Can we trust Web-page metadata?

Anders Ardö
Dept. of Electrical and Information Technology
Lund University, Box 118, SE-221 00 Lund, Sweden
e-mail: Anders.Ardo@eit.lth.se

September 24, 2009

Abstract

A statistical study of embedded metadata in a sample of more than 4 million HTML Web-pages is reported. The paper tries to determine and quantify the validity of this metadata. Of particular interest is to see if it is trustworthy enough for determining the topic of a Web-page. Datasets are collected by a Web crawler running both as a general and a focused crawler. Metadata fields 'title', 'author', 'keywords', 'description', and 'language' are analyzed in detail together with Dublin Core metadata. The study reveals problems with how metadata is created. Among the 75 % of all Web-pages that have interesting metadata, the field 'language' is the most trustworthy. All other metadata fields show a high degree of duplication thus degrading their usefulness. The strict answer to the title question is 'No', however there is a lot of meaningful and useful information, but it must be interpreted and used with care. The study also provides statistics on the usage of metadata today and how it has changed over time.

Keywords: Web metadata, metadata usage, metadata validity, metadata analysis, Dublin Core

1 Introduction

What metadata fields are used the most? How frequent is Dublin Core metadata? Is metadata in Web-pages relevant to the content? Does it describe the pages in a valid and reasonable way? Can it be useful for determining the topic of a Web-page? Some answers to these questions, and others, are found in this paper.

This paper attempts to make a contribution towards understanding metadata usage on the Web by investigating and determining the validity of a large sample (several millions) of HTML (World Wide Web Consortium, 1999) Web-pages with metadata embedded (<title> and <meta> elements). Metadata sources separate from the Web-page itself, like RDF and databases, are not utilized. Even though this type of metadata is relevant it is out of scope for this study. (Hillmann, Dushay, & Phipps, 2004) observes that metadata from repositories, possibly created by librarians, also have quality problems. Statistical methods is used on the sample in order to derive quality measures for the metadata. Almost no attempt of manual analysis or verification is done due to the labor and cost intensive nature. Metadata usage is also reported in detail.

Web page metadata play a key role in the Semantic Web framework (Herman, 2009). Though Semantic Web applications will not be analyzed here, it is important to know the quality of metadata before building applications that rely on it. For organizations writing metadata usage guides it is useful to know the most common mistakes and abuses. Of particular interest for this study is to see if metadata is trustworthy enough for determining topic of a Web-page, for example in order to create a vertical (topical) search engine. Metadata holds the promise to improve indexing in search engine databases, but does it live up to that promise? An earlier investigation to empirically determine the feasibility of using HTML tags to indicate page content (Hodgson, 2001) finds that 84 % of metadata fields and comments contain no semantic content. However this investigation is based on a small sample of 225 pages. Another study investigates usefulness of HTML structural elements and metadata in automated subject classification (Golub & Ardö, 2005). It is based on a collection of 1000 pages and concludes that all the structural elements and metadata are useful and contributes towards making accurate content classifications.

Unfortunately, many of the major search engines have stopped using metadata to improve relevance ranking, and some have even stopped indexing metadata because of the increase in spamming (the term given to the deliberate misuse of metadata in order to boost site ranking in search results, for example by repeating keywords hundreds of times) (FAO, 2007). Google only understands and uses the metatags "description" and "title" for describing content (Mueller, 2007), as do Microsoft's search engine Bing (Microsoft, 2009). Yahoo recommends the use of "title" and metatags "description" and "keyword" (Yahoo, 2007) and claims that it will improve ranking.

Two main sources of embedded metadata in Web-pages are

- manual creation by the author
- automatic generation by a program, like a HTML editor, word processor, or similar

Lazy authors, like me, can copy an old page, or a skeleton page, with metadata and then use it without changing the metadata. This can lead to many pages with identical metadata. As the analysis below show, a significant amount of duplicate metadata is found in the sample.

For the purpose of this study a Web-page is considered to consist of 3 parts, **title**, **metadata**, and **text**.

title is the content of the <title>-tag in the <head> section.

metadata is the name and content pair from the <meta ...>-tags in the <head> section

text is the content of the <body> section cleaned from HTML-tags, scripts and comments.

No attempt to correct misspellings or identify and map variants of common metatag field names is done, except where explicitly stated. Both field names and content are used exactly as they are in the Web-pages.

1.1 Datasets

The datasets used for this study are collected by a Web-crawler (Ardö, 2005) during the last few years. Web crawlers (Pant, Srinivasan, & Menczer, 2004) are programs that exploit the graph structure of the Web to move from page to page, retrieving them to a

local database. General purpose crawlers can process large portions of the Web without any specific focus, thus often leading to very large amounts of data (Google is a good example). In contrast a focused crawler cover specialized topics in more depth and keep the crawl more fresh, because there is less to cover for each crawler. A focused Web crawler (Chakrabarti, Berg, & Dom, 1999) is goal-directed using a classifier to evaluate the relevance of a Web-page with respect to focus topic. Such a crawler only stores relevant pages locally and only follows links from those relevant pages.

