Publishing on the Internet: The Internet as an Academic Information Source

Henriette Günther Soerensen

Soebakkevej 23, 5210 Odense NV, Denmark

e-mail: farkhes@moes.hum.aau.dk

Kaj Fredsgaard Rasmussen

Spobjergvej 72, 1, 1, 1+2, 8220 Brabrand Denmark

e-mail: farkkfr@moes.hum.aau.dk

Abstract

In this paper we will look into publishing and the Internet from various viewpoints. The focal point is the sharing and retrieving of scholarly publications. This concerns both edited and non-edited publication, as well as the need for a connection between these and structured languages.

Automated indexing of the information on the Internet is the *way of making sure that we do not get lost on the World Wide Web. This again calls for structured languages in the form of content specifying tags. There is no way to get around the use of these structuring aids, if we wish to make sure that the information thus formatted will be readable and retrievable in the future.*

Key words: sharing information, structured languages, online publishing, retrieving information, scholarly publishing

1. Introduction

Publishing on the Internet is to some extent already established in most sciences, even though the majority of scientific research never gets this far.

The printed word, in the minds of most scientists, is still the optimum media. This is due to the fact that the printed word is the more established of the two, concerning referencing and credibility. Accessibility to the Internet speaks in favour of publishing there. The only precondition is the connection.

In Denmark a large part of the archaeological material is never published because of the cost and labour requirements of traditional publishing. To overcome this, publishing on the Internet is yet again the solution, due to its lower cost.

2.1. Sharing information

In the striving towards a universal understanding of the past, information technology can provide a wider perspective and a tighter communication network through global exchange of information. This was made possible by the accelerated evolution of the processing powers of computers in combination with the growth of the Internet, its increasing speed and bandwidth (this means that more information can be transferred simultaneously). Bandwidth is one of the push-factors in generating the structured languages we will discuss later on.

Today, processing power provides a secure base for combining information with several multimedia-formats, which provides new ways of presenting data in a more efficient manner. This is an important point in archaeology, as we often need to present illustrations of parts of our publications to make these more comprehensible, and the printed media may not be the cheapest way to achieve this.

Another of Internet's positive influences on information exchange is the fact that the information presented there is accessible to anybody, anywhere, at a low or at no cost. In archaeology this is often a great help in providing a global perspective.

Sharing information via the Internet makes it possible to work in smaller or bigger groups across borders. This potential is based on the constantly increasing computer- and network capabilities. Publishing on the Internet is much more complex than described here. A large part of this complexity lies in the exact difference in the cost between traditional and other various online publishing methods. This will not be elaborated in full extent in this context.

2.2. Structured languages

Structured languages are often conceived as more complex than the normally used *word processing programs*. Transfer to various structured languages is often possible in newer versions of these programs, and there exists a range of programs available specifically for editing these types of documents.

Because of the need to preserve the original categories of formatting structured languages have become a necessity in online scholarly publishing. This provides the possibility to recreate not only the visual appearance, but also to some extent the meta-data (information *about* the contents of the data).

Structured languages are often a practical solution to publishing on the Internet, as transmission speeds on the Internet limit the size of files for most users. Structured languages refer to formatting in an efficient manner, resulting in smaller files and as a consequence faster re-assembly of the document locally.

By far the most common mark-up language is the *HyperText Markup Language* (HTML) which originally uses a predefined general *Document Type Definition* (DTD). This trait is inherited from the older SGML (see below).

In the DTD for HTML provided by the World Wide Web Consortium (http://www.W3C.org, Burnard 1995:chapter 5.1) only the formatting is defined and not the content. This means that HTML only defines the visual appearance of the text (apart from one or

two meagre content defining formats like “address” (see http://hotwired.lycos.com/webmonkey, http://www.W3C.org). This visual formatting is not even interpreted in the same manner by various browsers. This is due to the fact that the formatting does not define all aspects of typography - line height, letter spacing and the like are not defined universally.

