U.S. Department of Energy
Office of Civilian Radioactive Waste Management

YUCCA MOUNTAIN PROJECT

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

Presented to:
NRC/DOE Technical Exchange on Electronic Submissions

Presented by:
Jake Wooley
Deputy Director, Office of Information Management
U.S. Department of Energy
Office of Civilian Radioactive Waste Management
Dr. Tom Nartker, Dr. Kazem Taghva, & Julie Borsack
University of Nevada Las Vegas / Information Science Research Institute

June 26, 2002
Las Vegas, Nevada

Study of Retrievability

- Agenda
  - Introduction
  - Background
  - Document conversion system recommendations
  - Tests to measure document conversion performance
    - Text accuracy
    - Retrievability tests
  - Conclusion

Study of Retrievability

(Continued)

- Background
  - Who is Information Science Research Institute (ISRI)?
  - Licensing Support System (LSS) (1990 - Current)
  - Optical Character Reader (OCR) Conferences (1991 - 1995)
  - Contracted by M&O (1996 - 1999)
  - Contracted by DOE (1990 - 1995, 2000 - Current)
    - Current tasks for FY02
      - » Provide recommendations on DOE document conversion system
      - » Evaluate performance of DOE document conversion system

Study of Retrievability

(Continued)

- Document conversion system recommendations
  - Retrievability is a better performance metric than character accuracy
    - Not all characters are used by a retrieval system
  - Automatic zoning, followed by MANICURE, will produce retrievability equivalent to manual zoning
    - Manually zoned text
    - Automatic zoned text
    - MANICURE

Tests to Measure Document Conversion Performance

- Text accuracy
  - NRC Licensing Support Network Administrator (LSNA) target accuracy for OCR created text (Licensing Support Network Guidelines provided 1/02)
    - Goal is to have 99.5% accurate text
  - Text accuracy test
    - 17 documents (1253 pages, 164,483 non-stopwords, and 1,361,124 characters)
    - Non-stopword accuracy tests
      - » DOE word accuracy between 96.15% and 97.23%
    - Tests of character accuracy of non-stopwords
      - » DOE character accuracy between 98.83% and 99.30%

Tests to Measure Document Conversion Performance

(Continued)

- Retrievability tests
  - Test data
    - 1055 documents containing 75,236 Pages
    - 40 queries
    - Average number of relevancy judgements per query - 100
    - Autonomy Server™ v2.2.0
  - Retrievability metrics
    - Precision and recall

Tests to Measure Document Conversion Performance

(Continued)

- Automatic zoned text versus manually zoned text retrieval tests
  - Average precision
    - Manually zoned - 37.9%
    - Automatic zoned - 39.2%

Tests to Measure Document Conversion Performance

(Continued)

- Ranking of retrieved documents compared between the manually zoned text and the automatic zoned text
  - Importance of ranking in retrieval systems
  - Results of ranking tests
    - Correlation factor - .97
  - Ranking problems in information retrieval systems

Tests to Measure Document Conversion Performance

(Continued)

- Automatic zoned text versus 99.8% correct text retrieval tests
  - 1058 documents containing 46,731 Pages
  - 62 queries
  - Average number of relevant documents per query = 17
  - Average precision
    - 99.8% accurate text - 24.5%
    - Automatic zoned text - 24.2%

Study of Retrievability

- Conclusion
  - Character accuracy produced by DOE document conversion system close to NRC LSNA goal
  - MANICURE improves word and character accuracy
  - Average character accuracy of non-stopwords on DOE documents is between 98.83% and 99.30%
  - Retrievability is equivalent for automatic zoned text and manually zoned text
  - Ranking of query results is equivalent for automatic zoned text and manually zoned text
  - Retrievability is equivalent for automatic zoned text and 99.8% accurate text

Backup

1

Tertiary extension north of the Las Vegas Valley shear zone, Sheep and Desert Ranges, Clark County, Nevada

2

PETER L. GUTH* *Department of Earth and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139*

3

ABSTRACT

Detailed mapping reveals the presence of high-angle extensional faults and low-angle gravity slides on the west side of the Sheep Range. Three major high-angle faulting events each account for 20° of eastward rotation and accommodate extension between the Las Vegas Range and the Desert Range. Low-angle faults represent surficial slides in response to topography produced by extension on the high-angle faults. Faulting took place during the Miocene, synchronously with deposition of the Horse Spring Formation and with displacement on the Las Vegas Valley shear zone. The extension in the Sheep Range took place without volcanism, intrusion, or metamorphism of the Paleozoic sedimentary rocks.

