DO INDEXING AND ABSTRACTING HAVE A FUTURE?*

F. W. Lancaster

Professor Emeritus. Graduate School of Library and Information Science. University of Illinois at Urbana-Champaign.

Abstract: The World Wide Web presents information access problems orders of magnitude greater than any encountered before. These problems are identified and discussed. Although the intellectual indexing of the entire Web is impossible and, in any case, not worthwhile, selective indexing of high value resources (e.g., by means of specialized portals) is essential. The summarization of Web resources (i.e., abstracting) is also important although acceptable abstracts for some purposes may be automatically constructed. Despite advances in automatic procedures, there will probably be a need for indexing and abstracting skills for a very long time, although this work may take on somewhat different forms.

Keywords: Indexing; abstracting; World Wide Web; portals; classification; intranets.

Título: ¿TIENE FUTURO LA INDIZACIÓN Y EL RESUMEN?

Resumen: El *World Wide Web* presenta problemas de acceso a la información de una orden de magnitud mayor que cualquiera de los encontrados anteriormente. Se identifican y discuten estos problemas. Aunque es imposible la indización intelectual de todo el Web y, en todo caso, no merece la pena, es esencial la indización selectiva de recursos de alto valor (e. g., mediante portales especializados). También es importante el resumen de recursos Web aunque para algunos propósitos pueden construirse automáticamente resúmenes aceptables. A pesar de los avances en procedimientos automáticos, habrá probablemente una necesidad de habilidades de indización y resumen durante mucho tiempo, aunque este trabajo podría llevarse a cabo de diversas formas.

Palabras clave: Indización; resumen; *World Wide Web;* portales; clasificación; intranets.

Writing almost 50 years ago, Fairthorne (1958) claimed that "Indexing is the basic problem as well as the costliest bottleneck of information retrieval." Indexing is still the central problem of information access and Fairthorne's mind would surely have boggled at the immensity of the information access problems posed by the World Wide Web.

The major defect of the Internet as an information source, apart from its sheer size, is the fact that it lacks any form of quality control. That information services work reasonably effectively in the world of print on paper is due to the fact that various institutions are in place to perform a quality filtering function. Publishers of scholarly books and journals apply reviewing/refereeing procedures that are, at least to some extent, effective in eliminating the most worthless of what is written. The published indexing and abstracting services provide the next level of quality filtering, mostly by choosing the journals, report series, or other publications that they cover on a regular basis. Finally, libraries, particu-

* This article is based largely on a chapter in the third edition of the author's *Indexing and Abstracting in Theory and Practice*, to be published in 2003 by the Graduate School of Library and Information Science of the University of Illinois and by Facet Publishing, London.

larly those serving the research and scholarly communities, put into place the filters closest to the actual users by purchasing those materials judged of most value to these users and by arranging collections by levels of accessibility, the most accessible materials (physically and perhaps also intellectually) being those that users will be most likely to want frequently.

It is clear that the vast expanse of poorly organized resources that are accessible, at least in a theoretical sense, through the Internet make the construction of effective filters a daunting proposition, whether at individual or institutional levels. Moreover, we are assured that the situation will get much worse.

At the present time, it does not seem likely that the current chaotic situation caused by the "every man his own publisher" phenomenon is likely to be reversible. In other words, it is hard to visualize the possibility that governments could or would impose overall quality standards on network publication or distribution. Consequently, the viability of a vast network as an information resource must depend upon the imposition of quality filters similar to those of the print-on-paper world.

There is no doubt that the filtering function is just as important in the electronic environment as it was in a publishing environment dominated by print on paper. Because indexing and abstracting, in one form or another, are essential elements in information filtering, it follows that they must have a future. The only remaining questions, then, are:

1. What form will these activities take, and
2. Who will or should perform them?

It is interesting to note that Odlyzko, who for some years has predicted that both libraries and scholarly journals will become obsolete, at least in their traditional forms (see, for example, Odlyzko, 1995), is quite positive on the future of indexing and abstracting services. He claims (Odlyzko, 1999) that such services will survive because they make a substantial intellectual contribution and are comparatively inexpensive relative to that contribution.

