Kristina M. Irvin. Comparing Information Retrieval Effectiveness of Different Metadata Generation Methods. A Master's paper for the M.S. in I.S. degree. April, 2003. 28 pages. Advisor: Jane Greenberg.

This study describes an information retrieval experiment comparing the retrieval effectiveness (recall and precision) for queries run against professionally and automatically generated metadata records. The metadata records represented web pages from the National Institute of Environmental Health Sciences. The results of 10 queries were analyzed in terms of recall and precision for this small-scale study. The results of the study suggest that professionally generated metadata records are not significantly better in terms of information retrieval effectiveness than automatically generated metadata records.

Headings:

Metadata

Professionally generated Metadata

Automatically generated Metadata

Information Retrieval

COMPARING INFORMATION RETRIEVAL EFFECTIVENESS OF DIFFERENT METADATA GENERATION METHODS

by
Kristina M. Irvin

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Information Science.

Chapel Hill, North Carolina

April, 2003

Approved by:

____

Advisor

Table of Contents

Introduction

In today's world, it is well known that the sheer amount of information available is overwhelming. With the introduction of the World Wide Web (web), this problem has become even greater. In 1999, the web was estimated to contain 4.6 million unique sites, and it is ever increasing ("On the Size", 2001). In 2001, the web was estimated to contain 8.4 million unique sites ("On the Size", 2001), quite an increase for just 2 years. A persistent problem through the ever expanding amount of information available on the web is the retrieval of that information. Anyone who has searched the web is likely aware of how difficult and frustrating it can be to locate a document on the web. Having enormous amounts of information available on the web offers new opportunities, but if there is not an effective way to retrieve this information, the information becomes less useful. One of the ways that information professionals are attempting to improve information retrieval for the web is through the introduction and application of metadata standards. Metadata, as it is so often defined, is simply, "data about data". However, a more detailed definition for metadata is "structured data about an object that supports functions associated with the designated object" (Greenberg, 2002). Following this definition, a key function that metadata supports is information retrieval.

The current norm for storing metadata for web pages is in the form of HTML Meta tags. A number of studies have been conducted on the existence of HTML Meta tags in web sites, for example "The New Meta Tags Are Coming – Or Are They?" study conducted in 1997. In general, web pages do not contain Meta tags or they only contain keyword and description tags. One problem with the Meta tags is that they lack consistency and often are abused by the author to increase the chances of surfacing the web pages in web searches (Sullivan, 2002). A study by Drott (2002) of corporate web site use of meta tags showed that only slightly more than a third of the web sites were using Meta tags. Drott determined that if web sites increased their usage of Meta tags, the overall index coverage of the web could be improved.

While metadata could greatly improve the retrieval of information on the web (Marchiori, 1998), the issue of actually generating the metadata still remains. Because there are so many web pages already in existence without metadata, it would be an enormous undertaking to generate the metadata for all of these web sites, not only in time but also in money invested. Additionally, there are currently no requirements for people to include metadata so people do not generally take the time to generate metadata (Marchiori, 1998). For those who do generate their own metadata, there are issues of consistency and reliability (Jenkins, Jackson, Burden, & Wallis, 1999). Since there is not a single, agreed-upon standard for metadata for web content, it is not surprising that there are consistency and reliability issues. Additionally, the web page creators who generate their own metadata have not been trained as professionals. In other words, they are not likely to be aware of standards that they could follow to aid the process.

There are several ways to generate metadata including author generated, professionally generated and automatically generated (Greenberg, 2002). Considering the expense and time investment for metadata generation for the first two methods, automated metadata generation is appealing. Additionally, automated metadata generation can assist with the problems of unreliability and inconsistency (Jenkins et al., 1999). However, it is important that the automated metadata be as effective in facilitating information retrieval as the other two methods of generation. The proposed research for this paper will address the effectiveness of automated metadata generation for information retrieval by comparing it to professionally generated metadata.

This project is an extension of a larger research project, “Optimizing Metadata Creation: A Model for Integrating Human Automatic Processes.” This project addresses the questions of “who” and “how” for metadata creation by examining the quality of metadata created by resource authors and metadata professionals for the National Institute of Environmental Health Sciences (NIEHS). (“Metadata Generation Research Project”, 2002)

Literature Review

Compared to other topics in information retrieval such as indexing, query expansion, and searching algorithms, the application of metadata for digital resources as a way to improve information retrieval is still a fairly new concept. Much of the efforts for metadata thus far have been focused on developing standards and determining the best applications of metadata. As a result, not much effort has been focused on the best ways to generate the actual metadata (Greenberg, 2002).

Important research preceding automatic metadata generation is progress made in automatic indexing. Automatic indexing involves extracting key terms and subject areas, which are forms of metadata, for resources so that they may be located in information retrieval. The purpose of indexing is to point to or indicate the “content, meaning, purpose, and features of messages, texts, and documents" (Anderson & Perez-Carballo, 2001). Metadata shares similar goals in facilitating information retrieval.

The longstanding debate for human indexing versus automatic indexing relates to the quality of the index produced. In general, it is believed that human indexers are better than machines because humans have cognitive processing. However, automatic indexing is gaining popularity as research shows that it can be as equally effective as human indexing (Anderson & Perez-Carballo, 2001) for information retrieval.

