The state of metadata in open access journals: possibilities and restrictions

Francke, Helena

Published in:
ELPUB:2008

2008

The State of Metadata in Open Access Journals: Possibilities and Restrictions

Helena Francke

Department of Cultural Sciences, Lund University
SE-223 62 Lund, Sweden
and
Swedish School of Library and Information Science,
Göteborg University and University College of Borås, Sweden
e-mail: helena.francke@hb.se

Abstract

This paper reports on an inquiry into the use of metadata, publishing formats, and markup in editor-managed open access journals. It builds on findings from a study of the document architectures of open access journals, conducted through a survey of 265 journal web sites and a qualitative, descriptive analysis of 4 journal web sites. The journals' choices of publishing formats and the consistency of their markup are described as a background. The main investigation is of their inclusion of metadata. Framing the description is a discussion of whether the journals' metadata may be automatically retrieved by libraries and other information services in order to provide better tools for helping potential readers locate relevant journal articles.

Keywords: scholarly journals; metadata; markup; open access; information access

1. Introduction

This paper will report on an inquiry into the use of metadata, publishing formats, and markup, in editor-managed open access journals[1]. The open access movement endorses and is actively working towards the possibility for everyone with an Internet connection and sufficient information literacy to be able to access scholarly contributions on the Web. However, given the amount of documents and services on the Web, making content available is no guarantee that it will also be found by the intended target groups. Although there are several ways for authors and publishers of open access scholarly journals to address the problem of their products "being found", including Search Engine Optimization, many of them require a potential reader to either already be familiar with the journal or to enter a suitable search query into a search engine. The latter is presumably the most common locating tool that readers use [2]. Making the journal articles searchable through OAI-compliant repositories or library online catalogues can aid in bringing articles to the attention of potential readers, often with the additional perk that comes with positioning the articles within the context of the journal to a larger extent than is the case when individual files are found through a search engine. For small publishers of scholarly journals [3], particularly in cases where an open access journal is run on a low budget by an individual or an organization such as a university department or library [4],[5], it may be difficult to find the time and resources to promote the journal.

Libraries and other information services may provide help with collecting and making available article metadata from this group of journals in order to increase their visibility. Such projects already exist, e.g. the Lund University Library's DOAJ [6] or the University of Michigan's OAIster [7], but these services still require input from the journals in the form of harvestable metadata. If article metadata could be retrieved directly from the journal web sites without a need for the publishers to provide it in a specific

format, there would be better opportunity for libraries to work with publishers of small and local journals so as to help them target a world-wide audience [cf. e.g. 8]. In this paper, I will present findings concerning the use of metadata, publishing formats, and markup in editor-managed open access journals [5, p. 5] that can be of use for librarians, scholars, and computer scientists who are considering taking on such tasks. Focus in the paper is on which metadata are included and marked up in the journals; the choice of format and markup consistency are included because they constitute important prerequisites for how metadata may be reused.

2. Methodology

The data were collected through a combination of qualitative and quantitative methods. This allows for conclusions to be drawn both across journals and across the different issues and articles within individual journals. The document architectures of the journals were studied with regard to their choice of publishing format, their use of markup in cases where markup languages were used, and the marked up and visible metadata or bibliographic data included. The study looked at three levels of the journals: the start page, the table of contents pages, and the article pages. The quantitative study comprised 265 journals. The most recent issue and its first article were studied, and for some variables the first issue published online was also included. The qualitative study included four journals, which were investigated in greater detail, including all or most of the issues and a few articles for each issue. The margins of error for each variable in the statistical study were estimated with 95% confidence by using Jowett's method [9], [10].

2.1 Journals included in the study

The focus of the study was on journals that are published by small open access publishers. These journals are often run by individuals or groups of individuals, or sponsored by universities or university libraries, and they may be termed editor-managed journals [5, p. 5] because much of the publishing work is made by editors who are subject specialists rather than professional publishers. The journals included in the sampling frame were identified through the DOAJ [6] and Open J-Gate [11] databases and was restricted to those journals that were peer reviewed, published the web site in one of the languages Danish, English, French, German, Norwegian, or Swedish, that were open access, and could be considered editor-managed. From the sampling frame of approximately 700 journals (in spring 2006), a random sample of 265 journals was drawn. The majority of the journals in the sample, 70.2%, were published by university departments. Another 9.8% were published by university presses or e-journal initiatives, and 7.2% each by another type of non-profit organisation or under the journal's name. English was the most common language, with 85.3% of the journals having this as their main language. The journals represented every first level subject category included in DOAJ.

