Unified Access to Heterogeneous Data in Cultural Heritage

Marijn Koolen[1] **Avi Arampatzis**[1] **Jaap Kamps**[1,2]
Vincent de Keijzer[3] **Nir Nussbaum**[2]

[1]Archives and Information Studies, University of Amsterdam, The Netherlands
[2]ISLA, University of Amsterdam, The Netherlands
[3]Haags Gemeentemuseum, The Hague, The Netherlands

Abstract

This paper addresses the prototypical problem of a cultural heritage institution with the ambition to disclose all of its content in a single, unified system. Like other enterprises, these institutions have heterogeneous collections distributed over multiple legacy systems. Our approach is to turn the metadata retrieval problem into a free-text retrieval problem by an unconditional merging of the heterogeneous sub-collections and flattening of all metadata structures. We investigate the viability of the approach by an extensive case study of a large museum. Our main findings are as follows: First, by converting all digital content to text, and indexing it with a standard IR system, we can effectively build a unified system providing access to all data. Second, an initial empirical evaluation shows superior performance in comparison with the legacy systems currently in use by the institute. Therefore, our third and overall finding is that our approach is a viable option to give access to heterogeneous collections.

1 Introduction

This paper addresses the prototypical problem of an organization with the ambition to disclose all of its content in a single, unified system. All organizations have their digital content stored in one or more databases and shared file-systems, in order to be able to search, find it back, and use it on demand. But as the amount of digital content grows over time, the inherent difficulty arises of finding the right information, with respect to a task, in a collection of mostly irrelevant information. Typically there are different legacy systems, and each of the legacy systems has a propriety format tailor-made for certain types of documents, and allow users to search for specific documents within that system. These systems, although they allow quick search through all the descriptions on a specific field (e.g. title or creator), are cumbersome for end-users that do not have sufficient knowledge of the fields and vocabulary used and of the organization of the data. To make things worse, the passing of time has introduced different archiving methods, technologies, and even differences in opinions of human indexers.

In IR research, the main focus has been on documents that are more or less similar in nature: a single large collection of natural language texts of the same format[1]. So, at a first

[1]There are some notable exceptions, e.g., (Baraglia et al., 2005).

glance, traditional IR techniques cannot be straightforwardly applied on the digital collections of most institutions and companies for a number of reasons. First, there is no single collection but rather several collections, residing in several different systems and locations; there is no single system that can access everything. Second, documents are in various formats (e.g., propriety formats, plain text, MS Word, PDF, HTML), languages, and some of them may even not contain natural language but fielded descriptions, with terms selected from controlled vocabularies. In short, the total volume of stored data is *fragmented* over different database systems or file-systems, and it is *heterogeneous* in nature.

The paper proposes an IR approach to the problem of providing unified access to heterogeneous data, by simply treating all content as free text and applying the powerful methods of text retrieval. This approach raises the following two main questions:

1. Is a full-text IR solution suitable for unified access to all data and metadata?
2. If yes, how does the effectiveness of such a system compare to that of existing, separate, tailor-made expert systems?

In order to answer these questions we conducted an extensive case study based on the realistic context of an existing *cultural heritage institution* (CHI). The aforementioned problems are typical for a CHI with the ambition to disclose its digital content to external users: it is faced with several types of descriptions and types of documents (multimedia), accessible from several systems with different interfaces and different formats for storing their content. In short, cultural heritage data collections are fragmented and heterogeneous in nature, as well. Our case study uses the combined data of a large museum. As a first step, we modified a typical full-text IR system and loaded up all museum's data and metadata, flattening any existing structure (e.g., fields, links across documents, etc.) We created a test set of information requests targeting known-items in the collections, and experimented with the system. Although our case study focuses on cultural heritage, the problems to provide unified access to heterogeneous data is much broader, and our findings may also directly apply to other *enterprise search* scenarios (Hawking, 2004; Fagin et al., 2003; Hawking et al., 2002).

The rest of this paper is organized as follows. Section 2 provides background on cultural-heritage descriptions and metadata formats. Section 3 focuses on the data and current database retrieval systems of the museum used in the case study. In Section 4 we describe the baseline version of a unified system. Next, in Section 5, we conduct a comparative evaluation of the unified system and the legacy systems. Finally, conclusions are drawn in Section 6.

2 Cultural-Heritage Descriptions

Cultural heritage institutions have one or more systems to store information about the objects in their collection. Apart from these systems containing descriptions, there is often a file-system where documents are stored containing information on the day-to-day business like: questions from visitors, paperwork on acquisition, preservation and provenance of objects, press articles, etc. This is another vital part of the digital content of cultural heritage institutions, often providing context for the more bare resource descriptions.

