INTERDISCIPLINARY SCIENCE REVIEWS, Vol. 34 No. 2–3, 2009, 268–284

Information Retrieval in Cultural Heritage

Marijn Koolen

Archives and Information Studies, Faculty of Humanities, University of Amsterdam, The Netherlands

Jaap Kamps

Archives and Information Studies, Faculty of Humanities and ISLA, Faculty of Science, University of Amsterdam, The Netherlands

Vincent de Keijzer

Gemeentemuseum, The Hague, The Netherlands

This article discusses the opportunities and challenges of applying modern information retrieval techniques to the cultural heritage domain. Although the field of information retrieval is closely associated with computer science, it originally emerged from library science — also one of the main disciplines concerned with access to cultural heritage material. Hence we are, in a sense, exploring what happens if we bring these strands of research back together again. The article consists of three parts. In the first part, we explain the field of information retrieval and its multidisciplinary nature. In the second part, we discuss how and why the problem of providing access to cultural heritage can be cast naturally as an information retrieval problem. In the third and main part, we present a detailed case study of applying the modern information retrieval approach in practice within a museum.

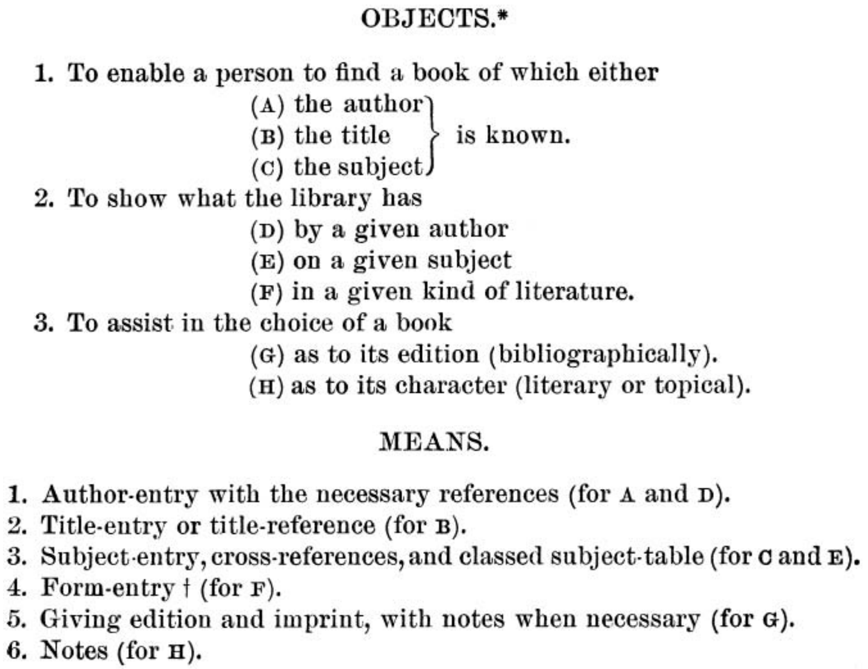
KEYWORDS Information retrieval, Cultural heritage, Gemeente museum, Library Science

Introduction

The field of information retrieval is now best known for the Internet search engines that give access to the endless amount of information on the Web, and that greatly impact our daily lives both professionally and personally. While modern information retrieval started in the 1950s, the underlying problem of bringing searchers and information sources together has been

© Institute of Materials, Minerals and Mining 2009
Published by Maney on behalf of the Institute

DOI 10.1179/174327909X441153



OBJECTS.*

1. To enable a person to find a book of which either
   - (A) the author
   - (B) the title
   - (C) the subject

   is known.
2. To show what the library has
   - (D) by a given author
   - (E) on a given subject
   - (F) in a given kind of literature.
3. To assist in the choice of a book
   - (G) as to its edition (bibliographically).
   - (H) as to its character (literary or topical).

MEANS.

1. Author-entry with the necessary references (for A and D).
2. Title-entry or title-reference (for B).
3. Subject-entry, cross-references, and classed subject-table (for C and E).
4. Form-entry † (for F).
5. Giving edition and imprint, with notes when necessary (for G).
6. Notes (for H).

FIGURE 1 Charles Ammi Cutter's famous objectives of a library catalogue (1876).

studied for centuries in library science. Ever since book collections grew to considerable sizes, librarians have organized collections of books on their subject, and compiled indexes and catalogues based on inverted lists of titles and subjects. Figure 1 shows the objectives of a library catalogue as explicitly formulated by Cutter (1876, 10). These objectives clearly exhibit the user-centred point of view so typical for modern information retrieval.

