2006

A Study of Metadata Element Co-Occurrence

Jin Zhang
University of Wisconsin, Milwaukee

Iris Jastram
University of Wisconsin, Milwaukee, ijastram@carleton.edu

A study of metadata element co-occurrence

Jin Zhang and Iris Jastram
School of Information Studies, University of Wisconsin, Milwaukee, Wisconsin, USA

Refereed article received
28 February 2006
Revision approved for
publication 15 April 2006

Abstract

Purpose – This paper aims to investigate the internet web page metadata usage behavior in terms of their metadata element co-occurrences. Metadata are designed to facilitate both web publishers/authors to organize their web pages and search engines to index the web pages accurately.

Design/methodology/approach – This study examines the types of metadata elements employed by different professional groups of web authors, the number of elements they prefer to use, and the types of element combinations they typically embed in their pages' HTML code.

Findings – The findings reveal that the "keyword" and "description" elements were the most popular single elements. The most popular combination of two elements was that of "keyword and description". Very few authors included combinations of five elements. This study also shows that preferences for element combinations varied by domains.

Originality/value – This approach will enhance the current understanding of metadata usage behavior and may help search engine designers as they continue their quest for improved indexing and retrieval of web pages.

Keywords Behaviour, Internet, Information organizations

Paper type Research paper

1. Introduction

The proliferation of information on the internet has made information retrieval from that resource a challenging discussion topic for researchers, search engine and subject index developers, and Internet users alike. Metadata could help this situation if it were used consistently and well. There is no centralized control over the form or content of embedded metadata however, which causes many to fear that it is too easily misused or abused. Given the vast potential metadata possesses to enhance internet information organization and retrieval, and given the equally vast potential for internet resource creators to misrepresent their pages through metadata (either accidentally or maliciously), researchers have focused their efforts either on the theoretical side or the practical side of metadata implementation: how metadata can or should be used and how metadata is being used.

The most commonly used metadata scheme on the Internet is the HTML "meta" tag. The researchers found that of the 2,400 pages visited, 62.83 percent included this type of metadata embedded in the HTML code. This is a much greater percentage than the 7.42 percent of pages containing Dublin Core and the 44.12 percent containing any other scheme of metadata. This scheme has no standardized element set, leaving the choice of the type and quantity of elements entirely up to the resource author. This type of metadata can be found in the source code of a web page in the format < META name ="[tag name, such as keywords]" content ="[metadata content, such as a list of keywords]"/>. Because the author has complete control over the type and quantity of

Online Information Review
Vol. 30 No. 4, 2006
pp. 428-453

1468-4527
DOI 10.1108/14684520610686319

metadata elements, this scheme allows metadata to be as simple or as complex as the author wishes.

It is this flexibility that causes much of the discussion among researchers and search engine developers. How much granularity is beneficial and how much dilutes the effectiveness of the scheme or renders the scheme too complex for the average web author? As Campbell points out, metadata is pulled in two directions: that of traditional information organization and bibliographic description on one side, and that of "the emerging standards that will form the web of the future" on the other (Campbell, 2002). On the one hand, metadata stems from a long history of describing resources using standardized formats and vocabularies. On the other hand, metadata of the future is as yet undetermined; this type of metadata is still evolving quietly on the web.

Those researchers who believe metadata should be governed by stricter standards argue that the lack of controlled vocabulary fundamentally dilutes metadata's effectiveness. Chepesuik, for example, argues that metadata is really "cataloging by another name" (Chepesuik, 1999). As such, he maintains, controlled vocabulary is necessary to fend off "bibliographic chaos" (Chepesuik, 1999). He quotes Michael Gorman as saying:

> There is no third way between cataloging, controlled vocabularies, etc. (expensive and effective) and the chaos of keyword searching on the web (inexpensive and utterly ineffective) (Chepesuik, 1999).

Other researchers also note that without some standardization and centralized control, metadata will have little value and therefore will not be used by search engines (see Sokvitne, 2000; Henshaw and Valauskas, 2001; Tennant, 2003, 2004). The stakes are therefore quite high. Lack of bibliographic control could lead to such inconsistent metadata that search engines completely disregard it, which would make the use of metadata by authors publishing pages to the open internet an exercise in futility.

Proponents of a looser metadata scheme, however, argue that if metadata is too difficult for the average web author to create, those authors will not use the scheme or will misuse the scheme, each of which could result in the ultimate demise of metadata as a tool of internet resource discovery. Carl Lagoze (2001), for example, argues that even though there is a place for greater granularity in metadata, there is also a strong argument for "pidgin" metadata on the internet. This type of metadata would be simple enough that multiple and diverse search algorithms could access its contents, and in this way the pidgin scheme would allow for basic cross-domain resource discovery (Lagoze, 2001). Diane Hillman (2003) agrees, saying that there are such differences in vocabulary preferences between spheres of knowledge that pidgin metadata schemes provide better cross-domain retrieval possibilities.

Those who give advice and do research on how to increase web page visibility seem to agree with Lagoze and Hillman. Most advocate using only "keyword", "description", or a combination of those two elements (see Richardson, 2003; Search Engine Optimization, n.d.; Search Engine Optimization 1-2-3, n.d.; Sullivan, 2003; Yahoo.com, n.d.). Other research indicates that the "keyword", "description", and "title" elements influence retrieval and ranking more than other elements do (Zhang and Dimitroff, 2005). This type of research tends to support the proposal to keep metadata simple. Creating metadata is expensive, requiring time and thought. Every element added costs money, so in an age of tightening profit margins metadata's strength is often seen

in its potential to enable resource discovery and visibility rather than resource description, a potential strength Turner and Brackbill (1998) recognized in their study on the effect of metadata on web page visibility.

Several studies have been conducted to determine the characteristics of certain metadata elements currently in use on the internet. Timothy Craven has conducted numerous studies to determine what types of information web authors include in their "description" and "title" elements (Craven, 2000, 2001a, b, c, d, 2002a, b, 2003, 2005). Lloyd Sokvitne (2000) has studied the quality of Dublin Core metadata elements in the web pages of large Australian organizations. However, no studies have been done to determine the types of elements and quantity of elements preferred by web authors, and no studies have been done to determine how a web author's chosen discipline affects metadata element use. These are crucial pieces of evidence for those seeking to determine how metadata is used and what the potential for future use might be. Examining this evidence may help to determine whether metadata will function descriptively, as a tool for resource discovery across domains and disciplines, or some combination of the two.

Understanding the type and quantity of elements used by web authors from different disciplines will not only shed light on current web publishing behavior and trends, but it will also provide concrete information for search engine designers. It will reveal the most popular metadata elements and element combinations, and it will therefore enable search engine designers to make informed decisions about the inclusion and relative importance of certain metadata elements in their search algorithms.

This study examines the metadata from 2,400 web pages, 600 from each of four predefined domains. These domains are Library and Information Science (LIS), Government Agencies and Non-profit Organizations (Gov/Org), Businesses and Industries (B&I), and Information Technology (IT). The researchers formulated queries to retrieve pages from each domain from search engines and selected subjects related to each domain from subject directories. Pages for analysis were then selected as randomly as possible from the search engine and subject directory result lists. Half of the pages selected for each domain came from search engines and half from subject directories.

The researchers then performed a pilot study to identify and define possible metadata elements for analysis. These elements were:

- *Title*. The name of the page as displayed prominently on the page or (if the title does not appear on the page) as it is displayed on the browser application's title bar.
- *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 primary language of the text 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", "content", "area", "destination", and others that were added to some sites but were not commonly chosen for inclusion in the metadata of most sites), and fields that are administrative in nature (such as "approved-by", "site-product-code", "terminator", "department", "expires", "template-id", or "revisit-after").

