

# Open Coding

---

Shahedul Huq Khandkar

## Introduction

We need to give names to our ideas and concepts to define, analyze and share with others. Once it's defined, we can begin to examine them comparatively and ask questions to systematically specify the states and to imply possible relations with others. It's also important that we name our concepts appropriately; because "people act toward things based on the meaning those things have for them; and these meanings are derived from social interaction and modified through interpretation." [1]

To build concepts from a textual data source, we need to open up the text and expose the meaning, idea and thoughts in it. One of the processes of analyzing textual content is *Open Coding*. Open Coding includes labeling concepts, defining and developing categories based on their properties and dimensions. It is used to analyze qualitative data and part of many Qualitative Data Analysis methodologies like Grounded Theory.

## Qualitative Data Analysis

Qualitative Data Analysis (QDA) consists of three parts: Noticing, Collecting and thinking about interesting things [2]. QDA is generally a non-linear process and often can be recursive. As you continue on collecting information, you may notice new things and need to think about them. As a result, you sometimes have to go back to old data and analyze them again.

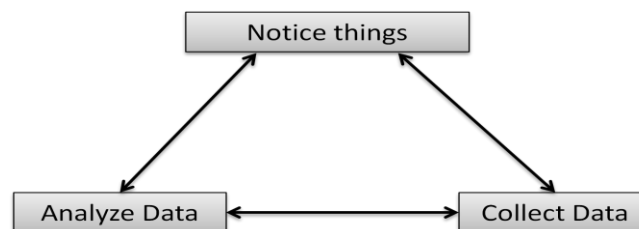


Figure 1: Workflow of Qualitative Data Analysis

In general, noticing means taking notes based on observation, recording events or interviews, gathering documents etc. In the analysis phase, when you are going through the data you often mark important sections and add descriptive name or 'code' to it. This is the first step of coding and called *Open Coding*. This article begins with a description of Open Coding and provides walk through on different open coding techniques.

## Building Concepts

The first step in qualitative data analysis is to go through the data (i.e. text) to break down in to pieces to examine closely, compare for relations, similarities and dissimilarities. Different parts of the data are marked with appropriate labels or 'codes' to identify them for further analysis.

A concept is a labeled section of data that a researcher identifies as significant to some facts that data represent. Concepts are abstract representations of events, objects, actions or interactions and they allow researchers to group similar information to better understand the data.

Concepts can be of various types; communication, storm or private company includes example of concepts. Concepts may incite certain natural imagery as they have their own properties. For example, we can think of data set representing telephone conversations between two participants and we can label them as 'Telephone Communication'. So a labeled thing is something that can be location and placed in a class of similar objects. Anything under a classification has one or more familiar properties or characteristics; like sending information is a property of communication. It is important to understand that concepts can be classified differently, it depends on the different properties of data the researcher is focusing on and how he/she is translating them.

## Abstracting the Concepts

As we continue to analyze the data by breaking down into distinct ideas, events or objects, we label any important information in the process. The name of the labels can be decided by the analyzer or can be taken from the content too. The later style is often called "in vivo codes" (Glaser & Strauss, 1967).

Naming Type	Description
In vivo codes	Wording that participants use in the interview
Constructed codes	Coded data from in vivo codes, Created by the researcher, Academic terms

Table 1: Different naming strategy for codes

While analyzing the data, we sometimes get events or objects with common characteristics yet other properties may separate them. We can use the common properties to group them under same concept. Different researchers may think of different names from the same data set but in general, it should be based on the context.

Following is a partial transcript of an interview with women in her early 20s and is about drug use by teens. The interviewer did not have preset questions to ask. He continued his questions based on the interviewee's response.

**Interviewer:** Tell me about teens and drug use.

**Respondent:** I think teens use drugs as a release from their parents [*"rebellious act"*]. Well, I don't know.

*I can only talk for myself. For me, it was an experience [“**experience**”] [in vivo code]. You hear a lot about drugs [“**drug talk**”]. You hear they are bad for you [“**negative connotation**” to the “**drug talk**”]. ...*

Source: **Basics of Qualitative Research**, Second Edition by Anselm Strauss & Juliet Corbin [6]

As you can see in the above interview transcript, we have grouped the similar information using abstract labels (i.e. drug talk). Some of the names for the labels are selected directly from the data (i.e. hard-core use). This process of going through line by line data to assign codes is called line-by-line coding.

