UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

Effective access to digital assets: An XML-based EAD search system

Zhang, J.; Fachry, K.N.; Kamps, J.

Publication date
2009
Document Version
Final published version
Published in
Proceedings of DigCCurr2009: Digital curation: Practice, promise and prospects

Link to publication

Citation for published version (APA):
Zhang, J., Fachry, K. N., & Kamps, J. (2009). Effective access to digital assets: An XML-based EAD search system. In H. R. Tibbo, C. Hank, C. A. Lee, & R. Clemens (Eds.), *Proceedings of DigCCurr2009: Digital curation: Practice, promise and prospects* (pp. 49-56). University of North Carolina at Chapel Hill, School of Information and Library Science. http://stores.lulu.com/DigCCurr2009

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, Singel 425, 1012 WP 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:04 Dec 2024

Effective Access to Digital Assets: An XML-based EAD Search System

Junte Zhang
Archives and Information Studies,
Faculty of Humanities,
University of Amsterdam
j.zhang@uva.nl

Khairun Nisa Fachry
Archives and Information Studies,
Faculty of Humanities,
University of Amsterdam
k.n.fachry@uva.nl

Jaap Kamps
Archives and Information Studies,
Faculty of Humanities & ISLA,
Faculty of Science, University of Amsterdam
kamps@uva.nl

ABSTRACT

This paper focuses on the question of effective access methods, by developing novel search tools that will be crucial on the massive scale of digital asset repositories. We illustrate concretely why XML matters in digital curation by describing an implementation of a baseline digital asset search system that is fully XML-driven. The system aims to provide better access to archival material through digital finding aids in the Encoded Archival Description (EAD) standard. Relevant (parts of) archival descriptions within often lengthy and complexly organized digital archival finding aids can be found faster and with more ease. A succinct walk-through of the process of design and implementation of such a system is given, from a higher-level conceptual and generic view, where we start from the actual digital archival finding aid to the eventual delivery of the fonds to the user. Beyond this baseline, we propose a method for automatically providing extra archival context through automatic link detection between archival finding aids. We relate our efforts with the Encoded Archival Context (EAC) initiative.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.3.4 [**Information Storage and Retrieval**]: Systems and Software; H.3.7 [**Information Storage and Retrieval**]: Digital Libraries.

Keywords

Encoded Archival Description (EAD), archival access, information retrieval, information context, Encoded Archival Context (EAC).

1. INTRODUCTION

Digital curation is a recent umbrella term for a comprehensive approach to digital asset management [31]. The essence of digital curation is that it covers the whole live-cycle of a digital asset, from its creation to its future use. The comprehensive approach requires, on the one hand, activities centered on the digital assets (such as appraisal and selection, preservation, and records management), and on the other hand, activities centered on the future use (such as continual enrichment or updating, and effective access methods). The integration of both these aspects is a distinct characteristic of digital curation activities. In this paper, we will focus on the question of effective access methods, by developing novel search tools that will be crucial on the massive scale of digital asset repositories. These new search tools that are tailored to the data at hand, in our case a large collection of digital finding aids, are build from generic components. These search tools are not only valuable for online users but also for digital curators themselves, allowing them to better explore their repository and understand potential use of their digital assets. We illustrate concretely why XML matters in digital curation as our approach is fully XML-driven.

Archives, libraries and museums are memory institutions [9], which store the memories of societies, increasingly also digital assets, and enable their access. The archives have an important usage for users such as historians, as the archives offer primary sources (personal letter, handwritten diary, etc), which are used to reconstruct history. Historians are also the most respected users of archives [28]. These are described in archival descriptions, traditionally in paper form, so the creator or someone else can easier find them again. The archival material consists of records. A comprehensive overview of electronic record management is presented in [6] with the different ontological, epistemological and axiological points of view. An archival finding aid not only represents these records, but also their logical relationships and recorded information about the records, and this all makes an archive accessible.

The archival descriptions are increasingly created digitally in Extensible Markup Language (XML)[1]. The archival descriptions can be considerable in length and numerous in numbers within a finding aid or fonds. The digital finding aids, which are digital assets repositories, are increasingly coded in the Encoded Archival Description (EAD) standard. This standard as described in [14] is the "SGML/XML based document type definition that archives, libraries, and museums are using to create, store, and distribute descriptions of their collections." This is possible, because XML is used to create parse-able and hierarchical object models, in our case EAD, and thus facilitates the sharing of structured data across different information systems, particularly via local networks and the Internet, and also between users and information systems. EAD is maintained by the Library of

This work is licensed under the Creative Commons Attribution-Noncommercial-No Derivative Works 3.0 Unported license. You are free to share this work (copy, distribute and transmit) under the following conditions: attribution, non-commercial, and no derivative works. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-nd/3.0/.

DigCCurr2009, April 1-3, 2009, Chapel Hill, NC, USA

[1] http://www.w3.org/XML/

Congress (LoC) in partnership with the Society of American Archivists (SAA) [18], and is compatible with ISAD(G) [10].

The Retrieving Encoded Archival Descriptions More Effectively (README) project aims to improve archival access by developing better computational methods for finding information in digital finding aids in EAD, such that more precise or direct, and faster access to the archival material is offered. On the one hand, we hope to contribute to archival science by deploying state-of-the-art search technology developed in the Information Retrieval (IR) field to improve access to archival material, and on the other hand we are shedding new lights on IR by testing and evaluating existing search technology on real, vast and steadily increasing amounts of richly structured cultural heritage data in the form of archival finding aids.

