

# Evaluation of the Current DOE Document Conversion System: A Study of Retrievability

Technical Report 2002-07  
Information Science Research Institute  
University of Nevada, Las Vegas

May 2002

## 1.0 Background

During the 2001/2002 fiscal year, the Information Science Research Institute (ISRI) at the University of Nevada, Las Vegas (UNLV) has been tasked to suggest improvements and evaluate the performance of the current DOE document conversion system.<sup>1</sup> This report gives a summary of the recommendations made by ISRI staff and a summary of the results of two types of performance tests.

There are two approaches to evaluating the performance of document conversion systems. One approach is to measure the accuracy of the textual output (i.e., average character accuracy) of the system. A second approach is to measure the performance of the system that will make use of the output text. In this case, textual output will be used to build the index for an Information Retrieval (IR) system that will aid in the task of finding documents of interest. The appropriate performance measure for IR systems is retrievability (i.e., precision and recall). Thus, to provide a thorough evaluation of system performance, two different studies (a character accuracy study and a retrievability study) have been conducted. [1, 2] Section 3.1 below, gives a summary of the results of accuracy tests and Section 3.2 summarizes the results of the retrievability tests.

## 2.0 Document Conversion System Recommendations

The task of document preparation for the LSN has two major components: *character recognition* and *page zoning*. The task of loading the text produced into an information retrieval system is, by comparison, straightforward and not error prone. Thus, in any document conversion system, character recognition and page zoning are performance-controlling operations.

### 2.1 The Importance of Measuring Retrievability

Although character recognition is typically measured by standard character accuracy, many characters in a document's text have no role in its retrievability. For example, punctuation marks, end-of-line hyphenation, and characters in stopwords<sup>2</sup> are ignored by an IR system. The top ten standard stopwords account for about 20 to 30 percent of all words in any collection. Thus, while character accuracy is related to retrievability, it is not a good measure of retrievability. For these reasons, and because retrievability is more important to users of the LSN, the retrievability testing described in Section 3.2 was recommended by ISRI staff.

---

<sup>1</sup> The system of concern is a computer driven character recognition system used to create an electronic (i.e. electronic text) copy of

paper documents. This system includes the Scansoft SDK 2000 OCR system running on IBM compatible equipment and is operated for the DOE by the Bechtel-SAIC Company, LLC (BSC) in Summerlin, Nevada.

2 Stopwords are common English words with no information content such as “the” and “and.”

## 2.2 The Use of Automatic Zoning

Page zoning can be done either manually by drawing a box around text to be captured or automatically by an OCR engine. Manual zoning not only necessitates thousands of hours of manpower, it also requires a pre-defined set of *zoning rules* which we have found to be error prone. [3] Automatic zoning on the other hand is performed by the recognition system and, although not always 100% accurate, captures all the data required for information retrieval.

During the 1990’s, ISRI conducted a series of experiments comparing retrievability from manually zoned collections to retrievability from automatically zoned collections. In every experiment, the use of automatic zoning followed by MANICURE [4] gave retrieval results equivalent to what one could expect from manually zoned pages, even from a nearly perfect collection. Based on this experience, ISRI recommended that DOE employ automatic zoning and the MANICURE post-processing system. The studies reported in Section 3 were performed to compare both automatic and manual zoning with and without MANICURE post-processing to support the ISRI recommendation.

## 3.0 Results of Tests to Measure Conversion Performance

The LSN is the discovery database to be used in the licensing proceedings and the DOE documents will be an important component of the LSN. The performance of the conversion system is therefore very important to the operation of the LSN. It is therefore appropriate to thoroughly test the DOE system to insure that users of the LSN will have the best technology available to enable them to find documents of interest.

To be realistic, such tests must be based on sample sets of actual documents that will be submitted to the LSN. To be reliable, tests must use well-accepted standards and scientific methods. To be statistically significant, tests must be based on reasonably sized, random samples of DOE documents.