During data collection, the above elements were collected from each selected web page and then recorded for later analysis.

2. Results and analysis

2.1 General element co-occurrence analysis

Because there are no mandatory restrictions or requirements associated with metadata use, it is entirely up to the web page authors and publishers to determine the type and number of elements to include. Web authors choose which elements to include based on their own preference and need. They can choose to include no metadata at all or anywhere from one to twelve of the elements described above.

The findings show that most web authors using metadata in the code of their pages include two elements. In other words, when the number of combined metadata elements increases (>2), the corresponding count decreases across all domains. When the number of combined metadata elements decreases (<2), the corresponding count decreases across all domains. This phenomenon is consistent both across all domains and within each individual domain. Also the rates for combinations of few elements were relatively high, and the rates for combinations of many elements were relatively low. Notice that in the LIS domain, no web pages contain nine or ten elements. None of the web pages analyzed in any domain included 11 or 12 elements in their metadata sets.

Figure 1 provides a visual representation of the number of times each domain exhibits specific co-occurrence values. The number of combined elements is along the X-axis, and the frequency of the co-occurrence is along the Y-axis. Here the peak of pages including two metadata elements is quite apparent.

2.2 Individual element distribution analysis

Since metadata is not required for web page publication, not all investigated web pages have embedded metadata, and not all web pages having metadata include all of the defined metadata elements. The metadata occurrence analysis provides an overview of the preferences each of the four professional domains exhibits for metadata inclusion and element selection. This analysis does not indicate the co-occurrence of elements.

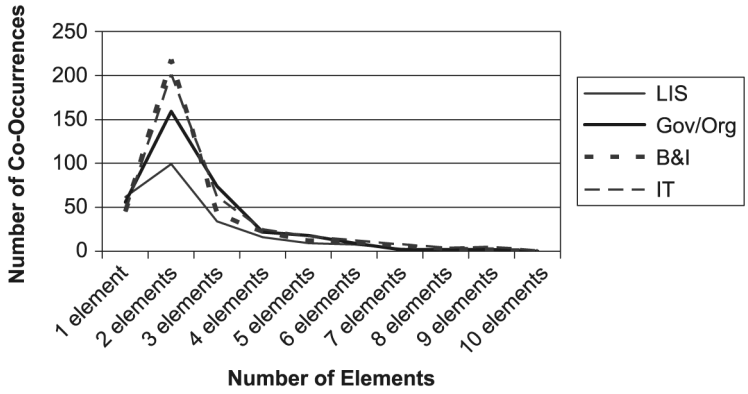

Figure 1.
Chart for co-occurrence analysis

Table I displays the distribution of each individual metadata element between and within the four domains. This analysis does not determine the number of elements in a given page, so while 1,318 web pages included metadata, this analysis examines each of the 3,366 individual elements found. In this table, "count" refers to the raw number of occurrences of a given metadata element in a given domain. For instance, LIS pages include the "author" element 76 times. The "percent of total" comparison is defined as the ratio of a certain metadata element within a domain to the total of all elements in all domains. For example, the "percent of total" for "author" element in the LIS domain is 2.3 percent ($76/3366 = 2.3$ percent, where 76 is the number of author fields found in the given domain and 3,366 is the total number of metadata fields found in all domains). Author fields therefore account for 2.3 percent of all the fields examined.

Figure 2 illustrates the results of element distribution. The X-axis displays the metadata elements, and the Y-axis shows the raw number of occurrences of each element.

Among the four domains, "subject/keyword" (34.2 percent), "description" (31.8 percent), "miscellaneous" (10.7 percent) and "author" (8.9 percent) were the most frequent metadata elements, while "title" (1.7 percent), "language" (1.8 percent), "resource type" (1.2 percent), "publisher" (1.1 percent) and "date" (0.8 percent) were the least frequently used. This suggests that web page publishers often pay more attention to subject-oriented metadata fields such as "subject/keyword" and "description". Search engine crawlers use these subject-oriented fields to index web pages by extracting keywords from these fields (Zhang and Dimitroff, 2005). However, according to Zhang and Dimitroff (2005), these crawlers also extract keywords from the "title" field, and this field is among the least frequently used in the web pages visited.

2.3 Two element co-occurrence analysis

Since there are 12 elements identified and defined in this study, there are 66 possible combinations of two elements when at least two elements are present in a given web page ($C_{12}^2 = 66$). Table II displays the detailed analysis of each of these combinations in the web pages analyzed for this study. In this table, "count" refers to the raw number of times each possible combination occurred in the pages analyzed. For example, the combination of "keyword" and "date" occurred six times in web pages having two or

Elements		Domain: Library and information science	Government agencies and non-profit organizations	Businesses and industries	Information technology	Total
Author	Count	76	86	61	78	301
	Percent of total	2.3	2.6	1.8	2.3	8.9
Publisher	Count	3	9	10	14	36
	Percent of total	0.1	0.3	0.3	0.4	1.1
Miscellaneous	Count	59	122	78	100	359
	Percent of total	1.8	3.6	2.3	3.0	10.7
Copyright	Count	21	13	30	44	108
	Percent of total	0.6	0.4	0.9	1.3	3.2
Rating	Count	11	19	20	37	87
	Percent of total	0.3	0.6	0.6	1.1	2.6
Resource type	Count	7	10	11	12	40
	Percent of total	0.2	0.3	0.3	0.4	1.2
Language	Count	7	17	13	22	59
	Percent of total	0.2	0.5	0.4	0.7	1.8
Distribution	Count	7	15	19	27	68
	Percent of total	0.2	0.4	0.6	0.8	2.0
Date	Count	10	0	12	5	27
	Percent of total	0.3	0.0	0.4	0.1	0.8
Keyword	Count	175	294	332	351	1152
	Percent of total	5.2	8.7	9.9	10.4	34.2
Description	Count	160	272	312	328	1072
	Percent of total	4.8	8.1	9.3	9.7	31.8
Title	Count	12	19	11	15	57
	Percent of total	0.4	0.6	0.3	0.4	1.7
Total	Count	548	876	909	1033	3366
	Percent of total	16.3	26.0	27.0	30.7	100.0

Table I.
Distributions of metadata elements

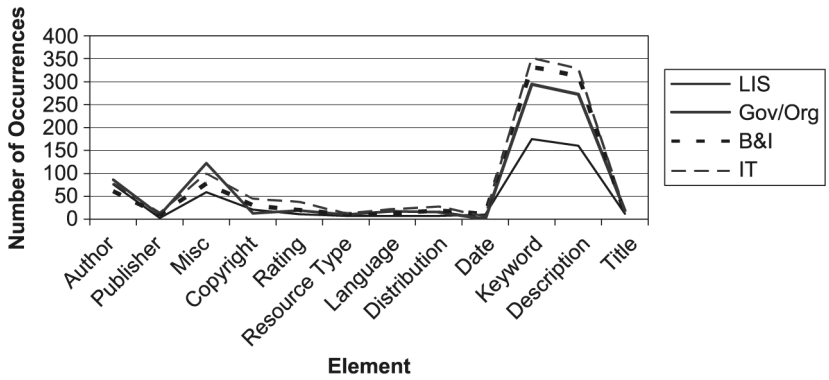

Figure 2.
Visual display of element distribution

more metadata elements. "Percent within combination" is the ratio of the number of times the combination of elements occurs in a given domain to the number of times that combination occurs across all four domains. By the same token, the combination of "keyword and date" in LIS' pages accounts for 33.3 percent of all the times "keyword and date" appeared together ($6/18 = 33.3$ percent). "Percent within domain" is the ratio of the number of times the combination of elements occurs in a given domain to the number of times there are at least two elements present in the web pages of that domain. Finally, the combination of "keyword and date" occurred in 1.0 percent of the LIS pages that had at least two embedded elements ($6/605 = 1.0$ percent).

