Multiple-collection Searching Using Metadata (MuSeUM)

University of Amsterdam
Gemeentemuseum Den Haag

NWO/Catch Program 2005

1 Summary of Research Proposal

This project addresses the prototypical problem of a cultural heritage institute with the ambition to disclose all of its content in a single, unified system. The institute has various legacy systems, each dealing with a small part of the collection, each constructed for different purposes, in different times, by different people, working in different traditions, based on different design principles, with different access methods, etcetera. In short, the cultural heritage institute is confronted with its *own* history. The proposed project will investigate theoretically transparent ways of combining modern information retrieval methods based on statistical language modeling with varying amounts of metadata and non-content features. Our approach to metadata is, in essence, the famous *dumb-down* principle: although metadata is based on a specific thesaurus or ontology, we can alway fall back on the description of the terms in ordinary language. In this way, we can directly employ the powerful methods of textual information retrieval. Concretely, we will address the following research questions: 1) What is the effectiveness of information retrieval techniques on a collection with varying degrees of metadata. 2) What is the retrieval effectiveness for various user types and task types? 3) What is the relative impact of techniques dealing with structure? 4) What is the relative impact of techniques dealing with multilingual content, metadata and information needs?

2 Description of the Proposed Research

2.a) Scientific Aspects

Motivation This project addresses the prototypical problem of a cultural heritage institute with the ambition to disclose all of its content in a single, unified system. The institute has various legacy systems, each dealing with a small part of the collection, each constructed for different purposes, in different times, by

different people, working in different traditions, based on different design principles, with different access methods, etcetera. In short, the cultural heritage institute is confronted with its *own* history. This problem keeps many cultural heritage institutes in a double bind: On the one hand, the various finding aids have to be converted into a single, uniform set of access points, typically requiring expensive manual or supervised assignment of a common set of metadata, and having limited indexing depth. On the other hand, optimal disclosure of each of the diverse subcollections would require specific access methods tailored to the particular content of the subcollection at hand.

The *Gemeentemuseum* in The Hague is an exemplary show-case since it covers all the major traditions in describing cultural heritage information. First and foremost, it is a *museum*, with over 100,000 detailed descriptions of museum objects. Second, it is a substantial *library*, with over 250,000 bibliographic descriptions (such as books, articles, multi-media objects). Thirdly, it is a huge *archive*, with some 500,000 process-related descriptions of activities involving museum objects (such as the acquisition, presentation, storage, preservation, loan, or use in expositions). The combined content of these three pillars of the *Gemeentemuseum* is tied together by the *Kroniek* system [4]. The *Kroniek* is based on the extensive exposition documentation that has been centrally registered since the 1970s. In its current version, the *Kroniek* is a shallow layer of common metadata using unqualified Dublin Core elements [3].

The direct motivation for the project comes from experiences with the current *Kroniek* system. The *Kroniek* contains a wealth of information that is of direct relevance to museum employees as well naive and expert users outside the museum. This potential of the *Kroniek* system has not been realized: currently, a search can only be conducted with the assistance of experienced search intermediaries. The system has evolved during its long history, and numerous changes occurred in the way cultural heritage descriptions are registered. No retrospective updates of existing descriptions have occurred due to limited availability of resources. The resulting system appears uniform and well-organized at first glance, but has many subtle but crucial incoherencies, incongruencies, and inconsistencies when subjected to greater scrutiny. As a consequence, the system requires substantial search experience and familiarity with its various peculiarities and exceptions. The differences within the system pose particular difficulties for complex queries that combine results from different parts of the system. There is a desperate need to provide uniform access to the information in the *Kroniek* without requiring intimate knowledge of system peculiarities, and of the used controlled vocabularies. The search and retrieval strategies should also be robust against the noisy nature of the existing descriptions. Yet at the same time, we should retain the specific advantages of expert searches exploiting specialized controlled vocabularies. Examples of thesauri and classifications used in the *Kroniek* are Art & Architecture Thesaurus (AAT), Hornbostel Sachs classification of musical instruments, Iconclass classification, RKDartists, and the Getty Thesaurus of Geographical Names (TGN).

Scientific Problem The crucial issue is to investigate the possibility of providing common access points for the whole cultural heritage collection, *without* sacrificing the in-depth disclosure badly needed by experts searching in a particular subcollection. Our approach to metadata is, in essence, the famous dumb-down principle [27]: although metadata is based on a specific thesaurus or ontology, we can alway fall back on the description of the terms in ordinary language. In this way, we can directly employ the powerful methods of textual information retrieval.