Automatic indexing offers several advantages over human indexing. Automatic indexing is more cost effective and continues to become even more so with technological advances while the cost of human indexing is rising (Anderson & Perez-Carballo, 2001). Automatic indexing can also be applied to extensive collections of resources like the web "where the volume of texts and constant change, both in individual texts and in the comparison of the collection as a whole, makes human indexing impractical, if not impossible" (Anderson & Perez-Carballo, 2001). This relates to the problems mentioned earlier about the rather large and expensive undertaking it would be for humans (authors and/or professional metadata creators) to generate metadata for the existing web. Additionally automatic indexing offers consistency. Human indexers interpret the text and thus are vulnerable to subjectivity due to their own experiences, culture, and even prejudices (Anderson & Perez-Carballo, 2001). Automatic indexers, on the other hand, use the same algorithm every time a document is examined, and thus, always produce consistent, unbiased results. This offers a potential advantage for information retrieval since the way in which a document is indexed can be understood by the user thus allowing the user to know how to search for documents.

This is not to say that automatic indexing is considered to be better than human indexing. Both automatic indexing and human indexing have advantages and disadvantages. Automatic indexing is attempting to mimic the human indexing process and until it is more successful across multiple domains, automatic processing will not be considered superior. Despite limitations, automatic indexing provides several advantages that may be important to the web such as saving the time and expense that would be required for human indexing.

Another area of research relevant to issues of automatic metadata generation, and overlapping with automatic indexing, is natural language processing which has also achieved significant progress in the last several decades and continues to improve. In a study by Wacholder, Evans, and Klavans, (2001) they found that "natural language processing techniques have reached the point of being able to reliably identify terms that are coherent enough to merit inclusion in a dynamic text browser." With all of the recent success in automatic indexing and improved natural language techniques, it seems plausible that automated metadata generation could reap similar benefits.

In fact, there has been some success with automated generation of metadata beyond just subjects and key terms in automated indexing. Several tools are being developed for the automated generation of metadata and are able to extract certain types of metadata well such as title, author, and subjects. In a study by Jenkins et al. (1999), they had some success with an automatic classifier that classified HTML documents according to Dewey Decimal Classifications and used the classifier to extract other metadata. Assignment of a classification is another form of metadata in addition to subject metadata. An advantage to this tool is that it will work regardless of when a web page was created or with what editing tool created it (Jenkins, 1999). In another study, Giuffrida, Shek, and Yang (2000) were able to automatically extract metadata from scientific conference paper PostScript files with promising success. They created a rule-based tool that extracts metadata based on the structure of the PostScript document. Their tool extracted the title with 92% accuracy, author(s) with 87% accuracy, affiliation(s) with 75% accuracy, author-affiliations with 71% accuracy and table of contents with 76% accuracy. While their tool still requires work in extracting sentences, paragraphs, and other phrases, it offers great promise for automated metadata extraction tools. The tool could be extended to work with other structured file-types such as HTML (Giuffrida, Shek, & Yang, 2000).

Additionally, there are a number of automated metadata generation tools that generate metadata for web pages. These tools accept a URL as input and produce metadata elements that can then be manually edited for greater accuracy. These tools are successful in extracting the title, content-type, and subject areas for web pages, but lack in extraction of other types of elements such as author. Two such tools include Klarity and DC-Dot, metadata extraction tools created for generating Dublin Core metadata. The Dublin Core metadata standard is a standard for web resources that includes 15 elements (Dublin Core Metadata Element Set, Version 1.1, 2003). For more information on the Dublin Core, please reference the Dublin Core Metadata Initiative web site, http://www.dublincore.org.

Automated generation of metadata appears promising. However, it is useful to note that the success of automated generation is limited by the lack of conformance to a single metadata standard. Lawrence and Giles (1999) found that only 0.3% of web pages contained Dublin Core metadata in a study of 2500 web servers. Computers work with prescribed algorithms in order to produce information. Each automatic generator must, therefore, be schema-specific (Greenberg, 2002). Manual indexers have an advantage in this area as they are able to learn several different schemas and adapt easily to changes in schemas. The schema-specific limitation also means that automatic generators are likely to be created for simple schemas and more complex schemas are likely to be ignored (Greenberg, 2002). Another challenge to automated generation of metadata is the need for people to trust the metadata that is generated. Without this trust, automated generation cannot be successful.

A way to increase the trust of automated metadata generation is to show that information retrieval based on automatically generated metadata is as effective as information retrieval based on professionally generated metadata. The research on automatic indexing, natural language processing, and automatic classification (all forms of metadata) holds promising implications for the success of automated metadata generation that goes beyond subject and classification identification. As the automatic indexing techniques are shown to be as effective as human indexing, then automated metadata generation, too, could be as effective as professionally generated metadata. The research for this proposal is motivated by the several noted advantages that automated generation of metadata has to offer including faster creation of the metadata, lower cost, and improved compliance with standards. While it is known that automatically generated metadata may have poorer quality than professionally generated metadata, a key test beyond quality is to determine if information retrieval based on automatically generated metadata facilitates information retrieval as effectively as professionally generated metadata. Based on these ideas, the proposed research will examine the effectiveness of information retrieval based on automatically generated metadata by comparing it to the effectiveness of information retrieval based on professionally generated metadata. This comparison will be based largely on the average recall and precision scores.

