DOCUMENT RESUME

ED 411 858 IR 056 685

AUTHOR	Welsh, Sue
TITLE	OMNI--Alternative Approaches to Internet Metadata.
PUB DATE	1996-00-00
NOTE	8p.; In: Online Information 96. Proceedings of the International Online Information Meeting (20th, Olympia 2, London, England, United Kingdom, December 3-5, 1996); see IR 056 631.
PUB TYPE	Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Classification; Content Analysis; Evaluation Methods; Foreign Countries; Futures (of Society); Indexes; *Indexing; Information Needs; *Information Retrieval; *Internet; Library Technical Processes; *Online Searching; Online Systems; Reference Services; Relevance (Information Retrieval); Search Strategies; Technological Advancement; User Needs (Information); User Satisfaction (Information)
IDENTIFIERS	*Metadata; United Kingdom

ABSTRACT

The growth in the size of the Internet has resulted in much effort being spent on indexing its contents. The most popular solutions are created by automatic methods, and although offering impressive coverage, they are disappointing where precision of meaning is required. Alternative services created by human beings arrange and index resources according to concept and offer more relevant retrieval, but cannot hope to achieve 100% coverage. New developments are imminent which may support both the automated and non-automated approaches, and vastly improve the quality of Internet metadata. This paper demonstrates that the task of locating information on the Internet has not been and will not be accomplished by the use of sheer computing power alone. It reviews the topic of metadata creation, with special reference to the OMNI project (Organizing Medical Networked Information) and describes two encouraging new initiatives: PICS (Platform for Internet Content Selection) and the Dublin Core Metadata Set/Warwick Framework. All users need improved search tools with which to navigate the Internet. The answers to the present dilemma will be based on more than one tool, on both human intervention and intelligent automated data gathering. Metadata, as exemplified by the old catalog card, is set to become a key concept in the future of the new information world. (Contains 19 references.) (Author/SWC)

OMNI — alternative approaches to Internet metadata

Sue Welsh
The OMNI Project, National Institute for Medical Research, UK

Abstract: *The growth in size of the Internet has resulted in much effort being spent on indexing its contents. The most popular solutions are created by automatic methods and although offering impressive coverage they disappoint where precision of meaning is required. Alternative services created by human beings arrange and index resources according to concept and offer more relevant retrieval, but cannot hope to achieve 100% coverage. New developments are imminent which may support both the automated and non-automated approaches, and vastly improve the quality of Internet metadata.*

1. Introduction

The quality of information on the Internet is today as much a talking point as the Internet itself. The World Wide Web (WWW) in particular has given everyone a chance to publish, and finding what you need in the resulting free-for-all is becoming increasingly difficult.

Why should this be when there is no shortage of tools designed for this task? The most comprehensive solutions are, by necessity, computer generated. This total approach to Internet indexing as practised by Alta Vista (Ref 1) and Lycos (Ref 2) is in sharp contrast to the subject-based information gateways (SBIGs) such as OMNI (Ref 3), SOSIG (Ref 4) and EEVL (Ref 5), and electronic communities such as PharmWeb (Ref 6) and BioMedNet (Ref 7). In between these two extremes, indexes and directories abound. Why aren't we being well served with all this choice?

In an article entitled 'Seek and Ye Shall Find (Maybe)' in the Internet style magazine *Wired*, Steve Steinburg joked

> '... information retrieval is really only a problem for people in library science. If some computer scientists were to put their heads together, they'd probably have it solved before lunchtime' (Ref 8).

but the problem is clearly still with us despite occupying some of the Internet's brightest minds for several years.

This paper will seek to demonstrate that the task of locating information on the Internet has not been and will not be accomplished by the use of sheer computing power alone. It will review the topic of metadata creation with special reference to the OMNI project and look forward to two encouraging new initiatives: PICS and the Dublin Core/Warwick Framework.

2. Metadata, what is it and why is it necessary?

2.1. What is metadata?

Metadata is data about data. Its most familiar form is the descriptive catalogue record used to describe printed books, as defined by AACR2 (the Anglo American Cataloguing Rules) and expressed in MARC format.

2.2. Metadata formats

In contrast to the very well defined standards in existence for printed material, there is no standard form of metadata for Internet objects. Projects creating it have used various formats and some have created their own: for example NISS (Networked Information Systems and Services) (Ref 9) developed the NISS template for its gateway to networked resources (Ref 10).

OCLC has used MARC-like records in NetFirst (Ref 11) which has recently been made available to the academic community in the United Kingdom by CHEST.

Metadata can range far wider than the traditional descriptive catalogue record; administrative data (e.g. pricing, administrative contacts, authentication instructions) and content ratings (e.g. stars or marks out of ten) are two such examples.

