Searching Titles with Initial Articles in Library Catalogs

A Case Study and Search Behavior Analysis

By Clément Arsenault and Elaine Ménard

This study examines problems caused by initial articles in library catalogs. The problematic records observed are those whose titles begin with a word erroneously considered to be an article at the retrieval stage. Many retrieval algorithms edit queries by removing initial words corresponding to articles found in an exclusion list even whether the initial word is an article or not. Consequently, a certain number of documents remain more difficult to find. The study also examines user behavior during known-item retrieval using the title index in library catalogs, concentrating on the problems caused by the presence of an initial article or of a word homograph to an article. Measures of success and effectiveness are taken to determine if retrieval is affected in such cases.

Clément Arsenault (clement.arsenault@umontreal.ca) is Assistant Professor, École de bibliothéconomie et des sciences de l'information (EBSI), Université de Montréal, Montréal (Québec), Canada. **Elaine Ménard** (elaine.menard@umontreal.ca) is a Ph.D. candidate at EBSI.

This paper has received generous financial support from the Social Sciences and Humanities Research Council of Canada as well as from the Association pour l'avancement des sciences et techniques de la documentation (Québec) to whom the authors are grateful. The authors also wish to thank the Collège de Maisonneuve Library (Montréal) for granting access to their premises. Thanks are also due to Carole Paradis of the Systems Office staff of the Université de Montréal libraries for providing the transaction logs of the Atrium catalog.

Submitted August 1, 2006; accepted for publication October 9, 2006, pending revision; revision submitted October 13, 2006, and accepted for publication.

When filing entries alphabetically in an index, ignoring initial definite and indefinite articles is customary.[1] For instance, the book titled *The Earth and Its Inhabitants* is normally filed under the letter "e." This procedure is used almost universally because initial articles "tend to be used intermittently," and also because, due to the high occurrences of initial articles in titles, it would otherwise produce very large groupings of entries beginning with the same word, thus losing the desired alphabetical dispersion of entries within the index.[2] In the current version of the MARC 21 standard, this procedure can be achieved, for the first index subfield in some fields, by using a numerical indicator (the non-filing characters indicator) corresponding to the number of initial characters to be ignored at the beginning of the string being indexed. In the above example, the non-filing indicator of field 245 (title) would be set to 4, indicating that the first four characters (t-h-e and the space) are to be ignored for indexing.[3] Using this technique allows the initial article to be retained in the title field and used for display, without being taken into account in the browse index.

Because the non-filing indicator is not available for all the fields in which articles and other non-filing elements occur, and also because non-filing data elements do not always occur at the beginning of a field, a new technique, setting off the non-filing zone by means of control characters, was approved in 1999 as a result of American Library Association (ALA) Machine-Readable Bibliographic Information (MARBI) Committees Proposal 98-16R.[4] Guidelines for use of the new non-filing control characters were discussed in two discussion papers, DP118 (June 1999) and 2002-DP05 (January 2002), and finally published in

2004 by the Network Development and MARC Standards Office of the Library of Congress.[5] This procedure offers more flexibility, as it allows the cataloger to identify non-sorting zones virtually anywhere in the record and tag them with the use of special control characters whose function is to delimit the beginning and the end of the non-filing elements. As far as data representation is concerned, there are fairly standardized, documented, and efficient ways of dealing with initial definite and indefinite articles in data elements; however, the MARC coding controls only the way initial articles are to be indexed, not the way the retrieval is done.[6] Less standardization is found at the retrieval stage and this is what is investigated in this study.

All systems preprocess search strings to some extent (e.g., ignoring case distinction, omitting punctuation or replacing it with spaces, ignoring diacritics) before sending them to the index. When a user launches a browse-title search in a library catalog, the retrieval module may activate an algorithm to detect the presence of an inopportune initial article at the beginning of the query string. Because most initial articles are removed from the entries when indexing the title strings, even if a user includes an initial article in his or her query, the algorithm will automatically eliminate the word/article and bring the user to the correct entry point in the index. This procedure may prove very useful in some cases. For instance, if the user retains the initial article in a search query (for example, ti=the earth and its inhabitants), the algorithm detects the initial article and automatically suppresses it from the search query before it is sent to the index. In this example, the system therefore will bring the user the index of titles beginning with the letter "E" rather than the letter "T."

Nonetheless, most of these algorithms are not sophisticated enough to detect some linguistic subtleties, which can result in retrieval problems. This automatic detection of initial articles in search queries poses a number of problems, particularly in multilingual environments.[7] The cataloger's decision to declare an initial word as an article to be ignored must be based on several factors, among which the language comes first, since it can be reasonably assumed that an initial article in one language will have a corresponding legitimate non-article equivalent in another language. This is the case, for instance, in German with the article *"die,"* which is homographic to (i.e., spelled with the same sequence of letters as) the English verb "to die." It would not be correct to file the title *Die Another Day* under the letter "A". In some cases, it is even necessary to grammatically analyze the titles in order to avoid incorrect assumptions within a language. In French, for instance, the definite article *"la"* is homographic (albeit the diacritic) to the adverb of place *"là"* ('there'); and the word *"un"* can either be an indefinite article, as in *Un destin tragique,* a pronoun, as in *L'un d'entre eux,* or a number, as in *Un, deux, trois, partez!* It can even be part of an adverbial locution, as in *Un peu de fatigue.* That is not counting the fact that it also is the homograph of the acronym form for United Nations (UN). Therefore, processing titles case by case is essential. Also, sentences (and titles) can begin with only one article, so it makes no sense grammatically to remove two or more words from the beginning of a title search query. Yet, the algorithms tested in this project will remove any number of words that appear at the beginning of a search query that match the words in their exclusion list. For instance, in Atrium (the Université de Montréal catalog), the query *"un thé au Sahara"* will be transposed to *"au sahara"* because the "un" matches a French article and the "thé," when transposed to "the," matches an English article.

The detection algorithms included in most information retrieval systems are not sophisticated enough to detect these linguistic subtleties, which are the cause of some retrieval problems. Some homographic non-article words might be erroneously removed from the queries. This is the case for a title such as *Las Vegas, The Success of Excess*. This title will be correctly filed in the index under letter "L" since the word "Las" is part of a place name, but if the word *"Las"* is included in the exclusion list of the algorithm, it will be interpreted as the Spanish definite article and automatically stripped of the query string, and the user will be misguided to the letter "V" in the index where the entry is nowhere to be found.