Objectives

The purpose of this research project is to examine the difference between automatically generated metadata and professionally generated metadata with respect to information retrieval effectiveness. The goal of this research is to determine if automatically generated metadata can sufficiently be used in place of professionally generated metadata in information retrieval without a loss of effectiveness.

In other words:

- Does automatically generated metadata perform similarly to professionally generated metadata in terms of information retrieval effectiveness?

Methodology

The primary method used is an experiment that tests queries against metadata records generated by both automated means and professional cataloguers. The metadata records are document surrogates for a set of web pages. The methodology for this project has been developed by following the guidelines outlined in “The Pragmatics of Information Retrieval Experimentation, Revisited” by Jean Tague-Sutcliffe (1992).

The metadata records for this project have already been generated as part of the “Optimizing Metadata Creation: A Model for Integrating Human Automatic Processes

Project" ("Metadata Generation Project", 2002), mentioned earlier. The metadata records included in the testing sample were created from a set of web pages that were produced by scientists at the National Institute of Environmental Health Sciences (NIEHS). The web page sample falls into the following categories: organizational information (14 web pages), personnel information (4 web pages), products/services (2 web pages), publications (2 web pages), research information (11 web pages) and educational (1 web page). The majority of the web pages are organizational and research information web pages. For the purpose of this study the web pages were also categorized into two categories based on the amount of textual content on the pages. Web pages were either grouped as containing predominantly textual information in paragraph form or as table of content pages, containing mostly hyperlinks to other information. The web pages include a range of material focused around environmental and health science issues and research. Topics include toxicology, proteins, genetics, safety, mutation, and reproduction.

The first set of metadata records were generated by three professional cataloguers who examined the 34 web pages and produced a metadata record for each page (See Appendix A for a sample record). These metadata records were created according to the NIEHS application profile (Harper, Greenberg, Robertson, & Leadem, 2002), which is based on the Dublin Core metadata standard for web resources ("Dublin Core Metadata Initiative", 2003). The second set of metadata records was produced by DC-Dot and Klarity, Dublin Core generators and editors. These programs accept a URL and generate the metadata in Dublin Core format ("DC-Dot", 2002, "What is Klarity", 2003). Therefore, these records were created according to the Dublin Core metadata standard rather than the NIEHS application profile. Both sets of metadata, the automatically generated records and the professionally generated records have been incorporated into a Microsoft Access database against which the queries for this experiment will be executed.

In order to compare the effectiveness of information retrieval for the different types of metadata, a set of artificial queries was generated by an NIEHS librarian. The librarian was given the URLs for the 34 web pages as well as the categorical breakdown of the pages into the six categories mentioned earlier, e.g. organizational, research information, etc. Based on this information, she generated 20 queries in the form of questions that the NIEHS library could likely receive from the educated lay person or member of the public, regarding the type of research conducted at NIEHS. The list of questions used in this experiment is included in Appendix B. From the list of 20 questions, 10 were randomly selected for this experiment. The selected queries were then converted into SQL statements that could be executed against the Microsoft Access database. To convert the questions to queries, the distinguishing keywords of the question were selected and the query was constructed as follows:

> "select url from elements where (content like '*keyword1*' or content like '*keyword2*' ……"

For example, the question,

> Is NIEHS conducting any research in the area of HIV-related proteins?

was converted into the following SQL query:

```
select distinct url from  elements where (content like '*HIV*' or content like
'*protein*')
```

The keywords HIV and protein were extracted from the original question. NIEHS was not chosen as a keyword since all of the web pages were created at the NIEHS.

The keywords were linked by "or" as opposed to "and" due to the nature of the metadata records and this is often the default for most information retrieval systems. A metadata record often distributes keywords into separate metadata elements. Therefore, the use of "and" would most often return no results which would not be accurate.

Second, the relevance for the web pages represented in the metadata records for each query was pre-determined. Three evaluators with a background in health science and environmental issues pre-determined the relevance of the web pages to the queries. Two of the evaluators were graduate students at the University of North Carolina at Chapel Hill and currently work in the Health Sciences library on campus. The third evaluator was in the last semester for completing a B.S. in Animal Science at North Carolina State University. This evaluator also worked in the VetMed Department at the NIEHS during the summer of 2002. The evaluators examined each web page in reference to each query. Each evaluator gave each web page a relevance rating of Y(Yes) or N(No) for each query. The response majority was used as the final relevance judgment for the web page-query combination. For example, if two students indicated the web page was not relevant and one student indicated the page as relevant, the web page was recorded as not relevant for that query. The final relevance judgments are included in Appendix C.

After the relevance judgments were determined, the queries were executed against the two metadata record sets in the Microsoft Access database. Each query was executed twice. The first execution of the query contained a “where” clause that limited the search to automatically generated metadata records. The second query contained a “where” clause that limited the search to professionally generated metadata records.

Each query returned a set of URLs. The number of URLs returned and the actual URLs were recorded.

The independent variable for this study is the method of metadata generation for the document surrogates to represent the web pages. There are two methods: automatically generated and professionally generated. The dependent variables are the recall and precision scores. Recall is the proportion of relevant documents retrieved out of the total number of expected relevant documents in the entire collection. Precision is the ratio of relevant documents retrieved to the number of documents retrieved.

