Can a system make novice users experts?
Important factors for automatic metadata generation systems

Sue Yeon Syn
School of Information Sciences
University of Pittsburgh
USA
sus16@pitt.edu

Michael B. Spring
School of Information Sciences
University of Pittsburgh
USA
spring@pitt.edu

Abstract

The description of Web resources is one of the problems related to the development of the Semantic Web. The major problem in metadata generation is due to the lack of experts and the tremendous amount of Web resources. It is expected that an automated system would encourage creation of metadata for Web resources. This work is focused on suggesting four classes of descriptive elements – bibliographic, semantic, keywords, structure – in designing an automated metadata generation system. In addition, this study tries to assess how the level of knowledge or skill impacts the quality of the metadata generated and, based on the results, it suggests factors that automatic metadata generation tools should suggest to users.

Keywords: metadata; metadata generation; evaluation.

1. Introduction

The World Wide Web (WWW) makes it possible for users to post resources in a distributed way and find resources by following links. This basic structure makes the WWW become a pool of tremendous information but also makes it difficult to locate all the resources relevant to a given topic or query. Search engines have emerged to help users search for Web resources by full-text indexing. Regardless of algorithmic improvements such as in ranking and clustering, full-text indexing problems include synonymy and polysemy as well as semantic connectivity. Services such as Yahoo! create directories based on content analysis done by humans. While directory services can provide more precise classification of Web resources, human intervention raises the cost and causes scalability problems. Both approaches, full-text indexing and directory services, have problems related to the churn in Web pages (new pages appearing, old pages changing or being removed) and the increasing use of programmatic links (CGI programs and Web services) that “hide” the content of pages. As Web 2.0 technologies such as AJAX and RSS take hold, these problems will be compounded.

In efforts to provide a better way to find proper resources on the WWW, research has been undertaken to analyze Web resource content and to create metadata automatically of a quality equal to that generated by humans but without the cost and scalability problems. Metadata of Web resources generally provide bibliographic information about the resource in a structured way and allow us to locate the resources in a more effective way. Having metadata for Web resources also contributes to the development of the Semantic Web. The Semantic Web promises to “bring structure to the meaningful content of Web pages, creating an environment where software agents roaming from page to page can readily carry out sophisticated tasks for users” (Berners-Lee, Hendler & Lassila, 2001). Having descriptions of Web resources would make it possible to structure the content in a machine understandable way.

Although the needs and benefits are certain, it is still considered to be difficult to generate metadata for Web resources. One of the challenges of metadata generation is that someone (most of the time, the creator of the Web resource or a Web master)has to create the metadata as well since it is impossible for professional catalogers to classify every Web resource. The problem has three parts, Firstly, most users fail to see the benefit of taking the time to create resource descriptions. Secondly, most users are not skilled in developing good descriptions. Finally, there are billions of pages that have already been created that need to be classified. Conceptually, once the metadata for Web resources are created, it is easier to locate and use the Web resources. How to easily create good metadata becomes an issue given the fact that creating good metadata manually takes time and expertise.

With the recognition of the issue, this paper studies how a tool might make it easier to generate metadata for Web resources in a more automatic way. It concentrates on finding factors that will let metadata generating systems be designed to be simple, distributed, extensible, and scalable in developing metadata and be adapted to many end-users to let them take a part in the metadata generating process.

2. Approaches in Automated Metadata Generation

There have been two main approaches in automatic generation of metadata: harvesting and extraction (Greenberg, 2004). Harvesting draws information from existing metadata in the resources such as meta tags in HTML (Greenberg, 2004; Jenkins et al., 1999). There are tools for harvesting information for Web resources such as DC-DOT. Many well-known HTML editor applications such as Front Page, Dreamweaver, and even Microsoft Word automatically input meta tags with some basic bibliographic information when creating HTML files. However, this approach mainly concentrates on simple bibliographic information not considering other kinds of information such as the semantics of the contents.

Extraction generates information automatically using an algorithm based on the attributes or content of the resources (Greenberg, 2004). Information extraction and natural language processing techniques are used for extraction. There have been many research projects that have attempted to generate metadata based on extraction. MetaExtract automatically assigns Dublin Core and GEM metadata using extraction techniques from natural language processing (Yilmazel, Finneram & Liddy, 2004). The Simple Indexing Interface is a framework for automatic metadata generation for learning objects and extracts metadata based on information from document content, context, usage, and composite documents structure (Cardinaels, Meire & Duval, 2005). Both studies demonstrate that extraction based automatic generation of resource description could create metadata of the quality of manually generated metadata, at least when assigning to specific schemas such as DC and GEM. When the extraction approach is an effort to let machines understand the resource content with less human involvement, it, however, still fails to understand the semantic connectivity of the content since it only extracts information from what is expressed in that content.

