Computer-Aided Thematic Analysis™: Useful Technique for Analyzing Non-Quantitative Data

by M. E. Kabay, PhD, CISSP-ISSMP
mkabay@norwich.edu
Program Director, Master of Science in Information Assurance (MSIA)
School of Graduate Studies
Norwich University, Northfield, VT

Version 21.

Abstract

A method using a computer and any program providing sort functions can help anyone trying to make sense of large amounts of qualitative information.

1 Introduction

All writers face the problem of organizing written material, whether scientific papers, reports on competing products or business strategies, literary criticism, or their own insights.

It was to solve the problem of ordering information without a pre-existing analytical framework that I stumbled on the following method called *Computer-Aided Thematic Analysis™* or CATA™ . CATA is usable by anyone with a computer and any software that can sort lines of text; e.g., spreadsheet programs, databases and many word-processing programs.

In preparing this brief report, I found reports of similar ideas and methods in the field of ethnography. Although using techniques very similar to CATA, the ethnographic software is highly specialized and the methods poorly known outside the field. CATA is a technique that is easily applied and accessible across disciplines. The use of techniques similar to CATA by ethnographers suggests that it is an idea whose time has come. This simple computer application promises to help the organization of knowledge in any field to which it is applied. It is particularly helpful to anyone involved in a literature review, such as researchers and also students writing term papers.

2 How CATA Developed

In 1987, I was asked to evaluate the status of a research project in a discipline about which I was ignorant, community development.[1] I was selected because my lack of experience would supposedly help me avoid bias and preconceptions.

I interviewed 55 people for this analysis. During the interviews, I took notes on a portable computer. At the end of the data collection process, I had 450 pages of single-spaced transcript.

How was I to make sense of this daunting mass of qualitative information?

Following the interview series, I extracted brief themes from each interview. For example, the following paragraphs are the from the original notes for part of an interview:

> There are two areas I find weak: economic development and political development. I sometimes wonder if these are ignored on purpose. In practice, we find they're indispensable. Political and economic development return self-esteem and independence to the people. In our experience here, these are the most important aspects right now, while cultural and social development are taking place.
>
> The other programs: Elders' Program. I have talked about his with Xxxx. I still have many doubts there from the medical end. There's something missing there. From the nutritional end, also, I have to discuss things with him (he's the nutritionist, not me). And also there's the cultural thing again. But focusing on the elders is a wonderful idea, beyond doubt.
>
> Spirit of the Rainbow: excellent. No doubt about them. On personal note, however, whenever I've seen them, they seem overtired. Their lives don't seem balanced for their age. Their responsibilities seem huge, and I'm afraid they would wear themselves out. And they're outstanding, they're excellent.

I extracted the following themes from this section of the interview and typed them into a spreadsheet program with one line per theme. In a column assigned to represent the particular interview, I marked the page number where the theme occurred so I could find the reference quickly.

```
====================================================================
THEME                            Interview #
--------------------------------------------------------------------
weak: economic & political dev't                            43
wonder if economics & politics left out on purpose          43
political & economic dev't important here now               43
elders' program: doubts from medical end                    44
but elder focus wonderful               44
spirit of rainbow: excellent.  no doubt about them          44
s.o.r. staff seem overtired                                 44
s.o.r. responsibilities huge; could wear themselves out            44
====================================================================
```

Thematic extraction from the entire 450 pages of transcript resulted in 923 themes and 15 pages of printout. Next, I identified a few general areas into which the themes seem to fall. These categories, in no particular order, were numbered as follows:

1	ORGANIZATION
2	MATERIALS
3	APPROACH
4	HOSTILITY

I next classified each of the 923 themes using numbers assigned to these crude preliminary categories. For example, the themes above were classified as follows:

```
=================================================================
3   weak: economic & political dev't
3   wonder if economics & politics left out on purpose
3   political & economic dev't important here now
2   elder's program: doubts from medical end
2   but elder focus wonderful
2   spirit of rainbow: excellent.  no doubt about them
2   s.o.r. staff seem overtired
2   s.o.r. responsibilities huge; could wear themselves out
=================================================================
```