Suppose a user needs to find the work by Michel Leiris entitled *À cor et à cri*. Browsing through the title index normally would be done with the standard query "a cor et a cri." Unfortunately, if the initial article detection algorithm is activated, the user will be misguided to letter "C" in the index since the initial "a" of the query will be, in this case, wrongly interpreted as the English indefinite article "a" and the query text will be truncated, often without the user being aware of it, becoming "cor et a cri." The title having been correctly indexed under letter "A," the user will be wrongly positioned in the index as illustrated (figure 1) and may wrongly assume that the title is not in the collection. This lack of system feedback most probably has a negative impact on end users learning to use the catalog.

In the catalog (the University of Toronto catalog) in figure 1, the user has to choose between two search modes: either the keywords mode (*containing*), or the browse mode (*starting with*). If the *starting with* option is chosen, the user will probably draw the conclusion that the document being sought is not in the catalog, since the title is not displayed in the results. The record nonetheless can still be retrieved using a keyword search. Choosing the *containing* option presents another difficulty. In the catalog of the University of Toronto (in April 2006), querying "cor" produces 1,160 records, which must then be painstakingly examined one by one; querying "cri" produces 219 records, which is better.

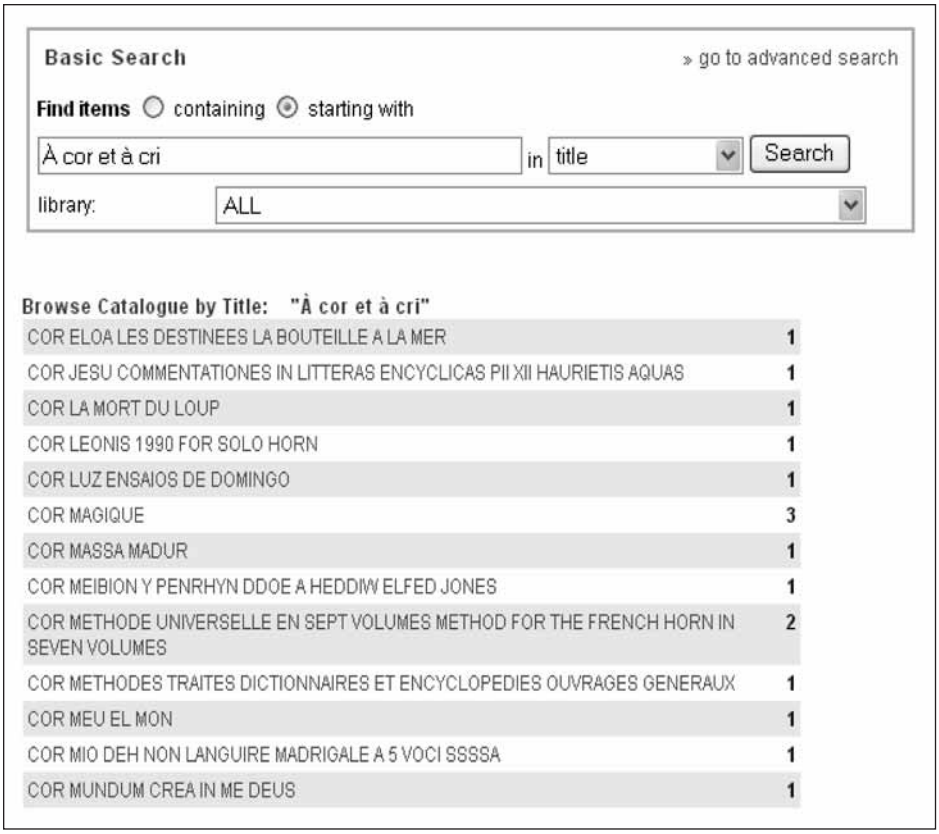

Figure 1. Example of an unsuccessful search in browse mode

This is still high, especially considering that the search is for a single known title. Querying "cor cri" (with an implicit Boolean AND) produces five records, which is more acceptable. Nonetheless, some titles only offer very limited terms when searched in the keywords mode—for example, *À la française* or *À tous*. Such searches in keywords mode lead to very large search results sets that are virtually unusable—6,608 and 2,093 results respectively (in the University of Toronto catalog).

A more efficient solution may be to deactivate the initial article detection algorithm in the search module and to replace it by providing the end users with clear instructions on omitting initial articles in queries. Taylor reports that if the instructions are clearly positioned (see figure 2 for an example) users will follow the instruction: "users tend to follow this advice *if* the instruction is noticeable and can be seen from the search box."[8]

Given these observations, one may question the usefulness of an initial article detection algorithm based on an exclusion list in a library catalog since its use may cause as many problems as it solves. On the one hand, the use of an exclusion list affords some help to the naive searcher by participating in the formulation of his or her queries. Such users thus can cut the electronic reference retrieved and paste it directly in the Search dialog box of the catalog without concerning themselves about anything else. If the title begins with a word or a series of words that are contained in the exclusions list, the search algorithm will remove the unnecessary words from the query without a user's knowledge. On the other hand, this very exclusion list has several drawbacks and can disadvantage the users. One may ask, therefore, what course to follow. A professional librarian may be expected to know how to get around this type of retrieval problem, but this is not the case with end users, who are increasingly independent in their bibliographic searches.

Research Objectives

The goal of the first stage of this research was to examine the extent of the retrieval problems caused by erroneous initial article detection at the retrieval stage in library catalogs. Consequently, two specific objectives were defined:

- Identify which initial articles have the potential to cause the most problems due to interference with non-article homographs
- Estimate the proportion (i.e., number of records with affected titles divided by total number of monographic records in the database) of bibliographic records (monographs) that are affected because of these non-article homograph words at the beginning of the title field.

The goal of the second stage of the project was to study the extent of the above-mentioned retrieval problems from the point of view of the user. To achieve this, four other specific objectives were defined:

- Determine whether end users tend to keep or omit initial articles from titles in their browse queries
- Identify which search mode is used by end users when they search the title index of the library catalog, when the titles they look for begin with an article
- Verify whether the success rate (i.e., the proportion of

retrieved records) when searching in the title index is affected by the presence of a non-article word, which is homographic to an article
- Establish whether or not the identified problem (homographic confusion between a non-article initial word in a title and an initial article) affects the efficiency level (time and effort required to perform a search task) in title-based retrieval.

