Knowledge Management on the Web: Global Anarchy or Global Standardization?

Prof. György Sebestyén, Ph.D.
Eötvös Loránd University of Budapest
Department of Library and Information Science
E-mail: lion@ludens.elte.hu

Abstract

In the justified euphoria over the Web information retrieval, the extremely valuable experiences of librarian science, accumulated since the most remote antiquity, seem to be utterly neglected. Even the most elementary rules, such as the primary and secondary level of information sources, are often confounded. The spreading of information literacy is very useful, but this process does not necessarily mean that the graduates have acquired all the cultural and scientific knowledge indispensable for efficient information seeking performances. Although information providers cannot assume the whole responsibility of the educational system, but by setting up fully automated interactive communicative search systems that are coupled with the reference librarians' constant personal assistance, a great deal of search mistakes can be eliminated. In this context, metadata systems should play a more important role because their content and knowledge organizational functions not only help non-specialist Web users to formulate competent queries, but they efficiently overcome paradigmal barriers hindering intercultural information exchange. Web users in their information retrieval attempts are also greatly hampered by the fact that search engines, databases and content providers offer an astonishingly disparate array of search tools and interfaces, despite the existence of the CCL (Common Command Language). Until the time when a genuinely global standardization can take place – encompassing the whole scale and scope of information retrieval, it will be impossible to remedy the above-mentioned detrimental conditions merely by fostering new inventions.

1. Introduction

Human nature is constantly fascinated by the possibility to make revolutionary inventions which will bring tremendous benefits never anticipated before. Unconsciously, this constant fascination gives rise to a deep-rooted discontent: "so little has been achieved so far and so much remains to be done". In our discipline, this kind of discontent seems to me as a particularly dangerous trap that we should avoid carefully. The present paper aims at making us aware of how numerous and how great results we have achieved and the main problem is that both information providers and users do not realize the enormous potentialities of the capabilities that we have developed by the beginning of the third millennium.

Instead of using these capabilities coherently to build up a genuinely global system, we have caused a great deal of anarchic fragmentation. No one thinks seriously that the TCP/IP (Transmission Control Protocol/Internet Protocol), the HTML (Hypertext Markup Language), and all the other protocols have solved the globalization of knowledge management on the Web. In the era of globalization, in a time when everything gets connected into networks, the actors of the knowledge based society – instead of integrating all the available knowledge – tend to disintegrate it, dividing the rift between the resources and the users.

But the facts stated above represent only one facet of the problem. Unfortunately, there is another dangerous tendency: while key-issues of sophisticated software developments are being scrutinized with incessant, unrelenting energy, (no matter how exorbitant their costs may be), some simple but crucial solutions – not only available, but visible to the naked eye – appear to be abandoned.

The above phenomena, together with their brief analysis as well as the proposed remedies are as follows:

2. The predominantly secondary nature of indexing

Disappointingly, the extremely valuable experiences of the several millennium-old librarianship are astonishingly neglected in terms of information search on the Web. Even the basic distinction between primary and secondary information/document is too often forgotten.

One can retrieve only what has been indexed (appropriately or wrongly). Well, the indexing services do a secondary level job, e.g. the bibliographic resources cannot be taken as granted for substantial information in the primary documents, the contents of which may lack quality, updating, timeliness and even authentic topicality. We can conclude that the identification of relevant material on the secondary level may lack reliability.

3. Information literacy does not solve everything

No one denies that information literacy is the appropriate educational response to the paradigm shift in the digital era. At present, there are already millions of ECDL (European Computer Driving Licence) graduates and a constantly increasing number of people have become the users of both the Internet and the other electronic information resources. Still, the problems concerning satisfactory information seeking performance – instead of decreasing – seem to be accumulating. Who or what is to blame?

A number of experts are convinced that the widespread phenomenon to receive irrelevant or inessential responses to inquiries is one that is more ascribable to the weaknesses of information users than of information providers. A great number of users simply have not acquired the sufficient information seeking skills, or even prior to the search itself, they are unable to specify their needs or purposes properly, therefore they formulate more or less wrong search questions because they do not even know what they want. The main reason for this situation is that the users do not always dispose of the cultural maturity indispensable to understand the nature of the required information. They wrongly believe that if they just know where to find the required information, they will undoubtedly succeed with absolute infallibility.[1] This misconception is particularly conspicuous in connection with the Web which is considered by a very large audience as an omniscient answering machine, providing relevant information at one or two clicks.