Results

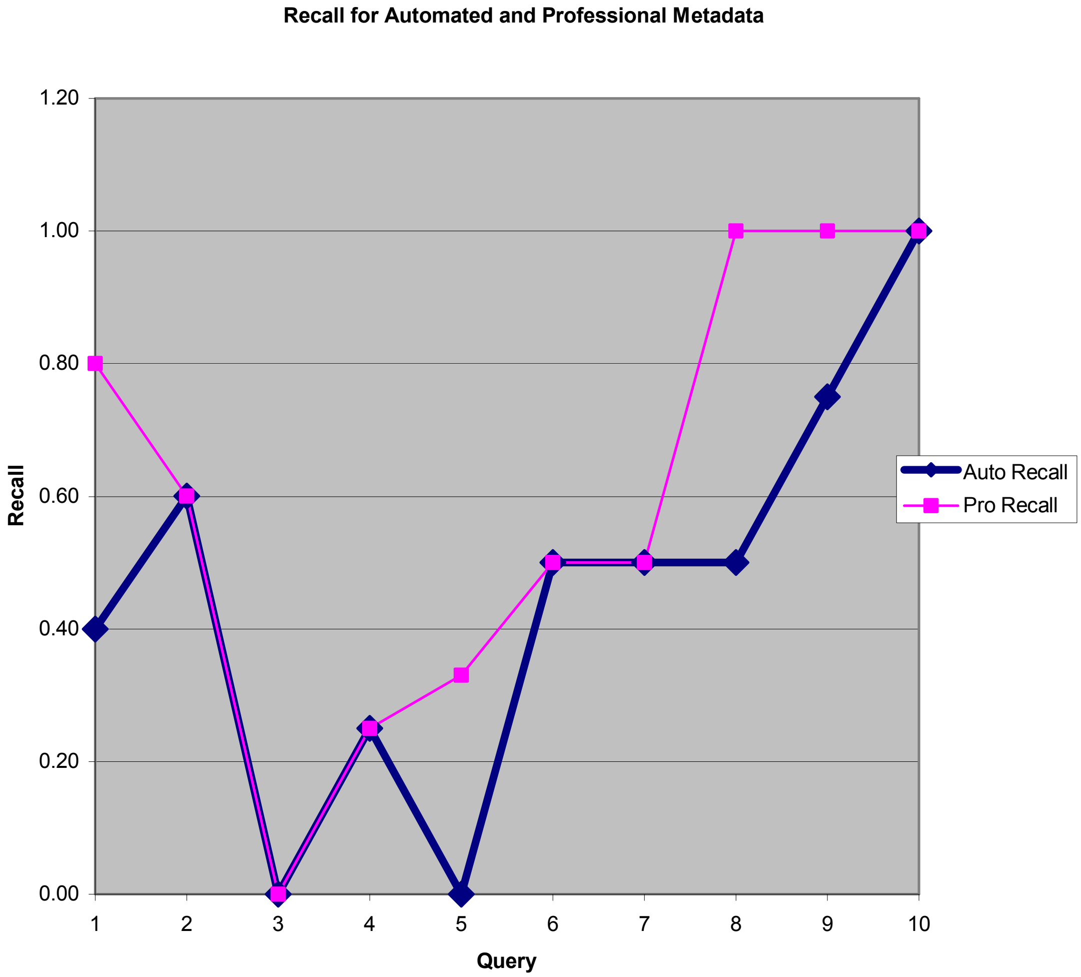
The relevance and precision scores for each query for each metadata treatment were calculated based on the pre-determined relevance judgments. Because of the relatively small size of the collection, true recall scores were able to be calculated. Therefore, the recall scores were calculated as the number of relevant documents retrieved divided by the total number of relevant documents in the database. Precision was calculated as the total number of relevant documents retrieved divided by the number of documents retrieved. The results are summarized below:

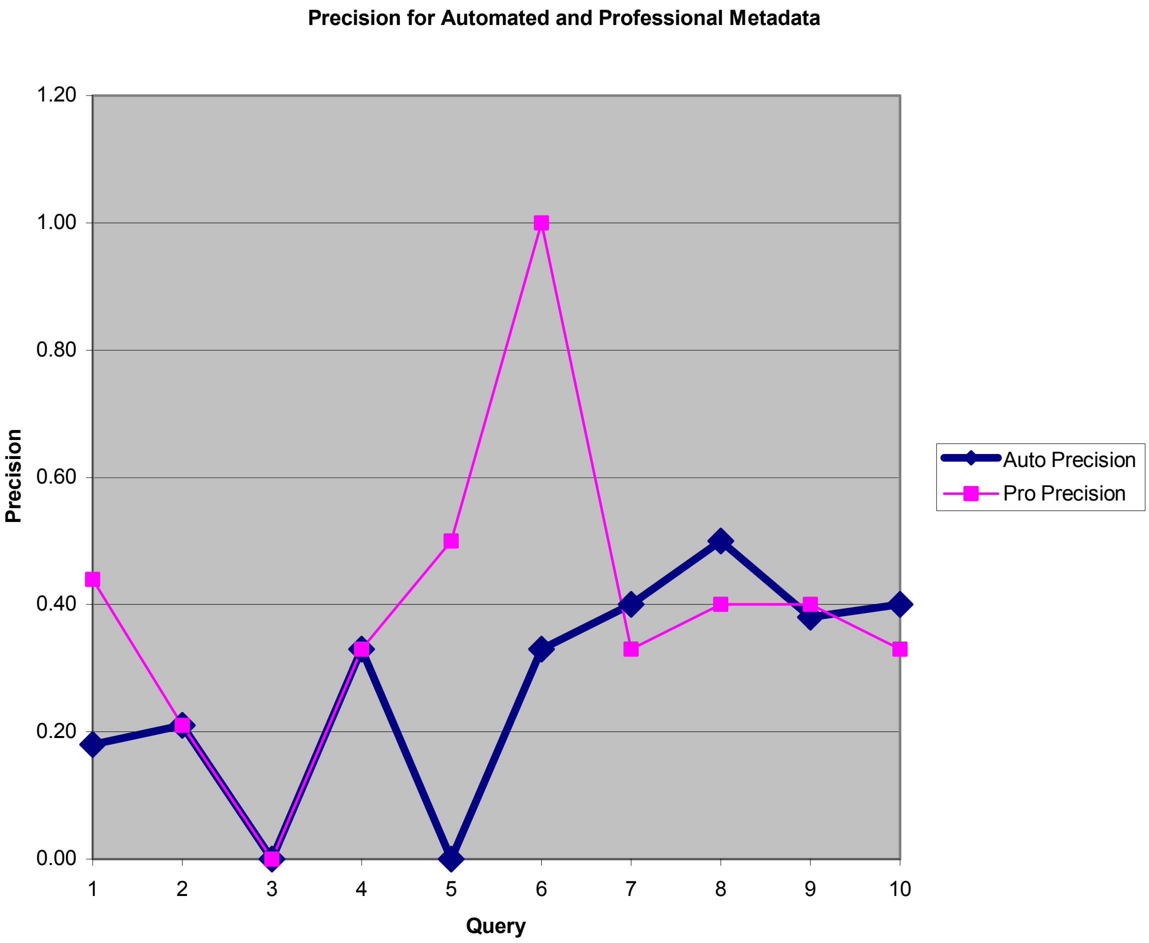
Table I. Recall and Precision Results for Each Query

QUERY	Auto Recall	Pro Recall	Auto Precision	Pro Precision
Q1	0.40	0.80	0.18	0.44
Q2	0.60	0.60	0.21	0.21
Q3	0.00	0.00	0.00	0.00
Q4	0.25	0.25	0.33	0.33
Q5	0.00	0.33	0.00	0.50
Q6	0.50	0.50	0.33	1.00
Q7	0.50	0.50	0.40	0.33
Q8	0.50	1.00	0.50	0.40
Q9	0.75	1.00	0.38	0.40
Q10	1.00	1.00	0.40	0.33
MEAN Values	**.45**	**.60**	**.27**	**.39**

Note: Auto=Automatic and Pro=Professional

The following graphs illustrate the variation of the recall and precision scores between the two treatments.





As can be observed in the graphs above, the professionally generated metadata did not always have higher scores than the automatically generated metadata. In some instances, the automatically generated metadata scores are higher. Additionally, sometimes the recall and precision scores are identical.

To analyze the data, descriptive statistics were also applied. According to Tague-Sutcliffe (1992), for measurements, the mean should be calculated for central tendency and the standard deviation should be calculated for variation. Therefore, these values were calculated for both the recall and precision scores for each treatment.

These values are displayed below:

Table II. Summary of Recall Scores

Treatment	Mean	Standard Deviation	Range
Automated	.45	.29	.16-.74
Professional	.60	.33	.27-.93

Table III. Summary of Precision Scores

Treatment	Mean	Standard Deviation	Range
Automated	.27	.16	.11-.43
Professional	.39	.29	.10-.68

Ideally, both the recall and precision scores will be close to one, but this is rare. Often, precision and recall values inversely affect each other (Salton, 1986). However, in this study, the individual scores themselves are not being examined, but rather the scores are being compared to each other. The mean recall and the mean precision scores are higher for professionally generated metadata than automatically generated metadata. The mean recall score is .15 higher and the mean precision score is .12 higher. The differences between the recall and precision scores are very close. Taking into consideration the standard deviation, the range for the recall for automatically generated metadata is .16 to .74 while the range for professionally generated metadata is .27 to .93. This means a range of .27 to .74 is possible for the average recall of both the automatically and professionally generated metadata. The range for precision for automatically generated metadata is .11 to .43 while the range for professionally generated metadata is .10 to .68. This means a range of .11 to .43 is possible for the average precision of both the automatically and professionally generated metadata. The average precision range for automatically generated metadata is actually included in the range for the professionally generated metadata.

Discussion

The results for this study indicate that the professionally generated metadata records are not necessarily superior to the automatically generated metadata records in terms of information retrieval effectiveness. The recall and precision scores suggest that the performance of the two generation methods is somewhat similar. Based on the results from this study, as the standard deviations for the recall and precision scores are quite large, it seems that one treatment cannot be considered outstanding in performance to the other. This is, however, a relatively small sample consisting of only 34 web pages and it focuses on a single domain which may affect the results. Another issue to take into consideration when interpreting the results is that the professional cataloguers were following the NIEHS application profile while the automated generators were following only the Dublin Core metadata standard and did not include all the elements. Because the application profile provides slightly more content for the metadata, this may have affected the results. If the automated generators were customized to generate metadata according to the NIEHS application profile, the recall and precision scores may have been more similar between the automatically generated metadata and the professionally generated metadata.

