MuSeUM: Unified Access to the State of the Art

Avi Arampatzis
Archives and Information Science
Media Studies
University of Amsterdam
avi@science.uva.nl

Jaap Kamps
Archives and Information Science
Media Studies
University of Amsterdam
kamps@science.uva.nl

Marijn Koolen
Archives and Information Science
Media Studies
University of Amsterdam
mkoolen@science.uva.nl

Nir Nussbaum
Intelligent Systems Lab Amsterdam
Infomratics Institute
University of Amsterdam
nir.nussbaum@gmail.com

ABSTRACT

This work addresses the prototypical problem of a cultural heritage institution with the ambition to disclose all of its content in a single, unified system. Like enterprises, these institutions often have heterogeneous collections distributed over multiple legacy systems. The brute-force approach taken here involves a mostly unconditional merging of the heterogeneous sub-collections and flattening of all metadata structures, effectively turning the problem to free-text retrieval. Our main findings are as follows: First, by converting all digital content from several systems of one cultural heritage institution to text, and indexing it with a standard IR system, we show that a unified approach is a viable option to give access to heterogeneous collections. Second, although our approach is simplistic, the initial empirical evaluation validates its superior performance against the legacy fragmented systems currently in use by the institute. Third, in a user study, with test persons ranging from expert users (such as internal employees) to naive users (such as potential visitors of the institution's web-site), we find that all test person's preferred the unified system—even those who work with the existing propriety system on a day-to-day basis.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]

1. INTRODUCTION

Many institutions and enterprises have their digital content stored in (one or more) databases and (shared) filesystems, 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. 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.

At a first 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., 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 filesystems, and it is *heterogeneous* in nature. These problems play an important role in Web retrieval as well [8].

Legacy systems are often 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 usually cumbersome for end-users due the required knowledge on the fields and vocabulary used and 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. For instance, a *cultural heritage institution* (CHI) with the ambition to disclose its digital content to external users, will find itself 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.

This paper reports the results of the pilot study of the *MuSeUM* (*Mu*ltiple collection *Se*arching *U*sing *M*etadata) project [1] which seeks to provide an IR approach to the problem of disclosing multiple data and meta-data collections, each with its own characteristics, in a single, unified

DIR 2007, March 28-29, 2007, Leuven, Belgium.

[1] `http://www.nwo.nl/catch/museum/`

system. The research is being conducted in the realistic context of an existing cultural heritage institution, namely, the *Gemeentemuseum* in the Hague. However, the approach taken and the findings may also directly apply to the *enterprise search* problem [2, 3, 4]. At this stage, three main issues are addressed:

1. Is a full-text IR solution suitable for unified access to all data and metadata?

2. How does the effectiveness of such a system compare to that of existing, separate, tailor-made expert systems?

3. What would the usability of such a system be? Would end-users prefer it?

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.) This simple approach will serve as a baseline to build upon. We created a test set of information requests targeting known items in the collections, and experimented with the system. Moreover, we conducted a user study, with test persons ranging from expert users (such as museum's employees) to naive users (such as potential visitors of the museum's web-site).

The rest of this paper is structured as follows. In Section 2, we detail the current state-of-affairs in the institution. We focus both on the currently available systems (Section 2.1), as well as on the heritage descriptions and other data available (Section 2.2). We continue, in Section 3, with the pilot project proper. We discuss the design of the initial version of a unified system, which we named "CatchUp." (Section 3.1). We also perform a comparative evaluation of the existing systems and the unified CatchUp system (Section 3.2), both in term of retrieval effectiveness as well as in terms of user satisfaction. Finally, in Section 4, we draw some initial conclusions.

2. STATE OF AFFAIRS

2.1 Systems

Currently, the institution has several systems, one containing descriptions of museum objects, one system describing bibliographical objects, and one describing processes-related documents involving the institution (such as an exhibition archive, acquisition, loan, or image rights). We will refer to these systems as the *Kroniek*. Figure 1 shows a screen-shot of one of these systems. Each of these expert systems has a complex interface, allowing for sophisticated querying, but requiring extensive knowledge to use them properly.

