

Development of a Computerized Keyword Index/Thesaurus
to the Model Building Codes

Henry J. Lawrence, Jr.

BISON Corporation
1921 Sunderland Place, NW.
Washington, D.C. 20036, USA

KEYWORDS

Building Codes, Keyword Index, Thesaurus

ABSTRACT

Lack of a consistent language among building codes in the United States makes it difficult to locate all relevant provisions for certain topics. BISON Corporation recently completed a research project under a grant from the National Science Foundation, to develop a controlled vocabulary which will serve as a first step toward making code information accessible.

The primary objectives of this research project were to:

- o develop a preliminary keyword index/thesaurus to the model building codes;
- o design a model for information access based on similar provisions from the codes;
- o test the information access system for use by design professionals and building officials; and
- o seek input from a multi-disciplinary review panel for refinement of the scope and objectives for Phase II.

Research methodologies included state-of-the-art text scanning techniques and manual classification of keyterms.

Demonstrated was the feasibility of developing an online thesaurus/keyword index to facilitate building codes searches.

La Développement d'Un Index Computerisé des Mot-Clés
Pour Les Codes De Construction

Henry J. Lawrence, Jr. Esq.

Bison Corporation

MOTS-CLÉS

Codes de Construction, Recherche du bâtiment, Index des Mots-Clés, Trésor

SOMMAIRE

À défaut d'un langage compatible entre les codes de construction aux États-Unis, c'est difficile à trouver les mesures prises à propos par les sujets spécifiques. BISON Corporation a fini récemment un projet de recherche avec une allocation du National Science Foundation. Ce projet a développé un langage contrôlé qui servira comme le début pour mettre à la disposition d'information des codes.

Les objectifs premiers de cette recherche étaient:

- développer un index des mots-clés préliminaire pour les codes de construction modèles
- imaginer un modèle pour l'accès à l'information fondé sur les modèles ressemblant
- éprouver le système pour l'accès à l'information pour être utilisé par les professionnels de l'architecture et la construction
- chercher les réponses d'un tableau multidisciplinaire pour l'affinage des objectifs et l'entendue de Phase II.

Les méthodes de recherche ont compris "text-scanning" courant, et la classification manuel de mot-clés.

La praticabilité de la développement d'un index computerisé des mots-clés était démontré.

1. PROBLEM STATEMENT

A. Model Building Code Overview

Building codes have come a long way since the 1800s when most jurisdictions in the United States had limited building regulations to protect the health and safety of building users. The situation has improved greatly, however, with the introduction of model building codes. Model building codes are developed with the intent that jurisdictions will either adopt them as their local building requirements or use them as the base to develop their own codes. They do not carry the force of law until they are adopted by states or local jurisdictions. Most jurisdictions do not have sufficient finances to develop their own codes, hence, use of model codes is widespread.

Currently there are three organizations that develop model building codes: the International Conference of Building Officials (ICBO), that publishes the Uniform Building Code which is used primarily in the western U.S.; the Building Officials and Code Administrators (BOCA), publishers of the Basic/National Building Code which is used in many of the northeastern and mid-atlantic states; and the Southern Building Code Congress International (SBCCI), publishers of the Standard Building Code which is used mainly in the South.

Although the development of model building codes has contributed to a more uniform building regulatory system in the United States, inconsistent code development, interpretation and enforcement remains a major problem. Consistency among codes is difficult to achieve because a rapid turnover of information is necessary to keep pace with changing building technology. Every year an average of three hundred changes is proposed to each of the three model building codes, which influence much of the construction in this country. These changes are proposed by anyone concerned with problems, such as builders, manufacturers, architects, engineers, and plumbers. Therefore, the level of technical expression varies among provisions.

B. Development of a Controlled Vocabulary

Expanded application of computer technology to manage and locate information in building codes and regulations is needed. Though searching for code provisions in a single code can be achieved by scanning the computer-based code text with available software, it is not a viable method to address the language differences within a code or among several codes. Searching the codes manually can be an exasperating experience because no common format exists and the code indices are inadequate as a searching tool. Therefore, a cross-reference thesaurus is needed.