Other projects have tried other methods to help automatic metadata generation. The Final Report of the AMeGA (Automatic Metadata Generation Applications) Project recommends some functionalities for automatic metadata generation applications organized in categories including system configuration, metadata identification/gathering, support for human metadata generation, metadata enhancement/refinement and publishing, etc. (Greenberg, Spurgin & Crystals, 2004). It also emphasizes the need for satisfying the users, i.e. the metadata experts. Fedora (Fedora Project; Lagoze et al., 2005) is an extensible framework for the storage, management, and dissemination of complex objects and the relationships among them. By supporting aggregation and association of digital objects, Fedora gives flexibility, meaning interoperability and extensibility (Payette et al., 1999), to users in terms of representation and management of objects. Fedora adopts the ABC ontology model (Lagoze & Hunter, 2001) and the Warwick Framework (Lagoze, Lynch & Daniel, 1996) that provide a common conceptual model to facilitate interoperability and extensibility for different domains. This effort focuses on sharing existing metadata therefore making it easier to create metadata for resources. However, Fedora is only focusing on digital objects in a repository or related repositories that are not distributed resources. CREAM (CREAting Metadata) is an annotation framework to create relational metadata, i.e.

metadata that instantiates interrelated definitions of classes in a domain ontology rather than a comparatively rigid template-like schema such as the Dublin Core (Handschuh & Staab, 2002). This study stresses extensibility in metadata creation based on the recognition that human involvement is somewhat necessary to generate high quality metadata and tries to reduce, not eliminate, the human effort in metadata generation which still requires expert users.

Some studies have tried to involve non-expert users in the metadata generating process in so-called collaborative tagging that lets users add tags to certain Web resources. Many Web services especially related to multi-media contents allow the addition of tags to represent the content. For example: (1) Flickr (http://www.flickr.com) is a Web-based photograph sharing service that lets users add tags to their photos and share with other users; (2) del.icio.us (http://del.icio.us/) is a collaborative bookmarking system that lets users tag their favorites and share them; and (3) YouTube (http://www.youtube.com/) is a video sharing system that lets users tag what the video is about. These systems and services make it simple to organize and position their resources regardless of various resource types by letting users be involved in the metadata generation process (Macgregor & McCulloch, 2006). This approach shows the possibility of reducing costs and increasing the scalability of metadata generation processes. However, since novice users have less knowledge of how metadata should be created and used, the quality of generated metadata is not guaranteed.

This study focuses on how a system can help novice users create metadata similar in quality to that created by expert users. With this goal, this study also suggests some factors for automatic metadata generation tools targeting novice users.

3. Methodology

Human-generated metadata developed by skilled classifiers are generally considered to be more precise than system-generated metadata. However, it is not possible for qualified humans to create metadata for every Web resource. Therefore, tools are needed that automatically generate the metadata or that enable less skilled classifiers to generate quality metadata. In this study, two conditions are considered: expert-generated metadata (manually generated metadata) and system-generated metadata with human intervention (including both experts and novices). These conditions are considered based on the observation that automatic systems cannot match the accuracy of human counterparts (Greenberg, 2001). The purpose of this study is to develop preliminary data to assess the effectiveness and efficiency of semi-automatic metadata generation systems and suggest possible factors for automatic metadata generation system development.

The study seeks to develop base data that will allow us to answer three questions:

1. How much of the metadata typically associated with a Web resource can be generated automatically?
2. To what extent are novice users able to review and correct automatically generated metadata?
3. What factors would extend the quality of the metadata generated by novice users?

For this study the Website of the Andy Warhol Museum in Pittsburgh (http://www.warhol.org/) was used as a target resource to create metadata. The Andy Warhol Museum exhibits the work of Andy Warhol as well as some other collections. The collections include art, pictures, and videos. The Andy Warhol Museum Website was selected because it is a Website with a variety of content in a well-organized structure. It contains bibliographic information for some of the collections and important raw data, such as images and pictures. Twenty-nine pages within the site were used for the study. For each page, metadata were generated.

The first part of the study had experts create metadata to test how human experts classify the Web pages and generate metadata for them. The subjects were librarians or library science students at the masters and PhD level with at least three years of professional librarian experience and for the purposes of this study they were considered to be expert catalogers. They classified the Web pages and created metadata for the given Web page set. The average time to generate metadata and agreed metadata attributes were measured. The values were considered to be in agreement if more than half of the experts agreed. Since there was no public benchmark available to apply to the current study, the data were used as a benchmark for the remaining analysis.

