Available online at www.sciencedirect.com
SCIENCE DIRECT®
Information Processing and Management 42 (2006) 1099–1122
INFORMATION PROCESSING & MANAGEMENT
www.elsevier.com/locate/infoproman

A study of the metadata creation behavior of different user groups on the Internet

Jin Zhang *, Iris Jastram

School of Information Studies, University of Wisconsin—Milwaukee, Milwaukee, WI 53211, United States

Received 28 February 2005; accepted 2 May 2005
Available online 20 June 2005

Abstract

Metadata is designed to improve information organization and information retrieval effectiveness and efficiency on the Internet. The way web publishers respond to metadata and the way they use it when publishing their web pages, however, is still a mystery. The authors of this paper aim to solve this mystery by defining different professional publisher groups, examining the behaviors of these user groups, and identifying the characteristics of their metadata use. This study will enhance the current understanding of metadata application behavior and provide evidence useful to researchers, web publishers, and search engine designers.

Keywords: Metadata; Metadata evaluation; Internet information organization; Web publishing behavior; Author-generated metadata

1. Introduction

In the morass of vast Internet retrieval sets, many researchers place their hope in metadata as they work to improve search engine performance. If web pages' contents were accurately represented in metadata fields, and if search engines used these metadata fields to influence the retrieval and ranking of pages, precision could increase and retrieval sets could be reduced to manageable levels and ranked more accurately. Theoretically, searchers would also be able to search by author, title, subject, and keyword as they do in other information retrieval systems. Ideally, then, web pages would all be embedded with metadata elements in much the same way that records in OPACs indicate the origination, instantiation, and content of books in a library setting.

* Corresponding author. Tel.: +1 414 229 2712.
E-mail addresses: jzhang@uwm.edu (J. Zhang), irisjastram@hotmail.com (I. Jastram).

0306-4573/$ - see front matter
doi:10.1016/j.ipm.2005.05.002

Achieving this ideal has proven difficult, however. Research and debate are on going and have gravitated around two main questions: how metadata is structured and how it is used by web publishers and search engines? These questions each have their roots in the twin characteristics of the ideal metadata system: consistent accuracy and consistent use. The question of structure is a debate over what constitutes *ideal accuracy* and how to achieve this accuracy, while the question of utilization deals with the practicalities of *ideal use*.

Though the debate surrounding the structure of metadata often seems to be a dispute over standards, it is fundamentally a debate over the purpose of metadata. On one side are the minimalists, who contend that metadata should be a very simple set of only a few elements so that it is equally useful across domains and resource types (Campbell, 2002; Lagoze, 2001). Proponents of this type of simple metadata argue that as metadata standards become more narrowly defined or require greater semantic complexity, they run the risk of becoming less rather than more useful. Search engines may have greater difficulty collating information from diverse sources (Lagoze, 2001), and creators may have greater difficulty describing their sites through metadata efficiently or effectively (Hillman, 2003).

On the other side of this debate, those in favor of stricter standards and more complex element sets argue that in order for search engines to perform either the locating or the collocating function for which they are designed, the metadata elements must be *consistent* (Chepesuik, 1999; Sokvitne, 2000; Tennant, 2004). Sokvitne (2000) points out that without some authority control there will be inconsistencies that automated systems cannot process effectively. Chepesuik (1999) and Tennant (2004) argue much more vehemently that without controlled vocabulary there can only be bibliographic chaos.

While these scholars debate the theoretical purpose and uses of metadata, other researchers look at how metadata is currently employed. These researchers study web publishers to see how metadata is created, and they study search engines to see how that created metadata is used and how it influences web page visibility. Visibility is one of the primary concerns of web page publishers, many of whom hire consultants to write titles, descriptions, and keyword lists that will increase the chances that their pages will rank near the top of search engine result lists (Richardson, 2003). Hundreds of companies offer services ranging from advice to consultations on how to improve customers' rank in search engine result lists, a process called search engine optimization (Zhang & Dimitroff, 2005a). (For examples, see "Search Engine Optimization 1-2-3", "Search Engine Optimization", Sullivan (2003), Yahoo.com.) The advice offered by these companies, however, is generally based on conventional wisdom rather than proof. Actual research in this area indicates that for most search engines those pages embedded with metadata achieve greater visibility than those that are not embedded with metadata (Turner & Brackbill, 1998; Zhang & Dimitroff, 2005b). Of the search engines tested by Zhang and Dimitroff (2004), only Alta Vista and AllTheWeb did not increase the rank of pages that had metadata over those that did not. Other search engines favored sites having metadata, especially those sites having Keyword, Title, or Description fields (Zhang & Dimitroff, 2005a, 2005b).

Based on the current accuracy of the metadata embedded in web pages, though, what is the potential that search engines will be able to accurately rank their result lists? Researchers interested in this question, such as Craven and Sokvitne, focus their studies on determining the type, amount, and quality of the metadata produced by those posting web pages on the Internet. Craven studies metadata use in general and the Description and Title fields in particular (Craven, 2000, 2001a, 2001b, 2001c, 2001d, 2002a, 2002b, 2003). This research shows that the content of the Description field is very similar to traditional abstracts in terms of language characteristics (Craven, 2000) and that descriptions often change over time as the site is updated (Craven, 2001a). Craven (2000) also notes that the Description tags he examined tended to include information about products and services, and he surmises that this must be because a high proportion of the sites he downloaded were commercial sites. Based on his analysis, which indicates that most authors do not blatantly misrepresent their pages through their Description fields (Craven, 2000), Craven (2000) concludes that general metadata quality is good enough to be useful to search engines as they index and display web search results.

Sokvitne's (2000) paper analyzes the descriptive metadata embedded in 100 Australian governmental and educational web pages. He looks specifically at the quality of metadata use, measuring it against standards of indexing as identified. Through this study, Sokvitne concludes that government and educational websites in Australia apply metadata inconsistently, describing their sites with widely varying degrees of success.

Where Sokvitne and Craven analyze the quality of *produced* metadata, other researchers examine attitudes toward metadata *production* and how these attitudes in turn influence the way metadata itself is perceived. Greenberg, Pattuelli, and Robertson (2001), for example, says that author-generated metadata is perceived to be of poor quality. Greenberg et al. (2001) tests this assumption and finds that, in fact, the majority of the authors she studied were able to create acceptable metadata after only minimal training.

The perception that authors misrepresent their pages through their metadata is so rampant among researchers and writers, however, that it is not thought necessary to prove the point or to cite proof. Richardson (2003) and Sherman (2002) use this perception to explain why search engines no longer rely on Keyword tags. Doctorow (2001) cites author ignorance and dishonesty as the primary reasons that metadata will never realize its full potential as an aid to resource discovery on the Internet. Yet, these authors do not cite studies showing the amount of metadata abuse that actually occurs.

Current research on metadata usage lacks a comprehensive investigation of the quality and characteristics of individual metadata fields as they are understood and employed by specific user groups. It also fails to recognize the usage trends of different categories of web publishers and authors. Because of this, it often excludes whole categories of publishers based on untested assumptions. And finally, it often draws conclusions about metadata use based on data samples that are either too undefined or too limited in size and scope to provide reliable information about actual metadata use on the Internet. Just as studying the nutritional content of food is less meaningful if the types of foods studied are not defined and characterized, so the study of metadata use on the Internet is less meaningful if the types of users are not defined and characterized.

Research on metadata, its purpose, and its utilization is important not only to web page authors and publishers wishing to increase the visibility of their websites, but it is also important to web search engine designers as they develop and modify their algorithms so that retrieval and ranking of result lists provide users with the most relevant web sites possible. Two important questions have not yet been answered, however. The first is, what is the current level of metadata quality and accuracy on the Web? In other words, do authors commonly misrepresent their pages through their metadata either intentionally or unintentionally? To date, no comprehensive study has been done to answer this important question even though the common perception is that author-generated metadata is suspect at best. The second question is, what are the trends and patterns in metadata quality and usage based on user group? For example, do sites maintained by information professionals actually have higher quality metadata (as Sokvitne assumed)? Without answers to these questions, search engine optimization services and search engine designers will have no concrete information about the actual state of metadata quality and accuracy on which to base their advice and algorithms. Henshaw and Valauskas (2001) assert that metadata is vital to the success of search engines. Metadata is meaningless, however, unless search engine designers know to what extent they can rely upon it and what usage trends currently exist on the Web.

Through this research we hope to gain

(a) A better understanding of the strengths and weaknesses of metadata as it is currently employed in Internet publishing.
(b) Evidence about current metadata use preferences and subject content analysis habits that may help to inform future improvements in search engine indexing and ranking algorithms.
(c) Evidence useful for future metadata standards revisions because of its comprehensive analysis of current usage patterns and trends.

