UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

System evaluation of archival description and access

Zhang, J.

Publication date
2011

Link to publication

Citation for published version (APA):
Zhang, J. (2011). *System evaluation of archival description and access*. [Thesis, fully internal, Universiteit van Amsterdam]. IR Publications.

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:18 Nov 2026

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.