Offset thrust faults suggest that the area west of the Sheep Range extended almost 100% during the Miocene, while the corresponding area south of the Las Vegas Valley shear zone did not extend significantly. The shear zone bounded the extending terrane on the south, acting as a transform fault. This extension west of the Sheep Range may in part balance that mapped by Anderson (1971) in the Eldorado Mountains. The Las Vegas Valley shear zone and the Lake Mead fault system may have acted together to compensate for areas of localized extension between the Colorado Plateau and the vicinity of the Specter Range.

INTRODUCTION

The importance of extensional faulting in the development of structure and physiography in the Great Basin has been evident since the pioneering observations of Gilbert (1874). Recently Stewart (1971, 1980) reviewed and summarized the horst and graben, tilted-block, and listric-fault models commonly applied to the Great Basin. Stewart (1971) calculated that the horst and graben model requires about 10% extension across the entire Great Basin, assuming that 60° dips exist on range-front faults. The tilted-block model of Morton and Black (1975) seems to require 20-30% extension across the entire Great Basin, but local extension might exceed 100% (Stewart, 1980). Inferred listric-fault geometry leads to local estimates of 30% to over 100% extension (Anderson, 1971; Wright and Troxel, 1973; Proffett, 1977), although the model might not apply to the entire Great Basin.

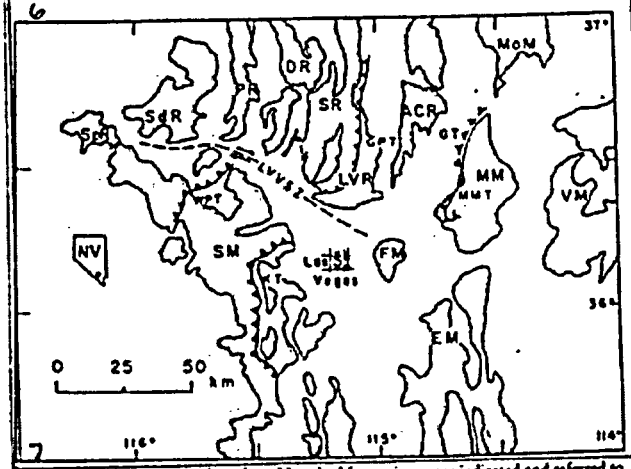
Strike-slip faults are related to extensional faulting as boundaries of domains with differences in style or magnitude of extension (Davis, 1979; Stewart, 1980). Strike-slip faults may be transform boundaries between regions of differential extension, such as the Garlock fault of southern California (Hamilton and Myers, 1966; Davis and Burchfiel, 1973). Recent work suggests this model for faulting in the Lake Mead area of southern Nevada (Bohannon, 1979a). Both strike-slip and extensional faulting appear to be dominantly late Tertiary events in the Great Basin.

Armstrong (1972) reviewed widespread low-angle denudational faulting in the eastern Great Basin. The faults generally place younger rocks on older, and Armstrong advocated a Tertiary gravity mechanism distinct from Mesozoic compression. In con-

4

5

6



7

Figure 1. Locality map of southern Nevada. Mountain ranges indicated and referred to in the text are: ACR, Arrow Canyon Range; DR, Desert Range; EM, Eldorado Mountains; FM, Frenchman Mountain; LVR, Las Vegas Range; MM, Muddy Mountains; MoM, Mormon Mountains; PR, Pintwater Range; SR, Sheep Range; SdR, Spotted Range; SpR, Specter Range; and VM, Virgin Mountains. Thrust faults indicated are the Wheeler Pass (WPT), Keystone (KT), Gass Peak (GPT), Muddy Mountain (MMT), and Glendale (GT).

8 * Present address: B Company, 34th Engineer Battalion (Combat) (Heavy), Fort Riley, Kansas 66442.