2.3. Automated metadata

Both NISS and OCLC create rich metadata records by relying on human input. Automated indexes also create metadata but are far more limited in this respect as:

- they rely on the internal structure of the HTML (HyperText Markup Language) document;
- they cannot glean any information from non-textual objects such as images or sound;
- they are hampered by poor practice from HTML authors, e.g. lack of <title></title> in the head of the document;
- they can be fooled by simple tricks such as word spamming (the practice of including multiple repetitions of a keyword in a HTML document in such a way that is invisible to the person viewing with a standard Web browser but picked up by crawlers/robots. It has been used by unscrupulous Web authors to ensure their documents float to the top of a ranked list of resources).

Even putting these problems aside, retrieval from automated indexes suffers because they index the full text and not concepts (but see Excite later.) This means that false drops due to homonyms (a single word with multiple meanings) are common and reasonably comprehensive searching is likely to require knowledge of too many synonyms (multiple words with the same meaning) to be possible. This makes automated search engines good for searching for a very specific needle in a large haystack, but poor for searches on topics with any breadth.

2.4. Automated concept indexing

Can concept indexing be automated? The makers of Excite (Ref 12), one of the more innovative search engines would like us to think so. 'Excite ... has the intelligence to search through reams of information and bring you the good stuff,' claims the publicity, but in fact there is no intelligence at work here, just more computing power applied with flair. Using its full text index, Excite clusters documents that have similar profiles (many keywords in common) and creates a classification scheme as a result. How this affects the quality of results returned probably depends on the type of search but false drops and inexplicable hits seem to occur with Excite as with the other automated indexes. At the present time, content indexing is firmly in the hands of the AI researcher.

In a nutshell, automated indexes fail to satisfy all our search requirements because they use processing might to take the place of a decent index or high quality metadata.

3. Creating metadata, a case study

How can we create a better index? If robots are not good at creating metadata, can it be created for the Web by hand? A group of projects under the eLib (Ref 13) umbrella are investigating a different model. A case study of the OMNI project follows, illustrating the processes and problems involved.

3.1. Background to the OMNI project

The OMNI project is part of the Access to Networked Resources group (ANR) of the Electronic Libraries Programme (eLib). OMNI is compiling a gateway to high quality biomedical information available via the Internet and also produces a comprehensive listing of UK sites in this area. Funded by the UK's Higher Education Funding Councils for a two year period, OMNI began work in July 1997 and launched the first edition of its service in November 1995.

3.2. The OMNI philosophy on metadata

We have seen that the most widely used search and retrieval tools are unsuited to certain types of search query because of the way in which their metadata is constructed. This is particularly true of subject-based inquiries. They also offer no quality control over resources included: indeed it is their aim to include everything. The challenge faced by OMNI was to seek an alternative approach, which must differ from what existed previously by:

(1) being tailored to the requirements of the research and academic community;

(2) offering access to resources in one subject area, biomedicine;

(3) selecting on the basis of quality;

(4) providing metadata that would effectively guide users to the right resource, eliminating the need for directionless 'surfing';

(5) paying special attention to the needs of the UK community.

Each of these five key issues was also a point against the use of the traditional automated methods. Neither restricted audience focus, subject requirements or the effect of geographical location on user needs can be expressed in such a way that makes it possible to program a robot. While it is possible to direct an automated process towards resources in the academic/research domains (e.g. *.ac.uk* or *.edu*) or the UK domain alone (*.co.uk*, *.ac.uk*, *.gov.uk* etc.) this is an oversimplification of the requirements expressed above. Serving the information needs of the UK community is not the same as giving access to all that that community publishes on the Web, neither does collecting together the combined outputs of the various educational domains provide all that

that community may require.

Quality is also in the eye of the beholder. There is no way to assess it other than establishing criteria which must be judged by a real person: a robot cannot judge content, usability, intended audience and so forth.

Finally the kind of metadata that is required was not available, and is not yet present in large quantities, to be picked up from the Internet itself. Robots can only pick up what is there. Compare and contrast the OMNI metadata for the home page of the Center for Complex Systems Research, at the Beckman Institute for Advanced Science and Technology (University of Illinois at Urbana-Champaign), with the automatically created entry from a major search engine (Figure 1). A rather extreme example which serves to illustrate many of the problems with automatically gathered metadata!

OMNI metadata (selection)

Title: Center for Complex Systems Research, University of Illinois at Urbana-Champaign.