2.1 Descriptions Preserve Resources

One aim of describing resources is to preserve them. For example, in a museum, many resources are fragile and deteriorate through handling and exposure to light. By providing information about resources, e.g. date of last restoration, descriptions reduce the need to get resources out of their often specially constructed storage.

In CHIs, detailed descriptions are made for many different resources. Books, pamphlets, paintings, sculptures, clothing, jewelry, coins, persons, events, all have their own characteristics. Books have authors and a certain number of pages, whereas coins and ballroom dresses obviously have not. But what they do share is that they have spatial dimensions and are made of a certain type of material. However, events have no width and are not made out of any material.

Clearly, there is no universal set of characteristics to capture all the different aspects of different objects and phenomena. Therefore, *metadata* schemes have been made for descriptions of a certain type, like EAD (Pitti, 1999) for archival descriptions, MARC (MARC, 2006) for bibliographic descriptions or CRM (CIDOC, 2006) for museum object descriptions. A metadata scheme for archival descriptions, has elements that cannot be used for object descriptions. But there are several different schemes in use across CHIs. One institution might create descriptions that are not in accordance with the metadata scheme used by another similar institution.

2.2 Descriptions Provide Access to Resources

Due to the lack of descriptor uniformity, systems containing descriptions are not directly interoperable (Iannella and Waugh, 1997; Dempsey and Heery, 1998). But the descriptions are meaningful and can be unique identifiers of a resource. Without their descriptions, resources will inevitably get lost in the countless depots that contain them.

An initiative to make descriptions interoperable is the Dublin Core Metadata Initiative (DCMI, 2006). Its aim is to provide a set of metadata elements usable for many different resources. However, this set only contains the most used and common fields. It only provides a shallow layer on top of the more specific detailed descriptions underneath. Sokvitne (2000) studied the effectiveness of Dublin Core metadata in retrieval. Results showed that the metadata descriptions were nearly useless for retrieval because the quality of the metadata descriptions was poor. However, in the research of Sokvitne, the metadata was used to retrieve the actual data that was described by the metadata descriptions. In our case, the metadata descriptions itself is what we want to retrieve. The resources described are mostly physical objects, and we are interested in the information in the descriptions. Employees in CHIs are trained in making these descriptions. They don't add these description as an afterthought. Describing resources is their main work, plausibly leading to description of much higher quality.

Due to large scale projects in recent years (DigiCULT, 2005), digitization of large cultural heritage collections has matured, making the disclosure of this information a goal in itself. The main problem with disclosing multi-faceted cultural heritage information is that the current expert system in service usually provides no way to access all information in a unified way. Every collection has its own system, with few experts knowing how to really use it effectively.

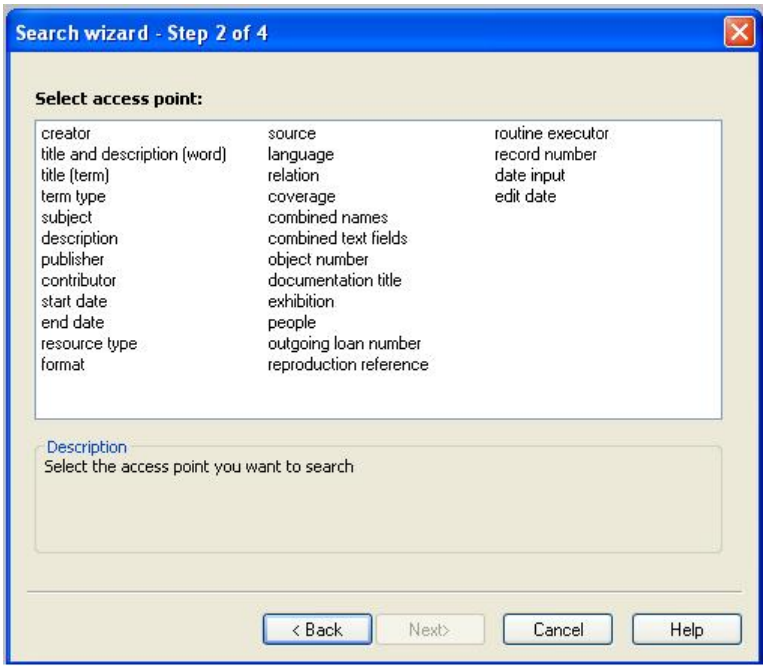

Figure 1: An example description from Kroniek's museum module (partial view).