The prehistory of modern information retrieval puts the field in an interesting relation to cultural heritage — broadly speaking, the material considered worth preserving by past or current generations — available in libraries, archives, and museums around the world. On the one hand, the traditional cataloguing and indexing that led to the field of modern information retrieval is still dominating the cultural heritage sector. Library catalogues, archival finding aids, and museum registers are still based on traditional descriptions generated by cataloguers and indexers. On the other hand, advances in modern information retrieval have led to highly accurate retrieval methods that work well on any type of document representation. That is, they do not require or assume a particular cataloguing method, but work on any document surrogate (either using controlled vocabularies, document free text, or even full text if the source document is digitally available).

This suggests employing modern information retrieval methods on currently available descriptions of cultural heritage material — to bring the fruits of information retrieval back to its founding discipline, as it were. The precise outcome is not clear since there are still many open questions on the disclosure of documents using metadata (Svenonius 1986). How well does this work out in practice? That is, can a modern information retrieval system provide effective access to a heterogeneous set of cultural heritage descriptions? Since traditional descriptions cater for professional searchers, is this equally effective for expert searchers and non-expert searchers? And are there ways to preserve the structure of the original descriptions, and can this help answering complex search requests?

The rest of this article is structured as follows. In *Information retrieval* section, we explain the field of information retrieval and its multidisciplinary nature. *Cultural heritage* section discusses the problem of providing access to cultural heritage, and how and why it can be cast naturally as an information retrieval problem. Then, we present a detailed case study of applying the information retrieval approach in practice within a museum. We end by drawing some conclusions on information retrieval in cultural heritage.

Information retrieval

In this section, we will explain the field of information retrieval, its history and multidisciplinary nature, and the role of document structure in information retrieval.

Definition of information retrieval

Modern information retrieval is an inherently multidisciplinary field. Its roots are in library science, most notably the work on evaluating different languages for manually indexing and searching scientific literature, but with the spread of computing, it quickly expanded to computer science, dealing with methods and data structures for automated indexing and retrieving of documents. In addition, there are definite links with linguistics and statistics (matching keywords using natural language processing techniques), and cognitive psychology (studying how and why humans interact with search systems). The prototypical problem is the retrieval of the 'right' documents in response to a user's query or topic statement. Essentially, a user wants to access information for some reason — she has an information need — and the system should give her access to the digital objects (and only those objects) that satisfy her information need, regardless of how she expresses herself. This notion of 'user satisfaction', however, is a vague, problematic one and a point of contention for information retrieval researchers. User satisfaction is usually approximated by relevance, though this too is not strictly defined. Side-stepping the complexities of the issue, we point out simply that this is in sharp contrast with the related field of databases, where the answer set of a database query is defined as all records that match the keywords in the specified fields.

Multidisciplinary nature of information retrieval

Modern information retrieval research is generally considered to have started in the 1950s (Robertson 2008), when it focused on how to evaluate search systems and on measuring the effectiveness and efficiency of automated indexing techniques for literature search systems. An important aspect of library science is subject classification, the process of assigning keywords that properly describe what a literary work is about. By indexing these keywords, one can easily look up which documents cover a certain topic. Cyril Cleverdon, a librarian at the Cranfield College of Aeronautics at the time, conducted experiments to compare four different classification schemes, representing four opposing views of how to organize information (Cleverdon 1962, 1967). Experts in each scheme were asked to index documents using their indexing system and then do their own searching. These Cranfield experiments were conducted without any aid of computers, relying on card indexes and human searchers.

While the results that were the original object of these experiments may have lost their relevance, the important, long-lasting outcome was the shaping of an information retrieval evaluation methodology, which is still the dominant form of evaluation today. First, a number of search requests (representing the information needs of users) were formulated, after which the experimenters, for each request, selected documents to be judged as relevant or irrelevant by the user with the information need. Once the documents relevant to a query were identified, each classification scheme could be evaluated by counting how many of the relevant documents it retrieved (the notion of *recall*) and how many of the retrieved documents were in fact relevant (the notion of *precision*). This method relies on three simplifying assumptions:

1. Relevance is approximated by topic similarity, i.e., if a document is about the same topic as a search request, it is relevant and, therefore, all relevant documents are equally desirable, the relevance of one document is independent of the relevance of any other document, the user's information need is static
2. A single set of judgements is representative for the user population
3. The list of relevant documents is complete, that is, all relevant documents are known.

The introduction of the computer into the information retrieval process, in the late 1950s (Robertson 1994), also introduced another academic field into information retrieval research: computer science. One of the conclusions of the experiments at Cranfield was that indexing and searching documents using natural language worked better than using some formal, restricted indexing language. This facilitated the transition from manual indexing and searching to automatic methods of indexing and searching: while it is hard for computers to interpret the text of the document and choose terms from the indexing vocabulary, it is relatively straightforward to segment the text into words to use as index terms, as the words in the text inherently encapsulate the subject matter of the document. Another product of computerized search is the use of a *scoring function* to measure how well a document matches the query entered by the user, then order the documents by their score.