Description: The home page of the Center for Complex Systems Research (CCSR), at the Beckman Institute for Advanced Science and Technology (University of Illinois at Urbana-Champaign). The CCSR is an interdisciplinary group investigating a variety of complex dynamic processes including biological intelligence (endodynamics, neural networks and metabolic networks). This site offers access to background information about the Center and the full text of technical reports generated by the group.

Keywords: neural networks (computer); acclimatization; metabolism

Accessible by: http://www.ccsr.uiuc.edu/

Automated metadata

No title

This page has moved to here.

http://www.ccsr.uiuc.edu/CCSRHome.html - *size 132 bytes - 27 Mar 96*

Figure 1: Internet metadata.

Having decided that human intervention was essential it remained to choose a metadata format to work with. There is no standard for Internet metadata, although there are several potential candidates which may become the standard of the future. One of these is the IAFA (Internet Anonymous File Archive) template, which was originally proposed as a method of describing the contents of FTP servers and now includes formats for a variety of other Internet objects.

3.3. ROADS and IAFA

In common with the majority of ANR projects including the established and well regarded SOSIG (social science) service, OMNI chose to use the ROADS (Ref 14) database software. ROADS (Resource Organisation and Discovery in Subject-Based Services) is another eLib project developing a user-oriented resource discovery system. ROADS is based around the IAFA (Internet Anonymous FTP Archive) metadata format (Ref 15), illustrated in Figure 2. IAFA records consist of a list of attribute:value pairs and exist for various types of Internet objects, software, services, documents, images etc. Although ROADS is committed to developing IAFA it recognises that other formats may become more important during the lifetime of the project and interoperability must be ensured.

OMNI is currently using the first version of the ROADS software which allows creation, editing, searching and browsing of IAFA records. Future ROADS developments will facilitate simultaneous access to distributed ROADS databases using the WHOIS++ directory service protocol.

```
Template-Type: DOCUMENT
Handle: 814014233-2699
Category:
Title: Medical/Clinical/Occupational Toxicology Professional Groups
URI-v1: http://www.pitt.edu/~martint/pages/motoxorg.htm
URI-v2: http://www.pitt.edu/~martint/pages/motoxorg.htm
Admin-Name-v1: Thomas G. Martin
Admin-Work-Postal-v1:
Admin-Job-Title-v1:
Admin-Department-v1:
Admin-Email-v1: TGM@MED.PITT.EDU
Admin-Handle-v1:
Admin—v1:
Description: A list of (mostly US) professional groups of interest to toxicologists. A short description of each organi-
        sation is given, as well as addresses, phone numbers and contact names.
Citation:
Publisher-Handle-v1:
Publisher-Name-v1: Toxicology Treatment Program
Publisher-Postal-v1: University of Pittsburgh Medical Center,
Pittsburgh, Pennsylvania, USA
Publisher-City-v1:
Publisher-State-v1:
Publisher-Country-v1:
Publisher-Email-v1:
Publisher-Phone-v1:
Publisher-Fax-v1:
Keywords: toxicology; poison control centers
Version-v1:
Version-v2:
Format-v1:
Format-v2:
Size-v1:
Size-v2:
Language-v1:
Language-v2:
Subject-Descriptor-v1: 615.9
Subject-Descriptor-v2: QV600
Subject-Descriptor-Scheme-v1: UDC
Subject-Descriptor-Scheme-v2: NLM
Destination: OMNIworld
Record-Last-Modified-Date: Thu, 18 Oct 1995 11:06:31 +0000
Record-Last-Modified-Email: unknown@nimp377.nimr.mrc.ac.uk
Record-Created-Date: Thu, 18 Oct 1995 11:03:53 +0000
Record-Created-Email: unknown@nimp377.nimr.mrc.ac.uk
```

Figure 2: An IAFA record from the OMNI database.

3.4. Creating quality metadata

All OMNI metadata is created by hand and occupies three members of staff and a band of over 80 volunteers from the bioscience community. By the end of its initial period of funding the project aims to have 2000 completed records. Not all the IAFA attributes are used and only four are displayed to the public (title, description, keywords and URL) at present.

A record entering the OMNI database follows the following path:

- discovery e.g. announcement via e-mail or newsgroup, serendipitous discovery or as part of organised scanning of major sites;
- evaluation according to OMNI's guidelines, which are available for public scrutiny and comment;
- description: either a partial record is composed by an OMNI volunteer and completed and checked by OMNI staff or the entire record is authored by OMNI. Descriptions include a summary of the content of the object, MESH (Medical Subject Headings) indexing and classification by NLM (National Library of Medicine) and UDC (Universal Decimal Classification);
- dissemination: OMNI users are alerted periodically that new resources are available by e-mail and via the WWW.

