain specific test collections.
  - Analysis of search behavior in yet unexplored structured documents.
  - Tailoring IR evaluation to specific searcher stereotypes.
- *Archival Science:*
  - Investigation of retrieval aspects of access to EAD finding aids.
  - Done a large scale system evaluation of EAD retrieval.
  - Insight in the online consultation of EAD finding aids.
  - Found support for the relative effectiveness of traditional archival structure in EAD finding aids for access.

7.2 Future Work

There are also limitations of our research in terms of scope, and could be considered research opportunities in the future.

- We only studied EAD finding aids based on traditional inventories of paper archives. These finding aids will remain essential to give access to historical records, because even when parts of the materials are digitized, they are not yet machine-readable. There are also alternative forms of archival description with different types of finding aids or surrogates, for example based on transcriptions or tagging. These finding aids are outside of the current scope, but system evaluation can be applied on these surrogates as well.
- We have not investigated the evaluation in terms of user interface, interaction design, or system efficiency. We looked at the IR effectiveness only by using test collections. In the future, these other evaluation methodologies can complement the system evaluation as illustrated in this dissertation.
- While the log-based test collections are reusable to evaluate new systems under the same experimental conditions (i.e. same document collections, same topics and relevance assessments), the reusability of the evaluation is limited by the bias of the *Nationaal Archief* system. The evaluation may underestimate the performance of a radically different ranking model, and hence care must taken when evaluating systems that deviate substantially from standard text retrieval systems as explored in our experiments.

Bibliography

Agichtein, E., Brill, E., and Dumais, S. (2006). Improving Web search ranking by incorporating user behavior information. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 19–26, New York, NY, USA. ACM. (Cited on page 116.)

Allen, R. B. (2005). Using information visualization to support access to archival records. *Journal of Archival Organization*, 3(1):37–49. (Cited on page 31.)

Alonso, O., Rose, D. E., and Stewart, B. (2008). Crowdsourcing for relevance evaluation. *SIGIR Forum*, 42:9–15. (Cited on page 138.)

Anderson, I. (2004). Are you being served? historians and the search for primary sources. *Archivaria*, 58(1):81–129. (Cited on page 86.)

Aula, A., Khan, R. M., and Guan, Z. (2010). How does search behavior change as search becomes more difficult? In *Proceedings of the 28th international conference on Human factors in computing systems*, CHI '10, pages 35–44, New York, NY, USA. ACM. (Cited on page 84.)

Baeza-Yates, R. A. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA. (Cited on page 2.)

Bailey, P., Craswell, N., and Hawking, D. (2003). Engineering a multi-purpose test collection for Web retrieval experiments. *Inf. Process. Manage.*, 39(6):853–871. (Cited on page 36.)

Balatsoukas, P. and Demian, P. (2010). Effects of granularity of search results on the relevance judgment behavior of engineers: Building systems for retrieval and understanding of context. *J. Am. Soc. Inf. Sci. Technol.*, 61:453–467. (Cited on page 3.)

Barrett, A. (2005). The information-seeking habits of graduate student researchers in the humanities. *The Journal of Academic Librarianship*, 31(4):324 – 331. (Cited on page 86.)

Bates, M. J. (1990). Where should the person stop and the information search interface start? *Information Processing & Management*, 26(5):575 – 591. (Cited on pages 104 and 116.)

Bearman, D. (1979). Automated access to archival information: Accessing systems. *American Archivist*, 42(2):179–190. (Cited on pages 2, 10, 14, and 58.)

Bearman, D. (1992). Documenting documentation. *Archivaria*, 34(1):33–49. (Cited on page 2.)

Bearman, D. and Lytle, R. (1985). The power of the principle of provenance. *Archivaria*, 21(1). (Cited on pages 69 and 79.)

Beg, M. S. and Ahmad, N. (2007). Web search enhancement by mining user actions. *Information Sciences*, 177(23):5203 – 5218. (Cited on page 116.)

Belkin, N. J. (1993). Interaction with texts: Information retrieval as information-seeking behavior. In *Information Retrieval'93*, pages 55–66, Konstanz. Universitetsverlag Konstanz. (Cited on pages 81 and 82.)

Belkin, N. J. and Croft, W. B. (1992). Information filtering and information retrieval: two sides of the same coin? *Commun. ACM*, 35:29–38. (Cited on page 44.)

Belkin, N. J., Oddy, R. N., and Brooks, H. M. (1982). ASK for Information Retrieval: Part I. Background and Theory. *Journal of Documentation*, 38(2):61–71. (Cited on page 81.)

Berners-Lee, T., Cailliau, R., Luotonen, A., Nielsen, H. F., and Secret, A. (1994). The world-wide Web. *Commun. ACM*, 37:76–82. (Cited on pages 9 and 10.)

Boag, S., Chamberlin, D., Fernández, M., Florescu, D., Robie, J., and Siméon, J. (2002). XQuery 1.0: An XML query language. W3C Working Draft. `http://www.w3.org/TR/xquery/`. (Cited on page 19.)

Boles, F. (1982). Disrespecting original order. *American Archivist*, 45(1):26–33. (Cited on page 74.)

Boncz, P. A., Grust, T., van Keulen, M., Manegold, S., Rittinger, J., and Teubner, J. (2006). MonetDB/XQuery: A Fast XQuery Processor Powered by a Relational Engine. In *SIGMOD '06*, pages 479–490. ACM. (Cited on page 23.)

Borgman, C. L., Gallagher, A. L., Hirsh, S. G., and Walter, V. A. (1995). Children's searching behavior on browsing and keyword online catalogs: the science library catalog project. *J. Am. Soc. Inf. Sci.*, 46:663–684. (Cited on page 85.)

Bosak, J. and Bray, T. (1999). XML and the Second-Generation Web. *Scientific American*, 280(5):89–93. (Cited on pages 9, 10, and 65.)

Brand-Gruwel, S., Wopereis, I., and Vermetten, Y. (2005). Information problem solving by experts and novices: analysis of a complex cognitive skill. *Computers in Human Behavior*, 21(3):487–508. (Cited on pages 115 and 128.)

Bray, T., Paoli, J., Sperberg-McQueen, C. M., Maler, E., and Yergeau, F. (2008). Extensible Markup Language (XML) 1.0 (Fifth Edition). (Cited on pages 9 and 65.)

Campbell, R. R. (1967). Automation and information retrieval in archives—the broad concepts. *American Archivist*, 30(2):279–286. (Cited on pages 69 and 79.)

Chamberlin, D. (2003). Xquery: a query language for xml. In *Proceedings of the 2003 ACM SIGMOD international conference on Management of data*, SIGMOD '03, pages 682–682, New York, NY, USA. ACM. (Cited on page 19.)

Chandler, R. L. (2002). Museums in the online archive of california (moac): Building digital collections across libraries and museums. *First Monday*, 7(5). (Cited on page 11.)

Chapman, C. N., Love, E., Milham, R. P., ElRif, P., and Alford, J. L. (2008). Quantitative evaluation of personas as information. In *Human Factors and Ergonomics Society Annual Meeting Proceedings*, volume 52, pages 1107–1111. Human Factors and Ergonomics Society. (Cited on page 114.)

Chen, S. F. and Goodman, J. (1996). An empirical study of smoothing techniques for language modeling. In *ACL*, pages 310–318. (Cited on page 44.)

Clark, J. and DeRose, S. (1999). Xml path language (xpath) version 1.0. Technical report, World Wide Web Consortium. (Cited on page 16.)

