

# Retrievability of Documents Produced by the Current DOE Document Conversion System

Kazem Taghva, Julie Borsack, Steve Lumos, and Allen Condit

Technical Report 2002-05  
Information Science Research Institute  
University of Nevada, Las Vegas

April 25, 2002

## 1 Introduction

The Licensing Support Network (LSN), managed by the Nuclear Regulatory Commission (NRC), will provide information to all interested parties that is potentially relevant to the licensing of the high-level radioactive waste repository proposed for Yucca Mountain. There are several contributors of documents and information, the largest of which is the Department of Energy (DOE). Since the DOE will contribute over a million hard copy documents to the LSN, the most accurate and efficient method for capturing these documents in electronic form is crucial to producing this information in a timely way.

What's more, effective retrievability should be of the utmost concern. Whatever capture method is selected, it should produce the expected results. For many years, the information retrieval community has employed two well-accepted measures for determining retrieval effectiveness: *recall* and *precision*. These are defined and described in Appendix A. We apply these same measures in this report to explain and clarify the results of our studies. The retrievability studies we perform evaluate and compare automatic and manual methods that could be used for document conversion. We report on:

- Manual zoning vs. automatic zoning (Section 2).
- Post-processing applied to improve the OCR text produced (Section 3).
- The DOE's document conversion methods and compare them to nearly perfect (99.8% correct) text (Section 4).

These tests, in conjunction with the results reported in [2], will provide in-depth analysis of the performance of the current DOE document conversion system.

## 2 Manual Zoning vs. Automatic Zoning

### 2.1 Environment

The zoning experiments described in this report required that ISRI produce an environment that duplicates the systems and procedures applied by the DOE to prepare their documents for the LSN. Further, the documents and queries we use for these experiments should be a good representation of the expected LSN collection and its anticipated use. The system and procedures we applied in all the following tests are exactly what the DOE and NRC had in place at the date of testing:

**OCR:** Scansoft v10 with the MTX OCR module

Document count	1055
Number of pages	75,236
Query Count	40
Average number of relevant documents/query	100
Median number of relevant documents/query	64
Fewest relevant documents for a query	2
Most relevant documents for a query	608

Table 1: Dataset statistics

Recall Points	Manual	Auto
0.00	0.78	0.84
0.10	0.58	0.61
0.20	0.50	0.54
0.30	0.44	0.46
0.40	0.41	0.42
0.50	0.36	0.35
0.60	0.31	0.31
0.70	0.27	0.26
0.80	0.22	0.22
0.90	0.18	0.18
1.00	0.12	0.12
Avg	0.379	0.392

Table 2: 11-Point Precision for Manually and Automatically Zoned Sets

**Post Processing:** MANICURE v1.7

**Retrieval Engine:** Autonomy Knowledge Server v2.1

The test collection that we use to compare manual vs. automatic zoning consists of 1055 documents that were selected from documents in the RIS with the document type “Report,” “Plan,” “Design Document,” or “Correspondence.” Manual zoning information for each page was made available to us from the zoning procedures conducted by the Yucca Mountain Project management and operations contractor prior to 1999. Forty queries, with relevancy judgments were produced by UNLV geology students who were familiar with the RIS collection. Table 1 shows some statistics for this dataset.

## 2.2 Recall/Precision Results

Two collections were produced: one applying the manual zoning information described above (call it `manual`), the other applying automatic zoning performed by Scansoft (call it `auto`). Both collections were loaded and indexed into Knowledge Server. All 40 queries were batch run against these two datasets in exactly the same way. Table 2 shows average precision at 11 recall points for both collections.

Recall and precision are the accepted measures applied in the IR community for comparing retrieval results. For a more complete discussion of these measures, see Appendix A. Note that the 11-point average for `auto` is 3.5% better than for `manual`. This higher average return for `auto` indicates that running these queries against this data set gives better results from automatic zoning than one could expect from manual zoning. This difference is not statistically significant though. To be statistically significant for this size collection, the difference would have to be 5% or greater. What we can learn from these results is that in general, automatic zoning gives results as good as those obtained from manually-zoned OCR.