This chart shows that among the four domains, the five most frequent combinations of elements were "keyword and description" (24.6 percent), "keyword and miscellaneous" (7.0 percent), "description and miscellaneous" (6.5 percent), "keyword and author" (5.9 percent), and "author and description" (5.4 percent). It is significant that the combination of "keyword and description" appears most frequently across the four domains because these are highly subject-oriented elements from which search engines can glean keywords. In fact, the combination of "keyword and description" is not only the most common combination across the four domains, but it is also the most common combination in each of the four individual domains. There are telling differences, however, between the combinations that rank just below this one in the individual domains.

The ten most frequent combinations of two elements in the LIS group were "keyword and description" (23.1 percent), "keyword and author" (7.8 percent), "description and author" (7.6 percent), "keyword and miscellaneous" (6.3 percent), "description and miscellaneous" (5.8 percent), "author and miscellaneous" (4.0 percent), "keyword and copyright" (2.8 percent), "author and copyright" (2.6 percent), keyword and distribution' (2.3 percent), and "description and copyright" (2.1 percent).

The ten most frequent combinations of two elements for the Gov/Org group were "keyword and description" (25.2 percent), "keyword and miscellaneous" (9.8 percent), "description and miscellaneous" (8.7 percent), "keyword and author" (7.0 percent), "description and author" (6.0 percent), "author and miscellaneous" (4.2 percent), "keyword and distribution" (2.8 percent), "keyword and title" (1.8 percent), "keyword and language" (1.7 percent), and "keyword and rating" (1.6 percent).

The ten most frequent combinations of two elements for the B&I group were "keyword and description" (28.1 percent), "keyword and miscellaneous" (6.1 percent),

Combinations		Domain				Total
		Library and information science	Government agencies and non-profit organizations	Businesses and industries	Information technology	
Keyword and date	Count	6	0	8	4	18
	Percent within combination	33.3	0.0	44.4	22.2	100.0
	Percent within domain	1.0	0.0	0.7	0.3	0.4
Keyword and distribution	Count	14	28	38	52	132
	Percent within combination	10.6	21.2	28.8	39.4	100.0
	Percent within domain	2.3	2.8	3.5	3.7	3.2
Keyword and language	Count	7	17	13	20	57
	Percent within combination	12.3	29.8	22.8	35.1	100.0
	Percent within domain	1.2	1.7	1.2	1.4	1.4
Keyword and resource type	Count	5	8	9	11	33
	Percent within combination	15.2	24.2	27.3	33.3	100.0
	Percent within domain	0.8	0.8	0.8	0.8	0.8
Keyword and rating	Count	10	16	19	37	82
	Percent within combination	12.2	19.5	23.2	45.1	100.0
	Percent within domain	1.7	1.6	1.8	2.6	2.0
Keyword and copyright	Count	17	11	30	41	99
	Percent within combination	17.2	11.1	30.3	41.4	100.0
	Percent within domain	2.8	1.1	2.8	2.9	2.4
Keyword and miscellaneous	Count	38	99	65	85	287
	Percent within combination	13.2	34.5	22.6	29.6	100.0
	Percent within domain	6.3	9.8	6.1	6.0	7.0
Keyword and publisher	Count	3	9	10	14	36
	Percent within combination	8.3	25.0	27.8	38.9	100.0
	Percent within domain	0.5	0.9	0.9	1.0	0.9
Keyword and author	Count	47	71	55	71	244
	Percent within combination	19.3	29.1	22.5	29.1	100.0
	Percent within domain	7.8	7.0	5.1	5.0	5.9
Keyword and title	Count	9	18	11	14	52
	Percent within combination	17.3	34.6	21.2	26.9	100.0
	Percent within domain	1.5	1.8	1.0	1.0	1.3

(*continued*)

Table II.
Distribution for combinations of two metadata elements

Table II.

Combinations		Domain: Library and information science	Government agencies and non-profit organizations	Businesses and industries	Information technology	Total
Keyword and description	Count	140	255	301	314	1,010
	Percent within combination	13.9	25.2	29.8	31.1	100.0
	Percent within domain	23.1	25.2	28.1	22.1	24.6
Date and distribution	Count	1	0	0	1	2
	Percent within combination	50.0	0.0	0.0	50.0	100.0
	Percent within domain	0.2	0.0	0.0	0.1	0.0
Date and language	Count	0	0	3	2	5
	Percent within combination	0.0	0.0	60.0	40.0	100.0
	Percent within domain	0.0	0.0	0.3	0.1	0.1
Date and resource type	Count	1	0	0	2	3
	Percent within combination	33.3	0.0	0.0	66.7	100.0
	Percent within domain	0.2	0.0	0.0	0.1	0.1
Date and rating	Count	1	0	0	1	2
	Percent within combination	50.0	0.0	0.0	50.0	100.0
	Percent within domain	0.2	0.0	0.0	0.1	0.0
Date and copyright	Count	2	0	1	1	4
	Percent within combination	50.0	0.0	25.0	25.0	100.0
	Percent within domain	0.3	0.0	0.1	0.1	0.1
Date and miscellaneous	Count	6	0	8	4	18
	Percent within combination	33.3	0.0	44.4	22.2	100.0
	Percent within domain	1.0	0.0	0.7	0.3	0.4
Date and publisher	Count	1	0	2	0	3
	Percent within combination	33.3	0.0	66.7	0.0	100.0
	Percent within domain	0.2	0.0	0.2	0.0	0.1
Date and author	Count	8	0	4	3	15
	Percent within combination	53.3	0.0	26.7	20.0	100.0
	Percent within domain	1.3	0.0	0.4	0.2	0.4
Date and title	Count	3	0	3	0	6
	Percent within combination	50.0	0.0	50.0	0.0	100.0
	Percent within domain	0.5	0.0	0.3	0.0	0.1

(*continued*)

Table II.

Combinations		Domain: Library and information science	Government agencies and non-profit organizations	Businesses and industries	Information technology	Total
Date and description	Count	7	0	7	4	18
	Percent within combination	38.9	0.0	38.9	22.2	100.0
	Percent within domain	1.2	0.0	0.7	0.3	0.4
Distribution and language	Count	3	4	7	9	23
	Percent within combination	13.0	17.4	30.4	39.1	100.0
	Percent within domain	0.5	0.4	0.7	0.6	0.6
Distribution and resource type	Count	2	7	6	8	23
	Percent within combination	8.7	30.4	26.1	34.8	100.0
	Percent within domain	0.3	0.7	0.6	0.6	0.6
Distribution and rating	Count	4	9	10	18	41
	Percent within combination	9.80	22.0	24.4	43.9	100.0
	Percent within domain	0.7	0.9	0.9	1.3	1.0
Distribution and copyright	Count	3	7	11	15	36
	Percent within combination	8.3	19.4	30.6	41.7	100.0
	Percent within domain	0.5	0.7	1.0	1.1	0.9
Distribution and miscellaneous	Count	7	11	17	23	58
	Percent within combination	12.1	19.0	29.3	39.7	100.0
	Percent within domain	1.2	1.1	1.6	1.6	1.4
Distribution and publisher	Count	0	2	1	8	11
	Percent within combination	0.0	18.2	9.1	72.7	100.0
	Percent within domain	0.0	0.2	0.1	0.6	0.3
Distribution and author	Count	4	6	9	14	33
	Percent within combination	12.1	18.2	27.3	42.4	100.0
	Percent within domain	0.7	0.6	0.8	1.0	0.8
Distribution and title	Count	3	2	0	2	7
	Percent within combination	42.9	28.6	0.0	28.60	100.0
	Percent within domain	0.5	0.2	0.0	0.1	0.2
Distribution and description	Count	5	13	19	26	63
	Percent within combination	7.9	20.6	30.2	41.3	100.0
	Percent within domain	0.8	1.3	1.8	1.8	1.5

(*continued*)

Table II.