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences (2nd Edition)*. Routledge Academic, 2 edition. (Cited on page 99.)

Cook, T. (1997). What is past is prologue: A history of archival ideas since 1898, and the future paradigm shift. *Archivaria*, 43(1):17–63. (Cited on page 58.)

Cox, R. (2008). Revisiting the archival finding aid. *Journal of Archival Organization*, 5(4):134–140. (Cited on pages 59 and 78.)

Crymble, A. (2010). An analysis of Twitter and Facebook use by the archival community. *Archivaria*, 70(2):125–151. (Cited on page 7.)

Daniels, M. G. and Yakel, E. (2010). Seek and you may find: Successful search in online finding aid systems. *American Archivist*, 73(2):535–568. (Cited on pages 37 and 112.)

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3):pp. 319–340. (Cited on page 115.)

Denoyer, L. and Gallinari, P. (2006). The Wikipedia XML corpus. *SIGIR Forum*, 40:64–69. (Cited on page 15.)

DeRose, S. J. (1997). Navigation, access, and control using structured information. *American Archivist*, 60(3):298–309. (Cited on pages 12, 16, and 63.)

Duff, W. M., Craig, B., and Cherry, J. (2004). Historians' use of archival sources: Promises and pitfalls of the digital age. *The Public Historian*, 26(2):7–22. (Cited on page 2.)

Duff, W. M. and Johnson, C. A. (2002). Accidentally found on purpose: information-seeking behavior of historians in archives. *Library Quarterly*, 72:472–96. (Cited on pages 86, 110, and 140.)

Duff, W. M. and Johnson, C. A. (2003). Where is the list with all the names? information-seeking behavior of genealogists. *American Archivist*, 66(1):79–95. (Cited on pages 110, 112, 114, 117, 118, 122, 128, and 140.)

Duff, W. M. and Stoyanova, P. (1998). Transforming the crazy quilt: Archival displays from a user's point of view. *Archivaria*, 45(1):44–79. (Cited on pages 18, 33, 65, and 112.)

Duff, W. M., Yakel, E., Tibbo, H. R., Cherry, J. M., McKay, A., Krause, M. G., and Sheffield, R. (2010). The development, testing, and evaluation of the Archival Metrics Toolkits. *American Archivist*, 73(2):569–599. (Cited on page 112.)

Dumais, S., Joachims, T., Bharat, K., and Weigend, A. (2003). SIGIR 2003 workshop report: implicit measures of user interests and preferences. *SIGIR Forum*, 37(2):50–54. (Cited on pages 36, 111, 112, and 114.)

Dupret, G. and Liao, C. (2010). A model to estimate intrinsic document relevance from the clickthrough logs of a Web search engine. In *Proceedings of the third ACM international conference on Web search and data mining*, WSDM '10, pages 181–190, New York, NY, USA. ACM. (Cited on pages 36 and 138.)

Duranti, L. (1992). Origin and development of the concept of archival description. *Archivaria*, 35(1). (Cited on page 58.)

Duranti, L. (2001). The impact of digital technology on archival science. *Archival Science*, 1:39–55. (Cited on page 8.)

Ellis, D. (1989). A behavioral approach to information retrieval system design. *Journal of Documentation*, 45(3):171–212. (Cited on pages 114, 115, and 117.)

Ellis, D. (1997). Modelling the information seeking patterns of engineers and research scientists in an industrial environment. *Journal of Documentation*, 53(4):384–403. (Cited on page 115.)

Ellis, D., Cox, D., and Hall, K. (1993). A comparison of the information seeking patterns of researchers in the physical and social sciences. *Journal of Documentation*, pages 356–369. (Cited on page 115.)

Evans, J. and Rouche, N. (2004). Utilizing Systems Development Methods in Archival Systems Research: Building a Metadata Schema Registry. *Archival Science*, 4(3-4):315–334. (Cited on pages 8 and 136.)

Evans, M. J. (1986). Authority control: An alternative to the record group concept. *American Archivist*, 49(3):249–261. (Cited on pages 78 and 79.)

Fachry, K. N., Kamps, J., and Zhang, J. (2008). Access to archival material in context. In *IIiX '08*, pages 102–109, New York, NY, USA. ACM. (Cited on page 31.)

Fang, W. (2007). Using Google Analytics for improving library Website content and design: A case study. *Library Philosophy and Practice 2007: Special Issue on Google and Libraries*. (Cited on page 114.)

Feeney, K. (1999). Retrieval of archival finding aids using World-Wide-Web search engines. *American Archivist*, 62(2):206–228. (Cited on page 37.)

Forde, H. (2005). Access and preservation in the 21st century: What has changed? *Journal of the Society of Archivists*, 26(2):193–200. (Cited on page 10.)

Fox, E. A. and Sornil, O. (1999). Digital libraries. In *Modern Information Retrieval*, pages 415–432. Addison Wesley, 1st edition. (Cited on page 33.)

Fuhr, N., Kamps, J., Lalmas, M., Malik, S., and Trotman, A. (2007). Overview of the INEX 2007 ad hoc track. In *INEX*, pages 1–23. Springer. (Cited on page 15.)

Fuhr, N., Malik, S., and Lalmas, M. (2004). Overview of the initiative for the evaluation of XML retrieval (INEX) 2003. In *INEX*, pages 15–17. (Cited on page 15.)

Gilliland-Swetland, A. J. (2001). Popularizing the finding aid: exploiting EAD to enhance online discovery and retrieval in archival information systems by diverse user groups. *Journal of Internet Cataloging*, 4(3):199–225. (Cited on pages 2, 58, 110, 113, 117, and 140.)

Gilliland-Swetland, A. J. (2005). Electronic records management. *Annual Review of Information Science and Technology*, 39(1):219–253. (Cited on pages 30, 58, and 59.)

Gilliland-Swetland, A. J. and Mckemmish, S. (2004). Building an infrastructure for archival research. *Archival Science*, 4(3):149–197. (Cited on pages 8 and 136.)

Goldberg, K. H. (2008). *XML: Visual QuickStart Guide*. Peachpit Press, Berkeley, CA, USA, 2nd edition. (Cited on pages 9 and 10.)

Goldfarb, C. F. (1984). The standard generalized markup language: Basic concepts. In *Offene Multifunktionale Büroarbeitsplätze und Bildschirmtext*, pages 132–140. (Cited on page 9.)

Gövert, N. and Kazai, G. (2002a). Overview of the initiative for the evaluation of XML retrieval (INEX) 2002. In *INEX*, pages 1–17. (Cited on pages 14 and 15.)

Gövert, N. and Kazai, G. (2002b). Overview of the initiative for the evaluation of XML retrieval (INEX) 2002. In *INEX*, pages 1–17. (Cited on page 36.)