## Notes for concepts or codes

Sometimes a name with few words is not enough to describe an entire concept. In such scenario, we can write notes against a concept that we call “Memo”. A memo can contain a paragraph or even more if needed. If we take a closer look in to the line-by-line coding example, the answer of first question of the interview has more meaning than we have expressed in the code. With memo, we can record that information like this:

**Interviewer:** *Tell me about teens and drug use.*

**Respondent:** *I think teens use drugs as a release from their parents*

**Memo:** *The first thing that strikes me in this sentence is the work “use”. This is a strange term because, when taken out of the context of drug taking, the work means that an object or a person is being employed for some purpose. It implies a willful and directed act. In making a comparison, when I think about a computer, I think about employing it to accomplish a task. I think of it as being at my disposal.*

Source: **Basics of Qualitative Research**, Second Edition by Anselm Strauss & Juliet Corbin [6]

Glaser (1978) offered guidelines for preparing effective memos to generate substantive theory including the following [3]:

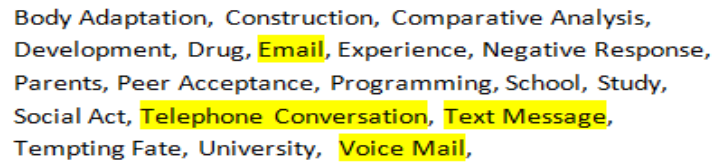
- Keep memos separate from data
- Stop coding when an idea for memo occurs, so as not to lose the thoughts
- A memo can be brought to you by literally forcing it, by beginning to write about the code
- When a lot of memos on different code appear similar, compare the codes for any differences that may have been missed. If the codes still seem the same, collapse to codes into one code.
- When you have two ideas, add two separate memos to avoid confusion.

Source: **Nursing research: principles and methods** by Denise F. Polit & Cheryl Tatano Beck

## Defining Categories

As we continue to create codes for new concepts, it’s not unexpected to come to a point when we will have more than few pages of codes. At that stage, we should analyze the codes to find the similarities and group them into categories based on their common properties. We may also consider dimensions of the codes that represent the location of the property along a continuum or range. The name of the

category can be different from the codes to express its scope better and if necessary, we can also create sub-categories from the codes then link to categories.



Body Adaptation, Construction, Comparative Analysis, Development, Drug, Email, Experience, Negative Response, Parents, Peer Acceptance, Programming, School, Study, Social Act, Telephone Conversation, Text Message, Tempting Fate, University, Voice Mail.

Figure 2: A sample set of codes generated from a qualitative analysis

From the above set of codes, we can group the concepts: 'Email', 'Telephone Conversation', 'Text message' & 'Voice Mail' into a category and name it 'Communication'.

## When to stop line-by-line coding?

As we can see, the line-by-line coding is a very time consuming and tedious work but at the same time it also helps to build detail structured conceptual data model. When we are not really finding any new concepts but only repeating the existing labels, we can stop doing this very detailed analysis. But to discover more information from the data, we have to continue our analysis. At this stage, we can use analytic tools to break the data and collect more information. This process is called "microanalysis".

## Types of Open Coding

There are a number of ways to do Open Coding. In our previous examples, we have analyzed the data line by line, every sentence and even word by word. This process of coding is called line-by-line coding which is important to build concepts and categories. But based on the research requirement, we can also look into a bit broader scale and code against a sentence, paragraph, chapter etc. There can be situation where we may just need to define concepts for an entire document.

## Research Team Size

When doing open coding, it's better to do in a group at the beginning. Sarker from Washington State University and Lau & Sahay from University of Alberta found that problems can occur if the team members do the initial coding separately [8]. Researchers from Institut für Informatik, Freie Universität Berlin also found some key benefits of pair coding that includes:

- Group conversation helps to take important decisions (i.e. single out phenomena for coding, decide which existing concepts to use for coding or when to create new concept) [Berlin]
- Concept definitions become more exact and differentiations get more precise
- The data perspective is maintained more consistently
- Generally, more number of phenomena are discovered and processed.