Apart from the systems with heritage descriptions or metadata, there are a lot of digital documents related to these objects or processes. Some of the descriptions have links to related documents, but many documents are not (yet) linked to. Only a few of the museum employees really know how to use these systems, and they act as intermediaries between these systems and people looking for information.

2.2 Data

As a first step, we have extracted all the data from the museum's systems. There are three main systems:

- The **Museum** collection containing 116,846 descriptions of museum objects.
- The **Library** collection containing 277,870 bibliographic descriptions (such as books, articles, or multi-media objects).
- The **Archive** collection containing 728,710 descriptions of process-related information (such as exhibitions, loans, or acquisitions).

In total, there are more than 1 million of these descriptions. They have been exported to XML to maintain the structure of the metadata fields, and to allow flexible storage and use.

In addition to the metadata description, there are 29,135 MS word files (there are also a lot of pdfs, jpgs, html files, etc., but these are not used in the pilot project), scattered over the file system, created by employees of the museum, containing information on many different aspects of the museum.

		Average	
Module	# documents	size	# fields
1. Museum	116,846	1417	32.83
2. Library	277,870	745	17.30
3. Archive	728,710	769	21.04
4. Other	29,124	4559	–
1+2+3	1,123,426	831	21.34
1+2+3+4	1,152,550	925	–

Table 1: Document collection statistics.

Table 1 gives some statistics of the document collection. 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.

During this process of data extraction, we encountered a number of interesting problems. Within the collection, there is some information that should not be made accessible, like insurance values of art works, security protocols, wages of employees, etc. There are also a lot of duplicate documents and documents concerning internal workflow, that are of little interest to people outside the museum. There is no clear separation between these data, and other, less sensitive and more interesting data. One of the challenges is to find a good way to distinguish between these types of documents. Our current solution is to treat all the descriptions and the documents that are linked to them as suitable for external access, the public part of the collection. The rest of the documents can only be accessed by museum employees.

3. PILOT PROJECT

3.1 CatchUp

After extracting the data, we modified a standard version of [7] to index the entire collection. The *CatchUp* system is a first version, albeit very primitive, of the final unified system. CatchUp gives all users, internal or external, expert or non-expert, easy access to the full digital cultural heritage content of the museum. For our ranking, we use either a vector-space retrieval model or a language model [5].

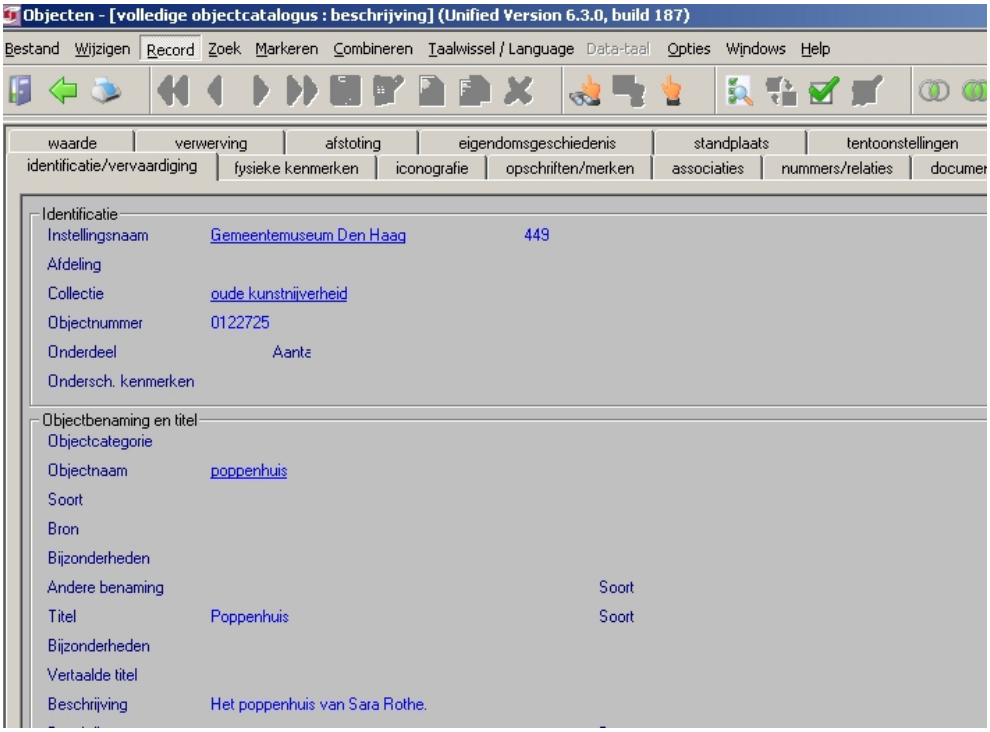

Figure 1: Screenshot of one of the three main systems (objects database) currently in use at the institution.