Hammer-Aebi, B., Christensen, K. W., Lund, H., and Larsen, B. (2006). Users, structured documents and overlap: interactive searching of elements and the influence of context on search behaviour. In *Proceedings of the 1st international conference on Information interaction in context*, IIiX, pages 46–55, New York, NY, USA. ACM. (Cited

on pages 3, 14, 16, and 24.)
Hatcher, E. and Gospodnetic, O. (2004). *Lucene in Action*. Manning Publications Co., Greenwich, CT, USA. (Cited on page 19.)
Haworth, K. M. (2001). Archival description: Content and context in search of structure. *Journal of Internet Cataloging*, 4:1091–1367. (Cited on pages 1, 10, 26, 27, 58, 59, and 65.)
Hedstrom, M. (2002). Archives, memory, and interfaces with the past. *Archival Science*, 2:21–43. 10.1007/BF02435629. (Cited on page 31.)
Hiemstra, D., Rode, H., van Os, R., and Flokstra, J. (2006). PF/Tijah: text search in an XML database system. In *OSIR '06*, pages 12–17. (Cited on pages 23 and 73.)
Hill, A., Stockting, B., and Higgins, S. (2005). Different strokes for different folks: Presenting EAD in three UK online catalogues. *Journal of Archival Organization*, 3(2-3):183–206. (Cited on page 11.)
Hjørland, B. (2000). Documents, memory institutions and information science. *Journal of Documentation*, 56:27–41. (Cited on page 11.)
Hodkinson, C., Kiel, G., and McColl-Kennedy, J. R. (2000). Consumer Web search behaviour: diagrammatic illustration of wayfinding on the Web. *Int. J. Hum.-Comput. Stud.*, 52:805–830. (Cited on pages 82 and 87.)
Holscher, C. and Strube, G. (2000). Web search behavior of internet experts and newbies. *Computer Networks*, 33(1-6):337–346. (Cited on pages 115, 127, and 128.)
Horsman, P. (2002). The last dance of the phoenix or the de-discovery of the archival fonds. *Archivaria*, 54(1):1–23. (Cited on page 57.)
Horsman, P., Ketelaar, F., and Thomassen, T. (2005). *Tekst en context van de handleiding voor het ordenen en beschrijven van archieven van 1898*. Verloren. (Cited on page 57.)
Hsieh-Yee, I. (2001). Research on Web search behavior. *Library & Information Science Research*, 23(2):167–185. (Cited on pages 110 and 140.)
Huffman, N. G. (2008). Search features and other characteristics of XML retrieval systems for EAD finding aids: A content analysis. MSc thesis. University of North Carolina, Chapel Hill. (Cited on page 18.)
Hutchinson, T. (1997). Strategies for searching online finding aids: A retrieval experiment. *Archivaria*, 44(2):72–101. (Cited on pages 36 and 112.)
Ingwersen, P. and Järvelin, K. (2005). *The Turn: Integration of Information Seeking and Retrieval in Context (The Information Retrieval Series)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA. (Cited on page 82.)
International Council on Archives (1999). ISAD(G): General International Standard Archival Description, Second edition. Technical report, International Council on Archives, Ottawa. (Cited on page 8.)
Islamaj Dogan, R., Murray, G. C., Névéol, A., and Lu, Z. (2009). Understanding PubMed user search behavior through log analysis. *Database : the journal of biological databases and curation*, 2009(0):18. (Cited on page 84.)
Jansen, B. J. (2006). Search log analysis: What it is, what's been done, how to do it. *Library & Information Science Research*, 28(3):407–432. (Cited on pages 35, 84, 111, 112, 114, and 118.)
Jansen, B. J. and Spink, A. (2006). How are we searching the World Wide Web? A comparison of nine search engine transaction logs. *Information Processing & Management*, 42(1):248 – 263. (Cited on page 84.)

Jansen, B. J., Spink, A., Bateman, J., and Saracevic, T. (1998). Real life information retrieval: a study of user queries on the Web. *SIGIR Forum*, 32(1):5–17. (Cited on pages 114, 117, and 118.)

Jansen, B. J., Spink, A., Blakely, C., and Koshman, S. (2007). Defining a session on Web search engines. *J. Am. Soc. Inf. Sci. Technol.*, 58(6):862–871. (Cited on pages 114, 117, and 121.)

Jansen, B. J., Tapia, A., and Spink, A. (2010). Searching for salvation: An analysis of US religious searching on the World Wide Web. *Religion*, 40(1):39 – 52. (Cited on page 84.)

Jardine, N. and van Rijsbergen, C. (1971). The use of hierarchical clustering in information retrieval. *Information Storage and Retrieval*, 7:217–240. (Cited on page 74.)

Jenkinson, H. (1944). Reflections of an Archivist. *Contemporary Review*, 165:355–361. (Cited on page 74.)

Joachims, T., Granka, L., Pan, B., Hembrooke, H., and Gay, G. (2005). Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '05, pages 154–161, New York, NY, USA. ACM. (Cited on page 36.)

Jones, S., Cunningham, S. J., McNab, R. J., and Boddie, S. J. (2000). A transaction log analysis of a digital library. *Int. J. on Digital Libraries*, 3(2):152–169. (Cited on pages 35, 114, 117, 118, 121, and 125.)

Joty, S. and Sadid-Al-Hasan, S. (2007). Advances in focused retrieval: A general review. In *Computer and information technology, 2007. iccit 2007. 10th international conference on*, pages 1–5. (Cited on page 12.)

Kamps, J. and Sigurbjörnsson, B. (2005). What do users think of an XML element retrieval system? In *INEX*, pages 411–421. (Cited on page 3.)

Kazai, G., Lalmas, M., and de Vries, A. P. (2004). The overlap problem in content-oriented XML retrieval evaluation. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '04, pages 72–79, New York, NY, USA. ACM. (Cited on pages 16, 24, 27, and 75.)

Kelly, D. (2009). Methods for evaluating interactive information retrieval systems with users. *Found. Trends Inf. Retr.*, 3:1–224. (Cited on page 82.)

Kelly, D. and Teevan, J. (2003). Implicit feedback for inferring user preference: a bibliography. *SIGIR Forum*, 37:18–28. (Cited on page 36.)

Ketelaar, E. (1996). Archival theory and the Dutch manual. *Archivaria*, 41(1). (Cited on pages 57, 58, 59, and 74.)

Ketelaar, E. (2008). Archives as spaces of memory. *Journal of the Society of Archivists*, 29(1):9–27. (Cited on page 1.)

Kieras, D. and Polson, P. G. (1985). An approach to the formal analysis of user complexity. *International Journal of Man-Machine Studies*, 22(4):365 – 394. (Cited on page 82.)

Kiesling, K. (1997). EAD as an Archival Descriptive Standard. *American Archivist*, 60(3):344–354. (Cited on page 11.)

Kiesling, K. (2001). Metadata, metadata, everywhere - but where is the hook? *OCLC Systems & Services*, 17:84–88. (Cited on pages 16 and 63.)

Kim, J., Oard, D. W., and Romanik, K. (2001). User modeling for information filtering based on implicit feedback. In *In ISKO-France 2001*. (Cited on page 36.)

Kinsella, J. and Bryant, P. (1987). Online public access catalog research in the united kingdom: An overview. *Library Trends*, 35(4):619–630. (Cited on page 113.)

Knuth, D. E. (1964). backus normal form vs. backus naur form. *Commun. ACM*, 7:735–736. (Cited on page 9.)

Koenemann, J. and Belkin, N. J. (1996). 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: common ground*, CHI '96, pages 205–212, New York, NY, USA. ACM. (Cited on page 115.)

Kraaij, W. (2004). *Variations on language modeling for information retrieval*. PhD thesis, University of Twente, Enschede. (Cited on page 44.)

Kuhlthau, C. C. (1991). Inside the search process: Information seeking from the user's perspective. *JASIS*, 42(5):361–371. (Cited on page 81.)

Lalmas, M. (2010). XML Information Retrieval. *Encyclopedia of Library and Information Sciences, Third Edition*, pages 5700–5712. (Cited on pages 3, 12, and 36.)

Larsen, B., Tombros, A., and Malik, S. (2006). Is XML retrieval meaningful to users?: searcher preferences for full documents vs. elements. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, pages 663–664, New York, NY, USA. ACM. (Cited on pages 3 and 14.)

