Abstract: Open system protocols for search and retrieval have not provided explicit ways in which to implement thesaurus-aided searching. A number of different approaches within the existing protocols, as well as a proposed service, are evaluated. A general approach to implementing thesaurus-aided searching, particularly during consultation of a thesaurus, requires an entirely new service, whose main features are described.

1. Introduction

Thesaurus-aided searching is the explicit use of a thesaurus by a searcher to enhance information retrieval in an automated system. Thesaurus-aided searching has been used in many information retrieval (IR) systems to enhance retrieval, usually for subject access, but sometimes for other hierarchically-structured authority files such as those for personal or corporate names. However, problems have arisen when trying to generalize this function for information retrieval standards such as the ANSI Z39.50 Search and Retrieval Protocol (ANSI/NISO, 1995) or the associated international Search and Retrieval Protocol ISO 10162 and 10163 (ISO, 1992). The ANSI Z39.50 and the ISO 10162 and 10163 standards (hereafter referred to simply as the SR protocols) describe in detail the interaction between client and server systems (referred to in the standards as origin and target systems) in providing distributed search services. Though designed to be "comprehensive information search and retrieval protocol[sl]" (St. Pierre, 1996), these standards contain as yet no explicit thesaurus-aided search capability, largely because of the difficulty in agreeing on the way in which thesauri can be used to enhance the search process. This situation is increasingly restrictive in the open systems world of internetworking, and has discouraged developers of SR origin or target software from considering implementation of full thesaurus-aided searching.

This paper briefly describes previous efforts to implement thesaurus-aided searching using current SR protocols, and a recently proposed service based upon the model of the thesaurus as a database separate from the principal searchable database. In order to provide a generalized, simple and comprehensive approach to thesaurus-aided searching, however, another new service is also required. The main features of this new service are described, along with several scenarios describing how it could be used to support thesaurus-aided searching.

Broadly speaking, there are two distinct phases of user-thesaurus interaction during thesaurus-aided searching: a consultation phase, in which the searcher (i.e., the end user) explores the thesaurus, collecting information which can be used in formulating a search strategy; and a query phase, in which the searcher submits a search strategy with explicit instructions to enhance that search using the structure of a thesaurus. (The enhancement of a search without the knowledge of the user is excluded from this discussion, since it depends only the target software, and has a minimal effect on a search and retrieval protocol.) Each of these two phases is examined separately.
2. Thesaurus Consultation

Thesaurus consultation is the phase of thesaurus-aided searching in which the user consults the thesaurus in order to gather information that will help to develop the strategy actually submitted during the query phase. Currently there are a number of possible approaches to implementing thesaurus consultation within the SR protocols.

The ability to learn about possible search keys in a database is provided in the SR protocols by the Scan Service of the Browse Facility, an addition to Version 3 of the protocols. This service was modeled on a feature supported by many IR systems which provides the user with a list of possible search keys, usually extracted from the inverted or index file which provides access to the records stored within the database. In formulating a Scan request, the origin system supplies a number of data elements, including the field with which the index is associated, the search key to be used as the entry point into the index, the maximum number of index keys to be returned, and the relative position within the list at which the search key is or would be found. In response to this request, the target system returns an ordered set of search keys, embedded in a very simple data structure. The ordering of the search keys, and the ability of the origin to provide a relative position of the search key, allow the origin to request lists of keys preceding or following the current list, in effect providing the means for the searcher to "page up" or "page down" through the entire list of possible search keys in the index.

This simple Scan service was clearly intended for sorted key lists that have no particular structure, such as title keywords, and in this role it serves very well. However, two different techniques have been used as a means of passing structured thesaurus information using this service. The first technique is to make use of an undefined data element called OtherInformation found in the simple data structure used to return keys in the target's Scan response. This data element was designed to allow targets to include, with each key, certain specific pieces of information that apply to that key, and can be used to graft another data structure onto the returned index key. Using this data structure, target systems can pass back to origins a complete thesaurus record structure including scope notes, history notes and other language versions of descriptors as well as a full range of thesaurus relationships including Broader, Narrower, and Related Term relations.

