**Financial Statement Notes Text- Data Dictionary**

**General description**

This document describes the data file that contains information on financial statement notes. The data is extracted using XBRL. Without XBRL, extracting financial statement notes is challenging because the terminology used to describe notes varies across firms. Using XBRL, companies must tag each footnote in its entirety using a “TextBlock” tag, which essentially captures a block of text in HTML format.[[1]](#footnote-1) Hence, using XBRL it is possible to extract footnote data and perform comparisons across companies that use the same “TextBlock” tags.

To aid researchers, we provide data in a SAS format, which includes readable Text (after removing tables and HTML tags) of select footnotes, the count of words in each footnote, and the count of numbers in each footnote (not including numbers in tables, and numbers representing years, addresses etc.). Since SAS has a field size limitation in rare instances (less than 0.3%) the text is truncated. We indicate these instances in the dataset.[[2]](#footnote-2)

In addition, because there is some flexibility in choosing XBRL tags from the FASB U.S. GAAP taxonomy, and because not all footnotes contain the same information, we group multiple tags into categories of footnotes. The data includes the original name of the TextBlock tags as well as our own grouping.[[3]](#footnote-3)

**Sample**

The available data is only for selected footnote categories that are frequent within our database based on data extracted from 10-K filings with fiscal years between 2011 and 2021. We do not consider XBRL tags that are not within the selected footnote categories. We also do not consider tags with a very low frequency or extended (custom) tags that companies create and do not appear on the FASB U.S. GAAP taxonomy. We do not include them because they are likely not common and challenging to compare across firms. We only present footnotes that contain at least 20 words of readable text.

**The data we provide includes the following fields:**

|  |  |
| --- | --- |
| **Variables About an EDGAR Submission** | |
| *ADSH* | Accession Number. The 20-character string formed from the 18-digit number assigned by the SEC to each EDGAR submission. |
| *CIK* | Central Index Key (CIK). Ten digit number assigned by the SEC to each registrant that submits filings. |
| *Fiscal\_Period\_End* | Fiscal period end (day, month, year) equivalent to the *DATADATE* in Compustat |
| *SIC* | Standard Industrial Classification (SIC). Four-digit code assigned by the SEC as of the filing date, indicating the registrant's type of business. |
| *ARC* | The count of distinct (within each disclosure) monetary XBRL tags in an SEC filing – used to capture firm/accounting complexity |
| *Date\_Filed* | The actual date of company filing |
| *Filer\_Status* | The type of filer, accepting the following values: 1-LAF=Large Accelerated, 2-ACC=Accelerated, 3-SRA=Smaller Reporting Accelerated, 4-NON=NonAccelerated, 5-SML=Smaller, Reporting Filer, NULL=not assigned |
| *ID* | Unique internally generated ID for each TextBlock /ADSH. |
| *tag\_name* | The XBRL TextBlock tag name. |
| *Note category­* | We generate a “normalized” list of note categories because of the flexibility of firms to select TextBlock tags in XBRL. We also provide a description of each XBRL tag based on the FASB U.S. GAAP taxonomy. **A detailed list of notes categories and mapping to XBRL tags, the TextBlock name, and the description is available in a separate Excel file (TextBlock Categories.xlsx).** |
| *Readable\_Text* | Contains the entire text in the financial statement note after the removal of HTML tagging. The field contains textual disclosures that do not appear in tables. Although numbers are included in the text, the count of words excludes all numbers. |
| *Word\_count* | The count of non-numeric words in each footnote.[[4]](#footnote-4) |
| *Number\_count* | The count of numbers in each footnote (excluding numbers in tables) |
| *Truncated* | Because SAS limits the field size to 32,767 characters, few observations are truncated. If the Truncated field equals one, the field was truncated. Note, that because the count of words is generated before the data is imported to SAS it is **complete** and is not affected by the truncation. |

1. In few exceptions, most commonly in the lease footnotes, firms use several TextBlocks to tag a single footnote. [↑](#footnote-ref-1)
2. If you want the full HTML files please reach out to us. [↑](#footnote-ref-2)
3. We provide an Excel file with our own financial statement notes categories. For example, some firms use the “BusinessCombinationDisclosureTextBlock” or “MergersAcquisitionsAndDispositionsDisclosuresTextBlock” to describe their M&A activity. We group these two Textblocks into one “Business\_Combination” category. [↑](#footnote-ref-3)
4. Our word count is based on text outside of tables, which we believe is the cleanest approach to capture words. While it is possible that some firms describe text within tables, using traditional heuristic rules to classify tables into those that contain more/less text is often inaccurate. [↑](#footnote-ref-4)