The breakthrough came when the 923 themes were sorted by category. All the themes tagged with the same category code were instantly brought together into the same area of the list. Simply by inspection, a new level of classification seemed reasonable among the categories. For convenience, these new categories were identified by adding a digit to the original numerical code. For example, the themes dealing with organization were all tagged with the number "1"; they seemed to fall easily into the following subdivisions:

10	ORGANIZATION
11	STAFF
12	FUNDING
13	EFFICIENCY
14	SUGGESTIONS.

All 923 themes were thus numbered with this new level of precision. Thematic analysis continued iteratively and recursively with more detailed classifications as required. For example, section 11, dealing with the staff within the organization, was easy to classify further as follows:

110	STAFF
111	LEADER
112	OPENNESS & FLEXIBILITY
113	COMMITMENT
114	PRACTICAL, DOWN TO EARTH
115	GENEROUS WITH TIME, MATERIALS
116	XX ARE ROLE MODELS.

By this point, if the themes from the passage shown above had been sorted together, they would have looked like this:

```
=================================================================
308      weak: economic & political dev't
308      wonder if economics & politics left out on purpose
308      political & economic dev't important here now
204      elder's program: doubts from medical end
204      but elder focus wonderful
211      spirit of rainbow: excellent.  no doubt about them
211      s.o.r. staff seem overtired
```

```
211       s.o.r. responsibilities huge; could wear themselves out
=================================================================
```

By the end of the cycle, themes were ordered into a comprehensive structure reflecting the many comments made by respondents. Throughout the sorting and resorting, the references to origin and location followed the themes. By the end of CATA, not only could I locate every reference at once, but I could see patterns showing clusters of interviews that discussed similar themes. The clustering helped me make more sense of the complex information collected.

Example from Neurological Research Paper

My wife, neuropsychiatrist Dr Deborah N. Black, wrote a review paper using CATA.[2] Here are notes from her spreadsheet during CATA on two pages from a particular research paper to illustrate how she used the technique:

paper & page	code	notes
a71	e	unknown whether transient paraplegia spinal or peripheral or vascular in origin
a71	h	great diversity in immed. effects shock individual threshold; amperage; ?state of brain--sleep, anesthesia...)
a72	b1	intracranial bleeding, oedema, single or multiple vascular lesions responsible for neurol. sequelae
a72	c1	ms-like picture
a72	c1	hemiplegia, aphasia, choreoathetosis (rare), headache, giddiness, insomnia, forgetfulness, epilepsy (rare)

Here is Deborah’s final set of categories.

A	GENERAL
B	PATHOLOGY
B1	BRAIN PATH
B2	CORD PATH
B3	PNS PATH
C1	CLINICAL SYNDROMES--BRAIN
C1.A	EARLY BRAIN SEQUELAE
C1.B	LATE BRAIN SEQUELAE
C1.B.1	PARKINSONIAN SYNDROME
C1.C	NEUROPSYCH SEQUELAE
C2	CLINICAL SYNDROMES--CORD
C2.A	EARLY CORD SEQUELAE
C2.B	LATE CORD SEQUELAE
C3	CLINICAL SYNDROMES--PNS
D	DNB
E	PATHOGENESIS
F	EPIDEMIOLOGY
G	PHYSICS
H	VARIABILITY OF OUTCOME
I	COMPARISONS TO OTHER CONDITIONS

This example illustrates the flexibility of CATA: there is no need for rigid standards of classification or notation. Furthermore, different sections can easily be subdivided to greater or lesser degrees.

3 Discussion

Language determines what we see and think. For example, Whorf [3] emphasized that ignorance of categories leads to their invisibility. Inuit people recognize many distinct categories of what more southern people call "snow;" mushroom hunters see gill mushrooms, bracket fungi, and boletes where the untutored see either "toadstools" or nothing at all.

In CATA, the initial step is to define working categories. These initial categories do not seem to be as important as simply beginning to group the data. This observation is consistent with evidence from experimental psychology that organizing information enlarges one's capacity to remember that information [4]. For example, arranging large numbers of words into categories enhances experimental subjects' ability to remember the words [5], and the larger the number of categories, the better the recall.

