Transcription Collection

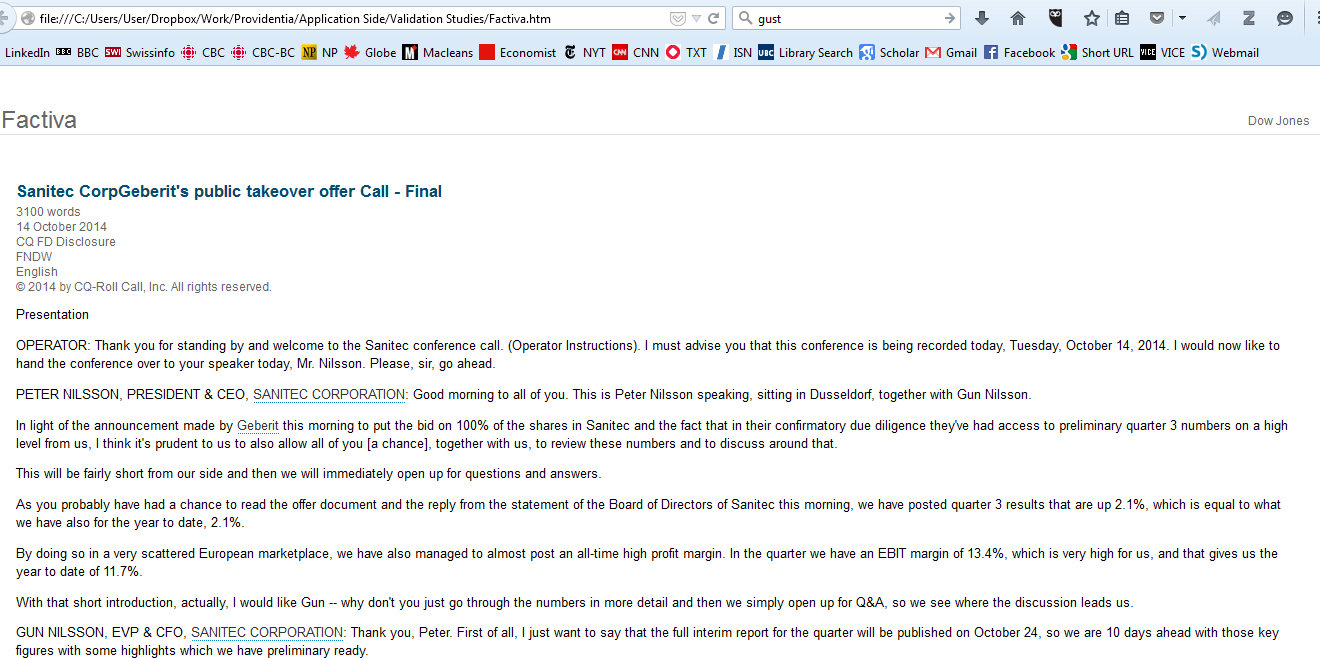
Ryan Cross ([ryan.w.cross@gmail.com](mailto:ryan.w.cross@gmail.com)) April 17, 2015

# Background and Purpose

Publically listed companies regularly provide information to analysts who trade their stocks. This allows analysts and traders to understand the company and question the executive team; transcripts of are available afterwards for review. The need with this project is to create a tool to assemble a local dataset of [earnings calls](http://en.wikipedia.org/wiki/Earnings_call) and other affiliated transcripts. The assembled transcripts will become a repository for analysis.

