Subject authority control in the world of Internet

Micco, Mary

1996

Micco, M. (1996). Subject authority control in the world of Internet. Library and Information Science Research E-Journal, 6(3). https://dx.doi.org/10.32655/LIBRES.1996.3.2

https://hdl.handle.net/10356/152356

https://doi.org/10.32655/LIBRES.1996.3.2

LIBRES: Library and Information Science Research
Electronic Journal ISSN 1058-6768
1996 Volume 6 Issue 3; September
Quarterly LIBRE6N3 MICCO_1
DOI: https://doi.org/10.32655/LIBRES.1996.3.2

Subject Authority Control in the World of Internet

Mary Micco

This article has been written in two parts. Part One deals with the question of authority control on the Internet in general and the difficulties involved. Part Two deals with the use of classification in subject authority across the world of the Internet to improve filtering and precision.

Abstract:

When we talk about subject authority control today, we must see it as a system that will support searching across the vast domain of the Internet. While we recognize that there are many difficulties, it is time to take a fresh look and to explore new kinds of solutions. Automatically generated graphical displays that identify information objects at many different levels but are organized by broad subject area should be designed as navigation tools. This means we need to take advantage of the new methods of document production by encouraging authors to add in subject keywords and most importantly broad class numbers with the help of expert systems software. This initial cataloging and classification can be refined subsequently by professionals but at least it will form a base for more sophisticated manipulation and indexing from the point of entry into the system. The significant problem today is how to filter out what you dont want and restrict your search to documents likely to be useful.

Part I. Subject Authority Control must be seen in broader context

A. Goal: Building computer systems to support effective information retrieval. Doing so with minimal human intervention. For the purposes of this discussion I have chosen to limit my comments to subject authority control. I will be discussing subject authority control not in terms of the library catalog for books and non-print materials but rather from the point of view of the user with an identified information need. He/she is seeking information via the Internet which provides access not only to library resources but to a whole array of less traditional resources including the "invisible college" of his/her peers. The true purpose of authority control should be to help the user move effortlessly from his/her terminology (natural language) to the terms in use (controlled vocabulary) of the system and to locate all materials (objects) that are relevant regardless of what database they are stored in or the form in which they are presented. If he/she searches for heart attack the authority control system should link him/her to myocardial infarction and other synonyms as well as to all

the relevant or appropriate sources of information. An equally important requirement for any subject authority control system is that it should be automated and self sustaining with minimal human intervention.

If all periodical databases, electronic lists and journals, CD-ROMs, ftp directories, gopher menus, WWW home pages were classified as objects, assigned a class number (by subject) and identified by type it would be relatively simple to focus a user's search to the subject area of interest instead of doing a brute force search of the entire world's resources.

B. Domain: Recorded Wisdom of Mankind. Generally the user is interested in whatever might be available and wishes to begin his/her search from a desktop workstation. Ideally he/shewould like to have access to the entire recorded wisdom of mankind but only in so far as it pertains to his/her interest. In other words he/she wants to be able to search the whole haystack but to filter out whatever is not relevant to the particular need. He/she wants the best of what is available rather than the first 200 hits. We need tools that will collect a user profile to assist in filtering out what is not appropriate. This would suggest that we also need to implement automated weighting to assist in ranking the hits. Obviously if the hit is found in the title or controlled keywords it should be weighted more heavily than if it simply appeared in the text of one paragraph. The implication is that the user should be able to use a hierarchical classification to enter the system at the desired level of specificity in the topic of their choice with the option of broadening or narrowing a search that is not succesful. Current tools do not permit this and will not be able to unless some form of classification with a hierarchical organization is implemented. It is interesting to note how many classified catalogs are appearing on the Internet to help guide people through the maze. The success of the Yahoo server (www.yahoo.com) clearly shows that people appreciate a hierarchical organization by subject.

C. Materials: All available The Internet has greatly expanded the ways in which information can be relayed to others. Electronic mail, listservers, bulletin boards and usenet services provide easy access to a wealth of informal current communication on every imaginable topic. FTP has made possible the informal publication and distribution of materials both print and multimedia very rapidly and at low cost. Gophers have given us the ability to offer access to the entire online information of campuses, government organizations and businesses in a linked network of sites. World Wide Web has further expanded this capability by providing hypertext linking for multimedia so that we can browse graphical images, videoclips, animations and a host of other resources quickly and easily.