There are two objections to this approach. First, both target and origin must agree on the structure that will be passed back to the origin with the search key: while the developer that provides both origin and target software to customers can arbitrarily determine such a structure, this works against the notion of interoperability which is at the heart of the SR protocols. Second, any definition of this OtherInformation data element requires a fixed, pre-determined structure, and a structure sufficiently general to accommodate a wide range of thesauri could very quickly become large, complex and difficult to modify for unanticipated future requirements.

A second approach to expanding the current Scan service to support thesaurus consultation has been to use an SR construct called attribute sets to request the target to return information about thesaurus associations. Attribute sets are a Z39.50 construct that allow the specification of a number of characteristics to be associated with a search term: for example, a Use attribute such as "Date of publication" indicates that the origin hopes to find a specific search key used in a date of publication field, and a Relation attribute, such as "less than", indicates that the origin wishes to search for values less than the corresponding search term. A basic attribute set called Bib-1 defines common characteristics bibliographic data and has been responsible for much of the success of the standard in ensuring interoperability between different Z39.50 origins and targets. Another attribute set called the Scientific and Technical Attribute Set (STAS-1) was developed to include data elements appropriate to scientific and technical information; this attribute set includes in its Relationship attributes the basic thesaurus relationships, as well as values not
nonnally considered relationships, e.g., Scope Notes, and combinations of various relationships, e.g., All (STAS Maintenance Agency, 1995). Though designed principally for searching, by attaching attributes for these kinds of relationships to a Scan Request the origin could in effect request that the target return certain associated values for each search key, i.e., keys and text strings that had a specific kind of relationship to the main index key.

The use of attribute sets in this context seems somewhat non-intuitive, especially for identifying thesaurus data elements that do not apply directly to searching at all, such as definitions, scope notes and history notes. The size and complexity of attribute sets has already been noted as a problem in SR protocol development (Lynch, 1996): including in each attribute set data elements necessary for the display of information from any associated controlled vocabulary will only increase their size and complexity. Finally, like the use of Other Information, the use of attribute sets only allows a single sequence of controlled terms, whereas some thesauri could profitably be consulted in both an alphabetic and systematic sequence.

While both of these approaches allow enhancement of the Scan Service to include some information about a thesaurus, neither of them is alone sufficient for full thesaurus consultation. In particular neither provides a model for viewing only certain selected thesaurus terms. A common requirement during thesaurus consultation involves a request to view only a particular subset of thesaurus terms, such as all descriptors that contain a particular keyword; the resulting terms, and associated data elements are returned to the user but not necessarily presented in any particular order. This behavior is difficult to model using the Scan service which is implicitly based on a sequential display of all search keys from an index file. As a result of some of these concerns, a proposal has recently been advanced within the Z39.50 Implementors' Group to develop a new service within the Browse Facility, separate from the Scan service.

This proposed new service, called Structured Vocabulary Browse or simply SVB (Kunze, 1995), recognizes that the consultation of a structured vocabulary such as a thesaurus is more complex than a simple "scan" of index keys. The underlying model for this new service is not that of a list of index keys, but rather a database search, where the database being searched is not a bibliographic database, but a thesaurus or other structured vocabulary: the origin submits a specific keyword, and receives back a number of records that match the request. This proposed new service has several advantages. First, it allows for only a subset of thesaurus terms being returned in response to a consultation, by allowing developers to make use of some of the same software and functions already developed to support bibliographic searching in which results are usually a subset of records provided without an explicit sequence. Second, the proposal suggests returning information in a flexible structure that depends upon the thesaurus in use by the target, in the same or similar record syntaxes already cited within the SR protocols for the return of bibliographic information. This use of these existing record syntaxes (such as MARC authority formats, or flexible, general syntaxes such as the General Record Syntax GRS-1), can accommodate a variety of different types of thesaurus structures and hence promotes interoperability.