3 Case Study: A Cultural Heritage Institution

As a case study, we consider the combined descriptions of a large museum, the *Gemeentemuseum, Den Haag*.[2] This museum is a excellent case study, since its combined descriptions cover all three traditions of cultural heritage:

1. It is a **museum**, with 116,846 detailed descriptions of museum objects.
2. It is a **library**, with 277,870 bibliographic descriptions for books, articles, multimedia objects, and others.
3. It is an **archive**, with 728,710 process-related descriptions of activities involving museum objects such as the acquisition, presentation, storage, preservation, loan, or use in expositions.

The descriptions are created, stored and can be retrieved using several separate modules of *Kroniek*, a database management system specifically designed for archives, libraries, and museums. Each module allows the users to search on a specific field of the descriptions (see figure 3). As the figure shows, there are a large number of fields to choose from, which is useful when the user is looking for something very specific and knows which field to use. But the user doesn't always have this knowledge, and may not understand what the fields mean and how they are related. In that case, fielded search might require a lot of trial and error before a satisfactory result is given. Even for expert users, the enormous number of fields, many of which are rarely used, is off-putting.

[2] The *Gemeentemuseum* is most famous for its collections of modern art, prints and poster, fashion, and musical instruments. See, e.g., `http://en.wikipedia.org/wiki/Municipal_Museum_The_Hague` or `http://www.gemeentemuseum.nl`.

		Average	
Module	# documents	doc. size	# fields per doc.
1. Museum	116,846	1,417	32.83
2. Library	277,870	745	17.30
3. Archive	728,710	769	21.04
4. Other	29,124	4,559	–
1+2+3	1,123,426	831	21.34
1+2+3+4	1,152,550	925	–

Table 1: Average size and number of fields per document.

Added to the descriptions, there are 29,124 MS-Word files consisting of press releases, letters on loans and acquisitions, internal communication and meeting notes among others. There are also a large number of PDF and HTML files, images, videos and audio files, but in this stage of the project we have only included the Word files in the test collection. These files form one of the problems in the museum. They contain a lot of interesting and important information, for both internal and external users. However, the file-systems allow search on the file-name only, and the legacy systems have no direct access to them. In some cases, the museum employees extend descriptions by adding links to relevant files on the file system, but these links have to be added manually. Adding links to all relevant files is too much work, and even with a link, a user cannot search on the content of these files.

As shown in Table 1, the document collection consists of more than 1 million descriptions from the 3 modules, and close to 30,000 Word files. The document sizes are in number of characters. The Word files are much bigger than the descriptions. The museum descriptions contain twice as much characters as the descriptions from the archive and library modules; they also contain more fields on average.

To be able to index and access the descriptions through one system, the descriptions were exported from their respective modules in XML (2006). This allows us to maintain their structure and make them readable for other systems as well. The text was extracted from the Word documents using AbiWord (2007). A similar thing can be done for the other (text based) file types.

3.1 Data Sensitivity and Access Rights

During the extraction process, a few important aspects of digital CHI data has come to light:

- **Confidential vs. non-confidential data**: some information of CHIs should not be shared with the general public, like security protocols, insurance values of museum objects, home addresses of employees, donors who wish to remain anonymous, etc. A separation has to be made between confidential, private or sensitive data, and other data.
- **Internal communication vs. cultural heritage data**: the file-system of the *Gemeentemuseum* has many press releases containing very interesting information on exhibitions, artists, styles and periods, but it also contains numerous documents concerning internal communication and other organisational data not directly related to cultural heritage.

Clearly, the confidential and organisational data cannot easily be separated from the "interesting" cultural heritage information.

As a case in point, after deciding to leave out the files and folders of the security department and financial administration, we asked the museum employees to compile a short list of questions to use as queries in our baseline system for which the correct answer should not be returned. One of the questions was: "What is salary of the director?" The correct answer was found at rank 1 in the result list. As a sanity check, it shows that our system seems to work properly, but the data separation problem requires further investigation before making the collection available to external users.

Instead of creating two collections, one for internal users only, and one for all users, we have created different roles for users, internal or external, and give access to the collection based on access rights. Internal users have access to all data, whereas external users only have access to data judged as non-sensitive by the *Gemeentemuseum*. This is a generic solution, which can be applied to different institutions as well. If the collection is extended to include information from other institutions, a specific role for access to the sensitive new data can be easily added, and the required role can be added to this data at indexing time.

We have masked certain fields in the descriptions containing sensitive information: insurance value, storage location, current location, and acquisition source. Furthermore, we have removed the archival descriptions that contain letters or e-mails from people asking questions about the museum from the public part of the collection. The reason for removing the latter is that these descriptions often contain the names, addresses and telephone numbers of people who may expect from the museum to treat their communication as confidential. There are 15,520 of these descriptions (2% of all the archival descriptions). For the MS-word files, we made a similar distinction. There are three sets of files. One set of files with internal communication, which is judged as possibly incomplete and unreliable to provide to the general public. The second set of files contains correspondence between the museum and individuals or other museums, with many names, addresses and telephone numbers. These two sets, numbering 25,957 documents (89% of all word documents) are accessible only by the museum employees. The remaining set of 3,169 MS-word files (11%) form the public part of the collection. In the case of the *Gemeentemuseum*, these sets are identified by their location on the file-systems.

