15. OCR ACCURACY

15.1 Guidance

Text accuracy is subject to the following guidance:

- For non-OCR'd documents, such as native word processing or HTML rendered directly from the authoring package, the target is 99.95% character accuracy.

- For OCR'd documents, the target is 95% word accuracy on those words that are not stop words. Where feasible, un-edited OCR output that contains header and footer information should be properly zoned to remove this information in order not to inhibit proximity searches for text at the top and bottom of scanned pages.

15.2 Methodology

Word accuracy is defined as follows:

$$\text{Word Accuracy} = \frac{\text{Total Words} - \text{Number of Incorrect Words}}{\text{Total Words}}$$

In this calculation, "words" are defined as any words that are not stop words[1] as set forth in the list included in LSN Guideline 21. To determine the word accuracy achieved, all stop words are subtracted from the count of "Total Words" as well as from the count of "Number of Incorrect Words." The calculation results in an accuracy rate representative of words that are not stop words.

LSN participants may use any metric of their choosing in performing their internal quality assurance assessments on their conversion processes. However, the objective of having an OCR standard is to support the most effective operation of the LSN text search and retrieval capability. Autonomy, the software package used by the LSN to provide text search and retrieval, operates on words, which is why word accuracy was identified as the metric for performance assessment. The LSN Administrator will state text accuracy assessments using word accuracy as defined above in reporting to the participants on sampling efforts, as well as for internal reports to management and the Commission.

[1] Stop words are those words that text retrieval database software ignores and does not index.