One remedy for the missing definitions of content and the lack of typographical design-tools in HTML has been, and still is, the use of Cascading Style Sheets (CSS) (Meyer 1999). These are created in order to provide the possibility of defining new formats and rules for their use. In using CSS you gain a far better control of the formatting, as well as a possibility of defining more meaningful mark-up tags, which can help the reading and indexing processes. These tags, if used, describe the content of the formatted text. An important point in this connection is that the use of meaningful tags is not obligatory. CSS are often “misused” for pure formatting reasons. Hereby HTML combined with CSS is a halfway clone towards the *eXtended Markup Language* (XML) described below.

The *Standard Generalized Mark-up Language* (defined before the advent of HTML) and the newer *eXtended Markup Language* are both defined specific and flexible (Burnard 1995). In these the user can create new DTD containing definitions of formatting, content types and rules for formatting with the defined syntax. Programs for formatting with the use of these languages are traditionally rather expensive (this applies especially to SGML tools) or difficult to obtain for other reasons.

Today we see more and more of the original HTML-editing tools extended with some XML-capabilities. The lack of programs supporting XML 1.0 as defined by the World Wide Web Consortium in 1998 (Bray et al. 1998: front page and page 1), is probably caused by the rather recent arrival of the final standard. Developers have simply not had enough time to provide full compatibility at a reasonable price.

In addition, the need for Document Type Definitions demands functionality far beyond the typical WYSIWYG structured language editor. This calls for either substantial restructuring of programs, or altogether new programs. These programs help in defining the syntax and the rules for the use of syntax: e.g. the order of statements, whether statements concern whole paragraphs or may exist inline within other paragraphs. This is slightly more complex than what the ordinary text editor can present in a comprehensible manner (Bünte 2000:200-213).

Conversion of SGML or XML formatted text is a new issue that arose from the extensive formatting possibilities in these two languages. Not all browsers, viewers, printers, WAP-phones or other viewing devices provide the same capabilities for presenting the defined formats. There are two solutions to this problem: Either the authors provide DTD’s for each conceivable kind of device or they provide a conversion routine for the required adjustments. The latter are provided by the eXtensible Stylesheet Language Transformations (XSLT) currently in development by the W3C. This new language defines the rules for conversion to the *Public Domain Format* (PDF), regular HTML –code or the like (Bünte 2000:202). Such automation is important in preserving the information encapsulated in various languages, thus providing compatibility in the future.

3.1. Online publishing

Publishing on the Internet results in lower costs compared to traditional publishing, and as we often need to illustrate archaeological finds and findings, the Internet might be a preferable media to the traditionally labour intensive, slow and expensive publishing.

Today we see two different ways of publishing on the Internet. One is based on the editing of the publications, the other on publishing non-edited texts. The truth is, of course, that no publications belong strictly to one of the two.

The tradition behind edited online publications is the printed word, and the content of such documents is often easier to understand, compared to the non-edited publications. The drawback is that these are labour intensive and consequently slower due to the editing and administrative phases.

In the end it is the user who pays the eventual cost of producing these publications. The high price often connected with scholarly publishing online by publishers outside the economic safety of institutions is likely to create a barrier between those who can, and those who can not afford access to these publications.

The bulk of the savings from the use of online publishing, as compared to printed publication in scientific circles, to a great extent lies in the process of printing, translating and distributing. To give an example let us look at the periodical “Journal of Danish Archaeology”: in this case the printing is prepaid by donations from foundations. Additional editing and layout is left to university employees as a part of their daily work.

As the publisher (the University) has no need of making a profit, migration of this kind of periodical to the Internet should be easily achieved, as it indeed should be the case for several university presses.

A completely different problem concerning edited publications is that the editing boards are by virtue of office in a position to control the free flow of information. They decide what is worth publishing and what is not. There exists a danger of this could to some extent lead to the monopolisation of information. On the other hand, the editing process insures a certain level of quality of the published material.

Editing on the Internet is a question of credibility and reliability. These can be used to give the reader of online documents a chance to verify the origin and the content. The editing board provides a security for the readers, that the documents are in the original form. Furthermore the editing can be expected to guarantee the quality of references and content.

In the case of non-edited publication there is no editing board to guarantee credibility and reliability. These documents can therefore be considered of a lower quality. The content is prone to contain various errors. These can occur as weak scientific argumentation, referential mistakes or bad syntax. As a result the reader must be more aware of, and critical towards these publications.

