Innovational Research Incentives Scheme
VIDI scheme 2006

1 Registration form (basic details)

1.1 Title of research proposal

Retrieving Encoded Archival Descriptions More Effectively (README)

1.2 Summary of research proposal

XML retrieval is a very active branch of Information Retrieval addressing the focused retrieval of semi-structured data. One of the main open problems in XML retrieval is to understand the information seeking behavior of real element retrieval users. Such understanding is crucial for further progress, since retrieval models, and even metrics, make explicit assumptions on user preferences—assumptions that are unwarranted by our current knowledge. In archival science, modern archival finding aids are encoded in XML using the *Encoded Archival Description* standard. However, up to now, archivists have focused almost entirely on the "supply" side rather than the "client" side: Web sites with digital finding aids often have no search functionality, and users often fail to locate the relevant archival material, frustrated by the intricacies of the lengthy and complexly structured finding aids.

The proposed project aims to investigate the contextual, focused retrieval of archival material in digital finding aids, from both a user-centered (key objective 1) and a system-centered (key objective 2) perspective. This will make dual contributions to both XML retrieval and to archival access. The user studies (key objective 1) will provide insight into user information seeking behavior in XML element retrieval, crucial for validating retrieval models and evaluation metrics. At the same time, these studies will provide crucial feedback on the strengths and weaknesses of digital finding aids for users in search of archival material. The developed retrieval techniques (key objective 2) will significantly extend current XML retrieval models by building on the results of the user studies, and by taking into account the user's profile and context, and the textual context of the unit to return. At the same time, the resulting search engine will greatly enhance user access to archival material in digital finding aids.

Keywords: Information Retrieval, Semi-structured data, XML retrieval, Archival Finding Aids, EAD.

2 Research proposal

2.1 Description of the proposed research

2.2 Research topic

Motivation The direct motivation for the proposal is the perfect match of two closely related problems in two different disciplines.

First, in the field of Information Retrieval (IR), there is the very active branch of XML retrieval, studying the focused retrieval of semi-structured data. XML retrieval is a radical departure from standard IR. In traditional document retrieval, the unit of retrieval is fixed and known: only the entire document can be returned to the user. In XML element retrieval, each individual component can be retrieved, ranging from the full-blown article to individual paragraphs or italicized phrases [12]. One of the main open problems in XML retrieval is to understand the information seeking behavior of real element retrieval users [45, 46]. Such understanding is crucial for further progress, since retrieval models, and even metrics, make explicit assumptions on user preferences—assumptions that are unwarranted by our current knowledge.

Second, in archival science, modern archival finding aids are digital, using SGML or XML based standards like the *Encoded Archival Description* [EAD, 8]. Each finding aid following the *International Standard for Archival Description* [ISAD, 15] is a multilevel description proceeding in a top-down fashion, first describing the whole archive, then its major sub-components, and so on. It results in a hierarchical structure that can be naturally encoded in the XML structure of EAD [32].[1] Having digital finding aids holds the promise to significantly improve access to archival material. However, up to now, archivists have focused almost entirely on the "supply" side rather than the "client" side [34]. Web sites with digital finding aids often have no search functionality, making it a difficult task to locate the appropriate finding aids [25]. Moreover, once a finding aid is found, users often fail to locate the relevant archival material, frustrated by the intricacies of the lengthy and complexly structured finding aids [48]. In short, the full potential of digital finding aids fails to be realized.

In sum, archival access is in need of the focused search techniques for digital finding aids that XML retrieval can provide, and XML retrieval is in need of the real-world application of XML element retrieval that archival finding aids can provide.

Overall Aim and Key Objectives The overall aim of the proposed project is

- to investigate the contextual, focused retrieval of archival material in digital finding aids, from both a user-centered and a system-centered perspective.

[1] Technically speaking, EAD is a fairly complex *document type definition* (DTD) with 166 elements, of which 150 can have one or more attributes [8, 9].