The second part of the study assessed human intervention on tool-generated metadata. Sixteen subjects were divided into two groups: experts and novices in cataloging and classification. The subjects were graduate students at the School of Information Science, University of Pittsburgh. They were asked to create metadata for the Andy Warhol Museum Website with the assistance of a metadata generation system. All of them were familiar with information retrieval. None of them had experiences in using any kind of metadata generation tools. The expert group had over 3 years' experience on average in working on cataloging or classification related tasks. This part of the study provided an opportunity to assess how the level of knowledge or skill impacts the quality of the metadata generated. The condition of the tool was varied in this study. In one condition, all system-generated values were included. In the other, only high confidence system-generated values were included. This variation provided some preliminary data on how users assess system-generated information and determine what kinds of factors could improve the quality of metadata when assistance was provided. The measurement methods were precision, recall and time. The results were compared to the results obtained in the first part of the study to provide some base data to serve as benchmarks for use in analyzing the level of effect of human intervention in generating metadata.

Following the experiment, the subjects completed a survey to gather some preliminary data on user satisfaction with system performance and interface design. Questions were asked pertaining to each module of the system in addition to an overall evaluation. It was expected that suggestions and comments would provide insights for possible future enhancements of the system.

4. Factors for Metadata Suggestion

Given a more "Semantic Web" as the goal, it is necessary to have information about Web pages and their contents structured into a certain format to provide updated summarization of the Web pages. Taking into account the characteristics of Web pages and difficulties in placing their contents to the related domain properly, a system was designed and implemented to collect and store Web page metadata. The tool used for this study, the Metadata Generation System (MGS), is a modified spider that catalogs a site. MGS harvests/extracts as much metadata as possible about the resources on the site for presentation and assessment by a human. The resulting metadata is then stored in an XML file as a modified RDF Site Summary for further processing. MGS is designed to be extensible. In its current form, it works on four categories of information about Web pages: bibliographic, semantic, keyword, and structure. The bibliographic module gathers basic bibliographic information about the Web page. The semantic module indicates relevant semantics of the content of the page based on a classification schema or ontology. The keyword module addresses the content of the page. The structure module represents how the page is structured. Each of these modules is described in more detail below.

The bibliographic module gathers basic information on the documents. Existing mechanisms for bibliographic information on Web pages, such as the Dublin Core elements are used. This component declares a bibliographic scheme based on the type of the pages since a previous study showed that the relationship with genre of the resource influences the correctness of description (Greenberg, 2004). For each type of Web page, a different scheme is applied according to its characteristics. This component suggests values by harvesting HTML tags. The bibliographic information provides basic metadata for the Web pages.

The semantic module uses existing classification schemata or ontologies, the Dewey Decimal Classification (DDC) and the Universal Decimal Classification (UDC) at the current time, to define the domain or topic of the Web page. Any schema that is described in XML can be used.

Therefore for each Web page, any ontology or classification schema related to the specific domain or topic can be applied. Search within classification schemata and text input functions are included to give flexibility to users. If a certain class is chosen, the parents of it are also chosen to represent the document. The information collected from this module is significant since it declares what domain the Web pages belong to.

The keyword module suggests keywords based on term frequency and anchor text. This component works on the content of the Web page. Finding frequently used words and anchor texts after eliminating stop words from the page makes it easier to determine the subject of the page. At most, ten terms are suggested and users can easily add or remove keywords.

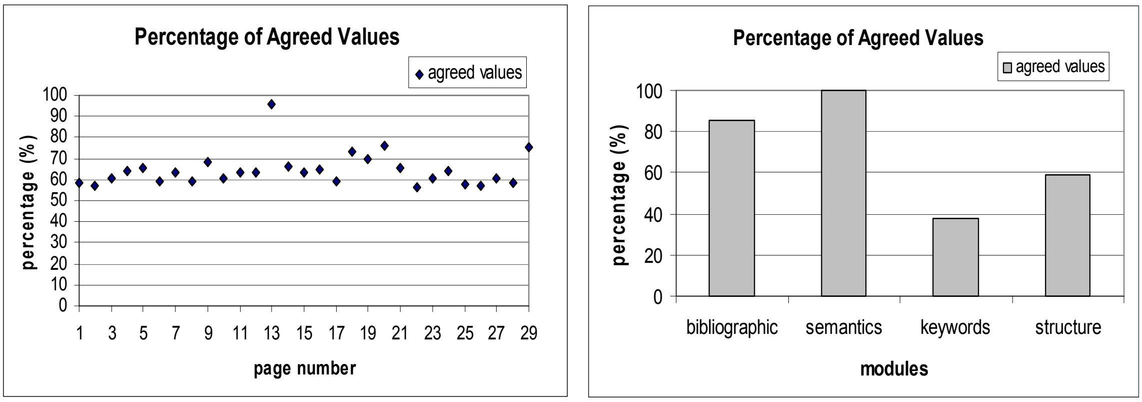
Structural information about a document is considered significant given that structural information can be used in a variety of ways and can be generated automatically as well as on previous survey results (Greenberg, Spurgin & Crystal, 2006).