In this way, we end up with one collection of 2 different roles, internal or external, and 4 sub-collections: archive, library and museum descriptions, and MS word documents. The external role gives access to all descriptions apart from the information requests, and all the MS word documents that are linked to from the descriptions. The internal role gives access to the entire collection, including any MS word document from the file-systems.

4 CatchUp

For retrieval we use a standard version of Lucene (2005) to index the entire collection, because it is a widely available and often used system. Unlike (Graupmann et al., 2005), where everything is converted to XML and semantically annotated, and which provides a complex query language, we index the extracted text from the MS-word files as plain text and only the content of the XML elements. The standard Lucene uses a vector space model for indexing and retrieval (Salton and McGill, 1983), and has a simple keyword-based query language. The system, called *CatchUp*, is a primitive first version of a unified

system. *CatchUp* gives all users, internal or external, expert or non-expert, easy access to the full digital cultural heritage content of the *Gemeentemuseum*.

4.1 Retrieval Models

For our ranking, we use either a vector-space retrieval model or a language model (Hiemstra, 2001).

Our vector space model is the default similarity measure in Lucene, i.e., for a collection D, document d and query q:

$$sim(q,d) = \sum_{t \in q} \frac{tf_{t,q} \cdot idf_t}{norm_q} \cdot \frac{tf_{t,d} \cdot idf_t}{norm_d} \cdot coord_{q,d} \cdot weight_t, \quad (1)$$

where

$$tf_{t,X} = \sqrt{\text{freq}(t,X)} \quad (2)$$

$$idf_t = 1 + \log \frac{|D|}{\text{freq}(t,D)} \quad (3)$$

$$norm_q = \sqrt{\sum_{t \in q} tf_{t,q} \cdot idf_t{}^2} \quad (4)$$

$$norm_d = \sqrt{|d|} \quad (5)$$

$$coord_{q,d} = \frac{|q \cap d|}{|q|} \quad (6)$$

and $weight_t$ an explicit term weight (a query operator that is not used in our experiments).

Our language model is an extension to Lucene (ILPS, 2005), i.e., for a collection D, document d and query q:

$$P(d|q) = P(d) \cdot \prod_{t \in q} \left((1-\lambda) \cdot P(t|D) + \lambda \cdot P(t|d) \right), \quad (7)$$

where

$$P(t|d) = \frac{\text{freq}(t,d)}{|d|} \quad (8)$$

$$P(t|D) = \frac{\text{freq}(t,D)}{\sum_{d' \in D} |d|} \quad (9)$$

$$P(d) = \frac{|d|}{\sum_{d' \in D} |d|} \quad (10)$$

The smoothing parameter λ regulates the importance of a query term being present in a document, and helps deal with data-sparseness. The standard value for the smoothing parameter λ is 0.15.

We indexed all text from the documents and all the text in the fields of the descriptions. We left out the field names (i.e. the tags) in the index.

We apply post-filtering of the result list for external users, to remove documents from the list that are not part of the public collection. A user can select one or several sub-collections,

