LECTURE 8.2b READING 1

This is the seminal reading on request-oriented indexing and it is still relevant today. Of course, today one would use computer searching rather than edge-notched cards (one of the few mechanisms to implement combination searching in 1954). This reading includes only those sections of the paper that deal with developing the index language and applying it in indexing. File created by OCR to Word

From Casey, Robert S., et al., ed.
Punched cards. Their applications in science and industry. 2. ed. New York, NY: Reinhold;
1958.

Chapter 15

A CASE HISTORY OF A ZATOCODING INFORMATION RETRIEVAL SYSTEM

CLAUDE W. BRENNER
Allied Research Associates, Inc. Boston, Mass.

AND

Calvin N. Mooers
Zator Company, Cambridge, Mass.

The Problem

A rapidly growing collection of research reports presented an acute reference problem to Allied Research Associates, Inc., Boston, Massachusetts, in 1954. This organization of engineers and scientists, doing research, engineering, and development in the aeronautical and physical sciences, had been expanding since its beginning in 1951, with three engineers on its stall. At first, personal files of reports were sufficient for the company’s information filing and retrieval needs. Later, project files were set up, and reports touching on the different projects were segregated into these files. As more contracts were undertaken, the engineering staff increased, and the rate of influx of technical reports and papers steadily mounted. By late 1951 the company had a staff of fifty, and the bulging files held about 3000 reports, with more coming in every day.

It was evident that the files would very soon become unmanageable.

[Creating a conventional card catalog would be labor-intensive and still not offer the search power needed. A new approach based on searching by combining concepts promised to solve the problem.

The Descriptor Dictionary System

The heart of the new system, its most important part and intellectual foundation, is the descriptor dictionary system. It is called a descriptor dictionary system because it is not merely a list of subject words. Instead, it comprises several different kinds of lists, each having a definite function. It is the intellectual tool that couples the mind of the information searcher to the hardware of the Zatocoding system in such a manner that the hardware does the work of selecting the desired subject matter from the file.

TABLE 15-1. A portion of the alphabetically arranged scope notes. Descriptors are preceded by asterisks. Terms not descriptors (n.d.) are cross- referenced to descriptors.

**Stability* 135 12-11:39-35

In aeronautical engineering, pertains to the study of aircraft stability as used in conjunction with *Static, *Dynamic, *Lateral, * Longitudinal. Also refers to instability, such as buckling or other structural instabilities. Use with * Derivatives instability and control studies. For Lateral-longitudinal Stability Coupling (n.d.), use *Stability plus *Lateral plus *Longitudinal plus *Interference.

**Stall and Buffet* 136 38 16: 2 6 - 7

Stall pertains to the condition of partially or wholly separated (low on air foils at high angles of attack. Buffet is the disturbance due to periodic boundary layer separation on a surface or the motion of a surface in a fluctuating wake.

**Static* 137 38 3 :2 8 -2 5

With *Stability, pertains to static stability studies.

Statistical Mechanics (n.d.)

Use *Thermodynamics

Statistics (n.d.)

Use * Probability.

Stick Force (n.d.)

Use *Control plus *Biology.

Strain Gaye (n.d.)

Use *Stress and Strain plus * Instrumentation.

**Stress and Strain* 138 15-3: 33 8

Any process involving the loading and deflection of structures, e .g., bending of beams, deflections of plates, theoretical elastic y studies, elastic behavior. Use also for Torsion (n.d.). With *Instrumentation it means Strain Cage (n.d.).

A descriptor is something like a “subject heading” in library practice, though it is usually much broader in meaning. For instance, a subject heading might be “oils—–effect of temperature on viscosity.” In descriptor analysis, the separate descriptors “oil,” “thermal,” and “viscosity” would be used together to delineate this meaning. Each descriptor is a word- symbol standing for an idea or concept, generally of a rather broad scope. The particular scope of meaning for a descriptor is *assigned* in such a way that the descriptor will be most useful for retrieving information in a specified collection. Thus, the assignment of meanings at Allied Research is in part quite different from those assigned in other Zatocoding systems. Retrieval meanings need not conform strictly to standard technological usage of the word chosen to be the descriptor symbol. Because the meanngs are often slightly different from the ordinary usage, it is essential that the descriptor dictionary system include a list of “scope notes,” with a scope note for each descriptor. An alphabetically arranged list of scope notes such as shown in Table 15-1 then makes the full range of assigned meanings easily accessible to anyone desiring to use the Zatocoding system. These special descriptor meanings are private, for use in retrieval only, and there is **110** intent (nor likelihood) of imposing them upon ordinary speech or technical writing within or outside the company.

Deriving the Schedule of Descriptors

The schedule of descriptors is the most important component of the dictionary system. At Allied Research, a panel of four top engineers and physicists worked together in deriving a schedule of descriptors. With this panel of top personnel, problems of scientific and company policy as they affected the future use of the retrieval system could be settled on the spot. Thus, in areas where the company expected to embark on a new line of endeavor, the group was anxious to make sure that appropriate descriptors were obtained.