However, automatic indexing involves more than segmenting the text into words. Manual indexing relied on elaborate controlled vocabularies to try to control different ways of expression. The purpose of indexing is to create representations of the documents against which the user queries can be matched, regardless of surface differences between the words in the queries and the words in the documents. A document containing the word *canine* may well be relevant to a query with the word *dog*, as might documents containing the words *dogs*, *doggy* or *labrador* (but perhaps not documents that use the verb *to dog*). Ideally, some linguistic manipulation should be carried out to deal with synonymy (canine), morphology (dogs), hypernymy/hyponymy (labrador) and homonymy (to dog). Grefenstette and Tapanainen (1994) give a detailed discussion on linguistic processing. Moreover, should compound terms, e.g., *long stay car park*, be indexed as a whole or as their individual components? Clearly, linguistics is another field that plays an important role in information retrieval.

Using the evaluation methodology described above, any retrieval method could be evaluated using the same set of documents, search requests and relevance judgements, in a laboratory setting, ignoring all user related aspects such as interfaces and interaction. In the 1970s, a number of researchers criticized this system-oriented evaluation, and promoted a more user-oriented approach (e.g. Belkin 1980). This starts from the fact that the system is situated in the real world, with users having different background knowledge, different interpretations of what is relevant, information needs that gradually change by interacting with the system, different ways of expressing those information needs and different levels of affinity with retrieval systems. Currently, information retrieval research is typically divided into these two system-centred and user-centred perspectives, with the vast majority of researchers focusing on the former, where a clear evaluation methodology and reusable test collections make it easier to conduct experiments.

Retrieval on structured documents

What sets information retrieval apart from the neighbouring field of database retrieval are the notions of *ranking* and *best match*. In contrast with database systems that consider all documents matching the query equally relevant, an information retrieval system has to determine to what extent a document matches the query. If not all terms of the query occur in a document, the document might still be considered useful by the searcher. The task of information retrieval is to order the documents by relevance to the user's information need, regardless of how it is expressed in the query.

Another related difference is that databases use data in a structured way, whereas information retrieval largely ignores structure, thereby keeping the system generic enough to deal with any kind of document representation, regardless of the specific structure.

However, this is changing due to the recently increasing attention to XML retrieval. XML stands for eXtensible Markup Language, which is a markup language similar, but more general than, HTML, which is perhaps the most well-known markup language, widely used in the World Wide Web. XML is often used to explicitly structure text. Research in XML retrieval aims to investigate how retrieval systems can exploit the structure inherent in many document collections. This is especially done at the INEX forum (INEX 2009).

These new retrieval methods hold a great potential to couple the flexibility of retrieval systems with the expressiveness of database query languages.

Wrap Up

In this section, we have given a short introduction to the field of information retrieval, its history and multidisciplinary nature, and the role of document structure in information retrieval. When viewed as part of computer science, information retrieval has an unusual human-centred focus: there are no *a priori* correct answers, and human judgements on the usefulness of a result are ultimately decisive. There is hence a strong emphasis on evaluation and experimental research. Research in information retrieval traditionally ignored document structure — to ensure that results apply to any text — but with the wide availability of structured documents in generic formats like HTML and XML, structured retrieval is now recognized as an important research topic. In the next section, we will discuss information access in the cultural heritage domain.

Cultural heritage

In this section, we will discuss cultural heritage, and how cultural heritage material is described traditionally in memory institutions like libraries, archives, and museums. Then we detail the case of the Gemeentemuseum, The Hague, and discuss similarities and differences between non-expert and expert users of cultural heritage data.

Definition of cultural heritage

The Continuous Access To Cultural Heritage (CATCH) research programme in the Netherlands addresses the problem of improving access to digital cultural heritage — broadly speaking the material considered worth preserving by past or current generations. It can be defined, according to the Wikipedia, in the following way (Wikipedia 2009):

> Cultural heritage ('national heritage' or just 'heritage') is the legacy of physical artefacts and intangible attributes of a group or society that are inherited from past generations, maintained in the present and bestowed for the benefit of future generations. Often though, what is considered cultural heritage by one generation may be rejected by the next generation, only to be revived by a succeeding generation.

Cultural heritage material is usually curated by our memory institutions, i.e. our libraries, archives, and museums. From the quotation, it is clear that cultural heritage encompasses a vast range of different phenomena, ranging from fine arts to archival records.

Describing cultural heritage material

For many centuries, cultural heritage institutions have spent their efforts on collecting and describing artefacts and social phenomena to preserve and give access to our cultural heritage, and have dealt with problems of information storage and retrieval since their beginnings.