The four journals in the qualitative section were selected mainly because they use web technology in an innovative or interesting fashion. This was of relevance to other parts of the study than those reported in this paper. The journals were all from the humanities or education, namely: *assemblage: the Sheffield graduate journal of archaeology*, *The Journal of Interactive Media in Education (JIME)*, *The Journal of Music and Meaning (JMM)*, and *The International Review of Research in Open and Distance Learning (IRRODL)*.

3. Results

From the study outlined above, data have been selected for presentation that concern three different areas: the publishing formats of the journals, their use of (X)HTML markup, and their inclusion of metadata. Focus is on marked up metadata included in the journal files at the various journal levels. To what extent

do editor-managed open access journals include marked up metadata, and are the text strings that are marked up in this way potentially useful for various forms of automatic collection of metadata into a system? However, marked up metadata requires a file format based on a markup language of some sort. This motivates an initial look at the publishing formats used in the journals at the various journal levels. The usefulness of the metadata, as well as other marked up text is also to some extent restricted by how the markup has been performed. Therefore, the predictability and validity of the journal's markup is also discussed before turning to a more thorough report of the inclusion of marked up metadata.

3.1 Publishing formats

The start page of a Web-based journal is often intended to be a mutable space where news and updates are added regularly. The page also often functions as a portal with a collection of hyperlinks to the other parts of the journal web site. It is therefore not surprising that the start pages of all the journals in the sample publish through some version of (X)HTML. Most journals also have separate table of contents pages for each issue. These pages have a higher degree of permanency than the start pages, because they are generally not updated once the issue has been published. In most cases, their primary function is to direct the visitor to one of the issue's articles. When these pages exist separately, they are (X)HTML based, but in 5 of the journals the issue is published as a single unit in PDF or DOC, and the table of contents is placed at the beginning of that file.

At the article level, the variety of file formats is much wider, but (X)HTML and PDF are by far the most common ones. As many as 67.1 to 78.1% of the journals in the population publish the articles in their latest issue in PDF, whereas between 36.6 and 48.8% of the journals use (X)HTML. The articles in somewhere around one fifth of the journals are actually made available in more than one file format, and it is often the case that both PDF and (X)HTML are used. Furthermore, the proportion of journals with PDF as the publishing format for the articles is higher in the latest issues than in the first ones, with a corresponding decline in the popularity of (X)HTML. There are many reasons that could account for why PDF has become more popular. These include a desire on the part of the journals to use a file format that indicates permanency, something that is often associated with credibility; the ease of using the same file for derivatives in several media (notably print and Web); and a wish to facilitate for readers who print the articles before reading.

Publishing format	1st issue	Latest issue
HTML non-specified	82	53
HTML 2.0	2	--
HTML 3.2	8	3
HTML 4.01 Transitional	26	29
HTML 4.01 Frameset	2	3
XHTML 1.0	18	25
(X)HTML Total	*139*	*113*
PDF	158	193
PostScript	10	9
MS Word	5	4
RTF	3	1
DVI	5	5
Hyperdvi	1	--
DjVu	2	2
TeX	3	3
ASCII/txt	4	--
WordPerfect	1	--
Mp3	1	1
PNG	1	--
EPS	--	1

Table 1: Frequency of publishing formats in the journals, including journals that publish their articles inmore than one format. First peer reviewed article in the first and most recent issue published on the journal web site.

Other file formats found occasionally at the article level are various LaTeX output formats such as DVI, Hyperdvi and TeX, as well as PostScript and DjVu. Apart from the latter format, these exist solely in journals within the areas of mathematics and computer science. A few occurrences of MS Word, RDF, and TXT were noted, and one journal – *IRRODL* – contained MP3 versions of some of its articles (see also Table 1).

One of the consequences of the dominance of PDF and the decline of the use of (X)HTML at article level is that fewer journals provide the possibility of including marked up metadata at article level. Rather, the issue level becomes more important as a potential location for metadata, even for metadata describing an article rather than an issue. Some journals also offer a "paratext page", generally positioned between the issue's table of contents page and the article page, where they include non-marked up metadata (or paratexts) describing the article. This page can include information on the author(s) and the journal, and various descriptions of the article such as title, abstract, keywords, and sometimes even references. As these paratext pages are generally in (X)HTML, this can be a spot to also identify marked up metadata. However, a consequence of the limited use of (X)HTML at article level is that the places to look for marked up metadata in the journals varies depending on which file formats are used at which levels.

