Department of Energy

Office of Civilian Radioactive Waste Management
Office of Repository Development
P.O. Box 364629
North Las Vegas, NV 89036-8629

QA: N/A

OVERNIGHT MAIL

MAR 05 2003

Daniel J. Graser, LSN Administrator
Atomic Safety and Licensing Board Panel
U. S. Nuclear Regulatory Commission
Two White Flint North
Rockville, MD 20852

COMMENTS ON JANUARY 31, 2003, DRAFT REVISION TO LICENSING SUPPORT NETWORK (LSN) GUIDELINES

The following comments are provided by the U.S. Department of Energy (DOE) in response to a U.S. Nuclear Regulatory Commission (NRC) proposed revision to the LSN Guidelines. The DOE's LSN contractor, CACI, and the University of Nevada, Las Vegas, (UNLV) Information Science Research Institute (ISRI) have jointly participated in helping the DOE prepare the following comments and rationale to retain header and footer content in the Optical Character Recognition (OCR) generated text. It is recommended that the NRC allow the retention of the headers and footers in the OCR generated text of documents in the LSN.

1. The removal of headers and footers from the un-edited OCR output would be a significant burden to impose at this point in the DOE schedule for implementing the LSN. At this time there is no commercially available product to automatically perform this editing function.

2. In CACI's experience, and as supported by UNLV, the information captured from headers and footers during the OCR process is very helpful and is often used for document retrieval. For example, headers and footers often contain company names, personnel names, addresses, and other information pertinent to some searches. Information in headers and footers is valuable and legible; therefore, it enhances the retrievability of those pieces of data.

3. Additionally, it is a relatively rare occurrence that terms important enough to be used in a proximity search are both split across multiple pages and do not appear more than once in a document – so it is a similarly rare occurrence that either a header or footer would prevent finding the terms via proximity searching. In CACI's experience it is unlikely that the term or phrase being searched would occur only once in a document, even if the terms were separated by a header or footer. The UNLV/ISRI's research supports this assertion: In their test of Autonomy with 1058 documents (the one to which NRC refers), ISRI used two versions of the same collection. The first version was prepared using the current DOE OCR procedure; headers and footers were not removed. The second version used an OCR procedure that is 99.8% correct, prepared using manual zoning; the zoning process included header and footer removal. The ISRI's retrieval test showed no statistical difference for recall/precision level for the two versions of the collection.

4. Autonomy is a concept-based engine. The engine identifies concepts based upon detailed statistical and natural language routines that are immune to occasional misplacement of a few words that may cross a header or footer.

5. Further, search techniques such as widening the search proximity or searching separately for the terms and then conducting a subordinate search will normally allow finding the term groups.

If NRC determines that removal of headers and footers is still required, then we recommend the following alternative language so the requirement is stated without a specification constraining how it must be accomplished: "Un-edited OCR output that contains header and footer information should be removed in order not to inhibit proximity searched for text at the top and bottom of scanned pages."

The first comment is in Section 15.1, second paragraph:

Comment 1:
Remove second sentence:

> "Un-edited OCR output that contains header and footer information should be properly zoned to remove this information in order not to inhibit proximity searched for text at the top and bottom of scanned pages."

The second comment is in Section 15.3, last paragraph:

Comment 2:
Remove the following from the last sentence:

> "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."

If you have any questions concerning these comments, please contact Harry E. Leake at (702) 794-1457 or Sheryl A. Morris at (702) 794-5487.

for Joseph D. Ziegler, Acting Director
Office of License Application and Strategy

OLA&S:SAM-0785

cc:
J. M. Harding, BSC, Las Vegas, NV
CMS Coordinator, BSC, Las Vegas, NV
Veronica Hubbard, CACI, Las Vegas, NV
Marshall Bishop, MTS, Las Vegas, NV
E. R. Jorgensen, RSIS, Las Vegas, NV
A. V. Gil, DOE/ORD (RW-40W), Las Vegas, NV
H. E. Leake, DOE/ORD (RW-32W), Las Vegas, NV
S. A. Morris, DOE/ORD (RW-40W), Las Vegas, NV
D. R. Williams, DOE/ORD (RW-40W), Las Vegas, NV
J. W. Wooley, DOE/ORD (RW-32W), Las Vegas, NV
J. D. Ziegler, DOE/ORD (RW-40W), Las Vegas, NV
OLA&S Library
Records Processing Center = "6"