Preserving and giving access to cultural heritage is done through collecting information about cultural heritage objects (or archaeological sites or people's lives), stored and organized in information systems like library catalogues, archival finding aids and museum registers. These systems do not give direct access to the phenomena themselves. Whereas Internet search engines can give access to Web pages directly by providing hyperlinks to pages that contain words in the query, the objects in cultural heritage collections are not directly accessible, so that information systems have to deal with (textual) object representations, often in the form of object records. One of the key activities of cultural heritage institutions is thus to make detailed descriptions of their objects in library catalogues, archival finding aids, and museum registers. Without such descriptions, organized in some systematic way, an object is almost completely inaccessible.

The classical approach to this problem depends on precise and consistent descriptions, made by experienced scholars, where all the information has a dedicated place in a single, fully defined and structured classification system or ontology. Such descriptions were originally published as printed lists or books, which evolved into card catalogues that allowed relative ease of updating them continuously. When computers became available, the distribution and printing of card records was greatly facilitated by storing the descriptions in a digital format, in so-called MARC (Machine Readable Cataloguing) records. Digital catalogues were created by storing these records in a database system, leading to a wide availability of public access catalogues. Even today, the majority of descriptions and systems in the cultural heritage sector relies on a database framework, therefore on consistency, correctness, and completeness. Legacy systems require experience with and knowledge of evolution of the system and data structures.

In theory, these descriptions are a show-case of rigorousness and consistency. There are multi-volume books on how to describe books; ontologies, taxonomies, thesauri and other controlled vocabularies ensure consistency in terminology and deal with synonymy and homonymy; authority files are used to disambiguate between items with similar names or titles and to group multiple versions of a given work under a uniform title. In practice, upon scrutinizing actual collection records, this idea of consistency and rigorousness seems difficult to uphold. There are several reasons for lack of consistency. First, there are various standards within museums, libraries, and archives for describing their collections, so object collections from two museums, or even two sub-collections within the same museum, might be described using different standards. Second, when new objects are described, there is only a very limited budget for editorial support to ensure this is done consistently and without typing and spelling errors. Third, ideas and principles of describing change over time, with cultural heritage institutions having little to no budget to update old material. Fourth, there is no single correct way of assigning keywords to items. If you ask two persons to assign keywords to the same item, they will come up with different and equally applicable keywords. In fact, if you ask the same person at different times to describe the same item, you might end up with different keywords.

Case of the Gemeentemuseum

The Multiple-collection Searching Using Metadata (MuSeUM) project casts the CATCH problem as an information retrieval problem. Using the large digital collection of the Gemeentemuseum, The Hague (Gemeentemuseum 2009), which combines substantial collections from all three major traditions in cultural heritage — museum, library and archive — we investigate ways to give unified access to all this information through a single information retrieval system.

The Gemeentemuseum is most famous for its large collection of Mondrian paintings, but also has large collections of modern art, prints and posters, ceramics, fashion, and musical instruments. This museum is an excellent case study, since its combined descriptions cover all three traditions of cultural heritage:

1. First and foremost, it is a *museum*, with well over 100,000 detailed descriptions of museum objects
2. But it also houses a substantial *library*, with over 250,000 bibliographic descriptions for books, articles, multimedia objects, and so on, typically related to the works of art in the museum
3. Moreover, it is also an *archive*, with almost 750,000 process-related descriptions of activities involving museum objects such as the acquisition, presentation, storage, preservation, loan, or use in expositions.

Although these collections contain a lot of interesting information, the disclosure of them for the public is not the main focus of the museum.

First and for all, the Gemeentemuseum is an institution that organizes exhibitions. In these presentations, a selection of the museum collection is shown with objects from other museums. A lot of effort is put into presenting the objects in the best possible way, often accompanied by a catalogue of the exhibition, providing rich contextual information to show the link between objects and the theme of the exhibition. This main focus on exhibitions will not change overnight, especially since the subsidiaries mainly judge the success of the museum on the amount of visitors that literally come to the building.

Like other museums, the Gemeentemuseum comes from a history of organization by sub-collections. The whole cake, as it were, was divided into parts (paintings, works on paper, fashion, applied arts, musical instruments, etc.). Every part had its own staff with the curator at the top. He or she was responsible for everything related to this specific part of the collection: storage, loans, acquisitions, exhibitions, restoration, etc.

Currently different descriptions are stored in different databases, and the heterogeneous nature of these data makes it hard to envisage a single huge database that could be used to easily search through all these different sub-collections (Koolen *et al.* 2007). The lack of uniformity in structure and terminology in metadata calls for a more robust approach that can give access to information regardless of structure and consistency.