If these objectives could be carried out, it would be possible to empirically measure the extent of the retrieval problems identified. During preparation of this project, the authors noted that literature on this subject is scant; this paper aims to study this phenomenon in greater depth.[9] Title searching is still one of the most frequent types of search in library catalogs. Making it as efficient as possible is, therefore, advisable. Broadbent's failure analysis study revealed that around 40 percent of her survey participants came to the library looking for known items (either author or title search).[10] Larson's study on OPAC use also showed that, during his data collection phase (1986) in a specific catalog, the number of known-item searches (author and title) exceeded topical searches.[11] More specifically, in 1987 Kaske measured that more than 27.5 percent of searches in a specific catalog were title searches.[12] Matsushita's analysis of the OPAC log at the Kunitachi College of Music Library, Tokyo (Japan) in 2000 also revealed that the most frequently used access keys are names and titles.[13]

Research Method

The research was carried out in two phases. The first phase of the study analyzed more than 6,000 bibliographic records from the Atrium catalog (Université de Montréal). For the second phase of the study, a controlled experimental method to collect data was adopted, which made measuring the extent of the problem in one specific catalog (the University of Toronto catalog) possible. The means at the authors' disposal being limited, this study only explored one specific catalog and prepared for a more comprehensive study of several different catalogs.

Phase 1: Case Study

For the first part of this study, the decision was made to focus on Atrium, the Université de Montréal Library catalog, as a case study. Research was further limited to monographic titles by selecting entries found in the following MARC 21 fields: 240, 245, 246, 700, 710, 711, 730, and 740, thus excluding series titles. Time and money constraints made excluding them from the sample necessary.

To meet the first two objectives, the following research questions were formulated:

Question 1: Which of the articles on Atrium's exclusion list have the most entries beginning with that string of letters when not used as an article?
Question 2: What proportion (i.e., number of records with affected titles divided by total number of monographic records in the database) of records is affected by the deficient retrieval algorithm in Atrium?

Data collection began by identifying the 41 articles in the exclusion list used by Atrium's initial article detection algorithm. The list is reproduced in table 1.

This list was developed locally for internal purposes and for the needs of the collection. It represents only a fraction of all initial articles listed in Annex E of the *Règles de catalogage anglo-Américaines*.[14] The local list was used since research could only be performed on the articles already in the exclusion list. It should be noted that, due to system limitations, investigating the French article "l'" was not possible. This resulted in a total of 40 articles under investigation.

To answer the authors' first research question, each article was searched individually in browse-title mode. The title index was then systematically and thoroughly scanned in order to find all the entries beginning with a non-article word homographic to an initial article. This was done by typing the article in the search box. It

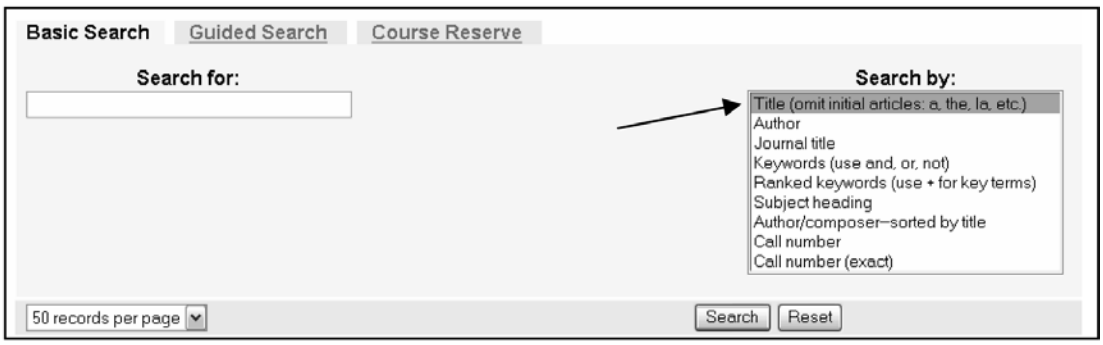

Figure 2. Example of clear instructions in the search interface

should be explained here that if only the article is included in the search string (and no other words), the system does not strip the "article" and positions the user at the beginning of the title index for that word. This is how it was possible to thoroughly scan the index for each article. For each entry thus identified, the corresponding MARC record was examined to find out which field contained the problematic word. Problematic entries were recorded in a spreadsheet; those entries that resulted from the inevitable miscoding of the non-filing indicators were not retained. While examining the MARC record, the record number in field 001 was also recorded to allow the total number of affected records to be determined. This number was less than the number of problematic entries found in the index, because any given record could contain more than one title and therefore generate two or more problematic entries in the index.

To provide an answer to the authors' second research question, the total number of affected records (those that generate at least one problematic entry in the title index) was compiled and compared to the total number of monographic records contained in Atrium at the time of research (summer 2004). It was then possible to obtain this data from the Université de Montréal Library systems office.

Phase 2: Search Behavior Analysis

For the second part of the study, a controlled experiment involving real users was prepared. Given the exploratory nature of this paper, and the limited means at the authors' disposal, the decision was made to use Atrium, the University of Toronto catalog, as a case study. This catalog was chosen because the retrieval module integrates a detection algorithm designed to detect the presence of the three English articles: "a," "an" and "the," and offers a search interface on which it is possible, at the first level, to select a specific search mode (browse or keywords). Atrium automatically defaults to keyword searches and this is the reason why it could not be used for this part of the study. Some transaction logs of queries entered into Atrium nonetheless were used, along with the data collected from the University of Toronto search sessions, to provide additional data for question 3.

To meet the four objectives defined for this part of the study, the following four research questions were formulated:

> **Question 3:** Do users usually keep the initial articles in their queries when searching the title index in browse mode or do they leave them out?
> **Question 4:** When users search for known titles, which mode do they usually use: "browse" or "keywords"?
> **Question 5:** What is the proportion (number of problematic records found divided by total number of problematic records searched) of monographic titles containing a word wrongly processed as an initial article by the detection algorithm that is actually retrieved by the end users, and is this proportion the same for titles not affected by this problem?
> **Question 6:** Are monographic titles containing a non-article word homograph to an initial article usually harder to retrieve than other titles, in terms of time and effort?