The remainder of this paper will deal with both issues, and is setup as follow: first, we enumerate the different topics that frame our research; second, we present the baseline README system and approach; and third, we discuss the horizon beyond the baseline with more research challenges or opportunities, such as with the Encoded Archival Context (EAC) initiative.

2. RESEARCH FRAMEWORKS

Archival material and access

The importance of work processes in archival science is explained in [27]. Resulting from these work processed are for example online digital finding aids in EAD. Initiatives have been taken to facilitate the creation of the finding aids. An instance of an open-source project that deals with creating EAD files is the project Make EAD (proMEAD)[2], which is a web-based native EAD editor, developed in collaboration with the National Archives of the Netherlands. Another web-based editor for EAD is ICA-Atom[3] that is multi-lingual and supports multi-repository collections. Other (commercial) XML editors are also used to create digital archival descriptions in EAD, and hence advancing the 'digitization' of archival materials via digital finding aids, both online as well as offline. These editors use forms, effectively this means that creators and editors do not have to face and thus deal with the actual XML code directly.

In terms of archival access, the importance of user needs is stressed in [20], because the users eventually seek access to the online archival resources. It was argued that studying navigational features and contextual information is important, because these features better help users to understand the archives. This argument is advanced in [30], which suggest that interfaces need to provide a way to a navigational aid that supports users in providing local detail and global view of the finding aids. This suggestion emerged because it was found that the users were lost in the hierarchy, especially in the full text view. Moreover, when engaging with finding aids, users search for archival material from the bottom up and the fullest description necessary at those levels needs to be provided [25].

It is pointed out in [14] that it is in the nature of librarians and archivists to organize things in metadata such as Dublin Core, MARC and EAD. As such, there is no shortage of metadata in finding aids, but "it is a matter of finding the right hook to make them more accessible."

Information Retrieval

A general view

Information retrieval (IR) deals with the representation, storage, organization of, and access to information items [1]. In [24] a succinct overview of the history of Information Retrieval (IR) research is given. IR research consists of two parts: automated indexing and automated retrieval. This research has been done for fifty year, and has become increasingly solid [24]. However, the impact on operational library and information systems has been slow and uneven, an area where we (and this paper) contribute to.

There is an active sub-field within IR called *Focused retrieval*. Focused retrieval goes further than standard IR as it tries to remove the burden on the end-user by providing more direct access to relevant information within a document [11, 23]. For lengthy and complexly structured EADs, it would save users time and effort in locating the archive they want to access.

Focused retrieval on archival finding aids

There is a range of applications within focused retrieval, such as retrieving text passages, retrieving answers to questions, and XML element retrieval by retrieving arbitrary parts of XML files. The latter is an application of focused retrieval that resembles most strongly with the approach as discussed in this paper and attempts to use the XML markup of documents to the fullest. This markup is used to represent the different levels of granularity or complexity (see Fig. 1) of possible interesting text objects. The EAD markup is mostly logical, but EAD also has document-centric features as the markup is also used for the presentation and layout. This granularity can be explicitly seen as structural hints,

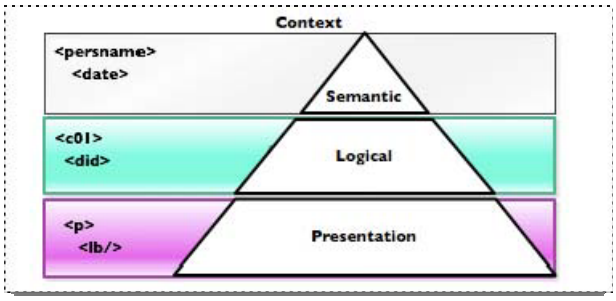

Figure 1: Figure 1: EAD/XML markup and granularity: presentation, logical, semantic aspects.

and used to improve the retrieval of the actual text objects.

An example is the work of [23] with XML Element retrieval on mostly scientific articles from the publisher IEEE. As archival finding aids are richly structured documents, with a complex model of information organization, finding relevant text objects in the files can be difficult. Not only because of the complexity of the organization of the archival material, but also because of the length of the archival descriptions. The quest to provide better access to EADs could be aided by technology such as XML element retrieval. Besides focused retrieval of archival material and other archival information within a finding aid, we can also contribute to improving the archival intelligence [29] of users and visitors of the archives, in other words, enhancing the

[2] http://www.promead.org/

[3] http://ica-atom.org/

understanding of the archival material and the approach of working with these resources through improved usability, resulting not only in focused, but also effective access.

Importance of context

Context is a major concept for archival finding aids. The context of a finding aid partly makes content data significant and of (high) quality, besides also the form and structure. If the structure and context is detached from the actual information, then a finding aid is de-contextualized, and loses its value. Without the (logical) relationships, an archive can facilely degrade to just a collection of historical documents, or as put it in [27]:

> Reliable information becomes unreliable information, high quality information degenerates to information of poorer quality; archives degenerate to documentary collections, evidence turns into documentation, documents into loose data.

Therefore, the main problem in the retrieval and presentation of content data within a finding aid is not only the actual retrieval of the desired information, but also not de-contextualizing the information at the same time. This is one of the major axioms within archival science, and one that we keep in regard. Context is also a relevant feature in IR, and can be used as a common denominator to bridge the gulf.