(d) A new methodology for similar research. This method recognizes the diversity of web professional groups and integrates this recognition into the research methodology.

Toward this aim, the present study examines the following five hypotheses:

[1] There are no statistically significant differences with respect to Keyword accuracy among library and information science professionals, government agencies and major non-profit organizations, businesses and industries, and information technology professionals.
[2] There are no statistically significant differences with respect to Keyword characteristics among library and information science professionals, government agencies and major non-profit organizations, businesses and industries, and information technology professionals.
[3] There are no statistically significant differences with respect to metadata Description accuracy among library and information science professionals, government agencies and major non-profit organizations, businesses and industries, and information technology professionals.
[4] There are no statistically significant differences with respect to metadata Description characteristics among library and information science professionals, government agencies and major non-profit organizations, businesses and industries, and information technology professionals.
[5] There are no statistically significant differences with respect to metadata Title characteristics among library and information science professionals, government agencies and major non-profit organizations, businesses and industries, and information technology professionals.

Keyword metadata, description metadata, and title metadata of different professional groups are examined in the proposed hypotheses because these metadata are central to description of a webpage and have a strong impact on webpage visibility in search engine search results lists. For keyword metadata and description metadata both accuracy and characteristics are addressed separately.

2. Research method

2.1. Identifying professional domains

People in different professions often have different educational backgrounds and different professional goals. This will cause them to have different information organization and retrieval expertise and needs as well as different publishing emphasis and preferences. Their Internet publishing behaviors and awareness of metadata applications may therefore vary. It is important to recognize these differences, to identify and define user groups based on different professions, and to examine them separately in investigation and data analysis. In this way, research results will be more comprehensive and sound.

With this in mind, this study identifies, defines, and examines four distinct domains: library and information science (LIS), government agencies and major non-profit organizations (Gov/Org), businesses and industries (B&I), and information technology (IT). These domains are discussed in greater detail below.

One might assume that professionals in library and information science would highly value information organization and retrieval and would subsequently be aware of metadata and use it consistently and well when publishing pages on the Internet. This study will investigate the validity of this assumption by comparing the metadata of this domain to that of other domains. This group includes such publishers as public libraries, academic libraries, special libraries, information agencies, archives, metadata production professionals, and information centers.

Government agencies and major non-profit organizations are great potential users of the Internet. These organizations conjure images of accuracy and consistency both because they place great value in structured

inter-organizational information discovery and because they often have access to the personnel, research, and funding needed to create good metadata. This study will test these assumptions and determine the impact of these perceived characteristics on actual metadata production. This category includes federal, state, and local governments, governmental agencies, major non-profit organizations (such as the Red Cross and the United Nations), and military branches.

Businesses and industries are also important Internet user groups. They understand the potential impact of the Internet as a marketing and sales tool, so visibility becomes a high priority for these web publishers. This motive for web publishing is inherently suspect, however, because people assume that publishers manipulate the metadata in order to increase their visibility and sales. While many researchers have commented on this group's potential abuse of metadata, none have examined the metadata created by this group to determine the existence or extent of abuse or to compare the number of inaccuracies with those produced by web authors from other domains. This category includes large, medium, and small firms, businesses, and financial institutions.

The information technology domain has its own unique distinctions: people in this domain engage in the research, development, and application of the technology that powers the Internet, among other pursuits. They fully understand the importance of Internet technology and are often assumed to be aware of metadata and proficient in its use. People may not associate them, however, with knowledge of and practice in indexing and describing their sites through metadata. This category includes technology designers, programmers, researchers, and information technology related companies.

Institutions or agencies can straddle two or more of the domains defined above. For instance, the Library of Congress can be classified as both a library and a government agency. For the purposes of this study, when a site can fit into more than one domain, it is treated as a member of the first possible domain in the following hierarchy:

a. Businesses and industries (B&I),
b. Library and information science (LIS),
c. Government agencies and non-profit organizations (Gov/Org),
d. Information technology (IT).

This hierarchy attempts to take into account the primary motivations and expertise of those who are principally concerned in the essential functions of the site. For example, the Library of Congress is categorized in LIS domain rather than in the Gov/Org domain, giving preference to the information professionals who are engaged in the principle work of the Library of Congress rather than to the branch of government that supports the library's work.

2.2. Metadata elements examined

There are several metadata schemes currently in use, but two of the most common are Dublin Core and generic markup tags in the format <META name = "[*tag name*]" content = "[*metadata content*]"/>. This study examines the generic markup tags, which are much less structured than Dublin Core tags. Web authors can create new elements for this generic type of metadata as needed, and there is no centralized control system that defines or approves metadata elements. This means that the scheme can be as simple or as complex as the author wishes.[1]

[1] This generic metadata is also used much more frequently than is Dublin Core metadata (62.8% of the time as opposed to 7.4% of the time, according to the researchers' study of 2400 web pages).

Since the generic metadata structure does not have a predefined metadata element set, a pilot study was conducted to determine a potential metadata element set. This pilot study examined 800 web pages (200 from each domain) to see which tags were used consistently and what information they contained. The pilot study revealed the following elements, which were then investigated further in the full study.

Title: the distinguishing name metadata description in that page.
Author: the person or other entity responsible for the content of the site.
Publisher: the name of the person or other entity responsible for making the site's content available to the public.
Copyright: information about the person or other entity that holds the rights to the site's intellectual content, information about rights reserved to that person or entity, and/or when those rights became effective.
Rating: a description of the appropriateness of the site for different users, such as children or the general public.
Resource Type: the nature of the contents of the page, including terms that describe the general categories, functions, genres, or aggregate levels of the page.
Language: the language metadata description of the page.
Distribution: the intended scope of the resource described in terms of geographical location or jurisdiction.
Date: dates associated with the site (such as the date of creation, modification, publication, etc.).
Keyword/subject: words or phrases chosen to represent the content of the site.
Description: an account or summary of the page's content.
Miscellaneous: fields that are nonstandard (such as "owner", "area", and "destination") and fields that are administrative in nature (such as "approved-by", "site-product-code", "terminator", "department", "expires", "template-id", or "revisit-after").

2.3. Selection of web pages

After the four domains were defined, two methods were used to select 600 web pages from each domain, resulting in a total of 2400 selected web sites. The first method was to employ existing subject directories on the Internet, such as the Yahoo directory, to lead to lists of related web pages. The second method was to form basic search queries designed for high recall within the specified domains and then to use this search string to query major Internet search engines, such as Google. We did not examine all the result items from any given page of the result list and did not always begin with the first page of results or view consecutive pages of the result list. Half of the examined web pages in each domain were selected from subject directories and half from major search engines.

In order to avoid bias and generate valid results, we tried to randomly select web pages for four groups. For instance, we randomly selected web pages from a search engine result list, and randomly picked up directories of Yahoo and websites as well.

2.4. Examination of selected pages

After selecting the web pages, we examined the metadata embedded in the pages' markup contents both for their semantic and their syntax characteristics. For each selected page, metadata elements were examined, analyzed, and recorded. Although all available metadata elements were examined during this investigation, the Title, Keyword, and Description fields received particular scrutiny. According to previous studies (Zhang & Dimitroff, 2004, 2005a, 2005b), search engines are more sensitive to these elements than

to other elements, causing these fields to impact the visibility of web pages more dramatically than do other metadata fields. It is therefore important that the content of these fields be accurate. Choosing poor quality terms to include here may negatively influence search engines' abilities to retrieve relevant items in response to end user queries. These fields also contain subjective representations of each page's contents, representations influenced by the author's preferences, background, and knowledge and expertise in indexing and information retrieval.

In order to effectively measure and record the quality of the metadata embedded in the examined web pages, we developed evaluative criteria for each element. The Keyword and Description fields were each assigned a value from a 5-point Likert scale to enable statistical comparison. In this case 1, 2, 3, 4, and 5 stand for "very accurate", "accurate", "minimally accurate", "not very accurate", and "inaccurate". Descriptions and Keyword terms were also analyzed to determine their granularity as compared to the specificity of the contents of the web page. Each term in these fields was characterized either as narrow, broad, incorrect, correct, or as duplicate. Duplicate in these fields means that the same term appears in the same field. Finally, the contents of the Title fields were compared with the titles prominently displayed on the web page or on the browser application's title bar. Each Title field was then defined as correct, partially correct, or incorrect. The assigned values and characteristics therefore become the measurements investigated below. In order to ensure that these measurements were consistently applied to every field evaluated, one researcher did all of the information evaluation.