3.2 Markup

The predictability and validity of the markup of (X)HTML pages may affect the possibilities to make use of the markup in various ways. If elements are correctly and consistently marked up it is easier to identify and extract them for specific purposes. This includes identifying an article's title through a <title> tag, finding words occurring in headings, block quotes, or image texts, and the use of XPath to locate a specific position in a document. Among the journals that were studied, very few made use of valid (X)HTML markup. Among start pages, 6.8% passed validation and the corresponding figure at the article level was 8.0%. Due to the low proportion of articles that were published in (X)HTML, this means that between 1.6 and 6.4% of all journals can be expected to publish articles with valid (X)HTML markup. It should be acknowledged that validation of the pages was made automatically, using the fairly strict W3C validator, and that no evaluation was made in the survey of the types of errors that it reported. A closer inspection of the types of errors that came up in the validation of one of the journals in the qualitative study illustrates how attempts to accommodate various (older) web browsers can cause the markup to break W3C recommendations. Thus, a conscious choice may in some cases have been made that has resulted in a minor violation of the recommendations.

It was clear in the sample that a majority of the valid (X)HTML pages were found among start pages and articles where XHTML 1.0 was the HTML version used; this was the case in two thirds of the valid pages. With one exception, the remaining third of the valid pages were HTML 4.01 Transitional. A concern with validity (or the use of editor software that generates more correct markup) was thus found primarily among those web sites that use newer versions of (X)HTML. At the same time, only half of the start pages and article pages in the sample that used XHTML 1.0 had valid markup.

So far, the (X)HTML validators are not intelligent in the sense that they take into account whether or not the marked up content of the elements fit the logic for which they are marked up. It is, for instance, quite possible to mark up a section of the body text as a heading, such as <h3>, and this is sometimes done in order to achieve a specific visual effect. However, if one wishes to use markup for identifying and retrieving content, it is of importance both that the markup is used for a text string of the content type indicated by that markup element and that all the content of that type is marked up with the correct element and not with other elements. For instance, if one wishes to use the element <blockquote> in order to locate and extract any block quotes in the articles of a journal, this will only be successful if block quotes have in fact been marked up as such and not as, e.g., <dir><dir><font size=-1>, and if <blockquote> has not been used

to achieve a desired visual appearance for, say, the abstracts.

The markup of three types of content was studied in the survey, namely headings, block quotes, and the inclusion of alternative text as an attribute in image elements. These three types were chosen because headings are a common element on a web page and may contain terms that are significant to describe an article's content, block quotes have close ties to the scholarly article as a genre and indicate a reference to somebody other than the article author(s), and the "alt" attribute could give an indication of what an image represents through means that are possible to use in text – rather than image – retrieval. Of these, block quotes was the element that was used correctly most often, namely in 51.3% of the cases. On the other hand, because many of the journals publish in other formats than (X)HTML and given that block quotes are less common than, for instance, headings, only between 11.3 and 20.4% of journals contain correctly marked up block quotes. Some journals that do not mark quotes using the <blockquote> element nevertheless indicates the function of the string of text by including block quote as a class, name, or ID attribute.

All articles can be expected to contain headings, if nothing else then at least an article title, which would presumably be marked up as a heading of the highest degree. Just under half of the journals in the sample with articles in (X)HTML use <h> for headings, and slightly more than 40% of these journals have headings marked up according to hierarchy beginning at the topmost level and downwards. A further 15.7% adhere to hierarchy but do not begin with <h1>. In total, between 7.5 and 15.3% of all the journals can be expected to use <h> to identify headings hierarchically. The "alt" attribute to the image element – optional in earlier versions of HTML but compulsory in later versions – was included in slightly under one third of the articles that contained the <img> element, which was 75.2% of the journals in the sample publishing articles in (X)HTML. A few articles contained the "alt" attribute, but it was left without content. This means that the total proportion of journals with "alt" attributes that could be used for various purposes is between 4.4 and 11.0%.