3. SYSTEMS AND APPROACH

Motivation

Objectives

An effective approach to focused retrieval of archival material, which could enhance archival access, is an intricate challenge. Therefore, we are addressing the following two research objectives.

1. Study effective retrieval techniques tailored to focused retrieval on archival finding aids, taking into account the user's profile and context, the structural context, and the contextual content, of the unit to return.
2. Enhance user access to archival material through digital finding aids from multiple sources.

This paper contributes to the research conducted to fundamental approaches dealing with focused retrieval and focused presentation of archival data. We address the objectives by implementing and testing a search system that offers more focused archival access.

Requirements

Archival practices and principles. The system needs to be compliant with existing archival practices. A key archival principle is *respect des fonds* or the Principle of Provenance; all records of one creator are kept together. Another key principle is Respect for Original Order; all records are maintained in the order the creator had them. It is important that the autonomy of the fonds is respected.

Generalizability. The aim of this article is to give system recommendations and best practice guidelines with the README approach. Henceforth, this approach should be generalizable by other researchers and practitioners in this field as well. Moreover, we validate our approach by buckling it down to different collections from different institutions, which each have different characteristics despite using the same EAD standard.

Open-source. The software and resources that were used should be freely available. We also plan to release our tools and scripts open-source as well. It further facilitates realizable replication of our approach, making our process and results as transparent and creditable as possible. Wherever possible, we stick with state-of-the-art software that is yet to mature, but illustrate the latest (technological) possibilities. Moreover, it means our approach and achieved results can be replicated without any financial investment in software.

Overview of System Architecture

We detail the design of a state-of-the-art vertical search engine, README, for archival descriptions. An overview of the design of the README architecture is depicted in Fig. 2, in which we

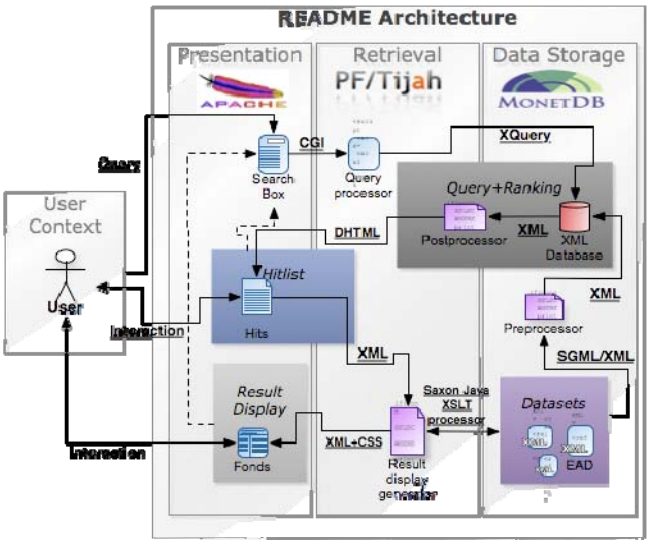

Figure 2: README System Architecture.

follow the conventional 3-tier approach of data storage, retrieval, and the eventual presentation to the user.

The README systems are developed in an out-of-the- box Fedora Core Linux operating system environment, and it is running in this environment as well. The software is also running under the Apache web server. The PC that we use is a standard desktop computer with a dual-core Intel(R) Pentium (R) processor 3.00GHz (no hyperthreads), 200Gb hard-drive, and with 2GB physical main memory.

Data

The digital asset repositories are collections of digital archival finding aids from different institutions, which also differ in length, complexity of structure, and language. The bulk of these finding aids were collected from National Archives of the Netherlands (NA), the International Institute of Social History (ISSH) located in Amsterdam (the Netherlands), and the Archives Hub (AH) in the UK. Moreover, on a smaller scale, we obtained over a hundred of finding aids from the University Libraries of

the University of Amsterdam (UBA) and the Leiden University (UBL).

Instit.	Files	File size (bytes)			Lang.
		Min	Max	Mean	
AH	3,119	1,697	889,218	9,301	English
IISH	2,866	2,048	2,922,445	35,362	Multi.
NA	2,174	5,787	10,720,767	205,749	Dutch
UBL	109	5,577	2,616,931	136,775	Multi.
UBA	60	8,984	51,677	19,196	Dutch

Table 1: General statistics of the finding aids: number of files, file size, and language.

Both libraries have adopted EAD for their special collections and have a relatively small but valuable sample of EADs. The International Institute of Social History and the Dutch National Archives are one of the few institutions in the Netherlands that have numerous full-sized and very complete EADs.

The finding aids from the NA are completely Dutch, those from the AH are completely English, and the EADs from the IISH are a mix of languages, mostly Dutch (about two-third of total), but for instance also German and English. Topic-wise, many finding aids from the Dutch National Archives are about Dutch government agencies, whereas the finding aids from the IISH can be related to topics about social-economic history such as archives about communists and socialists, the Archives Hub's finding aids detail the collections of libraries and museums in the UK.

The sum size of the 8000+ finding aids is 654.5 MB. The finding aids from the Dutch National Archive are significantly larger and lengthier than those from the other two institutions.