Table 1. Articles in Atrium's exclusion list

a	der	ein	enas	hena	l'	li	um	uno
ai	die	eine	gli	henas	la	lo	uma	
al	e	eis	hai	het	las	los	un	
an	een	el	heis	hoi	le	oi	una	
das	eene	ena	hen	i	les	the	une	

Note: "ai" is a contracted article in Italian. As such, it should not in theory be included in this list.

To answer the first of these four research questions, user behavior when searching the title index of a library catalog was analyzed. The transaction logs provided by the systems office of the Université de Montréal libraries were initially examined for the searches in browse mode in the title index of the Atrium catalog for the duration of one month (October 2005). With these data in hand, checking whether users usually keep the initial articles in their title queries, or whether they leave them out, was possible.

To answer the three remaining research questions, the authors first compiled all titles containing a word that might be erroneously considered as an initial article in the University of Toronto catalog. The exclusion list used at the University of Toronto catalog consists of only the three English articles. The authors built a file of all titles that might be difficult to retrieve—in other words, the documents whose title begins with the word "a," "an," or "the" when this word is not an article (for example: *À bout portant*; *An der Wegscheide*; *Thé ou café, Monsieur le Ministre?*)—and obtained 4,384 such document titles. In order to create the data sample, only those titles were kept that were in French or in English (i.e., 1,545 titles), because participants in the study were only fluent in these two languages.

From this set of problematic titles, 24 lists were prepared, with 30 titles in each, all titles being selected at random. In order not to influence the search behavior of the participants in the study, the authors mixed different types of titles in each list. Each list of 30 titles was made up of 3 groups of titles as follows:

> **Group 1:** Five titles beginning with an "ordinary" word, i.e., neither an article, nor homographic to an article (for example, *Out after dark*).

Group 2: Ten titles beginning with a real article (for example, *A very profitable war*).
Group 3: Fifteen "problematic" titles, i.e., beginning with a non-article word homograph to an initial article (for example, *À la plage*).

Throughout the rest of this paper, the first two groups of titles are usually referred to as the "non-problematic" titles and the third group of titles as the "problematic" titles.

All titles included in the lists were cataloged in the University of Toronto Library catalog, and were therefore, in principle, retrievable. The order of presentation of the titles in the lists to be searched was determined randomly, and it was modified for each list to minimize the learning factor. An example of such a list appears in the appendix.

Once these lists were prepared, 24 students at the pre-university level (first or second year of Cégep [Collèges d'enseignement général et professionnel]), enrolled in the pre-university profile (in Québec, Cégep is a required step between high school and university), were asked to try and locate the bibliographic records for the titles listed on one list. Each participant received a different list so that there would be no contamination effect. The main reason for selecting college students was to have a rather homogenous group from the point of view of exposure and experience with bibliographic searching in catalogs. Each student was requested to search all titles on his or her list using one or the other of the two search options *containing* or *starting with* as shown in figure 3.

At the start of each session, the two search options were alternated, selecting the *containing* option initially for one half of the participants, and the *starting with* option for the other half, to avoid a bias in favor of either of the two search modes, at least at the beginning of the search process. The participants were completely free to use either of the two modes at any time during the search session. The *title* index was preselected and the participants were not allowed to change it. Each of the search sessions was recorded using Camtasia, a software application designed to record all the operations performed on screen and to create a video that reproduces the search sessions faithfully.

Once they retrieved a record, the participants had to write down the call number on the form (see appendix),

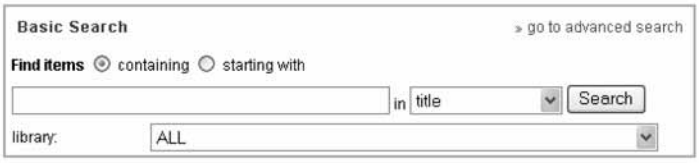

Figure 3. Basic search interface of the University of Toronto catalog

which made it possible to easily ascertain the success rate. Their answers were double-checked by replaying each video. The following information was also recorded for each title:

- Starting time: the moment when the user executes his query by clicking on the Search button
- End time: the moment when the user displays the right record (if found)
- Search mode for each query: *containing* (keyword mode) or *starting with* (browse mode)
- Number of results: in the case of keyword searches, the number of results retrieved
- Initial article inclusion or omission in string search, for titles beginning with an article

Observations and Analysis

For the first phase of the study, data collection was performed between July 5 and August 6, 2004. The authors were able to identify 6,360 problematic entries in the title index and believe the results would have been higher if series titles had been included because these titles often contain initial articles.

Question 1

Which of the articles on Atrium's exclusion list have the most entries beginning with that string of letters when not used as an article?

Table 2 shows the total number of affected entries for each surveyed article and their origin in the record. A rapid survey of the data in table 2 clearly shows that some of the articles potentially were much more problematic than others. Almost half of the articles in the exclusion list never generated any problematic entries. Conversely, the article "a" alone generated 4,230 problematic entries in the index (66 percent of the total). The main explanation is that the article "a" is very common in English (and in other languages as well) and it is also a very frequently used preposition in French. For example, 1,205 documents with a title beginning with "À propos de . . ." ("All about . . .") were noted. All these entries were the cause of retrieval problems in browse mode searches. It was also noted that problems occurred when the title began with the initial of a first name beginning as "A" (*A.B.C. contre Poirot*), and also for many acronyms (A.A.C.R., A.B.B.), or the many works beginning with A.B.C. (*A.B.C. de la lecture,* for example). The high proportion of French language works in the Atrium catalog, as compared to catalogs in

English-speaking institutions, magnified the problem in this case. For instance, the University of Toronto catalog (being much larger than Atrium) has a lesser proportion of problematic records, quite probably because the proportion of French language resources in the former is lower than in the latter.

Question 2

What proportion (i.e., number of records with affected titles divided by total number of monographic records in the database) of records is affected by the deficient retrieval algorithm in Atrium?