Larson, R. R. and Sanderson, R. (2006). Cheshire3: retrieving from tera-scale grid-based digital libraries. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, pages 730–730, New York, NY, USA. ACM. (Cited on page 18.)

Lazonder, A. W., Biemans, H. J. A., and Wopereis, I. G. J. H. (2000). Differences between novice and experienced users in searching information on the World Wide Web. *Journal of the American Society for Information Science*, 51:576–581. (Cited on pages 15, 127, and 128.)

Lesk, M. (2004). *Understanding Digital Libraries, Second Edition (The Morgan Kaufmann Series in Multimedia and Information Systems)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. (Cited on page 15.)

Lowenthal, D. (2005). Natural and cultural heritage. *International Journal of Heritage Studies*, 11(1):81–92. (Cited on page 85.)

Lytle, R. H. (1980). Intellectual Access to Archives: I. Provenance and Content Indexing Methods of Subject Retrieval. *American Archivist*, 43(1):64–75. (Cited on pages 14, 36, 74, and 85.)

MacNeil, H. (2008). Archivalterity: Rethinking original order. *Archivaria*, 66(2). (Cited on page 74.)

Malik, S., Kazai, G., Lalmas, M., and Fuhr, N. (2005). Overview of INEX 2005. In *INEX*, pages 1–15. Springer. (Cited on page 15.)

Malik, S., Lalmas, M., and Fuhr, N. (2004). Overview of INEX 2004. In *INEX*, pages 1–15. Springer. (Cited on page 15.)

Malik, S., Trotman, A., Lalmas, M., and Fuhr, N. (2006). Overview of INEX 2006. In *INEX*, pages 1–11. Springer. (Cited on page 15.)

Meehan, J. (2010). Rethinking original order and personal records. *Archivaria*, 70(2). (Cited on page 74.)

Meier, W. (2003). eXist: An open source native XML database. In *Web, Web-Services, and Database Systems*, volume 2593 of *LNCS*, pages 169–183. Springer. (Cited on page

18.)
Menne-Haritz, A. (2001). Access: The reformulation of an archival paradigm. *Archival Science*, 1:57–82. (Cited on page 58.)
Navarro-Prieto, R., Scaife, M., and Rogers, Y. (1999). Cognitive strategies in Web searching. *Proceedings of the 5th Conference on Human Factors the Web*, pages 1–13. (Cited on page 93.)
Nicholas, D., Huntington, P., Jamali, H. R., and Watkinson, A. (2006). The information seeking behaviour of the users of digital scholarly journals. *Inf. Process. Manage.*, 42:1345–1365. (Cited on page 83.)
O'Reilly, T. (2007). What is Web 2.0: Design patterns and business models for the next generation of software. MPRA Paper 4578, University Library of Munich, Germany. (Cited on page 7.)
Ottosson, P.-G. (2005). EAC and the development of national and European gateways to archives. *Journal of Archival Organization*, 3(2/3):261–274. (Cited on page 79.)
Pearce-Moses, R. (2005). A glossary of archival and records terminology. Retrieved 2010/05/28 from http://www.archivists.org/glossary/. (Cited on pages 1, 2, 25, and 57.)
Peters, T. (1993). The history and development of transaction log analysis. *Library Hi Tech*, 42(11):41–66. (Cited on pages 35, 112, 113, and 114.)
Pharo, N. (2008). The effect of granularity and order in XML element retrieval. *Inf. Process. Manage.*, 44:1732–1740. (Cited on pages 3 and 14.)
Pitti, D. V. (1997). Encoded Archival Description: The development of an encoding standard for archival finding aids. *American Archivist*, 60(3):268–283. (Cited on pages 2, 10, and 77.)
Pitti, D. V. (1999). Encoded Archival Description: An introduction and overview. *D-Lib Magazine*, 5(11). (Cited on pages 11, 22, 36, 63, 65, 67, 69, and 74.)
Ponte, J. M. and Croft, W. B. (1998). A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, pages 275–281, New York, NY, USA. ACM. (Cited on pages 23, 44, and 75.)
Prom, C. (2002). The EAD Cookbook: A Survey and Usability Study. *American Archivist*, 65(2):257–275. (Cited on page 19.)
Prom, C. J. (2004). User interactions with electronic finding aids in a controlled setting. *American Archivist*, 67(2):234–268. (Cited on page 82.)
Prom, C. J. (2011). Using web analytics to improve online access to archival resources. *American Archivist*, 74(1):158–184. (Cited on pages 4, 86, and 114.)
Prom, C. J., Rishel, C. A., Schwartz, S. W., and Fox, K. J. (2007). A unified platform for archival description and access. In *JCDL '07*, pages 157–166, New York, NY, USA. ACM. (Cited on page 18.)
Ribeiro, F. (1996). Subject Indexing and Authority Control in Archives: The Need for Subject Indexing in Archives and for an Indexing Policy Using Controlled Language. *Journal of the Society of Archivists*, 17(1):27–54. (Cited on page 36.)
Rich, E. (1979). User modeling via stereotypes. *Cognitive Science*, 3(4):329–354. (Cited on page 114.)
Robertson, S. E. (2008). On the history of evaluation in IR. *J. Inf. Sci.*, 34(4):439–456. (Cited on pages 33 and 36.)

Robertson, S. E. and Hancock-Beaulieu, M. M. (1992). On the evaluation of IR systems. *Inf. Process. Manage.*, 28(4):457–466. (Cited on pages 33, 34, 111, 112, and 113.)

Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M., and Gatford, M. (1994). Okapi at TREC-3. In *TREC*, pages 109–126. (Cited on page 44.)

Rose, D. E. (2006). Reconciling information-seeking behavior with search user interfaces for the Web. *Journal of the American Society for Information Science and Technology*, 57(6):797–799. (Cited on page 116.)

Ruth, J. E. (1997). Encoded Archival Description: A structural overview. *American Archivist*, 60(3):310–329. (Cited on pages 11, 16, and 65.)

Samouelian, M. (2009). Embracing Web 2.0: Archives and the Newest Generation of Web Applications. *American Archivist*, 72(1):42–71. (Cited on page 7.)

Saracevic, T. (1975). Relevance: a review of and a framework for the thinking on the notion in information science. *Journal of the American Society for Information Science*, 26(6):321–343. (Cited on pages 114 and 131.)

Saving, T. R. (1970). Concentration ratios and the degree of monopoly. *International Economic Review*, 11(1):139–46. (Cited on page 69.)

Savoy, J. (2003). Report on CLEF 2002 experiments: Combining multiple sources of evidence. In *Advances in Cross-Language Information Retrieval*, volume 2785 of *Lecture Notes in Computer Science*, pages 66–90. Springer, Berlin, Heidelberg. (Cited on page 49.)

Schellenberg, T. R. (1965). *The management of archives*. Columbia University Press, New York,. (Cited on page 68.)

Schwartz, S. W., Prom, C. J., Rishel, C. A., and Fox, K. J. (2007). Archon: A unified information storage and retrieval system for lone archivists, special collections librarians and curators. *Canadian Journal of Library and Information Practice and Research*, 2(2). (Cited on page 18.)

Scott, P. J. (1966). The record group concept: A case for abandonment. *American Archivist*, 29(4):493–504. (Cited on page 78.)

Sévigny, M. and Clavaud, F. (2003). Pleade: EAD for the Web. *DigitCULT.Info*, 6:16–18. (Cited on page 18.)