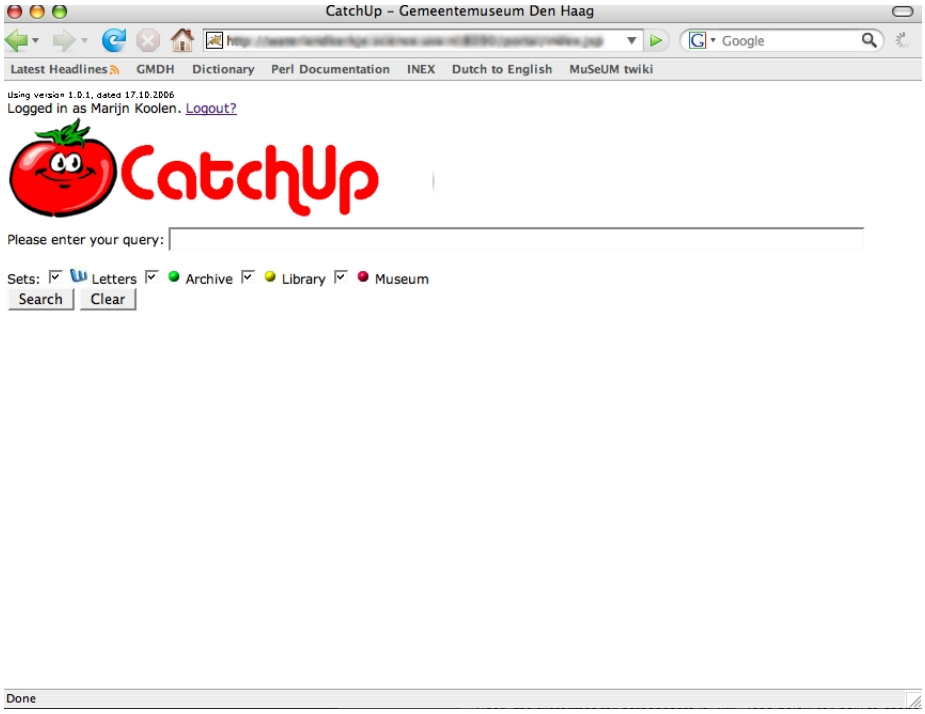

Figure 2: Search box of the CatchUp system providing unified access to all data.

maintaining the option to search only in the archival descriptions for instance. If a user knows what kind of document she is looking for, this option allows her to narrow down the search. Even if all the sub-collections are selected, there is a colored indicator for every result in the ranked list indicating the source sub-collection of the document (i.e. archival descriptions are indicated with a green dot, object descriptions with a red dot, bibliographic description with a yellow dot, and documents from the file-system with a blue 'w', see Figure 2.)

The search result is a standard ranked list, based on the selected sub-collections. If all sub-collections are selected, which is the default, documents are retrieved from and terms are weighted on the entire collection. But giving unified access is only part of disclosing all content. 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?

5 Experiments and Evaluation

Some form of evaluation is required to be able to judge if 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 systems, and second, we need a method to measure how well both systems perform.

5.1 Experimental Setup

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 systems.

We considered using questions send to the museum as topics, but these are often too complex, asking about details that are not described ('What is the title of the painting of a dog in that room with the small bench?'). Therefore we have opted to make create the topics ourselves, but taking the types of questions that visitors ask into account.

We have constructed 66 known-item topics based on documents from all 4 parts of the collection. There are 10 topics for documents of the archive module, 16 for documents of the library module, 23 for documents of the museum module, and 17 topics for documents of the shared file-system. We used the public part of the collection for *CatchUp* on all 66 topics. The *Kroniek* search functions have no direct access to the documents on the file-system, so the 17 topics based on these documents will not have a positive result using *Kroniek*. 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.

Using the known-items topic set, we compared retrieval effectiveness of *CatchUp* with that of 3 modules of *Kroniek* using two measures for evaluation:

- *Success@10*, i.e. the percentage of topics for which the known-item is found in the first 10 results, and
- *MRR* (Mean Reciprocal Rank), i.e. the mean of reciprocal ranks of all known-item topics. The reciprocal rank of a known-item topic is $\frac{1}{r}$ if the known-item (the first retrieved relevant document) is found at rank r.[3]

We experimented with both the vector space model and the language model. To determine whether the observed differences between two retrieval approaches are statistically significant, we used the bootstrap method, a non-parametric inference test (Efron, 1979; Savoy, 1997). We take 100,000 resamples, and look for significant improvements (one-tailed) at significance levels of 0.95 (*), 0.99 (**), and 0.999 (***). For these significance tests, we use *Kroniek* as our baseline, and test both the vector space and language model.

5.2 Results

The results for the success rates are shown in table 2, the result for the MRR are given in table 3. The results for *Other* are for the topics based on documents from the file-system, which is not part of *Kroniek*. The results for *Kroniek* show a clear distinction in performance for the different topic categories. The library module scores much better than the other two modules, both in success rate and reciprocal rank. For 2 of the 10 Archive topics the known

[3]For practical reasons, since the legacy systems have to be operated by hand, we only look at the top 10 results. To allow for a fair comparison, all results reported in the paper assign a reciprocal rank of 0 for all topics not found by the 10th rank.

	# queries	Kroniek (baseline)	CatchUp VSM	CatchUp LM
1. Museum	23	34.78	73.91 **	86.96 ***
2. Library	16	62.50	81.25 *	81.25 *
3. Archive	10	20.00	80.00 ***	70.00 ***
4. Other	17	–	76.47	88.24
1+2+3	49	40.82	77.55 ***	81.63 ***
1+2+3+4	66	30.30	77.27 ***	83.33 ***

Table 2: Success rate by 10th rank for 66 known-item topics.