In the survey, the markup was studied in the first peer reviewed article in the most recent issue of each journal. The qualitative studies indicate that there can be large variations to how markup validity and predictability are handled between different issues of the same journal and even between articles in the same issue. At the moment, this makes the use of markup an unreliable means to identify specific logical elements in the articles.

3.3 Metadata

The journals' start pages, issue pages, paratext pages, and article pages contain data that describe the articles and the journal in various ways. This information can be divided into that which is marked up according to its content type and that which is not marked up but whose content type can be identified by a person or, in some cases, automatically through an algorithm that can identify such specific features as a copyright sign or a phone number. Focus here will be primarily on marked up, machine-readable metadata, which is the type most easily usable in, for instance, various projects for automated data collection (for more results concerning the non-marked up type, see [1]). The types of machine-readable metadata that will be discussed are those marked up by the <title> and <meta> elements, including <meta> elements that make use of elements from the Dublin Core Metadata Element Set. The occurrence of RSS feeds will also be briefly discussed. Three things are of particular interest in this context:

1. to what extent are various types of marked up metadata included at various levels in the journals?
2. what content is entered into the metadata elements? and
3. which levels of the journal do these metadata describe (journal, issue, article)?

Type of content in the <title> element	% of journals – issue level (n=265)	% of journals – article level (n=265)	% of journals with <title> – article level (n=112)
Journal title	38.9	7.6	44.6
No/vol. of issue	5.3	--	24.1
"Current issue" or similar	1.1	n.a.	--
2 of the above	40.4	1.9	n.a.
Article title	n.a.	7.9	53.6
Name of author	n.a.	2.3	33.9
Article title and author	n.a.	3.4	n.a.
More or other of the above	n.a.	18.1	n.a.
Other	10.9	1.1	1.1

Table 2: Types of content included in the <title> element at issue and article level. In the right-most column, the composite values have been broken down into single values. [1, p. 247]

In the presentation of findings that follows, it may be good to keep in mind that the file formats that the journals use vary at the different levels. All the journals use (X)HTML on their start pages and almost all (98.1%) for the table of contents pages (the issue level). The use of (X)HTML is less common at article level, where it is found in the most recent issues of 42.6% of the journals. This means that when the article level is discussed below, only this smaller sample of (X)HTML files has formed the basis for the results.

The most commonly occurring metadata type is the <title> element, which can be found at the start page and issue levels in at least 95% of the journals and at the article level of a minimum of 93% of the journals publishing in (X)HTML at this level. The journal title is the most commonly included information in the <title> element at the issue level, occurring in between 73.9 and 84.0% of the journals. Information on the issue and/or volume number, or a text that indicates that it is the "current issue" occurs in between 40.7 and 53.0% of the journals. Both the journal title and the issue/volume number also occur fairly frequently in the <title> element at the article level – the title in just below half of the journals and the issue/volume number in about a quarter of them. Approximately as common – slightly more common in the sample, in fact – are the article title and the name of the author(s). Between 43.9 and 63.1% of the journals include the article title in the <title> element of the article files. However, at this level, it is not entirely uncommon for the <title> element to contain a number of different types of information. The figures of the most common content types are listed in Table 2.

Some variety can also be found among the words listed in <title> – many are quite generic, such as "Article/s", "contributions", "Mainpage", or "Default Normal Template", whereas others provide additional information that may be used to identify the journal, support its credentials, or advertise the journal, such as the name of the publisher or the ISSN. Very few of the <title> elements contain nonsensical text.

A particular problem can be caused by journals that use frames. In many cases, frames mean that if the content of a <title> element on a page is to be used for some purpose, a decision has to be made with regard to which file (and <title> element) should be preferred over the others. Perhaps the most likely candidates are the frameset file and the file which contains the article text. However, these can have different text in their <title> elements. One of the journals in the qualitative study illustrates this, and also that there was some inconsistency in what content was included in the <title> elements of similarly positioned files in different issues (this was also the case in another of the journals in the qualitative study). The differences are by no means very large, but it is not uncommon for the content to be formulated according to varying patterns (abbreviations, notation, order, etc.) or to contain slightly different types of content. Overall, the variety of exactly what the <title> element contains is quite wide and covers many more types of content than, for instance, the main heading of the pages.