The data presented in table 2 indicates that the total number of problematic entries in the title index was 6,360. While matching this data with the record numbers collected while doing data collection, these entries were found to be coming from 5,111 distinct bibliographic records in the catalog. Again, one should remember that any one record can contribute more than one entry in the title index. For instance, a record might have two problematic titles, one in field 245 and one in field 740.

The total number of monographic records in Atrium at the time of the data collection was estimated to be approximately 1,318,000. It may, therefore, be estimated that the proportion of monographic records affected by the initial article detection algorithm was slightly less than 0.4 percent (table 3). This proportion concerns only those titles found in six MARC fields, and this number would probably be higher if series titles had been considered.

For the second phase of the authors' research, a log of Atrium browse-title queries was captured for the month of October 2005. For the part involving participants, data were collected at Collège de Maisonneuve (Montréal, Canada), between January 30 and February 6, 2006. Recruiting was done through posters explaining the tasks to be performed, the estimated time required (roughly 45 minutes), and the remuneration offered ($20).

Question 3

Do users usually keep the initial articles in their queries when searching the title index in browse mode or do they leave them out?

Analysis of the queries collected in the transaction log of the Atrium catalog indicated that users seemed to retain the initial article in their queries in approximately two cases out of three (table 4). Out of the 12,216 queries recorded in the transaction log, 1,468 queries (approximately 12 percent) were queries made to search works whose titles began with an article. This was estimated to the best of the authors' knowledge by examining each query on a case-by-case basis, but it was not always possible to be 100 percent certain whether the title of the resource sought by the end user actually began with an article. From these queries, it was observed that the initial article was omitted in only 36.8 percent of all cases, leading the authors to believe that end users usually would rather leave the initial articles in their queries. Comparing these data with other catalogs where

Table 2. Number of index entries affected for each article

	MARC field*							
Article	**245**	**740**	**700**	**730**	**246**	**240**	**Total**	**%**
a	3,519	368	157	5	172	9	4,230	66.5
e	391	101	5	0	34	29	560	8.8
i	377	63	53	0	36	4	533	8.4
la	185	55	21	3	5	0	269	4.2
le	222	5	1	0	1	0	229	3.6
an	121	6	28	0	3	5	163	2.6
al	78	8	2	1	1	9	99	1.6
el	56	0	0	0	0	0	56	0.9
ai	33	19	1	0	1	0	54	0.8
los	40	0	0	0	0	0	40	0.6
un	28	12	0	0	0	0	40	0.6
um	24	0	0	0	0	0	24	0.4
las	11	2	5	0	2	3	23	0.4
li	17	3	3	0	0	0	23	0.4
the	5	0	0	0	0	0	5	0.1
het	1	3	0	0	0	0	4	0.1
uno	4	0	0	0	0	0	4	0.1
der	1	0	0	0	0	0	1	<0.1
ein	0	0	0	0	0	1	1	<0.1
eis	1	0	0	0	0	0	1	<0.1
les	1	0	0	0	0	0	1	<0.1
19 others	0	0	0	0	0	0	0	0
Total	**5,360**	**1,385**	**976**	**739**	**501**	**300**	**6,360**	**100**

*No problematic entries were found from fields 710 and 711.

Table 3. Number of affected entries and records

Category	Number	Percent of total records
Monographic records in Atrium	1,318,000*	
Problematic entries in the title index	6,360	
Affected records	5,111	0.3888

*Number is approximate

this feature is not present would be interesting, because Atrium users may have learned over time, from using the catalog, that they do not need to pay attention to the initial article. Nonetheless, in their study of known-item queries in OPACs, Kan and Poo noted the same behavior observed in this study.[14]

Similar proportions were observed when the video-recorded search sessions in the University of Toronto catalog were analyzed (see table 4). Out of 54 queries made in browse mode to find titles with an initial article, 37 queries (68.5 percent) contained the initial article, while the user had not included the article in the other 17 queries (31.5 percent). The authors' analysis revealed that the queries where the initial article was omitted were more successful. Of 37 queries in which the initial article was retained, only 18 (48.6 percent) successfully retrieved the desired record. This rises to 88.2 percent when the initial article was removed from the query.

Table 4. Analysis of user searches

Data source	Type of query	Number	%	Successful queries (number)	Successful queries (%)
Atrium transaction log (Oct. 2005)*	Queries in browse-title mode (total)	12,216			
	Queries for titles with an initial article	1,468	100		
	Queries with the initial article kept	928	63.2		
	Queries with the initial article omitted	540	36.8		
Video-recorded search sessions in Univ. of Toronto catalog (Jan. 30–Feb. 6, 2006)	Queries in browse-title mode (total)	213			
	Queries for titles with an initial article	54	100		
	Queries with the initial article kept	37	68.5	18	48.6
	Queries with the initial article omitted	17	31.5	15	88.2

*Transaction logs do not report query success

Table 5. Analysis of searching modes

	Browse mode		Keywords mode		Total	
	Number	%	Number	%	Number	%
Total queries	234	23.1	778	76.9	1,102	100
First query issued for each titles	128	17.8	592	82.2	720	100
Last query issued for each title found	64	9.6	600	9.4	664	100

Question 4

When users search for known titles, which mode do they usually use: browse or keywords?

The compilation of the total number of queries made by the 24 participants to find their 30 titles indicates that more than three quarters of the queries were issued using the keywords mode (see table 5). This proportion rises to 80.2 percent if only the first query is counted for each title. Following these observations, one might assume that the users' preferred mode is the keywords mode, but it must be remembered that the title samples submitted to the participants consisted of 50 percent problematic titles, which is not at all representative of the proportion of problematic titles in a catalog (less than 0.4 percent, according to the authors' previous analysis). Because of the initial article automatic detection algorithm, retrieving these titles in browse mode is nearly impossible. The authors' data reveal that none of the 360 problematic titles (15 titles on each of the 24 lists) could be retrieved using the browse mode. Analysis for all titles reveals that the last query—the query that successfully retrieved the record—was made in keywords mode in 9 times out of 10 (table 5). It is not surprising, therefore, that users ended up choosing this mode most of the time.