	# queries	Kroniek (baseline)	CatchUp VSM	CatchUp LM
1. Museum	23	0.1560	0.4416 **	0.5389 ***
2. Library	16	0.5938	0.5833	0.6719
3. Archive	10	0.2000	0.4004	0.3625
4. Other	17	–	0.4761	0.5795
1+2+3	49	0.3079	0.4795 *	0.5463 ***
1+2+3+4	66	0.2286	0.4786 ***	0.5549 ***

Table 3: Mean Reciprocal Rank for 66 known-item topics.

item is found in the first 10 results, and in both cases as the first result since the success rate is equal to the reciprocal rank. For the Museum topics, more known items are found, but at lower ranks. Assuming that the right module is known for each topic, we can calculate the average score over all three modules. The performance of the *Kroniek* is agreeable in terms of the reciprocal rank: the average rank is roughly three, but less attractive in terms of the success rate: only 40 percent of the topics is satisfied in the top 10.

We now turn to the unified system. For both *CatchUp* versions, where we use the entire collection for all the runs, the success rates across the topic categories are relatively stable. The only notable exception is the lower success rate of the language model on the Archive topics. The MRR scores show that the scores for the Library topics are higher than the scores for the Museum topics, which in turn are higher than the Archive topics. The language model performs better than the vector space model, apart from the Archive topics, for both the success rate and the MRR. Note that the unified system also contains documents from the file-system, for which no metadata exists in one of the modules of the *Kroniek*. For these Other topics, we see very comparable performance to the other sub-collections. This clearly demonstrates the effectiveness of *CatchUp* to retrieve documents for which no manual descriptions are available.

When we compare the scores of *Kroniek with the scores of* CatchUp, both on the individual modules and on the combination of the three modules, we see from the success rates that *CatchUp* retrieves more known items. This holds for both the vector space model and the language model, and all the improvements are significant. When including the *Other* topics, for which *Kroniek* cannot retrieve the right documents, the performance of *CatchUp* is only improving. For the reciprocal rank scores, the improvements of the Library and Archive are not significant. The improvements for the Museum topics and the combined topics $1+2+3$, the improvements are significant. The only part where the *Kroniek* comes close is in the library module. For the topics on bibliographic descriptions *Kroniek* ranks the known items

	Frequency	
Type	Top 10	Top 1,000
Museum	2.36	350
Library	1.91	172
Archive	1.97	143
Other	3.76	190

Table 4: Average distribution of the sub-collection documents in the top-10 and top-1000.

better than the vector space model version of *CatchUp*. The language model outperforms *Kroniek* on all subsets of the topics. Summarising, the results show that *CatchUp* clearly outperforms *Kroniek*, with both the vector space model and the language model approach, and the improvements are significant in most cases. This is a non-trivial accomplishment, given that *CatchUp* is searching the entire, combined collection and *Kroniek* is searching only on the appropriate sub-collection.

5.3 Analysis

How can these results be explained? The archival descriptions are very short, so a bias in our retrieval models towards longer documents might lead to a skewed distribution of results with very little archival descriptions. However, the library descriptions are even shorter on average, and as mentioned above, the library topic scores are much better. Now, if we look at the distribution of the top 10 and top 1,000 results (see Table 4) we see that the archival and library description types appear in the top 10 with more or less the same frequency. We show the distribution of result types for the language model run, the distribution of results for the vector space model run is almost exactly the same.

The word files are retrieved much more often in the top 10, while they form only a small part of the collection (in the public part, only 0.3% of the documents are word files.) One possible reason for this phenomenon is that the documents on the file-system contain more text, with higher term frequencies for many query terms. In documents about Mondriaan, the keyword 'Mondriaan' often occurs many times. Both the Vector Space model and the Language Model use the term frequency, t_f in equation (2) and $freq(t, d)$ in equation (8), as an indicator of relevance. For these models, there seems to be a bias towards these natural language documents, because in making metadata descriptions, people are careful to enter terms only when necessary, leading to lower term frequencies. If we look at the top 1000, the museum descriptions are more frequently retrieved than word documents, which is not strange, since there are far more museum object descriptions than word files. The museum object descriptions are also retrieved more often than the library and archive descriptions, which is very probably caused by the fact that museum descriptions are much longer.

So, as there seems to be a preference for longer documents, the lower score on the archive topics might be because of their small size. But why then, do the library topics score so much better? One aspect that can influence retrieval performance is query length (Table 5). Two of the library queries, and two of the museum queries contain words that are removed through stopword removal. These words are excluded from the numbers in Table 5. The average query length per category shows that the queries for the library and archive module and for the word files, for which performance is better than for the queries for the archival

Category	# queries	Query terms	
		Total #	Avg. # (%)
All	66	179	2.71
Library	16	40	2.50
Archive	10	24	2.40
Museum	23	62	2.70
Other	17	53	3.12

Table 5: Average query length per sub-collection.

descriptions, deviate from the average over all queries. The library and archive queries are shorter than average, while the *other* queries are longer than average. The library queries are slightly longer than the archive queries on average, but the Archive topic scores are much lower than the Library topic scores. There is no clear effect of query length on retrieval performance.