A comparison of the content in the <title> element and that marked up as DC.title (only few journals make use of the Dublin Core title element, 12 journals at the issue level and 14 at the article level) shows that the content is similar in most cases (10 journals at the issue level, 8 at the article level). In the few other cases, the Dublin Core elements sometimes contain more precise content in the form of the article title where the <title> equivalent has more types of content, and sometimes the Dublin Core element contains generic content such as "Article". However, since the Dublin Core title element is much less common than the <title> element, and in many cases contains the same information, it does not seem to be particularly useful to target specifically.

A type of markup that is of specific interest in this case is the <meta> element available in (X)HTML, which can be used for marking up various types of metadata – in the words of the HTML 4.01 specification, "generic metainformation" [12, sect. 7.4.4]. The attributes *name* and *http-equiv* are used to describe the type of metadata (or property) that is included, and the attribute *content* to include the metadata text itself (the value). As the HTML specification does not restrict the properties that are possible to use, some variety in properties is likely to be encountered, but some properties have emerged as more common than others. Among the 90% of the journals in the sample that included a <meta> element, most used the technically oriented *http-equiv* with various properties. The details of this attribute were not included in the study.

Among the properties associated with the *name* attribute, the most commonly used were keywords, description, and generator (see Table 3). Keywords and description, in particular, have emerged as quite frequently found on the web sites. Apart from some of the journals that include *http-equiv*, files often contain more than one <meta> element. Combinations of the properties keyword, description, and http-equiv and of http-equiv and generator are the most common (the two latter properties are likely to be included by the software employed and seldom requires the person marking up the text to fill out the values).

Type of metadata	Journal level (n=265)		Issue level (n=260)		Article level (n=113)	
http-equiv	206	(77.7%)	199	(76.5%)	91	(80.5%)
keywords	98	(37.9%)	74	(28.5%)	30	(26.5%)
description	93	(35.1%)	74	(28.5%)	31	(27.4%)
generator	66	(24.9%)	71	(27.3%)	40	(35.4%)
author	36	(13.6%)	30	(11.5%)	20	(17.7%)
robots	16	(6.0%)	11	(4.2%)	4	(3.5%)
copyright	11	(4.2%)	10	(3.8%)	6	(5.3%)
title	3	(1.1%)	2	(0.8%)	5	(4.4%)
date	2	(0.8%)	2	(0.8%)	2	(1.8%)

Table 3: Types of metadata in the <meta> element, in frequency and proportion of the (X)HTML files. [1,p. 251] The discrepancy in the number of (X)HTML files at article level compared to Table 2 is due to inconsistencies in the study.

Small variations can be seen in the sample when it comes to the frequency of the various properties at different journal levels, but generally they show a similar pattern. The differences need to be treated with caution, as they are not statistically significant for the population at large. The generator property is slightly more common at article level in the sample, as is the case with author. That the author property would not be more common at article level is perhaps a bit surprising, as it is generally easier to identify the particular author(s) of an article than decide who should be listed in that position for the journal at large. The fact that the keyword and description properties are more common on the start pages than on the table of contents or article level could have to do with the fact that it is easy to enter the values to these properties once on the start page when creating the site, but requires certain routines if they are to be entered for each new table of contents page and article.

The qualitative studies, where more attention was placed on the values included in the <meta> elements, provide examples of how journals try to counter the fact that in the general use of the <meta> element properties one does not adhere to a specific vocabulary, by offering various versions of suitable keywords. Anticipated variations in how users will search for certain words with regard to number, spelling, and synonyms were met by including alternative keywords, e.g. university, universities – archaeology, archeology – and journal, periodical. Some journals also explore the fact that search engines can as easily search through post- as pre-coordination. They include quite unexpected phrases among the keywords, phrases that one would perhaps not expect potential readers to search for but where separate terms can still be retrieved.

In the fairly rare cases in the sample where the <meta> element is used to mark up a more regulated set of properties, namely those from the Dublin Core Metadata Element Set, the number of properties that are included is quite extensive, ranging from four to 14, with a median of 7 or 8 (depending on journal level). Between 3.8 and 11.4% of the journals contain Dublin Core metadata. In the sample, 18 journals were found to include this metadata type at the journal level, 17 at the issue level, and 20 at the article level. Only the Dublin Core properties that contained a value to the *content* attribute were included in the study. The practice of including subject (keywords) and description remain fairly strong at all levels, but even more commonly used are properties that may be easier to include (and in some cases to inherit from a template), such as DC.Type, DC.Format, and DC.Language. The Dublin Core elements are also used to indicate the originator to quite a large degree through such properties as DC.Creator, DC.Publisher, DC.Rights, and DC.Identifier. The only other property that occurs in more than 10 journals on at least one of the levels is DC.Title (cf. above).