However, this new service is deliberately limited in that it does not attempt to replace Scan in terms of providing access to a full thesaurus in an explicit sorted order; nor does it address the limitations of both of the previously mentioned approaches towards using Scan for thesaurus information. It also introduces an evident duplication, since in one service (Scan) thesaurus information is presented in a fixed, pre-determined, private format, or in the form of an attribute set, whereas in another service (SVB), the same information associated with each term is presented in a record syntax. Origin software wishing to support thesaurus-aided searching across a variety of target systems and databases must now be able to process essentially the same data in three different ways: once as extended information to the Scan response (using perhaps several different database-specific structures), once in the form of an attribute set, and once in the format
of a database record in one of the recognized SR record syntaxes, such as GRS-1 or a MARC authority format.

The Structured Vocabulary Browse service represents a move towards a more general and flexible support for thesaurus consultation by modeling the thesaurus as a database, and returning records in a record syntax. These features need to be extended to the sequential viewing of thesaurus contents. The problems evident in extending the Scan service to thesaurus information suggests the need for a new SR service that (like the Structured Vocabulary Browse service) is implicitly based on the notion that consulting a thesaurus is like consulting a database. This notion should be familiar to many IR system developers because in many systems the controlled vocabulary is stored physically as a database, in a file separate from the index file containing keys used for searching the main bibliographic database. A complete thesaurus consultation process would make use of the existing SR Explain service to discover information about the thesaurus, the proposed Structured Vocabulary Browse service to identify specific unsorted subsets of thesaurus terms, and a new proposed service, which we will call Structured Vocabulary Scan, to provide access to sorted segments of thesaurus terms. With the use of these three services, a complete and generalized interface for thesaurus-aided consultation encompassing all navigation of thesauri by the end user of the thesaurus could be implemented in the SR protocols.

First, an origin can discover the name of a structured vocabulary database by requesting information about the field in the main bibliographic database using the SR Explain service. The Explain information on a given field could easily be extended to include the name of the database that contains the structured vocabulary associated with that field. The origin can then submit a second Explain request against the structured vocabulary database (i.e., the name of the thesaurus database), to obtain information about the structure of the particular thesaurus, such as the number and nature of different thesaurus relationships (BT, NT, RT), of different language versions of a term (English, French, Spanish), and of different purely informational fields (scope notes, history notes). The only additional piece of information required by this thesaurus use of the Explain service is an indicator of which fields represent relationships (e.g., BT, NT, RT) which can be used either in searching the bibliographic database or in navigating the thesaurus, and which fields represent purely information data elements, such as scope notes.

Second, an origin can request a search of the thesaurus database, using (for example) a keyword taken from a descriptor. This request would use the proposed Structured Vocabulary Browse service described above, in which the origin would request a search of the thesaurus database using some search criteria (e.g., "all terms with Canada in them"); in response, the target would usually reply with a number of thesaurus records, returned in a record syntax. The origin could then display those records to the searcher in a format which would allow the searcher to decide which term to use in another search (thus ending the thesaurus consultation) or would allow the searcher to decide what record or records provided an interesting point for further consultation.

Third, an origin can use the new proposed service, the Structured Vocabulary Scan service, to browse the thesaurus. In this new Service, the origin would send a Structured Vocabulary Scan request to the target, and the target would return a Structured Vocabulary Scan response. The request would include a number of specific data elements including the origin's preferences for the sequence of the thesaurus records returned, a starting term value, and a preferred record syntax. The response would include many of the same elements, indicating what the target was able to supply, usually with a sorted set of thesaurus records returned in response to that request. A brief description of the key data elements in the request and response follows.
Database name is the name of the structured vocabulary (thesaurus) database from which the records would be drawn. The origin would determine the database name through the Explain service.

Term is the thesaurus term which provides the entry point into browsing the thesaurus. The origin supplies this value; no response is required from the target.