Expert and non-expert searchers

Do museum curators, who have a high level of expertise in their domain and knowledge of the collection, have similar information needs to non-expert users, and do they express these needs in similar ways? To answer this question, we asked a group of four non-experts and a group of seven curators of the Gemeentemuseum to provide search requests. This resulted in 40 topics created by the non-experts and 44 topics created by the museum curators.

Analysing the differences between these groups of search requests will give us an indication whether the current system will be able to cope with both non-expert and expert users, and provide pointers to possible approaches dealing with the more expert search requests.

We analysed the sets of non-expert and expert search requests. We classified the topics according to the kinds of aspects that are used as "conditions" that the objects have to satisfy. That is, topics asking about works from a certain creator are classified as containing a *creator* condition. Table 1 shows the results of this analysis. The two most striking differences are that 1) the experts use far more varied aspects of the objects, and 2) non-experts mainly use terms describing what the object depicts — i.e. all objects depicting trees — whereas the curators use a much broader range of aspects in their requests.

Table 2 shows a breakdown of the topics over the number of conditions contained in them. The non-expert requests contain mostly one or two aspects

TABLE 1

DISTRIBUTION OF NON-EXPERT AND EXPERT TOPICS OVER CLASSES OF CONDITIONS

Queries / aspect	**Non-expert queries**		**Expert queries**	
	#	%	#	%
acquisition	0	0	5	11.4
condition	1	2.5	1	2.3
creator	7	17.5	16	36.4
depiction	33	82.5	4	9.1
loan	1	2.5	3	6.8
location	1	2.5	9	20.5
material	4	10	5	11.4
period	1	2.5	8	18.2
style	1	2.5	5	11.4
technique	1	2.5	0	0
title	0	0	1	2.3
type	13	32.5	24	54.5
other	0	0	3	6.8

TABLE 2

DISTRIBUTION OF NON-EXPERT AND EXPERT TOPICS OVER NUMBER OF CONDITIONS

Aspects / query	Non-expert queries		Expert queries	
	#	%	#	%
1	17	42.5	14	31.8
2	22	55	16	36.4
3	1	2.5	13	29.5
4	0	0	1	2.3

— object type and what is depicted — whereas a substantial percentage of the expert topics contain three or more aspects. This signals that expert searchers have more complex information needs, and have a good understanding of the types of information that are available in the descriptions. Under these circumstances, preserving the original record structure may benefit the expert searchers, allowing them to articulate their information needs better by including references to the fields of the record structure.

Wrap Up

In this section, we discussed traditional descriptions of cultural heritage material in memory institutions like libraries, archives, and museums. These descriptions are the work of human cataloguers and indexers, and rely on control and consistency by using strict formats, with strict rules, and elaborate controlled vocabularies. These high standards are difficult to live up to in practice — if only due to how the documentation standards have changed over time. Different types of documentation — even specific sub-collections — have their own standards and systems, making it very hard to search across the entire collection. We looked in detail at the case of the Gemeentemuseum, The Hague, whose documentation covers all three traditions in cultural heritage: object descriptions in the museum's register, bibliographic descriptions in the library catalogue, and process-related descriptions in archival finding aids. We analysed similarities and differences between non-expert/expert users of cultural heritage information, and found that experts have more complex information needs: they use a much wider range of aspects, and more frequently combine different aspects. The above suggests that a text-retrieval approach — that does not rely on particular structure or vocabularies, but potentially can profit from them if available — can be an attractive alternative to currently used systems. We now turn to a case study of applying an information retrieval framework to the heterogeneous collections of documents in a large museum.

Information retrieval in cultural heritage

This section consists of four parts. First, we will describe our approach to give unified access via a simple information retrieval system to the combined digital collections of the Gemeentemuseum. Second, we will evaluate our approach by comparing the performance of a single retrieval system providing unified access against a more traditional approach of access through multiple legacy database systems. Third, we take a closer look at how an information retrieval system can deal with the complex queries posed by expert users and exploit the available structure in cultural heritage data.

Unified access

The database management system of the Gemeentemuseum is called *Kroniek* (in English: Chronicle) and consists of separate modules specifically designed for museum, library and archival descriptions. Each module allows the users to search on specific fields in the descriptions. To be able to index and access the descriptions through one system, the descriptions were exported from their respective modules as textual XML documents. This allows us to maintain their structure and make them readable for other systems as well.

For retrieval, we use Lucene (2009), a general purpose search engine, to index the entire collection, because it is a widely available and often used system. The standard Lucene uses a vector space model for indexing and retrieval (Salton and McGill 1983), and has a simple keyword-based query language. A home-grown extension to Lucene allows the use of another ranking model based on statistical language models (Hiemstra 2001), which we will use in our experiments below. The resulting system, called *CatchUp*, is a primitive first version of a unified system. By ignoring all structure and simply treating description records as plain text documents, *CatchUp* gives all users, internal or external, expert or non-expert, easy access to the full digital cultural heritage content of the Gemeentemuseum.