The Web crawler used here can work both main modes, either as a focused crawler or as a general crawler. Topic focus is maintained by an automated topic classification technique using a pre-defined topic definition, containing a controlled vocabulary of topical terms, to determine topical relevance (Ardö & Koch, 1999; Koch & Ardö, 2000).

There are 8 sub-datasets, 6 collected using focused crawls and 2 using general crawls. Statistics for these are in table 1. The 8 sub-datasets are:

GB Focused crawl, topic 'artist Gunnar Brusewitz', multiple languages (being small and very narrow this sub-dataset is dominated by one large site)

SE Focused crawl, topic 'search engines', English

Alg Focused crawl, topic 'algebra', English

CP Focused crawl, topic 'Carnivorous plants', multiple languages

Ei Focused crawl, topic 'Engineering' (based on the Ei thesaurus (Milstead, 1995), English

MS Focused crawl, topic 'Material Science', English

Delos General crawl, all pages from partners in the Delos EU project (Delos, 2009), multiple languages

Gen General crawl, no restrictions, multiple languages

The metadata fields 'content-type' and 'character-encoding' are excluded from the statistics below. Only documents of content type 'text/html' are considered. Furthermore the crawler obeys the 'Robots Exclusion Protocol', where a site or a Web-page can ask to be excluded from crawling, so such Web-pages are never crawled.

	Total	**Pages with**		
sub-dataset	**no of pages**	**meta fields (%)**	**title (%)**	**meta* fields (%)**
GB	17907	89.1	99.9	88.9
SE	911923	87.3	99.5	81.9
Alg	60481	58.1	97.3	47.1
CP	169787	86.4	97.8	72.9
Gen	491377	76.0	98.1	57.1
Ei	282578	81.3	99.1	72.8
Delos	74722	54.6	94.1	26.5
MS	2098052	85.8	99.3	79.7
Total	4106827	83.7	99.0	75.3

Table 1: Statistics for the sub-datasets. (meta* is interesting metadata, see text)

The detailed analysis is done on a subset of all metadata, a selection of interesting metadata fields, which from our perspective is defined as any metadata field whose

name contains any of the words: 'description', 'abstract', 'subject', 'classification', 'keyword', 'topic', 'category', 'author', 'creator', 'language', or 'title'. When the text refers to 'interesting metadata' or '**meta***' it's this subset that's referred to. 'Interesting metadata' implies interesting for determining topic of the content, and does not imply that the other metadata is uninteresting.

All sub-datasets together comprise 3093267 pages, from 925820 different Web-sites, with interesting metadata (meta*). Those pages contain on average 2.6 metadata instances per page, ranging from 1.7 up to 3.3 for the different sub-datasets. Metadata in these pages are analyzed in detail below.

A focused crawl, as done here, favors pages with metadata, thus the dataset might be biased towards Web-pages with metadata. As can be seen from table 1 the variation is large and datasets Gen, Alg and specially Delos have a lower proportion interesting metadata (even though Delos is a Digital Library project!). Thus a certain caution is recommended using these numbers to characterize the Web in general.

The crawls where made independently from each other, which can lead to the same Web-page being included in several sub-datasets. Using MD5 checksums based on the total content of a Web-page it was determined that a total of 19039 (less than 0.5 %) duplicates indeed were found in the total dataset. Deduplication made sure that these pages are only counted once in the detailed analysis below.

2 Background

Earlier studies of metadata use are summarized in table 2. NWIse and NWIdk (Koch, 2000) come from the now closed Nordic Web Index project (Ardö, Arvidsson, Hammer, Holmlund, & Lundberg, 1998) and were made some 10 years ago. From the last 4 years there are 2 other studies, MAMA (Wilson, 2008) and Google (Google, 2005).

	NWIse	**NWIdk**	**Google**	**MAMA**	**This**
year	1998	1999	2005	2008	2009
dataset size	2000000	3630172	1000000000	3509180	4106827
any metadata (%)	20.2		(87)		83.7
useful meta* (%)	7.5	13	(65)	77.2	75.3
title (%)		91.7	98	98.6	99.0
Dublin Core (%)		1.3			1.6
Most used metatags (popularity ranking)					
kewords	2	1	1	1	1
description	4	3	2	2	2
robots	13		3	4	3
generator	1		4	3	4
author	3	2	5	5	5
content-language			7		6
copyright	21	6	8	7	7
revisit-after			6	6	8
distribution	8		12	9	9
language	17	4	14	11	10

Table 2: Metadata usage studies. Numbers in parenthesis are estimated.

The different results in table 2 are not always directly comparable due to different ways of calculating the numbers. Despite this the studies show similar results. There is even a striking similarity for the metatags relative usage between this study and the Google study. It is interesting to note that the use of metadata have increased considerably since the earlier studies 10 years ago. This observation is confirmed by (O'Neill, Lavoie, & Bennett, 2003) who reports that the part of Web-pages that have metadata have grown from 45 % in 1998 to 70 % in 2002. In the same time interval the mean number of metadata instances per page is reported to be in the range of 2.2 to 2.8 which is well in accordance with our value of 2.6. This is of course due to the huge popularity of the two fields "keywords" and "description" (see table 3).