## Use of Open Coding

Open Coding is generally the initial stage of Qualitative Data Analysis. After completing the Open Coding, depending on the methodology we use, we can do Axial Coding and Selective Coding. At later stage of the research, these coding help us to build theories in an inductive process (i.e. Grounded Theory). Open Coding can be used with inductive, deductive or verification modes of inquiry too.

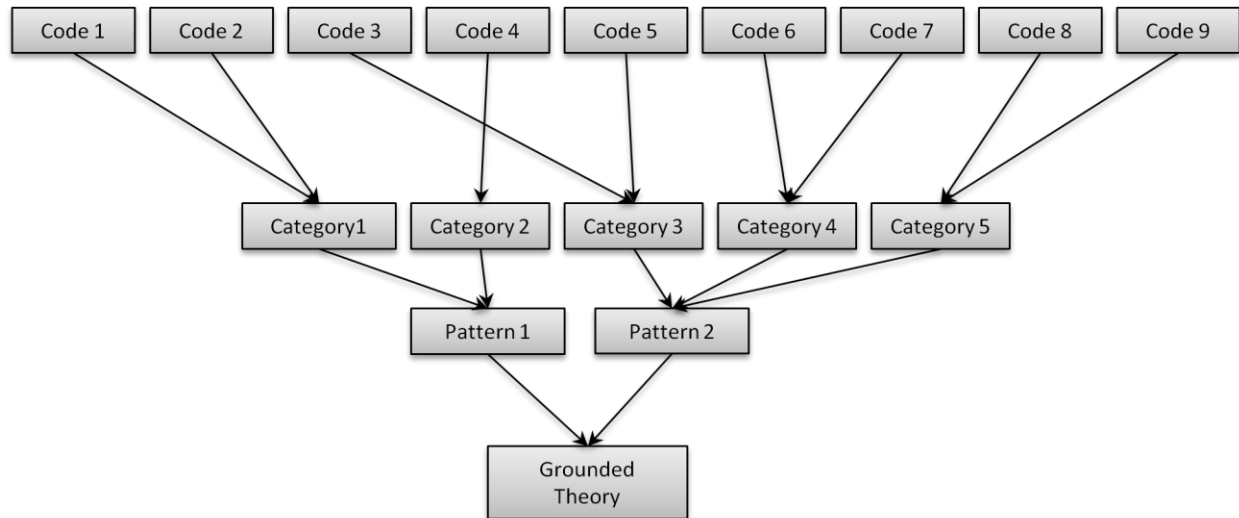


Figure 3: Workflow of Grounded Theory methodology

## Exercise

Now that we know how to do Open Coding, let's try to use it. Following is part of an interview transcript with a woman in her 20s and is about drug use by teens. We would like to use open coding to analyze the data.

**Interviewer:** Tell me about teens and drug use.

**Respondent:** I think teens use drugs as a release from their parents. Well, I don't know. I can only talk for myself. For me, it was an experience. You hear a lot about drugs. You hear they are bad for you. There is a lot of them around. You just get into them because they're accessible and because it's kind of a new thing. It's cool! You know, it's something that is bad for you, taboo, a "no". Everyone is against it. If you are a teenager, the first thing you are going to do is try them.

**Interviewer:** Do teens experiment a lot with drugs?

**Respondent:** Most just try a few. It depends on where you are and how accessible they are. Most don't really get into in hard-core. A lot of teens are into pot, hash, a little organic stuff. It depends on what phase of life you are at. It's kind of progressive. You start off with the basic drugs like pot. Then you go on to try more intense drugs like hallucinogens.

**Interviewer:** Are drugs easily accessible?

**Respondent:** You can get them anywhere. You just talk to people. You go to parties, and they are passed around. You can get them at school. You ask people, and they direct you as to who might be able to supply you.

