Case Study

Development of the PRESNET Subject Descriptor Thesaurus

WILLIAM H. McNITT

Abstract: The PRESNET Thesaurus is a controlled vocabulary of subject descriptors employed in indexing the descriptions of archival materials entered in PRESNET (the Presidential Libraries Information Network). An archivist involved in its development and use discusses the decision to employ a thesaurus, the steps involved in designing and creating it, subsequent modification of the thesaurus after usage began, and the experience of the Gerald R. Ford Library in using it.

About the author: *William H. McNitt has been an archivist with the Gerald R. Ford Library since 1977 and serves as the PRESNET system operator and thesaurus usage supervisor.* *The author would like to thank his colleague Dennis A. Daellenbach for assistance with and comments on this article and to acknowledge the contributions of fellow thesaurus constructors Robert D. Bohanan and Thomas F. Soapes of the National Archives and Records Administration. This article is a revised version of a paper presented at the Society of American Archivists meeting in Atlanta in September 1988.*

IN PLANNING AN ARCHIVAL description database, one major concern is authority control over indexing terms that provide subject access. Many information retrieval systems provide such authority control through the use of a thesaurus. An information retrieval thesaurus can be defined as "a compilation of words and phrases showing synonyms, hierarchical, and other relationships and dependencies, the function of which is to provide a standardized vocabulary for information storage and retrieval."[1]

Although similar in many ways, thesauri are not identical to library subject heading lists. Thesaurus terms represent single concepts, while subject headings often represent multiple concepts or employ subdivisions. Thesauri are based on sets of rules, while subject heading lists, such as the Library of Congress Subject Headings, contain many inconsistencies in term forms and relationships.[2]

Many archivists are aware of deficiencies in the Library of Congress Subject Headings, especially in their use with automated information retrieval systems, but continue to use them due to the lack of a thesaurus suitable for use with large national databases.[3] Nevertheless, archival institutions developing local description databases should consider the construction of smaller, specialized thesauri as a viable alternative.

PRESNET and the Development of the PRESNET Thesaurus

The National Archives and Records Administration's Office of Presidential Libraries contracted with American Management Systems, Inc. (AMS) in March 1983 to study whether an automated system could improve access to the holdings of the presidential libraries, standardize descriptive practices among the libraries, and improve productivity of archival work. Over the next three years, AMS produced reports covering functional requirements, cost estimates, system concepts, and a system design for the proposed Presidential Libraries Information Network (PRESNET).

After completing a prototype of PRESNET's Manuscript Processing and Reference subsystem in the spring of 1986, AMS installed it for testing at the Gerald R. Ford Library in Ann Arbor, Michigan. The prototype automated a variety of processes involved in solicitation, accessioning, description, arrangement, and reference activities. A three-month field test identified several "bugs" to fix and enhancements to add. While PRESNET continues to undergo refinement and enhancement, the system has been operational for some time and currently is fully integrated into the library's processing and reference activities.[4]

Since early in the planning process PRESNET has included a thesaurus to enhance subject access. Some staff members expressed initial skepticism of the plan to use subject descriptors, believing that searching on folder titles in a free-text mode

[1]American National Standards Institute, *American National Standard Guidelines for Thesaurus Structure, Construction, and Use* (New York: American National Standards Institute, 1980), 9.

[2]Mary Dykstra, "LC Subject Headings Disguised as a Thesaurus," *Library Journal* 113 (1 March 1988): 42-46; Tze-chung Li, *An Introduction to Online Searching* (Westport, CT: Greenwood Press, 1985), 34.

[3]Avra Michelson, "Description and Reference in the Age of Automation," *American Archivist* 50 (Spring 1987): 197-198; Jean E. Dryden, "Subject Headings: The PAASH Experience," *Archivaria* 24 (Summer 1987): 175-176.

[4]PRESNET is based on General Physics Corporation's SEEK database software and PRIME INFORMATION database management software and currently runs on a PRIME 2455 minicomputer. Modules to automate audiovisual processing and reference, museum object tracking, and document declassification tracking may be added in the future. While PRESNET is currently in use only at the Gerald R. Ford Library, it is scheduled for installation at the Jimmy Carter Library early in 1990 with the possibility of expanding it to other libraries as resources permit.

usually would suffice. Their lack of experience with descriptors prevented the staff from appreciating the advantages that a controlled vocabulary could provide in overcoming certain problems involved with online searching. For example, folder titles lack standardization from one collection to the next and often do not adequately describe the folder's contents. In addition, the terminology applied to various national problems has changed significantly during the presidencies covered by the presidential libraries (1928 to the present). These potential searching problems led AMS to recommend the use of subject descriptors in PRESNET.