3 Analysis of <meta>-tag usage

There are more than 17000 unique metadata field names in the total dataset. The most used are detailed in table 3. Only metadata fields in the meta* selection are shown. Popularity is ranking according to usage of all metadata fields, i.e. including those not in the meta* selection. The <title>-tag in the <head> section is analyzed in section 3.1.

			% of	
Popularity	**metadata field**	**No of pages**	**all pages**	**meta* pages**
1	keywords	2850991	69.4	92.2
2	description	2838132	69.1	91.8
5	author	676324	16.5	21.9
6	content-language	500192	12.2	16.2
11	language	337009	8.2	10.9
13	classification	220439	5.4	7.1
19	title	97468	2.4	3.2
26	abstract	53081	1.3	1.7
28	subject	49796	1.2	1.6
30	dc.title	44052	1.1	1.4
34	dc.language	35946	0.9	1.2

Table 3: Usage for the most popular meta* fields.

The metadata fields **keywords** and **description** dominate, each being present in more than 90 % of all pages containing interesting metadata. Such a page normally contains several (average 2.6) instances of metadata. The popularity of the keywords field is most likely due to it's use by search engines such as Infoseek and AltaVista, while description is still used by many search engines. There is a big jump to the next one (author) which is present only in 22 %. There are ca 100 unique metadata field names that are present in more than 0.1 % of the pages containing meta* fields. Most of the tags are very uncommon and are either variants (misspellings, etc) of the common ones or have unknown meaning (like vw96.objecttype, y_key, icbm, or pd).

3.1 Title

The <title>-tag in the <head> section is the most common metadata and used in almost all Web-pages. There are 4064108 pages with title (99 % of the total dataset). It might

seem impressive that so many Web-pages actually have a title. However, 1293082 (32 %) of those are duplicates. Table 4 list the most common of these. The length distribution of titles looks reasonable (see figures 1 and 2) with 96 % being 20 words or shorter.

Instances	Title
24979	resor media shopping datorer hårdvara at 14k-gold-auctions.info
10324	- Mp3 Download, Biography and Discography.
6866	IHB Internationale Holzboerse
3178	Untitled Document
3095	Photos de plantes, fleurs et vegetaux - Florum
2794	In Search of Arctic Birds: Richard Vaughan, Gunnar Brusewitz: Amazon.co.uk: Books
2630	Dionee Association Francophone des Amateurs de Plantes Carnivores
2367	Online Technologies, Inc. | Audio Video Products, A/V, Presentation Equipment, Little Rock Audio Visual Audio Video Little Rock Projectors, Powerpoint Projector, Rental, Plasma, Home Theater
2211	dionaea
2140	Amazon.com: Hunting : hunters, game, weapons, and hunting methods from the remote past to the present day: Gunnar Brusewitz, Walstan Wheeler: Books
1965	Home

Table 4: Most common duplicated titles.

It seems that titles are often reused. This can have several reasons:

- using templates and not editing the title.
- copying an existing page to create a new page and not changing the old title.
- using a Web content managing system (WCMS) to publish a Web site. Such a system often generate identical titles automatically for all pages (unless title is explicitly given).

This reuse leads to a large proportion ($\approx$ 32 %) of titles that are identical for 2 or more pages. Thus they must be general in nature, not providing detailed, page specific information about the content, or in the worst case giving directly misleading information. While possibly being relevant to the general topic they will not help us differentiate between similar pages from the same Web-site.

3.2 Author

Web-pages with metadata field names containing any of the words 'author', 'dc.creator', 'creator' or variants anywhere in the field name, was selected and analyzed for author information. There are 732524 such Web-pages (23.7 % of the analyzed pages). There were 98132 unique values, which means that there are a lot of pages with the same author information.

Manual categorization of a random sample of 300 authors show that they fall in 3 roughly equally sized categories: personal names, company/departement/organization names, and others which includes things like emails and URLs. The two first categories are legitimate uses but the third (others, emails/URLs) might cause problems when the author field is used for searching.

Given the number of duplicates and the noise introduced by the third category (others), it seems hard to use this information without some external discriminator to determine at least the category (like personal name, company, email, department, or organization).

3.3 Keywords and description metadata

The two most common metadata fields (not counting title) are keywords and description. For this analysis Web-pages with nonempty descriptions (metadata field names that contain any of the words 'description', 'dc.description', 'abstract', or variants of those words) and nonempty keywords (metadata names that contain any of the words 'keyword', 'keywords', 'subject', 'dc.subject', 'classification', 'topic', or variants of those words) are selected. In the case there are several instances of such fields in a Web-page they are aggregated before the analysis. Most often, in 2606214 Web-pages, both description and keyword metadata are present.