Docid	Manual Rank	Automatic Rank
MOL.19990701.0270	6	3
MOL.19981008.0009	1	3
HQO.19950224.0009	5	3
NNA.19870625.0060	5	3
NNA.19920504.0221	35	3
MOL.19981009.0176	2	3
MOL.19980609.0061	10	3
MOL.19980123.0860	19	3
NNA.19870331.0563	3	3
MOL.19980122.0032	11	3
MOL.19990702.0236	5	3
MOL.19981008.0006	3	3
NNA.19920528.0154	135	3
MOL.19980716.0493	2	3
MOL.19981008.0002	3	3
MOL.19980729.0051	5	3
MOL.19980729.0047	3	3
MOL.19980724.0391	9	3

Table 3: Automatic and Manual Ranks

### 2.3 Ranked Query Analysis

The 11-point precision average indicates that there is no difference in query results for the methods used for collection preparation. On the micro level, we felt it important to investigate what exactly happens to individual query rankings. In other words, if a relevant document was ranked, say 25, in the automatic results, what would be the rank of the same document in the manual version? By reviewing the query-by-query results for both versions, we observed that there was no significant variation between rankings. The following paragraphs are the technical details supporting our observation.

The collection has 1055 documents. Hence, for a specific query, a relevant document can be ranked between 1 and 1055. Obviously, we would like to see the relevant documents ranked as close to 1 as possible. Now consider all the relevant documents that were ranked, say 3, for the automatic version. We may ask, what are the rankings of these document in the manual version? Table 3 shows all the documents ranked 3 in automatic and the corresponding rank of the same documents in the manually-zoned version. We can represent these points as as (3,6), (3,1), (3,5), etc.

If we continue this process for all the ranks and plot these points, we will discover the scatter plot in Figure 1. This graph exhibits the relationship between the corresponding rankings. In other words, for a fixed rank  $m$  on the  $x$  axis (Automatic), then the  $y$  values (Manual) represent the corresponding manual ranks for the same documents. We can summarize this graph by average ranks, standard deviation (SD), and the correlation coefficient  $r$  as shown in Figure 1. The  $r$  value shows the strength of the association between the two variables. The  $r$  value ranges between 0 and 1. The closer  $r$  is to 1, the stronger the association.

We can use this plot to draw the *regression line* and use the *regression method* to predict the rank of the manual ranks from the automatic ranks. The solid line in Figure 1 represents the regression line.

As can be seen in the scatter plot, the points in this graph are tightly clustered around the regression line. This clustering indicates a strong linear association between the two variables. In general, we use Equation 1 for calculating the predicted manual ranking of the same document in the manual version.

Table 4 shows examples of the ranks of documents in the manually-zoned set when the automatically-zoned rank is known.

$$manual\_average - \left[ \left( \frac{auto\_average - auto\_rank}{auto\_SD} \right) (r)(manual\_SD) \right] = manual\_rank \quad (1)$$

Auto Average	Auto SD	Manual Average	Manual SD
289	258	289	258
correlation coefficient $r = 0.97$			