This is clearly a time consuming process and OMNI is a service at the opposite end of the scale to Alta Vista and Lycos. The subject-based information gateways in general are characterised by:

- small, well defined areas of subject interest;
- selection of resources based on quality;
- added value in the form of description, indexing and classification.

Only time will tell if this approach can fill the gaps left by the automated search engines. Meanwhile, some exciting developments might make the task of providing both types of service easier.

4. Future developments

Human intervention results in better metadata but is time consuming. The amount of information flooding into the public domain via the Internet is so great that services like OMNI can never hope to create rich metadata records for every document within their subject area: they must select using some set of criteria (in OMNI's case selection is made following assessment of the quality of the object according to our own quality criteria.) Selection itself, though, takes more time. How can this whole process be speeded and what can we do about the resources which will never be considered important enough to catalogued in this way?

4.1. The Dublin Core

A key initiative attempting to answer these questions has put forward suggestions for author-generated metadata. The Dublin Core Metadata Element Set, proposed at a meeting in Dublin, Ohio in 1995 (Ref 16) and consolidated at a second conference in Warwick, UK (Ref 17) is a set of 13 key metadata elements for describing document-like objects on the Internet (see Figure 3.)

Subject: The topic addressed by the object

Title: The name of the object

Author: The person(s) primarily responsible for the intellectual content of the object

Publisher: The agent or agency responsible for making the object available

Other Agent: The person(s) such as editors and transcribers, who have made other significant intellectual contribution to the work

Date: The date of publication

Object Type: The genre of the object, such as novel, poem or dictionary

Form: The data representation of the object, such as Postscript file

Identifier: String or number used to uniquely identify the object

Relation: Relationship to other objects

Source: Objects, either print or electronic, from which this object is derived.

Language: the Language of the intellectual content

Coverage: The spatial locations and temporal duration characteristic of the object.

Figure 3: The Dublin Core, taken from http://cs-tr.cs.cornell.edu/~lagoze/warwick/warwick.doc.

The Ohio meeting, which first proposed the 13 elements, was organised by OCLC and NCSA (the National Center for Supercomputing Applications). The Warwick Conference (sponsored by UKOLN, the UK Office for Library Networking) took this simple idea forward by suggesting a mechanism for carriage of metadata (which may be Dublin Core or some other metadata set). Of most immediate interest is a specified special case where metadata is embedded in the HTML document it describes. Put simply, this offers Web authors the chance to catalogue their own material in a standard format. It will then be a simple matter for a robot to pick up this metadata and pass it onto an index. Automated indexes would be able to gather in better metadata and authors achieve better visibility for their work; selective sites can automate some proportion of their metadata creation; everyone benefits.

The Dublin Core will have a major impact on Web indexing only if authors and publishers can be persuaded to use it, and there are proposals to incorporate a mechanism for easy addition of the required data into common HTML authoring tools. The accuracy of author generated metadata is another problem altogether and it is likely we will continue to see services like the subject-based information gateways employing editors to check the veracity of the data being harvested.

4.2. The PICS standard

The PICS (Platform for Internet Content Selection) (Ref 18) standard is designed to facilitate provision of a specific type of metadata — content rating. The idea of scoring or ranking Internet sites by some agreed scheme is not new. So called evaluative search services such as Magellan (Ref 19) are already doing just that. Magellan reviewers award sites a marks out of ten for depth, ease of exploration and 'net appeal', combining these scores to produce an overall star rating of one to five stars. When you browse or search using Magellan your results will be returned with star ratings.

Evaluative services seem at first to be an excellent way of finding good material but are aimed at a popular, general audience (the concept of 'net appeal' is not useful to the average clinician, for example). Their major problem lies with the breadth of the material on the Web and the variety of people using it. How well does a single set of rating criteria deal with the range of information available? How can a single reviewer award a rating that makes sense to everyone from Star Trek enthusiasts to molecular biologists? Surely we need different ratings for different subject areas and communities? Again we return to a key characteristic of this information we call metadata. Each community has a clear and different idea of what metadata is and each community is right: your idea of useful metadata depends on the data you use and what use you make of it.

The avowed purpose of the PICS standard is to stem the rising tide of concern amongst parents that their children are accessing unsuitable material via the WWW. The idea is simple: PICS compliant browsers will point at Web resources via a ratings label. The label may be published by a trusted third party providing ratings information for a particular grouping in the form of a database, or information providers may label their own resources. Parents will be able to set up their browser to deny access to material without a suitable rating. Parents with different attitudes to 'offensive' material will be able to point at different trusted third parties or deny access to different levels of material. PICS has been described by its proponents as value-neutral, as it enables the implementation of context-specific access rather than global restrictions which they believe would destroy the freedom of expression that makes the Internet so popular.