The project has the following *key objectives*:

- *Key objective 1* (User-centered approach): Study the information seeking behavior of users of archival finding aids, for a range of user profiles having varying degrees of expertise on archival descriptions.
- *Key objective 2* (System-centered approach): Study effective retrieval techniques tailored to XML element retrieval on archival finding aids, taking into account the user's profile and context, and the textual context of the unit to return.

Scientific Background and Innovative Elements XML retrieval is a very active branch of IR addressing the focused retrieval of semi-structured data. The main thrust, since 2002, is the annual INitiative for the Evaluation of XML retrieval [INEX, 12]. Although much progress has been made [e.g., 3, 10, 16], the time has come to take stock and reflect upon the XML retrieval task. As it turns out, very little is known about XML element retrieval in action: How do users interact with an actual XML element retrieval system? Which elements do users find most useful? Etc. One of the main open problems in XML retrieval is to understand the information seeking behavior of real element retrieval users. Such understanding is crucial for further progress, since retrieval models, and even metrics, make explicit assumptions on user preferences—assumptions that are unwarranted by our current knowledge.

In INEX, a collection of full-text scientific articles is used. This test collection has served us well, but is perhaps not the most suitable for taking the next steps in XML element retrieval. Rather than appreciating all elements equally, users seem to have a bias for the full article—much against the very spirit of XML retrieval [20]. As Trotman [45, p.64] has it:

> For element retrieval to be useful it is necessary to identify the environment in which it might be used. Given there is a user who wants the said technology, it's possible to identify the characteristics of the document collection they are using. The document collection must be in a markup language that contains elements. This might be XML, SGML or any other mark-up language. The documents must contain several disparate parts (elements) that, while atomic in themselves, are also atomic in the context of the document.

In fact, Trotman [45, p.69] seems not very optimistic that such a collection exists:

> Identifying an application of element retrieval is a vital first step. If it isn't possible to identify such an application, such an application may not exit. Unless the community can collectively identify such an application methodological issues will continue plague the research.

Archival finding aids are a perfect XML element retrieval application. Modern archival finding aids are digital, using SGML or XML based standards like the *Encoded Archival Description* [EAD, 8]. An archive is no arbitrary collection of documents, but the unique records of a particular person, family, or corporate body. Archival documents are basically the paper trail of individuals living their lives, or corporate bodies carrying out their functions. Individual documents, which may be in any form or medium, are interrelated in complex ways, sharing a common provenance. The object of interest is not so much an individual item, but the items within their original context. Only in this way the legal or historical evidence can be preserved [7]. Hence, each finding aid following the *International Standard for Archival Description* [ISAD, 15] is a multilevel description, proceeding in a top-down fashion, first describing the whole archive (or *fonds*), then its major sub-components, and so on. It results in a hierarchical structure ranging from the *fonds*, to *subfonds*, *series*, *subseries*, *files*, until the individual *items*. Each of these subdivisions corresponds to a natural grouping of archival material, for example, to administrative subdivisions of the originating organization, to geographical areas, to chronological periods, or to a particular personal or corporate activity, etc. Hence, depending on the user's information need, each of these groupings on each of these levels could be a natural unit to return. Recall, the object of interest is often derived from parts of the paper trail, for example, how a particular personnel department functioned during war-time may be derived from their recruitment decisions. In this sense, each archive has implicitly a multitude of stories to tell [24].

Digital finding aids, structured in EAD, naturally support the hierarchical arrangement using SGML or XML [32]. Having digital finding aids holds the promise to significantly improve access to archival material. However, their full potential fails to be realized. In a recent survey of EAD finding aids on the Web, Kim [25, p.53] notes

> Search functions are a growing necessity on EAD sites. However, only seven web sites out of seventeen allowed users to search for EAD finding aids.