The *research problem* of this proposal is to investigate theoretically transparent ways of combining modern information retrieval methods based on statistical language modeling with varying amounts of metadata and non-content features. Concretely, we will address the following research questions:

1. What is the effectiveness of information retrieval techniques on a collection with varying degrees of metadata. That is: What if we ignore all metadata? What if we use only the heterogeneous metadata of the original subcollections? What if we use only the common metadata? What if we use all available metadata?

2. What is the retrieval effectiveness for various user types and task types?

3. What is the relative impact of techniques dealing with structure?

4. What is the relative impact of techniques dealing with multilingual content, metadata and information needs?

Research Method We will frame our research issues as an *information retrieval* problem: a user wants to access cultural heritage content for some reason—she has an information need—and the system should provide her with the digital objects that are relevant for her information need, regardless of how she expresses herself. Our research methodology is based on three main principles:

User-centric We want to identify real users, with varying degrees of domain knowledge, and their natural information needs. We will investigate different user types, ranging from *naive users*, who generally prefer a straightforward natural language interface, to *expert users*, who generally prefer an advanced interface allowing the search of restricted parts of the collection, or specific access-points such as creator, subject, etc. We plan to deal with a rich set of user profiles, ranging from internal usage by employees to external visitors, and a variety of task profiles, ranging from casual visits to the museum's web site to the educational dissemination of cultural heritage content.

Evaluation We will set up a proper evaluation test-suite for a Cultural Heritage Retrieval System. This will be based on a static *Kroniek* snapshot taken

at the start of the project, containing over 650,000 description records as well as the associated digital content. We are interested in users that want to access cultural heritage information *regardless* of whether it is part of a museum, library, or archive collection. Ideally, we need three of these different collections that deal with related content. The combined systems of the *Kroniek* provide us with exactly that. 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 retrieval effectiveness of cultural heritage finding aids.

Collection-independent We will focus exclusively on methods that can (1) deal with *large* data volumes (scalable, automatic), (2) deal with data that is *heterogeneous* both in content and assigned metadata, (3) deal with *multilingual* content, (4) exploit *metadata* and *collection/document structure* if available, and (5) are based on open source and open standards. That is, the system should make very few assumptions on the presence of particular metadata fields, or on the content of those fields, but should exploit them if available. Collection managers should be able to freely add metadata, or relations and links between documents. This approach will ensure that our findings transcend the particular situation of the *Kroniek*, and can be employed—in whole or in part—by other cultural heritage institutes.

Against this methodological background, we will employ a wide range of techniques within the statistical language modeling framework, where we build a variety of document representations, apply mixture language models that combine the evidence from the various sources available, and experiment with the prior probability of retrieval for various non-content features of documents. Specifically, we will adopt the following strategies in our research and development:

- First, based on our extensive experience in exploiting metadata for retrieval [1, 7, 12, 26], we will build particular document representations based on the assigned metadata. Here, we can take semantic relations between metadata terms into account, either based on thesauri (which exist only for the original subcollections) or on the usage of terms in the collection itself (which can be applied to heterogeneous metadata [7]).

- Second, by simply viewing documents with metadata as semi-structured documents, or documents with particular fields or mark-up, we can directly employ methods from semi-structured retrieval [5]. We are a leading group in XML retrieval [9, 10, 13, 14, 22, 23]. Much of the functionality and control desired by advanced users can be catered for by using particular features of query languages such as XPath. Note that these techniques carry over to currently proposed semantic web languages, which are all based on XML.

- Third, of particular interest are the relations and links between digital objects. These can be exploited in a variety of ways, either to provide additional retrieval cues, or to derive non-content features of documents. An example of the first is to locate a non-textual object such as an image based on the text of documents linking to it. An example of the latter is to use the links and relations as an indication of the importance of a digital object. These techniques are common in web retrieval [15, 17, 20].

- Fourth, we will view the semantic interoperability problem as a translation problem between different languages, and apply translation and result combination methods well-known in cross-lingual information retrieval [6, 8, 11, 16, 18, 19].

2.b) Innovation

Scientific Significance The relative effectiveness of retrieval using controlled language or free text is one of the long standing debates in information retrieval. This debate can be traced back to the nineteenth century, when rules for title term indexing in classed catalogs were introduced. It has been high on the information retrieval agenda, ever since the first large scale experimental evaluations of retrieval effectiveness of various indexing languages [2]. The majority of evidence shows that automatic indexing based on the document's own text can be at least as effective as relying on manually assigned terms from a controlled vocabulary [24]. Since both approaches to indexing have different strengths and weaknesses, there are still many open questions on the disclosure of documents using metadata [25]. Recent studies [7, 21] have shown that free-text and metadata can fruitfully be combined, leading to methods that "get the best of both worlds." The current project will contribute to our understanding of the effectiveness of metadata in information retrieval, by going beyond bibliographic descriptions, by looking at heterogeneous metadata, and by explicitly considering various user types and task scenarios.

