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_2
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 dealt with the question of authority control on the Internet in general and the difficulties involved. Part Two presented here 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 at 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 II The Use of Classification in Subject authority across the world of the Internet to Improve Filtering and Precision.

I.Creating Effective Filters (subject authority)
Let us agree that one of our major problems is to develop an effective filtering system that will restrict the domain of search to relevant general areas while increasing the depth and variety of index terms within that group. This being the case, a number of steps can be taken to improve searching. The first of these is toensure that a rudimentary class number identifying the "aboutness" of the document, the type of document (information package) and the key concepts involved are included as part of the file information for every document type being added to the Internet at the point of entry. There are many kinds of filters that can be established with minimal difficulty.

A. Determining the "aboutness"or subject of the document.
Most authors are used to creating strings or titles to describe the subject of their document.

It should not be difficult to improve the quality of this indexing by encouraging the use of meaningful keywords and subtitles. When combined with major keywords drawn from the document the expert system search software should be able to match this document with others on the same topic in the system.

B. Identifying the appropriate information package.
Clearly concepts or ideas evolve and mature into established knowledge over time becoming part of the information store or legacy that is transmitted to the next generation. Along the way a great deal is filtered out, consolidated, modified and otherwise manipulated. The information goes through a much repackaging. It may start out as an item in a listserver. Then go on to become a paper on ftp, before maturing into a conference paper publised in a journal. In some cases new terms will be assigned to the concept and it will become part of the technical jargon of a discipline. It may be many years before eventually it appears as a chapter in a book..or an article in an encyclopedia or a definition in a dictionary. Only if it is very significant will it become worthy of an entire book. The final step occurs when it becomes incorporated in a textbook for use in teaching. Relatively few concepts mature to this point. Most disappear along the way. When looking at vocabulary control the question becomes at what point does a concept become mature enough to warrant the effort and cost of processing by a trained cataloger and how can we utilize our technology to provide some level of control automatically for ephemeral materials.

Since we have determined that information can be wrapped in many different packages depending on its evolutionary stage and that each of these has significance, it becomes an important way to filter information. Restrict this search to refereed journals. Search for encyclopedia articles on... What are they saying in Usenet about...Has this appeared in any conference proceedings? Search in company annual reports. . Just as important is to realize that we have the basic unit..the document in many packages but we also have compound packages which include multiple documents such as conference proceedings and periodical databases. To assist us we also have finder tools such as telephone directories, reference books, indexes, dictionaries and thesauri, information packages which assist us in the process of locating the information required. Each document should be coded by information package type to assist in the filtering process.

B. Naming the intended audience "To:"
Although librarians have for years provided a tag for the intellectual level of the document there has been considerable reluctance to make this judgement. Instead we are proposing to use the To: field. Most information is written with a specific audience in mind. It may be a research report on a new drug therapy for myorcardial infarction intended for the general practitioner or a guide for patients written in layman's terms on how to prevent clogged arteries through exercise and diet. Even if there were only 5 or so general categories such as Juvenile, Layman, Professional, Scholarly, Technical it would assist in filtering unwanted material.

C. Extracting major concepts: What, Why
Most authors can readily produce an abstract that explains what they have written and why. These keywords should be placed in the identifying information.

D. Utilizing facets such as When, Where, Who, How Equally useful for filtering are the answers to the when, where, who and how questions. If these are established as filters they can be used in addition to subject keywords to help improve the precision of the search. If a name authority control system were in place on the Internet, authors could check that they were using the correct form of the name with an expert system. If the name was not in the database it could be submitted for verification and possible inclusion.

E. Assigning the classification number: where it fits in the existingknowledge tree:
This is probably the most difficult intellectual task required of the author when describing the document. In many cases it can be newsgroup. All periodicals are classified already so that any article being submitted can automatically adopt this more general class number. FTP sites are rather more difficult because they are dumping grounds. Nevertheless they should be encouraged to group similar materials together in directories. Well maintained sites could provide class numbers for directories without a superhuman effort. Another possibility is that the same software designed to assist users to identify matches for their requests could be used by authors to locate similar materials deriving the class numbers to use from the closest matches.