And even if a user is able to locate a relevant finding aid, which may range in length from a few pages up to the size of a multi-volume book, there are more problems ahead. The results of an EAD user study where "not encouraging": users got lost in the hierarchy of the finding aid, where overwhelmed by the length of the finding-aid, and stumbled over unfamiliar terms used to label parts of the finding aid [48, 49]. Archivists have, so far, focused almost entirely on the "supply" side rather than the "client" side [34]. Recent review articles [4, 6] reveal how little is known about the information seeking behavior of users of digital finding aids. As early as 1998, Tatum [43] warned that ignoring users would cause problems with EAD development and its promotion and acceptance.

The proposed project will make dual contributions to both XML retrieval and to archival access. The user studies (key objective 1) will provide insight into

user information seeking behavior in XML element retrieval, crucial for validating retrieval models and evaluation metrics. At the same time, these studies will provide crucial feedback on the strengths and weaknesses of digital finding aids for users in search of archival material. The developed retrieval techniques (key objective 2) will significantly extend current XML retrieval models by building on the results of the user studies, and by taking into account the user's profile and context, and the textual context of the unit to return. At the same time, the resulting search engine will greatly enhance the user access to archival material through digital finding aids.

2.3 Approach

We will frame the research of the proposal as an information retrieval problem: a user wants to access archival material for some reason—she has an information need—and the system should provide her with the information relevant for her information need, regardless of how she expresses herself. Our approach relies heavily on user-centered and system-centered approaches in IR [35, 41].

Key objective 1: User-centered approach We will study the information seeking behavior of real XML element retrieval users, with varying degrees of domain knowledge, and their natural information needs.

- **Collection**. We have already started building up a collection of EAD finding aids. Fortunately, many record holding institutions allow for downloading digital finding aids from their sites—in fact, they may even be legally obliged to provide public access to them. We also plan to collect query log files from some of the major institutions serving digital finding aids on the web.

- **Evaluation**. We will set up a proper evaluation test-suite for a large collection of digital finding aids. We will construct an evaluation test suite, consisting of a document collection, a set of search topics, and user judgments on the relevance of documents for these topics. This investment will create a reusable test suite for evaluating the effectiveness of retrieval techniques for archival finding aids.

- **User Study**. Continuing our earlier efforts [18], we will conduct a series of user studies and interactive experiments [42]. We want investigate different user types, ranging from *naive users*, such as casual visitors of an archive's web site, to *expert users*, such as historians and archivists. We plan to deal with a rich set of user profiles and task types. Additionally, we intend to collaborate with an archival institution to conduct a operational study of their user population. We have already established initial contacts with a number of archives holding EAD based digital finding aids.

We will certainly consider the embedding of these activities within the INEX initiative (where the applicant is one of the organizers responsible for retrieval

tasks). Note that finding staff or students to help with topic development and assessments may be easier on this type of collection for INEX participants active in Library and Information Science schools. Precisely these participants tend to be the most interested in conducting user studies as part of the current INEX Interactive Track [27, 44].

Key objective 2: System-centered approach Our modeling framework will be based in so-called *statistical language models* for information retrieval [11]. Language models can be transparently combined with information from other models, or with particular non-content biases [26].

The bottle-neck for providing more focused retrieval is in the shallowness on the client-side, i.e., users who provide no more than a few keywords to express their complex information needs. An approach would be to let users articulate their search request in a structured query language [22, 38]. Here, our main approach is to try to avoid that the user has to provide explicit references to the internal document structure, at least not in an initial query. We aim to use the textual context provided by the structure of the archival finding aid, and to use the user's context implicitly and explicitly.

- **Document's context**. The textual context of the element to return may contain vital retrieval cues. This is generally the case in XML retrieval, but caries special importance in the case of finding aids. In archival science, along with content and structure, context is one of the three fundamental aspects of a record—again, a perfect match with XML retrieval. In language models for document retrieval, the context of the whole collection is taken into account by smoothing document models in order to account for data sparseness [50]. Inspired by this, Sigurbjörnsson et al. [37] introduced a three level mixture model for XML retrieval—consisting of the collection, the full article, and the XML element to be returned—and showed that the article context was very effective for zooming in on the desired XML elements. The resulting 'document pivot' was also shown to improve other retrieval models [31]. Recently, more generalized contextualization models have been proposed [2]. We will significantly extend current models, by studying fine-grained contextual models, taking into account the special structure of archival finding aids, and based on insights from user navigation within archival finding aids.