Comparing CatchUp and traditional approaches

To be able to compare the performance of unified access through an information retrieval system and traditional access through legacy database systems, we use a test-collection containing search requests targeting all parts of the museum data. The expert systems at the Gemeentemuseum have been specifically designed to retrieve highly relevant information. The database oriented approach of fielded-search often leads to high precision. How does our general purpose retrieval engine compare to these expert systems?

Some form of evaluation is required to be able to judge if simple, unified access is indeed a step forward. If retrieval performance with *CatchUp* is significantly worse than with the expert systems, perhaps this kind of unified access is not suitable for disclosing the particular heterogeneous collections of the Gemeentemuseum. But how can we compare the retrieval effectiveness of a full-text retrieval system with that of multiple legacy systems? First of all, we need a task which is natural for both types of system, and second, we need a method to measure how well both systems perform.

As a natural task, we used known-item retrieval, i.e., the user is looking for a specific document, which is known to be in the collection. The employees in the museum use the *Kroniek* for such a task on a daily basis. This wouldn't change if they were to switch to a full-text retrieval system, showing that the task makes sense for both types of system. Using the evaluation methodology, we have constructed 49 known-item topics based on descriptions from all three modules of *Kroniek*. That is, for each of the 49 descriptions from the collection, we created a query aimed at retrieving that specific description. Among the 49 topics, there are 10 topics for documents of the archive module, 16 for documents of the library module and 23 for documents of the museum module. We assume perfect knowledge of the appropriate module for the other topics. Thus, topics based on archival descriptions are only used on the archive module, etc. For each module, we have searched using the most important — according to the museum experts — field. For the library module, we have entered the query in the *title* field, the *description* field for the museum module, and the *title+description* field for the archive module.

What we want evaluate for known-item retrieval is how well both systems perform on retrieving and ranking the requested document. Assuming that users read the list with returned results from the top down, whether they are ranked by relevance or simply by the order in which they are found in a database, we can express the effort needed to find the known item in this list by a number between zero and one. A score of one reflects the least effort, i.e., the known-item is returned as the first result in the list. A score of zero reflects an unsuccessful effort, where the requested document is not returned by the system and the user thus wastes all effort on locating it in the results list. For all 49 topics, we can compute the reciprocal ranks as 1 divided by the rank of the known item in the results list. Thus, if the requested document is returned as the seventh result in the list, the reciprocal rank of that topic is $\frac{1}{7}$. The Mean Reciprocal Rank (MRR) is the average of the reciprocal ranks over all 49 topics. The results are shown in Table 3.

First, we will discuss the results for *Kroniek*. We see that *Kroniek* performs better on the library module than on the other two modules. One reason for this could be that the library records of the Gemeentemuseum are very short and contain most of the information in only one field, the *title* field. The archive descriptions tend to be short as well, but are much larger in number. Also, there are many archival descriptions related to the same topic. For an exhibition, there are often archival records describing loan requests, correspondence, press coverage and the opening of the exhibition. With

TABLE 3

MEAN RECIPROCAL RANK FOR 66 KNOWN-ITEM TOPICS

	# **queries**	**Kroniek (baseline)**	**CatchUp**
1. Museum	23	0.1560	0.5389
2. Library	16	0.5938	0.6719
3. Archive	10	0.2000	0.3625
1+2+3	49	0.3079	0.5463

multiple records related to a single topic, there are more results to return for a query on this topic, and thus a good chance that the user has to spend more time locating the record they are looking for.

For *CatchUp*, we see something similar. It performs best on the library documents. Given that *CatchUp* searches through all three collections, this difference in performance cannot be caused by a smaller number of records searched for those topics. What can be an explanation is that the multiple archival descriptions on the same topic make it harder to find one specific record. With *CatchUp*, the difference between library and museum topics is much smaller than with *Kroniek*. One explanation might be that museum descriptions generally contain much more text, making it easier for the system to rank them according to their similarity to the query.

If we compare the two systems, we see that *CatchUp* easily outperforms *Kroniek* on all three topic sets. This shows that the framework of ignoring structure and searching for query terms in the whole descriptions, and subsequently ranking the descriptions according to their similarity to the query, is a viable and effective approach to provide access to cultural heritage data.

There are, however, some limitations to this form of evaluation. First, by using known-item retrieval, where we measure how well a system is able to find a single specific document, we get no indication of how well a system performs more general information retrieval tasks where there are many documents relevant to a query. To evaluate this, we can use the topics discussed in the *expert and non-expert searchers* section, which require the system to return lists of relevant objects. We are in the process of collecting relevance judgements for the 44 expert topics. More seriously, we have seen in the previous section that expert users tend to have more complex information needs that might require the system to make use of the available structure of cultural heritage data. Although the *CatchUp* system seems to perform well for the reasonable simple, non-expert known-item topics, the choice of ignoring the field structure of the description records may not be the best strategy for more complex search requests typical for expert users. Therefore, we next discuss how the standard text retrieval approach might be extended to take both the structure of the data and the structure of complex queries into account.