A chronological analysis of the queries indicates that at the beginning of the session, users were using the browse mode more often. Seventeen out of the 24 participants (71 percent) used this mode to issue their very first query, in spite of the authors taking care to preselect the keywords mode as the starting selection for half of them. In figure 4, that behavior can be seen at the beginning of the session. For the first 5 titles, both modes scored approximately the same—they were equally used. As the session continued, users progressively abandoned the browse mode for the keywords mode (only 2 percent of the queries in browse mode for the last 5 titles searched) in spite of the fact that

the browse mode is known to be more efficient for locating a known document. Affirming that users prefer the keywords mode is difficult, because the overrepresentation of problematic titles in the sample gave the participants the misleading impression that the browse index mode was less efficient. A Web catalog analysis by Halcoussis and colleagues revealed that when asked to rate the level of satisfaction regarding organization of a Web catalog based on a variety of criteria, "browse-title" ranked as one of the highest search types, with a coefficient estimate of 0.919 (subject searches being set to zero as a control category), while "keywords-in-title" ranked as the lowest search type, with a coefficient estimate of -1.711.[16]

Question 5

What is the proportion (number of problematic records found divided by total number of problematic records searched) of monographic titles containing a word wrongly processed as an initial article by the detection algorithm that is actually retrieved by the end users, and is this proportion the same for titles not affected by this problem?

The search for a known document (known-item search) for which the end user has the exact title is one of the easiest imaginable task in any catalog. The success rate should be near 100 percent. This is what was observed for all the titles in the samples that were not problematic (with articles and without articles combined). However, for the titles considered problematic because of the presence at the beginning of the field of a non-article homograph to an article, 2 titles out of 15 were not retrieved on average (see table 6). A *t* test comparison of the averages obtained reveals that the differences observed are significant ($p < .0005$). The authors have, therefore, concluded that titles that are considered problematic because of the presence of a word erroneously treated as an initial article by the detection algorithm are more difficult to retrieve.

Question 6

Are monographic titles containing a non-article word homograph to an initial article usually harder to retrieve than other titles, in terms of time and effort?

The time measured was from the moment the user pressed a key to launch his query and the moment the record displayed on screen. The time for keying-in the query was not counted, since titles can vary in length. System response time was noted to be minimal at all times; the time measured here corresponded mainly to the time it took for the user to recognize the correct record and select it. Titles that were not found were excluded from the average.

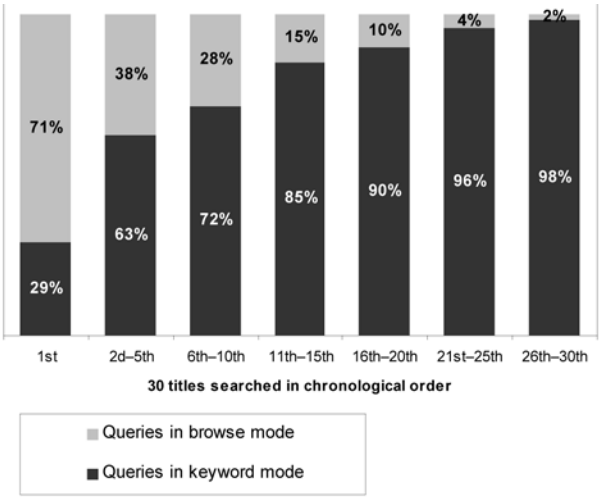

Figure 4. Search mode used for the first query issued for each title in chronological order

Table 6. Number of titles found on average

	Average		Standard deviation
	Number	%	
Non-problematic titles ($N = 15$)	14.7	97.88	0.56
Problematic titles ($N = 15$)	13.0	86.7	1.69

Analysis of the time necessary to find the records reveals that problematic titles have taken much more time on average (see table 7). Finding the titles containing an initial article took more time, compared to those without such an article, but the statistical analysis reveals that this difference is not significant ($p = .062$). Statistical analysis of problematic titles compared with the other two title groups combined shows that the differences observed are, in this case, meaningful ($p < .0005$).

In this study, in addition to time, two measurements were used to represent the effort invested by the participants to locate a title: the mean number of queries used and the mean size of the retrieved sets (for the queries issued in keywords mode) were measured (see table 7). On average, more queries were necessary to find the titles containing an initial article than to find those that did not, but the statistical analysis shows that the difference is non-significant ($p = .489$). Conversely, the statistical analysis comparing problematic titles with titles of the two other groups combined shows that the differences are significant ($p < .0005$).

On average, the users retrieved slightly larger sets (offering less precision) to find titles containing an initial article than to find those without an article. The statistical analysis, however, reveals that the difference is non-significant ($p = .763$) (see table 7). Nonetheless, the statistical analysis of the problematic titles compared with each of the 2 other title groups (with and without article) shows that the differences are significant in this case ($p < .005$ and $p < .011$ respectively). These two measurements, therefore, indicate that (on average) more time and more effort (number of queries and size of sets to browse) were necessary to locate a problematic title.

Table 7. Time and effort to find a title

	Mean time (in seconds) to find a title		Mean number of queries per title		Mean size of the sets per title	
	Average	**St. dev.**	**Average**	**St. dev.**	**Average**	**St. dev.**
Titles without an article ($N = 5$)	5.58	6.28	1.18	0.37	3.11	4.56
Titles with an initial article ($N = 10$)	9.32	5.99	1.25	0.28	3.31	3.08
Problematic titles ($N = 15$)	19.76	10.14	1.66	0.33	54.85	77.57

Conclusions and Future Research

This exploratory study has supplied empirical data that are valuable if there is to be a better understanding of the phenomena of title retrieval with regard to initial articles in automated information retrieval systems. While preparing for this project, the authors' review of existing literature revealed that, while the problem is well documented on the data representation side, it is seldom examined on the retrieval side. Title searches are still one of the most, if not the most, common search type in library catalogs. It is, therefore, desirable that they be made more effective and more efficient. The results of this study show that applying an initial article detection algorithm to queries negatively affects only a small proportion of records (less than 0.4 percent of all bibliographic records in Atrium). This proportion may seem so small as to be negligible, but in reality it is

some 5,000 records that are thus less visible when a browse search is performed in the title index. This is not a negligible number if the acquisition and processing costs of these items are considered.