The structure module gathers information about the structure of the Web page. Structure means the elements of the documents and the techniques used in the document. This feature can be useful when a search engine finds a specific type or format of information such as images. For this system, specific descriptions of tables, lists, images are taken into consideration along with stylesheets and scripts. The structural information is gathered using extraction techniques.

For the study, the system was designed in two versions of the tool with different levels of confidence values. Tool 1 suggests every system generated value regardless of accuracy and quantity. Tool 2 suggests only high confidence values that were determined by experts' agreements. Both tool 1 and tool 2 are designed with the four modules described above.

5. Evaluation

This study developed preliminary data on a system to improve automated metadata generation. The first part of the study generated values for resource descriptions of selected pages from the Andy Warhol Museum Website. Experts were provided a simple interface with input boxes where they could input appropriate values for metadata attributes. *Agreed values* were those where more than 50% of the experts assigned the same value for the attributes. *Agreed values* were used as benchmark values for this study and were considered to be the *high-confidence values*. Overall 64.37% of the attributes were agreed values (FIG. 1-a). Average time spent per page to generate metadata was 13.29 minutes. Comparing agreed values of each module (FIG. 1-b), bibliographic and semantic modules were highly agreed since they were based more on professional knowledge. Keywords varied depending on selection of words, numbers, etc. causing a lower percentage of agreed values. Structure attributes differed depending on experts' decisions as to how detailed the metadata should be.



(a) Agreed Values of Pages **(b) Agreed Vales of Modules**

FIG. 1. Percentages of Agreed Values Among Experts.

For the second part of the study, the two versions of the tools were tested with expert and novice subjects. Tool 1 provided every value generated by the system. Tool 2 only provided values that were matched with agreed values from the benchmark, i.e. *high confidence values*. The concept of confidence level was introduced in Cardinaels et al. study (2005) to provide options of system suggested values. In Cardinaels' study, a confidence value indicates the degree of certainty for the purpose of solving conflicts among system- suggested values. In this study, to assess the impact of including lower confidence data on the human review of the generated metadata, the confidence level was set by comparing the generated values to the *agreed values* of the benchmark. If system generated values did not match the *agreed values* they were not displayed in tool 2 whereas tool 1 displayed all values generated by the system. For each tool, four experts and four novices participated in generating resource descriptions.

To evaluate the metadata generation process and its results, we treated the metadata values like documents in a retrieval set, where the "high confidence values" represent all the correct values and the "correct values" represent the number of correct values retrieved. Thus, precision and recall were measured as follows.

$$\text{Precision} = \frac{\textit{Number of Correct Values}}{\textit{Number of Filled-in Values}} \qquad \textit{(Eq1)}$$