- **User's context**. Current general purpose search engines follow a "one size fits all" approach, de-emphasizing the differences between user profiles, and task scenarios. There will exist major differences in the types of information needs of domain experts such as historians and archivists, and casual visitors of an archival web site. Information about the user's profile might come implicitly from matching her searching behavior with stored user profiles, or explicitly from user interaction. Based on an initial query returning documents from different subdivisions of the finding aid, one

could also engage in fine-grained interaction similar to Google's spelling suggestions, for instance "Do you want to focus on *Amsterdam*?", "Do you want to focus on documents from around 11 September 2001?", "Do you want to focus on documents from Europe?" [33]. Exploiting the structure of the archival finding aids, we will develop retrieval models aware of the user's context [13, 14, 28].

- **Contextual Language Models**. Language models for retrieval are easily combined with information from other models. Our approach to contextual models is related to relevance models [29, 30] in which the set of initially retrieved documents is included in the model as a layer between the document and the collection model. Based on the user's profile, her earlier search results, or simply on the structure of the finding aid, we will be able to derive topical models, i.e., models of the language typically used in major subdivisions of the archive. For instance, these models could correspond to particular administrative units or activities, geographical areas, or temporal periods. These models can be used to provide a more targeted search. Insights into user seeking behavior, as well as specific user preferences can be straightforwardly modeled as prior probabilities [16].

Note that, although we focus on archival finding aids, our findings will be carefully cross-validated against other collections such as those used in the INEX initiative [12]. Our approach to XML retrieval has always focused on techniques that could be, in principle, applied to arbitrary document-oriented XML [e.g., 19, 37, 39, 40].

2.4 Innovation

The proposed project will make a number of contributions to XML retrieval:

- First, an XML element retrieval collection tied to a natural application.
- Second, user studies shedding light on the information seeking behavior of XML element retrieval users. This will provide crucial validation of user assumptions in XML retrieval models and metrics.
- Third, contextual retrieval techniques that take into account the user's profile and context, and the context of the result elements.

Moreover, the proposed project will make a number of contributions to archival access:

- First, a scientific, reusable experimental test-suite to evaluate the effectiveness of retrieving archival material in finding aids, consisting of a large collection of finding aids; a set of test topics and relevance judgments.

- Second, provide a state-of-the-art search engine for archival finding aids, not only directing the user to the relevant finding aids but pin-pointing her directly to the relevant parts.

- Third, user studies shedding light on the information seeking behavior of user of archival finding aids. This will provide crucial feedback on the strengths and weaknesses of current digital finding aids.

Significance: potential contributions for science, technology, and/or society
The wealth of information available on the Internet creates an urgent need for methods that can help organize, cluster and classify information, and improve ways of accessing it (such as information disclosure and knowledge distillery). The current rate of growth of information on the Internet leads to ineffectiveness and myopia, unless it is paralleled with a similar growth in effectiveness of information retrieval techniques. Part of the answer will be in information pinpointing approaches such as XML retrieval and question answering, that provide more focused retrieval than the traditional retrieval of whole, atomic documents. In question answering [47] the task is to return exact answers to questions, rather than complete documents potentially containing the answer. In XML retrieval, the internal structure of document is exploited in order to identify the relevant information in it. The proposed research holds the promise to significantly advance current XML element retrieval techniques.

