UNITED STATES
NUCLEAR REGULATORY COMMISSION
WASHINGTON, DC 20555 - 0001

April 22, 2003

TO: Members of the Licensing Support Network Advisory Review Panel

FROM: Daniel J. Graser /**RA**/
LSN Administrator

RE: Revised LSN Guideline on OCR Accuracy

Comments on the proposed draft revision to the LSN Administrator (LSNA) Guideline on OCR accuracy for full text were received from the U.S. Department of Energy (DOE) and the staff of the U.S. Nuclear Regulatory Commission (NRC). No other comments were received.

In summary, NRC staff suggested that character-based statistics be used when stating accuracy rates, since OCR engines recognize and output characters. They also commented that a word misspelled in the original that is "correctly" OCR'd is technically a correct representation of the original misspelled word.

DOE made five observations and two recommendations, summarized as follows:

Comments:

1. Zoning to remove headers and footers imposes a cost burden as there is no commercially available product to automatically perform this function
2. Information contained in headers and footers is valuable and enhances retrievability
3. There is low probability that headers/footers will actually split up a phrase worthy of retrieval that does not appear elsewhere in the document
4. The LSN search engine is concept based and should be immune to the occasional misplacement of words that cross a header or footer
5. Other search techniques such as widening proximity operators or performing iterative subordinate searches can compensate.

Recommendations:

1. Remove second sentence in second paragraph of Section 15.1
2. Modify the last sentence of the last paragraph in Section 15.3 by removing "however, it was found that un-edited OCR output that is not properly zoned to remove header and footer information will prevent proximity searches for text at the top and bottom of scanned pages."

Response

The purpose of an LSNA Guideline is to provide guidance to LSN participants regarding design and operational matters that will make the LSN system operate effectively. A guideline is not

intended as a direction to the participants as to how to meet the requirements of 10 C.F.R. Part 2, Subpart J. The objective of the suggested OCR guideline is to support effective operation of LSN text search and retrieval capability based on Autonomy™ text retrieval software.

Because this software package operates on words, word accuracy was identified as the metric in the draft guidance. The LSN administrator will use the word accuracy metric in reporting back to the participants, agency management, and the Commission on sampling efforts. Moreover, as a practical consideration, a statistically valid sample of 1000 words is easier to identify and select than a statistically valid sample of 10,000 characters (assuming that in each case you use an order of magnitude larger sample than the decimal position in the target numerator). Measuring the quality of any OCR output implies that some level of post production review is necessary in order to generate the metric data. Electing to use a word count of 1000 (which can be accomplished by use of Grammatik™ Analysis - Basic Counts feature or similar products) seems preferable to introducing tens of thousands of "counts" to identify a sample character set.

Regardless of whether words or characters are used by a participant, however, in order to consistently attain these accuracy rates, participants will have to utilize some sort of effective post-processing assessment and feedback as part of the participant's conversion efforts, which ultimately is the goal sought relative to this guideline. Therefore, the parties are afforded the latitude to use any metric of their choosing, which addresses the comments submitted by NRC staff.

In connection with the DOE comments, after further consultation with the Autonomy™ software vendor, it does not appear that the presence of spurious header and footer text within the content text will impede the operation of the software. Therefore, the draft recommendation for zoning documents has been revised to indicate that this is an objective that should be pursued, where feasible.

Based upon the two above-mentioned changes to the guideline, Section 15.3, Discussion, in the preliminary draft has been removed as it no longer adds value to an understanding of the background for the OCR guideline established.

Copies:

A. Bates, SECY
M. Janney, ASLBP
M. Schmit, ASLBP
J. Turner, ASLBP