Out of the 40 articles that were examined in this study, the English article "a" is responsible for two thirds of the problems encountered. It seems that a large proportion of the problems could be solved by merely removing this article from the exclusion list. Moreover, eliminating the exclusion list altogether would eliminate the retrieval problems from the start. One may argue, however, that completely eliminating the exclusion list might introduce other problems in the searches; specifically if users inadvertently or unknowingly include the initial articles in their queries when doing a title search. The authors' log analysis revealed that, in browse searches, only one third of the queries for titles that start with a definite or an indefinite article did not contain an article. It was observed that in about 2 cases out of 3, users kept the initial article in their query, even when these articles were ignored in the indexing process. At the time of printed catalogs (index cards, for instance), removing initial articles was mandatory in order to locate a title at the right place. End users no longer seem instinctively to remove the initial articles from their queries. In this computerized world many queries likely are generated by using the cut and paste function, which may partially explain why initial articles are retained in the queries. End users also seem to believe that keeping or omitting the initial article will have no effect on retrieval because that is the case for most of the general search engines on the Web. Using automatic detection algorithms could, therefore, be regarded as a way to adapt to the changing search behaviors of end users. Regretfully these algorithms, as shown in this study, are not terribly sophisticated, and have major caveats, especially in multilingual environments.

The results of this study indicate that applying an exclusion list has a negative effect on a small but not negligible proportion of records, from the point of view of their visibility in the title index. The authors have observed that the success rate in finding these titles is significantly lower than the success rate in finding the other titles, since the problematic titles cannot be retrieved using the browse mode. The keywords mode is a good substitute in many cases, but retrieval may become tricky or simply impossible for short titles and for keywords with a high occurrence in the catalog. The authors' analysis has revealed also that retrieving problematic titles is more difficult in terms of time and effort needed. On average, more queries were necessary to retrieve any one title, and the precision of the sets retrieved was lower when a keywords search was used, because sets retrieved were generally larger. This analysis confirms that both search modes, browse and keywords, are, as mentioned by Frost and colleagues in their study on browse and search patterns, useful and necessary.[17] When one of them is not functional, the success and efficiency rates of the search are affected. Initial article detection algorithms can be useful if users keep the articles in their queries, but they slow down the search in browse mode for certain titles, and this seems to have negative repercussions on the retrieval of these titles. Therefore, the authors recommend that an alternative method be developed to eliminate this problem. One possible option is to initiate some form of interaction with the end user. For example, following a search on "ti=UN resolution 435" the system could provide a feedback such as "Do you want to search *un resolution 435* or *resolution 435*?" instead of keeping the whole procedure completely invisible.

This research could be extended to other catalogs in the future or to other environments, or be used to measure the impact on the user in a real research situation. The results of this study can be used for developing better retrieval algorithms in order to improve title searching in multilingual information systems. Since library catalogs are the entry point to many document collections, configuring the systems to maximize retrieval efficiency and success rate and, therefore, to improve customer satisfaction is essential.

The authors advocate against using detection algorithms based solely on exclusion lists since, in many cases, these mechanisms appear detrimental to end users title searches. It is preferable to include clear and highly visible instructions in the search interface, instructing end users to omit the initial article in their search. Regrettably, users are often not adequately trained or properly instructed for information retrieval in library catalogs. Before computerized catalogs existed, it was assumed that users knew that they had to remove the initial articles to find a title. Why should it be different today? It is a simple rule to learn. An alternate solution to using exclusion lists would be to ease the filing rules and allow a title containing an initial article to be filed under the article and also under the first significant word. This option, for entries starting with "The," is recognized as a "win-win" solution by Browne.[18] This indexing method, suggested by Nielsen and Pyle in 1995 and again, more recently, by Corrado, is already applied in some library catalogs.[19] However, to be completely efficient, it would require recording initial articles in all MARC fields where they appear, including fields 130, 240, 246, 247, 700/710/711 (subfield t) and 730. The implementation of the non-filing control characters within the data, as proposed in *Discussion Paper 2002-DP05*, would certainly make this possible.[20] The double entry solution may bulk up the title index a little but this technique makes the use of initial article detection algorithms unnecessary, because finding the titles either way (with or without the article in the query) becomes possible.

Because there are apparent advantages and disadvantages of using initial article detection algorithms, the dilemma between keeping and eliminating the exclusion list remains. Either the exclusion list is kept, allowing the cor-

rect redirection of queries containing an initial article, or it is eliminated to avoid losing track of titles beginning with a non-article word that is homographic to an article in the list. The authors hope the empirical data provided in this paper will help system designers and managers make better decisions regarding the use of such features in their catalogs.