Preprocessing of the data

The data that we obtained were unverified preliminary full drafts of the archival descriptions. As a result, we had to pre-process these files in order to make them machine readable as XML. This is a prerequisite, because our approach is fully XML-driven and we can only process data that is at least well formed XML.

For instance, the finding aids from the Archives Hub were in SGML, which had to be converted to XML. Although the finding aids from the NA and the IISH were in essence XML, a considerable subset of their files was not truly well-formed XML as some elements were not properly closed, or valid XML given the EAD specification in the Document Type Definition (DTD) or the XML Schema. Clearly, different expressions by different institutions of the EAD standard are possible, resulting in different XML code, and our approach can deal with these variations robustly. However, some uniformity such as the same set of elements as specified in the EAD standard is necessary. The uniformity is effectuated by pre-processing the files from the different institutions.

Indexing and search

Archival data encoded in EAD is structured data. Commonly used relational databases do not provide a perfect solution to store this type of data. XML databases are developed instead to provide a better solution to capture and preserve the richness of the structure in a data-structure. There are several open-source solutions available, such as eXist [17]. Other alternatives tailored specifically to archival finding aids in EAD are PLEADE (EAD on the Web) [22], Cheshire3 [15] as used by the Archives Hub in the UK, Archon developed at the University of Illinois [21], or the Digital Library eXtension Service (DLXS) software of the University of Michigan[4]. However, the README systems are based on another open-source solution: MonetDB [2] with the XQuery front-end Pathfinder [26] and its information retrieval implementation PF/Tijah [8].

The archival finding aids from the different institutions are indexed in a single main memory database, but in different indexes, where the 8000+ finding aids were processed and stored within minutes. The indexes are built without removal of stop words. Morphological normalization was applied on the words though by using a language-dependent stemmer for each finding aid. The document structure and order is fully preserved in this database, important information that is needed for focused retrieval of the finding aids and dealing with their context.

The queries are processed with XQuery templates. Different templates were used for each of the three README systems. Currently, we do not support yet the use of Boolean query operators (i.e. ‘and’, ‘or’, ‘not’) that is common in conventional information search systems. It is possible to do faceted search by restricting a query to a certain field like <TITLEPROPER> and selecting the collection that one wants to search exclusively in.

Ranking

A core task of IR is the matching process, i.e. given the information need of the user as expressed in a query, and a set of documents where this information can be found, what is the best (or exact) match between this query and a subset of these documents? This matching process is modeled mathematically or statistically, which is then called an *information retrieval model*.

The matching processes of the README systems are based on a unifying model that is called *Language Modeling* (LM) [19]. The essential idea in LM is that given a corpus of paired discourses, A and B, correlations can be established between the features of A and the features of B, so that for a new A, a new B can be estimated [24]. In IR, this means A is the query and B is a relevant document.

LM is an active area of research within IR and other research fields as well, because this general technique is effective for retrieval. We used the standard LM implementation of PF/Tijah as it was available and works in conjunction with our data storage component MonetDB. Using LM, we compute matching scores, which are used to rank the results in descending order according to relevance. As we work with XML files, the system returns any and arbitrary parts (depending on the focus of the granularity) of an XML file and rank these parts separately.

Presentation

Context as interface technique

The importance of context as an interface technique for making documents more understandable is discussed in [7]. Context as an interface technique for IR means that the set of found documents by a system is placed in the environment of other information types. Explicitly, context means showing the relationship of the

[4] http://www.dlxs.org/

finding aids with keywords of a search, collection overviews, descriptive metadata, hyperlink structure, document structure, and the relationships to other documents within the set of finding aids.

Users are getting lost in the hierarchical structure of archival finding aids [30], and to solve this problem, the idea of a user interface that could provide contextual navigation was floated. Such a presentation would support users by providing both the local detail and a global view of the relevant information. Ideally, this would make archival finding aids no longer barriers, but more boundary spanners. It is important to show relevant information in context [3, 16]. The findings in [12], where a study was conducted using a scientific collection of documents (not EAD), also suggest that users appreciate presenting information in context more.

Document Order- Structure- Depth Model

The presentation of focused retrieval of archival material remains an open question. That is why we propose in Fig. 3 our Document Order- Structure- Depth (DOSD) model, which captures our assumptions comprehensively. We use this model as a principle to present and display each retrieved result from an EAD/XML file in context in a user interface, given the document order, the structure and the depth.

A (part of the) screen can be represented as a Cartesian plane, with on the X-axis the depth, and on the Y-axis the structure (granularity, complexity) of the fonds. For example, retrieved text objects that appear in the second quadrant have little depth, little structure, and are in the top of the archival finding aid. Our supposition is that this model could intuitively give focused

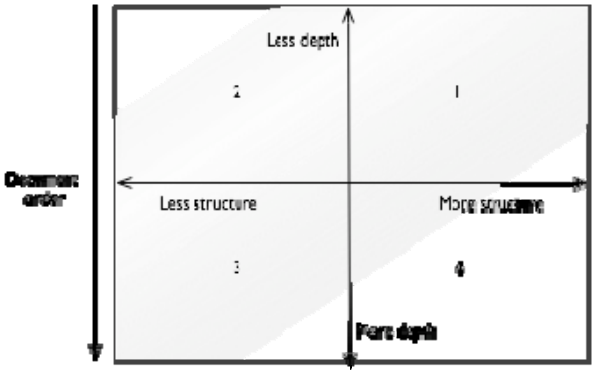

Figure 3: Document Order-Structure-Depth (DOSD) Model.

access to archival material in a natural way. More future research is needed to effectively discover the potential merits and inadequacies of this model.