F. Automatic weighting of terms based on sgml tags.
There remains the question of indexing the full text of the document. There is considerable interest in making full text searchable databases a reality and many institutions are already scanning in voluminous collections of technical papers and legal briefs in anticipation of greatly enhancing their search capabilities. The basic document description is still needed as a guidepost and will enable the indexing of large sets of full text documents so that users can quickly locate the texts of interest to them. In future systems the user may begin searching in just the file descriptions then switch to full text searching once the appropriate sets of records have been selected. Even so, it is not unlikely that users will be able to retrieve more hits than they know what to do with in full text systems unless some form of weighting is applied. For instance if a full text search for "myocardial infarction" turns up 3000 hits in a full text medical database, the user will want to use weighting to get the best of those hits. If the term appears in the file description (Weight 1) or title or abstract (Weight 2) it will obviously be more relevant than a paragraph heading (Weight 3) or a passing reference in the text of a paragraph (Weight 4). This weighting can be fully automated by noting the SGML tag associated with each term. In our prototype system we ranked the weighted terms and selected only the highest weight term when duplicates appeared. Once we have all of these in place and users have become accustomed to adding this sort of information to their document headers, a great deal of automatic indexing could be done on new items entering the system.

II Developing Information Management Tools-Expert Systems

The question of vocabulary control then has become very complex since we are talking about a system that will stand independently from the information objects in the system straddling all available databases and yet will assist users to gain access to the material type they want, in the subject area of their choice at the correct level of specificity in the relevant databases. Subject area maps will need to be created as objects in their own right and linked to each other in a tree like structure much like our road maps. If we want to

cross from California to New York City we will use summary maps that show mainly the major interstate highways. When we arrive in New York State we will probably want state and local highways.. Then when we arrive in New York City we will want a detailed street map to help us arrive at our destination. The same will be true for searching. Users should be able to zoom in and out at will using hypertext links. We will need expert systems to help in building and managing these maps as dynamic displays reflecting our evolving knowledge bases accurately and assisting users to locate areas on interest.

A. Using a classification scheme for vocabulary control Classification is an important tool in vocabulary control because it represents an organization of knowledge and can be used to determine where the subject fits in the broader context. It provides a framework for an infrastructure. It can show relationships between terms unlike alphabetic lists of terms or the bird's eye view provided by brute force keyword searching. There are several major classification systems in use in libraries for organizing the sum of human knowledge. Library of Congress Classification is widely used as a shelving system for books in the U.S. but is poorly designed for machine manipulation and is not a hierarchical scheme therefore it cannot support broadening or narrowing a topic, nor can the class numbers be exploded to capture all narrower concepts. The Dewey Classification Scheme (R) and UDC are both ideal for machine manipulation. DDC already has a machine readable version that can be displayed and manipulated by end users. It uses decimal subdivision, numerical notation and standard subdivisions with mnemonic features. Text strings are available for the numbersm, serving as a guide to the layman and providing additional keywords for searching. UDC is used in Europe but does not seem to have the recognition and widespread use enjoyed by Dewey. It is proposed that one of these be selected as the basis for a network wide classification system. In experiments conducted with the ILSA project (Micco) it was clear that even a 3 digit class number is helpful in filtering unwanted material and has the additional benefit that users can determine what aspect of the topic is being treated although 5 or more digits are preferred.

B. Desirable characteristics of a classification scheme

1. Hierarchical numbering scheme. In a decimal system if each set from 0 to 9 is reserved for a topic, each number can then be subdivided by 10 which can again be subdivided. This ties in well with the fact that we can deal only with 7 plus or minus 2 items in our short term memory. This means that each topic has a range of numbers which is known and which can conform to standard subdivisions.