Jacsó (2002) disagrees somewhat about the services, but he remains a firm believer in the need for abstracts within the Web:

> *The increasing availability of full-text databases has decreased the importance of A&I databases in the past 10 to 15 years, but not the need for abstracts. Full-text databases require abstracts for efficient use. The obvious reason for this is that scanning the search-results lists with short abstracts helps tremendously to select the most promising source documents, even when the abstracts leave much to be desired.*

He goes on to say:

> *The less obvious reason for having abstracts with these databases is that limiting a search in a full-text database to the abstract field is guaranteed to make it more precise than searching hundreds of thousands of full-text documents (Page 22).*

Of course, Jacsó is not necessarily referring to humanly prepared abstracts but to abstracts or extracts that are automatically prepared. In fact, his article reviews commercially available software designed for "document summarization".

Proposals concerning indexing of Web resources cover an extremely wide range, including claims that it is not possible at all. For example, Wellisch (1994) claimed that

"Electronic journals are unlikely to be indexed because of the instability of their texts".Since most sources on the Internet are much less stable than the journals, he presumably feels that the whole enterprise--i.e., indexing text that is subject to frequent change--is a lost cause.

It is obvious that human professional indexing of the entire Web is completely impractical. Even if it were, much of what appears on the Web is too impermanent or of too low quality to merit such indexing attention. Selective professional indexing, of course, is possible. Owen (1994) and Weinberg (1996) are two writers who have advocated professional indexing on a selective basis. Weinberg specifically recommended back-of-the-book type of indexing and this kind of indexing could certainly be applied to individual Web sites. Indeed, it has already been applied in this way, and Browne (2001) has discussed and illustrated the processes. Casey (1999) recognizes that her dream of a full "analytical index" to the Web (i.e., one that indexes below the level of the Web site) is utopian and that "small, focused indexes may be the best solution".

Ellis et al. (1998) suggest that a major problem in any approach to Web indexing is the fact that the indexer will always be very remote from the user:

> *. . . [in] the World Wide Web ... there is no closeness at all between designer or creator (which could be anyone) and potential user (which could be anyone or everyone). This is compounded by the lack of any clear understanding on the part of most searchers as to what it is the various search engines are actually doing when they search. So that the real source of problems in searching distributed online or Internet sources arises not from technical indexing problems but from the easy access provided by online services and the World Wide Web to information selected, structured and indexed for one group of users (with one set of characteristics and information requirements) by quite different sorts of users with quite different characteristics and requirements.*
>
> *This may be expected to exacerbate existing problems of indexer-user concept matching as users encounter many different files or sites, with differing characteristics, indexing practices and vocabularies, none of which can be expected to meet all, or even some, of the needs of any potential user or user group. This is a key issue, for the more distant users are, in characteristics and information needs, from the types of user conceived of and catered for by those creating or indexing a database, the more likely there are to be problems in accessing relevant information by users from that database. The problem is that of indexing for the unknown user (Page 44).*

Another obvious problem in indexing is the fact that some Web documents are "virtual"--documents "for which no persistent state exists"(they are created on the way to the user) (Watters, 1999).

PROFESSIONAL APPROACHES

Two major approaches to providing intellectual access to the more important Web resources are already in place. A major initiative for applying metadata to Web resources was established as CORC (Cooperative Online Resource Catalog), a joint program of OCLC and a cadre of participating libraries. In 2002, the program was renamed "Connexion".Participating libraries select the Web resources they believe to be most valuable and then catalog them, including Dewey Decimal Classification numbers to provide subject access. As of October 2002, around 700,000 records had been contributed by about 500 institutions.