Shaw, W. M., Wood, J. B., Wood, R. E., and Tibbo, H. R. (1991). The cystic fibrosis database: content and research opportunities. *Library and Information Science Research*, 13:347–366. (Cited on page 36.)

Sigurbjörnsson, B. (2006). *Focused Information Access Using XML Element Retrieval*. Doctoral dissertation, University of Amsterdam. (Cited on page 3.)

Skov, M. and Ingwersen, P. (2008). Exploring information seeking behaviour in a digital museum context. In *Proceedings of the second international symposium on Information interaction in context*, IIiX '08, pages 110–115, New York, NY, USA. ACM. (Cited on page 85.)

Spärck Jones, K. (1981). The cranfield tests. In Spärck Jones, K., editor, *Information Retrieval Experiment*, pages 256–284. Butterworth. (Cited on pages 81, 111, and 113.)

Spärck Jones, K. and van Rijsbergen, C. J. (1976). Information retrieval test collections. *J. of Documentation*, 32(1):59–75. (Cited on pages 36 and 130.)

Stibbe, H. L. P. (1992). Implementing the concept of fonds: Primary access point, multilevel description and authority control. *Archivaria*, 34(1):109–137. (Cited on page 60.)

Szary, R. V. (2005). Encoded Archival Context (EAC) and archival description : Rationale and background. *Journal of Archival Organization*, 3(2/3):217–227. (Cited on page 79.)

Tague-Sutcliffe, J. (1992). The pragmatics of information retrieval experimentation, revisited. *Inf. Process. Manage.*, 28:467–490. (Cited on page 34.)

Talja, S. and Maula, H. (2003). Reasons for the use and non-use of electronic journals and databases - A domain analytic study in four scholarly disciplines. *Journal of Documentation*, 59(6):673–691. (Cited on page 115.)

Thomassen, T. (2001). A first introduction to archival science. *Archival Science*, 1(4):373–385. (Cited on pages 30 and 58.)

Tibbo, H. R. (2003). Primarily History in America: How U. S. historians search for primary materials at the dawn of the digital age: Users and archival research. *American Archivist*, 66(1):9–50. (Cited on page 86.)

Tibbo, H. R. and Meho, L. I. (2001). Finding finding aids on the World Wide Web. *American Archivist*, 64(1):61–77. (Cited on page 37.)

Tolle, J. E. (1983). Transaction log analysis online catalogs. In *Proceedings of the 6th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '83, pages 147–160, New York, NY, USA. ACM. (Cited on page 113.)

Trotman, A., Geva, S., and Kamps, J. (2007). Report on the SIGIR 2007 workshop on focused retrieval. *SIGIR Forum*, 41(2):97–103. (Cited on pages 12 and 15.)

Trotman, A. and Sigurbjörnsson, B. (2004). Narrowed Extended XPath I (NEXI). In *INEX*, pages 16–40. Springer. (Cited on page 24.)

Tsikrika, T. (2009). Aggregation-based structured text retrieval. In *Encyclopedia of Database Systems*, pages 63–71. Springer, Berlin, Heidelberg. (Cited on page 16.)

Vakkari, P., Pennanen, M., and Serola, S. (2003). Changes of search terms and tactics while writing a research proposal: A longitudinal case study. *Information Processing & Management*, 39(3):445 – 463. (Cited on page 85.)

Vitali, S. (2005). What are the boundaries of archival context? : The SIASFI Project and the Online Guide to the Florence State Archives, Italy. *Journal of Archival Organization*, 3(2/3):243–260. (Cited on page 79.)

Voorhees, E. M. and Harman, D. K. (2005). *TREC: Experiment and Evaluation in Information Retrieval*. Digital Libraries and Electronic Publishing. MIT Press. (Cited on page 36.)

Westbrook, B. D., Mandell, L., Shepherd, K., Stevens, B., and Varghese, J. (2007). The Archivists' Toolkit: Another Step Toward Streamlined Archival Processing. *Journal of Archival Organization*, 4(1):229–253. (Cited on page 19.)

White, R. W. and Drucker, S. M. (2007). Investigating behavioral variability in Web search. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 21–30, New York, NY, USA. ACM. (Cited on page 84.)

White, R. W., Dumais, S., and Teevan, J. (2008). How medical expertise influences Web search interaction. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 791–792, New York, NY, USA. ACM. (Cited on page 15.)

White, R. W. and Dumais, S. T. (2009). Characterizing and predicting search engine switching behavior. In *Proceeding of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 87–96, New York, NY, USA. ACM. (Cited on page

84.)
White, R. W. and Morris, D. (2007). Investigating the querying and browsing behavior of advanced search engine users. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, pages 255–262, New York, NY, USA. ACM. (Cited on page 35.)
Wilbur, W. J. (1998). A comparison of group and individual performance among subject experts and untrained workers at the document retrieval task. *J. Am. Soc. Inf. Sci.*, 49:517–529. (Cited on page 115.)
Wilkinson, R. (1994). Effective retrieval of structured documents. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '94, pages 311–317, New York, NY, USA. Springer-Verlag New York, Inc. (Cited on page 12.)
Wilson, M. L., schraefel, M. C., and White, R. W. (2009). Evaluating advanced search interfaces using established information-seeking models. *J. Am. Soc. Inf. Sci. Technol.*, 60:1407–1422. (Cited on page 116.)
Wilson, T. D. (1999). Models in information behaviour research. *Journal of Documentation*, 55(3):249–270. (Cited on page 81.)
Wirth, N. (1977). What can we do about the unnecessary diversity of notation for syntactic definitions? *Commun. ACM*, 20:822–823. (Cited on page 10.)
Wolfram, D. (2008). Search characteristics in different types of Web-based IR environments: Are they the same? *Information Processing & Management*, 44(3):1279 – 1292. (Cited on page 84.)
Yaco, S. (2008). It's Complicated: Barriers to EAD Implementation. *American Archivist*, 71(2):456–475. (Cited on page 18.)
Yakel, E. (2004). Encoded Archival Description: Are finding aids boundary spanners or barriers for users? *Journal of Archival Organization*, 2(1):63–77. (Cited on pages 15, 33, and 112.)
Yakel, E. and Kim, J. (2005). Adoption and diffusion of Encoded Archival Description. *J. Am. Soc. Inf. Sci. Technol.*, 56:1427–1437. (Cited on page 18.)
Yakel, E., Shaw, S., and Reynolds, P. (2007). Creating the next generation of archival finding aids. *D-Lib Magazine*, 13(5/6). (Cited on page 7.)
Yakel, E. and Torres, D. A. (2003). AI: Archival intelligence and user expertise. *American Archivist*, 66(1):51–78. (Cited on pages 60, 112, 114, 122, 126, and 127.)
Yuan, X. and Belkin, N. J. (2010a). Evaluating an integrated system supporting multiple information-seeking strategies. *J. Am. Soc. Inf. Sci. Technol.*, 61:1987–2010. (Cited on page 116.)
Yuan, X. and Belkin, N. J. (2010b). Investigating information retrieval support techniques for different information-seeking strategies. *J. Am. Soc. Inf. Sci. Technol.*, 61:1543–1563. (Cited on page 116.)
Zach, L. (2005). When is "enough" enough? Modeling the information-seeking and stopping behavior of senior arts administrators: Research articles. *J. Am. Soc. Inf. Sci. Technol.*, 56:23–35. (Cited on page 85.)
Zhou, X. (2007). Examining search functions of EAD finding aids Web sites. *Journal of Archival Organization*, 4(3):99–118. (Cited on page 18.)
Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*. Addison-Wesley (Reading MA). (Cited on pages 12 and 83.)