2.5. Data analysis

In this study, we assume that the involved dependent variables (Keyword accuracy, Description accuracy, Keyword characteristics, Description characteristics, Title characteristics) are normally distributed, the population variances of the dependent variable are the same for all cells, the case represents random samples, and the values of the dependent variables are independent of each other. The independent variable is the professional domain of the web page.

The significance level (p) for all tests is 0.05. Regardless of the specific statistical approach used, if p is smaller than 0.05, the finding is statistically significant and the null hypothesis is rejected.

This study employs *ANOVA* and *Chi-Square* methods, depending on the nature of the measurement and the compared objects, to examine the proposed hypotheses. The *ANOVA* technique produces an analysis of variance for a quantitative dependent variable by independent variables. Analysis of variance is used to test the hypothesis that several means are equal. The *chi-square* test procedure tabulates a variable into categories and computes a chi-square statistic. This test compares the observed and expected frequencies in each category to test either that all categories contain the same proportion of values or that each category contains a user-specified proportion of values.

3. Result analysis

3.1. General descriptive analysis of the investigated web pages

3.1.1. Distribution of metadata element occurrence in the four defined domains

Since metadata is not required for web page publication, not all investigated web pages had embedded metadata. In fact, only 51.17% of LIS pages, 66.67% of Gov/Org pages, 67% of B&I pages, and 66.5% of IT pages contain embedded metadata. In total, 62.83% of all examined web sites contained embedded HTML metadata. What is more, of those pages having metadata, not all pages used all of the defined metadata elements. Table 1 provides an overview of the preferences each of the four professional domains exhibits for metadata element selection. This analysis does not indicate the quality of metadata implementation.

Table 1
Distributions of metadata elements

Elements		Domain				Total
		LIS	Gov/Org	B&I	IT	
Author	Count	76	86	61	78	301
	% within elements	25.2	28.6	20.3	25.9	100.0
	% within domain	13.9	9.8	6.7	7.6	8.9
	% of total	2.3	2.6	1.8	2.3	8.9
Publisher	Count	3	9	10	14	36
	% within elements	8.3	25.0	27.8	38.9	100.0
	% within domain	0.5	1.0	1.1	1.4	1.1
	% of total	0.1	0.3	0.3	0.4	1.1
Miscellaneous	Count	59	122	78	100	359
	% within elements	16.4	34.0	21.7	27.9	100.0
	% within domain	10.8	13.9	8.6	9.7	10.7
	% of total	1.8	3.6	2.3	3.0	10.7
Copyright	Count	21	13	30	44	108
	% within elements	19.4	12.0	27.8	40.7	100.0
	% within domain	3.8	1.5	3.3	4.3	3.2
	% of total	0.6	0.4	0.9	1.3	3.2
Rating	Count	11	19	20	37	87
	% within elements	12.6	21.8	23.0	42.5	100.0
	% within domain	2.0	2.2	2.2	3.6	2.6
	% of total	0.3	0.6	0.6	1.1	2.6
Resource type	Count	7	10	11	12	40
	% within elements	17.5	25.0	27.5	30.0	100.0
	% within domain	1.3	1.1	1.2	1.2	1.2
	% of total	0.2	0.3	0.3	0.4	1.2
Language	Count	7	17	13	22	59
	% within elements	11.9	28.8	22.0	37.3	100.0
	% within domain	1.3	1.9	1.4	2.1	1.8
	% of total	0.2	0.5	0.4	0.7	1.8
Distribution	Count	7	15	19	27	68
	% within elements	10.3	22.1	27.9	39.7	100.0
	% within domain	1.3	1.7	2.1	2.6	2.0
	% of total	0.2	0.4	0.6	0.8	2.0
Date	Count	10	0	12	5	27
	% within elements	37.0	0.0	44.4	18.5	100.0
	% within domain	1.8	0.0	1.3	0.5	0.8
	% of total	0.3	0.0	0.4	0.1	0.8
Keyword	Count	175	294	332	351	1152
	% within elements	15.2	25.5	28.8	30.5	100.0
	% within domain	31.9	33.6	36.5	34.0	34.2
	% of total	5.2	8.7	9.9	10.4	34.2
Description	Count	160	272	312	328	1072
	% within elements	14.9	25.4	29.1	30.6	100.0
	% within domain	29.2	31.1	34.3	31.8	31.8
	% of total	4.8	8.1	9.3	9.7	31.8

Table 1 (*continued*)

Elements		Domain				Total
		LIS	Gov/Org	B&I	IT	
Title	Count	12	19	11	15	57
	% within elements	21.1	33.3	19.3	26.3	100.0
	% within domain	2.2	2.2	1.2	1.5	1.7
	% of total	0.4	0.6	0.3	0.4	1.7
Total	Count	548	876	909	1033	3366
	% within elements	16.3	26.0	27.0	30.7	100.0
	% within domain	100.0	100.0	100.0	100.0	100.0
	% of total	16.3	26.0	27.0	30.7	100.0

In Table 1, ''Count'' refers to the number of pages containing a given metadata element in a given domain. For instance, the Author element appeared in 76 of the 600 LIS pages visited. The ''% within elements'' analysis shows the ratio of the occurrence of a given metadata element in a given domain to the occurrence of that metadata element across all four domains. For instance, the number of Author elements found in the LIS domain accounts for 25.2% of the Author fields found in all four domain (76/301 = 25.2%, where 76 is the number of author elements found in the given domain and 301 is total number of author elements across all four domains). The ''% within domain'' comparison shows the ratio of the occurrence of a given metadata element in a given domain to the occurrence of all metadata elements in that particular domain. For instance, the Author element in the LIS domain accounts for 13.9% of all metadata elements found in LIS pages (76/548 = 13.9%, where 76 is the number of author elements found in the given domain and 548 is the number of all the metadata fields found in the library and information science domain). The ''% of total'' is the ratio of a given metadata element within a given domain to the total of all elements in all domains. For example, the Author element in the LIS domain accounts for 2.3% of all the metadata fields found in all the domains during the course of this study (76/3366 = 2.3%, where 76 is the number of author fields found in the given domain and 3366 is the total number of metadata fields found in all domains).

Table 1 reveals certain preferences that web authors have regarding metadata. In general, publishers prefer to include a Keyword, Description, Miscellaneous, or Author field. The Title, Language, Resource Type, Publisher, and Date fields, on the other hand, are used comparatively rarely. This suggests that web publishers assume that subject-oriented fields are more important to the search engines are than fact-oriented fields.[2]

Within each individual domain, other domain-specific preference becomes evident. The LIS domain favors Keyword, Description, and Author fields while paying much less attention to the Publisher, Language, and Distribution elements. The Gov/Org domain prefers to use the Keyword, Description, and Miscellaneous elements while Publisher and Resources Type were hardly ever used, and Date was never used. The B&I domain, like the Gov/Org domain, preferred Keyword, Description, and Miscellaneous fields, but this domain often ignored the Publisher, Resources Type, and Title elements. And just as in the Gov/Org and B&I domains, the IT domain gave distinct preference to the Keyword, Description, and Miscellaneous fields. This domain chose not to use the Date, Resources Type, and Publisher fields frequently.

Notice that both Keyword and Description are always preferred in all 4 domains. This is consistent with the high value associated with these elements when considered from the information retrieval perspective. Date, Publisher, and Resources Type ranked the lowest for all four domains. For the purposes of

[2] It is interesting to note that the title element is one of the least frequently applied metadata elements (occurring only 1.7% of the time) even though it is very important to web indexers. One explanation for this phenomenon may be that most web pages have HTML Title tags, and that web authors assume that adding title information to the metadata elements is redundant.

information retrieval, these fact-oriented elements are rarely used as direct access points. This may cause web authors to pay little attention to these fields when selecting metadata elements for inclusion in the web page's code. Web publishers in the LIS domain were the only publishers to place the author element among their top three elements, possibly because it is a traditional access point for those trained in an environment that often locates, sorts, and gathers materials based on the name of the author.

A different measure of the relative value of individual elements to the four domains appears in the "% within elements" comparison in Table 1. This shows that Gov/Org pages contained more Author elements, Miscellaneous fields, and Title fields than did any other domain. B&I pages had a higher percentage of Date fields than any other domain. IT pages contained a higher percentage of all the other fields (Description, Keyword, Distribution, Language, Resource Type, Rating, Copyright, and Publisher). Interestingly, LIS pages did not achieve the highest score in any of the investigated metadata fields.