Combinations		Domain: Library and information science	Government agencies and non-profit organizations	Businesses and industries	Information technology	Total
Language and resource type	Count	2	2	3	5	12
	Percent within combination	16.7	16.7	25.0	41.7	100.0
	Percent within domain	0.3	0.2	0.3	0.4	0.3
Language and rating	Count	3	5	7	8	23
	Percent within combination	13.0	21.7	30.4	34.8	100.0
	Percent within domain	0.5	0.5	0.7	0.6	0.6
Language and copyright	Count	4	5	8	13	30
	Percent within combination	13.3	16.7	26.7	43.3	100.0
	Percent within domain	0.7	0.5	0.7	0.9	0.7
Language and miscellaneous	Count	5	15	12	20	52
	Percent within combination	9.6	28.8	23.1	38.5	100.0
	Percent within domain	0.8	1.5	1.1	1.4	1.3
Language and publisher	Count	0	4	4	6	14
	Percent within combination	0.0	28.6	28.6	42.9	100.0
	Percent within domain	0.0	0.4	0.4	0.4	0.3
Language and author	Count	5	12	4	13	34
	Percent within combination	14.7	35.3	11.8	38.2	100.0
	Percent within domain	0.8	1.2	0.4	0.9	0.8
Language and title	Count	1	5	3	5	14
	Percent within combination	7.1	35.7	21.4	35.7	100.0
	Percent within domain	0.2	0.5	0.3	0.4	0.3
Language and description	Count	5	15	13	20	53
	Percent within combination	9.4	28.3	24.5	37.7	100.0
	Percent within domain	0.8	1.5	1.2	1.4	1.3
Resource type and rating	Count	2	5	3	7	17
	Percent within combination	11.8	29.4	17.6	41.2	100.0
	Percent within domain	0.3	0.5	0.3	0.5	0.4
Resource type and copyright	Count	1	4	7	7	19
	Percent within combination	5.3	21.1	36.8	36.8	100.0
	Percent within domain	0.2	0.4	0.7	0.5	0.5

(*continued*)

Table II.

Combinations		Domain: Library and information science	Government agencies and non-profit organizations	Businesses and industries	Information technology	Total
Resource type and miscellaneous	Count	4	6	7	8	25
	Percent within combination	16.0	24.0	28.0	32.0	100.0
	Percent within domain	0.7	0.6	0.7	0.6	0.6
Resource type and publisher	Count	0	0	1	1	2
	Percent within combination	0.0	0.0	50.0	50.0	100.0
	Percent within domain	0.0	0.0	0.1	0.1	0.0
Resource type and author	Count	6	4	5	6	21
	Percent within combination	28.6	19.0	23.8	28.6	100.0
	Percent within domain	1.0	0.4	0.5	0.4	0.5
Resource type and title	Count	1	0	0	2	3
	Percent within combination	33.3	0.0	0.0	66.7	100.0
	Percent within domain	0.2	0.0	0.0	0.1	0.1
Resource type and description	Count	5	7	9	10	31
	Percent within combination	16.1	22.6	29.0	32.3	100.0
	Percent within domain	0.8	0.7	0.8	0.7	0.8
Rating and copyright	Count	6	5	8	16	35
	Percent within combination	17.1	14.3	22.9	45.7	100.0
	Percent within domain	1.0	0.5	0.7	1.1	0.9
Rating and miscellaneous	Count	9	14	18	29	70
	Percent within combination	12.9	20.0	25.7	41.4	100.0
	Percent within domain	1.5	1.4	1.7	2.0	1.7
Rating and publisher	Count	0	2	1	8	11
	Percent within combination	0.0	18.2	9.1	72.7	100.0
	Percent within domain	0.0	0.2	0.1	0.6	0.3
Rating and author	Count	8	8	10	18	44
	Percent within combination	18.2	18.2	22.7	40.9	100.0
	Percent within domain	1.3	0.8	0.9	1.3	1.1
Rating and title	Count	2	2	2	3	9
	Percent within combination	22.2	22.2	22.2	33.3	100.0
	Percent within domain	0.3	0.2	0.2	0.2	0.2

(*continued*)

Table II.

Combinations		Domain: Library and information science	Government agencies and non-profit organizations	Businesses and industries	Information technology	Total
Rating and description	Count	8	15	19	37	79
	Percent within combination	10.1	19.0	24.1	46.8	100.0
	Percent within domain	1.3	1.5	1.8	2.6	1.9
Copyright and miscellaneous	Count	8	11	19	31	69
	Percent within combination	11.6	15.9	27.5	44.9	100.0
	Percent within domain	1.3	1.1	1.8	2.2	1.7
Copyright and publisher	Count	0	2	8	9	19
	Percent within combination	0.0	10.5	42.1	47.4	100.0
	Percent within domain	0.0	0.2	0.7	0.6	0.5
Copyright and author	Count	16	7	21	28	72
	Percent within combination	22.2	9.7	29.2	38.9	100.0
	Percent within domain	2.6	0.7	2.0	2.0	1.8
Copyright and title	Count	2	2	3	6	13
	Percent within combination	15.4	15.4	23.1	46.2	100.0
	Percent within domain	0.3	0.2	0.3	0.4	0.3
Copyright and description	Count	13	9	29	40	91
	Percent within combination	14.3	9.9	31.9	44.0	100.0
	Percent within domain	2.1	0.9	2.7	2.8	2.2
Miscellaneous and publisher	Count	2	9	7	14	32
	Percent within combination	6.3	28.1	21.9	43.8	100.0
	Percent within domain	0.3	0.9	0.7	1.0	0.8
Miscellaneous and author	Count	24	42	24	41	131
	Percent within combination	18.3	32.1	18.3	31.3	100.0
	Percent within domain	4.0	4.2	2.2	2.9	3.2
Miscellaneous and title	Count	9	14	6	10	39
	Percent within combination	23.1	35.9	15.4	25.6	100.0
	Percent within domain	1.5	1.4	0.6	0.7	0.9
Miscellaneous and description	Count	35	88	61	82	266
	Percent within combination	13.2	33.1	22.9	30.8	100.0
	Percent within domain	5.8	8.7	5.7	5.8	6.5

(*continued*)

Combinations		Domain				Total
		Library and information science	Government agencies and non-profit organizations	Businesses and industries	Information technology	
Publisher and author	Count	1	6	6	12	25
	Percent within combination	4.0	24.0	24.0	48.0	100.0
	Percent within domain	0.2	0.6	0.6	0.8	0.6
Publisher and title	Count	1	2	3	0	6
	Percent within combination	16.7	33.3	50.0	0.0	100.0
	Percent within domain	0.2	0.2	0.3	0.0	0.1
Publisher and description	Count	3	8	9	14	34
	Percent within combination	8.8	23.5	26.5	41.2	100.0
	Percent within domain	0.5	0.8	0.8	1.0	0.8
Author and title	Count	4	8	3	5	20
	Percent within combination	20.0	40.0	15.0	25.0	100.0
	Percent within domain	0.7	0.8	0.3	0.4	0.5
Author and description	Count	46	61	50	66	223
	Percent within combination	20.6	27.4	22.4	29.6	100.0
	Percent within domain	7.6	6.0	4.7	4.7	5.4
Title and description	Count	7	15	11	14	47
	Percent within combination	14.9	31.9	23.4	29.8	100.0
	Percent within domain	1.2	1.5	1.0	1.0	1.1
Total	Count	605	1,012	1,071	1,418	4,106
	Percent within combination	14.7	24.6	26.1	34.5	100.0
	Percent within domain	100.0	100.0	100.0	100.0	100.0

Table II.