### 3.1 Accuracy Tests

The technical requirements for document collections submitted to the LSN processed through an OCR system have included a “target character accuracy of 99.5%.” ISRI has designed and conducted a test to measure average character accuracy of documents processed by the conversion system. Because IR systems ignore stopwords and punctuation marks, this test focused on non-stopword accuracy and on the character accuracy of non-stopwords. MANICURE post-processing of the OCR output is a part of the conversion system. The text accuracies were measured both before and after application of the MANICURE system.

All accuracy tests were conducted with a random sample of 17 Microsoft word documents from the current DOE collection. The total number of non-stopwords in these documents is 164,483. The total number of characters in these words is 1,361,124. Because these documents were native Microsoft word files, they provided almost perfect images not typical of the DOE collection. To reflect realistic image quality, we also added first through fourth generation photocopies of these documents to our study. The error counting programs used are a modification of the OCR performance metrics developed by ISRI staff in the early 1990’s [5].

#### 3.1.1 Word (Non-stopword) Accuracies

The most important result of our word accuracy tests was that non-stopword accuracies produced by MANICURE (see Table 1) were uniformly higher than those produced by the OCR system alone. Because the average non-stopword is eight characters long, word accuracies are always lower than character accuracies. (If the average character accuracy for this collection were 99.5%, and if each character error were in a different word, the average word accuracy would be 95.9%.)

Another result is that the increased accuracy of MANICURE output over OCR output improves as page quality decreases (i.e., MANICURE helps the poor quality images more). Because our best judgement of the average print quality of DOE documents is between that of a first and second generation photocopy, these results indicate that the word accuracy produced is between 97.23% and 96.15% correct. It is also important to note that a 1% increase in word accuracy for a 97% correct page corresponds to correcting 33.3% of the word errors.

System	% Word Accuracy of	Orig.	Gen 0	Gen 1	Gen 2	Gen 3	Gen 4
Raw OCR Output	All non-stopwords	97.44	97.03	96.45	95.05	92.78	91.91
MANICURE Output	All non-stopwords	98.01	97.54	97.23	96.15	94.64	94.14

**TABLE 1. Average Non-stopword Accuracy for All 17 Documents**

### 3.1.2 Character Accuracy (of Non-stopwords)

As with word accuracies, the character accuracies of non-stopwords produced by MANICURE (see Table 2) were uniformly higher than those produced by the OCR system alone.

System	% Character Accuracy of	Orig.	Gen 0	Gen 1	Gen 2	Gen 3	Gen 4
Raw OCR Output	All non-stopwords	99.37	99.40	99.20	98.67	97.89	97.79
MANICURE Output	All non-stopwords	99.50	99.47	99.30	98.83	98.16	98.06

**TABLE 2. Average Character Accuracy for All 17 Documents**

Because our best judgement of the average print quality of DOE documents is between that of a first and second generation photocopy, these results indicate that the character produced is between 99.30% and 98.83% correct.<sup>3</sup>

## 3.2 Retrievability Tests

The second series of tests [2], was designed to address retrievability of documents produced by the current DOE conversion system from the Autonomy search system. In part one of this test, we designed an experiment to compare retrievability from document collections that had been manually zoned (**manual zoned**) to retrievability from the identical document set that had been automatically zoned (**auto zoned**). The idea here is to determine if retrievability from auto zoned collections (i.e., zoned by the OCR system) is as good as retrievability from collections that were zoned by human operators.

Another consideration is the order in which documents are returned by the Autonomy system. In part two of this test, we compared the ordering of retrieved documents from the manually zoned and from the automatically zoned collections. In this study, and in the first retrieval test described above, we used a 1055 document subset of the DOE collection with 40 queries, “typical” of queries likely issued to the LSN, and relevance judgements for each query for each document.

Finally, we designed a test to compare retrievability from a collection of documents that were 99.8% correct to retrievability from the same set that were auto zoned with MANICURE post-processing. The idea here is to determine if retrievability from collections produced with auto zoning & MANICURE is as good as retrievability from collections that are close to 100% correct (i.e., in this case 99.8% character accuracy). In this study, we used 1058 documents from the LSS prototype collection with 68 queries, again “typical” of queries likely issued to the LSN, and relevance judgements for each query for each document.

---

**3** In conducting the accuracy test [1], an additional level of character accuracy was measured. The accuracy of “unique” non-stopwords was slightly higher than these percentages.

