B. OCR ACCURACY REQUIREMENT RATIONALE

B.1 Background

When the NRC assumed responsibility for the design and implementation of the LSN from DOE in 1998 and the overall architecture was redefined to be an Internet-based World Wide Web environment, it was recognized that the "old" LSS Functional Requirements had to be updated. By April 2000 numerous revisions of the old requirements had been performed by the LSNARP TWG and the LSN Functional Requirements were condensed to reflect the following proposed quality standard for text accuracy:

> LSN 2.22.02 Text format shall comply with US.ISO_8859-1 and be a searchable full text representation of the document. Documents must be accurately represented with an overall error rate of no more than 5.0%.

In August 2000, the focus of the proposed Functional Requirements became text accuracy as an objective rather than a requirement. The proposed revision also placed the overall objective in the context of a participant's collection rather than an individual document:

> **LSN 2.25.02** - Text format shall be formatted to comply with the US.ISO_8859-1 character set or be in one of the following acceptable formats: plain text, native word processing (Word and WordPerfect versions as requested by participants), PDF Normal, or HTML. As a goal, textual documents should be accurately represented with an overall error rate of no more than 0.5% based on character accuracy and a per page error rate of no more than 1.5%. Documents converted through means other than Optical Character Recognition (OCR) should have an error rate of less than 0.05%.

This standard was proposed as a goal rather than as an absolute standard to acknowledge that document quality can cause variability in the OCR process output results. Characterizing it as a goal also affords the participants a measure of latitude in meeting their scheduled availability commitments. Having most documents submitted in the native word processing formats or in HTML rendition output from the authoring package (by definition, 100% accurate, and at least 99.95% accurate after transmission) should compensate for portions of the collection that may have an accuracy rate as low as 95% accuracy (unedited OCR output) within a document. The blend of clean native text and some non-edited ASCII should result in meeting a target of 98.5% for a collection as a whole. NRC further indicated to the TWG representatives that a final pronouncement would be forthcoming once a hardware/software suite of products was selected for LSN implementation.

B.2 Discussion

In 1988, as the LSS rule was being finalized, DOE's support contractor (SAIC) was generating a series of comprehensive requirements and design analyses documents, including the issue of full text accuracy. In DOE's 1988 cost-benefit analysis study on the LSS (Section 5.2.2.7, Variant VI - Full Text via Re-keying), DOE identified an alternative approach to acquiring full text to meet LSS requirements:

> In this variant, there is no automated text conversion (OCR) process. The conversion of hardcopy text to ASCII is accomplished by re-keying the document. An expected 99.8% accuracy of data via re-keying would be achieved by double keying the original source document.[B-1]

DOE recognized that this "expected" accuracy rate needed further study and validation, and included data accuracy impacts in its initial LSS Prototype system. In the LSS Prototype Test Report (Section 2.3.3.2, Accuracy), the following findings were reported by the DOE contractor:

> Based on the recognition that error rate is a significant consideration for the success of full-text retrieval systems, the required accuracy rate for the prototype database was set at 99.8%. This corresponds to two errors per 1000 characters or as many as five misspelled words on a typical LSS page. It is believed that a lower accuracy rate would erode user confidence in the data through missed retrievals. In addition, team members believed that higher error rates would tend to discourage users during searches and thereby make interpretation of user test results more difficult.[B-2]

DOE subsequently launched an internal pilot effort to deliver this accuracy for text conversion of its record materials via intake stations that would scan, OCR convert, edit/cleanup, and perform quality assurance on record processing. The solicitation document for acquiring those internal resources reflected the test results and recommendations in the LSS prototype test report. In the solicitation (Section C.3.2, Text Conversion Subsystem Requirements), DOE specified that:

> The Text Conversion Subsystem supports all tasks related to the conversion of document images to ASCII text files. These requirements deal directly with functional capabilities of the text conversion process that will support capture of all machine-readable text within a document and editing of text to assure 99.8% accuracy.[B-3]

[B-1] U.S. Department of Energy Office of Civilian Radioactive Waste Management, Office of Resource Management, "Licensing Support System Benefit-Cost Analysis," July 1988. p.48.

[B-2] Science Applications International Corporation (SAIC), "Licensing Support System Prototype Test Report," February 16, 1990.

[B-3] TRW Environmental Safety Systems, "Request for Proposal - Document Capture System, Prepared under Contract DE-AC01991RW00134; July 10, 1992. p.26.

The LSSARP subsequently took up the issue in the May 12, 1995 ARP meeting at which time the TWG indicated that the Level 1 and Level 2 Functional Requirements would address the issue of text accuracy. The LSS design Functional Requirements were significantly refined by the LSSARP TWG and forwarded to the ARP members for consideration on May 17, 1995. The technical standards for OCR quality presented to the ARP were incorporated in high level Functional Requirement LSS1-005 (and similarly in a second level requirement LSS2-004):

> The LSS shall provide the capability to recognize characters from the digital image of a document and convert these characters into a standard text representation of the document. This optical character recognition function shall achieve character recognition accuracies that are achievable with the best commercial products available at the time of the LSS system design [2.1003(a)(1) and 2.1003(b)(1)].