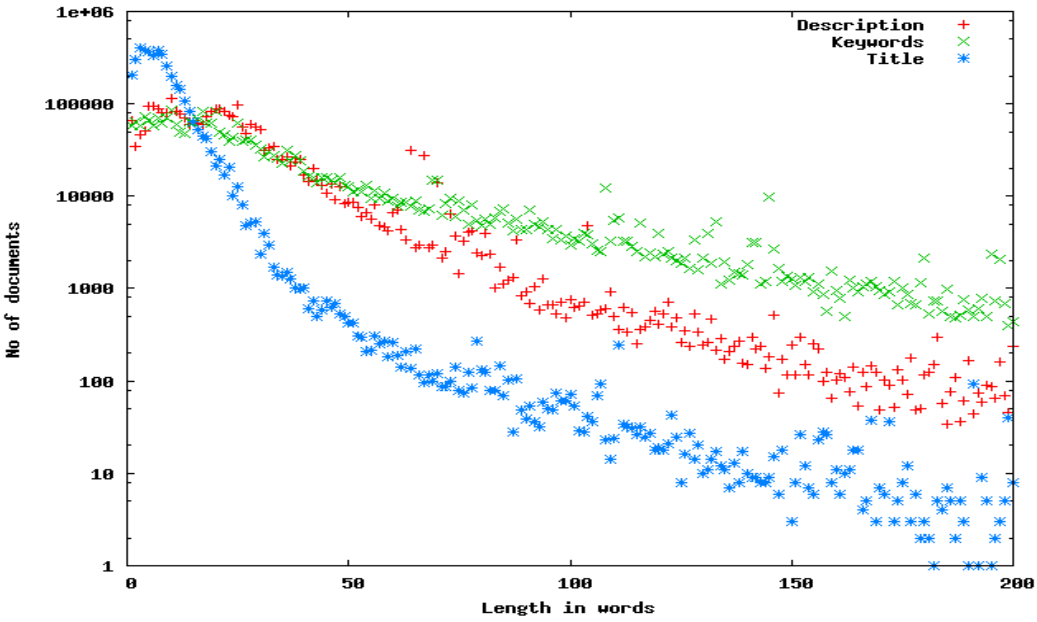

Figure 1: Distribution of title, description and keyword metadata length.

Somewhat surprising keyword metadata is on average almost 3 times as long as description metadata (73.5 words vs 24.6 words)! Details on the length distribution are shown in figures 1 and 2.

A manual inspection of a small random sample (200 items) of long keyword metadata (length > 100 words) indicates that it is indeed keyword lists, but that every imaginable word form of the relevant keywords or key-phrases are listed. This explains the lengthy keyword lists.

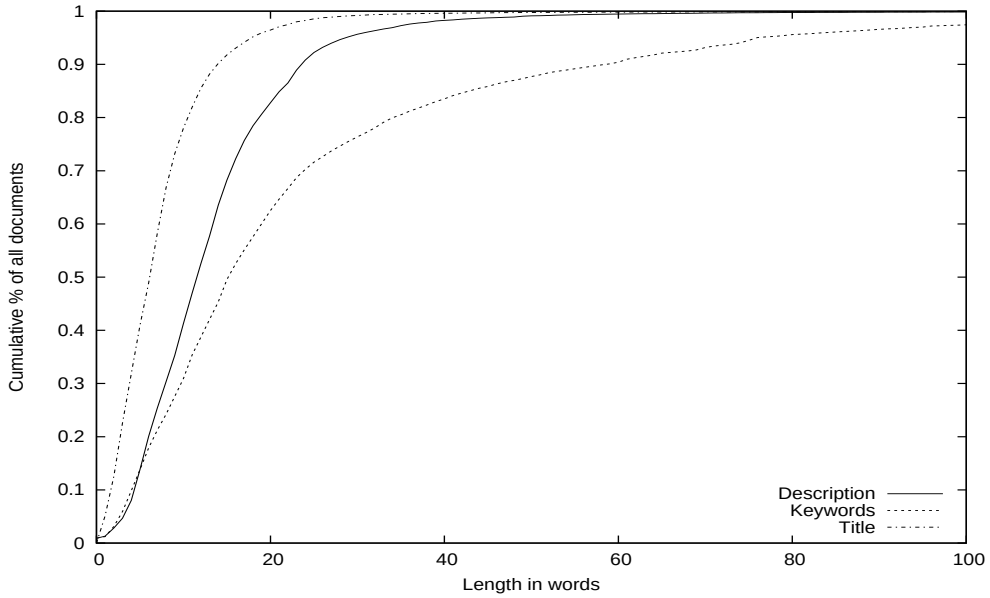

Figure 2: Cumulative length distribution for title, description and keywords.

Sometimes description metadata is actually a keyword list. In the dataset this is the case for 15 % of all description metadata pages. (A keyword list is here defined as a piece of text where the number of occurrences of the character ',' or ';' are more than 30 % of the number of words.) A manual inspection of long description metadata reveals that there is an even larger proportion of long descriptions that actually are keyword lists (32 % in the studied random sample of 200 items longer than 100 words).

It turns out that there is an alarmingly high proportion of duplicates, 58 % for description metadata and 61 % for keyword metadata. To be counted as a duplicate two instances have to have the same sequence of characters excluding whitespace characters. Reasons for this are hard to determine but similar effects as for title are likely culprits. Keywords for related or similar pages might be equal even when produced by a careful author, but descriptions for two similar pages should to show some difference, however small.

