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:17 Nov 2026

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.