Deriving the descriptors is a strictly empirical process. On four separate occasions the Allied Research panel met with the Zator representative. A stack of reports, giving a typical sample of their file, was brought out and placed on top of the conference table. The top report was taken, its title and abstract were read to the group, and it was passed around for a brief examination of its contents. Then the question was posed, “Why would anyone at Allied Research be interested in using this report?” The answer may have been that it was about *propellers*, that it was about propeller *aerodynamics*, and that it was a *wind tunnel* study. Each of these was taken as a presumptive descriptor and written down. The same empirical process was followed with the next report, and so on. On more than one occasion, sad experience has given convincing proof that descriptors “dreamed up” in an armchair without reference to actual reports are worthless.

Table 15-2.

A Portion (about one-quarter) of the descriptor schedule at Allied Research Associates, Inc., showing the grouping of the descriptors and the way in which leading questions are asked

What material was studied?

Metals
Gases
Plastics
Aluminum
Magnesium
Titanium
Air

Is the process dynamic (rather than static)?

Vibrations
Transient response
Impact
Stability
Velocity

***Are there specific aerodynamic lo*ads?**

Lift
Dra g
Moment
Gust Pressure
Center of application
e.g., aerod. center, center of pressure, etc.

Is structural strength and elasticity involved?-

Stress and strain
Plasticity
Failure
Ultimate properties
Material properties
Aeroelasticity
Flutter

What is the type *of fluid* flow-*?*

Fluid flow
Internal flow
Subsonic
Transonic
Supersonic
Hypersonic
Laminar
Turbulence
Slip flow
Compressibility
Viscosity
Vortices
Shock waves
Finite span

Is it a stability and control problem?

Stability
Control
Static
Dynamic = Trans, resp.
Longitudinal
Lateral
Derivatives
Damping
Weight and balance
e.g.. center of gravity, moments of inertia, etc.

Or is there another aerodynamics problem?

Boundary layer
Aeroelasticity
Flutter
Downwash
Stall and buffet
Interference
Hydraulics
Trajectory
Droplets
Modifying Technique
Performance

Is a thermal process involved

Thermodynamics
Thermodynamic constants
Combustion
Cooling
Convection
Conduction
Thermal
Radiation
Aerodynamic heating

This empirical procedure of discovering descriptors is surprisingly rapid. By the time that some fifty reports (selected to give an approximate cross-section of the company's interest) had been processed, more than 80 per cent of all the descriptors in the final schedule had been found.

In this stage of developing their system, the panel had many lively discussions about the theory and practice of using descriptors in information retrieval. These discussions were encouraged by the Zator representative, and the various points raised provided an excellent opportunity to bring up the? experiences of other Zator clients who had similar problems.

At the second meeting of the panel, about a week after the first meeting, the descriptors obtained so far were written down on a large sheet. This was the first draft of the descriptor schedule (see Table 15-2). Related descriptors were grouped together in the draft; duplicate descriptors were eliminated.

Additional reports were then analyzed in the same way. Now the panel began to use the draft schedule as a guide. A few more descriptors were added, and rough spots in the draft schedule were ironed out. During this stage, scope notes were being written on index cards (for later typing in list form). Decisions made by members of the panel about the usage of the descriptors were thus written down while the problem was fresh **111** their minds. At various times the Zator representative would ask questions or offer criticisms to make sure that the panel was aware of the consequences of their decisions. Except for the teaching and guidance of the Zator representative, the panel did all the work in deriving their schedule of descriptors.

The panel at Allied Research put in a total of less than 150 man-hours from the beginning of the process until the schedule was ready to hand over to their clerical staff for typing and code assignment. This time included the "homework" that was assigned to the various panel members between visits of the Zator representative.

During the entire operation of deriving the descriptors, it was stressed that the primary orientation of a retrieval system must be toward the requirements of the user. One of the most important consequences of user orientation is that the descriptors must be broad in meaning. When the descriptors are broad, the user's intellectual universe can be covered by a relatively small list of descriptors. At Allied Research, 250 descriptors are used. Because there are so few descriptors in the system, they are relatively easy to remember, which is a definite advantage. The very breadth of meaning of each descriptor makes it easy to decide its applicability to a given document. Descriptors with finely drawn distinctions between them are avoided. Precision is not lost b}^{r} using broad descriptors because ideas can always be synthesized by means of several descriptors.

With so few descriptors, it is easy to set them down on one big sheet, called the descriptor schedule. In this way, the analyst is able to see all the descriptors at once.