The LSSARP formally voted on and approved the design phase Functional Requirements at its July 6-7, 1995 meeting. The target for text accuracy acknowledged industry capabilities via regular reports from the Information Science Research Institute (ISRI) of the University of Nevada, Las Vegas (UNLV) to the LSSARP, which were based on information that the published in its Annual Report[B-4] on the results of OCR performance tests against a sample collection of DOE documents. Those findings were that the commercial products then available could generate in excess of 96% initial word accuracy and 97% initial character accuracy[B-5] against the collection of DOE sample documents.

It is not unreasonable to expect that the quality of initial OCR output will continue to improve, albeit recognizing that the best accuracy rates are achievable on clean, first generation print versions of documents not typical of most production scanning and OCR operations. Thus, it has been observed that:

> OCR technology has advanced to the point where today's systems are indeed useful for processing a large variety of machine-printed documents. Accuracies of 99% or more are routinely achieved on cleanly-printed pages.[B-6]

The standing requirement since 1995, therefore, has been to target OCR conversion initial output to the standard of the best commercially available OCR product. The last tested product comparisons of OCR products approached 98% character accuracy, as documented in ISRI's 1996 annual report. The document quality standard used in those tests was that of a first generation copy or print based on ISRI's DOE sample document collection.

[B-4] Information Science Research Institute, University of Nevada, Las Vegas, "1995 Annual Research Report", pp. 16 & 20.

[B-5] Caere OCR delivered overall word accuracy of 96.11% and 97.76% character accuracy; HP Labs OCR delivered overall word accuracy of 96.62% and 97.72% character accuracy.

[B-6] Rice, Stephen V., George Nagy, and Thomas A. Nartker, "Optical Character Recognition: An Illustrated Guide to the Frontier". Boston: Kluwer Academic Publishers, 1999. p.2.

To design and implement the LSN, NRC has selected a vendor that has proposed the use the Autonomy™ software product to address full text search and retrieval requirements. Autonomy offers most of the features originally defined in the LSS requirements for user interface expectations, including both Boolean and natural language search capabilities. Additionally, Autonomy has other enhancements that make the system much easier for non-ADP-professional users to work with, such as relevancy ranking.

Autonomy, however, is not a classically structured text retrieval package (i.e., it does not build a word list index of pointers to text occurrence locations), but rather relies upon Bayesian probability theory and Shannon's communication theory (semantic frequency of occurrence to infer the responsiveness of a given document to a query). Shannon's communication theory, in short, says that the ***fewer*** times a unit of communication occurs, the ***more relevancy*** should be accorded to it. Data streams generated by OCR software contain large quantities of misread characters, together with special characters marking such occurrences in a data stream. These misreads and marking characters are, for the most part, unique[B-7], and may occur only once in a data stream, so the question of "dirty data" from OCR output becomes critical to LSN performance. Even if the marking characters used by OCR software to identify suspect words are "turned off" (i.e., not placed in the ASCII stream) the software's "best guesses" at the misread characters are still scattered throughout the text. And, because there is no index of terms occurring in a central document, there is no index list in the text engine which could then subsequently be administered by a Database Administrator (DBA) for the cleanup of spurious characters.

The LSNA also considered compensating for possible search engine sensitivities by pre-categorizing documents and providing users with a more robust "browsing" environment so users would have less reliance on relevancy ranked query results. It is demonstrated, at least for small collections, that in multinomial probabilistic models using naive Bayesian assumptions, OCR errors have no effect on text categorization so long as dimensionality (number of categories) is constrained.[B-8] This means that categorization works well if there are relatively few broad categories to which the software has been trained and "tuned" by a system administrator. There is some concern over keeping a small enough number of categories to avoid excessive "paging down" through multiple Internet pages. Additionally, there is concern over the ability of the LSN, without imposing some prejudice, to define and then impose a limited number of categories that are able to be browsed on LSN searchers. Limiting categories for browsing shifts the burden of retrievability to the text search capability, specifically the reliance on a software package's precision and recall performance.[B-9]

[B-7] See Rice, et al, p.50. A mis-read of the term "Nye County" with a stray mark partially obliterating the "y" in the word County was mis-read by three different OCR software packages, and embedded in the resulting ASCII streams as, variously: **Nge ('ouch** or **Njje Count?`** or **Nyc couue~**.

[B-8] Taghva, Kazem, Tom Nartker, Julie Borsack, Steve Lumos, Allen Condit and Ron Young, "Evaluating Text Categorization in the Presence of OCR Errors", ISRI, 1999.

[B-9] Autonomy has never been baselined for precision and recall performance via participation in the National Institute of Standards and Technology's (NIST's) Text REtrieval Conference (TREC) program.