Having decided to employ a thesaurus, the Office of Presidential Libraries considered several options. One set of possibilities included using an existing thesaurus, with or without modification, or combining elements from several existing thesauri. At the other extreme, an expensive and time-consuming alternative was the design of a new thesaurus from scratch. A study of existing thesauri identified none fully satisfactory for use with PRESNET. Many information retrieval thesauri cover specific scientific or engineering fields. Existing social science thesauri proved to be too narrow or specialized because modern presidents deal with such a wide range of issues.

The Office of Presidential Libraries therefore chose to construct a new thesaurus[5] and to base it on an existing classification scheme—the White House Central Files filing manual. For several decades the White House has employed a single filing system to assign documents to subject categories. The manual thus reflects most topics on which a presidential library is likely to hold materials. The Central Files manual contains both subject terms and related alphanumeric codes. The Office of Presidential Libraries considered using Central Files codes for indexing purposes but chose instead to develop subject descriptors because the codes are revised periodically and often vary from one administration to the next. A committee of archivists representing several presidential libraries planned the conversion of the Central Files manual into a thesaurus. Committee members examined thesaurus construction standards and existing thesauri, consulted experts, and drew up lists of terms unique to specific presidential administrations. One archivist then completed most of the actual construction work.

Compared to the extensive work involved in term selection for many other thesauri, the Central Files headings made compilation of a proposed list of terms for the PRESNET Thesaurus relatively easy. This eliminated the need for time-consuming test indexing to generate a list. The archivist constructing the thesaurus then examined the initial list of descriptors to determine whether they were specific enough for their planned usage.[6] Although the Central Files staff had provided sufficient specificity in many sections of its classification scheme, the archivist occasionally added terms associated with only a single presidency (e.g., *Nixon Pardon* or *Maya-*

[5]Thesauri designed specifically for use by archival institutions are relatively rare. For information on two, see Lawrence J. McCrank, ed., *Automating the Archives* (White Plains, NY: Knowledge Industry Publications, 1981), 73-84 and 177-188.

[6]The number of terms to include in a thesaurus depends on the scope and complexity of the subject field, the kind of data to be indexed, and the intended specificity and exhaustivity of indexing. While the thesaurus must be sufficiently specific to define topics narrowly, a larger number of terms may also mean that more subject expertise is required of indexers and searchers. See Dagobert Soergel, *Indexing Languages and Thesauri: Construction and Maintenance* (Los Angeles: Melville Publishing Co., 1974), 6; F. W. Lancaster, *Vocabulary Control for Information Retrieval* (Arlington, VA: Information Resources Press, 1986), 140; Carol Tenopir, "Searching by Controlled Vocabulary or Free Text?" *Library Journal* 112 (15 November 1987): 59.

guez Crisis for the Ford administration) or combined Central Files headings under a more general term.

Some potential terms proved to be synonyms or near synonyms. In other cases, the thesaurus construction archivist wished to include synonyms not on the list as cross references in the final vocabulary. He selected one synonym from each set as the postable term and made the others nonpostable by adding cross references in order to lead indexers and searchers to the appropriate term (e.g., *Airlines USE Civil Aviation*).

Thesaurus construction standards dictated the proper form (noun or verb, singular or plural, etc.) for each term.[7] For instance, standards specify the use of direct entry or natural word order instead of the inverted form. The library subject heading *Privacy, Right of* thus appears in the thesaurus as *Right of Privacy*. For ease of data entry and retrieval, terms were made as short as possible; *Campaign Debates* is used instead of the Central Files heading *Debates of Political Candidates*.

After completing term selection, the archivist determined the scope and definition of each term. In order to broaden or narrow the definition from its common meaning or give usage instructions, he added scope notes. A more difficult and time-consuming task involved the determination of *Broader Term, Narrower Term,* and *Related Term* relationships (see the sample entry in Figure 1). Terms have a Broader Term-Narrower Term relationship only if one names a class of concepts and the other represents a member of that class (e.g., *Aviation* is a Broader Term to *Civil Aviation*), although some specific exceptions are allowed.[8] As he determined such relationships, the archivist added the appropriate cross references under each term. He also connected many terms that did not meet the requirements for a Broader Term—Narrower Term relationship by making them into Related Terms (e.g., *Aviation* is a Related Term to *Aircraft*).