Abstract

How do archives provide access to their records and let users search? The answer is archival description. Encoded Archival Description (EAD) in Extensible Markup Language (XML) is the *de facto* technical standard for 'electronic' archival descriptions. It is now used to bridge the gulf between tangible records in archives and digital objects on the World Wide Web. These descriptions are finding aids, which are tools to search and find information about, or references to, archival records. The archival finding aids in EAD are left to searchers (out of sight and contact) to explore in unknown ways: how do searchers interact with these finding aids, and what type of retrieval system is needed to support them? The approach of this dissertation is to apply XML retrieval techniques to the EAD finding aids, develop system evaluation of EAD retrieval, and study information seeking behavior of archival search.

The first study involves the design and implementation of the archival search engine README. The README system attempts to incorporate current technologies with the archival structure in finding aids—such as XML retrieval—and simultaneously to uphold the archival principles where this structure is based upon. The system is the proof of concept.

Having established this baseline, the next study explores and tests the construction of an information retrieval (IR) test collection. A test collection is a key component in IR evaluation. The basis of this test collection are the queries and clicks on archival descriptions that can be found in the search log files of the *Nationaal Archief* . There is no readily-available test collection for evaluating the accuracy of the retrieval of archival descriptions of records by an archival search engine. Manually creating such a collection is expensive. The study shows that automatically creating a test collection seems a viable alternative.

Archival principles—such as provenance and original order—are deeply rooted in the arrangement and subsequent description of archival records. The investigation continues by shedding new light on them in a system evaluation. Additionally, the experiments probe XML retrieval-specific issues, such as the retrieval of certain elements. The study concludes by reflecting on the README archival search engine, which is the baseline of the probes in this

dissertation. How effective are certain archival principles for archival access in this digital age?

Using the archival search log files, the research focus shifts to the arrangement of records in EAD and user search behaviors using this arrangement. The sub-document clicks within the finding aids point to the online interaction of users within 'electronic' archival descriptions of records. The analysis of the interactions comprises of quantifying the search behavior. This results in a state diagram that captures different information search behaviors of different people. By analyzing real-world interaction, the discussion on the use of the finding aid in this digital age as access tool becomes more complete. The result is more understanding of online archival search behavior within EAD finding aids, which can be used to improve a search system adapted to existing 'electronic' archival descriptions.

Finally, the system evaluation deals with tailoring a search engine to the different user stereotypes, namely 'expert' and 'novice' groups based on the number of times that a user re-uses the system. The results show that although there are significant differences in terms of search behavior, this does not necessarily mean that for more effective retrieval of archival descriptions, the system needs to be adapted to improve access for these different user groups.

Samenvatting

Hoe zorgen archieven voor de ontsluiting van hun archiefstukken en hoe laten ze gebruikers zoeken? Het antwoord is met archiefbeschrijvingen. Encoded Archival Description (EAD) in Extensible Markup Language (XML) is *de facto* de technische standaard voor digitale archiefbeschrijvingen. Het wordt nu gebruikt om de kloof tussen de tastbare stukken in de archieven en digitale objecten op het internet te overbruggen. Deze beschrijvingen zijn toegangen om informatie over, of verwijzingen naar, archiefstukken te vinden. De archiefbeschrijvingen in EAD worden aan gebruikers overgelaten om te verkennen op onbekende manieren: hoe werken gebruikers met deze archiefbeschrijvingen en wat voor zoeksysteem is nodig om ze te ondersteunen? Dit proefschrift maakt gebruik van *XML retrieval* technieken voor EAD toegangen, ontwikkelt een systeem evaluatie voor *EAD retrieval*, en bestudeert het informatiezoekgedrag van archiefgebruikers.

De eerste studie beschrijft het ontwerp en implementatie van de zoekmachine README. Het README systeem probeert de huidige technologieën te gebruiken in combinatie met de structuur in toegangen, zoals informatieontsluiting met XML, en tegelijkertijd om archivistische beginselen te handhaven waarop deze structuur is gebaseerd. Het systeem is een *proof of concept*.

De volgende studie verkent en test de constructie van een test collectie, een belangrijk component, voor het evalueren van de informatieontsluiting. De grondslag zijn de zoekvragen en klikken op archiefbeschrijvingen die gevonden kunnen worden in de logbestanden van het Nationaal Archief. Er bestaat nog geen beschikbare test collectie voor het evalueren van de ontsluiting van archiefbeschrijvingen door een zoekmachine. Het handmatig creëren van zo'n collectie is kostbaar. Het automatisch creëren van een test collectie kan een alternatief zijn.

Archivistische beginselen, zoals het herkomstbeginsel en de oude ordening, zijn diepgeworteld in de rangschikking en de daarop volgende beschrijving van de archiefstukken. Het onderzoek gaat verder door deze beginselen in een nieuw daglicht te plaatsen d.m.v. een systeem evaluatie. We onderzoeken specifieke zaken m.b.t. XML retrieval, zoals het ontsluiten van bepaalde elementen. De studie sluit af door te reflecteren op de README zoekmachine. Hoe effectief zijn bepaalde archivistische beginselen voor het ontsluiten van archieven in dit digitaal tijdperk?

Gebruikmakend van de logbestanden, verschuift het onderzoek zich naar de rangschikking van archiefstukken in EAD en het zoekgedrag van gebruikers met deze rangschikking. De klikken binnen de toegangen verwijzen naar de interactie van gebruikers met de digitale archiefbeschrijvingen. De analyse van de interactie bestaat uit het kwantificeren van het zoekgedrag. Dit resulteert in een toestandsdiagram dat het verschillend zoekgedrag van verschillende mensen vastlegt. Door het analyseren van bestaande interactie wordt de discussie over het gebruik van EAD toegangen in dit digitaal tijdperk als toegangsmiddel completer. Het resultaat is meer inzicht in het archivistisch zoekgedrag op het internet binnen toegangen in EAD, wat gebruikt kan worden om een zoekmachine gedreven door archiefbeschrijvingen te verbeteren.

Tenslotte, introduceren we gebruikersstereotypen. Dit zijn 'expert' en 'beginners' groepen gebaseerd op het aantal keer dat het systeem opnieuw wordt gebruikt. De resultaten laten zien dat, hoewel er significante verschillen zijn in het zoekgedrag, het niet per se hoeft te betekenen dat voor effectievere ontsluiting, het systeem voor hen hoeft te worden aangepast.

Titles in the ILLC Dissertation Series

ILLC DS-2006-01: **Troy Lee**
Kolmogorov complexity and formula size lower bounds

ILLC DS-2006-02: **Nick Bezhanishvili**
Lattices of intermediate and cylindric modal logics

ILLC DS-2006-03: **Clemens Kupke**
Finitary coalgebraic logics

ILLC DS-2006-04: **Robert Špalek**
Quantum Algorithms, Lower Bounds, and Time-Space Tradeoffs

ILLC DS-2006-05: **Aline Honingh**
The Origin and Well-Formedness of Tonal Pitch Structures

ILLC DS-2006-06: **Merlijn Sevenster**
Branches of imperfect information: logic, games, and computation

ILLC DS-2006-07: **Marie Nilsenova**
Rises and Falls. Studies in the Semantics and Pragmatics of Intonation

ILLC DS-2006-08: **Darko Sarenac**
Products of Topological Modal Logics

ILLC DS-2007-01: **Rudi Cilibrasi**
Statistical Inference Through Data Compression

ILLC DS-2007-02: **Neta Spiro**
What contributes to the perception of musical phrases in western classical music?