The issue for the LSN design team, therefore, is to encourage non-unique but important words to naturally rise to the surface in relevancy ranking. That is to say that we want to minimize the potential for OCR'd but non-edited data (full of spurious characters) to cause truly relevant documents from "sinking from the surface" in relevancy ranking because of the occurrence of so many unique characters/words (e.g., the misread characters) that are more "interesting" to Shannon's model.[B-10] This could happen if the clean, recognized data in a text stream -- such as three occurrences of the term "radionuclide transport" -- are overwhelmed by an extremely high volume of very unique character formations resulting from OCR misreads. In effect, the one-time-occurring error strings are "unique" and, per Shannon, therefore more valuable than a term occurring repeatedly in the document. In generating the relevancy result set for a search on "radionuclide transport" the Autonomy software may determine that all the generated, unique clutter is what the document is really about, ignoring the mundane occurrence of the term three times in a given page of a document which in any other algorithms would be a key indicator of the richness and concentration of the requested term.

B.3 Analysis

The LSN will be implemented using software that might be sensitive to large volumes of inaccurate data. The sensitivity would be manifested in how the software characterizes a document's relevancy to the search query. This skewed reporting of a document's relevancy ranking could be caused by spurious characters introduced by non-cleaned-up OCR. For example, in a collection where perhaps 10,000 pages may contain the term "radionuclide transport," incorrect relevancy characterizations presented by the software in its output could be a significant problem. If the software cannot be relied upon to deliver consistent relevancy-ranked results to attorneys who require highly predictable precision and recall[B-11] in a homogeneous collection, they will lose confidence in the use of the system.

There are two design solutions available to the LSN design team:

- Turn off relevancy ranking; or,
- Eliminate, as much as possible, the occurrence of "dirty data" and adhere to the 98.5% overall accuracy standard for a participant's entire collection rather than lowering the standard to 95% as was originally proposed in the April 2000 draft of the LSN Functional Requirements.

[B-10] The problems with probabilistic models are due to length normalization. The length of the document is usually defined to be the number of unique index terms in a document or a term with the maximum frequency. OCR errors will produce a lot of unique terms. This can create problems for a search engine that has no mechanism to clean them out while using both frequency and uniqueness as a determinant of high relevancy in a "larger" document - exacerbated by a large number of unique terms that are "garbage." For example, OCR can recognize "the" as "thc" and the software being used may repeat this misread because of a flaw in its algorithm combined with the unique characteristics of the documents' typefaces. Here the word "thc" is not a stop word and numerous misreads results in a high frequency, which can cause problems in ranking the document in the result set.

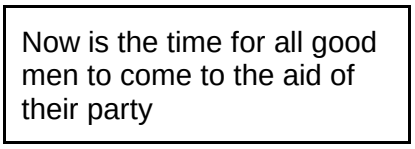
[B-11] See Dabney, "The Curse of Thamus: An Analysis of Full-Text Legal Document Retrieval" and the references to the Blair and Maron studies at http://www.yale.edu/lawweb/lawcrs/arc9798/lasdab1.htm# File Size and Retrieval Performance in Full-Text Systems.

Electing to strive for delivering data that more accurately represents the original document is consistent with the overall LSN goal of ensuring accuracy and integrity of information being used by participants. Additionally, it should have minimal impact for those organizations that have been adhering to the 1995 agreements, those that have not yet started processing and have relatively small collections, or those participants that already have substantial quantities of documents OCR'd at the 98% level and even greater amounts stored in native word processing formats.

B.4 Text Accuracy Evaluation

The NRC's LSN staff will routinely sample the text accuracy of documents posted on participant servers. Errors in text accuracy introduced by the OCR process are of two types: omission and commission (e.g. introduction of spurious characters by inaccurate substitution, misreads of splotches, skewness and kurtosis generated misreads, etc.). Although an argument can be made that introduced characters detract nothing from the subset of all other characters and words properly identified and therefore should not be counted against achieving the accuracy target, it has not been demonstrated that spurious characters have no effect on proximity searches.

In the example below of a document in which a text box has been used, the word "good" is right next to the word "men" as represented in the character string. It is also recognizable as a phrase "good men."



The text box introduces especially problematic vertical lines at the left and right sides that the OCR engine may interpret as a literal text string such as:

^l now is the time for all good ^1
^l men to come to the aid of ^1
^l their party ^1

Now, the OCR engine misreads the vertical from the box as a flag character and its guess as to what the character should be (e.g., the alpha character “l” or the number “1". If is assumed there is no editing on the raw OCR output to manually or automatically throw that data out, and that there is no stopword list that can account for all variations for spurious character combinations, the introduction of spurious characters may result in a proximity or phrase search for the adjacent **"good men"** that will no longer generate a hit. The word pattern it now is trying to match -- **“good ^1 ^l men”** -- is represented numerically something like 22686819003321357 and what the crawler and parser see is a string something like 226868190031003102332135 7. The garbage characters thus introduce data values that may lower the probability of a match on the term "good men" and also add word breaks.

The UNLV Information Science Research Institute is in the process of developing a new counting algorithm. In all likelihood, there may be three categories of errors that we report back to participants: character accuracy, word accuracy, and phrase accuracy. The LSNA will update participants on these recommendations as soon as they are available.

Until then, participants should assume that introduced, non-cleaned up spurious characters are counted.