Determining hierarchical relationships is often difficult, especially for social science terminology. Although a significant time commitment is required, it is extremely important to link semantically related terms. A good thesaurus is costly to draw up, but it gives more help in displaying useful relationships among terms and lightens the intellectual load on the indexer and searcher.[9]

About six months of construction work resulted in a draft thesaurus containing approximately 900 postable terms. Its small size (some thesauri have many thousands of terms) was due to the fact that the White House deals with many subjects in merely a general way, so very specific descriptors were needed for only selected areas of activity.

As with any new undertaking of this size, the first version of the PRESNET Thesaurus contained some imperfections that were not apparent until after usage began. These included:

1. the need to follow some thesaurus construction standards more closely;
2. broad descriptors requiring additional narrower terms to achieve necessary specificity;

[7]There are a number of sources for thesaurus construction standards. See International Organization for Standardization, *Documentation: Guidelines for the Establishment and Development of Monolingual Thesauri* (ISO 2788, 1986); British Standards Institution, *Guidelines for the Establishment and Development of Monolingual Thesauri* (BS 5723, 1987); American National Standards Institute, *American National Standard Guidelines for Thesaurus Structure, Construction, and Use* (ANSI Z39.19, 1980).

[8]These relationships are defined in the various thesaurus construction standards. A good synthesis of the portions of the standards concerning term relationships appears in Lancaster, *Vocabulary Control*, Chapter 6.

[9]Ibid., 151; Helen M. Townley and Ralph D. Gee, *Thesaurus-Making* (London: Andre Deutsch Limited, 1980), 113.

Figure 1

Sample Thesaurus Entry

Media	
SN	Use for the media in general. Use Broadcast Media or Press if the material specifically concerns those forms of media.
UF	Mass Media
	News Media
	Press Organizations
BT	Communications
NT	Broadcast Media
	Motion Pictures
	Press
	White House Press Corps
RT	Advertising
	Freedom of the Press
	Presidential News Summary
	Press Accreditation
	Press Conferences
	Press Interviews
	Press Releases
	Public Relations
	Records and Tapes
	White House Briefings

Note: The codes used in this entry are the standard thesaurus notation: *SN* for *Scope Note*, *UF* for *Used For*, *BT* for *Broader Term*, *NT* for *Narrower Term*, and *RT* for *Related Term*.

3. the need for more cross references and/or scope notes to lead users to correct postable terms;
4. a few topics overlooked during thesaurus design;
5. minor incompatibilities in form and punctuation between the PRESNET software and the thesaurus due to their simultaneous development.

Since 1986, two major thesaurus revisions and periodic smaller changes have greatly improved the PRESNET Thesaurus. The revisions consumed virtually as much staff time as the original construction, although the time spent revising it would have been significantly less had the first edition followed thesaurus construction standards more closely. Updating any thesaurus is, however, a continuing process. No matter how much time and money is invested in thesaurus design and construction, additional resources for revision will be necessary as flaws are revealed or as new topics develop.[10]

In 1988 the PRESNET Thesaurus became a building block in the development of a larger thesaurus. The National Archives enlarged the PRESNET Thesaurus through the addition of historical terms covering earlier periods in American history and also modified a small number of existing descriptors. PRESNET is now using this new National Archives Subject Reference Authority List, which may also be

[10]Dryden, "Subject Headings," 178 and 180; Soergel, *Indexing Languages*, 457. The eleventh edition of the *Thesaurus of ERIC Descriptors* (Phoenix: Oryx Press, 1987) reports adding 224 descriptors and 190 *USE* references plus modifying several hundred scope notes and cross references since the tenth edition was published in 1984.

employed with other automated description systems being developed throughout the National Archives and Records Administration.

Administration and Use of PRESNET and its Thesaurus

The creation of a quality thesaurus is a major factor in providing good subject access, but several equally important factors relate to its use. Oversight of indexers' work, indexing policy, and the indexer's personal "knowledge base" can greatly affect indexing quality. One must always keep in mind that high quality indexing allows for better retrieval in searching.[11]

Soon after the installation of the PRESNET prototype, the system administrator recognized the need to appoint a "Thesaurus Czar." This individual handles thesaurus modification, answers questions about specific indexing problems, and reviews descriptor usage by the staff.[12] Over time the thesaurus overseer has developed guidelines to help the staff recognize indexable topics, rules to promote inter-indexer consistency, and suggestions on the number of descriptors to be assigned to descriptions of the various collection-hierarchy levels. A formal indexing policy incorporating all of these elements is currently being drafted.