ILLC DS-2007-03: **Darrin Hindsill**
It's a Process and an Event: Perspectives in Event Semantics

ILLC DS-2007-04: **Katrin Schulz**
Minimal Models in Semantics and Pragmatics: Free Choice, Exhaustivity, and Conditionals

ILLC DS-2007-05: **Yoav Seginer**
Learning Syntactic Structure

ILLC DS-2008-01: **Stephanie Wehner**
Cryptography in a Quantum World

ILLC DS-2008-02: **Fenrong Liu**
Changing for the Better: Preference Dynamics and Agent Diversity

ILLC DS-2008-03: **Olivier Roy**
Thinking before Acting: Intentions, Logic, Rational Choice

ILLC DS-2008-04: **Patrick Girard**
Modal Logic for Belief and Preference Change

ILLC DS-2008-05: **Erik Rietveld**
Unreflective Action: A Philosophical Contribution to Integrative Neuroscience

ILLC DS-2008-06: **Falk Unger**
Noise in Quantum and Classical Computation and Non-locality

ILLC DS-2008-07: **Steven de Rooij**
Minimum Description Length Model Selection: Problems and Extensions

ILLC DS-2008-08: **Fabrice Nauze**
Modality in Typological Perspective

ILLC DS-2008-09: **Floris Roelofsen**
Anaphora Resolved

ILLC DS-2008-10: **Marian Counihan**
Looking for logic in all the wrong places: an investigation of language, literacy and logic in reasoning

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-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

Titles in the SIKS Dissertation Series

2009-01: **Rasa Jurgelenaite (RUN)**
Symmetric Causal Independence Models

2009-02: **Willem Robert van Hage (VU)**
Evaluating Ontology-Alignment Techniques

2009-03: **Hans Stol (UvT)**
A Framework for Evidence-based Policy Making Using IT

2009-04: **Josephine Nabukenya (RUN)**
Improving the Quality of Organisational Policy Making using Collaboration Engineering

2009-05: **Sietse Overbeek (RUN)**
Bridging Supply and Demand for Knowledge Intensive Tasks - Based on Knowledge, Cognition, and Quality

2009-06: **Muhammad Subianto (UU)**
Understanding Classification

2009-07: **Ronald Poppe (UT)**
Discriminative Vision-Based Recovery and Recognition of Human Motion

2009-08: **Volker Nannen (VU)**
Evolutionary Agent-Based Policy Analysis in Dynamic Environments

2009-09: **Benjamin Kanagwa (RUN)**
Design, Discovery and Construction of Service-oriented Systems

2009-10: **Jan Wielemaker (UVA)**
Logic programming for knowledge-intensive interactive applications

2009-11: **Alexander Boer (UVA)**
Legal Theory, Sources of Law & the Semantic Web

2009-12: **Peter Massuthe (TUE, Humboldt-Universitaet zu Berlin)**
Operating Guidelines for Services

2009-13: **Steven de Jong (UM)**
Fairness in Multi-Agent Systems

2009-14: **Maksym Korotkiy (VU)**
From ontology-enabled services to service-enabled ontologies (making ontologies work in e-science with ONTO-SOA)

2009-15: **Rinke Hoekstra (UVA)**
Ontology Representation - Design Patterns and Ontologies that Make Sense

2009-16: **Fritz Reul (UvT)**
New Architectures in Computer Chess

2009-17: **Laurens van der Maaten (UvT)**
Feature Extraction from Visual Data

2009-18: **Fabian Groffen (CWI)**
Armada, An Evolving Database System

2009-19: **Valentin Robu (CWI)**
Modeling Preferences, Strategic Reasoning and Collaboration in Agent-Mediated Electronic Markets

2009-20: **Bob van der Vecht (UU)**
Adjustable Autonomy: Controling Influences on Decision Making

2009-21: **Stijn Vanderlooy (UM)**
Ranking and Reliable Classification

2009-22: **Pavel Serdyukov (UT)**
Search For Expertise: Going beyond direct evidence

2009-23: **Peter Hofgesang (VU)**
Modelling Web Usage in a Changing Environment

2009-24: **Annerieke Heuvelink (VU)**
Cognitive Models for Training Simulations

2009-25: **Alex van Ballegooij (CWI)**
RAM: Array Database Management through Relational Mapping

2009-26: **Fernando Koch (UU)**
An Agent-Based Model for the Development of Intelligent Mobile Services

2009-27: **Christian Glahn (OU)**
Contextual Support of Social Engagement and Reflection on the Web

2009-28: **Sander Evers (UT)**
Sensor Data Management with Probabilistic Models

2009-29: **Stanislav Pokraev (UT)**
Model-Driven Semantic Integration of Service-Oriented Applications

2009-30: **Marcin Zukowski (CWI)**
Balancing vectorized query execution with bandwidth-optimized storage

2009-31: **Sofiya Katrenko (UVA)**
A Closer Look at Learning Relations from Text

2009-32: **Rik Farenhorst (VU) and Remco de Boer (VU)**
Architectural Knowledge Management: Supporting Architects and Auditors

2009-33: **Khiet Truong (UT)**
How Does Real Affect Affect Affect Recognition In Speech?

2009-34: **Inge van de Weerd (UU)**
Advancing in Software Product Management: An Incremental Method Engineering Approach

2009-35: **Wouter Koelewijn (UL)**
Privacy en Politiegegevens; Over geautomatiseerde normatieve informatie-uitwisseling

2009-36: **Marco Kalz (OU)**
Placement Support for Learners in Learning Networks

2009-37: **Hendrik Drachsler (OU)**
Navigation Support for Learners in Informal Learning Networks

2009-38: **Riina Vuorikari (OU)**
Tags and self-organisation: a metadata ecology for learning resources in a multilingual context

2009-39: **Christian Stahl (TUE, Humboldt-Universitaet zu Berlin)**
Service Substitution – A Behavioral Approach Based on Petri Nets

2009-40: **Stephan Raaijmakers (UvT)**
Multinomial Language Learning: Investigations into the Geometry of Language

2009-41: **Igor Berezhnyy (UvT)**
Digital Analysis of Paintings

2009-42: **Toine Bogers (UvT)**
Recommender Systems for Social Bookmarking

2009-43: **Virginia Nunes Leal Franqueira (UT)**
Finding Multi-step Attacks in Computer Networks Using Heuristic Search and Mobile Ambients

2009-44: **Roberto Santana Tapia (UT)**
Assessing Business-IT Alignment in Networked Organizations

2009-45: **Jilles Vreeken (UU)**
Making Pattern Mining Useful

2009-46: **Loredana Afanasiev (UVA)**
Querying XML: Benchmarks and Recursion

2010-01: **Matthijs van Leeuwen (UU)**
Patterns that Matter

2010-02: **Ingo Wassink (UT)**
Work flows in Life Science

2010-03: **Joost Geurts (CWI)**
A Document Engineering Model and Processing Framework for Multimedia documents

2010-04: **Olga Kulyk (UT)**
Do You Know What I Know? Situational Awareness of Co-located Teams in Multidisplay Environments

2010-05: **Claudia Hauff (UT)**
Predicting the Effectiveness of Queries and Retrieval Systems

2010-06: **Sander Bakkes (UvT)**
Rapid Adaptation of Video Game AI

2010-07: **Wim Fikkert (UT)**
Gesture Interaction at a Distance

2010-08: **Krzysztof Siewicz (UL)**
Towards an Improved Regulatory Framework of Free Software. Protecting user freedoms in a world of software communities and eGovernments