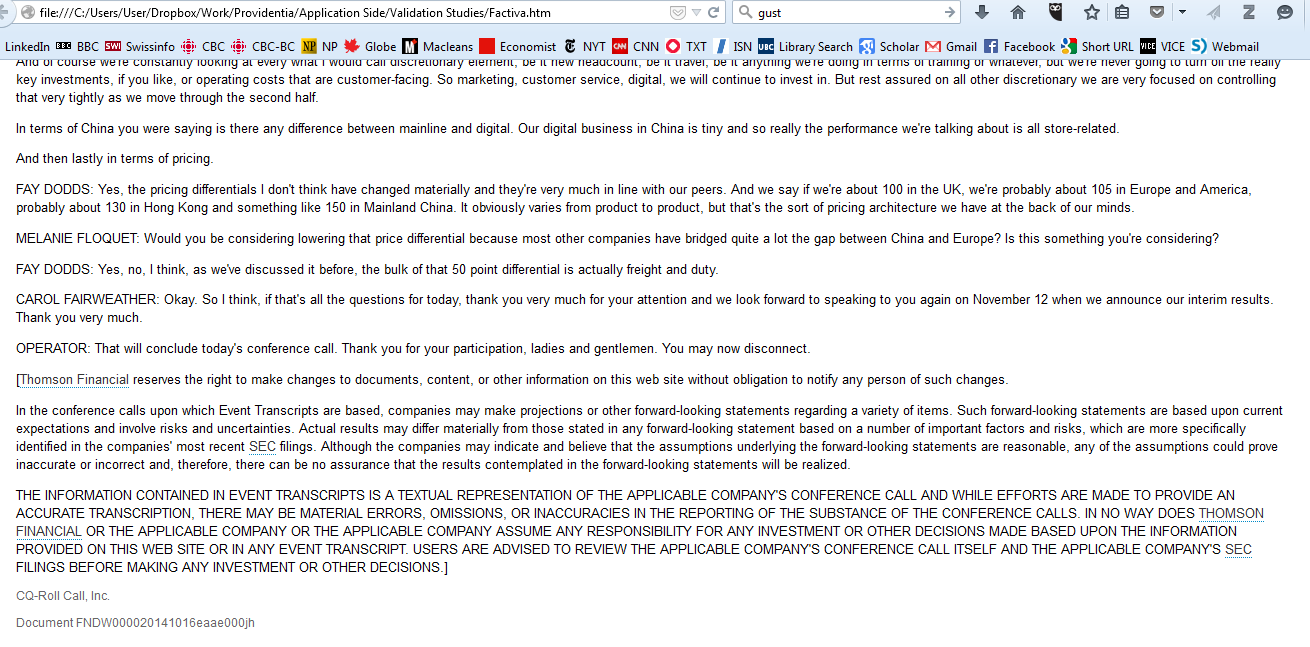
# Details

Many publically listed companies’ earnings calls and other key events are transcribed and made available. In some cases, for free via [Seeking Alpha](http://seekingalpha.com/analysis/transcripts/all), but also on subscription services such as [Factiva](http://en.wikipedia.org/wiki/Factiva). The information here is from Factiva. The transcripts are nearly always formatted in the same way. For example, the information in the small red box is (as far as I can tell) always identical at the start and end of each transcript (see following screen-grabs respectively).



The information within this box here always seems to be formatted in the identical way. It always contains the identifying information about the transcript.

Figure : Consistent Information at Start of Transcript



The information within this box here always seems to be identical way. If nothing else, the text at the very bottom left here (i.e., “Document FNDW…”) appears universally.

Figure : Consistent Information at End of Transcript

Transcripts contain paragraphs of text; *paragraphs are key for the organization of the data.* Each paragraph of each earnings call is the most elemental level of data required. In the above two figures, the paragraphs of text appear *outside* the red box’s.

One of the key aspects of what is needed is that I need to identify who said each specific paragraph. Likewise, I need to know which transcript each paragraph is associated with.

Here is an example of one specific transcript; I’ve identified the speaker of each paragraph:

* The text at the very top of the screen-grab is the material identified in Figure 1 (above) as being consistently formatted for each record. The material in the RED box *also* contains information which is relevant for this project. It contains information such as the name of the transcript and the date, which needs to be associated with each paragraph of text.
* The text in the PURPLE box is everything said by the speaker identified by “OPERATOR.”
* The text in the BLUE box is everything said by the speaker identified by “PETER NILSSON, PRESIDENT AND CEO, SANITEC CORPORTATION,”
* The text in the GREEN box is everything said by “GUN NILSSON, EVP & CFO, SANITEC CORPORTATION