Hitlist in context

The hitlist is the list that is returned by a system with ranked results; after the user has entered the query, and the system has computed the matching scores given the query and the EAD files. Since this is the first display that the user sees after entering the query, and the first stage of assessing the relevancy of the results, it is worthwhile to investigate not only what is returned, but also how and why. We believe we can provide more focused access to the archival material by showing relevant results directly (in

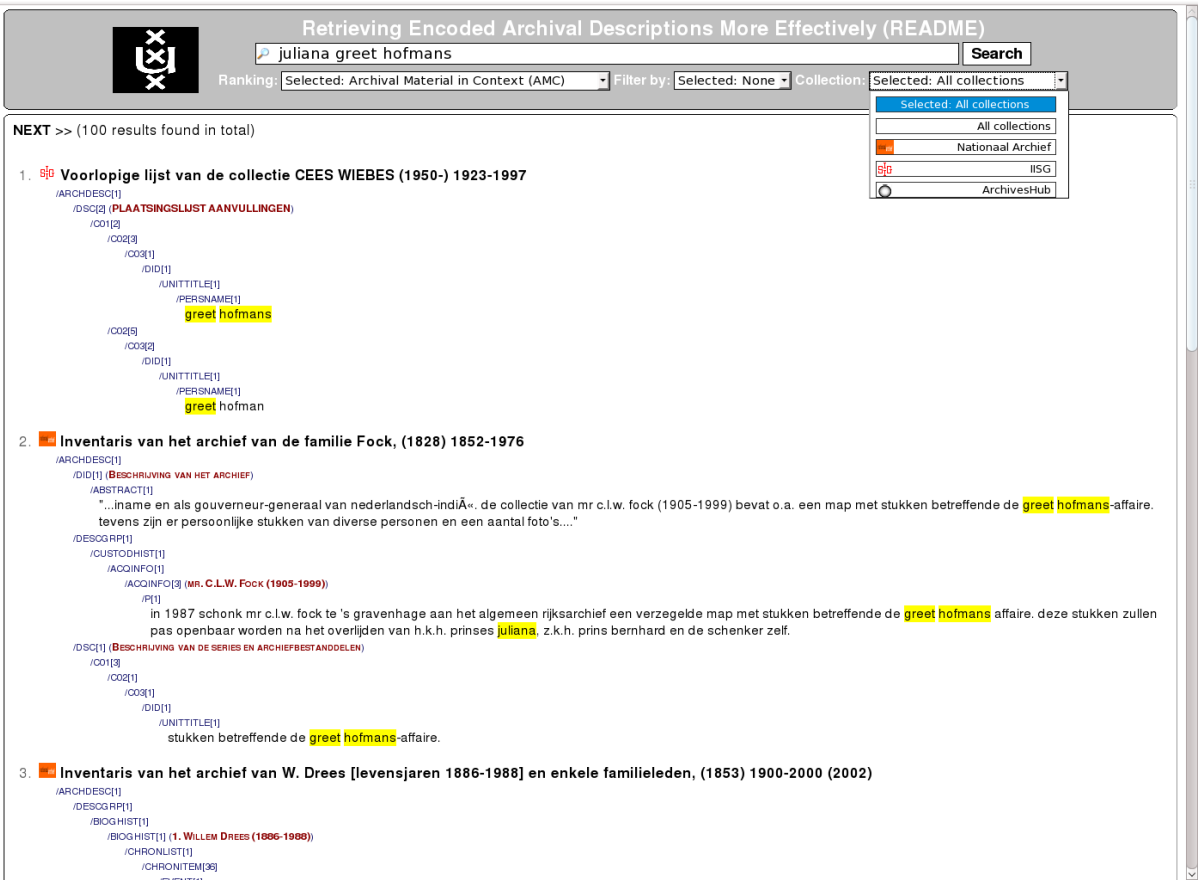

Figure 4: Archival Material in Context (AMC).

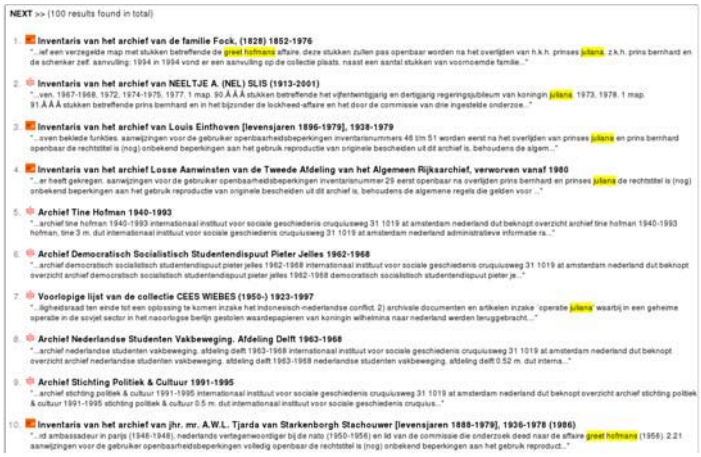

Figure 5: Whole Fonds (WF).