Description metadata and keyword metadata overlap significantly, and they share at least one word in almost all pages (96 %) that have both fields. Before calculating overlaps the text was cleaned from (English) stopwords, stemmed and all duplicate words within a field was removed. Overlap is then calculated as the percentage of the words in the smallest of the instances found in the other instance.

A detailed view of overlapping between description and keyword metadata is shown in figure 3. There is a strong but complex connection between the overlap and the size. Up to a size of around 25 words there is a negative correlation indicating that the longer metadata the larger the differences between them. But for sizes above 25 words the overlap is increasing with size indicating that the longer metadata we have the higher the probability that they are equal. Indeed for sizes above 100 words the probability for keyword and description metadata to be copies rapidly approaches 100 %.

Perhaps more interesting is the word overlap between keyword metadata and text in

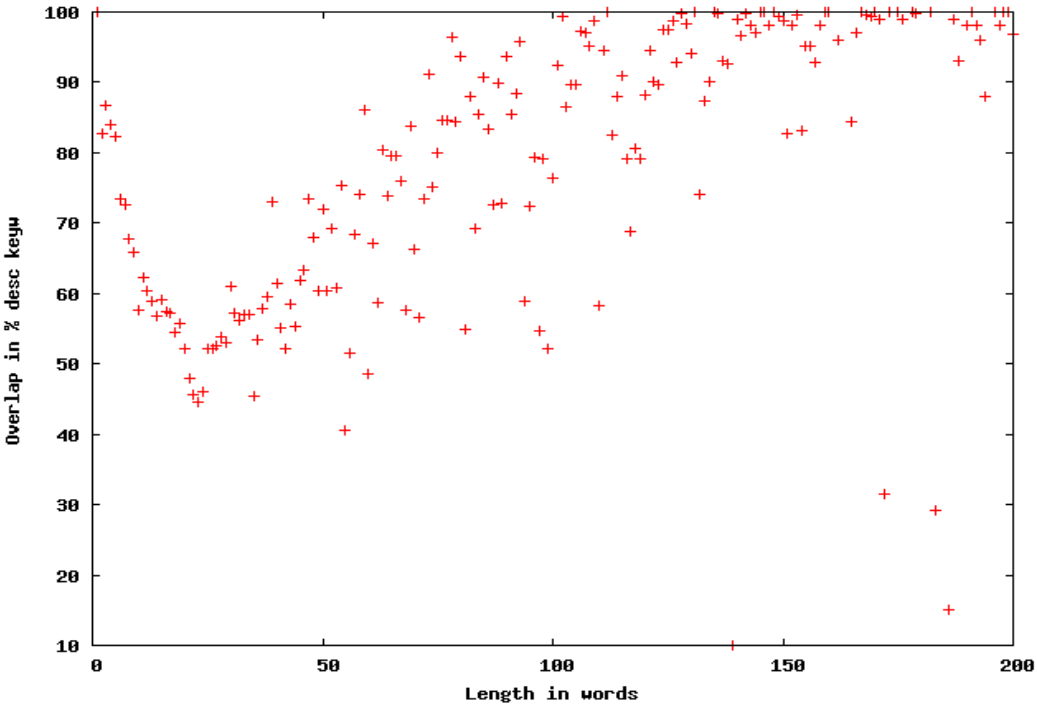

Figure 3: Keyword - description metadata overlap.

the Web-page. Figure 4 show a distinct relation between length of keyword metadata and overlap with the text. The shorter the keyword list is the more it overlaps with the text up to a length of around 50 words. For longer keyword metadata the curve levels out at 30 - 40 % overlap. Since most keyword metadata is below 50 words in length, the content of the Web-page is actually closely related to the keywords. The main observation is that the shorter the keyword metadata is the more it overlaps with the text in the Web-page. A qualitative study would reveal if the smaller overlap is due to smart selection of good keywords to complement the text or the use of non relevant keywords to attract a wider range of visitors.

3.4 Language

In the dataset there are 840060 Web-pages that have metadata with names that contain the word 'language', 'content-language', 'dc.language', or variants (excluding 'code-language') ie indicating the language of the content. These were compared to the result of a language identification program, Perl module Lingua::Identify (Simões, 2009), which uses a combination of 4 different methods (Small Word Technique, Prefix Analysis, Suffix Analysis, and N-gram Categorization) to determine the language. It supports 33 different languages. The methods are not perfect, but quite good, normally levels of 97 - 98 % correctness are cited. Each of the Web-pages with language metadata among the 33 supported languages and with a text of more than 100 characters (811935 pages) had their language identified with the language identifier and the result was compared to the given metadata. Results are detailed in table 5.

Language is important for Web-page topic classification since most automated topic classifier techniques are based on word lists, and therefore language dependent.