**Interviewer:** is there any stigma attached to using drugs?

**Respondent:** Not among your peers. If you're in a group of teenagers and everyone is doing it, if you don't use, you are frowned upon. You want to be able to say you've experienced it like the other people around you. Obviously, outsiders like older people will look down upon you. But within your own group of friends, it definitely is not a stigma.

**Interviewer:** You say you did drugs for the experience. Do kids talk about experience?

**Respondent:** it's more of sharing the experience rather than talking about the experience. You talk about doing drugs more than what it's like when you take drugs. It depends upon what level you are into it, I guess. Most kids are doing it because it is a trend in high school. They are not doing it because of the experience in some higher sense. They are doing it because they are following the crowd.

Source: **Basics of Qualitative Research**, Second Edition by Anselm Strauss & Juliet Corbin [6]

So, how are we going to do that? One option is to printout the content, high light the important concepts and write the codes. But a manual open coding approach is not a good process especially when we have to deal with large amount of data. As you are limited to search by reading only, it can also cause unwanted errors. This process is simply impractical for a large scale data analysis with open coding.

The image shows a screenshot of a document with interview transcripts. Handwritten annotations in black ink are present. On the right side, there are four main codes: 'Experience', 'Drug talk', 'negative talk', and 'easy access'. 'Experience' has an arrow pointing to the respondent's first paragraph. 'Drug talk' has an arrow pointing to the respondent's second paragraph. 'negative talk' has an arrow pointing to the respondent's third paragraph. 'easy access' has an arrow pointing to the respondent's fourth paragraph. In the middle of the document, there are two more codes: 'Challenge the adult negative stance' and 'Negative connection', both with arrows pointing to the respondent's second paragraph. At the top of the respondent's first paragraph, 'Rebellious act' is written with an arrow pointing to the phrase 'You hear they are bad for you'.

Figure 4: Open coding with pen and paper

We can use software applications like *Saturate* [4] or *Atlas.ti* [5] to do Open Coding. *Atlas.ti* is a commercial qualitative data analysis tool and *Saturate* is a free web based tool developed by Dr. Sillito

from University of Calgary. We will use this tool in our exercise. A video tutorial for the tool can be found here [4].

Now that we have software and know how to use it, let's start building concepts. Take a look into the response: *"I think teens use drugs as a release from their parents"*. This looks like an act of rebellion. So we can code it as "rebellious act". The term 'use' looks like meaning something more. If we take out the context of drug use for a second and think about it – it may mean that they are being used for some other reason which we are not sure at this state. So, we should take a note ("Memo") for future reference. We have to continue to analyze the data line-by-line and add codes as necessary as long as we find significantly new concepts. Tools like Saturate can help us in both improving the efficiency and better manage the data.

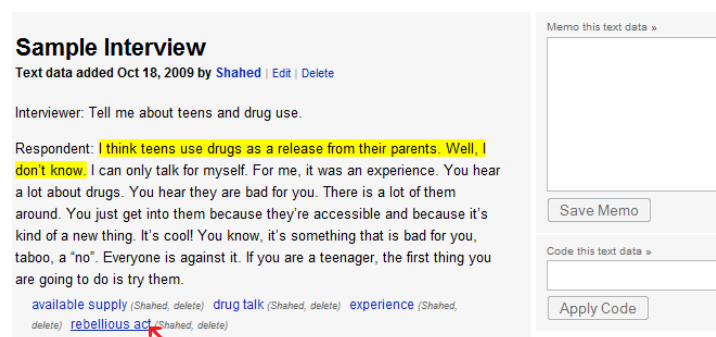


Figure 5: Adding codes using Saturate

As we continue on coding the data, we may find similar concepts and can classify them under common concepts (i.e. 'drug talk', 'negative connection'). Following code snippet shows a part of the interview transcript with codes.