As the main problem of publications is credibility one would expect that a common consensus on formatting and content would be advantageous. For the edited publications the editing boards grant this. Since this is not the case in non-edited publications, the need for a way of providing it arises. One solution is to include a standard section in the documents. This should contain at least

the name of the author, how to reach her/him, the year of publication and last update/check-up.

3.2. Layout

Consistent layout of documents enhances the content of a given text. In other words: if the text is well formed, understanding the content is much easier, due to the fact that we are used to the structured layout of the printed media.

Editing of publications may not be the most suitable solution to our specific publishing needs, as this method results in slow transfer of results within scholarly circles.

However, the use of non-edited publication on the Internet is not satisfactory either. This concept provides no way of ensuring common standardised formatting. If this is not used, reading scholarly publications *can* become a challenge.

Some kind of swift editing should therefore be the best solution providing the best of both worlds (the high speed of communication, good quality of the published material).

In this connection, the use of typographical formatting in accordance with structured languages, such as HTML, SGML and XML, can be seen as of a general interest to everybody; both to the edited and the non-edited online publications. The other advantages gained from these languages (as described in "Structured languages") underline the importance of this kind of formatting.

3.3. Retrieving information

Publishing on the Internet presents problems connected to information retrieval. Some indexing methods are more precise than others, this forces the reader and/or author to make certain choices based on their knowledge of the use of indexing methods. The need to gain knowledge of these methods in order to successfully use them may become a barrier for some readers.

Searching for information without this basic knowledge of indexing methods can be very time-consuming and render the Internet highly impractical.

The Internet *is* a more or less a chaotic place and you have to know how to use search engines as well as the important keywords in the quest for the desired information. The use of imprecise keywords leads to a multitude of results with little or no relevance. Relevance is an important factor in dealing with all kinds of information retrieval on the Internet (Lester 1997:21-22).

Basically there are two ways of retrieving information:

- One is based on automatic indexing by the software. The index search returns a million answers, but only a few of them are relevant. This originates in the way these data are collected. The indexing program has no way of telling whether a document *really* belongs to one category or the other. The solution is to count the use of specific words and/or word combinations, providing the possibility of returning not only documents containing them, but also a way of ordering the results according to their relevance.
- The other way of retrieving information is by searching lists or databases of relevant links collected manually. Such lists are maintained through the division of links into subjects. Rejecting supposedly irrelevant categories through browsing can result in the user's missing out an important information. A link list requires much care by the Webmaster in order to keep it up-to-date. Links to newer are information most likely not present in these lists. As the Internet is a vast collection of resources it is unlikely that all the categories are complete.

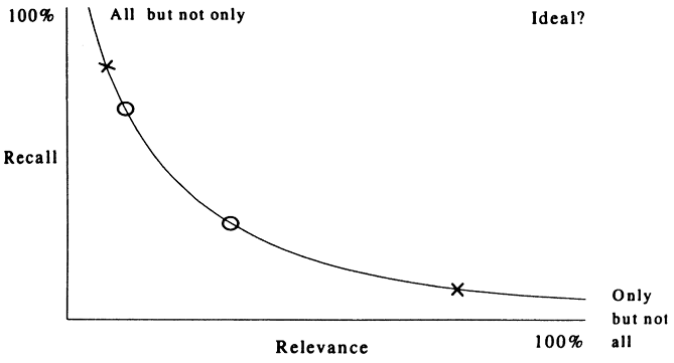

Figure 1: The recall-relevance / precision curve (Foskett 1996: fig. 2.3, p.19).

In this connection some type of stability in scholarly publishing would be advantageous, since this is the only way of ensuring that a link to the information placed here will still available after several years. This stability can be provided by the creation of online libraries maintained by public institutions, such as universities and libraries.

Databases on larger collections of links provide better possibilities in searching for cross-categories. This increases the chances of retrieving specific information on a given subject.

Index searching is most likely to retrieve all the relevant information and should be better than searching through collections of relevant links, even though it is time consuming.