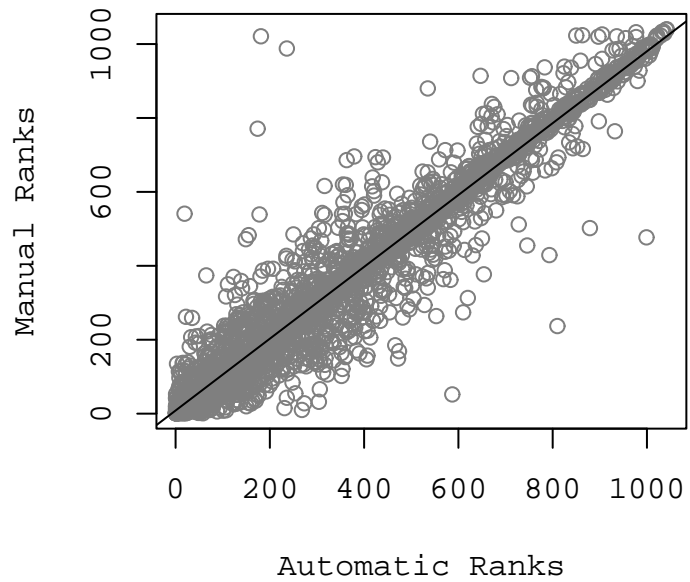


Figure 1: Manual vs. Automatic Rank Scatter Plot

Full Collection Ranks	
Auto Average 289 Automatic Rank	Manual Average 289 Manual Rank
200	203
150	154
100	106
50	56
10	19
350	349
400	397
First Quadrant Ranks	
Auto Average 103 Automatic Rank	Manual Average 109 Manual Rank
50	57
25	32
5	13
150	155

Table 4: Examples of Corresponding Ranks for Both Collections

We can use the above analysis to extrapolate the rank performance of the manual from the automatic version. As can be seen from the Table 4, the further away we are from the point of average the bigger difference we see in the ranks. This is the way the regression method works. To lend more credence to our analysis, we did the same calculation for the first quadrant ranks (i.e. only the ranks between 1 and 261), since these are the documents that the user will most likely evaluate. The graph in Figure 2 represents the correspondence between these ranks. The second half of Table 4 shows examples of the predicted ranks for the manually-zoned version for some rankings just in the first quadrant.

The regression method is a scientific way of comparing the ranking correspondence between the two collections, or in this case between two versions of the same collection. In our experiments, it can be seen that there is no significant difference in the ranking between the two methods of document preparation.

### 3 Verification of Procedures

Based on ISRI’s OCR and information retrieval research, several post-processing routines were built to improve the quality of OCR text loaded into a retrieval system. This set of processes eventually was streamlined into a system we call MANICURE[9]. Together with ISRI, DOE has been tuning this system specifically to LSN documents.

The major components of MANICURE include `ppsys`[4], a process that automatically corrects misspellings in the text, and `rmgarbage`, a process which removes “graphic text” and other non-retrieval strings from an automatically-zoned and OCR’d document. Previous experiments [5] have proven the effectiveness of `ppsys` but the efficacy of `rmgarbage` had yet to be tested. A simple means of testing its effects was to run the same experiment discussed in Section 2 except that the tested sets would be two versions of `auto`: `auto` with `rmgarbage` and `auto` without `rmgarbage`. All other processing steps remained the same. The results of this test appear in Table 5.

Reviewing these results, we see slightly increased precision at the highest recall values and average precision is nearly 40% when `rmgarbage` is used. And again, although not statistically significant, for this set of documents, a small but definite improvement is apparent.

These experiments show that with automatic zoning and MANICURE, users of the LSN will obtain retrieval results equivalent to what could be expected with manual zoning. Further, since in many cases non-stopwords [2] are corrected when MANICURE is applied, retrievability can potentially be improved.

Auto Average	Auto SD	Manual Average	Manual SD
103	76	109	91
correlation coefficient $r = 0.82$			