2. Explode feature. If you want to search all cancers (melanomas) you can locate the broad class number and simply explode it to capture all headings (represented by decimal subdivisons) that may be more specific. This is not possible in an alphabetic list of subject headings.

3. Specificity levels. You can select the level of specificity to fit the contents of the item. In Dewey for instance, a General encyclopedia would be classified under 000. Whereas a work on fishing banks in NewFoundland would be classified under 612.46 Fisheries.

4. Effective filtering. If you only want to see items about Fisheries you simply restrict your

filter to 612.46 and you should eliminate all other items.

5. Synonym control. There may be many other words used to describe these banks but the class number will pull together all the possible variations including translations.

6. Mnemonic features. can be built in. For instance it can be agreed that the 0 digits will represent 'general treatments' while the 9 digits will represent'historical treatments'.

7. Standard subdivisions. Standard subdivisions can be generated and used throughout e.g. geographic subdivisions for place and period can be used.

8. Efficient computer manipulation. Numbers, even very long ones, can be very efficiently manipulated by machine. However humans have some difficulty in understanding them. The solution adopted in Medlars is to map the numbers to descriptive phrases designed for human consumption which contain keywords. An additional benefit is that these phrases provide additional keywords for indexing.

9. Semi-automatic indexing possible. Using a Wais index it is possible to pull similar documents then extract indexing information from the best matches.

C. Developing an expert system to manage the information infrastructure.
Over 22 million books and periodicals already have class numbers assigned so that we already have a significant knowledge base in machine readable form to start from. Each of the records also contains subject headings which represent controlled terms collected and maintained by the Library of Congress for over 100 years. Using the information already contained in the Marc records we could generate three useful indexes.

1. Class Number and Associated Subject headings and keywords. If the controlled terms and any other natural language terms in the records (derived from title, notes or abstracts) were to be machine linked to the class numbers it should be possible to automatically rank lists of terms by frequency and associate them with each class number. This distribution could be displayed in subject area maps.

2. Index Term and Associated Class Numbers. It should also be possible to create an inverted index of all controlled and uncontrolled terms showing what class numbers they are associated with. The user will enter a term and will see what class numbers contain links to this term. Experience with the ILSA experimental project (Micco) with 100,000 Marc records yielded an average of 3.7 class numbers per term. If the class numbers themselves are assigned descriptive phrases such as History-Eastern Europe-Cold War Period-Berlin Wall then the user will be able to tell what aspect of the topic is being treated. In this way the user can select the subject area most likely to be useful.

3. In a third index, natural language terms should be linked to the controlled subject headings whenever they appear in the same record, thus providing a form of automated synonym control. If the user looks up a term, any links to other terms will be displayed. With such a large database to work from there is a very good likelihood that a user will get a match for his/her term and will be able from this entry point to determine what synonyms have been used and also to locate broader, narrower and related terms. If trained librarians

were to monitor this evolving database with the ability to quickly make adds, deletes and global changes authority control would be intimately linked to information retrieval and display and we would gain maximum benefit with minimal human intervention. If all objects recognized as finder tools (eg. thesauri, dictionaries, encyclopedias) were to be identified by codes in this primal database then it would be possible to link in more detailed subject maps and specialized finder tools such as thesauri for more in depth searching.

D. Periodical databases: These could be handled in much the same way. Here the periodicals have already been classified. We should be able to assign the periodical class number to each article derived from that periodical. Once again we propose linking the terms in the title and abstract to the class number and vice versa but here weighting will be more significant in helping to filter. For each database the thesaurus with its controlled terms and network of references should become an integral part of the search software.

E. Full Text databases are rather more difficult since they vary considerably in granularity. One the one hand you have a CD-ROM with the full text of the Grolier Encyclopedia, on the other you might have a database with 100,000 technical papers for the Department of Defence with each document having been assigned controlled terms from a DOD thesaurus. With each of these databases professional judgements will have to be made about the degree and depth of indexing and classification to be used. The indexing and authority control tools used by catalogers need to be carefully scrutinized and coordinated into a complete set of tools to be put online and made accessible to users and catalogers guided by expert system rules to help users select the appropriate level.