Sequence is the order in which thesaurus records are returned. The origin requests a preferred sequence, and the target responds with the sequence which has been supplied. Typical sequence values would be alphabetical (records in alphabetical order by term) or systematic (records in order according to an explicit or implicit classification scheme). If the origin requested a specific sequence and the target could not supply the requested sequence, the target would specify in the response the sequence that was being supplied.

Number of entries is the number of thesaurus records which the origin expects to receive back in the response to the request (in a request) or the number of records actually returned by the target (in a response).

Position in response is the position of the record corresponding to the requested term in the sequence of returned records. The origin indicates a preferred position, and the target indicates the actual position in the response.

Levels up and levels down indicate a number of hierarchical levels above or below to include in the response. The origin indicates the preferred number of levels to display, the target indicates the number of levels actually present in the records returned. This value allows the supply of hierarchical record displays, in which BT and NT terms are provided for a certain number of levels.

Element set name provides a way for the origin to indicate the desired fullness of records in the response. Element set names such as B for Brief record format or F for full format should be supported by all targets. If the origin intended to display only a limited amount of information about each term, such as the term itself, as would be found in a systematic listing in which no explicit BT, NT or RT relations or notes fields are supplied, a Brief record format would be preferred. If the origin intended to display more extensive information about the term, e.g., complete relationship information such as is found in many alphabetical displays, then a Full record format could be requested.

Record syntax is the syntax in which origin would prefer to have the records returned, or the record syntax in which the target actually returns the records. Record syntaxes would include Simple Unstructured Text records (SUTRS), General Record Syntax (GRS-1) records or MARC authority format records. A response in GRS-1 or a MARC authority format, together with the information about data elements in the thesaurus obtained through the Explain service, would allow the origin that made use of a Graphical User Interface (GUI) to format records, so that a user could click on a BT relation, automatically submit a request for another browse based on this particular term, and then see the selected term display. This would allow searchers to navigate simply and easily through multiple thesaurus displays searching for terms to add to their search strategies.

Finally, response records would the set of records returned by the target as part of the response.

Like the current Scan service, the new SVS Service allows the consultation of an ordered list of items. Unlike the current Scan service, the origin can request one of several different sequences, the thesaurus structure is determined by the target system based on the characteristics of the particular thesaurus used, and the results are returned in a record syntax both can express a wide range of thesaurus structures, but in such a way as to allow origins to interpret record contents and present them to the searcher in a useful way. Two or three simple scenarios,
describing the interaction between origin and target, will indicate how this new service could be used to support consultation of a thesaurus.

**Example 1:** The user wants to consult an alphabetic thesaurus such as the OECD Macrothesaurus, displaying records starting with the term "DEVELOPMENT AID". The client submits a request indicating the target's name for the Macrothesaurus database, term "DEVELOPMENT AID", sequence alphabetic, number of entries 5, position in response 2, levels up and levels down 1, F as the element set, and record syntax USMARC authority format. The target returns five records, starting with the record for "DEFENCE POLICY", (the term that immediately precedes "DEVELOPMENT AID" in the thesaurus) and continuing for the four records that follow "DEFENCE POLICY" in alphabetic order in the thesaurus. The records are provided in GRS-I syntax. The origin is able to extract BT and NT fields from this record syntax, identify them as selectable, and to format them on the screen so that the searcher can then click on the BT term "INTERNATIONAL COOPERATION"; the origin then submits a new request as before, but with search term "INTERNATIONAL COOPERATION", and receives back five more records in alphabetic sequence with the record for the descriptor "INTERNATIONAL COOPERATION" in the second position.

Of these five records, the first corresponds to the descriptor "INTERNATIONAL AGREEMENT" and the fifth corresponds to "INTERNATIONAL RELATIONS". The searcher issues a Page Up command, and the origin then requests term "INTERNATIONAL AGREEMENT", sequence alphabetical, number of entries 5, position in response 5. The server responds with a new sequence of 5 records, of which the 5th record corresponds to the descriptor "INTERNATIONAL AGREEMENT". The effect is to "page up" within the alphabetic sequence of the thesaurus.