So far, it is mainly the types of properties included in the <meta> that have been reported. However, as with the content of the <title> element, the <meta> elements are of little use if they do not contain values that may be used. For this reason, the quantitative study also included the various journal levels that the metadata describe. In order to discuss this, a distinction must be made between the level (journal/start page level, issue level, and article level) on which the file containing the <meta> element is placed and the level that the value of this <meta> element describes. I will refer to these as the levels where the metadata is placed and the level that the metadata describes.

The metadata (including Dublin Core elements) placed on the journals' start pages generally describe the journal at large. This is, however, also very often the case with metadata found at the issue and (to a smaller extent) article levels. When metadata at each of these levels does not (or not only) describe the journal level, it describes the level on which the metadata is placed. Thus, it is very rare for metadata placed on table of contents pages to describe individual articles, and for metadata placed in the article files to apply to the issue level. In fact, as can be seen in Figure 1, it is much more common at the issue level for the metadata to describe the journal than the issue. This further supports the hypothesis that metadata that can be entered once and continue to be valid, such as metadata describing the journal level, are more commonly included than metadata that needs to be updated for each new issue or article. The fact that some cases where found where the metadata had been copied from a previous issue or article without being changed indicates that when a new file is created based on a previous issue or article file, to change the marked up metadata could easily be forgotten. One of the journals in the qualitative study included quite a few <meta> and Dublin Core elements at its various levels. With a few exceptions at the article level, the values to each property were the same across the three levels, however. The metadata on this journal are thus site-specific rather than page-specific, which influences the granularity with which one can search for content from the journal.

Marked up metadata that are placed in a separate file are offered by 25 of the journals in the form of RSS feeds. This means that RSS files are available in between 5.6 and 13.6% (possibly as high as 17.8% at the article level) of the journals, at all three journal levels. 7 of these journals make use of a journal management

system (either PLONE or the Open Journal System), which has presumably made the inclusion of the feed easier. RSS feeds can provide marked up metadata that can be useful for various forms of reuse. Unlike the case with the <meta> element, the content of this metadata format is also more publicly visible, which could mean that the content is more carefully selected and entered.

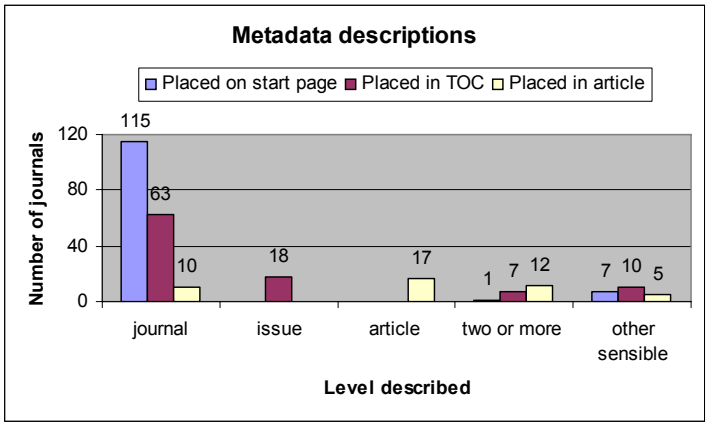

Figure 1: The levels of the journal described by the metadata (<meta> and Dublin Core) found in the files at the various journal levels, by number of journals. [1, p. 256]

4. Discussion and conclusions

Time is a valuable – and often scarce – resource for editorial staff of open access scholarly journals. A likely reason for the inconsequent use of marked up metadata that has come out as one of the results of this study is the lack of routines to follow when preparing an article for publication, both when a single person is responsible for the markup and design and when several people are involved. This results in great variations in what metadata are included in the various metadata elements as well as in how the metadata are notated and organized. The latter was shown to be the case in particular in the <title> element. As was illustrated in the qualitative studies, such variations occur not only between journals – where they are only to be expected – but also within journals and even within issues. Other problems that turned up in the study concern the reliability of metadata, such as when the values of the metadata elements are not updated when a new article file is created from an existing article or from a template. A certain lack of consistency was also found in one of the journals in the qualitative study that used frames. This raises the question of how to treat, and prioritize between, frames files when it comes to metadata. Thus, there are several potential problems with using existing metadata for various attempts at automatic collection of bibliographic data from the journals, even in the cases where there has been made an effort of including metadata elements. The great variety found in markup and metadata both between and within journals affects the possibilities for, for instance, libraries and other information services to retrieve data directly from the journal web sites in order to provide added value to the journals and their user communities.