"description and miscellaneous" (5.7 percent), "keyword and author" (5.1 percent), "description and author" (4.7 percent), "keyword and distribution" (3.5 percent), "keyword and copyright" (2.8 percent), "description and copyright" (2.7 percent), "author and miscellaneous" (2.2 percent), and "author and copyright" (2.0 percent).

The ten most frequent combinations of two elements in the IT group were "keyword and description" (22.1 percent), "keyword and miscellaneous" (6.0 percent), "description and miscellaneous" (5.8 percent), "keyword and author" (5.0 percent), "description and author" (4.7 percent), "keyword and distribution" (3.7 percent), "keyword and copyright" (2.9 percent), "author and miscellaneous" (2.9 percent), "description and copyright" (2.8 percent), and "keyword and rating" (2.6 percent).

If the "miscellaneous" category is not considered (because it does not have a specific meaningful definition), the combinations of two elements that appear in the top ten combinations of all four domains were "keyword and description", "keyword and author", "description and author", and "keyword and distribution". The combinations that appeared in the top ten of three individual domains were "keyword and copyright", and "copyright and description". The combinations that appeared in the top ten of two individual domains were "copyright and author", and "keyword and rating". The combinations that appeared in the top ten of only one domain were "keyword and title", "keyword and language". Both of these last combinations appeared in the Gov/Org domain.

Figure 3 provides a visual representation of the frequencies of each possible combination of two metadata elements.

2.4 Three element co-occurrence analysis in a selected element set

Theoretically, it is possible to analyze combinations of three, four, five, or more elements among the 12 elements identified for this study. As the number of elements in each combination increases, however, the number of possible combinations increases dramatically. For instance, there are 924 possible combinations of six elements ($C_{12}^{6} = [12 \times 11 \times 10 \times 9 \times 8 \times 7]/[6 \times 5 \times 4 \times 3 \times 2 \times 1] = 924$). This figure is too large for the scope of this paper. The researchers therefore selected the five most frequently used elements (other than the miscellaneous field), as identified in Figure 1, and the "title" field. Even though the title field is the seventh most frequently used element, it is included in place of the sixth (language) because of its significance as a major determinant of web page retrieval and ranking by search engines (Zhang and Dimitroff, 2005). These elements are "subject/keyword", "description", "author", "copyright", "rating", and "title".

These six elements can create 20 separate combinations of three elements each ($C_{6}^{3} = (6 \times 5 \times 4)/(3 \times 2 \times 1) = 20$). Table III shows the detailed analysis of each of these combinations in web pages having at least three elements.

As this table shows, the five most common element combinations were "author, description, and keyword" (25.9 percent); "copyright, description, and keyword" (10.8 percent); "description, keyword, and rating" (9.6 percent); "author, copyright, and keyword" (8.3 percent); and "author, copyright, and description" (7.8 percent). Again, those combinations that include "keyword", "description", or "author" are more common than those that do not. It is also interesting to note that pages from the IT domain include combinations of at least three elements more often than do pages from other domains.

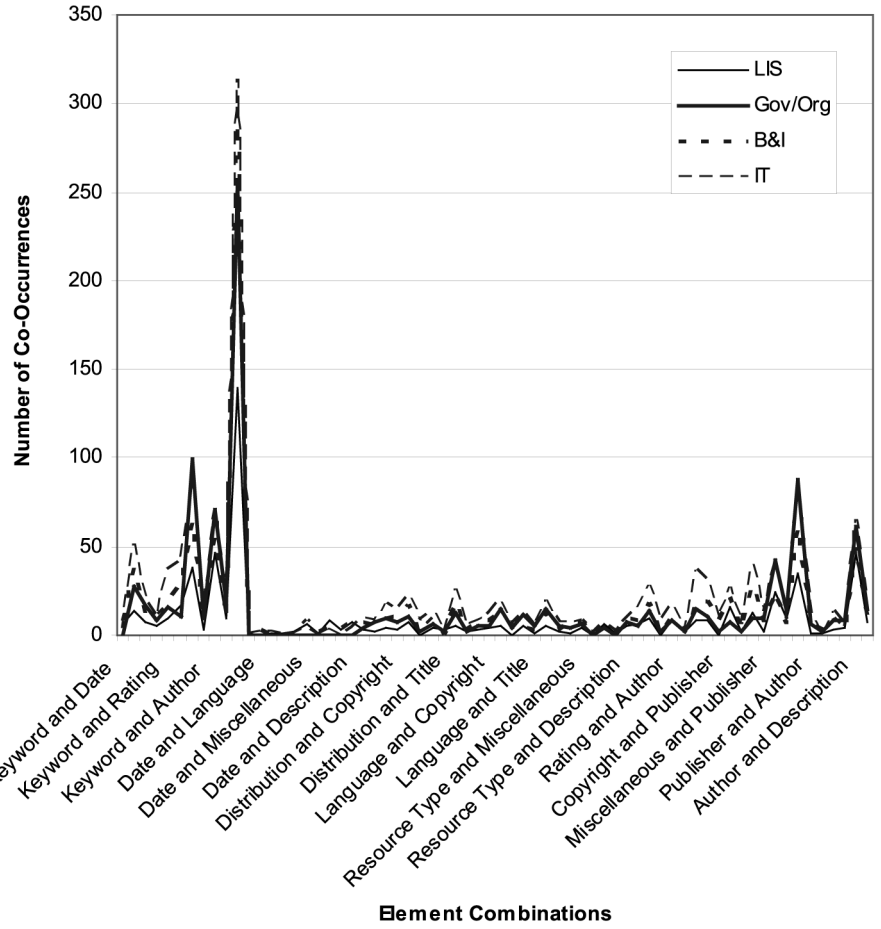

Figure 3.
Distribution of combinations of two elements

Figure 4 provides a visual representation of the frequency of each combination of three metadata elements.

2.5 Four element co-occurrence analysis in a selected element set

For the same reasons discussed in the previous section, the analysis of combinations of four metadata elements was conducted with the same selection of six metadata elements. In this case, there are 15 possible combinations of four elements ($C_6^4 = [6 \times 5 \times 4 \times 3]/[4 \times 3 \times 2 \times 1] = 15$). Table IV shows the results of this analysis, including the raw number of times each combination appears in pages having at least four embedded elements; how that count compares with pages from other domains having the same combinations; and how that count compares with other four-element combinations in the same domain. The top five combinations in this category were "author, copyright, description, and keyword" (25.8 percent); "author, description, keyword, and rating" (15.6 percent); "copyright, description, keyword, and rating" (13.1 percent); "author, copyright, keyword, and rating" (11.5 percent); and