Another possible explanation is that museum object and bibliographic descriptions target a unique resource, while looking for a specific description about an exhibition is a harder task, since events such as exhibitions are often described by multiple archival descriptions. If the system retrieves several descriptions about the same exhibition, one or more of them might be relevant for Ad Hoc topics, which often have more general information needs. With known-item topics, even though the system might retrieve several descriptions about the same event we're looking for, we consider only one specific description as relevant. One thing pointing in this direction is the big difference between the success rates and MRR scores for both the vector space and language model, which rank 80% and 70% of the known items in the top 10 results respectively, but rank them poorly compared to the known items in the other topic sets. To investigate this, we need to look at recall, which makes little sense for known item topics. We are currently working on constructing an Ad Hoc topic set, which should make it possible to study this aspect of archival descriptions more extensively.

To find an explanation for the *Kroniek* scores, we investigated the occurrence of query terms in specific fields in the *Kroniek* records. In the archive module, the title fields seems to be the most useful access point. The 10 known-item queries aiming archival descriptions contain 24 terms. Of these, 9 (38%) can be found in the *title* field of the known items. No query terms were found in the description field, indicating that using the *title+description* access point is not more useful than the *title* field alone. For the museum module, the 23 known-item queries contain 62 terms, and 20 of them (32%) are found in the *description* field, making it the most useful field. A few other fields are very useful as well. The *creator* and *notes* fields contain 11 query terms (18%). There are 16 queries about known-items in the library module, containing 38 terms. The *title* is extremely useful in this case, as 26 out the 38 query terms (68%) can be found in the *title* field of the known items. Another important field is the *internal_link_title*, containing 20 of the query terms (53%). The fields that the museum employees use indeed seem to be the most useful fields, although the *title+description* access point seems to offer no advantage over the *title* field alone. But the percentages of query terms found in these fields explain why the library module of *Kroniek* scores so much better on the known-item topics.

To sum up, *CatchUp* seems to be a bias towards natural language documents, as they are

retrieved more often, but this does not lead to a clear difference in performance between known item topics based on natural language documents and known item topics based on descriptions. Among descriptions, there are also some important aspects. Some fields are better access points than others. For the legacy systems, this is important information, whereas for a standard free-text retrieval system, the location of terms in the document plays no role. However, this information can possibly be used to push up the descriptions in the ranking.

6 Discussion and Conclusions

This paper proposed an unconditional IR approach to heterogeneous data, in which heterogeneous data from legacy systems is merged and all metadata structure is flattened, turning a metadata retrieval problem effectively into a free-text retrieval problem. Such an approach raises the following two research questions:

1. Is a full-text IR solution suitable for unified access to all data and metadata?
2. If yes, how does the effectiveness of such a system compare to that of existing, separate, tailor-made expert systems?

In order to answer these questions we conducted an extensive case study based on the realistic context of an existing *cultural heritage institution* (CHI).

In order to build a unified system, we exported propriety metadata formats into an open XML format, and extracted the text from the available full text documents (i.e., Word, PDF). By indexing the resulting documents as plain text using a standard search engine, the resulting system gives access to multiple collections, and can provide relevant information from each sub-collection. The main advantages of a single system is that it is easier to search for information, since it requires no knowledge of the particular sub-collections (i.e., which sub-collection contains the requested information?), nor knowledge of the particular format of the dedicated system (i.e., what search field contains the requested information?). There is a single access point for all combined data. Hence our answer to the first question is that, by converting all digital content to text and indexing it with a standard IR system, we can effectively build a unified system providing access to all data.

The cultural heritage institutions already have working systems providing specialized searching and retrieval tailored to the particular needs of the particular sub-collection. Hence, it is far from clear that a unified system searching all data can lead to comparable retrieval performance. Hence, we performed a comparative evaluation of the legacy systems and the unified system, based on a set of 66 known-item topics spanning all sub-collections. The results show that the performance of the unified system is significantly better than the legacy systems. Hence our answer to the second question is that the performance of a unified system can meet and exceed the performance of individual legacy systems.