F. Fitting it all together: To ease the problem of navigation and to assist users we will need an additional level of organization based on the classification. If every information object on the internet were to be assigned a class number it should be possible to generate maps which show what resources of all types are available in a given area with additional data showing the each object and whether or not it is searchable and or can be browsed.This of course could be machine generated from the information already collected in the file information.

G. Expert systems. To manage the complexity of such a network of sources and to assist users, expert systems should be developed which will first collect the user profile, negotiate the information need, and then assist the user to navigate through the Internet locating the informationpackages best suited to his/her need and more importantly filtering out material deemed irrelevant.. One of the more difficult tasks will be displaying the retrieved information in ways which will educate rather than confuse the earnest information seeker. The system should always leave the user in control, and be able to explain its decisions. Users should be able to take advantage of a system of well designed links between the many different finder tools, so that they will be more easily made aware of the variety of resources at their disposal and can narrow down their selections effortlessly. The ultimate goal of the system must be to make the best possible match between the user's information need and the resources in the system.

III. Conclusion.

Every item should have a standard header to assist information management and particularly filtering. If we are to build effective expert systems to help our information retrieval they need consistent placement of the critical information and consistent use of language. The work of creating the standard header can best be done by the author of the document providing that they have tools at their disposal to determine the class number and subject headings that seem best. We are assuming the use of SGML tags to provide weighting of terms in the system.

An information infrastructure based on an underlying classification structure must be developed made up of maps with pointers to search tools, finder tools, reference books, databases, vocabulary control tools, periodical databases, FTP sites, listservers, etc. Instead of one giant flat file of keywords we need to provide as rich a mixture of access tools including indexes and thesauri as we possibly can to meet different needs and expectations but these should be identified and flagged in the subject area maps so that users can switch into them easily.

Even within the library world there is disagreement about the use of classification in on line systems. However the very successful Medlars system has proved beyond a doubt, the value of having an underlying hierarchically structured numbering system. We are proposing that classification be revisited and used to provide some order in the chaos of the Internet.

Bibliography

1. Bowman, C. Mic 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."

2. Boynton, G.R. and Sheila D. Creth, eds. New Technologies and New Directions. Westport, CT: Meckler Publishing, 1993. 118p. "Nine articles from a symposium of University "scholarly" publishing, learning, creating and management of information using computers."

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." Susan K. Kubany , President of Omnet, Inc..quoted in.

4. Chan, Lois Mai. "Part II: Library of Congress Classification System. The Library of Congress in an Online Environment." Cataloging and Classification Quaterly 11 (1):7-25. "One of the great advantages of online retrieval systems is that access provisions need no longer be either/or propositions. The question now is, how can we make the best of our

battery of bibliographical access tools, with classification and the alphabetical approach used together to complement each other?" p 25.

5. Cochrane, Pauline A. and Karen Markey. "Preparing for the Use of Classification in Online Cataloging Systems and in Online Catalogs." Information Technology and Libraries 4 (June 1985):91- 111). For the online catalog user, library classification becomes a tool for augmenting subject access, providing browsing capabilities through the classed approach to subject searching in the schedules and the alphabetical approach in the index and enhancing the display of library materials' subject matter. 109

6. McMillan, Marilyn and Gregory Anderson. "The Prototyping Tank at MIT: "Come On In, the Water's Fine". Cause/Effect. Vol.17.No.3. Fall, 1994. pp51-54.

7. Micco, Mary and Rich Popp. " The ILSA Project: an Investigation into Techniques for Improving Subject Access : A Theory of Clustering based in Classification".. Library High Technology. June, 1994.

8.. 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 One of this article

This article was originally published in
_LIBRES: Library and Information Science
Electronic Journal_ (ISSN 1058-6768) September, 1996
Volume 6 Issue 3.

Mary Micco micco@grove.iup.edu

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