### 3.2.1 Retrievability from Manually vs. Automatically zoned Collections

Table 3 shows the average precision for the manual zoned and the auto zoned collections. Although average precision from the auto zoned collection differs by 3.5% from the manual zoned collection, this difference cannot be considered statistically significant. Basically, we can only conclude that retrievability between these two collections is equivalent.

Precision	Manual zoned	Auto zoned
Average	0.379	0.392

Table 3: Average Precision for Manually Zoned vs. Automatically Zoned Collections

### 3.2.2 Ranking from Automatically Zoned Collections

Another consideration, related to retrievability, is ranking. This is important because it makes a difference to the user if a relevant document is ranked 3rd or 300th. Thus, we also performed a detailed study to determine if any rank variability exists between retrieval results from manual zoned and auto zoned collections. Some IR systems available in the mid-1990's showed variability in ranking output of documents with automatically zoned collections. In fact, ISRI research during this period played a role in correcting this problem.

#### 3.2.2.1 Ranking Problems in Information Retrieval Systems

Ranking variability in optically recognized documents is typically due to the concept of *document length normalization*. IR engines use normalization to treat short and long documents equivalently. Long documents generally have more distinct words than short documents. Also, long and short documents about the same subject matter may have the same set of distinct words, but the frequencies of these words are much higher for longer documents. IR systems typically use the number of distinct words, or the maximum frequency of words, to adjust the weights of terms in the documents in order to give fair representation to words in shorter documents. In optically recognized documents, mis-recognized words inflate the number of distinct words and the maximum frequency. The fact that this inflation leads to ranking variability was first pointed out by Taghva in 1994 and 1996 [6, 7].

Fortunately, after this discovery, the concept of length normalization was revisited by Prof. Gerard Salton and his student Amit Singhal at Cornell University. Their efforts led to redefining length normalization [8]. New measures were defined that depend on the byte size of the document (i.e. the number of characters in the file) eliminating extreme rank variability. Modern IR engines either use the new measure, or a similar concept, which is not affected by misrecognized words.

#### 3.2.2.2 Results of Ranking Test

We summarized the ranking for both manual zoned and auto zoned collections by average relevant document rank and standard deviation. We further calculated the correlation coefficient between the ranks of the same documents for both the manual zoned and auto zoned versions. These values appear in Table 4. From the Table, we note that average relevant document rank and standard deviation for these sets are exactly equal. This is the first indication that document ranking in these two sets is similar. But it's actually the correlation coefficient  $r = 0.97$  that convinces us the ranking in these two sets is very close. The correlation coefficient ranges between 0 and 1, where the  $r$  value of 1 indicates identical correlation;  $r = 0.97$  says there is a very strong association between the ranked lists. A scatterplot pictorially shows this relationship. Figure 1 shows how the ranks of the manual zoned and the auto zoned tightly cluster around the regression line that begins at the origin.

<b>Average Rank Automatic</b>	<b>Standard Deviation Automatic</b>	<b>Average Rank Manual</b>	<b>Standard Deviation Manual</b>
289	258	289	258
Correlation coefficient $r = 0.97$			

Table 4. Average Rank and Standard Deviation for Auto Zoned and Manual Zoned documents

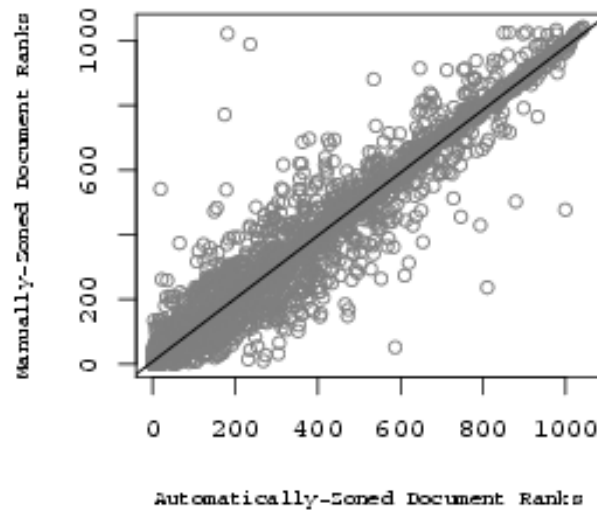


Figure 1. Scatter plot of Automatically Zoned vs. Manually Zoned Ranks for all Documents