Many sites in all domains included fields categorized as Miscellaneous. This may be because unlike the more structured Dublin Code metadata system, this less structured metadata system allows publishers to create or choose metadata elements at will. This flexibility may cause inconsistency and diversity of embedded metadata elements. These elements were often organization-specific, providing information useful for intra-organizational record keeping or processing.

3.1.2. Metadata element co-occurrence in the four defined domains

Because there are no standardized rules that mandate a set of required metadata elements, web page publishers can choose the number and kind of metadata elements based on their own preferences and needs. Since there are twelve different metadata elements, each web page containing metadata uses one to twelve of those elements. Table 2 and Fig. 1 show that most web pages include only two metadata elements. When the number of metadata elements increases beyond two elements, the corresponding count decreases across all the domains. Similarly, when the number of combined metadata elements decreases below two elements, the corresponding count decreases across all the domains.

Pages in the LIS domain generally contain fewer metadata elements than other domains. This could indicate that library and information science professionals prefer to use elements more economically.

Looking at individual elements, the Keyword, Author, and Description elements often appear together. The three elements appear together in 16.16% of the pages visited. The co-occurrence rates for the combinations of Author and Keyword, Author and Description, and Keyword and Description elements across the four domains are 18.51%, 16.92%, and 76.63% respectively. This means that over 75% of all pages having metadata include at least the Keyword and Description elements. They may add other elements to these, but these are the core elements of most embedded metadata.

Within individual domains, the most common combination of metadata elements is that of Keyword and Description, occurring in 60.34% of LIS sites, 74.13% of Gov/Org sites, 83.15% of B&I sites, and 82.63% of IT sites. The Author and Keyword elements appear in 20.26% of LIS pages, 20.64% of Gov/Org pages, 15.19% of B&I pages, and 18.68% of IT pages. The Author and Description elements appear in combination in 19.83% of LIS pages, 17.73% of Gov/Org pages, 13.81% of B&I pages, and 17.37% of IT pages. And finally, the Author, Description, and Keyword elements appear together in 16.81% of LIS pages, 17.44% of Gov/Org pages, 13.54% of B&I pages, and 17.11% of IT pages.

3.2. Analysis of the metadata Keyword element among the four domains

3.2.1. Analysis of Keyword accuracy

This comparison examines the hypothesis that there are no statistically significant differences with respect to Keyword accuracy among the four domains. The independent variable is the defined domain and the dependent variable is the accuracy of the Keyword field. For each page that contains the Keyword element, the list of keywords was examined as a whole and assigned a single accuracy value from 1 to 5 on a

Table 2
Elements co-occurrence analysis

Co-occurrences		Domain				Total
		LIS	Gov/Org	B&I	IT	
1 element	Count	62	56	45	45	208
	% within co-occurrence	29.8	26.9	21.6	21.6	100.0
	% within domain	26.7	16.3	12.4	11.8	15.8
2 elements	Count	99	159	218	201	677
	% within co-occurrence	14.6	23.5	32.2	29.7	100.0
	% within domain	42.7	46.2	60.2	52.9	51.4
3 elements	Count	34	74	43	62	213
	% within co-occurrence	16.0	34.7	20.2	29.1	100.0
	% within domain	14.7	21.5	11.9	16.3	16.2
4 elements	Count	16	22	22	25	85
	% within co-occurrence	18.8	25.9	25.9	29.4	100.0
	% within domain	6.9	6.4	6.1	6.6	6.4
5 elements	Count	9	18	12	17	56
	% within co-occurrence	16.1	32.1	21.4	30.4	100.0
	% within domain	3.9	5.2	3.3	4.5	4.2
6 elements	Count	8	9	10	12	39
	% within co-occurrence	20.5	23.1	25.6	30.8	100.0
	% within domain	3.4	2.6	2.8	3.2	3.0
7 elements	Count	3	2	7	8	20
	% within co-occurrence	15.0	10.0	35.0	40.0	100.0
	% within domain	1.3	0.6	1.9	2.1	1.5
8 elements	Count	1	2	3	4	10
	% within co-occurrence	10.0	20.0	30.0	40.0	100.0
	% within domain	0.4	0.6	0.8	1.1	0.8
9 elements	Count	0	2	2	5	9
	% within co-occurrence	0.0	22.2	22.2	55.6	100.0
	% within domain	0.0	0.6	0.6	1.3	0.7
10 elements	Count	0	0	0	1	1
	% within co-occurrence	0.0	0.0	0.0	100.0	100.0
	% within domain	0.0	0.0	0.0	0.3	0.1
11 elements	Count	0	0	0	0	0
	% within co-occurrence	100.0	100.0	100.0	100.0	100.0
	% within domain	0.0	0.0	0.0	0.0	0.0
12 elements	Count	0	0	0	0	0
	% within co-occurrence	100.0	100.0	100.0	100.0	100.0
	% within domain	0.0	0.0	0.0	0.0	0.0
Totals	Count	232	344	362	380	1318
	% within co-occurrence	17.6	26.1	27.5	28.8	100.0
	% within domain	100.0	100.0	100.0	100.0	100.0

5-point scale. The lower the value of the assigned number, the more accurate the keyword list is and vice versa. Since this analysis involves four comparison variables and nature of the measurement, an *ANOVA* test was used. Tables 3 and 4 show the results of the comparison.

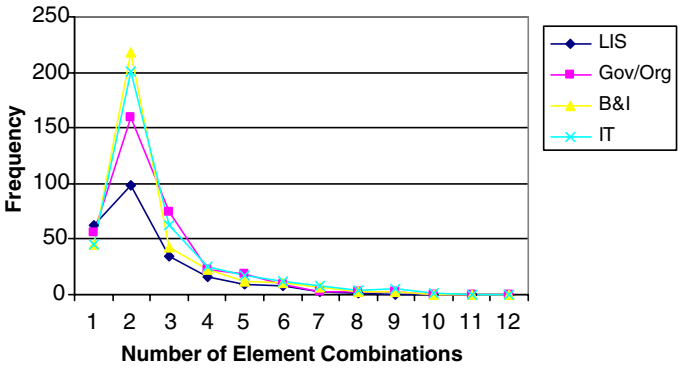

Fig. 1. Chart for co-occurrence analysis.

Table 3
Descriptive information for Keyword accuracy

	N	Mean	Std. deviation	Std. error	95% confidence interval for mean		Minimum	Maximum
					Lower bound	Upper bound		
LIS	181	2.5912	1.10992	0.08250	2.4284	2.7540	1.00	5.00
Gov/Org	296	2.5541	0.88132	0.05123	2.4532	2.6549	1.00	5.00
B&I	318	2.3931	0.76159	0.04271	2.3091	2.4771	1.00	5.00
IT	355	2.5775	0.80714	0.04284	2.4932	2.6617	1.00	5.00
Total	1150	2.5226	0.87162	0.02570	2.4722	2.5730	1.00	5.00

Table 4
ANOVA test result for Keyword accuracy

	Sum of squares	df	Mean square	*F*	Sig.
Between groups	7.547	3	2.516	3.331	0.019
Within groups	865.365	1146	0.755		
Total	872.912	1149			

Table 3 shows descriptive information regarding Keyword accuracy for the four groups respectively. In the table, the four groups had the same minimum and maximum values. LIS and B&I had the largest and smallest means respectively. LIS and B&I had the largest and smallest standard deviation values respectively.

In Table 4, since the *p*-value is 0.019 (<0.05, $F = 3.331$), this hypothesis was rejected. Because of this, post hoc multiple comparisons (*Tukey* honestly significant differences (HSD)) were conducted to evaluate pair-wise differences among the means. The data displayed in Table 5 indicate that the mean difference $(I - J)$ for B&I against IT is -0.1844^* which is negative and significant. This means that the B&I domain includes more accurate keywords, statistically, than does the IT domain. This clearly contributes to the hypothesis rejection. In fact, the B&I domain creates more accurate keywords than do either the LIS domain (−0.1981) Gov/Org domain (−0.1610). These differences are not statistically significant, however. Surprisingly, the LIS domain performed the worst across the four groups. All of the values in mean difference $(I - J)$ column were positive.

Table 5 shows multiple comparison regarding Keyword accuracy for the four groups respectively.