From the aspect of retrieving information with the methods and problems discussed in this chapter, both the correct selection of keywords, and the two extremes of search strategies, results in the situation shown in figure 1. The recall-relevance/precision curve as described by Foskett (see Foskett 1996: figure 2.3) shows that using precise keywords and searching through link collections returns only relevant links, but not all the information available (the lower right corner of the graph). In the upper left corner the results of index searching and imprecise keywords are located. Here we retrieve all relevant information, but these are showered with irrelevant hits.

The real world is not divided as strictly into Indexing/Link collection and precise/imprecise keywords as presented here. Boundaries between these categories are loose, of course, and most search available engines can be placed somewhere between Indexing and link collection, like the keywords in most user requests can be placed in-between precise and imprecise.

4. Conclusions

Formatting text through structured languages is a condition of advanced automatic indexing of content. In scholarly publishing a shared *Document Type Definition*, when using *eXtended Markup Language* or *Standardized General Markup Language*, could be of a great additional advantage. Exchange of information between

scientists would thereby be simplified due to fewer differences in document formatting. Combined with standard sections this could be a strong base for online publishing. Moreover, automated indexing is markedly enhanced, leading to improved access through search-engines.

Today one of the more popular ways of finding information is at publishers' online sites. This is because the Internet is as a whole too imprecise, unstable and not so trustworthy. The result is that the edited archaeological publications on the Internet are preferred to the non-edited ones. This could lead to some kind of economic monopoly over information, instead of leaving resources freely accessible on the Internet.

The edited and freely accessible Internet Archaeology (http://intarch.ac.uk/) is the exception that proves the rule of the impossibility of combining editing, Internet and free accessibility.

Redirecting development towards closed communities on the Internet would mean redirecting the focus from publishing for a small public in the supposedly everlasting printed media to publishing for the masses on the dangerously ever-changing Internet. If the problems concerning storage and conversion of files online should be resolved, there is a good chance that publishing here instead of publishing in the traditional way will become the long-term and permanent solution!

However, taking into consideration the problems dealt with in this article, which refer to the Internet as it is today, one can say that in general the Internet provides vast possibilities for publishing. Cheap publishing combined with good accessibility and an easy user interface (taking precautions against the problems connected with this issue) makes the Internet an open resource with a promising future as a supplement to the traditional ways of publishing.

Today the possibilities of the Internet are not well exploited regarding publishing. Many resources are wasted because the Internet is ruled by chaos and finding one's way around can cost a lot of time and patience.

Some sort of stability of scholarly publishing placed online in the form of online libraries is therefore essential. Larger libraries and universities should consider this a common goal, as these institutions are in the best position to provide it.

Acknowledgements

We are in debt to the following people for the help in preparing this text: Torsten Madsen, Lilian Ahlmann Johansen, Kristine Stub Precht and Asbjoern Romvig Thomsen.

References

FOSKETT, A.C., 1996. *The Subject Approach to Information*. Library Association Publishing, London. Fifth edition.

BRAY, T., PAOLI, J., SPERBERG-McQUEEN, C.M., 1998. Extensible Markup Language (XML) 1.0; W3C Recommendation 10-Feb-98. At: http://www.w3.org/TR/1998/REC-xml-19980210.pdf

BURNARD, L., 1995. What is SGML and How Does It Help? *Computers and The Humanities*, vol 29: 41-50. Also at: http://sable.ox.ac.uk/ota/teiedw25/

BÜNTE, O., 2000. XML auf dem Vormarsch. In *CT, Magazin für Computer Technik,* 10/2000. Verlag Heinz Heise Gmbh & Co KG: 200-213.

LESTR, R., 1997. The Need to Add Value. Towards a Worldwide Library: A Ten Year *Forecast.* In Helal, A.H. and Weiss, J.W. (eds.), *Veröffentlichungen der Universitätsbibliothek Essen* 21. Essen 1997: 13-31.

MEYER, E., 1999. CSS: If Not Now, When? At: http://webreview.com/wr/pub/1999/06/25/style/index.html *Web Review*. Miller Freeman.

Documents available on the World Wide Web:

The use of HTML and CSS: http://hotwired.lycos.com/webmonkey/

Definitions of DTD's, HTML, CSS, XML and the like: http://www.W3C.org

The online magazine *Internet Archaeology*: http://intarch.ac.uk/