The dataset is heavily biased towards English (92.8 %). In more than 90 % of Web-

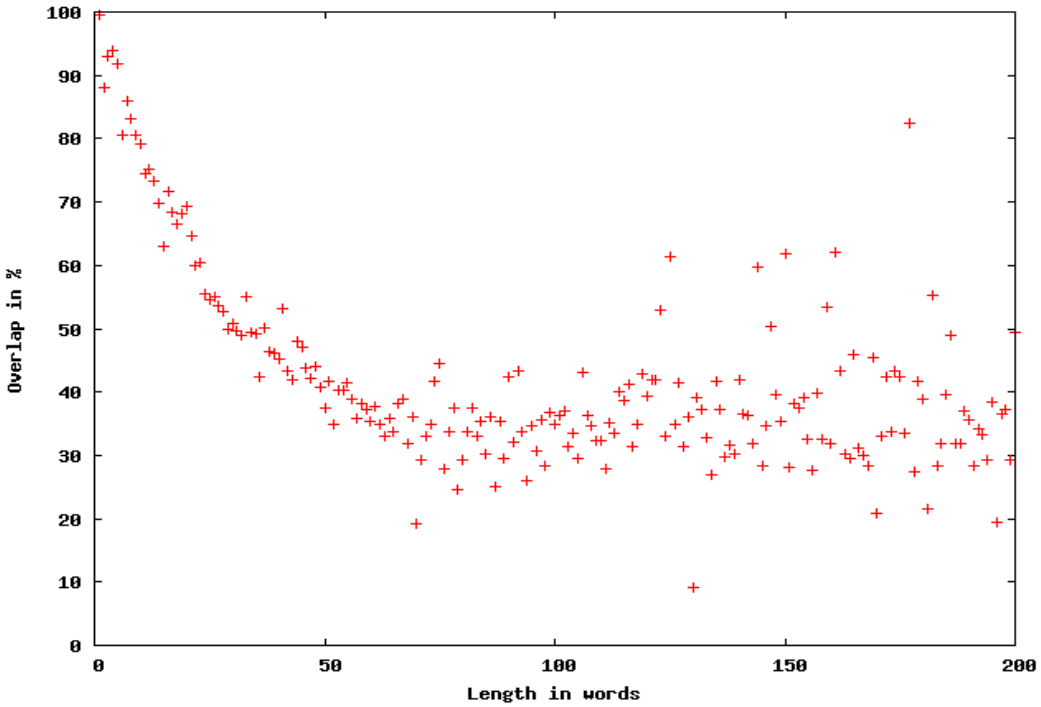

Figure 4: Keyword metadata and text overlap.

pages the two methods agreed, but with big variations between the sub-datasets. In general the multilingual datasets performed worse than the English only datasets. The most common mismatches are detailed in table 6(a).

Randomly selecting and manually analyzing 100 pages with mismatched languages it is clear that there is a number of different reasons for mismatches. Several pages had both languages (30 %) in them. The language identifier made errors on short pages and pages with lots of repetitions or non-words like part-numbers. Overall metadata was correct in 37 % of the cases while language identifier was correct in 20 %. In 13 % none of them were correct.

During analysis it was also noticed that some Web-pages had language metadata

Dataset	**Tested**	**OK (%)**	**Missmatch (%)**	**Language**
SE	197534	96.6	3.4	English
MS	493266	89.2	10.8	English
GB	967	80.9	19.1	multilingual
Alg	5873	90.2	9.8	English
Ei	33346	87.9	12.1	English
Delos	4370	72.6	27.4	multilingual
CP	24408	82.9	17.1	multilingual
Gen	52171	88.7	11.3	multilingual
Total	811935	90.7	9.3	

Table 5: Language analysis.

(a) Missmatches between Web-page metadata and identification program

Instances	Missmatches	
	metadata fields	identification program
30387	en	bg
9856	en	ru
4908	sv	en
3643	de	en
2432	sv	bg
2217	en	fr
1628	en	de
1517	fr	en
1491	en	da
996	it	en

(b) Statistics for metadata indicating more than one language

Different language metadata	
Instances	values
1619	cze,en
459	de,en
329	en,de
246	en,us
157	de,en,it,fr,es
157	de,en,it
157	de,en,it,fr,es,ru
157	de,en,it,fr
155	en,fr
134	en,cze

Table 6: Language verification.

that indicated several different languages. The most common of these specifications are given in table 6(b). It might of course be correct (a page with text in several languages) but for an automated system it is confusing and unclear which of the indicated languages to trust. On the other hand this only occurs in 0.6 % of Web-pages with language metadata so it is not a large problem.

In general the language metadata seems trustworthy to 95 %.

3.5 Dublin Core metadata

The Dublin Core Metadata Element Set v 1.1 (Dublin Core Metadata Initiative, 2008b) is a vocabulary of 15 properties for use in resource descriptions. Dublin Core (DC) - A Simple Content Description Model for Electronic Resources - is the most promising standard for providing metadata for electronic resources. DC is intended to facilitate discovery of WWW-pages and was originally conceived for author-generated descriptions. Here we concentrate on DC v1.1 (ie the dc: namespace) and do not attempt to analyze DCMI Metadata Terms (Dublin Core Metadata Initiative, 2008a) (the dcterms: namespace).