### 3.2.3 Retrievability from 99.8% Correct vs. Auto zoned & MANICURED Collections

Our final comparison test ties the OCR accuracy tests to retrieval. As mentioned in [1], building a large collection of OCR ground-truth data is an arduous task. The requirement for exact duplication in ASCII of an image of a page is difficult to obtain for even a small set of pages.

Fortunately, ISRI has in its collection a 99.8% correct set of documents with queries and relevancy judgments that had been prepared for the LSS Prototype. This is not equivalent to OCR ground-truth since carriage-returns, spacing, and columnization are not equivalent to the hard copy page. Still, the typed text was measured to be 99.8% correct. Since this is higher than the current goals set by the NRC, we felt by comparing the results of the auto zoned version to this 99.8% correct version, we could help clarify if improving character accuracy would improve retrieval results. The average precision for both sets appears in Table 5.

<b>Precision</b>	<b>99.8% correct</b>	<b>Auto zoned with MANICURE</b>
Average	0.245	0.242

Table 5: Average Precision for 99.8% Correct Text vs. Automatically Zoned and Recognized Text

Note that average precision differs by only 1.2% (i.e. retrievability is statistically equivalent). We can therefore conclude that the process used by DOE to prepare documents for the LSN will return results equivalent to a collection whose character accuracy was corrected to 99.8%.

#### 4.0 Summary

There are several things that can be concluded from these studies:

1. The character accuracy produced by the DOE conversion system is close to NRC requirements. For good quality images it is exactly 99.5% (see MANICURE character accuracy for the original image in Table 2).
2. The effect of MANICURE on character and word accuracies is uniformly positive.
3. Retrievability from automatically zoned collections is equivalent to retrievability from manually zoned collections.
4. Result ranking from automatically zoned collections is equivalent to ranking from manually zoned collections.
5. Retrievability from automatically zoned and MANICURED collections is equivalent to retrievability from 99.8% correct collections.

In conclusion, we believe that the combination of these accuracy tests and retrieval tests, demonstrate that the quality of the documents delivered by the DOE will give effective retrieval results for the users of the LSN.

#### References

- [1] T. Nartker and R. Young, OCR Accuracy Produced by the Current DOE Document Conversion System, Technical Report 2002-06, Information Science Research Institute, University of Nevada, Las Vegas, May 2002.
- [2] K. Taghva, J. Borsack, S. Lumos, and A. Condit, Retrievability of Documents Produced by the Current DOE Document Conversion System, Technical Report 2002-05, Information Science Research Institute, University of Nevada, Las Vegas, May 2002.
- [3] K. Taghva, J. Borsack, A. Condit, and S. Erva, The Effects of Noisy Data on Text Retrieval, *J. American Society for Information Science*, 45(1):50-58, January 1994.
- [4] K. Taghva, A. Condit, J. Borsack, J. Kilburg, C. Wu, and J. Gilbreth, The MANICURE Document Processing System, Proc. IS&T/SPIE 1998 Intl. Symp. on Electronic Imaging Science and Technology, 1998, San Jose, CA, January.
- [5] S. Rice, F. Jenkins, and T. Nartker, The Fifth Annual Test of OCR Accuracy, ISRI TR-96-01, April 1996.
- [6] K. Taghva, J. Borsack, and A. Condit, Results of Applying Probabilistic IR to OCR Text, Proc. 17th Intl. ACM/SIGIR Conf. On Research and Development in Information Retrieval, 1994, pp. 202-211.
- [7] K. Taghva, J. Borsack, and A. Condit, Effects of OCR Errors on Ranking and Feedback Using the Vector Space Model, *Inf. Proc. and Management*, 1996, 32(3), pp. 317-327.
- [8] A. Sinhal, G. Salton, and C. Buckley, Length Normalization in Degraded Text Collections, SDAIR 1996, pp. 149-162.