At the same time, many of the journals do include metadata in the form of <title> and <meta> elements, even though only keywords and description can be said to be properties that occur reasonably often in the journals. Below, some thoughts are offered on considerations to keep in mind for individual journal publishers and the editor-managed journal community as a whole – preferably in co-operation with the library community – when trying to develop simple improvements in the form of documented routines or even more long-term guidelines for improving metadata inclusion in the journals. The ambition here has been that the development and performance of such routines should require little technological know-how. However, if more consistency and predictability is found in the marked up metadata of the editor-managed open access journals, it would be more worthwhile to develop services that offer access to the journals through various forms of collections and through bibliographic control. Such initial improvement of the

metadata should be seen as a step towards the use of more advanced metadata systems, such as OAI-PMH. On the way towards the use of such systems, documented routines or guidelines can be developed that take into consideration the following aspects that emerged from the present study:

What level to describe in the metadata elements at various journal levels. At the moment, metadata placed in the table of contents and article files quite often describe the journal as a whole rather than the content of that particular file. This is particularly common at the issue level. It is often of great importance to include information about the journal not only on the start page but also in the files at the issue and article levels in order to highlight the connection between, for instance, an article and the journal in which it has been published, but such metadata is preferably supplemented with metadata describing the content of the file in which the elements are included. In particular, many article metadata are often included on the web site, even if they are not marked up. This includes the name of the author(s), article title, abstract, keywords, and date of publishing. In fact, not surprisingly, the first article in the latest issue of every journal in the survey displayed the author names and article title in the article file. Abstracts were included in 78.9% of the journals, either in the article file, on a paratext page, or on the table of contents page. The corresponding figure for keywords was 40.4% and for author affiliation 86.8%. Another property that is easily obtainable for the journal staff is the date of publishing. This suggests that these metadata are in many cases available, they are simply not included among the marked up metadata in the files.

At what journal level to place metadata describing the article. The article file seems to be the obvious place for metadata describing the article. However, in cases when the article is published in a file format that does not easily incorporate metadata for retrieval, an option can be to introduce a paratext page, a page situated between the table of contents page and the article page. When this is done, the paratext page generally serves the purpose of providing bibliographic data about the article that can help the potential reader to determine if it is relevant to download the article – possibly a practice that open access journals have inherited from closed access journals, but where cost rather than download time needs to be considered. Yet, the paratext page can also contain marked up metadata which can serve to direct a user to the article page itself. Another consideration to take into account is how much metadata describing the articles in an issue to include on the table of contents page. This was very rarely done in the journals in the survey. Associated with the issue of where to place metadata describing the article is the question of:

How to treat web sites with frames. In journals that use frames for displaying the web site, there are generally several options of where to place metadata that describe the article. The content of the <title> element displaying in the web browser's title bar will be that of the frameset file. As this file is most likely the same for the entire web site, it is in most cases not a likely candidate for where to place article level metadata. A careful choice needs to be made as to where to place them, taking the design of the site into account.

What metadata properties to include. It is easy to be ambitious when planning for metadata elements but sometimes difficult to maintain those ambitions in the daily work. In these cases, it is probably better to keep the number of metadata properties down and aim to update them for each new issue or article. However, some metadata are likely to be constant from issue to issue and from article to article, mainly the ones that concern the journal level and more technical aspects, such as file format and encoding. If the markup is copied from one article to the next, such metadata can remain unchanged. Among the journals in the survey, keywords, description, and author were among the more common <meta> element properties to be included. Title and date were much less frequent. Keywords and description are fairly established properties, but if one wishes to include more properties, there could be reason to use the Dublin Core elements in order to achieve consistency in property names. The Dublin Core Metadata Element Set, still used very seldom among the open access journals, supplies a standardized set of properties that may be beneficial, including the possibility of qualifying such ambiguous properties as "date".