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

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

where

$$\begin{aligned} tf_{t,X} &= \sqrt{\text{freq}(t,X)} \\ idf_t &= 1 + \log \frac{|D|}{\text{freq}(t,D)} \\ norm_q &= \sqrt{\sum_{t \in q} tf_{t,q} \cdot idf_t{}^2} \\ norm_d &= \sqrt{|d|} \\ coord_{q,d} &= \frac{|q \cap d|}{|q|} \end{aligned}$$

Our language model is an extension to Lucene [6], i.e., for a collection D, document d, query q and query term t:

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

where

$$\begin{aligned} P(t|d) &= \frac{tf_{t,d}}{|d|} \\ P(t|D) &= \frac{\text{doc\_freq}(t,D)}{\sum_{t' \in D} \text{doc\_freq}(t',D)} \\ P(d) &= \frac{|d|}{\sum_{d' \in D} |d|} \end{aligned}$$

The standard value for the smoothing parameter λ is 0.15.

Figure 2 shows a screen-shot of CatchUp. As should be immediately clear: simplicity was our main design principle for CatchUp. A user can select one or several sub-collections, 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 filesystem with a blue 'w'.)

The search result is a standard ranked list, based on the selected sub-collections. If all sub-collections are selected, 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 museum 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?

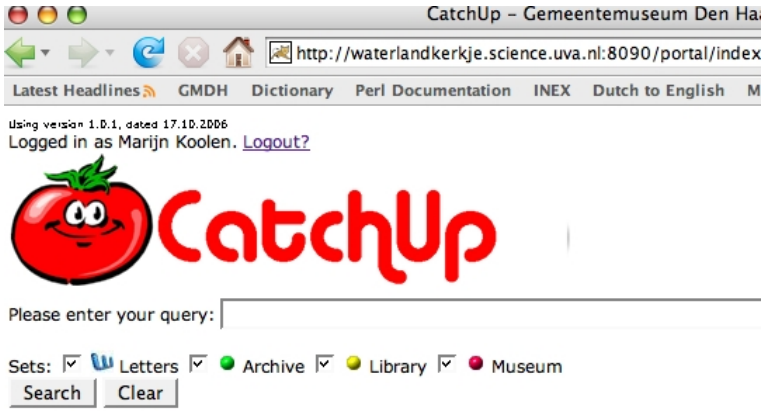

Figure 2: Screenshot of pilot project's unified system, CatchUp.

3.2 Experiments and Evaluation

To find out if a single system is preferred over the multiple expert systems, each specifically designed for the descriptions they contain, we have evaluated CatchUp on two aspects: retrieval performance and user satisfaction.

3.2.1 *Retrieval effectiveness*

To measure retrieval performance, a large test set of ad-hoc topics often gives stable results. However, creating an ad-hoc topic set, even a small one, is very time consuming. A less stable, but also less daunting approach is to use *known-item topics*, where a query is used to retrieve one specific document which is known to be in the collection. We have constructed 66 known-item topics based on documents from all 4 parts of the collection.

Using the known-items topic set, we compared retrieval performance of *CatchUp* with that of *Kroniek*. We experimented with both the standard vector space model (VSM) and the language model (LM). 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 filesystem. 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 filesystem, 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.

We have used 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 average of $1/Rank(K_i)$ of all known-items K_i; K_i's not found by the 10th rank are assigned a reciprocal rank of 0 for calculating the average.

As known-item topics have only one relevant document in the whole collection, the Success@10 score is the same as recall at rank 10. Also, for known-item topics, the MRR score is the same as mean average precision. In our case however, we use a cut-off at rank 10, treating any document at later ranks as not relevant.