DC metadata were found in 66951 Web-pages with a total of 369612 instances of metadata. This represents 2.2 % of all pages with interesting metadata and 1.6 % of all pages. Web-pages that contain DC have on the average 5.5 DC metadata instances per page. Authors using DC provide more than twice as much metadata as the average.

DC elements can be refined or qualified which normally is expressed in the dot notation by tacking on the qualification at the end, for example to qualify the date to be the creation date would be expressed as "dc.date.created". The guiding principle for the qualification of Dublin Core elements, colloquially known as the Dumb-Down Principle, is that a client should be able to ignore any qualifier and use the description as if it were unqualified.

The most common of the 638 unique DC names (including qualifications) found, are shown in table 7.

Name	No of pages	Percent of meta*	Percent of pages with DC	Instances
dc.title	44052	1.3	65.8	44223
dc.language	35946	1.0	53.7	38135
dc.subject	30339	0.9	45.3	37851
dc.description	31401	0.9	46.9	31504
dc.creator	27692	0.8	41.4	29818
dc.publisher	20082	0.6	30.0	21293
dc.identifier	19182	0.6	28.7	20649
dc.format	15543	0.5	23.2	17515
dc.rights	16046	0.5	24.0	16537
dc.type	14126	0.4	21.1	14768
dc.date	12034	0.3	18.0	12030
dc.date.created	9580	0.3	14.3	9678
dc.date.modified	8752	0.3	13.1	8994
dc.coverage.placename	5650	0.2	8.4	5823

Table 7: Statistics for DC metadata.

Name	Instances	% of all instances
dc.title	44838	12.1
dc.subject	42846	11.6
dc.language	38180	10.3
dc.creator	37140	10.1
dc.date	36120	9.8
dc.description	32744	8.9
dc.publisher	24479	6.6
dc.identifier	22596	6.1
dc.format	17980	4.9
dc.type	17034	4.6
dc.rights	16868	4.6
dc.coverage	14115	3.8
dc.source	5976	1.6
dc.contributor	5896	1.6
dc.relation	2948	0.8
dc.contributors	2561	0.7
dc.keywords	2090	0.6
dc.audience	1132	0.3
dc.author	960	0.3

Table 8: Statistics for DC metadata dumbed-down to top-level.

Looking at dumbed-down, unqualified DC names (ie top level only) there are 164 unique names in the dataset, top ones are shown in table 8. DC v1.1 defines 15 such names! There are a number of explanations to this, somewhat surprising result. One is of course misspellings, ie dc.titel for dc.title and dc.subjekt, dc.subjects for dc.subject. But the major reason is the authors ignorance of and unfamiliarity with DC specifications. Examples of the later type (which are the majority) are dc.webmaster, dc.robots, dc.mrg, dc.dcrlocation, etc.

There seems to be a reluctance among Web-page authors to adopt formal metadata schemes like DC. According to (O'Neill et al., 2003) the use seems to be increasing with time (from 0.5 % 1998 to 0.7 % 2002) which has increased to 1.6 % now. Despite the overall low use of DC, this seems to indicate that it's use is increasing, which points to an effort among authors to use a standardized, formal version of metadata. Though the abuse of DC top level names is disappointing.

Comparing contents of Web page <title>-tag with dc.title (44052 pages) we see that there are 10789 pages where they are identical, corresponding to (24.5 % duplicates). That 75 % of dc.title are different from <title>-tag indicate that more care have been exercised in selecting a Dublin Core title, and it is probably more reliable in case both are present.

4 Conclusion

There is a lot of metadata out there, and the part of Web pages that have embedded metadata and even meta* is growing. The most common metadata is the title, present in 99 % of all pages, and the keywords and description metadata fields, each present in $\approx$ 92 % of pages with interesting metadata (meta*). However the use of a well defined and structured metadata standard is very low, only 2 % of the pages with metadata have some form of Dublin Core v1.1. And even DC is abused – the dataset have 164 unique top level DC element names out of the the 15 defined!

The use of metadata is very varied with some practices that degrade the usefulness, like lots of duplicates. Even content of the two metadata fields keywords and description quite frequently are copies. And content of specific metadata fields are used in different ways, for example author metadata can be a personal name, a company name, an email address, an organization name, or a department name. Description metadata can be just long keyword lists or a piece of prose.

All together it is hard to unconditionally trust all metadata to provide detailed information about the content of the Web-page. So the answer to the title question '*Can we trust Web-page metadata?*' strictly is '*No*', however there is a lot useful information, but it must be interpreted with care and the following practical observations in mind:

- there are a lot of duplicate metadata for all metadata fields
- 32 % of all <title>-tags are general in nature, not providing detailed page specific information about the content
- the shorter the keyword metadata is the better it overlaps with the main text
- keyword lists often contain many different word forms for the same concept
- long description metadata frequently are keyword lists
- for sizes above 100 words the probability for keyword and description metadata to be copies rapidly approaches 100 %

- language metadata can be trusted in ≈ 95 % of the cases
- even DC v1.1 top-names are abused

5 Acknowledgments

This research was supported by the Internet fund, .SE (The Swedish Internet Infrastructure Foundation http://www.iis.se/en/se-ar-mer/internetfonden/), in the project 'Vertical Search Engines'. Many thanks to Traugott Koch for providing many useful comments which helped improve the paper considerably.