Table 5
Multiple comparison of Tukey HSD for Keyword accuracy

(*I*) Domain	(*J*) Domain	Mean difference (*I* − *J*)	Std. error	Sig.	95% confidence interval	
					Lower bound	Upper bound
LIS	Gov/Org	0.0371	0.08199	0.969	−0.1738	0.2481
	B&I	0.1981	0.08091	0.069	−0.0101	0.4062
	IT	0.0137	0.07937	0.998	−0.1905	0.2179
Gov/Org	LIS	−0.0371	0.08199	0.969	−0.2481	0.1738
	B&I	0.1610	0.07018	0.100	−0.0196	0.3415
	IT	−0.0234	0.06840	0.986	−0.1994	0.1526
B&I	LIS	−0.1981	0.08091	0.069	−0.4062	0.0101
	Gov/Org	−0.1610	0.07018	0.100	−0.3415	0.0196
	IT	−0.1844*	0.06709	0.031	−0.3570	−0.0118
IT	LIS	−0.0137	0.07937	0.998	−0.2179	0.1905
	Gov/Org	0.0234	0.06840	0.986	−0.1526	0.1994
	B&I	0.1844*	0.06709	0.031	0.0118	0.3570

* The mean difference is significant at the 0.05 level.

Table 6
Homogeneous subsets for Keyword accuracy

Domain	*N*	Subset for alpha = 0.05	
		1	2
B&I	318	2.3931	
Gov/Org	296	2.5541	2.5541
IT	355	2.5775	2.5775
LIS	181		2.5912
Sig.		0.0670	0.9600

Table 6 presents the post hoc test data by showing sets of means that differ significantly from each other. In this table, the means for the groups in homogeneous subsets are displayed. The Harmonic Mean Sample Size for this analysis is equal to 269.092. The group sizes are unequal so the harmonic mean of the group sizes is used. Type I error levels are not guaranteed. In this case, there are two homogeneous subsets. LIS, Gov/Org, and B&I are within the first homogeneous subset while Gov/Org, B&I, and IT are within the second.

In summary, the results indicated that there were significant differences regarding Keyword accuracy among the four groups.

3.2.2. Analysis of Keyword characteristics

This analysis examines the hypothesis that there are no statistically significant differences with respect to Keyword characteristics among the four domains. Each term contained in the Keyword fields found during this study was designated as a narrow, broad, incorrect, correct, or duplicate term. Many web pages containing the Keyword field contain some combination of these five categories of terms. In order to examine the hypothesis, the chi-square statistical method was employed to determine if there is a relationship between these five categorical variables.

Both the analysis of Keyword characteristics and the analysis of Keyword accuracy address the same problem but from different perspectives. They are complementary. A given keyword can be both "not very accurate" and narrow, for example. The following analysis cannot, however, determine the *degree* of broadness or narrowness for any given term. Table 7 shows the case processing summary for this analysis.

Table 7
Case processing summary for Keyword characteristics

	Cases					
	Valid		Missing		Total	
	N	Percent	*N*	Percent	*N*	Percent
Subject characteristics * domain	28,484	100.0	2	0.0	28,486	100.0

Table 8
Keyword characteristics * domain cross-tabulation

Keyword characteristics		Domain				Total
		LIS	Gov/Org	B&I	IT	
Narrow	Count	375	2968	2286	3561	9190
	Expected count	637.9	2554.6	3137.3	2860.2	9190.0
	% within subject characteristics	4.1	32.3	24.9	38.7	100.0
	% within domain	19.0	37.5	23.5	40.2	32.3
Broad	Count	1113	4320	6756	4503	16692
	Expected count	1158.5	4640.1	5698.4	5195.0	16692.0
	% within subject characteristics	6.7	25.9	40.5	27.0	100.0
	% within domain	56.3	54.6	69.5	50.8	58.6
Incorrect	Count	375	29	283	299	986
	Expected count	68.4	274.1	336.6	306.9	986.0
	% within subject characteristics	38.0	2.9	28.7	30.3	100.0
	% within domain	19.0	0.4	2.9	3.4	3.5
Correct	Count	104	567	273	297	1241
	Expected count	86.1	345.0	423.7	386.2	1241.0
	% within subject characteristics	8.4	45.7	22.0	23.9	100.0
	% within domain	5.3	7.2	2.8	3.4	4.4
Duplicate	Count	10	34	126	205	375
	Expected count	26.0	104.2	128.0	116.7	375.0
	% within subject characteristics	2.7	9.1	33.6	54.7	100.0
	% within domain	0.5	0.4	1.3	2.3	1.3
Total	Count	1977	7918	9724	8865	28484
	Expected count	1977.0	7918.0	9724.0	8865.0	28484.0
	% within subject characteristics	6.9	27.8	34.1	31.1	100.0
	% within domain	100.0	100.0	100.0	100.0	100.0

In Table 8, the definitions of "count", "% within subject characteristics", and "% within domain" are similar to those in previous sections. The "Expected count" reflects the results of Eq. (1). In Eq. (1), EC, RT, CT, and GT stand for "expected count", "row total", "column total", and "grand total" respectively. For instance, in Table 8, the expected count for narrow terms in the LIS group is equal to 637.9 (see calculation in Eq. (2) where RT = 9190, CT = 1977, GT = 28484). Eqs. (1) and (2) show how to calculate the expected count and an example respectively:

$$\mathrm{EC} = \frac{\mathrm{RT} \times \mathrm{CT}}{\mathrm{GT}}, \tag{1}$$

$$\mathrm{EC} = \frac{9190 \times 1977}{28,484} = 637.9. \tag{2}$$

Table 8 reveals certain preferences that each domain has regarding the specificity of its keywords. The LIS group, for example, has both the highest narrow term rate (38.7%) and the highest incorrect term rate (38.0%) compared to the other groups. The B&I group has the highest broad term rate (40.5%). Gov/Org domain has the highest correct term rate (45.7%) while the IT group has highest duplicate term rate (54.7%).

In all four domains, web page authors include more broad terms and fewer duplicate terms than any other category of terms. Within the LIS group, the broad and duplicate term rates are 56.3% and 0.5% respectively. In the Gov/Org domain 54.6% of the terms were broad while only 0.4% were duplicate terms. Similarly, in the B&I and IT domains the rates for broad and duplicate terms were 69.5% and 1.3% for B&I, and 50.8% and 2.3% for IT. When compared across the four domains, the percentages for narrow, broad, correct, and duplicate terms are 32.3%, 58.6%, 4.4%, and 1.3% (see "% within domain" in Table 8). Incorrect terms accounted for only 3.5% of all the keywords found in the course of this study.

Both broad terms and narrow terms dominate metadata keyword lists. This may be because web authors and publishers try to include terms that their target audience will use when searching for sites. In doing so, authors and publishers create lists that they hope will help to fulfill both the locating and gathering functions (i.e., lists that contain terms that are specific to the page and terms that are common to several related pages). This way, they hope to include terms that can be used either by a searcher trying to find a specific page or by a searcher trying to find a set of related pages. Moreover, there are relatively few terms that correctly describe any given web page, and even though those few terms would be sufficient to adequately describe the page, web authors and publishers may not wish to make searchers guess these particular terms. They often create keywords lists larger than might be beneficial for search engines in hopes of increasing the chances that searchers will find their sites. This may explain why there are relatively few correct terms in pages from each domain.

The incorrect terms found during this study appeared to be assigned both by accident and by design, although the majority of those found for this study appeared to have been assigned through error or carelessness rather than design. For example, terms may have been borrowed from another page in the web site (so, for example, they may describe a site's home page but not the "about us" page that is currently being examined), or terms may have been misspelled.

The results of the Pearson chi-square test demonstrates that there is a statistically significant difference between the four domains ($p = 0.000 < 0.05$, see Table 9). This implies that the present hypothesis should be rejected. The statistical result table may present a p value of 0.000 because the system can only produce approximate p values for calculation. In other words, numbers smaller than 0.0005 can be rounded down to 0.000.

Because multiple keywords were investigated for most pages, the sample may be somewhat lumpy which is not considered in the statistical tests.

3.2.3. Analysis of keyword number per web page

Most pages that include the Keyword metadata element use more than one keyword to describe the web page. The maximum number of keywords found was 314 while the minimum number of keywords (in pages

Table 9
Chi-square rest result for Keyword characteristics

	Value	df	Asymp. Sig. (2-sided)
Pearson chi-square	2836.268[a]	12	0.000
Likelihood ratio	2281.613	12	0.000
Linear-by-linear association	113.059	1	0.000
N of valid cases	28,484		

[a] 0 cells (0.0%) have expected count less than 5. The minimum expected count is 26.03.

having keywords) was one. The average number of metadata keywords included in pages having at least one keyword field was 24.98. The maximum metadata numbers for the LIS, Gov/Org, B&I, and IT domains were 76, 304, 314, and 240 respectively. The average numbers of metadata keywords for the LIS, Gov/Org, B&I, and IT domains were 11, 26.3, 29.9, and 26.2 respectively. The modes for the LIS, Gov/Org, B&I, and IT domains were 5, 12, 11, and 12 respectively.