The information provider community cannot take responsibility for the overall solution of the above-mentioned problems whose majority belong to the long-term issues of educational policy. Nevertheless, such library schemes as the VRD (Virtual Reference Desk)[2]can offer immediate solutions that are able to handle in this area a great number of difficulties. In fact, the VRD system gives a very clear rationale for solving the key-issues of the information seeking assistance for the users.

Why? Well, for the very simple reason that the VRD allows users and librarians to collaborate on all kinds of online reference source related problems by means of a chat proxy server, providing all the capabilities and functions for a perfect interactive communication. The users' wrong query formulations as well as the misinterpretations of their requirements can be almost instantly corrected by an immediate and constantly available feedback.

4. Metadata

Information retrieval is now undergoing a significant transformation catalyzed by the increasing uses of metadata information systems. Metadata have a number of applications to enhance the performance of information retrieval, especially in the following respects:

- achievement of a common understanding among the users of different culture-related resources, in particular when these resources belong to basically disparate paradigms, because appertaining to diverse times and/or spaces,
- metadata schemes can be considered as standardized content & knowledge description tools, in particular in the digital context,
- knowledge organization, content and knowledge management cannot do without resource-descriptive metadata tools because metadata – especially integrated with data warehouses – can aggregate vast masses information,[3]
- the above points suggest that metadata schemes incidentally, as a by-product, produce a sort of text filtering effect.

The above-mentioned features enable metadata schemes – primarily the extended Dublin Core metadata element sets – to meet the needs of local digital collections seeking large-scale retrievability on the Web.[4]At the same time, Dublin Core metadata can provide information to the widest range of Web users of all professions and levels, because its application is primarily intended to non-specialist indexing which will produce information comprehensible to non-specialist user-categories, too. In this context, applied together with the XML (eXtensible Markup Language) it can be very efficient.[5] With the XML not only most metadata requirements are met, but at the same time a correct description of the content and meaning of the information is obtained.[6]

5. Appropriate means to either narrow or broaden search results

How to improve search results, that is how to adjust the search when the original search strategy produced too few or too many hits, too little or too much information.

In terms of search results narrowing-broadening, we overview briefly the state of the art of the following search features and capabilities:

- logical operators
- proximity operators
- truncation
- time and space specification
- multiple search
- OPACs

5.1. Logical (Boolean) operators

At least the logical operators AND and OR are indispensable. The problem is that the daily, "normal" meaning of these words in many languages have quite the opposite meaning to the logical one. Rational explanations only increase the disturbances and often lead to total confusion, in particular when the search must be entered quickly. (I do not underestimate the abilities of Web users if I suppose that their overwhelming majority have never studied applied mathematics - and they never will.) According to my experiences, the best way to ensure their appropriate usage consists in some kind of visualization. The display of Van diagrams with the opportunity of clicking on the appropriate solution would be not only useful, but essential. Another option would be to draw the user's attention to the synonymity expressing function of OR – in fact, it connects terms expressing almost the same concept or having similar role in their subject.

5.2. Proximity operators

In this respect we can witness very different conditions when we compare the most important database providers. Dialog e.g. provides a full array of proximity operators – some affirm that this solution is a luxury because many users will never deal with them. I think this view cannot be proved. In sharp contrast with the richness of the Dialog, the Web of Science is much more frugal: its only proximity operator – the SAME (or SENT) – prescribes nothing more than the two search terms/expressions connected must be in the same sentence – regardless of their order. Maybe the two above examples represent two extremities, still, paradoxically, either of them has its own, indisputable advantage. The information provider & user community must decide one day whether the complete range or just the simplified versions are better in searching the Web.

5.3. Truncation

Controlled length truncation must be available at any time and anywhere. It is a common place that "ordinary" truncation of short, especially one-syllable words can result in hundreds of irrelevant records, because the first syllable will necessarily be completed with a number of quite different further syllables, producing a range of completely different words. These terms generally belong to different scientific areas:

Cat
Cat**s** (zoology)
Cat**hodes** (physics, electricity)
Cat**alogues** (library and information science)
Cat**hedrals** (history of art)

Nevertheless, the truncation of one-syllable words is very often not only needed, but absolutely inevitable – because of the plural forms. The solution lies in a means that ensures that only words with up to one additional character will be selected. Generally used truncation marks, wildcard characters (*, ?) do not allow this solution for the very simple reason that *a single* mark is used.

The Dialog command language has solved this problem in an exemplary way:

Cat? ?

This solution will retrieve words either with the stem or words with up to one more character. Otherwise, the number of the maximum additional characters will be equal to the number of the wildcard characters:

Librar*** will retrieve librar**y**, librar**ies**, librar**ian**, but not librar**ians** or librar**ianship.**

As we can see, at the age of global standardization a number of perplexing inconsistencies exist among the truncation systems of even the largest database providers. Just another example: Dialog does not allow embedded truncation for such cases as sul**f**ur/sul**ph**ur and prescribes the use of the OR operator. Well, in the Web of Science quite the contrary is to be done.