The proposed research will strengthen existing research ties between the fields of Computer Science and the Humanities. Although the field of Information Retrieval has always enjoyed interest from both the disciplines of computer science and humanities (library, archival, and museum science), direct collaborations between the two blood groups have been few and far between. The advent of digital cultural heritage—either in the form of digital descriptions or in the form of digital multimedia objects—have caused traditional barriers between the disciplines to erode. Due to web-based presentation, in combination with the current and shared interest in users of cultural heritage, the different traditions of describing cultural heritage in library catalogs, archival finding aids, and museum registers seem to be converging. A convergence of the content expertise in the traditional fields and the technical expertise in computer science is timely and urgent.

The importance of improving the means of access to archival material can hardly be underestimated. Archives are the memory of our society, providing not just information but also legal and historical evidence about our past, and preserving it for generations to come [36]. The advent of digital documents of varying media types is radically changing face of archival science [5]. Building an effective search engine for archival finding aids, as proposed in this project, seems a crucial precursor to the near future in which archives will contain predominantly digitally born documents. Here, full-text searching the electronic documents will provide crucial retrieval cues, that need to be put into perspective using the

context provided by the digital finding aids. The proposed techniques allow for the seamless integration of digital finding aids with digital documents.

2e. Literature references

[1] Archives Hub. Revolutionising access to the archives of UK universities and colleges, 2006. `http://www.archiveshub.ac.uk/`.

[2] P. Arvola, M. Junkkari, and J. Kekäläinen. Generalized contextualization method for XML information retrieval. In A. Chowdhury, N. Fuhr, M. Ronthaler, and H.-J. Schek, editors, *CIKM'05: Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, pages 20–27. ACM Press, New York NY, USA, 2005.

[3] D. Carmel, Y. S. Maarek, M. Mandelbrod, Y. Mass, and A. Soffer. Searching XML documents via XML fragments. In C. Clarke, G. Cormack, J. Callan, D. Hawking, and A. Smeaton, editors, *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 151–158. ACM Press, New York NY, USA, 2003.

[4] L. R. Coats. Users of EAD finding aids: Who are they and are they satisfied? *Journal of Archival Organization*, 2:25–39, 2004.

[5] T. Cook. What is past is prologue: A history of archival ideas since 1898, and the future paradigm shift. *Archivaria*, 43:38–39, 1997.

[6] K. Cruikshank, C. Daniels, D. Meissner, N. L. Nelson, and M. Shelstad. How do we show you what we've got? access to archival collections in the digital age. *Journal of the Association for History and Computing*, III(2), 2005.

[7] L. Duranti. Origin and development of the concept of archival description. *Archivaria*, 35:47–54, 1992.

[8] EAD. Encoded archival description version 2002, 2006. `http://www.loc.gov/ead/`.

[9] EAD Help Pages. Ead round table of the Society of American Archivists, 2006. `http://www.archivists.org/saagroups/ead/`.

[10] N. Fuhr and K. Großjohann. XIRQL: A query language for information retrieval in XML documents. In D. H. Kraft, W. B. Croft, D. J. Harper, and J. Zobel, editors, *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 172–180. ACM Press, New York NY, USA, 2001.

[11] D. Hiemstra. *Using Language Models for Information Retrieval*. PhD thesis, Center for Telematics and Information Technology, University of Twente, 2001.

[12] INEX. INitiative for the Evaluation of XML retrieval, 2006. `http://inex.is.informatik.uni-duisburg.de/`.

[13] P. Ingwersen and K. Järvelin. *The Turn: Integration of Information Seeking and Retrieval in Context*. The Kluwer International Series on Information Retrieval. Springer Verlag, Heidelberg, 2005.

[14] P. Ingwersen, K. Järvelin, and N. Belkin, editors. *Proceedings of the ACM SIGIR 2005 Workshop on Information Retrieval in Context (IRiX)*, 2005. Royal School of Library and Information Science, Copenhagen Denmark.

[15] ISAD(G). *General International Standard Archival Description*. International Council on Archives, Ottawa, second edition, 1999.

[16] J. Kamps, M. de Rijke, and B. Sigurbjörnsson. Length normalization in XML retrieval. In M. Sanderson, K. Järvelin, J. Allan, and P. Bruza, editors, *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 80–87. ACM Press, New York NY, USA, 2004.