In addition to examining the recall and precision scores, the web pages that were judged relevant for both the automatically generated metadata records and the professionally generated metadata records were placed into one of two categories based on the amount of textual content the web page contained. The web pages were categorized as either containing significant textual content such as paragraphs or categorized as table of contents web pages, mostly containing hyperlinks to other information. Examination of these two categories showed that roughly 50 percent of the web pages judged relevant have significant textual content and roughly 50 percent do not for both types of metadata generation. Therefore, the textual content does not seem to factor into the effectiveness of the metadata record produced for both automatically generated metadata as well as professionally generated metadata.

Conclusion

This study compared the information retrieval effectiveness of professionally generated metadata records and automatically generated metadata records for a small sample of web pages from the NIEHS. In this study, the results show that professionally generated metadata is not necessary to yield significantly better results in terms of recall and precision in information retrieval. These results indicate that taking into consideration the expense in terms of time and human resources for professionally generated metadata, automatically generated metadata could be used instead without a significant decrease in information retrieval effectiveness, at least for small collections of web resources. This research is important because it suggests that more metadata can be generated for the World Wide Web with fewer resources by using automated tools. As the use of metadata has been shown to improve retrieval effectiveness for web resources, this may alleviate some of the information retrieval issues currently experienced with the World Wide Web.

Further Research

This research examined the recall and precision scores for metadata records as document surrogates for web pages at the National Institute of Environmental Health Sciences. This test set is relatively small and focuses on health and environmental issues. It would be useful to extend this study to larger test sets with varied subject areas, much like the actual World Wide Web.

Another possibility for extending this study includes comparing the retrieval effectiveness of author generated metadata to the professionally and automatically generated metadata records. Author generated metadata is more expensive than automatically generated metadata in terms of time and human involvement, but is less resource intensive than involving metadata professionals.

Additionally, the research raised questions about automatic generation tools. It is possible that if automatic generation tools are customized for specific application profiles and domains, the results for information retrieval effectiveness using these metadata records could be improved. This suggests that further research needs to be explored related to the automatic generation tools themselves.

References

Anderson, J., & Perez-Carballo, J. (2001). The Nature of Indexing: How Humans and Machines Analyze Messages and Texts for Retrieval. Part I: Research, and the Nature of Human Indexing. Information Processing and Management, 37, 231-254.

DC-dot. (November 2002). http://www.ukoln.ac.uk/metadata/dcdot/.

Drott, C. M. (2002). Indexing Aids At Corporate Websites: The Use Of Robots.txt and META Tags. Information Processing and Management, 38, 209-219.

Dublin Core Metadata Initiative. (March 2003). http://dublincore.org/.

Dublin Core Metadata Element Set, Version 1.1. (2003). http://www.dublincore.org/documents/dces/.

Giuffrida, G., Shek, E., & Yang, J. (2000). Knowledge-based Metadata Extraction from PostScript Files. Proceedings of the Fifth ACM Conference on Digital Libraries, 77-84.

Greenberg, J., Patteulli, M., Parsia, B., & Robertson, W. (2001). Author-Generated Dublin Core Metadata for Web Resources: A Baseline Study in an Organization. Journal of Digital Information, 2(2), 38-46.

Greenberg, J. (2002). Metadata and the World Wide Web. Encyclopedia of Library and Information Science, 72(35), 244-261.

Griffin, L., & Thomas, C. (1999). Who will create the Metadata for the Internet? First Monday [On-line], 3(12). Available: http://www.firstmonday.dk/issues/issue3_12/thomas/index.html.

Harper, C., Greenberg, J., Robertson, W., & Leadem, E. (2002). Abstraction versus Implementation: Issues in Formalizing the NIEHS Application Profile.

Proceedings of DC-2002: Metadata for e-Communities: Supporting Diversity and Convergence, 213-215.

Jenkins, 1999, C., Jackson, M., Burden, P., & Wallis, J. (1999). Automatic RDF Metadata Generation for Resource Discovery. Computer Networks, 31, 1305-1320.

What is Klarity. (March 2003). http://www.klarity.com.au/.

Lawrence, S., & Giles, C. (1999). Accessibility of Information on the Web. Nature, 400, 107-109.

Marchiori, M. (1998). The Limits of Web Metadata, and Beyond. Computer Networks and ISDN Systems, 30, 1-9.

Metadata Generation Research Project. (2002). http://ils.unc.edu/~janeg/mgr/.

On the Size of the World Wide Web. (2001). http://www.pandia.com/sw-2001/57-websize.html.

Salton, G. (1986). Another Look At Automatic Text-Retrieval Systems. Communications of the ACM, 29(7), 648-656.

Sullivan, D. (2002). Death Of A Meta Tag. http://searchenginewatch.com/sereport/02/10-meta.html.

Tague-Sutcliffe, J. (1992). The Pragmatics of Information Retrieval Experimentation, Revisited. Information Processing and Management, 28(4), 467-490.