The B&I group uses more keywords in their metadata than any other group. This may indicate that they hope to increase the chances that searchers will find their sites, and subsequently purchase their products, by including as many searchable terms as possible. The LIS group, on the other hand, uses fewer metadata keywords compared to other groups. This may be influenced by traditional bibliographic indexing policies which limit the number of subject terms that can be assigned to any given item, or it may reflect knowledge that including more keywords does not necessarily benefit retrieval. See Fig. 2 for a summary of the number of keywords included in the examined web pages.

3.3. Analysis of the metadata Description element among the four domains

3.3.1. Analysis of Description accuracy

When web pages included the Description element in their metadata, that element was assigned a numeric value from a 5-point Likert scale designed to measure its accuracy. In this scale, lower values correspond to higher accuracy while higher values correspond to lower accuracy. In order to test the hypothesis that there are no statistically significant differences between the metadata description accuracy among the four domains, an *ANOVA* test was conducted to compare the four independent variables (the professional domains). The dependent variable in this test is the Description accuracy value assigned to each Description field found. A given web page could contain more than one Description field.

Table 10 shows the descriptive summary of the accuracy values assigned. The four groups had the same minimum and maximum values. LIS and Gov/Org had the largest and smallest means respectively. LIS and B&I had the largest and smallest standard deviation values respectively. Table 11 shows the results of the *ANOVA* test. The hypothesis was rejected because the p-value (0.000) was smaller than 0.05 ($F = 5.999$).

In order to understand what caused the hypothesis rejection, we carried out post hoc tests to compare all groups of subjects with each other. Table 12 shows the results of *Tukey's HSD*. In this table, each group of subjects was compared with the other three groups.

Table 12 shows the differences between group means for each pair of domains. This analysis indicates that the mean difference between the Gov/Org domain and LIS domain is -0.3893^{*} while the mean difference between the B&I domain and the LIS domain is -0.3517^{*}. Both of the differences are negative and statistically significant. In other words, both the Gov/Org group and the B&I group outperformed the LIS group in terms of Description accuracy. These differences led to the hypothesis rejection. Although

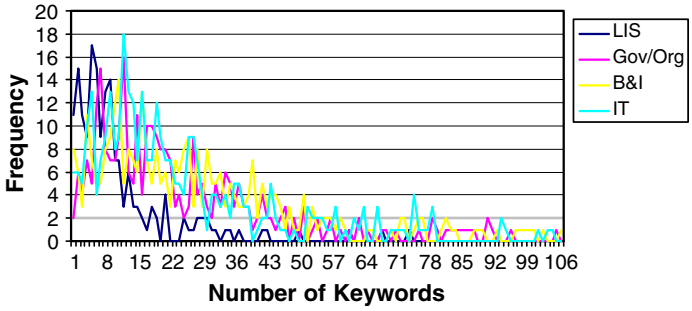

Fig. 2. Display of keyword number distribution.

Table 10
Descriptive summary for Description accuracy

	N	Mean	Std. deviation	Std. error	95% confidence interval for mean		Minimum	Maximum
					Lower bound	Upper bound		
LIS	169	2.3077	1.34960	0.10382	2.1027	2.5126	1.00	5.00
Gov/Org	282	1.9184	1.05224	0.06266	1.7951	2.0418	1.00	5.00
B&I	318	1.9560	0.97182	0.05450	1.8488	2.0632	1.00	5.00
IT	338	2.1243	1.02893	0.05597	2.0142	2.2343	1.00	5.00
Total	1107	2.0515	1.08168	0.03251	1.9877	2.1153	1.00	5.00

Table 11
ANOVA results for Description accuracy

	Sum of squares	df	Mean square	F	Sig.
Between groups	20.776	3	6.925	5.999	0.000
Within groups	1273.289	1103	1.154		
Total	1294.065	1106			

Table 12
Multiple comparisons of Tukey HSD for Description accuracy

(I) Domain	(J) Domain	Mean difference (I − J)	Std. error	Sig.	95% confidence interval	
					Lower bound	Upper bound
LIS	Gov/Org	0.3893*	0.10452	0.001	0.1203	0.6582
	B&I	0.3517*	0.10228	0.003	0.0886	0.6149
	IT	0.1834	0.10122	0.268	−0.0770	0.4439
Gov/Org	LIS	−0.3893*	0.10452	0.001	−0.6582	−0.1203
	B&I	−0.0375	0.08788	0.974	−0.2637	0.1886
	IT	−0.2058	0.08665	0.082	−0.4288	0.0171
B&I	LIS	−0.3517*	0.10228	0.003	−0.6149	−0.0886
	Gov/Org	0.0375	0.08788	0.974	−0.1886	0.2637
	IT	−0.1683	0.08394	0.187	−0.3843	0.0477
IT	LIS	−0.1834	0.10122	0.268	−0.4439	0.0770
	Gov/Org	0.2058	0.08665	0.082	−0.0171	0.4288
	B&I	0.1683	0.08394	0.187	−0.0477	0.3843

* The mean difference is significant at the 0.05 level.

the difference is not statistically significant, the IT group also performed better than the LIS group (mean difference $(I - J) = -0.1834$) in terms of Description accuracy. In fact, pages from the LIS domain were statistically less accurate than those from any other domain, while pages from the Gov/Org domain were statistically more accurate than those from any other domain.

Table 13 displays the means for the homogeneous subsets for Description accuracy. In Table 13 the harmonic mean sample size is equal to 256.962. The group sizes are unequal, so the harmonic mean of the group sizes is used. Type I error levels are not guaranteed. These tests display subsets of groups that have similar means. The *Tukey's* test creates two subsets of groups with statistically similar means. The first

Table 13
Homogeneous subsets for Description accuracy

Domain	*N*	Subset for alpha = 0.05	
		1	2
Gov/Org	282	1.9184	
B&I	318	1.9560	
IT	338	2.1243	2.1243
LIS	169		2.3077
Sig.		0.132	0.214

group includes the Gov/Org, B&I, and IT domains. The second subset includes the IT and LIS domains. This analysis shows that the first subset is statistically more accurate than the second because the first has relatively lower means.

3.3.2. Analysis of Description characteristics

This section examines the hypothesis that there are no statistically significant differences with respect to Description characteristics among the four domains. For those web pages that used the Description element, each description was characterized as narrow, broad, incorrect, correct, or duplicate. A given web page could contain more than one Description field. In order to compare these nominal values, the chi-square statistic method was employed.

Table 14 illustrates the case processing summary of the metadata Description characteristics.

Table 15 shows detailed information about the Description characteristics. In this table, the definitions of "count", "expected count", "% within description characteristics", and "% within domain" are similar to those in previous analysis results. Among the four groups, the broad, correct, narrow and incorrect, and duplicate rates were 48.9%, 37.7%, 9.7%, 3.7%, and 0.1% respectively. This suggests that web authors and publishers prefer broad descriptions when creating Description fields. The rate for narrow Description fields was not high (9.7%), suggesting that web authors do not want to exclude concepts from their Description fields. Notice that only the Gov/Org domain created more correct descriptions than incorrect, broad, narrow, or duplicate descriptions. In other domains, broad Description fields dominated all the other categories. The B&I group outperformed the other groups because it has the highest correct description rate (42.6%) and a relatively low incorrect description rate (2.8%). The IT domain performed the poorest because it has the lowest correct rate (22.9%) and a relatively high narrow rate (12.5%) and broad rate (57.1%) compared to its peer groups. It is also the only domain to include duplicate descriptions.