Even more powerful is our ability to assemble in one home page references to resources gathered from anywhere in the world that pertain to a particular interest or topic. We have greatly increased the number of places we can look for information without a commensurate increase in the sophistication of our search tools. Users should be able to specify a particular information package and we must begin to deal with the fact that certain packages represent compound structures with multiple documents e.g. Web home pages.

II. Problems: on Macro Scale

A. Information Explosion on Internet With over 3 million computers linked together with an unknown number of users and files within the past three years, it is small wonder that our capacity to manage and digest this information explosion has not kept up.

B. Lack of Cooperative Planning/ Governance The administrative structure of the Internet has focused understandably on connectivity and standards to achieve security and interoperability rather than on managing the information retrieval problems being created. Up until this point, librarians and information techonology professionals have not cooperated very actively in the design or administration of the Internet information retrieval services, although librarians have made active use of the backbone to distribute their own database services.

C. Lack of standards for describing information content Serious efforts are currently underway to redefine document management architectures and to provide more meaningful header information for files. A document will no longer be a single file but rather a book of pointers to text objects, images, fonts, sounds with summary information including author, title, keywords, version number, description and file statistics in addition to the traditional file descriptors (Reinhart). It is vital to the success of any authority control effort that the information needed be built into this document architecture in a standardized way. The selection of a class number and subject keywords must be made initially by the author of the document when it is being created and should be maintained as part of the file information in much the same way that abstracts are created for periodical articles. It would be prohibitively costly to pay intermediaries to do this. Tools including expert system components will need to be provided to assist in this process. Software that can extract matches for strings of 50 or more keywords and then rank the matches, already exists on the market. The user could simply ask for matches to their document and then select the closest ones appying the same controlled terms and class numbers to their own document. The assumption is that this technique will work very well for ephemeral material while material that is more substantive will be processed again higher up the information chain by trained professionals who can verify the choice of classification number and controlled terms.

D. Lack of any unifying information structure. Given that there are trillions of documents already and more being generated with increasing speed, we need a sophisticated information infrastructure in place to assist with the tracking and management of our information resources. Rather than treating the universe as one monolithic information service it would make much more sense to break it up into a series of hierarchically structured subject maps showing items grouped by material type.

Within each subject area map we should be able to identify critical special library collections, thesauri, encyclopedias, reference works, periodicals, databases of periodical articles, listservers, usenet groups, ftp sites and specialized gophers and home pages; all displayed in a general taxonomy showing how the subject is subdivided and how it fits into the larger scheme of things. Users should be able to switch rapidly from the subject map to the specific object of their choice. We need software that can dynamically build multitiered displays from the descriptions associated with each object in the system. Rather than the flat organization that currently exists where all keywords are treated as equal we need a tree

like organization, a classification system or taxonomy that will let the user come in at any subject level then broaden or narrow their search at will by moving up or down the tree of maps.

E. Culture Clashes. Technocrats vs. scholars vs salesmen. It is very evident that we have several different professional groups with ideas about how the resources of the internet should be organized and managed. If the general public is to be well served, these constituencies need to form cooperative workgroups and to listen to each other. They all need to agree on standards for information organization and retrieval that will provide maximum utility to all constituencies in the long run. One of the key issues in any document management system is how to identify the information so that it can be recovered on demand. There is an increasing need to be able to filter out what you don't need.

III. Problems: on a Micro scale

A. Existing Internet search tools inadequate: The current crop of search tools restrict themselves generally to searching filenames and their descriptors found in the directories of computing sites being indexed. There are obvious restrictions on the ammount of information that can be crammed into file descriptors. Without any standards or guidelines let alone authority control a filename like f-prot can and does have at least 7 possible variations. In systems that search with literal string matches this means that you have no guarantee that you will find all the instances that exist. Nor can you guarantee that the item being sought is not in the system.

1. List of Lists, Specialized ones such as Yarnoff's. Even a seemingly simple task such as locating a listserver on a particular topic becomes a major undertaking. A simple string search through the text file consisting of the names and descriptors of the lists that someone has collected is currently the best tool available.