5.4. Time and space specifications

Some specifications are simply forgotten. A good example is the neglect of some aspects of the dates that can be very important in the social sciences and in the arts and humanities (especially in the historical research, historiology, art history, etc). The problem is that in most systems time can be selected in terms of publication data, but as for the date limits of the period to be studied, I could not find any facilities at all.

5.5. Multiple search

Is the user allowed to select and use many databases? Professional database providers (e.g. Dialog) make it possible for the user to choose the most appropriate database or databases. In this option, the user is able either to single out one particular database or to carry out multiple searches (one search question is executed simultaneously in many databases). For the preparation of multiple searching a range of testing capabilities may be available: e.g. the user is allowed to scan and assess to what extent the tested databases are relevant to his search questions. In the latter case, the databases can be even ranked in order according to the number of the hits.

5.6. Immense disparities in the search modules of integrated library systems

This phenomenon cannot be exhaustively treated within the present framework and will be the topic of another paper. Nevertheless, I think that some concrete data should be cited here to reinforce the credibility of the above-stated points.

I consider that the list of the most important disparity areas will bring a number of cogent proofs. I emphasize that I intend to select data only from internationally known, large-scale integrated library systems. Even in this case, the disparities are enormous.

Some integrated systems provide both retrieval and browsing indexes, others provide only one of them. Even if both of these index-categories can be found, the *number* of the available retrieval and/or browsing indexes is quite different, according to the various integrated systems. Some systems offer the Boolean operators, others omit them; as for the proximity operators, their application is very rare, limited and inconsistent. Disparities occur in terms of truncation and graphic CCL (Common Command Language) capabilities, too.

Global networking revolutionized libraries primarily because their catalogues became electronic and thus accessible on the internet. But our satisfaction is far from being complete. Given the fact that each OPAC (Online Public Access Catalogue) belongs to a particular integrated system, (which the library purchased), its interface as well as the whole system of its search tools will differ from other OPACs'. Alas, how unpopular it may seem, I have to state that the virtual library world is not exempt from a certain anarchy. (That is the tax we pay for our freedom to choose on a large and free market.) It is not so difficult to imagine the chaotic circumstances in which the user is obliged to formulate the same search question in so many different OPAC environments. Fortunately, shared cataloguing is spreading fast and irresistibly.

6. Conclusion

I tried to cover a topic on which far too little has been written. I do hope that I managed to give a comprehensive overview of the state of the art of information retrieval capabilities – at least from the librarian's viewpoint. I tried to embrace a relatively broad spectrum of topics – *retrieval capabilities used predominantly by libraries but highly recommendable to improve information retrieval on the Web*. As a prerequisite for boosting further retrieval performances, fostering tools and schemes from many various – *but available* – resources and systems, their present performances have been evaluated and synthetized to reassess their applicability. The big challenge is the following: can we put an end to the fragmentation and can we use simultaneously all these tools in one, genuinely standardized information retrieval system? If we manage to standardize in such a way that each retrieval module receives the most efficient capabilities, gathered together from different systems, we will have an unimaginably powerful information retrieval system at our disposal. Having taken stock of our available means, having emphasized their impressive potentialities, I do not have the slightest intention to deny that, as knowledge providers, in the future, we shall meet many further key-challenges and new inventions will certainly play a key-role.

References

[1] Webb, T.J.: "Nice to know." *Business Information Review* 19 (3) Sep 2002, pp. 5-10.

[2] "Divine releases sixth version of Virtual Reference Desk." *Advanced Technology Libraries* 31 (10) Oct 2002, p. 7.x

[3] Lee H. Kim J. Kim T.: "A metadata oriented architecture for building datawarehouse."*Journal of Database Management* 12 (4) Oct-Dec 2001, pp. 15-25.

[4] Banski, E.: "Implementation of the Dublin Core at the University of Alberta Libraries." OCLC Systems and Services 18 (3) 2002. pp. 130-138.

[5] Kurahasi, E.: "Metadata and thesaurus: the library's challenge to remote access electronic materials 2." [In Japanese] *Pharmaceutical Library Bulletin (Yakugaku Toshokan)* 47 (2) 2002. pp. 158-162.

[6] Raisinghani, M.S.: "Extensible Markup Language: synthesis of key ideas and perspectives for management." *Information Management* 14 (3/4) Jul/Dec 2001, pp. 5, 18-19.