To determine whether the observed differences between two retrieval approaches are statistically significant, we used the bootstrap method, a non-parametric inference test [1, 9]. We take 100,000 resamples, and look for significant improvements (one-tailed) at significance levels of 0.95 (*), 0.99 (**), and 0.999 (***).

	# queries	Kroniek	C. VSM	C. 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+4	66	30.30	77.27 ***	83.33 ***
1+2+3	49	40.82	77.55 ***	81.63 ***

Table 2: Success @ 10 for 66 known-item topics.

Tables 2 and 3 summarize the results. The results for *Other* are for the topics based on documents from the filesystem. 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 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.

	# queries	Kroniek	C. VSM	C. LM
1. Museum	23	15.60	44.16 **	53.89 ***
2. Library	16	59.38	58.33	67.19
3. Archive	10	20.00	40.04	36.25
4. Other	17	–	47.61	57.95
1+2+3+4	66	22.86	47.86 ***	55.49 ***
1+2+3	49	30.79	47.95 *	54.63 ***

Table 3: MRR for 66 known-item topics.

	Frequency	
Type	Top 10	Top 1000
Museum	2.36	350
Library	1.91	172
Archive	1.97	143
Other	3.76	190

Table 4: Distribution of results for VSM and LM.

When we compare the scores of *Kroniek* with the scores of *CatchUp*, we see from the success rates that *CatchUp* retrieves many more known items, using either the vector space model or the language model, and all the improvements are significant. Even without the *Other* queries (i.e. 1+2+3 only), for which *Kroniek* cannot retrieve any documents, both *CatchUp* runs find more relevant documents and rank them better as well; although for the reciprocal rank scores, the improvements of the Library and Archive are not 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 better than the vector space model version of *CatchUp*. The language model performs better than the vector space model on most topics, both in terms of success rate and of reciprocal rank, apart from the archive topics.

Both retrieval models achieve a success rate around 80% for most categories. The vector space model scores better on the library and archive topics, while the language model scores better on the museum and *Other* topics. For the reciprocal rank we see some larger deviations. For the archive topics both models perform less than the other topics while for the library topics they perform better.

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 1000 results[2] (see table 4) we see that the archival and library description types appear in the top 10 with more or less the same frequency.

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 filesystem contain more text, with higher term frequencies for many query terms. In documents about Mondriaan, the keyword 'Mondriaan' often occurs many times. Although a high term occurrence frequency is a good indicator of relevance, 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 larger.

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? *Kroniek* scores better on the library topics as well. A further investigation of the known-items and the queries sheds some light on this observation.

First, we looked at the average query length (Table 5).

Category	**# queries**	**Query terms**	
		Total #	**Avg. # (%)**
All	66	183	2.77 (–)
Library	16	40	2.5 (-9.75%)
Archive	10	26	2.60 (-6.14%)
Museum	23	64	2.78 (-0.00%)
Other	17	53	3.12 (+12.64%)

Table 5: Average query length per sub-collection.

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 module and for the word files, for which performance is better than for the queries for the archival descriptions, deviate from the average over all queries. The library queries are shorter than average, while the *other* queries are longer than average. There is no clear effect of query length on retrieval performance.

Second, 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 26 terms. Of these, 9 (35%) 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 64 terms, and 20 of them (31%) 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 (17%). There are 16 queries about known-items in the library module, containing 40 terms. The *title* is extremely useful in this case, as 26 out the 40 query terms (65%) 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 (56%).

The fields that the museum employees use indeed seem

[2]We show the distribution of the language model run; the distribution of results for the vector space model run is almost exactly the same.

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.

This last fact also points to the inherent instability of known item topics in general, and the reason why the library topics score so much better in *CatchUp*. The library known-items match more query terms than the archive known-items. If the known item is stated in the same terms as many other documents, it is hard to single it out. In more general informational topics, this is no problem, as these other documents might be relevant as well.

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. There seems to be a bias towards natural language documents, as they are retrieved more often, but this does not lead to a 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.

3.2.2 User satisfaction

The second aspect of our evaluation is user satisfaction. We have conducted a small user study, where each user performs search tasks, one for each system, and gives feedback on his experiences through questionnaires. To avoid learning and ordering effects, each participant received a separate training task for each system, and the order in which the tasks and systems were used were rotated. The study was conducted with 4 museum employees and 4 external users.