Related Research An encyclopedic overview of related research is beyond the scope of the proposal. We focus here on the potential role that the proposed project could play within the overall CATCH program. Of particular interest is the relation with other projects in the "semantic interoperability through metadata" theme.

CHOICE The proposed research naturally complements the CHOICE project, by our primary focus on textual documents, and on retrieval based on textual search requests. Specific techniques for multimedia objects, as developed in CHOICE, are a valuable contribution to the experimental environment as envisaged in our proposal.

STITCH The proposed research naturally complements the STITCH project, by our focus on the integration of metadata and the textual content of documents. Specific techniques for semantically linking various controlled vocabularies could be naturally incorporated in the retrieval models used in our proposed project, along the lines suggested in [26].

2.c) Relevance for Cultural Heritage

Role of the Cultural Heritage Partner The cultural heritage partner will be responsible for framing the *problems*. First, by providing the combined collections in the *Kroniek*. Second, by signaling user problems with current and future versions of the *Kroniek*. This includes comprehension of complex search options by naive users; lack of control and indexing depth for advanced users; overwhelming numbers of links and relations between documents; and lack of consistency, style, and logic in metadata assignments (due to evolving views over the years, and due to limited supervision and authorization). Third, by providing the user profiles and task profiles to be studied. This includes a variety of users and tasks from both within the museum, such as general information requests; reproduction rights of images of museum objects; acquisition information; and so on, as well as from outside the museum, such as visitors to the web-site, use for external researchers, or use for educational purposes.

The university partner will be responsible for providing the *solutions*. By framing the research problem of the proposed research as an *information retrieval* problem, the scientific results, tools and techniques developed during the project will be directly applicable to the situation of the cultural heritage partner. The driving force of the whole project will be the natural information needs of various types of users, in various types of real-world task scenarios.

Desirable Results The main desirable results of the proposed project are:

- A scientific, reusable experimental test-suite to evaluate the retrieval effectiveness of cultural heritage finding aids. The resulting test collection will be made available to other CATCH participants (and could function as one of the integrators).
- A detailed assessment of retrieval techniques that exploit heterogeneous metadata based on the museum, library, and archival traditions of descriptions in each of the respective subcollections.
- A detailed assessment of the relative contribution on retrieval effectiveness of various shared metadata fields, allowing for cost/benefits analysis on the manual assignment of metadata and answering the question on how the scarce resources of cultural heritage institutes can be put to use most effectively.

- A toolbox of highly tunable, directly applicable retrieval techniques, catering for a range of users, varying from *naive users* (e.g., visitors of a cultural heritage web site) to *expert users* (e.g., information professionals employed by a cultural heritage institute).

3 Literature

[1] C. Caracciolo, W. van Hage, and M. de Rijke. Topic driven access to full text documents. In *Proceedings of the 8th European Conference on Digital Libraries (ECDL 2004)*, Lecture Notes in Computer Science, pages 495–500. Springer Verlag, Heidelberg, 2004.

[2] C. W. Cleverdon. Report on the testing and analysis of an investigation into the comparative efficiency of indexing systems. Technical report, College of Aeronautics, Cranfield UK, 1962.

[3] DCMI Usage Board. Dublin core metadata element set, version 1.1: Reference description. Technical report, Dublin Core Metadata Initiative (DCMI), 2004. `http://dublincore.org/documents/dces/`.

[4] V. de Keijzer. Gemeentemuseum Den Haag: gesloten voor registratie. In J. van der Starre and J. van Meeuwen, editors, *Nieuwe Media in Musea. Deel III: van collectieregistratie tot website*. Rijksbureau voor Kunsthistorische Documentatie (RDK), Den Haag, 1998.

[5] N. Fuhr. Information retrieval in digital libraries: Dealing with structure. In *DELOS Workshop: Information Seeking, Searching and Querying in Digital Libraries*, 2000.

[6] V. Hollink, J. Kamps, C. Monz, and M. de Rijke. Monolingual document retrieval for European languages. *Information Retrieval*, 7(1):33–52, 2004.

[7] J. Kamps. Improving retrieval effectiveness by reranking documents based on controlled vocabulary. In S. McDonald and J. Tait, editors, *Advances in Information Retrieval: 26th European Conference on IR Research (ECIR 2004)*, volume 2997 of *Lecture Notes in Computer Science*, pages 283–295. Springer-Verlag, Heidelberg, 2004.