2010-09: **Hugo Kielman (UL)**
Politiële gegevensverwerking en privacy. Naar een effectieve waarborging

2010-10: **Rebecca Ong (UL)**
Mobile Communication and Protection of Children

2010-11: **Adriaan Ter Mors (TUD)**
The world according to MARP: Multi-Agent Route Planning

2010-12: **Susan van den Braak (UU)**
Sensemaking software for crime analysis

2010-13: **Gianluigi Folino (RUN)**
High Performance Data Mining using Bio-inspired techniques

2010-14: **Sander van Splunter (VU)**
Automated Web Service Reconfiguration

2010-15: **Lianne Bodenstaff (UT)**
Managing Dependency Relations in Inter-Organizational Models

2010-16: **Sicco Verwer (TUD)**
Efficient Identification of Timed Automata, theory and practice

2010-17: **Spyros Kotoulas (VU)**
Scalable Discovery of Networked Resources: Algorithms, Infrastructure, Applications

2010-18: **Charlotte Gerritsen (VU)**
Caught in the Act: Investigating Crime by Agent-Based Simulation

2010-19: **Henriette Cramer (UVA)**
People's Responses to Autonomous and Adaptive Systems

2010-20: **Ivo Swartjes (UT)**
Whose Story Is It Anyway? How Improv Informs Agency and Authorship of Emergent Narrative

2010-21: **Harold van Heerde (UT)**
Privacy-aware data management by means of data degradation

2010-22: **Michiel Hildebrand (CWI)**
End-user Support for Access to Heterogeneous Linked Data

2010-23: **Bas Steunebrink (UU)**
The Logical Structure of Emotions

2010-24: **Dmytro Tykhonov (TUD)**
Designing Generic and Efficient Negotiation Strategies

2010-25: **Zulfiqar Ali Memon (VU)**
Modelling Human-Awareness for Ambient Agents: A Human Mindreading Perspective

2010-26: **Ying Zhang (CWI)**
XRPC: Efficient Distributed Query Processing on Heterogeneous XQuery Engines

2010-27: **Marten Voulon (UL)**
Automatisch contracteren

2010-28: **Arne Koopman (UU)**
Characteristic Relational Patterns

2010-29: **Stratos Idreos (CWI)**
Database Cracking: Towards Auto-tuning Database Kernels

2010-30: **Marieke van Erp (UvT)**
Accessing Natural History - Discoveries in data cleaning, structuring, and retrieval

2010-31: **Victor de Boer (UVA)**
Ontology Enrichment from Heterogeneous Sources on the Web

2010-32: **Marcel Hiel (UvT)**
An Adaptive Service Oriented Architecture: Automatically solving Interoperability Problems

2010-33: **Robin Aly (UT)**
Modeling Representation Uncertainty in Concept-Based Multimedia Retrieval

2010-34: **Teduh Dirgahayu (UT)**
Interaction Design in Service Compositions

2010-35: **Dolf Trieschnigg (UT)**
Proof of Concept: Concept-based Biomedical Information Retrieval

2010-36: **Jose Janssen (OU)**
Paving the Way for Lifelong Learning; Facilitating competence development through a learning path specification

2010-37: **Niels Lohmann (TUE)**
Correctness of services and their composition

2010-38: **Dirk Fahland (TUE)**
From Scenarios to components

2010-39: **Ghazanfar Farooq Siddiqui (VU)**
Integrative modeling of emotions in virtual agents

2010-40: **Mark van Assem (VU)**
Converting and Integrating Vocabularies for the Semantic Web

2010-41: **Guillaume Chaslot (UM)**
Monte-Carlo Tree Search

2010-42: **Sybren de Kinderen (VU)**
Needs-driven service bundling in a multi-supplier setting - the computational e3-service approach

2010-43: **Peter van Kranenburg (UU)**
A Computational Approach to Content-Based Retrieval of Folk Song Melodies

2010-44: **Pieter Bellekens (TUE)**
An Approach towards Context-sensitive and User-adapted Access to Heterogeneous Data Sources, Illustrated in the Television Domain

2010-45: **Vasilios Andrikopoulos (UvT)**
A theory and model for the evolution of software services

2010-46: **Vincent Pijpers (VU)**
e3alignment: Exploring Inter-Organizational Business-ICT Alignment

2010-47: **Chen Li (UT)**
Mining Process Model Variants: Challenges, Techniques, Examples

2010-48: **Milan Lovric (EUR)**
Behavioral Finance and Agent-Based Artificial Markets

2010-49: **Jahn-Takeshi Saito (UM)**
Solving difficult game positions

2010-50: **Bouke Huurnink (UVA)**
Search in Audiovisual Broadcast Archives

2010-51: **Alia Khairia Amin (CWI)**
Understanding and supporting information seeking tasks across multiple sources

2010-52: **Peter-Paul van Maanen (VU)**
Adaptive Support for Human-Computer Teams: Exploring the Use of Cognitive Models of Trust and Attention

2010-53: **Edgar Meij (UVA)**
Combining Concepts and Language Models for Information Access

2011-01: **Botond Cseke (RUN)**
Variational Algorithms for Bayesian Inference in Latent Gaussian Models

2011-02: **Nick Tinnemeier(UU)**
Work flows in Life Science

2011-03: **Jan Martijn van der Werf (TUE)**
Compositional Design and Verification of Component-Based Information Systems

2011-04: **Hado van Hasselt (UU)**
Insights in Reinforcement Learning; Formal analysis and empirical evaluation of temporal-difference learning algorithms

2011-05: **Base van der Raadt (VU)**
Enterprise Architecture Coming of Age - Increasing the Performance of an Emerging Discipline

2011-06: **Yiwen Wang (TUE)**
Semantically-Enhanced Recommendations in Cultural Heritage

2011-07: **Yujia Cao (UT)**
Multimodal Information Presentation for High Load Human Computer Interaction

2011-08: **Nieske Vergunst (UU)**
BDI-based Generation of Robust Task-Oriented Dialogues

2011-09: **Tim de Jong (OU)**
Contextualised Mobile Media for Learning

2011-10: **Bart Bogaert (UvT)**
Cloud Content Contention

2011-11: **Dhaval Vyas (UT)**
Designing for Awareness: An Experience-focused HCI Perspective

2011-12: **Carmen Bratosin (TUE)**
Grid Architecture for Distributed Process Mining

2011-13: **Xiaoyu Mao (UvT)**
Airport under Control. Multiagent Scheduling for Airport Ground Handling

2011-14: **Milan Lovric (EUR)**
Behavioral Finance and Agent-Based Artificial Markets

2011-15: **Marijn Koolen (UVA)**
The Meaning of Structure: the Value of Link Evidence for Information Retrieval

2011-16: **Maarten Schadd (UM)**
Selective Search in Games of Different Complexity

2011-17: **Jiyin He (UVA)**
Exploring Topic Structure: Coherence, Diversity and Relatedness

2011-18: **Mark Ponsen (UM)**
Strategic Decision-Making in complex games

2011-19: **Ellen Rusman (OU)**
The Mind ' s Eye on Personal Profiles

2011-20: **Qing Gu (VU)**
Guiding service-oriented software engineering - A view-based approach

2011-21: **Linda Terlouw (TUD)**
Modularization and Specification of Service-Oriented Systems

2011-22: **Junte Zhang (UVA)**
System Evaluation of Archival Description and Access