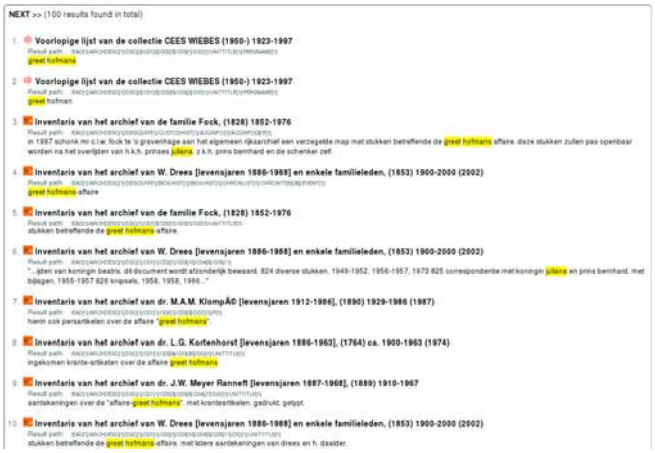

Figure 6: Individual Archival Material (IAM).

context); providing access to only the (beginning of an) entire fonds is therefore neither immediately necessary nor desired.

We materialized the **Archival Material in Context** (AMC) system as depicted in Fig. 4 with that idea in mind. It is an implementation of the DOSD model as discussed before. We used the query "juliana greet hofmans", with the intention to search for information related to former Dutch queen Juliana (1909-2004), her adviser Greet Hofmans (1894-1968), and the subsequent crisis in Netherlands in the 50's of the 20th century. We use this query as an example for all the three systems.

As reported in [32], there are three main principles of presenting a focused hitlist in context, namely *preserving provenance* by grouping most relevant individual items together per finding aid (and thus creator); *preserving document structure* and returning the individual archival items in the hierarchical document order, such that the local and global context of a finding aid can be combined and the archival bond of a fonds is kept in regard; and finally allowing *deep-linking and direct access* so that the user can get actual focused access to the individual items by optimally exploiting the full context.

Individual results can be put in context given the hierarchical XML tree by either showing its ancestors or descendents. The latter is however not always really usable from an IR point of view, because any information in the descendents is already known in the current node which results in overlap of information.

Alternative hitlists

Besides the AMC system, we developed two alternative versions (see Fig. 5 and 6) that retrieves and provides access to the finding aids on a different granularity level, namely on the file level (only top) and element level (anything between top and bottom).

Whole Fonds (WF) The Whole Fonds system as shown in Fig. 5 ranks and retrieves an entire finding aid (document), and is comparable to a conventional document retrieval system like Google or Yahoo. For each result, a title and a snippet (short preview of fonds) are presented.

Individual Archival Material (IAM) Fig. 6 shows the Individual Archival Material system, that retrieves XML element nodes as natural units, and it is therefore comparable to a standard XML element retrieval system that retrieves arbitrary parts of a XML document. Besides the title and the snippet of the element, we also show its result path in XPath.

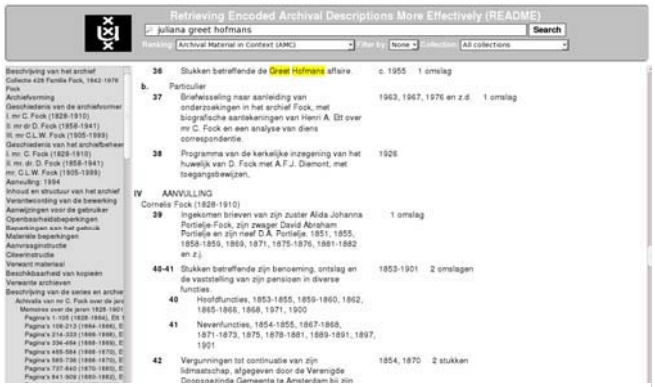

Figure 7: Result display of a complete fonds.

Fonds delivery and result display

Fig. 8 depicts the result display of a whole EAD file. The user gains access to this result display either from the start of the file when using the WF system, or gets directed to at any access point in the file given the result chosen in the IAM or AMC systems. This display is generated dynamically on the fly with XSLT and fully presented in CSS, with on the left side the table of contents (ToC) with the EAD headings <HEADING> and unit titles <UNITTITLE>, and on the right side the full presentation of the actual content of the fonds. Clicking on an item in the table of contents or using the scroll-bars in the browser navigates the user within the finding aid. The original keywords, as originally entered by the user, are highlighted in the fonds.

We do not transform the XML to XHTML, but render the presentation fully using CSS with minimal manipulation of the original XML file. CSS is sufficiently powerful to do this, for example, elements can be presented in tabular form or be filtered by hiding them. As such, we adhere to the original structure and respect the autonomy of the fonds when it is delivered to the user in full in the result display. Moreover, the global context is preserved.

Assessing systems in user study

In [4] we conducted a user study to assess the README system as outlined here. An empirical study was conducted with 9 test persons with sessions that lasted 1.5 hour on average for each participant. The AMC system was compared against a system that would return whole fonds (WF), and one that only returns the

individual archival materials (IAM). In both systems, the context is omitted, and using this comparison we can examine empirically the effects of the context in the hitlist. The experiment consisted of a series of questionnaires with random iterations of interaction with the three systems. Table 2 shows post-task questions and the responses toward features in the three different types of hitlists.