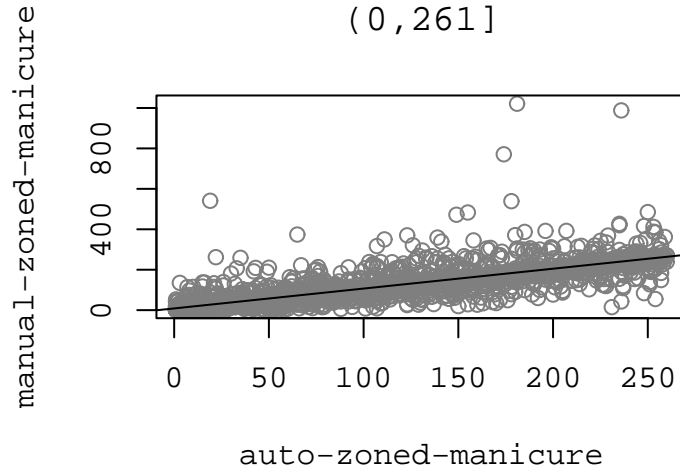


Figure 2: Manual vs. Automatic Rank Scatter Plot, First Quadrant

Recall Points	Auto with rmgarbage	Auto w/o rmgarbage
0.00	0.84	0.83
0.10	0.61	0.60
0.20	0.54	0.53
0.30	0.46	0.44
0.40	0.42	0.40
0.50	0.35	0.35
0.60	0.31	0.31
0.70	0.26	0.26
0.80	0.22	0.22
0.90	0.18	0.18
1.00	0.12	0.12
Avg	0.392	0.385

Table 5: 11-Point Precision for Auto-zoned with/without rmgarbage

Document count	1058
Number of pages	46,731
Query Count	62
Average number of relevant documents/query	17
Median number of relevant documents/query	9
Fewest relevant documents for a query	1
Most relevant documents for a query	99

Table 6: LSS Prototype Statistics

Recall Points	99.8% Correct	Auto-zoned w/MANICURE
0.00	0.55	0.54
0.10	0.46	0.45
0.20	0.35	0.34
0.30	0.29	0.30
0.40	0.26	0.26
0.50	0.22	0.22
0.60	0.18	0.18
0.70	0.14	0.14
0.80	0.12	0.11
0.90	0.08	0.07
1.00	0.06	0.05
Avg	0.245	0.242

Table 7: 11-Point Precision for 99.8% Correct Text and Automatically-zoned and Recognized

## 4 A Comparison Using the Prototype Collection

One might believe that the closer we get to 100% character accuracy, the better the retrieval results we will obtain from a search engine like Autonomy. In fact, one of the goals specified by the NRC is that collections submitted for the LSN should try to reach 99.5% character accuracy across the collection and 98.5% for any particular page. What this next experiment shows (and several other experiments performed at ISRI have shown)[8, 5, 1, 6, 7] is that close to 100% character accuracy may not be necessary for good retrieval performance.

We have a collection of 1058 documents, 62 queries, and 1104 relevancy judgments that we will use to answer this question. This collection is particularly well-suited for determining how character accuracy may affect retrieval performance for a couple of reasons. First, we have two versions of the collection: one version with 99.8% character accuracy[3] and another version that has been recognized and processed as describe in Section 2.1. Second, the documents and the queries in this collection were part of the original LSS Prototype Collection and so they should have similar characteristics and topic content as the planned LSN. Collection statistics appear in Table 6.

Again, as in our tests comparing manual vs. automatic zoning, we report on retrieval results using the standard measures of recall and precision. We loaded and indexed both collections into Knowledge Server and the 62 queries were batch run against these two datasets in exactly the same way. The recall/precision results appear in Table 7.

The difference between average precision for the two runs is less than 0.3%. As we pointed out for the manually-zoned vs. automatically-zoned runs, the difference is too small to be considered statistically significant. This test tells us that the process used by DOE to prepare the documents for the LSN will return results equivalent to a collection that was re-keyed to meet 99.8% character accuracy. With respect to retrievability, an artificially high character accuracy does not guarantee better results for the end user.

## 5 Conclusion

The tests in this report use well-accepted standards and scientific methods to measure and validate the current procedures used by the DOE to prepare documents for the LSN. We believe our investigation is unbiased and complete.

The aggregation of these results indicates that using automatic-zoning followed by MANICURE will give retrieval results equivalent to what one could expect from manually-zoned pages or even from a 99.8% correct collection. We have also shown that there is a strong linear association between ranked results. This association implies that for all practical purposes, the two ranked query result sets are statistically equivalent.

We believe these tests, in conjunction with the results reported in [2], provide a thorough analysis of the performance of the current DOE system for converting documents for inclusion in the LSN.