By reviewing the data in Table 15, certain trends become apparent. The Gov/Org domain produces a greater number of both narrow (39.6%) and incorrect (36.6%) Description fields than any other domain while the B&I group produces the greatest number of correct Description fields (32.6%). The IT group produces the greatest number of broad Description fields (35.1%). The comparison across domains demonstrates the same results that other cross-domain comparisons have shown: the B&I group produces the

Table 14
Case processing summary for Description characteristics

	Cases					
	Valid		Missing		Total	
	N	Percent	*N*	Percent	*N*	Percent
Description characteristics of * domain	1097	99.9	1	0.1	1098	100.0

Table 15
Description characteristics * domain cross-tabulation

Description characteristics		Domain				Total
		LIS	Gov/Org	B&I	IT	
Narrow	Count	13	42	10	41	106
	Expected count	16.3	27.2	30.6	31.8	106.0
	% within description characteristics	12.3	39.6	9.4	38.7	100.0
	% within domain	7.7	14.9	3.2	12.5	9.7
Broad	Count	75	110	163	188	536
	Expected count	82.5	137.7	154.7	161.0	536.0
	% within description characteristics	14.0	20.5	30.4	35.1	100.0
	% within domain	44.4	39.0	51.4	57.1	48.9
Incorrect	Count	12	15	9	5	41
	Expected count	6.3	10.5	11.8	12.3	41.0
	% within description characteristics	29.3	36.6	22.0	12.2	100.0
	% within domain	7.1	5.3	2.8	1.5	3.7
Correct	Count	69	115	135	95	414
	Expected count	63.7	106.3	119.5	124.2	414.0
	% within description characteristics	16.7	27.8	32.6	22.9	100.0
	% within domain	40.8	40.8	42.6	28.9	37.7
Duplicate	Count	0	0	0	1	1
	Expected count	0.2	0.3	0.3	0.3	1.0
	% within description characteristics	0.0	0.0	0.0	100.0	100.0
	% within domain	0.0	0.0	0.0	0.3	0.1
Total	Count	169	282	317	330	1098
	Expected count	169.0	282.0	317.0	330.0	1098.0
	% within description characteristics	15.4	25.7	28.9	30.0	100.0
	% within domain	100.0	100.0	100.0	100.0	100.0

most complete and accurate descriptive metadata while the LIS group produces the least complete and accurate descriptive metadata.

Table 16 shows that the proposed hypothesis was rejected because the significance value of the Pearson chi-square ($p = 0.000$) is smaller than 0.05. This means that the four domains exhibit different behaviors in terms of metadata description characteristics.

3.4. Analysis of metadata Title element characteristics among the four domains

The analysis in this section tests the hypothesis that there are no statistically significant differences with respect to metadata Title element characteristics among the four domains. The independent variables are

Table 16
Chi-square tests for Description characteristics

	Value	df	Asymp. Sig. (2-sided)
Pearson chi-square	58.528[a]	9	0.000
Likelihood ratio	62.909	9	0.000
Linear-by-linear association	10.356	1	0.001
N of valid cases	1097		

[a] 0 cells (0.0%) have expected count less than 5. The minimum expected count is 6.32.

Table 17
Case processing summary for Title characteristics

	Cases					
	Valid		Missing		Total	
	N	Percent	*N*	Percent	*N*	Percent
Title * domain	58	100.0	0	0.0	58	100.0

the defined domains and the dependent variables are the Title element characteristics. A given web page could contain more than one Title field. The Title characteristics comprise a nominal scale, which indicates whether the content of the field is correct, partially correct, or incorrect when compared to the title information prominently displayed in the visible portion of the web page. "Correct" means that there is a perfect match between the metadata and the visible page, "partially correct" means that the metadata and the visible page differ slightly, and "incorrect" means that the metadata and the visible page differ significantly in meaning and/or syntax. For example, if the visible title was "My Web Page" and the metadata Title field contained the words "Welcome to My Web Page" or "My Web Site", the field was designated "partially correct". If the Title field for that page only contained the words "Welcome" or "John Doe's Site", the field was designated "incorrect". Because the analysis involves four comparison variables, and because the measurement scale is nominal, a chi-square test was employed to test the hypothesis. Table 17 is the case processing summary for the metadata Title characteristics.

Table 18 shows the comparison of the Title characteristics within and between domains. This comparison reveals that the number of incorrect Title fields across domains is surprisingly high (17.2%) while the number of correct Title fields only accounts for 53.4% of the Title fields found during the study. This is especially surprising since no special knowledge or training is required in order to duplicate the text of the visible title and place it in the metadata Title field.

Comparing the domains, the Gov/Org domain produced the greatest number of correct Title fields (35.5%) while authors in the IT group produced only 12.9% of the correct Titles fields found. The IT group

Table 18
Title characteristics * domain cross-tabulation

Title characteristics		Domain				Total
		LIS	Gov/Org	B&I	IT	
Correct	Count	9	11	7	4	31
	Expected count	6.4	10.2	5.9	8.6	31.0
	% within Title	29.0	35.5	22.6	12.9	100.0
	% within domain	75.0	57.9	63.6	25.0	53.4
Partially correct	Count	3	5	3	6	17
	Expected count	3.5	5.6	3.2	4.7	17.0
	% within title	17.6	29.4	17.6	35.3	100.0
	% within domain	25.0	26.3	27.3	37.5	29.3
Incorrect	Count	0	3	1	6	10
	Expected count	2.1	3.3	1.9	2.8	10.0
	% within title	0.0	30.0	10.0	60.0	100.0
	% within domain	0.0	15.8	9.1	37.5	17.2
Total	Count	12	19	11	16	58
	Expected count	12.0	19.0	11.0	16.0	58.0
	% within title	20.7	32.8	19.0	27.6	100.0
	% within domain	100.0	100.0	100.0	100.0	100.0

Table 19
Chi-square tests for Title characteristics

	Value	df	Asymp. Sig. (2-sided)
Pearson chi-square	10.590[a]	6	0.102
Likelihood ratio	12.229	6	0.057
Linear-by-linear association	8.048	1	0.005
N of valid cases	58		

[a] 7 cells (58.3%) have expected count less than 5. The minimum expected count is 1.90.

created the greatest percentages of partially correct Title fields (35.3%) while pages in the LIS and B&I groups each contained only 17.6% of the partially correct Titles fields found. LIS pages contained 0% of the incorrect Title fields found, but IT pages have the dubious distinction of creating 60% of the incorrect Title fields found during this study.

Table 19 shows that the proposed hypothesis was accepted because the significance value ($p = 0.102$) was $p > 0.05$. This means that the use of the Title field does not differ significantly across the four domains. Notice that in this case, the sample size was not large. The hypothesis acceptance was expected because unlike Keyword and Description metadata elements, the Title metadata element does not require subject analysis of the web page. Web authors simply need to copy the original title of the web page to the metadata Title field. This reduces the possibility of introducing "noise" into the process, and requires no special knowledge or skill on the part of the metadata producer. What was not expected was that the domains would have such consistent difficulty applying such a simple metadata-creation process.

4. Conclusion

The nature of the Internet as it exists today presents special challenges to researchers, publishers, end users, and search engine designers alike. It is home to vast amounts of information, and this information is highly dynamic and ever increasing. Moreover, there is no centralized control over the quality or content of either the visible text or the embedded metadata. This makes many tasks that use, manipulate, or analyze aspects of the Internet difficult to plan and carry out. For example, web publishers work hard to increase the chances that search engines will retrieve their pages in response to relevant queries. The algorithms that search engines use to retrieve and rank result lists, however, are often proprietary. Web publishers therefore cannot study these algorithms and take advantage of that knowledge to increase the visibility of relevant web pages. This leaves web authors and publishers to guess which metadata elements are most important to search engine algorithms and what content is most influential in retrieval and ranking. Similarly, search engine designers attempt to make their search engines as useful to end users as possible. They are hampered, however, by limited concrete knowledge of how reliable and accurate author-produced metadata is. They are therefore forced to make decisions based on assumptions rather than on facts.

Our research indicates that many web authors respond to this information impasse by including massive numbers of keywords in the hopes that one or more might cause the search engine to retrieve the page in response to relevant queries. These keywords are often very broad or very narrow as authors include terms for whole categories of content or list every item available through the visible page. This presents a problem for search engines. As the number of keywords increases significantly, the relevance of each keyword decreases, and search engines will have difficulty deciding which of many keywords is the most relevant. Search engines would have to process all of the keywords, process only a select few of the keywords, or ignore the Keyword field altogether. None of these options ensure efficient and effective information retrieval.

Similarly, our analysis shows that web authors tend to choose only two of the three most popular metadata elements, selecting the two that they think will adequately describe their sites to the search engine and the end user. The three most popular of these descriptive elements are the Keyword, Description, and Author elements while the least popular are the Date, Publisher, and Resource Type elements. In other words, they choose elements that they believe describe the subjective and intellectual content of the page rather than the elements that do not directly reveal subject-oriented information. Previous study has shown that search engines also pay more attention to these subject-oriented metadata fields than to other fields (Zhang & Dimitroff, 2005a, 2005b). Retrieval will only be effective if authors include accurate information in their descriptive metadata.

Indeed, deciding what information to include in these fields and what information to retrieve from these fields requires tremendous amounts of guesswork, and this guesswork is often done based on assumptions rather than evidence. It is widely hoped that metadata has the potential to improve information organization and retrieval on the Internet. Yet it has been a mystery whether the Internet publishing community is aware of metadata's significance, widely accepts it, or uses it correctly. Although a few previous studies address this issue, their impact has been limited because of small sample sizes, undistinguished user groups, and investigations that focus on a small percentage of metadata elements. This does not give readers a clear picture of metadata creation behavior on the Internet and therefore does not provide adequate evidence to confirm or deny search engine developers' and researchers' assumptions about author-generated metadata.