The other approach is the "portal". Special libraries and information centers can provide an important service by identifying those Web resources of greatest relevance and value to their users, indexing those resources in some way, and developing a gateway that provides access to these resources through the metadata elements. A number of such gateways or portals are described and illustrated in Wells et al. (1999), who refer to them as "virtual libraries". Campbell (2000) has described his vision of a more general "scholars portal", and individual libraries can design and implement their own portals to Web resources (see, for example, Hurt and Potter, 2001, dealing with a general academic library, and Medeiros et al., 2001, dealing with an academic medical library). The importance of this type of activity for the public library was highlighted by Holt (1995) as follows:

> *. . . public library staff can save time for their constituents by organizing the mass of electronic information available on local, national, and international servers . . . [and] can develop electronic guides to help searchers through the metadata and megafiles with which they must deal online (Pages 555-556).*

He specifically mentions the importance of providing annotations for users, and sees the public library as an information clearinghouse staffed with "information agents".

All of these activities relate to the filtering of Web resources and they all imply some form of subject access provision through indexing or classification, and perhaps some form of abstracting. Trippe (2001) stresses the need for more classification of Web resources and Elrod (2000) summarizes an online discussion on the desirability of libraries assigning class numbers to the Web resources they provide access to (some already do).

ALTERNATIVE APPROACHES

Drott (2002) proposes a completely different solution. He has highlighted the Web indexing problem, as follows:

> *Finding information on specific topics on the web is hard and getting harder. New advances in automated web searching and algorithmic indexing have been largely offset by the enormous growth in the amount of material available. The estimates of search engine coverage of the web by Lawrence and Giles (1999) suggest the impossibility of using robots to index all of the web, and clearly, the more analytical time that a robot must devote to extracting index terms for a single page, the smaller the amount of the available material that can be indexed. Further, while strides are being made in improving the accuracy of automatic indexing, the fact remains that assigning index terms to a database as diverse as the web remains a problem with few promising solutions (Pages 209-210).*

He goes on to suggest, however, that, while the use of professional indexers may not be an economically attractive proposition, those responsible for creating Web pages should be able to do an acceptable job of indexing themselves:

> *Would encouraging web site creators to assign their own index terms be a good thing? The current model of indexing, such as that found in the major journal indexing services, is based on the use of skilled indexers with extensive training. There is, however, encouraging research by Coombs (1998) on indexing State Government web pages in Washington State. Coombs used the people who created and worked with the documents as indexers. The results of this study showed that, when the lay indexers share a common understanding of the content and uses of their documents, the keywords which they produce are reasonable subject location aids (Page 218).*

And, finally:

> *Our model of web indexing may well become one of "global chaos, local order" in which the author indexing of specific subject fields is adequate within narrow subject fields but only poorly integrated into any overall scheme of knowledge. This view suggests a two-tiered indexing system in which distributed processing of Meta tags by large number of computers running rather simple software is supported on the next level by more complex indexing robots. These robots should be designed not to extract specific content description from individual pages, but focus on placing groups of pages or entire sites into specific subject categories and leaving the details of content to the creators of the tags (Page 218).*

Another possibility is to promote the indexing of Web resources by their users. Besser (1997) discussed the need for this. Although he was dealing specifically with images on the Web, the approach is applicable to any resources:

> *If we can develop systems for user-assigned terminology, collection managers can rely upon users to assign terms or keywords to individual images. Under such a system, when a user finds an image, the system would ask them what words they might have used to search for this image. Those words are then entered into the retrieval system, and subsequent users searching on these words will find the image. As the number of people using such a system grows, so do the number of access points for many of the images.*
>
> *It is essential that such systems allow searches against officially-assigned terms both independently of user-contributed terms and in conjunction with them. We can expect two types of searches: one that only looks at terms assigned by catalogers, and the other that looks at both cataloger-assigned terms and at user-assigned terms. Systems like this will also be able to serve as aids to catalogers. One can envision a system where periodically user-contributed terms will be "upgraded" to officially assigned terms by a cataloger (and will then be retrievable by both methods).*
>
> *As systems like this grow, future users may want to limit their searches to terms assigned by people who they trust (perhaps because they come from the same field, or because they assign terms more reliably). So these systems will likely develop both a searchable "ownership" feature for each term assigned and a "confidence level" that a user can set which applies to a group of owners. Design of systems like this will also have to be sensitive to the privacy of term contributors. Users setting confidence levels for term-assigners may locate these people through*