## References

- [1] W. B. Croft, S. Harding, K. Taghva, and J. Borsack. An evaluation of information retrieval accuracy with simulated OCR output. In *Proc. 3rd Symposium on Document Analysis and Information Retrieval*, pages 115–126, Las Vegas, NV, April 1994.
- [2] Tom Nartker and Ron Young. OCR accuracy produced by the current DOE document conversion system. Technical Report 2002-06, Information Science Research Institute, University of Nevada, Las Vegas, April 2002.
- [3] Science Applications International Corporation. Capture station simulation lessons learned. Final report for the Licensing Support System, 1990. Final report for the Licensing Support System prepared under contract DE-AC01-87RW00084 for the U.S. Department of Energy, Office of Civilian Radioactive Waste Management, Washington D.C.
- [4] Kazem Taghva, Julie Borsack, and Allen Condit. An expert system for automatically correcting OCR output. In *Proc. IS&T/SPIE 1994 Intl. Symp. on Electronic Imaging Science and Technology*, pages 270–278, San Jose, CA, February 1994.
- [5] Kazem Taghva, Julie Borsack, and Allen Condit. Results of applying probabilistic IR to OCR text. In *Proc. 17th Intl. ACM/SIGIR Conf. on Research and Development in Information Retrieval*, pages 202–211, Dublin, Ireland, July 1994.
- [6] Kazem Taghva, Julie Borsack, and Allen Condit. Effects of OCR errors on ranking and feedback using the vector space model. *Inf. Proc. and Management*, 32(3):317–327, 1996.
- [7] Kazem Taghva, Julie Borsack, and Allen Condit. Evaluation of model-based retrieval effectiveness with OCR text. *ACM Transactions on Information Systems*, 14(1):64–93, January 1996.
- [8] Kazem Taghva, Julie Borsack, Allen Condit, and Srinivas Erva. The effects of noisy data on text retrieval. *J. American Soc. for Inf. Sci.*, 45(1):50–58, January 1994.
- [9] Kazem Taghva, Allen Condit, Julie Borsack, John Kilburg, Changshi Wu, and Jeff Gilbreth. The MANICURE document processing system. In *Proc. IS&T/SPIE 1998 Intl. Symp. on Electronic Imaging Science and Technology*, San Jose, CA, January 1998.

## A Recall and Precision Explained

For these tests to provide a quantitative measure of retrieval effectiveness, we must know in advance which documents are relevant to which queries. The *relevancy judgments*, or list of relevant documents to each query, give us this a priori information. We then apply standard quantitative measures to compare the list of documents retrieved by the system to the relevancy judgments.

The standard measures we use are *recall* (2) and *precision* (3). Recall is the percentage of the relevant documents in the collection that are responses to a query. Precision is the percentage of the responses that are relevant to the query. If you think of it from a users perspective, these are the assessments he would use as well, “Have I received all the relevant documents that are in this collection?” (recall). And, “How many documents do I have to look through to find the ones that are relevant?” (precision). Following are the mathematical formulas that calculate these two values:

$$recall = \frac{\#\_of\_relevant\_retrieved\_documents}{total \ \#\_of\_relevant\_documents} * 100 \quad (2)$$

$$precision = \frac{\#\_of\_relevant\_retrieved\_documents}{total \ \#\_of\_retrieved\_documents} * 100 \quad (3)$$

Averaging the precision values at specific recall points gives us a better perspective of the overall retrieved results. For example, if we look at Table 7, the system returns on average 35% of the relevant documents when it has returned 20% of the collection. The collective average, in the last row of this table, is just an average of the precision values at each recall point. As you can see, precision tends to decrease as more documents are returned by the system.

There is also a notion of *statistical significance* that we introduce in this report. This is important because slight differences in precision results may not necessarily indicate a fundamental and consistent difference between the result sets.

For us to make a general statement like “automatic zoning will return more relevant documents than manual zoning” the results must be statistically significant. Statistical significance is related to collection size as well as the number of queries used and the number of relevant documents for the queries. For the datasets we’ve used in these tests, a difference of 5% would be considered significant.