The New Meta Tags Are Coming – Or Are They? (1997). http://www.searchenginewatch.com/sereport/97/12-metatags.html.

Wacholder, N., Evans, D.K., & Klavans, J. (2001). Automatic Identification and Organization of Index Terms for Interactive Browsing. Proceedings of the First ACM/IEEE-CS Joint Conference on Digital Libraries, 126-134.

APPENDIX A

Sample Metadata Record

Element Type	Number	Content
AUDIENCE	1	Researchers
AUTHOR-CONTRIBUTOR	1	National Institute of Environmental Health Sciences (U.S.).
DATE_CREATED	1	2000s
DATE_MODIFIED	1	2000s
DESCRIPTION	1	One-page site that describes the research focus of the Cancer and Aging section of the Laboratory of Molecular Carcinogenesis of the Division of Intramural Research of the National Institute of Environmental Health Sciences. Includes email link to principal investigators.
LANGUAGE	1	English
RELATION:isPartOf	1	http://dir.niehs.nih.gov/dirlmc/ (National Institute of Environmental Health Sciences (U.S.). Division of Intramural Research. Laboratory of Molecular Carcinogenesis.
SUBJECT	1	Aging and cancer
TITLE	1	Cancer and Aging Section
TYPE	1	Text
URL	1	http://dir.niehs.nih.gov/dirlmc/cagrs.htm
AUDIENCE	2	NIEHS Employees
SUBJECT	2	Cell death - genetics
SUBJECT	3	Hormones and cancer
SUBJECT	4	Oxidative stress
SUBJECT	5	Malignant progression - physiology
SUBJECT	6	KAII (Metastasis suppressor gene)

Note: The number column is used to represent multi-valued elements associated with a single record.

APPENDIX B

Original Query Questions

Q1. Is NIEHS conducting any research in the area of HIV-related proteins?

Q2. Is anyone at NIEHS investigating the potential link between a person's genetic makeup and their predisposition to developing cancer or other diseases?

Q3. Anti-oxidants in the diet have been associated with reducing the risk of certain cancers. Is NIEHS conducting research in this area?

Q4. Are certain minority groups more susceptible to harm from environmental pollutants? Is NIEHS conducting any research in this area?

Q5. Has asthma been linked to any particular types of environmental pollutants?

Q6. What do we know about the relationship between estrogen and lupus?

Q7. Do NIEHS researchers hold theories as to how our bodies actually metabolize or excrete the toxins we are exposed to?

Q8. What kind of birth defects may be associated with maternal environmental exposures?

Q9. I live near high power electrical lines. Have they been associated with an increased risk of certain cancers?

Q10. What are the human health risks associated with dioxin exposure?

Actual Queries Submitted

Q1

```
select distinct url from  elements where (content like '*HIV*' or content like '*protein*')
and  (participant='DC-Dot' or participant='Klarity')
```

```
select distinct url from  elements where (content like '*HIV*' or content like '*protein*')
and  (participant='LoggerA' or participant='LoggerB' or participant='LoggerC')
```

Q2

```
select distinct url from elements where (content like '*genetic*' or content like '*cancer*'
or content like '*disease*') and  (participant='DC-Dot' or participant='Klarity')
```

```
select distinct url from elements where (content like '*genetic*' or content like '*cancer*'
or content like '*disease*') and  (participant='LoggerA' or participant='LoggerB' or
participant='LoggerC')
```

Q3

```
select distinct url from elements where (content like '*anti-oxidant*' or content like
'*diet*' or content like '*cancer*') and  (participant='DC-Dot' or participant='Klarity')
```

```
select distinct url from elements where (content like '*anti-oxidant*' or content like
'*diet*' or content like '*cancer*') and  (participant='LoggerA' or participant='LoggerB' or
participant='LoggerC')
```

Q4

```
select distinct url from elements where (content like '*minority*' or content like
'*environmental pollutant*'  and  (participant='DC-Dot' or participant='Klarity')
```

```
select distinct url from elements where (content like '*minority*' or content like
'*environmental pollutant*') and  (participant='LoggerA' or participant='LoggerB' or
participant='LoggerC')
```

Q5

```
select distinct url from elements where (content like '*asthma*' or content like
'*environmental pollutant*') and  (participant='DC-Dot' or participant='Klarity')
```

```
select distinct url from elements where (content like '*asthma*' or content like
'*environmental pollutant*' ) and  (participant='LoggerA' or participant='LoggerB' or
participant='LoggerC')
```

Q6

```
select distinct url from elements where (content like '*estrogen*' or content like '*lupus*')
and  (participant='DC-Dot' or participant='Klarity')
```

```
select distinct url from elements where (content like '*estrogen*' or content like '*lupus*')
and  (participant='LoggerA' or participant='LoggerB' or participant='LoggerC')
```

Q7

```
select distinct url from elements where (content like '*metabol*' or content like
'*excrete*' or content like '*toxin*') and  (participant='DC-Dot' or participant='Klarity')
```

```
select distinct url from elements where (content like '*metabol*' or content like
'*excrete*' or content like '*toxin*') and  (participant='LoggerA' or participant='LoggerB'
or participant='LoggerC')
```