> *basic profiles of their subject expertise and position (but not name), or they may locate them by finding correlations between other term-assigners and how the user him/herself assigns terms to other images . . . (Pages 24-25).*

AUTOMATIC APPROACHES

Software is available to perform some indexing or abstracting of Web resources automatically. Jacsó (2002) evaluates some commercially-available summarization programs, and Reamy (2002) refers to "auto-categorization" software (i.e., programs that put electronic resources into categories automatically) and predicts major developments in this area in the future.

CONCLUSION

From all of this, one might conclude that indexing and abstracting activities are increasing in importance rather than decreasing, and that professionals in these areas can make a substantial contribution at the level of the individual Web site or at broader levels such as portal design and implementation.

> *They could also have important roles to play in the operation of company intranets. In fact, Reamy (2002), a specialist in the area of knowledge management, while predicting the growth of "auto-categorization," presents a very compelling case for the need for professionals in intellectual access activities:*
>
> *Companies don't want to pay librarians to categorize their content because they think it's too expensive. They are wrong, at least when you factor in the time employees waste trying in vain to find that document that they must have in order to answer that customer's question, without which the customer will scram and go with a competitor who had the answer instead. Despite that, many companies still won't pay for humans to categorize their content, but they are more likely to pay anywhere from 250K to 750K for software that frequently does a less effective job (Page 18).*

He goes on as follows:

> *First and foremost, auto-categorization cannot completely replace a librarian or information architect although it can make them more productive, save them time, and produce a better end-product. The software itself, without some human rules-based categorization, cannot currently achieve more than about 90% accuracy--which sounds pretty good until you realize that one out of every ten documents listed in the results of a search or browse interface will be wrong. And more importantly, it will be wrong in inexplicable ways--ways that will cause users to lose confidence in the system.*
>
> *While it is much faster than a human categorizer and doesn't require vacation days and medical benefits, auto-categorization is still simply not as good as a human categorizer. It can't understand the subtleties of meaning like a human can, and it can't summarize like a human, because it doesn't understand things like implicit meaning in a document and because it doesn't bring the meaningful contexts*

> *that humans bring to the task of categorization. One thing that early AI efforts taught us is that while speed is significant, speed alone cannot make up for a lack of understanding of meaning (Page 21).*

And finally:

> *Rather than a danger to information professionals, auto-categorization can, in fact, not only enhance their ability to solve user's information problems, it may even elevate their status to something closer to the level it should be. Not only will librarians and information architects produce more and more economically, but they will have expensive software associated with the task and, as we all know, in today's corporations, unless there is expensive software involved, no one will think you're valuable.*
>
> *Well, OK, maybe that's a bit overstated, but auto-categorization software has the potential to highlight what should already be clear--that the information professional is engaged in a fundamental infrastructure activity. Information professionals are or should be involved in the creation and maintenance of the intellectual infrastructure of their organization. While technology and organizational infrastructures have received more attention and resources, some of the imbalance could be righted through the intelligent utilization and integration of new software, new methods of working with both content providers and content consumers, and new ways of presenting information.*
>
> *So, in conclusion, I think it's likely that auto-categorization will ultimately enhance both the power and the prestige of the information professional (Page 22).*

It seems clear that the continued growth of network-accessible information resources will make subject analysis activities of greater importance than ever before. Moreover, it is likely that more and more individuals will be involved in these functions. To be sure, methods for indexing and abstracting automatically will continue to improve. However, as Lancaster and Warner (2001) point out in their review of this area, it will probably be a very long time before machines are intelligent enough to completely replace humans in these important activities, if indeed they ever are.