[17] J. Kamps, M. de Rijke, and B. Sigurbjörnsson. The importance of length normalization for XML retrieval. *Information Retrieval*, 8:631–654, 2005.

[18] J. Kamps, M. de Rijke, and B. Sigurbjörnsson. University of Amsterdam at INEX 2005: Interactive track. In N. Fuhr, M. Lalmas, S. Malik, and G. Kazai, editors, *INEX 2005 Workshop Pre-Proceedings*, pages 327–332, 2005.

[19] J. Kamps, M. Marx, M. de Rijke, and B. Sigurbjörnsson. The importance of morphological normalization for XML retrieval. In N. Fuhr, N. Gövert, G. Kazai, and M. Lalmas, editors, *Proceedings of the First Workshop of the INitiative for the Evaluation of XML retrieval (INEX)*, pages 41–48. ERCIM Publications, 2003.

[20] J. Kamps, M. Marx, M. de Rijke, and B. Sigurbjörnsson. XML retrieval: What to retrieve? In C. Clarke, G. Cormack, J. Callan, D. Hawking, and A. Smeaton, editors, *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 409–410. ACM Press, New York NY, 2003.

[21] J. Kamps, M. Marx, M. de Rijke, and B. Sigurbjörnsson. Best-match querying from document-centric XML. In S. Amer-Yahia and L. Gravano, editors, *Proceedings of the Seventh International Workshop on the Web and Databases (WebDB 2004)*, pages 55–60, 2004.

[22] J. Kamps, M. Marx, M. de Rijke, and B. Sigurbjörnsson. Structured queries in XML retrieval. In A. Chowdhury, N. Fuhr, M. Ronthaler, and H.-J. Schek, editors, *CIKM'05: Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, pages 2–11. ACM Press, New York NY, USA, 2005.

[23] J. Kamps, M. Marx, M. de Rijke, and B. Sigurbjörnsson. Understanding content-and-structure. In A. Trotman, M. Lalmas, and N. Fuhr, editors, *Proceedings of the INEX 2005 Workshop on Element Retrieval Methodology*, pages 14–21. University of Otago, Dunedin New Zealand, 2005.

[24] E. Ketelaar. Tacit narratives: The meanings of archives. *Archival Science*, 1: 131–141, 2001.

[25] J. Kim. EAD encoding and display: A content analysis. *Journal of Archival Organization*, 2:41–55, 2004.

[26] W. Kraaij, T. Westerveld, and D. Hiemstra. The importance of prior probabilities for entry page search. In K. Järvelin, M. Beaulieu, R. Baeza-Yates, and S. H. Myaeng, editors, *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 27–34. ACM Press, New York NY, USA, 2002.

[27] B. Larsen, S. Malik, and T. Tombros. The interactive track at INEX 2005. In N. Fuhr, M. Lalmas, S. Malik, and G. Kazai, editors, *INEX 2005 Workshop Pre-Proceedings*, pages 313–327, 2005.

[28] B. Larssen, editor. *ACM SIGIR 2004 workshop on Information Retrieval in Context*, 2004.

[29] V. Lavrenko and W. Croft. Relevance-based language models. In *Proceedings of the 24th ACM Conference on Research and Development in Information Retrieval (SIGIR'01)*, pages 120–128, 2001.

[30] V. Lavrenko and W. Croft. Relevance models in information retrieval. In W. Croft and J. Lafferty, editors, *Language Modeling for Information Retrieval*, pages 11–56. Kluwer Academic Publishers, 2003.

[31] Y. Mass and M. Mandelbrod. Component ranking and automatic query refinement for xml retrieval. In N. Fuhr, M. Lalmas, S. Malik, and Z. Szlávik, editors, *Advances in XML Information Retrieval. Third Workshop of the INitiative for the Evaluation of XML Retrieval, INEX 2004*, volume 3493 of *Lecture Notes in Computer Science*, pages 73–84. Springer Verlag, Heidelberg, 2005.