**Respondent:** *I think teens use drugs as a release from their parents [**"rebellious act"**]. Well, I don't know. I can only talk for myself. For me, it was an experience [**"experience"**] [in vivo code]. You hear a lot about drugs [**"drug talk"**]. You hear they are bad for you [**"negative connotation"** to the **"drug talk"**]. There is a lot of them around [**"available supply"**]. You just get into them because they're accessible [**"easy access"**] and because it's kind of a new thing [**"novel experience"**]. It's cool! You know, it's something that is bad for you, taboo, a "no" [**"negative connection"**]. Everyone is against it [**"adult negative stance"**]. If you are a teenager, the first thing you are going to do is try them [**"challenge the adult negative stance"**].*

**Interviewer:** *Do teens experiment a lot with drugs?*

**Respondent:** *Most just try a few [**"limited experimenting"**]. It depends on where you are and how accessible they are [**"degree of accessibility"**]. Most don't really get into in hard-core [good in vivo concept] [**"hard-core use"** vs **"limited experimenting"**]. A lot of teens are into pot, hash, a little organic stuff [**"soft core drug types"**]. It depends on what phase of life you are at [**"personal developmental stage"**]. It's kind of progressive [**"progressive using"**]. You start off with the basic drugs like pot [**"basic drugs"**]. Then you go on to try more intense drugs like hallucinogens [**"intense drugs"**] [in vivo code].*

In the process of line-by-line coding, we will soon be able to group the concepts into categories like 'drug use' for concepts like 'hard-core use' and 'soft core'. Once we start getting too many old concepts,

we can stop labeling and move on to next step (i.e. selective coding, axial coding) based on our research methodology.

## **Benefits of Open Coding**

In the process of Open Coding, the concepts emerge from the raw data and later grouped into conceptual categories. The goal is to build a descriptive, multi-dimensional preliminary framework for later analysis. As its build directly from the raw data, its process itself ensures the validity of the work.

## **Problems**

Although Open Coding is an important tool for Qualitative Data Analysis but it's also a very time consuming and tedious work. Sometimes it's hard to decide when to stop line-by-line coding and if the researcher misses any important concept, he/she may have to restart the boring task again.

## **Annotated Bibliography:**

### **John V. Seidel. Qualitative Data Analysis [2]**

This document was originally part of the manual for the Ethnograph v4. It explains the process of Open Coding and also Qualitative Data Analysis in a broad sense.

### **Michael Nunes, Saul Greenberg, Carman Neustaedter. Using physical memorabilia as opportunities to move into collocated digital photo-sharing [9]**

A study on how physical memorabilia can be used as opportunities to move into home-based collocated digital photo-sharing. The researcher used semi-structured contextual interviews, each approximately one hour long. They used open coding as part of the data collection and analysis.

### **Sarker, S. Lau, F. Sahay, S. Building Inductive Theory of Collaboration in Virtual Teams: An Adapted Grounded Theory Approach [8]**

This paper outlined how the grounded theory was adapted to develop a theory of collaboration in virtual teams. The researchers studied virtual teams composed of students from two different universities and engaged in a 14 week long systems development projects. They analyzed the data using adapted versions of open coding, axial coding and selective coding.

## **References**

[1] **Symbolic Interactionism.** Bulmer H. (1969) [\[Link to Google Books\]](#)

[2] **Qualitative Data Analysis.** John V. Seidel [\[Link\]](#)

[3] Page 582, **Nursing research: principles and methods** by Denise F. Polit, Cheryl Tatano Beck [\[Link to Google Books\]](#)



[4] **Saturate**, a web-based Open Coding tool developed by Dr. Sillito. University of Calgary  
<http://www.saturateapp.com>

[5] **Atlas.Ti**, A commercial desktop application for Qualitative Data Analysis. <http://www.atlasti.com/>

[6] Chapter 8, **Basics of Qualitative Research**, Second Edition by Anselm Strauss & Juliet Corbin [\[Link\]](#)

[7] **A Coding Scheme Development Methodology Using Grounded Theory for Qualitative Analysis of Pair Programming**. Institut für Informatik, Freie Universität Berlin. [\[Link\]](#)

[8] **Building Inductive Theory of Collaboration in Virtual Teams: An Adapted Grounded Theory Approach**. Sarker, S. Lau, F. Sahay, S. [\[Link\]](#)

[9] **Using physical memorabilia as opportunities to move into collocated digital photo-sharing**. Michael Nunes, Saul Greenberg, Carman Neustaedter. [\[Link\]](#)