Q3.13 How satisfied were you with the information provided in the hitlist?
Q3.14: Was the overview of results clear?
Q3.15: Was it easy to select the most promising result?

	Q3.13	Q3.14	Q3.15
WF	3.78 (0.67)	3.67 (0.87)	3.44 (0.88)
IAM	3.11 (0.78)	2.89 (0.93)	3.11 (1.17)
AMC	3.33 (0.87)	3.22 (0.67)	4.11 (0.78)

Table 2: Questions and responses on hitlist: mean scores and standard deviations (in brackets).

The overview of the results was found most clearly in the WF system (Q3.13), likely because of its simplicity and it is conventional (and thus familiar) presentation. Henceforth, the test persons tend to be most satisfied with the information provided on the hitlist of the WF system (Q3.14). However, they found it easiest to select the most promising result in the AMC system (Q3.15). The IAM system was least appreciated. The results of the user study show that AMC system is not optimal, but achieves its objective of offering users focused archival access. The study gave concrete suggestions on how to improve the user interface by presenting the context in a more intuitive way, which we will explore in future research. Effectively, it means combining the best of the WF and AMC interfaces.

4. Concluding Discussion: Beyond Baseline

This paper focused on the question of effective access methods, by developing novel search tools that will be crucial on the massive scale of digital asset repositories. We illustrated concretely why XML matters in digital curation by presenting a fully XML-driven system description for digital assets. Some of the challenges that we faced to improve information access in the archives were identified. We proposed an approach to deal with these challenges.

However, there are still roadblocks lying ahead in terms of providing information access with EAD. For example, the ranking of the results, especially on the element level, has not been optimized yet in the IR model – crucial in providing focused access. To optimize the ranking of the results, we will conduct experiments to discover optimal settings in our retrieval models for retrieving desired archival descriptions more effectively - at least the ones that are available to our research by creating an EAD test collection.

The research in this paper has been centered on the retrieval and presentation of the archival descriptions from a document-centric and hierarchical structural point of view. Intrinsically, other views exist with additional applications of XML. For example, a promising direction is to help enrich EADs with link detection methods, and provide access to the archival descriptions by exploiting additional *relational structures* besides the *hierarchical structure*, which we have done so far. In other words, certain texts in a finding aid can be clicked and directs a user to a different finding aid or a different point in the same finding aid. There could be special use for automatically generated links within a fonds itself, specifically the result display as illustrated in Section 3.7.5. In case the user chooses to go beyond the hitlist, usually in the case of serendipitous information seeking ('browsing') task, then EADs enriched with links could provide additional focused access to the archival material by saving the user browsing time.

In [33] we set the first steps in this direction by presenting preliminary work on this topic, where we showed we could automatically detect occurrences of person names with high accuracy, both in and between archival descriptions. This allows us to create (pseudo) encoded archival context descriptions that provide novel means of navigation, improving access to the vast amounts of archival data not only through the inventories, but also through the actors. This means that besides discovering relationships between the fonds in one collection, we can also detect them between the fonds in the same collection, and even between different institutions. The concept of *parallel provenance* is strongly related to this, and is addressed by Ketelaar [13], which he paraphrased as "two or more entities residing in a different context as establishing the provenance of a record, even when they are involved in different kinds of action, for example creation and control."

Archival context may be constructed through the use of authority records that capture information about the record creators or actors (corporations, persons, or families) and the context of the record creation. By separating the record creator's descriptions from the records or resources descriptions themselves, we can automatically create 'links' from all occurrences of the creators to this context. The resulting descriptions of record creators can be encoded in XML and matched using the emerging Encoded Archival Context (EAC) standard.

Currently, EAC has only been applied experimentally. One of the main barriers to adoption is that it requires substantial effort to adopt EAC. The information for the creator's authority record is usually available in some form (for example, EAD descriptions usually have a detailed field <BIOGHIST> about the archive's creator). However, linking such a context description to occurrences of the creator in the archival descriptions requires more explicit structure than that is available in legacy data.

Having established these relations, we can create physical links by directly linking two or more fonds together, for example in XLink. We can also extract information existing in another fonds to create pseudo archival context descriptions, or we can even automatically construct an authority record in EAC by discovering co-references. These are all steps towards even more effective information access using EAD.

5. ACKNOWLEDGMENTS

Junte Zhang gratefully acknowledges Henny van Schie of the National Archives of the Netherlands for his support and insightful discussions on EAD. This research was supported by the Netherlands Organization for Scientific Research (NWO) Innovative Research Grants Program (VIDI Scheme) under project number 639.072.601.

6. REFERENCES

[1] Baeza-Yates, R., and Ribeiro-Neto, B. Modern Information Retrieval. Addison Wesley, May 1999.

[2] Boncz, P.A., Grust, T., van Keulen, M.,Manegold, S., Rittinger, J., and Teubner, J. MonetDB/XQuery: A Fast XQuery Processor Powered by a Relational Engine. In Proceedings of the ACM SIGMOD International Conference on Management of Data (Chicago, IL, USA, 2006).

[3] Dumais, S., Cutrell, E., and Chen, H. Optimizing Search by Showing Results in Context. In Proceedings of the ACM SIGCHI Conference (Seattle, WA, April 2001), pp. 277–284.

[4] Fachry, K.N., Kamps, J., and Zhang, J. Access to archival material in context. In Proceedings of the 2nd Symposium on Information Interaction in Context (2008), ACM, pp. 102-109.