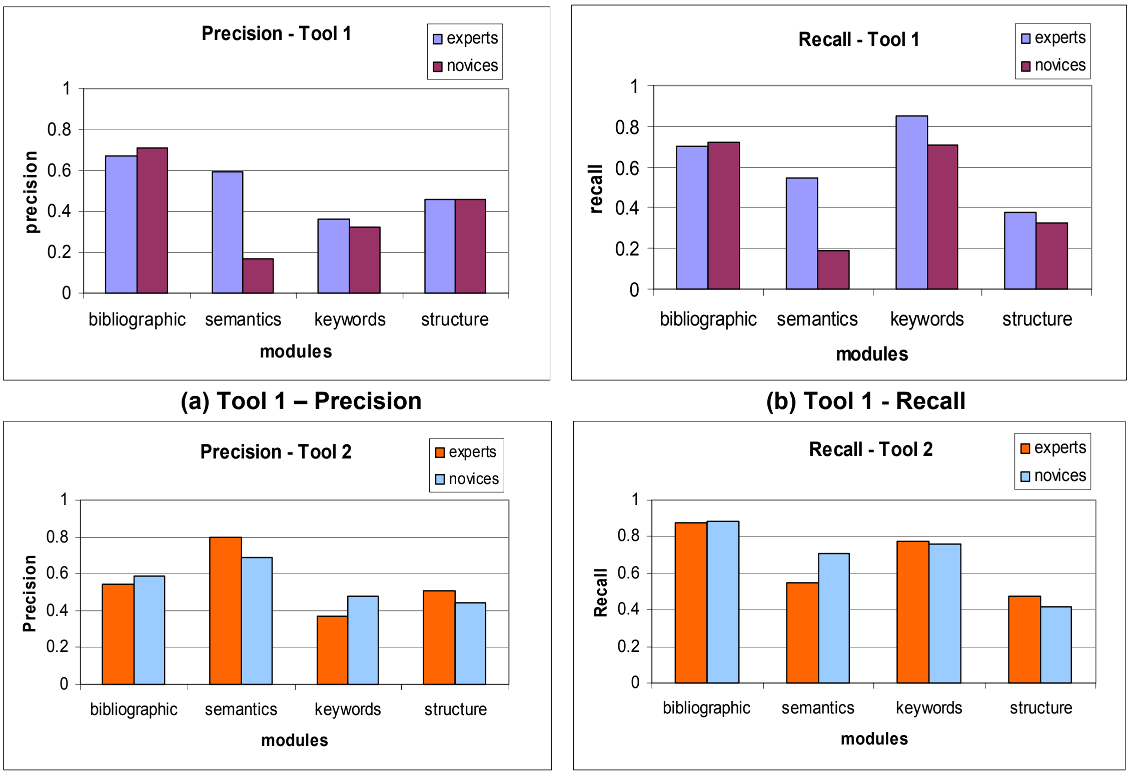
$$\text{Recall} = \frac{\textit{Number of Correct Values}}{\textit{Number of High Confidence Values}} \qquad \textit{(Eq2)}$$

Here the number of *filled-in values* is the number of attribute values that were input by subjects. The number of *correct values* is the number of input attribute values that matched the *agreed values*.

FIG. 2 shows the precision and recall for both tools. First of all, in comparing two tools, it can be observed that overall, the tool 2 has higher precision and recall. It would appear that when the system provides only *high confidence values* users input more *correct values*. For statistical support, a two-way between-subjects ANOVA was performed showing a significant difference on the precision scores between the tools average across experts and novices, $F(1,12)= 26.185$, $p< .001$, partial $\eta^2= .686$. To find the pattern of differences in precision, the differences between tool 1 and tool 2 for experts and novices were tested. There was a significant difference in precision between tool 1 and tool 2 for expert users, $F(1,12)= 9.008$, $p= .011$. There was also a significant difference in precision between tool 1 and tool 2 for novice users, $F(1,12)= 17.939$, $p= .001$. For both experts and novices, the precision of tool 2 was significantly higher than the precision of tool 1. This is also indicated in FIG. 2. Expert users did not show much difference between tools. Compared to novices, experts were more constant as they input values based on their knowledge rather than tool suggestion. Novice users were more influenced when the values suggested by the tool changed. Tool 2 (with *high confidence values*) was more trusted by novice users, whose amount of correct input values was higher with this tool.

Each of the four modules was separately evaluated. For the bibliographic module, novice users tend to input more *correct values*. This is due to the number of *filled-in values*. Novice users fill in only certain values whereas expert users tend to fill in as much information as possible which in turn results in fewer *correct values*. For novices the semantic module was considered the most difficult. Also the novices showed a tendency to assign one page to several classifications whereas experts assign a page to one classification similar to library classification tasks. This tendency caused lower precision for novices and higher precision for experts for the semantic module. Both keyword and structure modules have comparatively lower precision values. Users were free to input as many keywords they wanted. As a result, the number of *filled-in values* increased and the precision is low. The structure module caused confusion. For example, the system suggests tables when table tags were used in the pages without considering their purpose. Tables were often used for layout in the Web pages rather than to present data. Some users, especially experts, only described tables with meaningful information. Some described all tables even if they did not contain any meaningful information. This suggests that the structure module may require further algorithm development, e.g. it might be refined to exclude tables and other structures that are simply used for formatting purposes.

It is interesting to find differences in precision especially in the semantic and keyword modules where novices performed worse than in the other modules. From FIG. 2, it was indicated that the semantic and keyword modules had larger differences between tools. With ANOVA test, it was found that there is a significant difference in precision of the semantic module between tool 1 and tool 2 for novices, *F*(1,12)= 9.932, *p*= .008. There was a significant difference in precision of the keyword module between tool 1 and tool 2 for novices, *F*(1,12)= 7.087, *p*= .021. From this analysis, it appears that tool 2 helped novice users improve quality of metadata generated especially in semantic and keyword modules.