C. Problems Presented by Inconsistencies

The lack of a consistent language among the codes creates several

potential problems. It can be a major problem for the architect or engineer who has to use two, three or more codes as minimum design guidelines. Domestic and foreign construction product manufacturers have problems marketing products on a national scale because technical requirements for their products may vary from region to region, or even among cities. This lack of a common language can lead to impediments to the technical comparison of code provisions, increased costs for building design and construction and regulatory and economic barriers to the introduction of new technologies.

2. RESEARCH METHODOLOGY

The main component of this research phase was to investigate the feasibility of developing a preliminary keyword index/thesaurus for building code searches. The keyword index/thesaurus is of an "explicit" nature in that it is developed from keyterms that appear in the code text. In short, Phase I was merely a "trial run" version of Phase II in that Phase I entailed development of a preliminary thesaurus by indexing a section common to the major model codes, creation of a database, and testing by a group of experts. Phase II will then entail expanded development of the thesaurus. Working with us is William Brennehan of Computer Systems Resource, and Howard Coleman, of Charlottesville, Virginia.

The Phase I project consisted of six major tasks:

A. Preliminary Research and Development

Our first task was to choose code sections to include in the database. The major problem we encountered was the need to limit the amount of material included in Phase I, to stay within the National Science Foundation contractual time limitations.

To keep Phase I focused on our objective of using text scanning techniques on building codes, the conclusion was to approach the project with reference to a particular "building component" that would yield sufficient data for a study and illustrate the use of a thesaurus/keyword indexing method to show inconsistent code terminology. We chose "building foundations" and concluded that this would give us a usable database size with major differences in code terminology.

B. Development of a Keyword List/Thesaurus for Language Consistency

Code text making up the first database included a total of approximately 90 pages on "foundations" from the three model building codes. After the text was keyed into the database, a number of computer scans were made. Next an exclusion list was developed to delete from the text those words that were not needed to develop a keyword list. Our lexicographer examined those and selected words to be added to the exclusion list.

We then listed all of the different words in the code text which were not in the preliminary exclusion list. Once the keywords were chosen, an explicit index of the foundation sections of the codes was generated by indexing software.

C. Database Design and Implementation

1. Conceptual Design

One of the goals of this project was to design a computer database for accessing and comparing the information in the three model building codes. The system as designed and implemented enables access by means of the keywords identified by the methodology described above. The information accessed is displayed on a terminal as either groups of keywords describing a section of the code or the actual text of a section.

The strong point of the database is that because of the characteristics of the design and quality of the data developed using our methods, a complete search of all explicit occurrences of requested keyterms is guaranteed. In addition, the online thesaurus gives access to implicit references (e.g., "timber" is to "wood" as "rat" is to "rodent").

2. Implementation

For this project the system was implemented on a Texas Instruments 990 mini-computer under its DX10 operating system. The data dictionary and data management programs are written in OLADS (a high level, COBOL oriented language). The search algorithms are written in FORTRAN and can be called from OLADS. The design does not depend on specific hardware or these languages for its implementation.

3. System Design

The screen interactive inquiry system gives access to model code sections by keywords which can be combined in Boolean logic to target search inputs. Certain capabilities were specifically added to facilitate comparison of the codes.

The data structure consists of text which is split into small logical modules that can be easily displayed and comprehended on a screen. These modules correspond roughly to subsections of the codes and are maintained in a random access file to facilitate presentation. The keyterm file is used for display and is cross-referenced to the text by a section identification number. A variation of the keyterm file is stored separately for use by the search algorithms.

A vertical split screen format enables two separate sets of information (i.e., comparisons of related provisions from two different model codes) to be presented simultaneously.

The system provides the capability of searching the codes through the use of keyterms and presenting the results on a screen.

4. System Features

Searches can be performed by entering a keyword instructing the system to find all model code sections which contain that keyword. Whenever a keyword is entered it is checked against the list of valid keywords. The system will inform the user if it is not a valid keyword. If the word is an inflected form (e.g. plural) of a valid keyword the system will replace the entered word with the proper inflected form, and notify the user that this has occurred.