In the questionnaires, we asked the users to compare CatchUp with Kroniek on the aspects such as *ease of use*, *presentation of results*, *relevance of results*, *suitability for the tasks* and *responsiveness*, among others. Some example questions were:

- are you familiar with the subject?
- have you used this system before?
- are you satisfied with the results?
- was the system suitable for the task?
- how relevant was the information found?
- was the presentation helpful?
- which system had the best information?
- which system did you use most often?
- did you revise the query?
- what is good/bad about the system?
- which interface do you prefer?
- which presentation do you prefer?
- which system is more responsive ?

Results show that all participants, internal and external, prefer the single, intuitive interface of CatchUp over the fielded search through multiple interfaces of Kroniek. All participants declared to use search engines several times a day. They like the similarity of CatchUp to many well-known internet search engines, to the point where they failed to see the reason for doing a training task. 87.5% prefers the unified access to all the descriptions and word documents. Both systems are very responsive, but 62.5% of the participants find CatchUp to be quicker and 75% find that CatchUp gives better results on average. Also, 62.5% of the participants think the result presentation of CatchUp is better, especially because the result list has good indicators for the source of the documents (sources are archive, library and museum descriptions and word documents).

An important finding is that people find many documents hard to read. In Kroniek, there are too many tabs and fields to get a overview of the data. In CatchUp the same problem holds, all the fields are presented in one big list, so users have to scan for the right information. CatchUp also shows small snippets from the documents in the result list. Although useful for the MS word files, the snippets of the descriptions are hard to read and not very helpful. Apart from being hard to read, many users indicated that many of the descriptions are not very informative. The bibliographic descriptions for instance, contain no information, apart from the title, on what the bibliographic object it is describing is about. Most of the museum objects are described with one short sentence. Additionally, 25% of the participants indicated that there was not enough relevant information in the collection to properly perform the tasks.

4. CONCLUSIONS

The MuSeUM project's pilot study ran from April until September 2006. In this period, the aim was to build and evaluate an initial version of a unified system, that can serve as a baseline for the rest of the MuSeUM project.

Our achievements were the following. First, we extracted over 1 million metadata descriptions from various propriety systems, and converted them into an open XML format. Second, we created a baseline version of a unified system, combining data from multiple collections. Using a simple interface, any user can get easy access to all data. Third, we have done a comparative evaluation of the unified system and the existing propriety systems, using a set of known-item search topics targeting all sub-collections. We found that the unified system is significantly more effective. Fourth, we have conducted a small user study, with test persons ranging from expert users (such as museum's employees) to naive users (such as potential visitors of the museum's web-site). We found that all test persons preferred the unified system—even those that work with the existing propriety system on a day-to-day basis.

Although users prefered the unified approach and the ability to search in whole descriptions instead of single fields, there is still ample room for improvement. The presentation of the results and the document content can be improved upon. By unfolding the links in the descriptions, to include the text of linked descriptions and documents, the bias towards word documents might be reduced, possibly improving retrieval performance on the descriptions. Note that the baseline version of a unified system as developed in the pilot project, is tying together the descriptions from

the three traditional cultural heritage pillars: descriptions of museum objects, of bibliographic documents, and of process-related documents. In this light, our results transcend the particular context of the institution and hold the potential to generalize to a range of institutions in the cultural heritage field.

5. ACKNOWLEDGMENTS

This research was supported by the Netherlands Organization for Scientific Research (NWO) Catch program under project number 640.001.501. We would like to thank all users that participated in the user study.

References

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

[2] 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.

[3] D. Hawking. Challenges in enterprise search. In *Proceedings of the Australasian Database Conference ADC2004*, pages 15–26, Dunedin, New Zealand, January 2004. Invited paper: `http://es.csiro.au/pubs/hawking_adc04keynote.pdf`.

[4] 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. `http://es.csiro.au/pubs/hawking_se02talk.pdf`.

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

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

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

[8] F. S. Ranieri Baraglia, Domenico Laforenza. Sigir workshop report: the sigir heterogeneous and distributed information retrieval workshop. *SIGIR Forum*, 39(2):19–24, 2005.

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