Complex requests and structured queries

Many requests, either explicitly or implicitly contain structural constraints. For the query *paintings Mondrian*, the system should probably not simply return any description containing the word *painting*, but only the descriptions of objects that actually *are* paintings. With fielded records, it is easy to add this restriction to a query. Can we incorporate the notion of queries with structural constraints in an information retrieval system as well?

Indexing the records with structural information is relatively straightforward. Most current retrieval engine support fielded data, allowing searchers to restrict their attention to results that match the query in particular fields.

For example, a Lucene index always uses fields but stores the complete content in the default ‘text’ field. With a modified preprocessor, all content can be indexed in the original record’s field. To illustrate, the word *Mondrian* occurring in the *creator* field could be indexed preserving its field. This allows for later queries like creator:mondrian to match only these occurrences. Emerging XML retrieval engines like PFTijah/MonetDB (PF/Tijah 2009) support complex query languages like XQuery and NEXI (Trotman and Sigurbjörnsson 2005), which allow the user to submit queries with explicit structural conditions.

To show how the indexing of structure allows us to submit structured queries, we will look at two example expert topics:

- *Show me all* Art Nouveau *works of Jan Toorop between 1890 and 1905*
- *List all paintings from Jewish painters bought by the museum between 1933 and 1940 (for a national investigation of the sale of Jewish art to Dutch museums during the Nazi reign before the Second World War).*

We can express these information requests as structured NEXI queries as follows:

```
//object[./creator=‘Jan Toorop’ AND ./created.date.start >= ‘1890’ AND ./created.date.end <= ‘1905’ AND about(., ‘Art Nouveau’)]
//object[./creator[about (., ‘Jewish painters’)] AND object.type=‘painting’ AND acquisition.method=‘bought’ AND acquisition.date>=1933 AND acquisition.date<=1940]
```

The first query states that the system should return records with *creator* fields matching *Jan Toorop*, the date in the *date* field should lie between *1890* and *1905* and the record should be ‘about’ *Art Nouveau*. The second query requires the system to return objects that are *paintings* made by *Jewish painters* and are *bought* by the museum between *1933* and *1940*.

With these structured queries, users are able to specify that the system should match certain keywords in specified fields, thereby effectively disambiguating the query. The structural constraints allow for an easy filtering of objects created by a specific creator in a specific period, but the hard part is still to match the query words ‘Art Nouveau’ to objects in the Art Nouveau style.

Note that these complex queries are not easy to formulate and require substantial knowledge of the objects in the collection and of what the actual structure looks like. The curators in the museum have this knowledge and may wish to use this more complex query language to be able to use their knowledge of the collection while searching. But this group of experts represents only a small fraction of the potential users. The vast majority of users have limited to no knowledge of the collection, and plausibly will rather type much simpler queries, which *CatchUp* already handles well.

We are now building a test-collection of the expert and non-expert topics, with lists of relevant objects, to investigate the effectiveness of using the available structure for both expert and non-expert search requests. For this, we use an updated version of *CatchUp* that takes the structure of documents and queries into account.

Wrap Up

In this section, we have applied a standard text retrieval method to the digital collections of the Gemeentemuseum and compared the effectiveness of this approach with the traditional legacy database system *Kroniek* — the system currently in use in the museum — using a set of non-expert known-item search requests. Without exploiting any of the available structure, our system, *CatchUp*, outperforms *Kroniek* on this set of topics. We also looked at possible ways to deal with the complex queries of domain experts in a more structured way. Through recent developments in XML retrieval, multiple retrieval systems are now available that can index both content and structure of documents and use a more complex query language to allow users to express structural requirements. Although this allows expert users to exploit their knowledge of the data, structured queries are much harder to formulate. The majority of users have limited knowledge of the collection, and might prefer more simple interaction with the system.

Discussion and conclusions

We have shown in this paper that the problem of disclosing cultural heritage information can be naturally presented as an information retrieval problem. Information retrieval is an inherently multidisciplinary research area, addressing the problem of providing the user with documents relevant to their information need, regardless of how they express themselves. The data in cultural heritage institutions is often highly structured and organized in specific sub-collections, aiming for high degrees of rigorousness and consistency, but, in practice, may not live up to these high standards. Descriptions are often short and only provide a superficial flavour of the rich stories that could be told about many of the precious objects. Different descriptions may be inconsistent due to lack of money for editorial support, and due to the fundamental problem of subjectivity and interpretation when assigning describing objects.