BIBLIOGRAPHY

Besser, H. Image databases: the first decade, the present and the future. In*: Digital Image Access & Retrieval;* ed. by P. G. Heidorn and B. Sandore, pp. 11-28. Urbana-Champaign, University of Illinois, Graduate School of Library and Information Science, 1997.

Browne, G. M. Indexing Web sites: a practical guide. *Internet Reference Services Quarterly*, 5(3), 2001, 27-41.

Campbell, J. D. The case for creating a scholars portal to the Web. *ARL*, 211, August 2000, 1-4.

Casey, C. An analytical index to the Internet: dreams of Utopia. *College & Research Libraries*, 60, 1999, 586-595.

Drott, M. C. Indexing aids at corporate websites: the use of robots. txt and META tags. *Information Processing & Management*, 38, 2002, 209-219.

Ellis, D. et al. In search of the unknown user: indexing, hypertext and the World Wide Web. *Journal of Documentation*, 54, 1998, 28-47.

Elrod, J. M. Classification of Internet resources: an AUTOCAT discussion. *Cataloging & Classification Quarterly,* 29(4), 2000, 19-38.

Fairthorne, R. A. Automatic retrieval of recorded information, *Computer Journal*, 1(1), 1958, 36-41.

Holt, G. E. On becoming essential: an agenda for quality in twenty-first century public libraries. *Library Trends*, 44, 1995, 545-571.

Hurt, C. and Potter, W. G. CORC and the future of libraries. In: *CORC: New Tools and Possibilities for Cooperative Electronic Resource Description*; ed. by K. Calhoun and J. J. Riemer, pp. 17-27. Binghampton, NY, Haworth Press, 2001.

Jacsó, P. Document-summarization software*. Information Today*, 19(2), 2002, 22-23.

Lancaster, F. W. and Warner A. *Intelligent Technologies in Library and Information Service Applications.* Medford, NJ, Information Today, 2001.

Lawrence, S. and Giles, C. L. Accessibility of information on the Web. *Nature*, 400, 1999, 107-109.

Medeiros, N. et al. Utilizing CORC to develop and maintain access to biomedical Web sites. In: *CORC: New Tools and Possibilities for Cooperative Electronic Resource Description*; ed. by K. Calhoun and J. J. Riemer, pp. 111-121. Binghampton, NY, Haworth Press, 2001.

Odlyzko, A. M. Abstracting and reviewing in the digital era. *NFAIS Newsletter*, 41(6), 1999, 85, 90-92.

Odlyzko, A. M. Tragic loss or good riddance? The impending demise of traditional scholarly journals. *International Journal of Human-Computer Studies*, 42, 1995, 71-122.

Owen, P. Structured for success: the continuing role of quality indexing in intelligent information retrieval systems. In: *Online Information 94*, pp. 227-231. Medford, NJ, Learned Information, 1994.

Reamy, T. Auto-categorization – coming to a library or intranet near you! *EContent,* 25(11), 2002, 16-22.

Trippe, B. Taxonomies and topic maps: categorization steps forward. *EContent*, 24(6), 2001, 44-49.

Watters, C. Information retrieval and the virtual document. *Journal of the American Society for Information Science*, 50, 1999, 1028-1029.

Weinberg, B. H. Complexity in indexing systems – abandonment and failure: implications for organizing the Internet. *Proceedings of the American Society for Information Science*, 33, 1996, 84-90.

Wellisch, H. H. Book and periodical indexing. *Journal of the American Society for Information Science*, 45, 1994, 620-627.

Wells, A. T. et al. *The Amazing Internet Challenge*. Chicago, American Library Association, 1999.