Table III.
Three element combinations in a selected element set

Combination		Domain: Library and information science	Government agencies and non-profit organizations	Businesses and industry	Information technology	Total
Author, copyright, and description	Count	11	6	20	27	64
	Percent within combination	17.2	9.4	31.3	42.2	100.0
	Percent within domain	8.0	3.8	9.6	8.5	7.8
Author, copyright, and keyword	Count	14	7	21	26	68
	Percent within combination	20.6	10.3	30.9	38.2	100.0
	Percent within domain	10.1	4.4	10.1	8.2	8.3
Author, copyright, and rating	Count	6	3	6	13	28
	Percent within combination	21.4	10.7	21.4	46.4	100.0
	Percent within domain	4.3	1.9	2.9	4.1	3.4
Author, copyright, and title	Count	1	1	1	4	7
	Percent within combination	14.3	14.3	14.3	57.1	100.0
	Percent within domain	0.7	0.6	0.5	1.3	0.9
Author, description, and keyword	Count	39	60	49	65	213
	Percent within combination	18.3	28.2	23.0	30.5	100.0
	Percent within domain	28.3	37.7	23.6	20.4	25.9
Author, description, and rating	Count	5	5	10	18	38
	Percent within combination	13.2	13.2	26.3	47.4	100.0
	Percent within domain	3.6	3.1	4.8	5.7	4.6
Author, description, and title	Count	3	8	3	5	19
	Percent within combination	15.8	42.1	15.8	26.3	100.0
	Percent within domain	2.2	5.0	1.4	1.6	2.3
Author, keyword, and rating	Count	7	6	10	18	41
	Percent within combination	17.1	14.6	24.4	43.9	100.0
	Percent within domain	5.1	3.8	4.8	5.7	5.0
Author, keyword, and title	Count	4	8	3	4	19
	Percent within combination	21.1	42.1	15.8	21.1	100.0
	Percent within domain	2.9	5.0	1.4	1.3	2.3
Author, rating, and title	Count	4	0	0	0	4
	Percent within combination	100.0	0.0	0.0	0.0	100.0
	Percent within domain	2.9	0.0	0.0	0.0	0.5
Copyright, description, and keyword	Count	13	9	29	38	89
	Percent within combination	14.6	10.1	32.6	42.7	100.0
	Percent within domain	9.4	5.7	13.9	11.9	10.8

(*continued*)

Combination		Domain: Library and information science	Domain: Government agencies and non-profit organizations	Domain: Businesses and industry	Domain: Information technology	Total
Copyright, description, and rating	Count	4	4	8	16	32
	Percent within combination	12.5	12.5	25.0	50.0	100.0
	Percent within domain	2.9	2.5	3.8	5.0	3.9
Copyright, description, and title	Count	1	1	3	5	10
	Percent within combination	10.0	10.0	30.0	50.0	100.0
	Percent within domain	0.7	0.6	1.4	1.6	1.2
Copyright, keyword, and rating	Count	6	5	8	16	35
	Percent within combination	17.1	14.3	22.9	45.7	100.0
	Percent within domain	4.3	3.1	3.8	5.0	4.3
Copyright, keyword, and title	Count	2	2	3	5	12
	Percent within combination	16.7	16.7	25.0	41.7	100.0
	Percent within domain	1.4	1.3	1.4	1.6	1.5
Copyright, rating, and title	Count	0	0	0	2	2
	Percent within combination	0.0	0.0	0.0	100.0	100.0
	Percent within domain	0.0	0.0	0.0	0.6	0.2
Description, keyword, and rating	Count	8	15	19	37	79
	Percent within combination	10.1	19.0	24.1	46.8	100.0
	Percent within domain	5.8	9.4	9.1	11.6	9.6
Description, keyword, and title	Count	6	15	11	13	45
	Percent within combination	13.3	33.3	24.4	28.9	100.0
	Percent within domain	4.3	9.4	5.3	4.1	5.5
Description, rating, and title	Count	2	2	2	3	9
	Percent within combination	22.2	22.2	22.2	33.3	100.0
	Percent within domain	1.4	1.3	1.0	0.9	1.1
Keyword, rating, and title	Count	2	2	2	3	9
	Percent within combination	22.2	22.2	22.2	33.3	100.0
	Percent within domain	1.4	1.3	1.0	0.9	1.1
Total	Count	138	159	208	318	823
	Percent within combination	16.8	19.3	25.3	38.6	100.0
	Percent within domain	100.0	100.0	100.0	100.0	100.0

Table III.

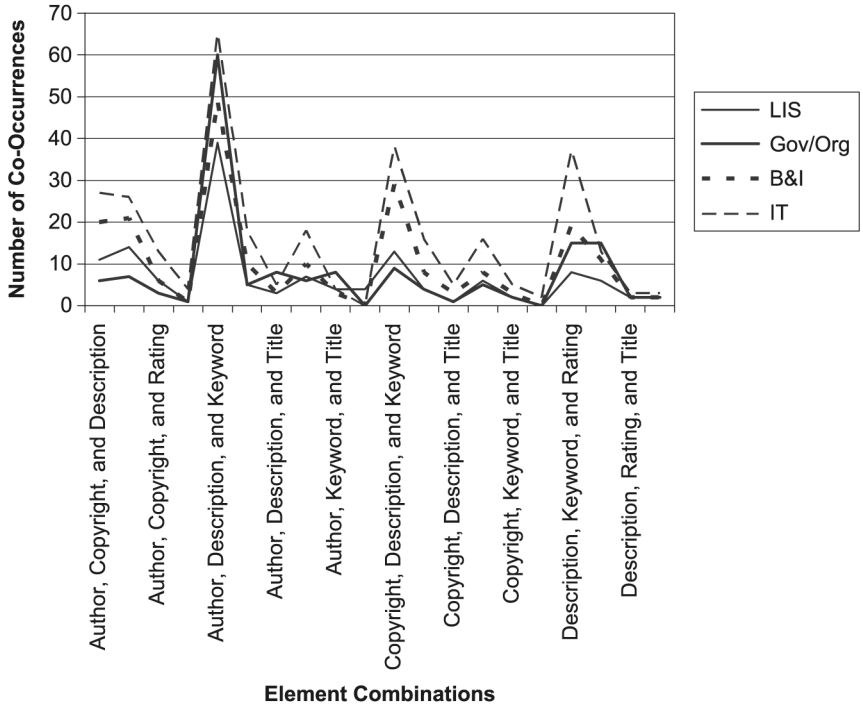

Figure 4.
Three element combination distribution in a selected element set

"author, copyright, description, and rating" (10.2 percent). Yet again, those combinations including "keyword", "description", or "author" were more common than those including other elements, and again the IT group appears to prefer combinations of at least four elements more often than do other groups (Figure 5).

2.6. Five and six element co-occurrence analysis in a selected element set

Similarly to the previous sections, the analysis of combinations of five and six metadata elements was conducted with the same selection of six metadata elements. There are six possible combinations of five elements ($C_6^5 = (6 \times 5 \times 4 \times 3 \times 2)/(5 \times 4 \times 3 \times 2 \times 1) = 6$), and there is one possible combination of six elements. Table V and Figure 6 show the detailed analysis of combinations found in pages having at least five and at least six metadata elements. The top combinations were "author, copyright, description, keyword, and rating" (65.8 percent); and "author, copyright, description, keyword, and title" (15.8 percent). Again, the IT domain appears to prefer combinations of elements more than do the other domains.

3. Conclusion

This study examines the co-occurrence of metadata elements in four different domains. It details the occurrence of pages having at least one or two of the 12 defined elements, and it examines the occurrence of pages having at least three, four, five, or six of the smaller six-element set defined in section 2.4. This analysis reveals that the "keyword" and "description" elements are the most popular single elements (occurring in 34.2 and

Combinations		Domain				
		Library and information science	Government agencies and non-profit organizations	Businesses and industries	Information technology	Total
Author, copyright, description, and keyword	Count	11	6	20	26	63
	Percentage within combination	17.5	9.5	31.7	41.3	100.0
	Percentage within domain	28.9	18.2	32.3	23.4	25.8
Author, copyright, description, and rating	Count	4	2	6	13	25
	Percentage within combination	16.0	8.0	24.0	52.0	100.0
	Percentage within domain	10.5	6.1	9.7	11.7	10.2
Author, copyright, description, and title	Count	1	1	1	4	7
	Percentage within combination	14.3	14.3	14.3	57.1	100.0
	Percentage within domain	2.6	3.0	1.6	3.6	2.9
Author, copyright, keyword, and rating	Count	6	3	6	13	28
	Percentage within combination	21.4	10.7	21.4	46.4	100.0
	Percentage within domain	15.8	9.1	9.7	11.7	11.5
Author, copyright, keyword, and title	Count	1	1	1	3	6
	Percentage within combination	16.7	16.7	16.7	50.0	100.0
	Percentage within domain	2.6	3.0	1.6	2.7	2.5
Author, copyright, rating, and title	Count	0	0	0	1	1
	Percentage within combination	0.0	0.0	0.0	100.0	100.0
	Percentage within domain	0.0	0.0	0.0	0.9	0.4
Author, description, keyword, and rating	Count	5	5	10	18	38
	Percentage within combination	13.2	13.2	26.3	47.4	100.0
	Percentage within domain	13.2	15.2	16.1	16.2	15.6
Author, description, keyword, and title	Count	3	8	3	4	18
	Percentage within combination	16.7	44.4	16.7	22.2	100.0
	Percentage within domain	7.9	24.2	4.8	3.6	7.4

(*continued*)

Table IV.
Four element co-occurrence analysis in a selected element set

Table IV.

