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 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.

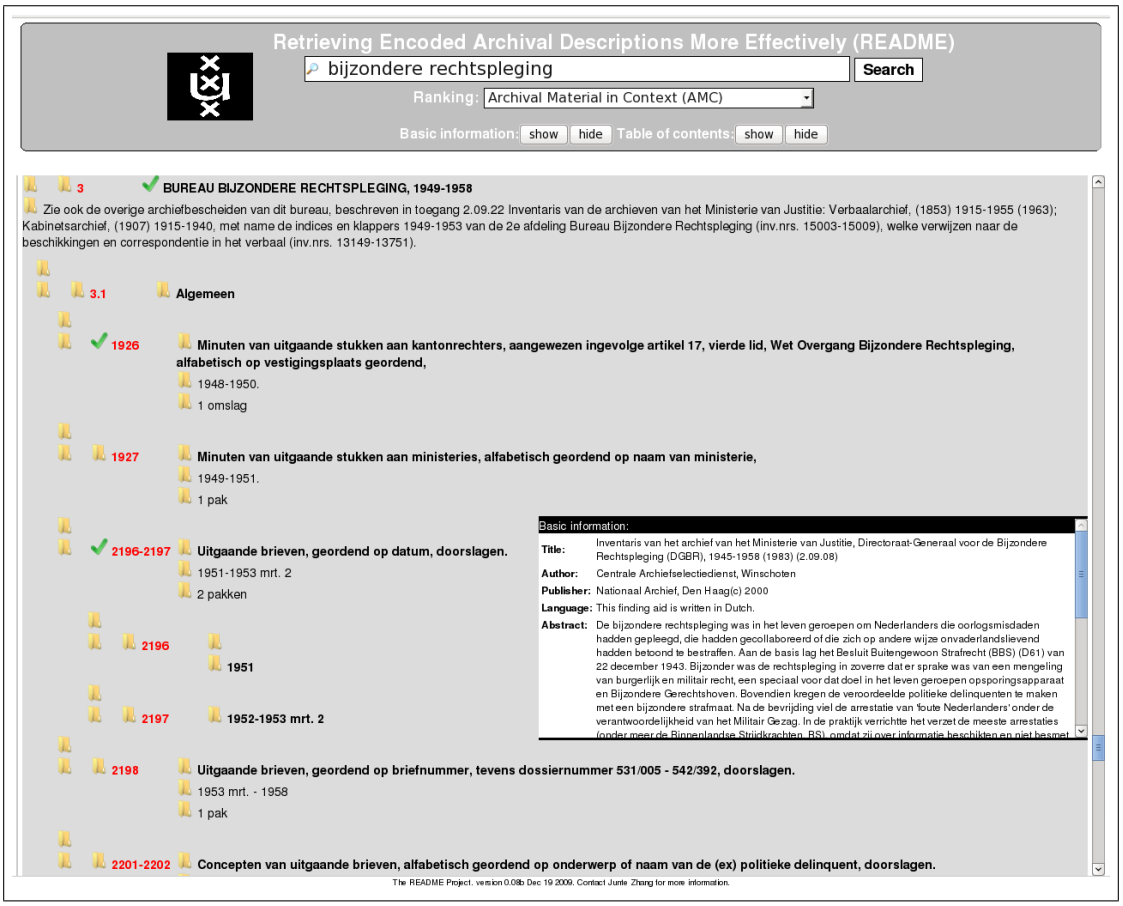


FIGURE 4.1: The README Assessment System for selecting useful or relevant elements within a fonds, which can be used to construct an EAD test collection.

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.

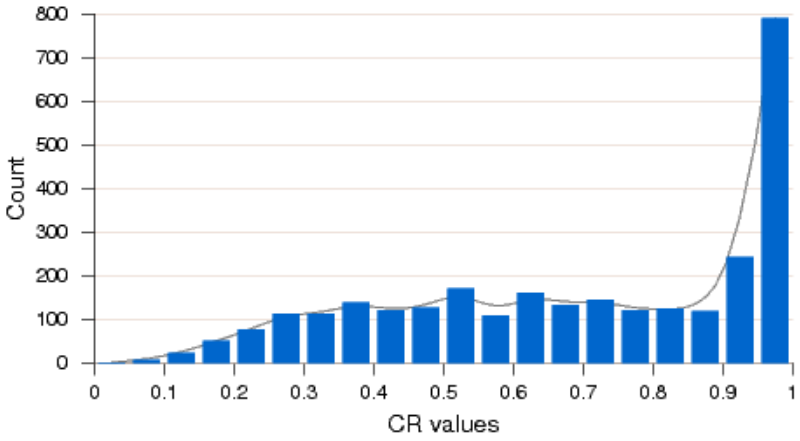
Next, we count the number of different EAD finding aids associated with a topic. We focus only on January 2009 in Table 4.9. The topic *suriname* (in English, *Surinam*), has an association with 39 different EADs, so users have used this range of EAD finding aids for that search request. Other popular 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.

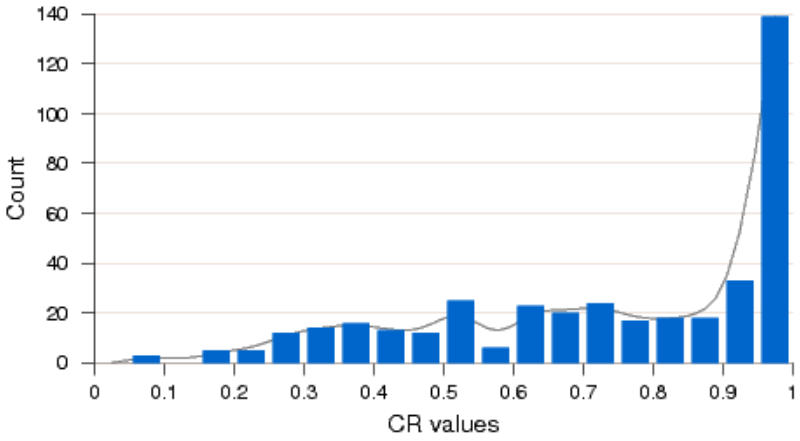
# EADs	Topics (incl. English Translation)	Topic Count
39	*"suriname"* (Surinam)	1
29	*"ptt"* (—)	1
26	*"financien"* (finances)	1
21	*"buitenlandse zaken"* (foreign affairs)	1
19	*"noordwijkerhout"* (—), *"4.miko"* (—)	2
17	*"japanse"* (Japanese), *"genie"* (military engineer), *"wetenschap"* (science), *"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?