(a) Tool 1 – Precision

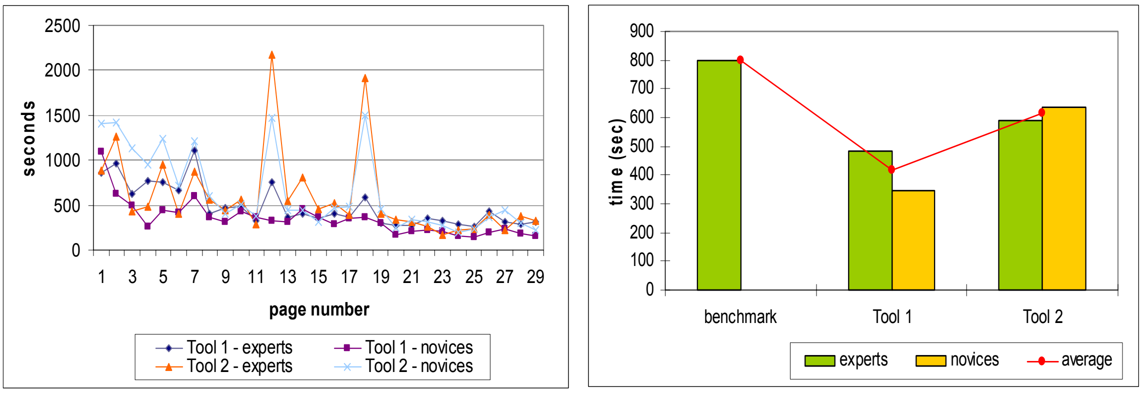
(b) Tool 1 - Recall

(c) Tool 2 – Precision

(d) Tool 2 - Recall

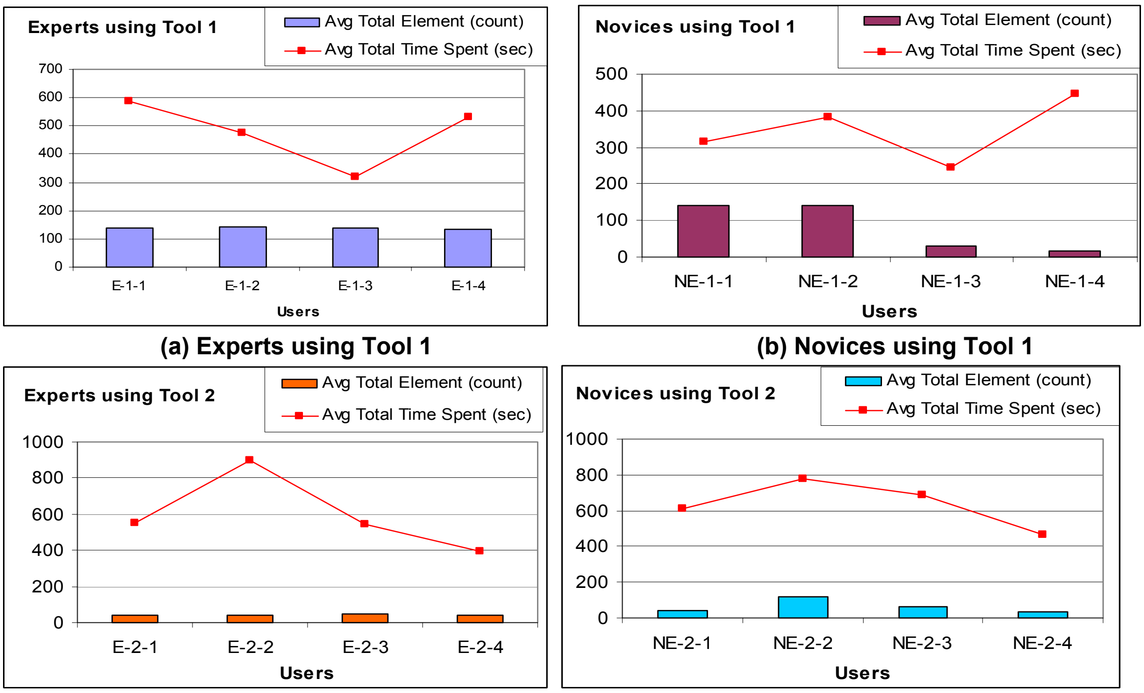
FIG. 2. Precision and Recall for Tool 1 and Tool 2 with Human Intervention.