The user can also enter a search strategy and the system will find all code sections which satisfy that request. A search strategy is a way of looking for several keywords in some combination. Once the system finds all the sections that satisfy the search combination requested it will tell the user how many it found in each model code. The user can then request a list of the HITS (sections which satisfy the search request), and then select any one of the HITS and display all of the keyterms which occur in that section. If it looks to the user like he is on the right track, he can then display the actual text of the related code sections from two model codes for comparison.

An alternative way to use the system is to enter a section number which is in one model code, and the system will display all of the keyterms which are in that section. The user can then select terms from the displayed list to define a search strategy by simply moving the cursor to the word and pushing a key. From that point the system acts as described above and searches all of the model codes to satisfy the request.

For comparison of code sections, various combinations of the half-screens can be displayed simultaneously through the use of console function keys. A copy of the screen can be stored for later analysis or sent directly to a printer upon command.

D. Technical Review Panel

A technical review panel composed of engineers, architects and building officials met at the onset of the project to provide guidance and suggest directions. The same group reconvened later in the project to discuss findings, future plans, and test the system.

We reaffirmed our goal which is not to rewrite what is already in the codes, but to just make access to the information easier. Ease of access will result in tools to show the technical differences among the codes and encourage more consistent and uniform development of codes nationally.

E. Test Projects

During our second meeting members conducted two hands-on tests of the system. In the first test architects and engineers were able to query the database with a personal computer telephone linkup from our meeting site in Washington, D.C., to the mini-computer in Charlottesville, Virginia. In the second test a building official conducted several searches.

F. Future Plan Development

Plans for Phase II are currently being developed. The focus of Phase II will be on an expanded model thesaurus and database.

3. FINDINGS AND RESULTS

A. Overall Findings

All five research objectives were met. We sought to:

- o design a model for information access;
- o develop a preliminary keyword index/thesaurus;
- o test it for use by design professionals and building officials;
- o seek input from a multi-disciplinary panel of experts; and
- o refine the scope and objectives for Phase II project development.

B. Indexing Methodology

Our indexing method consisting of state-of-the-art text scanning techniques and manual classification of the keyterms is appropriate to the task. Based on the recommendations of our review panel

Phase II will include an attempt to introduce a consistent classification system such as the industry-wide UCI masterformat used to organize specifications and product information. This may also help to transfer the research to other areas of design and construction.

C. Database Design

The system illustrates the usefulness of a thesaurus and online search techniques to manage and retrieve building code information.

D. Technical Implications

The technical implications of this research are many.

A model building code thesaurus with computerized access will demonstrate how an existing technology can manage code and regulatory information.

E. Technical Feasibility

This research phase demonstrates that development of consistent language for the model building codes is technically feasible. State-of-the-art text scanning methods, coupled with programming ingenuity and technical input can be a first step in achieving a more efficient building code system in the United States.

REFERENCES

Basic/National Building Code, Building Officials and Code Administrators, Country Club Hills, Illinois. 1984

Basic Reading Vocabularies, MacMillian, 1983.

"Building Regulation in the Twenty-First Century", Research and Innovation in the Building Regulatory Process. Milton, Hans J. NBS Special Publication # 518. 1978

CMEA Dictionary of Civil Engineering in Twelve Languages, Council of Mutual Economic Assistance, Moscow, USSR. 1983

"Common Format for the Model Building Codes: An Application of Advanced Techniques for Standards Analysis, Synthesis and Expression", Stahl, Fred I.. Research and Innovation in the Building Regulatory Process. 1984.

Developing Computer-Based Systems for Building Codes, A report on an Exploratory Research Effort, AIA Research Corp. 1974

Dictionary of Architecture and Construction, Stein, McGraw-Hill. 1975

Estimating Economic Impacts of Building Codes. NBS Special Publication #81-2402. 1981

International Bibliography of Multi-Lingual Building Vocabularies, Economic Council of Europe, Committee on Housing, Building and Planning. (ECE -HBP-36) 19__.

Report of the President's Commission on Housing, Washington, D.C. 1982

Standard Building Code, Southern Building Code Congress, International, Birmingham, Alabama. 1982

Uniform Building Code, International Conference of Building Officials, Whittier, California. 1982