2. Usenet This system of newsgroups offers only a primitive hierarchical subject organization as newsletters are grouped by broad topic categories with several levels of subdivision within each. But even this is useful. The concept of threads is also interesting. In that within a newsletter a particular topic or thread can be followed by means of filtering out other messages. The underlying organization is still chronological.

3. Archie [FTP]: The Archie software does not offer a great deal of sophistication or functionality but at the moment is the only way to search through Internet archive sites for items of interest. The software harvests the file directories of archive sites doing so many each night and completing the full circle once a month. The directory entries obtained are broken down into keywords then sorted alphabetically. It does have exact word/phrase searching but each term or phrase has to be done separately. You cannot AND two sets of results. It restricts itself to FTP sites.

4. Veronica (Very Easy Rodent oriented Netwide Index to Computerized Archives) This software will only search for topics on those Gophers on connected networks. Again the keywords are derived from filenames and descriptors entered in the directories. The user can use AND/ OR/ NOT boolean logic and there is also right truncation e.g. (native or aborigin*) (population* or people*).

Some Gophers limit the number of characters in a search string. You can narrow the search using a limiter "/". e.g. by the type of file. By default you only get the first 200 items although this can be changed.

5. WAIS: probabilistic retrieval. This is a more sophisticated searching tool designed for probabilistic retrieval with weighted terms in a group of full text databases. WAIS software needs more refinement and currently is being underutilized as it is being applied mainly to files containing directory entries and descriptors in menu items. You can select several databases to search from the set of databases offered, then key in your search string. Results are ranked..but again the ranking algorithms need more refinement.

6. Meta-indexes, Subject Oriented Lists, Specialized home pages. A number of groups have addressed the problems of locating items of interest on the Internet. Cern offers a meta-index organized by subject. The University of Minnesota has its library school students doing subject lists of all resources and these are made available for public distribution. Unfortunately as is typical of student projects, the coverage is not comprehensive, and there is no guarantee that the information is being updated or kept current. Specialized subject home pages offer more promise but again issues of coverage, quality control and maintenance present themselves.

7. Webcrawlers, Lycos, Aliweb.. A number of groups have developed automated tools to search URLS and index all documents found at each Web site. These are rather more powerful and since the updating is automatic more likely to be current. But again it is just a brute force key word search of the entire database with no weighting, practically no filtering.

B. Existing Library Tools restricted In most cases current library systems offer keyword searching with sophisticated boolean searching as well as minimal access to a subject heading browse. While a great deal of effort has been put into assigning classification numbers to all books and periodicals, this information is not yet being exploited or utilized effectively in current OPAC software.

1. OPAC software for books/AV materials. Most of the current systems offer only separate inverted index files for each of the identified tags. All keywords from subject headings are searchable. Or you can browse subject headings strings arranged alphabetically or you may combine keywords from title, subject headings abstracts and notes. Users are constrained to Boolean searching. If the word is not found the search fails. There is very little mapping of natural language to controlled terms or synonym control. If the user types in heart attack the system will not lead him/her to Myocardial infarction automatically. Classification is almost totally neglected. At best the user is offered a limited shelf browse of the authors and titles in the class number of interest. No captions are provided for class numbers and they cannot be browsed as a hierarchical system. The only authority control provided is for the subject headings. By and large this is the domain of the technical staff and few if any tools are provided online for users.

2. Keyword search engines (used for periodical databases) While these packages offer sophisticated boolean searching, there is little weighting of terms, and a very limited ability

to filter (except by language or year of publication) or rank or even sort the results. ILSA, an experimental prototype funded by the Council on Library Resources has demonstrated the feasibility and value of sorting results by class number thereby providing a very useful breakdown. In the case of a search for material on suicide some hits dealt with religion, others with sociology, still others with history.

In most systems you can only search one database at a time. With over 1000 periodical databases on line, it has become very expensive and difficult to guarantee finding everything on any topic. Problems with overlapping coverage and fragmentation of coverage add to the difficulties.

C. Lack of vocabulary control mechanisms While librarians have developed a number of vocabulary control tools including the Library of Congress Subject Headings and multiple thesauri for periodicals there has not been a lot of interest in or funding for efforts to enhance or automate these tools. In fact much of the research has shown that efforts to control vocabulary did not improve recall or precision at all in Boolean keyword programs within a particular database. However noone anticipated the explosion of databases that has occured. We now desperately need effective ways to filter the information to narrow the scope of our searches to the subset of interest before initiating boolean searching. We currently have no tools to do a 'brute force' search of the entire system and if we did we would retrieve more that we could possibly use. We should be exploring multi step search tools..In the first step you refine the subject area and the information objects. In the second step you delve down into full text searching with weighted terms.