The study also examined the time spent in generating metadata using the tools. The average time spent to generate metadata using tool 1 and tool 2 was about 7 minutes and 10 minutes per page respectively. Since tool 2 suggested fewer attribute values, subjects spent more time using tool 2. With both tools, the average time per page appears to decrease with experience (FIG. 3-a). Although the pages were provided in order some subjects did not follow the order for every page since the page process order was not restricted. While the page process order was generally followed by most subjects without any constraints, it can be observed that there are learning effects in generating metadata using the tools. Comparing both tools with the benchmark, it was found that there is significant difference in time spent to generate metadata manually versus the tools, *F*(2,12)= 7.986, *p*= .004, partial η^2= .484 (FIG. 3-b). The post hoc comparisons were performed using Scheffe adjustment to find patterns of differences. The results showed that tool 1 required significantly less time compared to benchmark (manual process), *p*= .005. No other significant differences were found on time spent.



(a) Time Spent per Page **(b) Time Comparison among tools**

FIG. 3. Spent Time Comparisons.



(a) Experts using Tool 1 **(b) Novices using Tool 1**

(a) Experts using Tool 2 **(b) Novices using Tool 2**

FIG. 4. Number of Elements/Attributes and Time Spent per User.

The time spent in generating metadata varied depending on the number of elements and attributes users filled in. The number of elements/attributes ranges from 9 to 242 per page. The number of elements/attributes created appeared to vary by user – some users liked to add as many values as possible and others liked to keep the metadata simple. However there was not a clear relationship between the number of elements input and the time. The results showed that there is some relation between the type of metadata information input and time. The semantic and keyword modules required more time per element, the bibliographic module took 7 seconds per element, the semantic module required 13.35 seconds, the keyword module required 9.92 seconds, and the structure module required 4.48 seconds. In addition, some observations can be made related to expert and novice differences in using the tools. For tool 1, experts consistently input about 138 elements with very little deviation. Elements input by novices ranged from 30 to 140. Using tool 2, experts input about 42 elements and novices input between 34 to 119 elements. Expert users tended to input a consistent number of elements when using the tool. Novices tended to vary widely in terms of number of input elements. As shown in FIG. 4, there does not seem to be a clear relationship among numbers of elements/attributes with time while the choice of tool does affect the time spent in generating metadata.

These results, while preliminary, suggest that recommending only *high confidence values* helped users, especially novices, generate more accurate metadata. It is very interesting to find this result since it supports possibilities that novices can generate good quality metadata with some help from a system. Novices appear to trust tool-generated values more than experts. This tendency causes novices to spend less time when using a tool with more values suggested, but creates metadata that is less accurate. However with only *high confidence values*, novices spend more time to input attribute values but create better metadata. These results suggest a need of compromise when developing a tool for metadata generation when non-experts are targeted as the primary user. For example, from the result it shows that experts generate good quality metadata using both tools, therefore selecting the tool that would reduce the time spent would be more reasonable. However, if the system suggested fewer values, novice users tend to spend more time. Although suggesting only high confidence values would result in better metadata, if this condition required too much time, it may not be a good factor to consider. This finding raises interesting questions in terms of balancing the number of suggested values and their quality.

To evaluate user satisfaction a post experiment questionnaire was administered to the sixteen subjects participating in the second part of the study. Subjects were asked to rate the usefulness of each module (on a scale of 1 to 5). Over 70% of the subject considered having four modules helpful in generating metadata. Each module was rated to be useful, especially by experts (Table 1): 87% for the bibliographic module, 53% for the semantic module, 100% for the keyword module, and 53% for the structure module.

TABLE 1. Mean and Standard Deviation of User Satisfaction Rate for Modules.

		Experts		Novices	
		Mean	St. Dev.	Mean	St. Dev.
Tool 1	bibliographic	4.25	0.50	3.25	1.50
	semantic	3.00	0.00	4.00	1.41
	keyword	4.75	0.50	4.75	0.50
	structure	3.75	1.26	4.50	1.00
Tool 2	bibliographic	4.50	0.58	4.25	0.96
	semantic	3.75	0.50	3.75	1.26
	keyword	5.00	0.00	4.00	0.00
	structure	3.50	1.29	3.00	0.82

From the comments provided by subjects, generally using the tool was considered to be good. However, users requested more control over the functionality of the tool. They thought the semantic module should have more options for the selection of classification schemata and ontologies. Subjects agreed that having metadata for structure was necessary, however, how detailed structural information should be recorded was an issue of confusion. For the keyword module, some suggested relating keywords with thesauri to provide more options for keywords selection. Many commented that the metadata generation for Web resources was important and indicated that they would be willing to use a good tool when existing to create metadata (71% of experts and 63% of novices).