At Allied Research, the booklet containing the scope notes (alphabetically arranged and printed), and the large sheet which is the descriptor schedule, are distributed to the engineers who are most active in using the system or who are on the team analyzing the incoming reports. To aid further in finding the correct descriptors, the scope notes have interpolated words and expressions in ordinary technical usage with cross-references to the proper descriptor.

Analysis of the Incoming Documents—The Filtering Technique

When the incoming reports arrive at Allied Research's document center, they are given a preliminary screening to determine which engineer analyst is to handle each report. About sixteen engineers and scientists are on the analytical team. Each person gets the reports most closely related to his specialty. This procedure has the added advantage that it also keeps the specialists cognizant of the latest work in their fields.

The procedures adopted for document analysis are also user oriented. No attempt is made in analysis to code the message of the document by writing a little abstract using descriptor words. The descriptors and their codes are used for retrieval only, and the message itself will always be available in the document. Neither is there an attempt to secure pin point precision with the descriptors. Excessively narrow descriptors will only frustrate the user when he attempts retrieval.

The user of a retrieval system has a difficult problem, he is confronted by nothing but a schedule of descriptors supplemented by the scope notes. He is not sure what the file contains. he frequently knows nothing about the finer details in the reports. Thus, with only the schedule and scope notes, he must be able to formulate a prescription that will retrieve information, the nature of which may in large part be unknown to him. His success will depend largely upon how well the analysts originally did their job.

In conformity to the philosophy of user orientation, the document analyst is asked to place himself in the user's position. He does so in this way. First he reads or skims over the document. Then he lays the document aside and concentrates upon the descriptor schedule. He works down the schedule, descriptor by descriptor exactly as if it were a check list. For each descriptor he asks, "Would anyone at Allied Research who is interested in the content of this document use this descriptor as a part of his retrieval prescription?" or, "Does the meaning of this descriptor touch in any way upon the message of the document?" If the answer is "yes" to any of these, the descriptor is chosen as one of those to characterize the document

This is known as a “filtering” technique, and according to this technique, the schedule of descriptors is filtered through the message of the document. Those that remain in the filter are the chosen descriptors. If there is any doubt about the applicability of any descriptor it is resol-ved by choosing the descriptor. A doubtful descriptor may be just the one that will be tried in a retrieval prescription by some eventual user.

This technique has proved to be invaluable in giving the retrieval system a consistent intellectual structure. Consistency is a real problem. There are as many as sixteen or more contributing analysts at Allied Research and this group continues to change over the years. Yet their efforts accumulate in the form of the Zatocard collection. These cards must be consistent to be usable, and rigorous application of the filtering technique has forced internal consistency of the system.

The filtering technique also has another advantage, it does not require the analyst to have a highly technical background. If there were no filtering method, heavy demands would be placed on his ability and imagination. I le would have to foresee all the possible uses of the document in order to decide which descriptors would apply. This is very difficult. However, the schedule of descriptors almost eliminates this problem because it serves as a check list of present and future contingencies as worked out by the top people in the laboratory. When the analyst uses the schedule as a check list, he only has to make very simple decisions.

The burden of using a schedule of 250 descriptors is eased by a simple process. About one-quarter of Allied Research’s descriptor schedule is shown in Table 15-2. The descriptors are grouped, with each group being composed of similar descriptors. At the top of each of the groups there is a question, such as, “Is there a type of fluid flow?” In using this kind of a schedule, the analyst first looks at the questions. If the answer to any of them is “yes,” then he picks out the one or more appropriate descriptors below the question. If the answer is “no,” he continues to the next question. Use of the filtering technique in the Zatocoding dictionary system involves going through a list of about 20 questions rather than through 250 individual descriptors. Carefully chosen “leading” questions, as in this example, make the incoming document analysis particularly easy.

This grouping of descriptors is not a scheme of hierarchal classification. There are no generic or specific terms. Any descriptor can be used with any other, and more than one descriptor from a single group can be used to characterize a document. A typical document in Allied Research’s collection has from six to fifteen descriptors in its characterization. It is sometimes convenient to place the same descriptor in two different groups. This is useful for a few of the descriptors that may appear in widely differing contexts. An alphabetically arranged list of descriptors is specifically not used for the analysis of documents because of its inferiority to the grouped arrangement in providing accurate analysis.

An actual analysis proceeds as follows. The first decision of the analyst is whether or not to include the document. Obviously worthless material must not be allowed to increase costs or to dilute the system. By the time the analyst sees the document, it has passed this threshold of utility. The analyst then skims or reads the document. Depending upon the obscurity of the writing, or the richness of the content (there is often an inverse correlation), this usually takes from 5 to 25 minutes. Fifteen minutes is not a pessimistic average for technical reports. The analyst then takes the descriptor schedule and reviews the check list of questions, writing down the chosen descriptor words on a Zatocard. This step of filtering and writing down the descriptors takes about two minutes. The card then goes to (he clerical staff).

The chapter goes on to discuss more detail not of interest for LIS 571