Q8

```
select distinct url from elements where (content like '*birth defect*' or content like
'*maternal*' or content like '*environmental exposure*') and  (participant='DC-Dot' or
participant='Klarity')
```

```
select distinct url from elements where (content like '*birth defect*' or content like
'*maternal*' or content like '*environmental exposure*') and  (participant='LoggerA' or
participant='LoggerB' or participant='LoggerC')
```

Q9

```
select distinct url from elements where (content like '*high power*' or content like
'*electrical*' or content like '*risk*' or content like '*cancer*') and  (participant='DC-Dot'
or participant='Klarity')
```

```
select distinct url from elements where (content like '*high power*' or content like
'*electrical*' or content like '*risk*' or content like '*cancer*') and
(participant='LoggerA' or participant='LoggerB' or participant='LoggerC')
```

Q10

```
select distinct url from elements where (content like '*risk*' or content like '*dioxin*')
and  (participant='DC-Dot' or participant='Klarity')
```

```
select distinct url from elements where (content content like '*risk*' or content like
'*dioxin*') and  (participant='LoggerA' or participant='LoggerB' or
participant='LoggerC')
```

APPENDIX C

Final Relevance Judgments

WEBPAGE	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
http://cerhr.niehs.nih.gov	N	N	N	N	Y	N	N	Y	N	Y
http://dir.niehs.nih.gov/direb/baird.htm	N	N	N	N	N	N	N	N	N	N
http://dir.niehs.nih.gov/direb/clu/home_clu.htm	N	N	N	Y	N	Y	N	N	N	N
http://dir.niehs.nih.gov/direb/london.htm	N	Y	N	N	Y	N	N	N	N	N
http://dir.niehs.nih.gov/dirlcbra/epidem.htm	N	N	N	N	N	N	Y	N	N	Y
http://dir.niehs.nih.gov/dirlecm/tcu/home.htm										
http://dir.niehs.nih.gov/dirlep/lcm.html	N	N	N	N	N	N	N	N	N	N
http://dir.niehs.nih.gov/dirlmc/	N	N	N	N	N	N	N	N	Y	N
http://dir.niehs.nih.gov/dirlmc/cagrs.htm										
http://dir.niehs.nih.gov/dirlmc/seqcore.htm	N	Y	N	N	N	N	N	N	N	N
http://dir.niehs.nih.gov/dirlmg/B_Copeland.html	Y	N	N	N	N	N	N	N	N	N
http://dir.niehs.nih.gov/dirlmg/home.htm	Y	Y	N	Y	Y	Y	N	Y	Y	N
http://dir.niehs.nih.gov/dirlmg/J_Mason.html	N	N	N	N	N	N	N	N	N	N
http://dir.niehs.nih.gov/dirlmg/R_Schaaper.html	N	N	N	N	N	N	N	N	N	N
http://dir.niehs.nih.gov/dirlpc/	N	N	N	N	N	N	Y	N	N	N
http://dir.niehs.nih.gov/dirlpc/chemmetab.htm	N	N	N	N	N	N	N	N	N	N
http://dir.niehs.nih.gov/dirlpc/intra.htm	N	N	N	N	N	N	Y	N	N	N
http://dir.niehs.nih.gov/dirlsb/hall_home.html	N	Y	N	N	N	N	N	N	N	N
http://dir.niehs.nih.gov/dirlsb/msprot.htm	N	N	N	N	N	N	N	N	N	N
http://dir.niehs.nih.gov/dirlsb/mssfacil.htm	N	N	N	N	N	N	N	N	N	N
http://dir.niehs.nih.gov/dirlsb/mssgroup.htm	Y	N	N	N	N	N	N	N	N	N
http://dir.niehs.nih.gov/dirlsb/msshome.htm	Y	N	N	N	N	N	N	N	N	N
http://dir.niehs.nih.gov/dirlsb/msspubs.htm	N	N	N	N	N	N	N	N	N	N
http://dir.niehs.nih.gov/dirlsb/msssumm.htm	Y	N	N	N	N	N	N	N	N	N
http://dir.niehs.nih.gov/dirlst/groups/obryan.htm	N	N	N	N	N	N	N	N	N	N
http://dir.niehs.nih.gov/dirlst/putney.htm	N	N	N	N	N	N	N	N	N	N
http://dir.niehs.nih.gov/faculty/	N	N	Y	N	N	N	Y	N	N	N
http://dir.niehs.nih.gov/proteomics/	N	N	N	N	N	N	N	N	N	N
http://www.niehs.nih.gov/dert/programs/special/specpops.htm	N	N	N	N	N	N	N	N	N	N
http://www.niehs.nih.gov/dert/programs/toxgenom.htm	N	Y	N	N	N	N	N	N	N	N
http://www.niehs.nih.gov/dert/programs/translat/cbpr/cbpr.htm	N	N	N	Y	N	N	N	N	N	N
http://www.niehs.nih.gov/dert/programs/translat/home.htm	N	N	N	N	N	N	N	N	Y	N
http://www.niehs.nih.gov/emfrapid/home.htm	N	N	N	N	N	N	N	N	Y	N
http://www.niehs.nih.gov/wetp	N	N	N	Y	N	N	N	N	N	N

*Note: The two web pages without ratings were no longer accessible and were thus excluded from the results calculated for the study.