[8] J. Kamps, S. F. Adafre, and M. de Rijke. Effective translation, tokenization and combination for cross-lingual retrieval. In C. Peters, P. Clough, J. Gonzalo, G. Jones, M. Kluck, and B. Magnini, editors, *Fifth Workshop of the Cross-Language Evaluation Forum (CLEF 2004)*, Lecture Notes in Computer Science. Springer Verlag, Heidelberg, 2005.

[9] 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.

[10] J. Kamps, M. de Rijke, and B. Sigurbjörnsson. The importance of length normalization for XML retrieval. *Information Retrieval*, 2005.

[11] J. Kamps and M. de Rijke. The effectiveness of combining information retrieval strategies for European languages. In H. M. Haddad, A. Omicini, R. L. Wainwright, and L. M. Liebrock, editors, *Proceedings of the 2004 ACM Symposium on Applied Computing*, pages 1073–1077. ACM Press, 2004.

[12] J. Kamps and M. Marx. Notions of indistinguishability for semantic web languages. In I. Horrocks and J. Hendler, editors, *The Semantic Web – ISWC 2002*, volume 2342 of *Lecture Notes in Computer Science*, pages 30–38. Springer, Berlin, 2002.

[13] 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.

[14] 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.

[15] J. Kamps, G. Mishne, and M. de Rijke. Language models for searching in web corpora. In *The Thirteenth Text REtrieval Conference (TREC 2004)*. National Institute of Standards and Technology, 2005.

[16] J. Kamps, C. Monz, and M. de Rijke. Combining evidence for cross-language information retrieval. In C. Peters, M. Braschler, J. Gonzalo, and M. Kluck, editors, *Advances in Cross-Language Information Retrieval*, volume 2785 of *Lecture Notes in Computer Science*, pages 111–126. Springer, 2003.

[17] J. Kamps, C. Monz, M. de Rijke, and B. Sigurbjörnsson. Approaches to robust and web retrieval. In E. M. Voorhees and L. P. Buckland, editors, *The Twelfth Text REtrieval Conference (TREC 2003)*, pages 594–599. National Institute of Standards and Technology. NIST Special Publication 500-255, 2004.

[18] J. Kamps, C. Monz, M. de Rijke, and B. Sigurbjörnsson. Language-dependent and language-independent approaches to cross-lingual text retrieval. In C. Peters, J. Gonzalo, M. Braschler, and M. Kluck, editors, *Comparative Evaluation of Multilingual Information Access Systems, CLEF 2003*, volume 3237 of *Lecture Notes in Computer Science*, pages 152–165. Springer, 2004.

[19] C. Monz and M. de Rijke. Shallow morphological analysis in monolingual information retrieval for Dutch, German and Italian. In C. Peters, M. Braschler, J. Gonzalo, and M. Kluck, editors, *Evaluation of Cross-Language Information Retrieval Systems, CLEF 2001*, volume 2406 of *Lecture Notes in Computer Science*, pages 262–277. Springer, 2002.

[20] C. Monz, J. Kamps, and M. de Rijke. The University of Amsterdam at

TREC 2002. In E. M. Voorhees and L. P. Buckland, editors, *The Eleventh Text REtrieval Conference (TREC 2002)*, pages 603–614. National Institute of Standards and Technology. NIST Special Publication 500-251, 2002.

[21] J. Savoy. Bibliographic database access using free-text and controlled vocabulary: an evaluation. *Information Processing and Management*, 41: 873–890, 2005.

[22] 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, 2004.

[23] 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*. Springer Verlag, Heidelberg, 2005.

[24] K. Sparck Jones. *Automatic Keyword Classification for Information Retrieval*. Butterworth, London, 1971.

[25] E. Svenonius. Unanswered questions in the design of controlled vocabularies. *Journal of the American Society for Information Science*, 37:331–340, 1986.

[26] W. van Hage, M. de Rijke, and M. Marx. Information retrieval support for ontology construction and use. In S. McIlraith, D. Plexousakis, and F. van Harmelen, editors, *Proceedings 3rd International Semantic Web Conference (ISWC 2004)*, volume 3298 of *Lecture Notes in Computer Science*, pages 518–533. Springer Verlag, Heidelberg, 2004.

[27] S. Weibel. Metadata: The foundations of resource description. *D-Lib Magazine*, 1(7), 1995. `http://www.dlib.org/dlib/July95/07weibel.html`.