6. Discussion and Conclusion

The results of the study suggest that the tool makes it possible to generate metadata information on a Web page more quickly and lets novice users generate metadata as accurately as experts. This study suggested some solutions for metadata generation tool development. When the tool is targeted at novice users, presenting only high confidence values lets users generate more accurate metadata. Considering that users who would generatemetadata for Web resources will not always be experts, this finding is important to those developing metadata generation tools. On the other hand, experts create metadata based on their professional knowledge and skills therefore the accuracy of the metadata generated was not affected by the values the tool suggests. However, using the tool for metadata generation will let expert users reduce the time spent in creating metadata.

The analyses described above as well as informal observation and questionnaire results suggest some further areas for study. Firstly, reviewing the data on the amount of the time required to generate metadata suggests that there is a learning effect. As indicated above, pages were generally reviewed in order but since such an order was not enforced we cannot be sure there is a learning effect. A study with a randomly assigned order in a controlled fashion might show the existence and magnitude in learning effect in using the tool. Secondly, we observed in the current study that experts and novices treated the data provided in the two tools differently. Two conditions of the tools were considered: all system-generated values and only high confidence values. Additional conditions might be of interest such as providing an indication of high and low confidence level of suggested values to users. Indicating the confidence level in the suggested values might affect the quality of the output metadata. Thirdly, better data harvesting/extraction and creation of high-confidence values from the Web resources automatically should be accomplished providing more accurate information and therefore reducing the workload of users and time required for each Web page. Adopting a better algorithm and increasing flexibility will improve the performance of the tools.

References

Berners-Lee, Tim, James Hendler, and Ora Lassila. (2001). The Semantic Web. *Scientific American, 284*(5), 35.

Cardinaels, Kris, Michael Meire, and Erik Duval. (2005, May). Automating metadata generation: The simple indexing interface. *Proceedings of the World Wide Web Conference Committee (IW3C2)* (pp. 548-556).

Fedora Project. (n.d.). *Fedora project: An open-source digital repository management system*. Retrieved, April 16, 2007, from http://www.fedora.info/.

Greenberg, Jane, Kistina Spurgin, and Abe Crystal. (2005). *Final report for the AMeGA (Automatic Metadata Generation Applications) project* (Technical Report).

Greenberg, Jane, Kistina Spurgin, and Abe Crystal. (2006). Functionalities for automatic metadata generation applications: A survey of metadata experts' opinions. *International Journal of Metadata, Semantics, and Ontologies, 1*(1), 3-20.

Greenberg, Jane. (2001). Automatic query expansion via lexical-semantic relationships. *Journal of the American Society for Information Science and Technology, 52*(5), 402-415.

Greenberg, Jane. (2004). Metadata extraction and harvesting: A comparison of two automatic metadata applications. *Journal of Internet Cataloging, 6*(4), 59-82.

Handschuh, Siegfried, and Steffen Staab. (2002, May). Authoring and annotation of Web pages in CREAM. *Proceedings of the 11th International Conference on World Wide Web (WWW2002), Honolulu, Hawaii, USA*.

Jenkins, Charlotte, Mike Jackson, Peter Burden, and Jon Wallis. (1999, May). Automatic RDF metadata generation for resource discovery. *Proceedings of the 8th International World Wide Web Conference, Toronto, Canada*.

Lagoze, Carl, and Jane Hunter. (2001, November). The ABC ontology and model. *Journal of Digital Information, 2*(2), 77.

Lagoze, Carl, Clifford A. Lynch, and Ron Daniel Jr. (1996, June). *The Warwick Framework: A container architecture for aggregating set of metadata* (Cornell Computer Science Technical Report TR96-1593).

Lagoze, Carl, Sandy Payette, Edwin Shin, and Chris Wilper. (2005). *Fedora: An architecture for complex objects and their relationships*. Retrieved April 16, 2007, from http://arxiv.org/abs/cs.DL/0501012.

Macgregor, George, and Emma McCulloch. (2006). Collaborative tagging as a knowledge organisation and resource discovery tool. *Library Review, 55*(5), 291-300.

Payette, Sandy, Christophe Blanchi, Carl Lagoze, and Edward Overly. (1999, May). Interoperability for digital objects and repositories: The Cornell/CNRI experiments. *D-Lib Magazine, 5*(5). Retrieved April 16, 2007, from http://www.dlib.org/dlib/may99/payette/05payette.html.

Yilmazel, Ozgur, Christina M. Finneram, and Elizabeth D. Liddy. (2004, June). MetaExtract: An NLP system to automatically assign metadata. *Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries (JCDL'04), Tuscon, AZ, USA* (pp. 241-242).