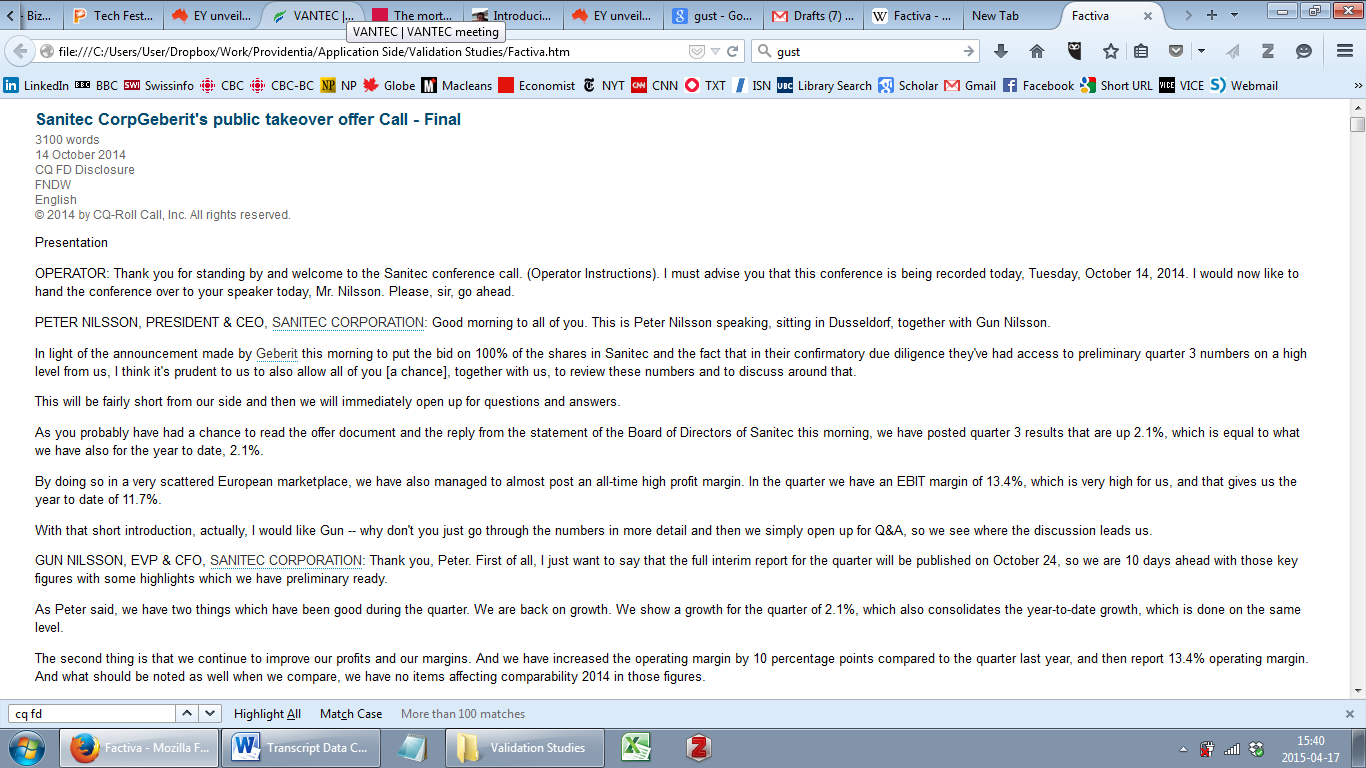


Figure : Transcript Paragraph Speaker Example

# Data Collection

I would ideally like to be able to run this database extraction "tool" on material I collect from Factiva on my own.

**Question:** *Would it be possible to build a solution that I can run myself? For example, if I were simply to save a results page from Factiva (such as the attached example) with a unique name in a folder, would it be possible to run this "tool" so that it automatically goes through all the files saved in that folder?*

Ideally, the material pulled into a database from the Factiva results listing page can either be put into a corresponding CSV file with the same name. I can manage the data once it has been “pulled” from the Factiva results listing page.

Here is a completed example from the original results listing as saved from Factiva, through to CSV file example of the first two transcripts that appear on this results listing page:

* Factiva Results Page, saved as “HTML Document (.htm).” [DropBox link here](https://www.dropbox.com/s/hvfxgxjdi6wakiw/Factiva.htm?dl=0) (also attached to email)
* PDF of first two Factiva Results Page transcripts, with notation on what I’ve excluded (i.e., some generic information which is irrelevant). DropBox link here (also attached to email)
* CSV file of the first two transcripts. Each of the columns in the CSV file is identified below (Table 1).

|  |  |  |
| --- | --- | --- |
| **Column** | **Header** | **Explanation** |
| A | Transcript\_Title | Title of the transcript as provided |
| B | Date | Date on as provided on transcript |
| C | Speaker\_Name | Name of Speaker |
| D | Paragraph\_Text | Text of paragraph |

Table 1: CSV Column Headers