How to achieve consistency in the metadata element values. Related to the question of how to find consistency in the choice of name for various properties is that of achieving consistency in the element values. There are two dimensions of interest here: how to be consistent in the type of metadata that are included in an element vs. how to be consistent in the notation of the element value, and consistency within a journal vs. consistency across journals. That the issue of what type of content to include in an element is difficult is illustrated by the great variety found in the content of the <title> element. It is also pointed out by Roberson and Dawson [8, p. 68], that of the four journals they worked with, there were three different interpretations as to what should be the value of the DC.Relation property. Simple documentation of routines can help make both the type of content and its notation and organization more consistent across all new pages of a journal web site. If there is time to go over existing pages to align them with the guidelines outlined in the documentation, the web site as a whole will be more useful.

One of the greatest challenges is to achieve such consistency across a number of journals while keeping the work both technologically simple and time efficient. At the same time, cross-journal consistency is only interesting if the machine-readable metadata are used, that is, if there is some benefit to be had from consistency. This is where journal editors and librarians/information specialists can work together to add value to and support services that increase the findability of open access journals published by small publishers. To create basic guidelines for the inclusion of marked up metadata is one way to begin such collaboration, but as with all things that require some form of performance, there also needs to be a reward, a reason for putting in the work.

5. Notes and References

[1] This paper builds on data that were collected as part of my dissertation work, which was reported in FRANCKE, H. *(Re)creations of Scholarly Journals: Document and Information Architecture in Open Access Journals*. Borås. Sweden : Valfrid, 2008. Also available from: <http://hdl.handle.net/2320/1815/> [cited 10 May 2008].

[2] HAGLUND, L. *et al*. Unga forskares behov av informationssökning och IT-stöd [Young Scientists' Need of Information Seeking and IT Support] [online]. Stockholm, Sweden: Karolinska Institutet/BIBSAM, 2006. Available from: <http://www.kb.se/BIBSAM/bidrag/projbidr/avslutade/2006/unga_forskares_behov_slutrapport.pdf> [cited 19 April 2007].

[3] The term *scholarly* is used in this paper to cover contributions from both the scholarly, scientific, and technological communities.

[4] HEDLUND, T.; GUSTAFSSON, T.; BJÖRK, B.-C. The Open Access Scientific Journal: An Empirical Study. *Learned Publishing*. 2004, vol. 17, no. 3, pp. 199-209.

[5] KAUFMAN-WILLS GROUP. *The Facts about Open Access : A Study of the Financial and Non-financial Effects of Alternative Business Models for Scholarly Journals* [online]. The Association of Learned and Professional Society Publishers, 2005. Available from: <http://www.alpsp.org/ForceDownload.asp?id=70> [cited 24 April 2007].

[6] The Directory of Open Access Journals is provided by Lund University Libraries at <http://www.doaj.org/>.

[7] OAIster is provided by the University of Michigan at <http://www.oaister.org/>.

[8] ROBERTSON, R. J.; DAWSON, A. An Easy Option? OAI Static Repositories as a Method of Exposing Publishers' Metadata to the Wider Information Environment. In MARTENS, B; DOBREVA, M. *ELPUB2006 : Digital Spectrum : Integrating Technology and Culture – Proceedings of the 10th International Conference on Electronic Publishing held in Bansko, Bulgaria 14-16 June 2006* [online]. pp. 59-70. Available from: <http://elpub.scix.net/data/works/att/261_elpub2006.content.pdf> [cited 12 January 2008].

[9] JOWETT, G. H. The Relationship Between the Binomial and F Distributions. *The Statistician*. 1963, vol. 13, no. 1, pp. 55-57.

[10] ELENIUS, M. (2004). *Några metoder att bestämma konfidensintervall för en binomialproportion : en litteratur- och simuleringsstudie* [*Some Methods for Determining Confidence Intervalls for a Binomial Proportion : A Literature and Simulation Study*]. Göteborg, Sweden: Department of Economics and Statistics, Göteborg University. C-essay in Statistics.

[11] Open J-Gate is provided by Informatics India Ltd at <http://www.openj-gate.com/>.

[12] RAGGETT, D.; LE HORS, A.; JACOBS, I., Eds. *HTML 4.01 Specification : W3C Recommendation 24 December 1999* [online]. W3C (World Wide Web Consortium), 1999. Available from: <http://www.w3.org/TR/html4/> [cited 9 May 2008].