References

1. Charles P. Bourne, "Initial Article Filing in Computer-Based Book Catalogs: Techniques, Problems, and Article Frequency," *Journal of Library Automation* 8, no. 3 (1975): 221–47; American Library Association, *ALA Filing Rules,* rule 4.2. (Chicago: ALA, 1980).
2. Library of Congress, *Discussion Paper No. 102* (1997). www.loc.gov/marc/marbi/dp/dp102.html (accessed Oct. 27, 2006).
3. Library of Congress, *MARC Standards*. www.loc.gov/marc (accessed Oct. 27, 2006).
4. Library of Congress, *Proposal 98-16R: Non-filing Characters in all MARC Formats* (Dec. 11, 1998). www.loc.gov/marc/marbi/1998/98-16r.html (accessed Oct. 27, 2006).
5. Library of Congress, *Discussion Paper No. 118: Non-filing Characters in MARC 21 Using the Control Character Technique* (June 1, 1999). www.loc.gov/marc/marbi/dp/dp118.html (accessed Oct. 27, 2006); Library of Congress, *Discussion Paper 2002-DP05: Guidelines for the Non-filing Control Character Technique in the MARC 21 Formats* (Dec. 18, 2001). www.loc.gov/marc/marbi/2002/2002-dp05.html (accessed Oct. 27, 2006); Library of Congress, *Network Development and MARC Standards Office Guidelines for the Non-Sorting Control Character Technique* (2004). www.loc.gov/marc/nonsorting.html (accessed Oct. 27, 2006).
6. Corey Seeman, "RE: Skipping Initial Articles," e-mail to the Innovative User's Group, June 11, 2002. http://innovativeusers.org/list/archives/2002/msg02463.html (accessed Oct. 27, 2006).
7. Bourne, "Initial Article Filing."
8. Arlene G. Taylor, *The Organization of Information,* 2nd ed. (Westport, Conn.: Libraries Unlimited, 2004), 121.
9. Bourne, "Initial Article Filing"; R. Nielsen and J. Pyle, "Lost Articles: Filing Problems with Initial Articles in Databases," *Library Resources & Technical Services* 39, no. 3 (1995): 221–22; Seeman, "RE: Skipping Initial Articles"; Min-Yen Kan and Danny C. C. Poo, "Detecting and Supporting Known Item Queries in Online Public Access Catalogs," in *International Conference on Digital Libraries Archive. Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries,* eds. M. Marlino, T. Sumner, and F. M. Shipman III, 91–99 (Denver: ACM, 2005).
10. Marianne Broadbent, "Who Wins? Who Loses? User Success and Failure in the State Library of Victoria," *Australian Academic and Research Libraries* 15, no. 2 (1984): 65–80.
11. Ray Larson, "The Decline of Subject Searching: Long-Term Trends and Pattern of Index Use in an Online Catalog," *Journal of the American Society for Information Science* 42, no. 3 (1991): 197–215.
12. Neil K. Kaske, "The Variability and Intensity over Time of Subject Searching in an Online Public Access Catalog," *Information Technology and Libraries* 7, no. 3 (1988): 273–87.
13. Hitoshi Matsushita, "Information Access Behavior of Music Researchers and Music Materials," *Journal of Information Science and Technology Association* (*Joho no Kagaku to Gijutsu*) 54, no. 7 (2004): 363–70.
14. *Règles de catalogage anglo-Américaines,* 2e éd., rév. de 1998, modifications de 2001–2005, élaborées sous la direction de: the Joint Steering Committee for Revision of AACR composé de délégués de the American Library Association . . . [et al.]; coordination de la version française, Pierre Manseau (Montréal: Éditions ASTED, 2005).
15. Kan and Poo, "Detecting and Supporting Known Item Queries."
16. Dennis Halcoussis et al., "An Empirical Analysis of Web Catalog User Experiences," *Information Technology and Libraries* 21, no. 4 (2002):148–57.
17. C. Olivia Frost et al., "Browse and Search Patterns in a Digital Image Database," *Information Retrieval* 1, no. 4 (2000): 287–313.
18. Glenda Browne, "The Definite Article: Acknowledging 'The' in Index Entries," *Indexer* 22, no. 3 (2001): 119–22.
19. Nielsen and Pyle, "Lost Articles"; Edward M. Corrado, "Initial Articles in Library Catalog Title Searches: An Impediment to Information Retrieval," in Andrew Grove, ed., *Proceedings of the American Society for Information Science and Technology* 43 (2006). http://dlist.sir.arizona.edu/1657 (accessed Nov. 21, 2006).
20. Library of Congress, *Discussion Paper 2002-DP05: Guidelines for the Non-filing Control Character Technique*.

Appendix. Example of One of the 24 Lists Given to Participants for the Retrieval Task

Nom : John Doe

Adresse : 123 Nowhere Drive, Montréal (Q.C)

Tél. : 514-555-1212 Courriel : john.doe@lalala.com

Âge : 17 Sexe : ☑ M ❑ F Langue maternelle : Français

Département : Sciences pures Cycle : Reg. Année : 2e

Liste des titres à rechercher

Liste 06

	Titre à chercher	Cote	
1.	Always a loser	PR9387.9 .A743 A75 1981	✓
2.	The most agreeable vice	huxley .H893 M685 1938 pam	✓
3.	A very profitable war	PQ2664 .A417 D4713 1994	✓
4.	The night after Christmas	JUV FIC S847ni	✓
5.	Jamais contente	PS9507 .U26 J3	✓
6.	À Vancouver sur le pouce : récit de voyage d'un étudiant à travers l'Amérique	HC D592.6a.v	✓
7.	A rire et à mourir : récits, paraboles et chansons du lointain pays, croquis, crocs, pointes très sèches, échos de la grande mort, cris et scies hors d'haleine	PQ2663 .H326 A2 1983	✓
8.	À la découverte de Shakespeare	PR2947 .D5 L38 T.1	✓
9.	Le beau baiser : roman	PQ2607 .E24 B4 1933	✓
10.	Sur le chemin Craig	PS9511 .E774 S87 1982	✓
11.	A la recherche de l'ambre baltique : l'expédition d'un chevalier romain sous Neron	HF377 .K65 1981	✓
12.	Les plus beaux de nos jours	PQ2601 .R55 P5 1944	✓

	Titre à chercher	Cote	
13.	A la rencontre de Marie Moret et de l'éducation au XIX^{e} siècle	LB675 .M67Z A43 1998	✓
14.	Un jour, je te tuerai : roman	PQ2664 .U68 J68 1999	✓
15.	Blessings of the table : mealtime prayers throughout the year	BV283 .G7 D39 1994	✗
16.	À la découverte des Îles du Saint-Laurent : de Cataracoui à Anticosti	F1050 .G83 2003	✓
17.	A la redécouverte de Patrice Emery Lumumba	DT658.2 .L85 A3 1996	✓
18.	À micro ouvert	PN1991.77 .I4 B65 1988	✓
19.	À la recherche de légitimités chrétiennes : représentations de l'espace et du temps dans l'Espagne médiévale, IX^{e}-XIII^{e} siècle	BR1024 .A253 2003	✓
20.	La mer au large : roman	PQ2670 .0745 M47 1987	✓
21.	A travers chants : études musicales, adorations boutades et critiques	ML60 .B45	✓
22.	À travers le verre du moyen âge à la renaissance	NK5108 .M87 ROMU	✓
23.	À outrance : de cent à zero et le contraire pour flûte et viola	MZ91 .P346 AZ 2000	✓
24.	Les femmes antillaises	HQ1525.85 .B4	✓
25.	Une mort très douce : récit	PQ2603 .E36Z M6 1986	✓
26.	Out after dark	PR1062 .E68 Z468 1989	✓
27.	A Istanbul et en Cappadoce	DR718 .W55 1986	✓
28.	A jamais la Bretagne	DC611 .B847 C483 1998	✓
29.	A B C on French Canada	cap 00714	✓
30.	An apple a day : a holistic health primer	R733 .B67	✓