Note that the unified system is searching the combined collection, where we assumed perfect sub-collection selection for the legacy systems. The results list contains results from all sub-collections, although the distribution of results seems biased to the longer free-text documents from the file system. This is not necessarily unwanted (e.g., Singhal et al., 1996), but will affect the relative performance of known-item topics from different categories. One solution would be to impose appropriate weights on the sub-collections, which could be estimated from the relative relevance over a large set of queries, as to ensure any

undesired bias (Kraaij et al., 2002; Kamps et al., 2004). Another solution is to exploit the structure preserved in the XML version of the data, by allowing users to formulate more expressive queries (Carmel et al., 2003; Kamps et al., 2006). We are currently investigating such improvements of the unified system.

Although our case study focused on cultural heritage, the problems to provide unified access to heterogeneous data is much broader, and our findings may also directly apply to other enterprise search scenarios (Hawking, 2004). Therefore, our overall finding is that an IR approach to turn a metadata retrieval problem into a free-text retrieval problem, is a viable option to give access to heterogeneous collections.

Acknowledgments

This research is part of the MUSEUM (MUltiple-collection SEarching Using Metadata; `http://www.nwo.nl/catch/museum/`) project of the CATCH (Continuous Access To Cultural Heritage) research program in the Netherlands.

This research was supported by the Netherlands Organization for Scientific Research (NWO, grants # 612.066.302, 612.066.513, 639.072.601, and 640.001.501), and by the E.U.'s 6th FP for RTD (project MultiMATCH contract IST-033104).

References

AbiWord. Word processing for everyone, 2007. `http://www.abisource.com/`.

R. Baraglia, D. Laforenza, and F. Silvestri. SIGIR workshop report: the SIGIR heterogeneous and distributed information retrieval workshop. *SIGIR Forum*, 39(2):19–24, 2005.

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.

CIDOC. The cidoc conceptual reference model, 2006. `http://cidoc.ics.forth.gr/`.

DCMI. Dublin core metadata initiative, 2006. `http://dublincore.org`.

L. Dempsey and R. Heery. Metadata: a current view of practice and use. *Journal of Documentation*, 54:145–172, may 1998. ISSN 0022-0418.

DigiCULT. Technology challenges for digital culture, 2005. `http://www.digicult.info/`.

B. Efron. Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7:1–26, 1979.

R. Fagin, R. Kumar, K. McCurley, J. Novak, D. Sivakumar, J. Tomlin, and D. Williamson. Searching the workplace web. In *Proceedings of the Twelfth International World Wide Web Conference, Budapest*, 2003.

J. Graupmann, R. Schenkel, and G. Weikum. The SphereSearch Engine for unified ranked retrieval of heterogeneous XML and web documents. In *VLDB '05: Proceedings of the 31st international conference on Very large data bases*, pages 529–540, Trondheim, Norway, 2005. VLDB Endowment.

D. Hawking. Challenges in enterprise search. In *Proceedings of the Australasian Database Conference ADC2004*, pages 15–26, Dunedin, New Zealand, January 2004.

D. Hawking, N. Craswell, F. Crimmins, and T. Upstill. Enterprise search: What works and what doesn't. In *Proceedings of the Infonortics Search Engines Meeting*, San Francisco, April 2002.

D. Hiemstra. *Using Language Models for Information Retrieval*. Thesis, University of Twente, 2001.

R. Iannella and A. Waugh. Metadata: enabling the Internet. In *The Information Professions and the Information Professional, Proceedings of CAUSE97*, pages 87–98. Distributed Systems Technology Centre, January 1997.

ILPS. The *ilps* extension of the *lucene* search engine, 2005. `http://ilps.science.uva.nl/Resources/`.

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.

J. Kamps, M. Marx, M. de Rijke, and B. Sigurbjörnsson. Articulating information needs in XML query languages. *Transactions on Information Systems*, 24:407–436, 2006.

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.

Lucene. The Lucene search engine, 2005. `http://jakarta.apache.org/lucene/`.

MARC. Understanding marc bibliographic: Machine-readable cataloging, 2006. URL `http://www.loc.gov/marc/umb`.

D. V. Pitti. Encoded archival description: An introduction and overview. *D-Lib Magazine*, 5(11), November 1999.

G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill computer science series. McGraw-Hill, New York, 1983.

J. Savoy. Statistical inference in retrieval effectiveness evaluation. *Information Processing and Management*, 33:495–512, 1997.

A. Singhal, G. Salton, M. Mitra, and C. Buckley. Document length normalization. *Information Processing and Management*, 32:619–633, 1996.

L. Sokvitne. An evaluation of the effectiveness of current Dublin Core metadata for retrieval. In *Proceedings of VALA2000*. Victorian Association for Library Automation Inc., 2000.

XML. Extensible markup language (XML) 1.1 second edition, 2006. `http://www.w3.org/TR/xml11/`.