The above are the typical problems studied in information retrieval research, which we address by treating all cultural heritage descriptions as essentially textual documents. As a first step, we used a simple keyword-based information retrieval system that ignores any structure in the descriptions. Although our initial evaluation showed that a standard text retrieval system can be a viable and effective way of giving simple access to digital cultural heritage collections, we also provided a comparison of non-expert and expert information needs and found the expert search requests to be more complex, which may require a less naive system that takes the available structure in cultural heritage descriptions into account.

While in this article, we detailed the interactions between the disciplines of information retrieval and cultural heritage, within the project this also required researchers from information retrieval and cultural heritage to collaborate on a day-to-day basis. The process of collaborating with the museum curators to obtain search requests and lists of objects relevant to these request has been very slow, but insightful. This slow pace is caused by a substantial gap between the information retrieval community and the daily practice of the museum community. As mentioned earlier, the digital collection of the museum was not created with the general public in mind, but rather as internal documentation. Scientific research aiming at improving access to heterogeneous information feels to the museum staff like a distant goal that has little to do with their daily work. In fact, for a lot of the (older) employees, the computer itself is something they fairly recently learned to use in their work. From the perspective of the information retrieval community, the problems of heterogeneous information pose an interesting challenge to provide intelligent information access no matter what data are available. Although the museum community is interested in having complete, correct and consistent data to work with — making their daily routine easier — they are less interested in the challenges that need to be faced in order to get there. Only considerable interaction over a long period of time, and a growing understanding of each other's worlds, made clear that the problems that the museum community encounters in organizing and accessing cultural heritage data, and the problems studied in information retrieval are indeed a close match.

Acknowledgments

This research was generously supported by the Netherlands Organisation for Scientific Research (NWO), within the Continuous Access To Cultural Heritage (CATCH) programme under project number 640.001.501.

Bibliography

Belkin, N. 1980. Anomalous states of knowledge as a basis for information retrieval. *Canadian Journal of Information Science* 5: 133–43.

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

Cleverdon, C.W. 1967. The Cranfield tests on index language devices. *Aslib* 19: 173–92.

Cutter, C.A. 1876. *Rules for a printed dictionary catalogue*. Washington: Government Printing Office.

Gemeentemuseum. 2009. Haags Gemeentemuseum. http://www.gemeentemuseum.nl/ (last checked: 05/03/2009).

Grefenstette, G. and P. Tapanainen. 1994. What is a word, what is a sentence? Problems of tokenization. In *The 3rd International Conference on Computational Lexicography*, 79–87, Budapest.

Hiemstra, D. 2001. *Using language models for information retrieval*. PhD diss., Center for Telematics and Information Technology, University of Twente.

INEX. 2009. The INitiative for Evaluation of Xml retrieval. http://www.inex.otago.ac.nz/ (last checked: 05/03/2009).

Koolen, M., A. Arampatzis, J. Kamps, V. de Keijzer, and N. Nussbaum. 2007. Unified access to heterogeneous data in cultural heritage. In *Proceedings of RIAO 2007: Large-Scale Semantic Access to Content (Text, Image, Video and Sound)*.

Lucene. 2009. The Lucene search engine. http://lucene.apache.org/ (last checked: 05/03/2009).

PF/Tijah. 2009. The INitiative for Evaluation of Xml retrieval. http://dbappl.cs.utwente.nl/pftijah/ (last checked: 05/03/2009).

Robertson, S. 1994. Computer retrieval. In *Fifty years of information progress*, ed. B. Vickery, 118–46. London: Aslib.

Robertson, S. 2008. On the history of evaluation in IR. *Journal of Information Science* 34(4): 439–56.

Salton, G. and M.J. McGill. 1983. *Introduction to modern information retrieval*. McGraw-Hill Computer Science Series. New York: McGraw-Hill.

Svenonius, E. 1986. Unanswered questions in the design of controlled vocabularies. *Journals of the American Society for Information Science* 37: 331–40.

Trotman, A. and B. Sigurbjörnsson. 2005. Narrowed Extended XPath I (NEXI). In N. Fuhr, M. Lalmas, S. Malik, and Z. Szlávik (Eds.), *Proceedings of the Initiative for the Evaluation of XML Retrieval (INEX 2004)*, Lecture Notes in Computer Science, 16–40. Heidelberg: Springer Verlag.

Wikipedia. 2009. Cultural heritage. http://en.wikipedia.org/wiki/Cultural heritage (last checked: 05/03/2009).

Notes on Contributors

Correspondence to: Jaap Kamps, ILPS-ISLA, University of Amsterdam, Science Park 107, 1098 XG Amsterdam, The Netherlands.

Email: kamps@uva.nl