Combinations		Domain: Library and information science	Domain: Government agencies and non-profit organizations	Domain: Businesses and industries	Domain: Information technology	Total
Copyright, description, rating, and title	Count	0	0	0	2	2
	Percentage within combination	0.0	0.0	0.0	100.0	100.0
	Percentage within domain	0.0	0.0	0.0	1.8	0.8
Author, description, rating, and title	Count	0	0	1	1	2
	Percentage within combination	0.0	0.0	50.0	50.0	100.0
	Percentage within domain	0.0	0.0	1.6	0.9	0.8
Author, keyword, rating, and title	Count	0	0	1	1	2
	Percentage within combination	0.0	0.0	50.0	50.0	100.0
	Percentage within domain	0.0	0.0	1.6	0.9	0.8
Copyright, description, keyword, and rating	Count	4	4	8	16	32
	Percentage within combination	12.5	12.5	25.0	50.0	100.0
	Percentage within domain	10.5	12.1	12.9	14.4	13.1
Copyright, description, keyword, and title	Count	1	1	3	4	9
	Percentage within combination	11.1	11.1	33.3	44.4	100.0
	Percentage within domain	2.6	3.0	4.8	3.6	3.7
Copyright, keyword, rating, and title	Count	0	0	0	2	2
	Percentage within combination	0.0	0.0	0.0	100.0	100.0
	Percentage within domain	0.0	0.0	0.0	1.8	0.8
Description, keyword, rating, and title	Count	2	2	2	3	9
	Percentage within combination	22.2	22.2	22.2	33.3	100.0
	Percentage within domain	5.3	6.1	3.2	2.7	3.7
Total	Count	38	33	62	111	244
	Percentage within combination	15.6	13.5	25.4	45.5	100.0
	Percentage within domain	100.0	100.0	100.0	100.0	100.0

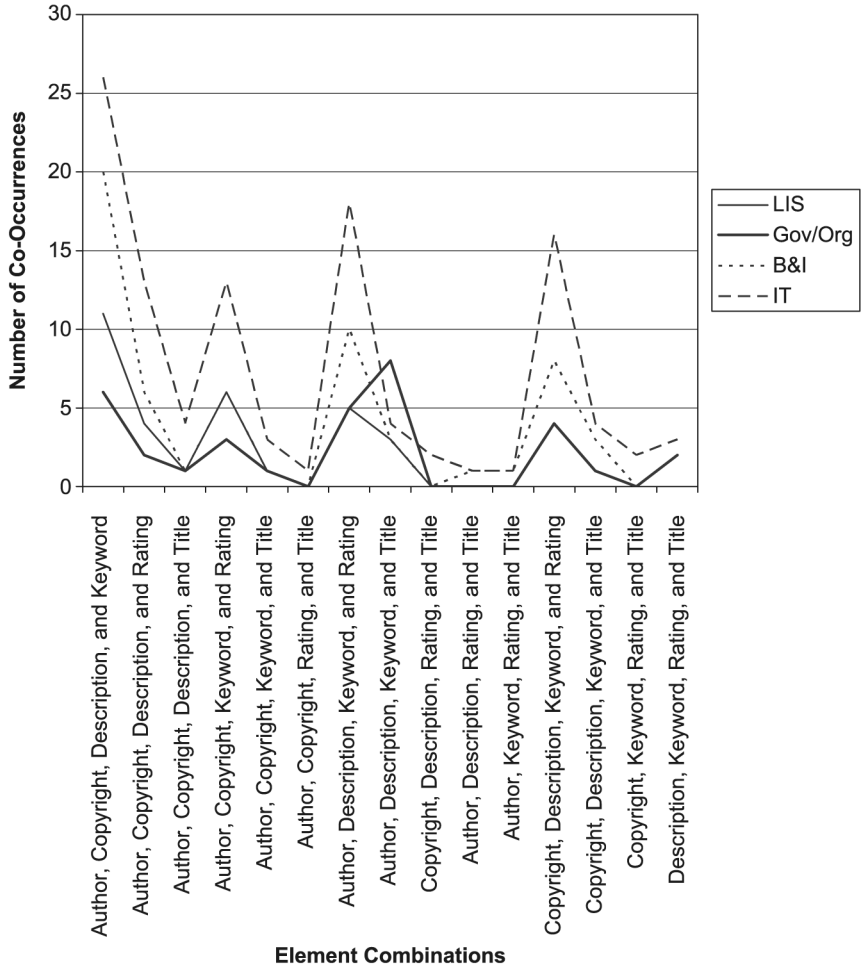

Figure 5.
Visual display of four element co-occurrence analysis

31.8 percent of the pages). The most popular combination of two elements was that of "keyword and description" (24.6 percent of the combinations of two elements). When combining at least three elements, web authors preferred "author, description, and keyword" (25.9 percent of the combinations of three elements), but when combining at least four elements they preferred "author, copyright, description, and keyword" (25.8 percent of the combinations of four elements).

Very few authors included combinations of five elements and only one used all six of the selected elements. Of those that used five elements, the most popular combination was "author, copyright, description, keyword, and rating" (65.8 percent of the combinations of five elements). Web authors using any combination of more than one element preferred to include "keyword and description" – elements that appear in the most popular of all combinations studied here. From an information retrieval perspective, this means that two of the most important and influential elements

Table V.
Five and six element co-occurrence analysis in a selected element set

Combination		Domain: Library and information science	Government agencies and non-profit organizations	Businesses and industries	Information technology	Total
All but title	Count	4	2	6	13	25
	Percentage within combination	16.0	8.0	24.0	52.0	100.0
	Percentage within domain	80.0	66.7	75.0	59.1	65.8
All but rating	Count	1	1	1	3	6
	Percentage within combination	16.7	16.7	16.7	50.0	100.0
	Percentage within domain	20.0	33.3	12.5	13.6	15.8
All but keyword	Count	0	0	0	1	1
	Percentage within combination	0.0	0.0	0.0	100.0	100.0
	Percentage within domain	0.0	0.0	0.0	4.5	2.6
All but description	Count	0	0	0	1	1
	Percentage within combination	0.0	0.0	0.0	100.0	100.0
	Percentage within domain	0.0	0.0	0.0	4.5	2.6
All but copyright	Count	0	0	1	1	2
	Percentage within combination	0.0	0.0	50.0	50.0	100.0
	Percentage within domain	0.0	0.0	12.5	4.5	5.3
All but author	Count	0	0	0	2	2
	Percentage within combination	0.0	0.0	0.0	100.0	100.0
	Percentage within domain	0.0	0.0	0.0	9.1	5.3
All six elements	Count	0	0	0	1	1
	Percentage within combination	0.0	0.0	0.0	100.0	100.0
	Percentage within domain	0.0	0.0	0.0	4.5	2.6
Total	Count	5	3	8	22	38
	Percentage within combination	13.2	7.9	21.1	57.9	100.0
	Percentage within domain	100.0	100.0	100.0	100.0	100.0