[5] Gilliland-Swetland, A. Electronic records management. Annual Review of Information Science and Technology 39, 1 (2005), 219–253.

[6] Hearst, M.A. User interfaces and visualization. In Modern Information Retrieval, R. Baeza-Yates and B. Ribeiro-Neto, Eds. Addison Wesley, 1999, pp. 257–323.

[7] Hiemstra, D., Rode, H., van Os, R., and Flokstra, J. PF/Tijah: text search in an XML database system. In Proceedings of the 2nd International Workshop on Open Source Information Retrieval (OSIR) (2006), pp. 12–17.

[8] HjØrland, B. Documents, memory institutions and information science. Journal of Documentation 56 (2000), 27–41.

[9] ISAD(G). General International Standard Archival Description, second ed. International Council on Archives, Ottawa, 1999.

[10] Joty, S.R., and Al-Hasan, S. Advances in focused retrieval: A general review. In Proceedings of the 10th IEEE International Conference on Computer and Information Technology (Dec. 2007), IEEE, pp. 1–5.

[11] Kamps, J., and Sigurbjörnsson, B. What do users think of an XML element retrieval system? In Advances in XML Information Retrieval and Evaluation (INEX 2005) (2006), vol. 3977 of LNCS, Springer, pp. 411–421.

[12] Ketelaar, E. Archives in the Digital Age: New Uses for an Old Science. Archives & Social Studies: A Journal of Interdisciplinary Research 1, 0 (2007), 167-191

[13] Kiesling, K. Metadata, metadata, everywhere - but where is the hook? OCLC Systems & Services 17 (2001), 84–88.

[14] Larson, R.R., and Sanderson, R. Cheshire3: retrieving from tera-scale grid-based digital libraries. Proceedings of the 29th annual international ACM SIGIR conference (New York, NY, USA, 2006), ACM, pp. 730–730.

[15] Lin, J., Quan, D., Sinha, V., Bakshi, K., Huynh, D., Katz, B., and Karger, D. R. The role of context in question answering systems. In CHI '03: CHI '03 extended abstracts on Human factors in computing systems (New York, NY, USA, 2003), ACM, pp. 1006–1007.

[16] Meier, W. eXist: An Open Source Native XML Database. In Revised Papers from the NODe 2002 Web and Database-Related Workshops on Web,Web-Services, and Database Systems (London, UK, 2003), Springer-Verlag, pp. 169–183.

[17] Pitti, D.V. Encoded archival description: An introduction and overview. D-Lib Magazine 5, 11 (1999). http://www.dlib.org/dlib/november99/11pitti.html.

[18] Ponte, J.M., and Croft, W.B. A Language Modeling Approach to Information Retrieval. Proceedings of the 21st ACM SIGIR Conference (1998), ACM, pp. 275-281

[19] Rosenbusch, A. Are our users being served? A report on online archival databases. Archives and Manuscripts 29 (2001), 44–61.

[20] Schwartz, S.W., Prom, C.J., Rishel, C.A., and Fox, K.J. Archon: A unified information storage and retrieval system for lone archivists, special collections librarians and curators. Canadian Journal of Library and Information Practice and Research 2, 2 (1997).

[21] Sévigny, M., and Clavaud, F. PLEADE – EAD for the Web. DigitCULT.Info, 6 (December 2003), 16–18.

[22] Sigurbjörnsson, B. Focused Information Access using XML Element Retrieval. SIKS dissertation series 2006-28, University of Amsterdam, 2006.

[23] Spärck Jones, K. Information retrieval and digital libraries: lessons of research. In IWRIDL '06: Proceedings of the 2006 international workshop on Research issues in digital libraries (New York, NY, USA, 2007), ACM, pp. 1–7.

[24] Stockting, B. Time to Settle Down? EAD Encoding Principles in the Access to Archives Programme (A2A) and the Research Library Group's Best Practice Guidelines. Journal of Archival Organization 3 (2004), 7–23.

[25] Teubner, J. Pathfinder: XQuery Compilation Techniques for Relational Database Targets. PhD thesis, Technische Universität München, Munich, Germany, 2006.

[26] Thomassen, T. A first introduction to archival science. Archival Science 1, 4 (2001), 373–385.

[27] Tibbo, H.R. Primarily history: historians and the search for primary source materials. In Proceedings of the 2nd JCDL (New York, NY, USA, 2002), ACM, pp. 1–10.

[28] Yakel, E and Torres, D.A., AI: Archival Intelligence and User Expertise. American Archivist 66, 1 (2003), 51-78.

[29] Yakel, E. Encoded archival description: Are finding aids boundary spanners or barriers for users? Journal of Archival Organization 2, 1-2 (2004), 63–77.

[30] Yakel, E. Digital Curation, OCLC Systems & Services: International digital library perspectives, 23 (2007), 335–340.

[31] Zhang, J., Fachry, K.N., and Kamps, J. Access to archival finding aids: Context matters. In ECDL (2008), vol. 5173 of Lecture Notes in Computer Science, Springer, pp. 455–457.

[32] Zhang, J., Fachry, K.N., and Kamps, J. Automatic link-detection in encoded archival descriptions. In Proceedings of Digital Humanities 2008 (Finland, 2008), University of Oulu, pp. 226–228.