The indexer's personal "knowledge base" is also important. Some experimentation with having a student employee assign descriptors confirmed the need for indexers to be experienced professional staff members. Indexers must have a detailed knowledge of the archival institution's holdings and the needs of the research clientele in order to recognize indexable topics consistently.[13]

Because the retrieval of information from the PRESNET database is rather complex, the Ford Library staff currently conducts all database searches for researchers. Search strategies are based on orientation interviews, search worksheets, and reference letters. Future plans, however, call for a terminal in the research room and the design of simplified search screens to allow researchers to conduct their own searches.

Although subject descriptors are a primary searching tool, full-text searches can be conducted on all description fields, of which there are thirty-nine at the collection level, fourteen at the series level, and ten at the folder level. Searchers enter queries in the form of phrases consisting of a field name, a relational connector (*equal to*, *not equal to*, *greater than* and *less than*), and a value. For example, to search for materials with the subject descriptor *Education* the phrase to employ is *SUBJ=EDUCATION*. Phrases can be combined with Boolean connectors (*and*, *or*, *xor*, and the modifier *not*) into more complex search strategies. The software also provides for the use of wild card characters, term truncation, and proximity searching.

System testing and debugging, thesaurus modification, and drafting policies and procedures consumed a great deal of time during the first two years. Even so, about 40 percent of the Library's open materials are now described in the database, including such key collections as the Presidential Handwriting File, the White House Central Files Subject File, and the files of Chief of Staff Richard Cheney and the Congressional Relations Office staff. Most of these collections contain some good materials on

[11]Townley and Gee, *Thesaurus-Making*, 113; Michelson, "Description and Reference," 194 and 196.

[12]The need for a staff member to perform such roles is also emphasized in Adele M. Newberger and Paul M. Rosenburger, "Automation and Access: Finding Aids for Urban Archives," *Drexel Library Quarterly* 13 (October 1977): 56-57.

[13]This point is also made in Lancaster, *Vocabulary Control*, 3.

a wide variety of Ford administration subjects, rather than large concentrations of information on a few narrow topics. The PRESNET database currently contains over eighteen thousand completed folder, series, and collection description records for more than 1,700 linear feet of material in about fifty manuscript collections. These description records employ an average of about 2.5 descriptors per folder record, five per series record, and six per collection record for a total of more than thirty-five thousand subject descriptors assigned.

During almost four years of experience with searching the PRESNET database, subject descriptors have proven to be very useful and have become the primary searching tool. Much of the research conducted at the Ford Library is subject oriented, so title or description free-text searching is employed primarily to supplement or complement subject descriptor searches. Researchers and staff frequently comment on the importance of PRESNET searches in identifying materials that they never would have found, even with a thorough examination of the traditional typed finding aids or an automated search on folder titles.

Lessons from the Ford Library's Thesaurus Experience

Designing and using the PRESNET Thesaurus has been an enlightening experience. Although the presidential libraries are unique in some ways, there are lessons to be learned from the development of this subject index system that apply to a wider range of archives:

1. Controlled-vocabulary subject terms can help to overcome inconsistencies and inadequacies of titles assigned by collection creators and changes in the terminology applied to an issue over time.
2. A thorough study of thesaurus construction literature, especially the national and international standards, is necessary before beginning construction of a thesaurus. Failure of the first edition of the PRESNET Thesaurus to adhere fully to thesaurus construction standards led to its extensive revision after usage began.
3. The time expended on thesaurus construction can be decreased by basing it on existing topic lists and classification schemes. Use of the White House Central Files filing manual as a source of terms eliminated the necessity of time-consuming test indexing to generate a list.
4. Thesaurus construction requires a significant allocation of resources, both for construction and subsequent updates. PRESNET Thesaurus construction and revision consumed the equivalent of almost a year of work by a single archivist.
5. No matter how thorough the original thesaurus construction work, a thesaurus is never complete. It must be revised periodically as new topics develop and flaws are revealed.
6. Inter-indexer consistency is important in improving the quality of retrieval, and can be promoted by developing an indexing policy and closely supervising all indexing work.
7. Indexing should be done by experienced archivists and not by clerical staff or students employees. A thorough knowledge of the institution's holdings and the needs of the research clientele improves indexing quality.

Some day the archival profession may have one or more broad thesauri suitable for large national databases. Until then narrow, specialized thesauri can serve a useful purpose, both as tools for indexing specific bodies of records and as potential building blocks in the development of a broader thesaurus.