References

Ardö, A. (2005). *Combine Web crawler.* [On-line]. Available: http://combine.it.lth.se/.

Ardö, A., Arvidsson, A., Hammer, S., Holmlund, K., & Lundberg, S. (1998). NWI II, an enhanced Nordic Web index: Final report. *NORDINFO Nytt*, *21*(3-4), 66-75.

Ardö, A., & Koch, T. (1999). Automatic classification applied to the full-text Internet documents in a robot-generated subject index. In *Online information 99, proceedings* (p. 239-246). [On-line]. Available: http://www.it.lth.se/anders/online99/.

Delos. (2009). *DELOS is a Network of Excellence on Digital Libraries.* [On-line]. Available: http://www.delos.info/.

Dublin Core Metadata Initiative. (2008a). *DCMI Metadata Terms.* [On-line]. Available: http://dublincore.org/documents/dcmi-terms/.

Dublin Core Metadata Initiative. (2008b). *Dublin Core Metadata Element Set, Version 1.1.* [On-line]. Available: http://dublincore.org/documents/dces/.

FAO. (2007). *Search engines and metadata.* [On-line]. Available: ftp://ftp.fao.org/gi/gil/gilws/aims/publications/papers/metata_and_search_engines.pdf.

Google. (2005). *Web Authoring Statistics.* [On-line]. Available: http://code.google.com/webstats/.

Microsoft. (2009). *Bing - New Features Relevant to Webmasters.* [On-line]. Available: http://download.microsoft.com/download/0/D/9/0D94EECB-C767-445E-B708-9C829275995F/Bing–NewFeaturesForWebmasters.pdf.

Yahoo. (2007). *How do I improve the ranking of my web site in the search results?* [On-line]. Available: http://help.yahoo.com/l/us/yahoo/search/ranking/ranking-02.html.

Chakrabarti, S., Berg, M. van den, & Dom, B. (1999). Focused crawling: a new approach to topic-specific Web resource discovery. *Computer Networks (Amsterdam, Netherlands: 1999)*, *31*(11–16), 1623–1640.

Golub, K., & Ardö, A. (2005, September). Importance of HTML Structural Elements in Automated Subject Classification. In A. Rauber, S. Christodoulakis, & A. M. Tjoa (Eds.), *9th European Conference*

on Research and Advanced Technology for Digital Libraries - ECDL 2005 (Vol. 3652, p. 368 - 378). Springer. [On-line]. Available: http://www.it.lth.se/Knowlib/publ/ECDL05.pdf.

Herman, I. (2009). *W3C Semantic Web Activity.* [On-line]. Available: http://www.w3.org/2001/sw/.

Hillmann, D., Dushay, N., & Phipps, J. (2004). Improving metadata quality: Augmentation and recombination. *International Conference on Dublin Core and Metadata Applications*. [On-line]. Available: http://dcpapers.dublincore.org/ojs/pubs/article/view/770.

Hodgson, J. (2001, Jan/Feb). Do html tags flag semantic content? *IEEE Internet Computing*, *5*(1), 20-25.

Koch, T. (2000). *Nordic Metadata usage.* [On-line]. Available: http://web.archive.org/web/20061211085728/http://www.lub.lu.se/metadata/Nordic-MDusage.html.

Koch, T., & Ardö, A. (2000). *Automatic classification of full-text HTML-documents from one specific subject area.* [On-line]. Available: http://www.it.lth.se/knowlib/publ/DESIRE36a-WP2.html.

Milstead, J. L. (Ed.). (1995). *Engineering Information Thesaurus* (revised, 2nd ed.). Engineering Information Inc.

Mueller, J. (2007). *Official Google Webmaster Central Blog: Answering more popular picks: meta tags and web search.* [On-line]. Available: http://googlewebmastercentral.blogspot.com/2007/12/answering-more-popular-picks-meta-tags.html.

O'Neill, E. T., Lavoie, B. F., & Bennett, R. (2003, April). Trends in the Evolution of the Public Web. *D-Lib Magazine*, *9*(4). [On-line]. Available: http://www.dlib.org/dlib/april03/lavoie/04lavoie.html.

Pant, G., Srinivasan, P., & Menczer, F. (2004). Crawling the web. In M. Levene & A. Poulovassilis (Eds.), *Web dynamics - adapting to change in content, size, topology and use* (p. 153-178). Springer.

Simões, A. M. B. (2009). *Lingua-Identify Perl module.* [On-line]. Available: http://search.cpan.org/~ambs/Lingua-Identify-0.23/.

Wilson, B. (2008). *MAMA: What is the Web made of?* [On-line]. Available: http://dev.opera.com/articles/view/mama/.

World Wide Web Consortium. (1999). *The global structure of an HTML document.* [On-line]. Available: http://www.w3.org/TR/html401/struct/global.html.