1. Users get little help in phrasing queries. An alphabetic browse of keywords or terms is not much help if you don't have the right word or subject area to start with. We need a set of bird's eye views to help us to zoom in on the subject of interest. There are currently no subject area maps to guide users. It would be particularly helpful if such maps could use color to show density of hits for the different related terms, while at the same time indicating the material types. For instance, periodical databases should be distinguished as separate links the user can jump to. Thesauri, encyclopedias and other useful reference tools should be highlighted to draw the users attention and then provided in fulltext through hypertext links. Even now online thesauri are sometimes available.but they are generally not an integral part of the system and in most cases they only show the controlled terms with broader, narrower and related terms. They do not link to the classification numbers or the uncontrolled terms of the abstracts or fulltext documents. Only in a few databases like Medlars are the controlled terms linked meaningfully to a classification scheme.

2. Users cannot determine what terms are in use By browsing a classification schedule showing the distribution of the literature (counts of hits), users could quickly see how a subject area was laid out and what subtopics were evolving. They could obtain a bird's eye view of the topic and then zoom into a particular subtopic for more detail. Such tools are not provided in today's systems.

3. Users are not provided with navigation tools. With the very rapid evolution of graphical user interfaces with hypertext links to full text documents we should be able to design much more flexible search screens that provide access to a whole series of useful

vocabulary control tools to assist in the formulation of good searches. At present most programs simply ask you to enter your search string and then submit the search.

Conclusion:

It is time to rethink our approach to authority control and in fact our whole approach to subject access. We can capture and should make sure that every document carries with it certain key descriptors, a class number, a type and the intended audience as well as the who what why and where of the contents to assist with the document filtering so badly needed when we are overwhelmed with the sheer volume of information on line. This should be done first by the author of the document with the help of expert systems. Pt. II. will address steps that can be taken to improve searching on the Internet including the use of classification.

Bibliography

1. Anon. "A Literature Search for information on Native Americans." Dialog(R) File 648: Trade & Industry ASAP (TM) 1994 15234861. Full text record. v17. No.2. p 45 (10)...subject access to electronic information on this network (Internet) is still primitive compared to the powerful command languages that searchers have become accustomed to with online services

2. Bowman, C. Michael et alia. "Scalable Internet Resource Discovery: Research Problems and Approaches". Communications of the ACM. August, 1994. Vol. 37. No.8. pp98-114. "Taxonomies allow a more uniform search space than is possible solely by content-indexing of documents". Expert system technology could be developed to match the key terms in the document against similar documents. The author could then select the class number and index terms that seemed most relevant adding more as needed. "We believe tools should be developed that allow authors to mark up their documents with classification terms from some selected set of taxonomies."

3. Broad, William J. "Doing Science on the Network: A Long Way from Gutenberg." New York Times, Tuesday, May 18, 1993, p.B10, col 1. "Much of the beauty and wonder of Internet and its resources...could become a horrific problem. Systems and people will shut down. I know people who have stopped using Internet because they get 500 messages a day."

4. Kubany Susan K., President of Omnet, Inc..quoted in Reinhart, Andy. "Managing the New Document" Byte. Vol. 19. No. 8. August, 1994. "A document will no longer be a single file but rather a book of pointers to text objects, data objects, images, fonts, and so on" p.93

Go to Part Two of the Article.

_LIBRES: Library and Information Science
Electronic Journal_ (ISSN 1058-6768) September, 1996
Volume 6 Issue 3.

Mary Micco micco@grove.iup.edu

To subscribe to LIBRES send e-mail message to
listproc@info.curtin.edu.au
with the text:
subscribe libres [your first name] [your last name]

Return to Contents Page
Return to Libres Home Page

This page is maintained by Derek Silvester, Dept of Information Studies, Curtin University of Technology, Perth, Western Australia.
Please sent comments and suggestions to Derek@biblio.curtin.edu.au
CRICOS provider code: 00301J