The next step in CATA is to bring similar ideas or observations into proximity. The contiguity of related information seems to foster further discrimination; it is easier to see patterns among ideas when one's thought is not interrupted by unrelated information. This impression of what happens in CATA is consistent with research showing that context influences perception. For example, ambiguous figures such as the well-known old/young woman drawing are interpreted as either an old woman or a young woman depending on which version of a less ambiguous rendition of the figure is shown first [6]. The context provided by grouping related information in CATA seems to spark recognition of hitherto-undetected patterns.

Why should the categorization inherent in CATA enhance our ability to work effectively with qualitative information? One factor may be our limited ability to hold in mind more than about 5 to 9 ideas simultaneously [7]. Faced with many more ideas than this practical limit, we tend to "chunk" the information into fewer but larger categories. "If we can pack the input more efficiently, we may squeeze more information into the same number of memorial units. The person who interprets the series 1 4 9 ... 81 as the 'the first nine squares' has done precisely this: he has recoded the inputs into larger units, sometimes called chunks. Each chunk imposes about the same load on memory as did each of the uncoded units that previously comprised it; but when eventually unpacked, it yields much more information" [8].

Ethnographers have for many years used a method described as "cut the interview into topical segments and sort" [9]. Specialized programs exist to help such social scientists classify and sort qualitative data [10]. The method described here parallels the evolution of ethnographic software but can be implemented by anyone having access to computerized sorting.

One additional note: when using a modern spreadsheet, it is easy to include extensive notes that are attached to specific cells. In these notes, which may contain pages of text, the user can insert extracts of quoted material and personal thoughts and comments about the theme indicated in that particular cell. If the user runs out of room, additional notes can be attached to cells in the same row corresponding to the points of origin of each theme.

4 Summary

CATA is a technique using simple and readily-available computer software to organize non-numerical information quickly. The core of the technique is iterative sorting and classification of notes. Sorting brings similar information together, allowing one to see relationships that are otherwise not evident. The iterative component allows progressive refinement of structure as additional categories intuitively become evident.

A computer allows one to sort and edit categories quickly. Such flexibility encourages experimentation, which may lead to unexpected juxtapositions that encourage new ways of thinking.

No external framework need be imposed upon the data. One does not need an a priori structure into which the data are forced; CATA lets one impose such a structure or not as one chooses. Reports virtually write themselves: their organization arises from the CATA categories and their content appears in the references.

This method will be useful to all computer-literate writers who wish to organize their thoughts, observations, and references in preparation for essays and reports [11].

5 END NOTES

[1] Five-Year Evaluation of the Four Worlds Development Project**.** University of Lethbridge.

[2] Deborah N. Black, MDCM, FRCP(C), Hôpital Louis-Hippolyte Lafontaine (Montréal); Vermont State Hospital (Waterbury, VT) and Central Vermont Hospital and Green Mountain Neurology (Berlin, VT).

[3] B. L. Whorf, in J. B. Carroll, Ed., *Language, Thought and Reality: Selected Writings of Benjamin Lee Whorf.* (MIT Press, Cambridge, MA, 1956).

[4] G. Mandler and Z. Pearlstone, *J. Verbal Learning and Verbal Behavior* **5**,126 (1966).

[5] G. Mandler, in *The Psychology of Learning and Motivation* , K. W. Spence and J. T. Spence, Eds. (Academic Press, New York, 1967). vol. 1, pp. 327-372.

[6] R. W. Leeper, *J. Genetic Psych*., **46**,41 (1935).

[7] G. A. Miller, *Psych. Rev*., **63**, 81 (1956).

[8] H. Gleitman, *Basic Psychology* (W. W. Norton, New York, ed. 2, 1987), p. 189; N. G. fielding and R. M. Lee, Eds, *Using Computers in Qualitative Research* (SAGE, London, 1991), pp. 4-5.

[9] M. Agar, in *Using Computers in Qualitative Research* , N. G. Fielding and R. M. Lee, Eds. (SAGE, New York, 1991), pp. 181-194.

[10] N. G. Fielding and R. M. Lee, Eds. *Op. cit*., pp. 195-199.

[11] The author thanks Drs Percy, Virginia and Deborah Black for helpful suggestions and editorial comments.