9 Geological Society of America Bulletin, Part I, v. 92, p. 763-771, 3 figs., 1 table, October 1981.

Zoned Page Image

Example Queries

- Find all documents that discuss fracture frequency data (fracture density and radial fracture density) in boreholes at Yucca Mountain
- Find all documents that discuss strategies for environmental restoration and remediation (Hanford Site)

Precision and Recall

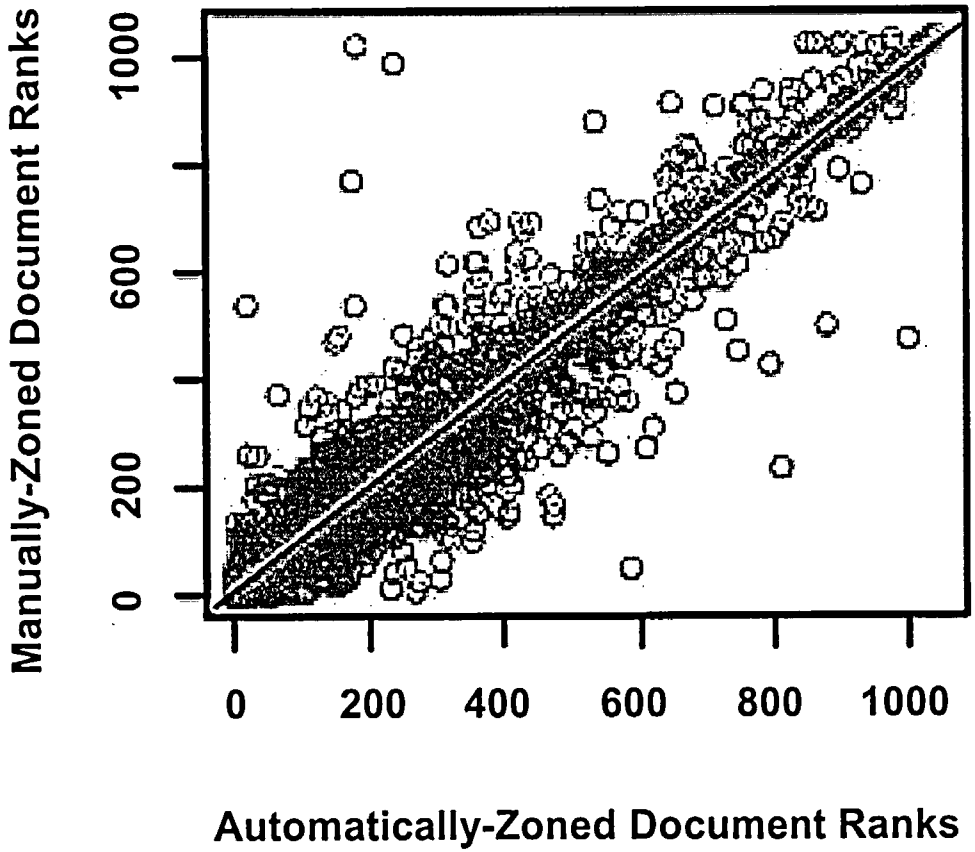
- Precision
  - # of relevant documents retrieved / # of retrieved documents
- Recall
  - # of relevant documents retrieved / total # of relevant documents
- Suppose there are 10 relevant documents for an example query
- And suppose the system returns 15 documents for this example query

Precision and Recall Example

Docid	Relevant	Recall	Precision
1	R	10%	100%
2	--	10%	50%
3	--	10%	33%
4	R	20%	50%
5	R	30%	60%
6	--	30%	50%
7	R	40%	57%
8	--	40%	50%
9	R	50%	55%
10	R	60%	60%
11	R	70%	64%
12	--	70%	58%
13	R	80%	62%
14	R	90%	64%
15	R	100%	67%

NOTE: Docid - Document identified

Calculating Average Precision

- Add precision values at 10%, 20%, . . . , 100% recall levels
  - 100+50+60+57+55+60+64+62+64+67 = 639
- Average total precision / # of recall levels
  - 639 / 10 = 63.9%

Scatter Plot of Automatic Zoned and Manual Zoned Ranks