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

[33] H. Rode and D. Hiemstra. Conceptual language models for context-aware text retrieval. In *Proceedings of the 13th Text Retrieval Conference (TREC)*, 2004.

[34] J. M. Roth. Serving up EAD: an exploratory study on the deployment and utilization of encoded archival description finding aids. *American Archivist*, 64:214–237, 2001.

[35] T. Saracevic. Information science. *Journal of the American Society for Information Science*, 50:1051–1063, 1999.

[36] T. Schellenberg. The appraisal of modern public records. In M. F. Daniels and T. Walch, editors, *Modern Archives Reader: Basic Readings on Archival Theory and Practice*, pages 57–70. National Archives and Records Service, 1984.

[37] B. Sigurbjörnsson, J. Kamps, and M. de Rijke. An element-based approach to XML retrieval. In N. Fuhr, S. Malik, and M. Lalmas, editors, *INEX 2003 Workshop Proceedings*, pages 19–26, 2004.

[38] B. Sigurbjörnsson, J. Kamps, and M. de Rijke. Processing content-oriented XPath queries. In *Proceedings of the Thirteenth ACM Conference on Information and Knowledge Management (CIKM 2004)*, pages 371–380. ACM Press, New York NY, USA, 2004.

[39] B. Sigurbjörnsson, J. Kamps, and M. de Rijke. Mixture models, overlap and structural hints in XML element retrieval. In N. Fuhr, M. Lalmas, S. Malik, and Z. Szlávik, editors, *Advances in XML Information Retrieval. Third Workshop of the INitiative for the Evaluation of XML Retrieval, INEX 2004*, volume 3493 of *Lecture Notes in Computer Science*, pages 196–210. Springer Verlag, Heidelberg, 2005.

[40] B. Sigurbjörnsson, J. Kamps, and M. de Rijke. University of Amsterdam at INEX 2005: Adhoc track. In N. Fuhr, M. Lalmas, S. Malik, and G. Kazai, editors, *INEX 2005 Workshop Pre-Proceedings*, pages 84–94, 2005.

[41] K. Sparck Jones and P. Willett, editors. *Readings in Information Retrieval*. Morgan Kaufmann, San Francisco, CA, 1997.

[42] J. Tague-Sutcliffe. The pragmatics of information retrieval experimentation, revisited. *Information Processing and Management*, 28:467–490, 1992.

[43] J. Tatum. EAD: Obstacles to implementation, opportunities for understanding. *Archival Issues*, 23:155–169, 1998.

[44] A. Tombros, B. Larsen, and S. Malik. The interactive track at INEX 2004. In N. Fuhr, M. Lalmas, S. Malik, and Z. Szlávik, editors, *Advances in XML Information Retrieval. Third Workshop of the INitiative for the Evaluation of XML Retrieval, INEX 2004*, volume 3493 of *Lecture Notes in Computer Science*, pages 410–423. Springer Verlag, Heidelberg, 2005.

[45] A. Trotman. Wanted: Element retrieval users. In A. Trotman, M. Lalmas, and N. Fuhr, editors, *Proceedings of the INEX 2005 Workshop on Element Retrieval Methodology*, pages 63–69. University of Otago, Dunedin New Zealand, 2005.

[46] A. Trotman and M. Lalmas. Report on the INEX 2005 workshop on element retrieval. *SIGIR Forum*, 39(2), 2005.

[47] E. M. Voorhees. Overview of the TREC 2001 question answering track. In E. M. Voorhees and D. K. Harman, editors, *The Tenth Text REtrieval Conference (TREC 2001)*, pages 42–51. National Institute for Standards and Technology. NIST Special Publication 500-250, 2002.

[48] E. Yakel. Listening to users. *Archival Issues: Journal of the Midwest Archives Conference*, 26:111–127, 2002.

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

[50] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 334–342, 2001.