**Example 2:** The searcher wants to consult a classified thesaurus, like the Art and Architecture Thesaurus, starting with the term "paint". The origin sends a request with database name for the AAT on the target system, term "paint", sequence systematic, number of entries 20, position in response 3, levels up/levels down 0, element set B (for brief), record syntax GRS-I. The target responds with 10 records, of which the first record corresponds to the descriptor "Maroger medium" (i.e., the term two terms before "paint" in the classified sequence of the thesaurus) passes through "megilp" and "paint", and finishes at "enamel paint", the descriptor 17 positions after "paint". Using information about the relative hierarchical level of each term returned within the record structure, the origin could, for example, format the descriptors one to a line in a 20-line window, with variable indentation to reflect their level within the thesaurus hierarchy.

**Example 3:** The searcher wants to consult an alphabetic thesaurus like the OECD Macrothesaurus using a hierarchical display. The origin sends a request with the database name known by the target for this thesaurus, the term "INDUSTRY", sequence alphabetical, number of entries 1, position in response 1, levels up 2, levels down 3, element set name F for Full, record syntax GRS-I. The target responds with a single record, pre-formatted by the target, displaying two levels of BT and 3 levels of NT embedded in the record. The origin could then show the searcher the one returned record in this "hierarchical" format of, however because of the simple unstructured format of the returned record, the searcher would likely have to type in another term appearing on the display to navigate to another part of the thesaurus.

3. The Query Phase

While thesaurus consultation requires some major changes in the SR protocols, the second phase of thesaurus-aided searching, the query phase, in which the user actually submits a search
against the source database, does not present the same modeling complexities. Once there is a mechanism in place for the origin to discover information about the data structure of the thesaurus, e.g., the number and nature of the thesaurus relationships between one term and another, then the origin can use these thesaurus relationships to explicitly request an enhancement of the search. The search request would have to be extended to allow, not only the specification of the descriptor to be searched, but also one or more expansion requests, where each expansion request consists of the specification of a thesaurus relationship (e.g., BT, NT, RT) and the number of degrees of relationship that would apply. For example, to request an expansion of the term "paint" to the descriptors below "paint" in the thesaurus, the search would consist of the term "paint", and an expansion specification of NT and level "1". To implement an "explode" operator, the origin would submit a search request with a term, and an expansion request of NT with the degrees of relationship either an arbitrary or an unnaturally high value such as 9999. With some modifications in terms of the means of specifying thesaurus relationships and in the additional specification of a number of degrees of relationship, the current Search Service could provide all the functionality required to use a thesaurus in searching a bibliographic database.

4. Summary

Generalized support for thesaurus-aided searching requires a mechanism for discovering information about a thesaurus, a means of consulting that thesaurus through a sequential display of records as well as a selected subset of records from that thesaurus; and a mechanism for using thesaurus relationships and differing degrees of relationship in the search process. Existing or proposed services in the SR protocols provide, with some relatively minor extensions or modification, three of those four requirements. However, the sequential display of thesaurus records demands a new SR service, which would present a number of advantages over the way in which thesaurus-aided searching can be currently implemented in SR protocols. By providing thesaurus information to be returned in a full record syntax, it would allow for a richer set of data elements to be returned, the use of already existing data standards (such as authority record formats) where appropriate, easier modification of the format of those data elements, and greater adaptability of the formats to accommodate a broad range of thesaurus and controlled vocabulary structures. The author is currently compiling a survey of thesaurus-aided searching supported in currently available information retrieval and library search software. The results of this work will be used to verify the generality and appropriateness of the proposed approach in meeting the needs of searchers using standard information retrieval protocols.

References


Lynch, Clifford. (1996). Defining and maintaining attribute sets for use with the Z39.50 protocol: a discussion paper. "Z3950W@NERVM.NERDC.UFL.EDU. Feb 2, 1996. Available from LISTSERV@NERVM.NERDC.UFL.EDU.