As seen in the study above, many people assume that there are many problems with author-generated metadata, including fraudulent or inaccurate use of metadata both in terms of syntax and semantics. For example, keyword terms can be syntactically improper if they are misspelled or semantically improper if they are far too narrow or too broad to accurately describe the site. The search engine community claims that web publishers can misuse metadata either intentionally or unintentionally, incorporating inaccurate, inappropriate, and duplicate keywords to promote their web sites. If this claim is true it may lead search engines to ignore embedded metadata altogether, negating any efforts toward metadata implementation. It is vital that researchers explore the validity of these assumptions.

In order to help provide evidence to confirm or deny these assumptions, the present authors tested five hypotheses to determine whether there are significant differences in the characteristics of metadata created by four separate user groups: library and information science (LIS), government agencies and non-profit organizations (Gov/Org), businesses and industries (B&I), and information technology (IT). This analysis would show whether there were significant inaccuracies in author-generated metadata, which user-group was most prone to inaccuracy, and which metadata elements contained the most inaccuracies. This analysis would also reveal current trends in metadata production.

This investigation found evidence that rejects four of the five proposed hypotheses. Contrary to the first four hypotheses, there were significant differences between the four defined domains in terms of Keyword accuracy, Keyword characteristics, Description accuracy, and Description characteristics. The fifth hypothesis, however, was accepted because there were no statistically significant differences with respect to metadata Title characteristics among the four domains. In other words, although web publishers produce similar Title fields, and although they generally prefer broad terms to narrow terms, publishers from different user groups display significantly different metadata creation behaviors. Web publishers for the B&I domain emphasize Keyword accuracy, while those in the LIS domain do not. The Gov/Org domain seems to value accurate Description fields more than other domains do, while publishers in the LIS group do not seem to believe accurate and specific Descriptions are as important as web publishers from other domains do. Indeed, LIS publishers produce below average metadata in almost every category, which is contrary to all expectations and assumptions. Similarly, very few web authors included incorrect or inappropriate information in their metadata, contradicting popular belief.

Also contrary to our expectations, the metadata Title element was not widely used by web authors in any of the four domains even though research has shown that this field has a significant impact on search engine

indexing and ranking algorithms (Zhang & Dimitroff, 2005a, 2005b). Web publishers might think that adding a metadata Title field is redundant because web pages already have an HTML Title field. Furthermore, among those web pages having metadata Title fields, the error rates were higher than expected. Many authors either shorten long titles or incorporate descriptive elements and/or welcoming words into the metadata version of their titles.

This research shows that it is important to take user groups into account when analyzing metadata creation. Each of the domains studied here displayed important differences in their preferences and use of metadata elements. Only 3.5% of Title fields, 3.7% of Description fields, and 3.5% of Keywords found during the course of the study were incorrect.

It is clear that researchers in semantic web technology and other fields which depend on metadata and its quality would benefit from the findings.

In order to further understand metadata creation behavior on the Internet, more research must be done. Future topics may include defining and including more domains into the investigation (such as medical groups, scientific groups, and educational groups); investigating Dublin Core creation behavior on the Internet; exploring the metadata creation behavior of web pages within the same domain but having different goals, such as informational, educational, or commercial; and examining metadata creation behavior in web pages that are not English-based.

References

Campbell, D. (2002). The use of the Dublin Core in web annotation programs. In *Proceedings of the international conference on Dublin Core and metadata for e-communities* (pp. 105–110).

Chepesuik, R. (1999). Organizing the Internet: the 'Core' of the challenge. *American Libraries* (January), 60–64.

Craven, T. (2000). The features of the description meta tags in public home pages. *Journal of Information Science, 26*(5), 303–311.

Craven, T. (2001a). Changes in metatag descriptions over time. *First Monday*. Available from http://www.firstmonday.dk/issues/issue6_10/craven/. Accessed 20 July 2004.

Craven, T. (2001b). 'DESCRIPTION' META tags in locally linked web pages. *Aslib Proceedings, 53*(6), 203–216. Available from http://dois.mimas.ac.uk/DoIS/data/Articles/julonkfhjy:2001:v:53:i:6:p:203-216.html. Accessed 10 May 2004.

Craven, T. (2001c). DESCRIPTION META tags in pages returned on different search engines. *Canadian Journal of Information and Library Science, 26*(1), 1–17.

Craven, T. (2001d). DESCRIPTION meta tags in public home and linked pages. *Libres, 11*(2). Available from http://libres.curtin.edu.au/LIBRE11N2/craven.htm. Accessed 9 May 2004.

Craven, T. (2002a). External descriptions of web pages: their features and their relationships to web page elements. *Libri, 52*(1), 36–47.

Craven, T. (2002b). What is the title of a web page? A study of the Webography practice. *Information Research, 7*(3). Available fromhttp://InformationR.net/ir/7-3/paper130.html. Accessed 7 June 2004.

Craven, T. (2003). HTML tags as extraction cues for web page description construction. *Informing Science Journal, 6*, 1–12. Available from http://inform.nu/Articles/Vol6/v6p001-012.pdf. Accessed 18 July 2004.

Doctorow, C. (2001, 26 August). Metacrap: Putting the torch to seven straw-men of the meta-utopia. *SearchEngineWatch.com*. Available from http://www.well.com/~doctorow/metacrap.htm#0. Accessed 21 July 2004.

Greenberg, J., Pattuelli, M., & Robertson, D. (2001). Author-generated Dublin Core metadata for web resources: a baseline study in an organization. *Journal of Digital Information, 2*(2). Article no. 78. Available from http://jodi.ecs.soton.ac.uk/Articles/v02/i02/Greenberg/. Accessed 11 May 2004.

Henshaw, R., & Valauskas, E. (2001). Metadata as a catalyst: experiments with metadata and search engines in the internet journal, First Monday. *Libri, 51*(2), 86–101.

Hillman, D. (2003, 23 August). Using Dublin Core. *DCMI*. Available from http://dublincore.org/documents/usageguide/. Accessed 13 May 2004.

Lagoze, C. (2001). Keeping Dublin Core simple: cross-domain discovery or resource description? *D-Lib Magazine, 7*(1). Available from http://www.dlib.org/dlib/january01/lagoze/01lagoze.html. Accessed 24 June 2004.

Richardson, T. (2003). Search engine savvy. *Canadian Business, 76*(24), 99–102, Online. Academic search elite, 30 April 2004.

Search engine optimization (2004). http://www.webmasterresources.com/search_engine_ranking/. Accessed 21 July 2004.

Search engine optimization 1-2-3 (2004). http://www.123searchengineoptimization.com/engines.html. Accessed 21 July 2004.

Sherman, C. (2002, 4 March). Metadata or metagarbage? *SearchEngineWatch.com*. Available from http://searchenginewatch.com/searchday/article.php/2159381. Accessed 19 July 2004.

Sokvitne, L. (2000). An evaluation of the effectiveness of current Dublin Core metadata for retrieval. *Presented at the VALA Conference, 2000*. Available from http://www.vala.org.au/vala2000/2000pdf/Sokvitne.PDF. Accessed 19 May 2004.

Sullivan, M. (2003). *Keyword magic: the truth about search engine optimization for your website* (pp. 1–8). Available from http://madiganpratt.com/KeywordMagic.pdf. Accessed 21 July 2004.

Tennant, R. (2004). Metadata's bitter harvest. *Library Journal, 129*(12), 32, 15 July 2004.

Turner, T., & Brackbill, L. (1998). Rising to the top: evaluating the use of the HTML META tag to improve retrieval of world wide web documents through Internet search engines. *Library Resources and Technology, 42*(4), 258–271.

Yahoo.com (2004). *How do I improve the ranking of my web site in the search results?* Available from http://help.yahoo.com/help/us/ysearch/ranking/ranking-02.html. Accessed 11 May 2004.

Zhang, J., & Dimitroff, A. (2004). Internet search engine's response to metadata Dublin Core implementation. *Journal of Information Science, 30*(4), 311–321.

Zhang, J., & Dimitroff, A. (2005a). The impact of webpage content characteristics on the webpage visibility in search engine results (Part I). *Information Processing & Management, 41*, 665–690.

Zhang, J., & Dimitroff, A. (2005b). The impact of metadata implementation on the webpage visibility in search engine results (Part II). *Information Processing & Management, 41*, 691–715.