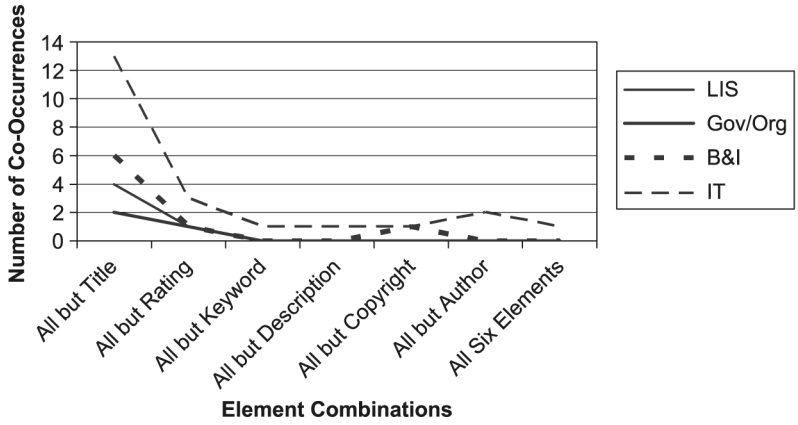

Figure 6.
Five and six element combination distribution in a selected element set

(according to Zhang and Dimitroff, 2005) are also the most commonly used by web authors and publishers.

Tellingly, web authors and publishers in all domains preferred to include only two elements in most of their web pages. There are more web pages that have only two elements than there are pages that have any other number of elements. A full 51.4 percent of web pages having metadata had two and only two elements. Even within single domains, combinations of two elements were more popular than combinations of any other number of elements, and in the B&I and IT domains, combinations of two elements were more popular than all the other possible combinations combined. This suggests that even when there are many choices of elements, from the general to the specific, most authors prefer minimal but effective metadata. In other words, they are not as interested in creating complete surrogates for their resources as they are in providing just enough meta-information for resource discovery.

This study also reveals that preferences for element combinations vary by domain. All domains preferred the "keyword" and "description" elements, but the LIS domain includes the "author" element in its second and third most common combinations of two elements, while the other three domains do not include the "author" element until their fourth most common combinations of two elements. The IT domain also favors combinations of greater numbers of elements more often than other domains do; while the frequency rates for combinations in the LIS domain dropped off the most dramatically of all the domains, after combinations of two elements. In fact, the LIS domain is the only domain to show a marked preference for including only one or two isolated elements. The Gov/Org domain prefers one, two, or three elements; the B&I domain prefers one to four elements; and the IT domain prefers up to four elements, but includes combinations of a greater numbers of elements than do the other domains.

The findings show that the four identified domains had similar metadata element co-occurrence patterns. But compared with the other three domains, the web authors in the information technology domain are more aggressive in terms of implementation of multiple metadata elements in their web pages. As the number of involved metadata elements increases, the differences between the information technology domain and the other three domains also increases.

These findings provide information about metadata use by web page authors, which may be useful for metadata standard development and revision. They identify user preferences as well as which elements are heavily used and could benefit from greater granularity, and which elements are rarely used and could be revised or even discarded. Search engine designers could also benefit from these findings. Understanding metadata and how metadata is actually used by web authors may enable them to develop accurate and user-centered search engine indexing algorithms.

Future research topics in this area include: investigating the co-occurrence of metadata elements in Dublin Core metadata, which is widely recognized in library and information professions; studying current trends in embedding web pages with more than one type of metadata; studying the use of element qualifiers in the metadata contexts; and exploring the co-occurrence of all twelve metadata elements in combinations of three or more elements.

References

Campbell, D. (2002), "The use of the Dublin Core in web annotation programs", *Proceedings of Int. Conference on Dublin Core and Metadata for e-Communities*, pp. 105-10.

Chepesuik, R. (1999), "Organizing the internet: the 'core' of the challenge", *American Libraries*, Vol. 30 No. 1, pp. 60-4.

Craven, T. (2000), "Features of the description meta tags in public home pages", *Journal of Information Science*, Vol. 26 No. 5, pp. 303-11.

Craven, T. (2001a), "Changes in metatag descriptions over time", *First Monday*, Vol. 6 No. 10, available at: www.firstmonday.dk/issues/issue6_10/craven/

Craven, T. (2001b), "Description meta tags in locally linked web pages", *Aslib Proceedings*, Vol. 53 No. 6, pp. 203-16.

Craven, T. (2001c), "Description meta tags in pages returned on different search engines", *Canadian Journal of Information and Library Science*, Vol. 26 No. 1, pp. 1-17.

Craven, T. (2001d), "Description meta tags in public home and linked pages", *Libres*, Vol. 11 No. 2, available at: http://libres.curtin.edu.au/LIBRE11N2/craven.htm

Craven, T. (2002a), "External descriptions of web pages: their features and their relationships to web page elements", *Libri*, Vol. 52 No. 1, pp. 36-47.

Craven, T. (2002b), "What is the title of a web page? A study of the webography practice", *Information Research*, Vol. 7 No. 3, available at: http://InformationR.net/ir/7-3/paper130.html

Craven, T. (2003), "HTML tags as extraction cues for web page description construction", *Informing Science Journal*, Vol. 6, pp. 1-12, available at: http://inform.nu/Articles/Vol6/v6p001-012.pdf

Craven, T. (2005), "Web authoring tools and meta tagging of page descriptions and keywords", *Online Information Review*, Vol. 29 No. 2, pp. 129-38.

Henshaw, R. and Valauskas, E. (2001), "Metadata as a catalyst: experiments with metadata and search engines in the internet journal, *First Monday*", *Libri*, Vol. 51 No. 2, pp. 86-101.

Hillman, D. (2003), "Using Dublin Core", *DCMI*, available at: http://dublincore.org/documents/usageguide/

Lagoze, C. (2001), "Keeping Dublin Core simple: cross-domain discovery or resource description?", *D-Lib Magazine*, Vol. 7 No. 1, available at: www.dlib.org/dlib/january01/lagoze/01lagoze.html

Richardson, T. (2003), "Search engine savvy", *Canadian Business*, Vol. 76 No. 24, pp. 99-102.

Search Engine Optimization (n.d.), available at: www.webmasterresources.com/search_engine_ranking/

Search Engine Optimization 1-2-3 (n.d.), available at: www.123-search-engine-optimization.com/engines.html.

Sokvitne, L. (2000), "An evaluation of the effectiveness of current Dublin Core metadata for retrieval", paper presented at the VALA Conference, 2000, available at: www.vala.org.au/vala2000/2000pdf/Sokvitne.PDF

Sullivan, M. (2003), "Keyword magic: the truth about search engine optimization for your website", available at: http://madiganpratt.com/KeywordMagic.pdf, pp. 1-8.

Tennant, R. (2003), "The engine of interoperability", *Library Journal*, Vol. 128 No. 20, p. 33.

Tennant, R. (2004), "Metadata's bitter harvest", *Library Journal*, Vol. 129 No. 12, p. 32.

Turner, T. and 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*, Vol. 42 No. 4, pp. 258-71.

Yahoo.com (n.d.), "How do I improve the ranking of my web site in the search results?", available at: http://help.yahoo.com/help/us/ysearch/ranking/ranking-02.html

Zhang, J. and Dimitroff, A. (2005), "The impact of metadata implementation on the webpage visibility in search engine results (Part II)", *Information Processing and Management*, Vol. 41, pp. 691-715.