PICS technology is being developed by the World Wide Web Consortium and (perhaps even more significantly) has won the support of industry leaders such as Netscape, Microsoft, AOL (America Online) and CompuServe. At the Fifth International WWW Conference in Paris this year a PICS compliant version of Microsoft's Internet Explorer was demonstrated and other browsers will not be far behind. It is less clear who will provide labelling services, but there are clear commercial opportunities and some sites are already offering PICS compatible rating systems.

5. Conclusion

Publishers on the WWW have a vested interest in making sure their site is visited. This may be because they have information there that they need people to see, or because they earn money though advertising and sponsorship (dependent on access statistics), or because they wish to sell you something themselves when you arrive.

Librarians and other information workers face a future where they must coordinate the old and new information domains and guide their users, as they have in the past, to information which is timely, accurate and conveniently located and priced.

Users, if they don't already, will want to use the Internet to save time rather than (as is now only too often the case) waste it.

All of us need improved search tools with which to navigate the Internet. I believe that the answers to our present dilemma will be based on more than one tool, on both human intervention and intelligent automated data gathering. Metadata, as exemplified by the fusty old catalogue card, is set to become a key concept in the future of the new information world.

Sue Welsh
OMNI Project Officer
National Institute for Medical Research
The Ridgeway
Mill Hill
London NW7 1AA
UK
E-mail: swelsh@nimr.mrc.ac.uk

References

[1] *Alta Vista Search* [online]. Digital Equipment Corporation. Available from http://www.altavista.digital.com/ [accessed 1 September 1996].

[2] *Lycos Search* [online]. Lycos, Inc. Available from http://www.lycos.com/ [accessed 1 September 1996].

[3] *OMNI: Organising Medical Networked Information* [online]. OMNI Consortium. Available from http://omni.ac.uk/ [accessed 1 September 1996].

[4] *Social Science Information Gateway — SOSIG* [online]. Bristol University: Centre for Computing in the Social Sciences. Available from http://sosig.ac.uk/ [accessed 1 September 1996].

[5] *Edinburgh Engineering Virtual Library (EEVL)* [online]. EEVL Consortium. Available from http://eevl.ac.uk/ [accessed 1 September 1996].

[6] *PharmWeb* [online]. Manchester University. Available from http://www.mcc.ac.uk/pharmweb/ [accessed 1 September 1996].

[7] *BioMedNet* [online]. London: Electronic Press. Available from http://www.biomednet.com/ [accessed 1 September 1996].

[8] Steinburg, S. (1996) Seek and ye shall find (maybe), *Wired*, May.

[9] *NISS Information Gateway* [online]. University of Bath: NISS (National Information Services and Systems). Available from http://www.niss.ac.uk/ [accessed 1 September 1996].

[10] *NISS Directory of Networked Resources* [online]. University of Bath: NISS (National Information Services and Systems). Available from http://www.niss.ac.uk/subject/index.html [accessed 1 September 1996].

[11] *NetFirst* [online]. OCLC Online Computer Library Center, Inc. Available from http://www.netfirst.ac.uk/ [accessed 1 September 1996].

[12] *Excite* [online]. Available from http://www.excite.com/ [accessed 25 July 1996].

[13] *eLib: The Electronic Libraries Programme* [online]. University of Bath: eLib. Available from http://ukoln.bath.ac.uk/elib/intro.html [accessed 1 September 1996].

[14] *ROADS, Resource Organisation And Discovery in Subject-Based Services* [online]. Loughborough University of Technology: Department of Computer Studies. Available from http://www.roads.lut.ac.uk/ [accessed 1 September 1996].

[15] Deutsch, P. *et al.* (1995) Publishing Information on the Internet with Anonymous FTP [online]. Available from http://info.webcrawler.com/mak/projects/iafa/iafa.txt [accessed 1 September 1996].

[16] Weibel, S. (1995) Metadata: the foundations of resource description, *D-Lib Magazine* [online], 7. Available from http://ukoln.bath.ac.uk/dlib/dlib/July95/07weibel.html [accessed 1 September 1996].

[17] Lagoze, C. *et al.* (1996) The Warwick framework: a container architecture for aggregating sets of metadata. Available from http://cs-tr.cs.cornell.edu/~lagoze/warwick/warwick.doc [accessed 1 September 1996].

[18] *Platform for Internet* Content [online]. World Wide Web Consortium. Available from http://www.w3.org/PICS/ [accessed 1 September 1996].

[19] *Magellan Search* [online]. McKinley. Available from http://www.mckinley.com/ [accessed 1 September 1996].