Rescuing Legacy Transportation Data @ NTL

Lisa Curtin, USDOT Intern & MSIS Candidate 2020
University of Tennessee School of Information Sciences
lisa.curtin2@gmail.com

Mary Moulton, Digital Librarian
National Transportation Library, USDOT
mary.moulton@dot.gov,

U.S. Department of Transportation

RECOMMENDED CITATION:
Curtin, Lisa https://orcid.org/0000-0003-1137-7789 and Mary E. Moulton https://orcid.org/0000-0002-1791-068X. 2020. "Rescuing Legacy Transportation Data @ NTL." Transportation Research Board 99th Annual Meeting. Washington, D.C., USA. https://doi.org/10.21949/1506099

ABSTRACT

Legacy data—data collected or compiled in the past and stored in obsolete formats (e.g., paper and floppy disks)—can be made accessible via the process of data rescue, which may be as simple as scanning documents or, to achieve machine readability, as complex as developing artificial intelligence algorithms. Machine readability not only facilitates the discoverability of data and their reuse for new purposes but also expands the coverage of data sets and supports access for screen reader users. At the National Transportation Library, legacy data exist primarily as scanned-to-PDF documents that, albeit digitized, are not machine-readable due to the quality of scanning. Using Adobe Acrobat Pro's text-recognition function, the Library's current ability to rescue machine-readable data from PDF images has an error rate exceeding 25%. Consequently, in an average table of data, at least 25% of cells require manual correction, which though feasible for occasionally rescuing small datasets is infeasible for large-scale data rescue. As an informal survey of other federal information agencies and a literature review revealed, however, lower error rates are possible with alternative methods of data rescue, including using commercial optical character recognition software and contracting external data rescue service providers.

WHY RESCUE OLD DATA?

- Expand temporal and spatial coverage of existing datasets
- Meet open government and open data priorities
- Make data reusable and reduce duplication of effort
- Make data full-text searchable & findable
- Improve accessibility for screen-reader users

HOW TO RESCUE DATA

- Use commercial Optical Character Recognition (OCR) software such as ABBYY or OmniPage
- Use open-source or design DIY solutions to correct for OCR errors
- Pursue cutting-edge machine learning solutions for data rescue and error correction
- Crowdsourcing to correct OCR errors

LEGACY DATA ARE DATA COLLECTED OR COMPILED IN THE PAST, STORED IN AN OBSOLETE FORMAT.

DATA RESCUE IS THE PROCESS OF MAKING LEGACY DATA ACCESSIBLE.

WHAT DO OTHER AGENCIES DO?

The National Science Foundation funds legacy data rescue projects

The Library of Congress and the Smithsonian utilise public crowdsourcing platforms

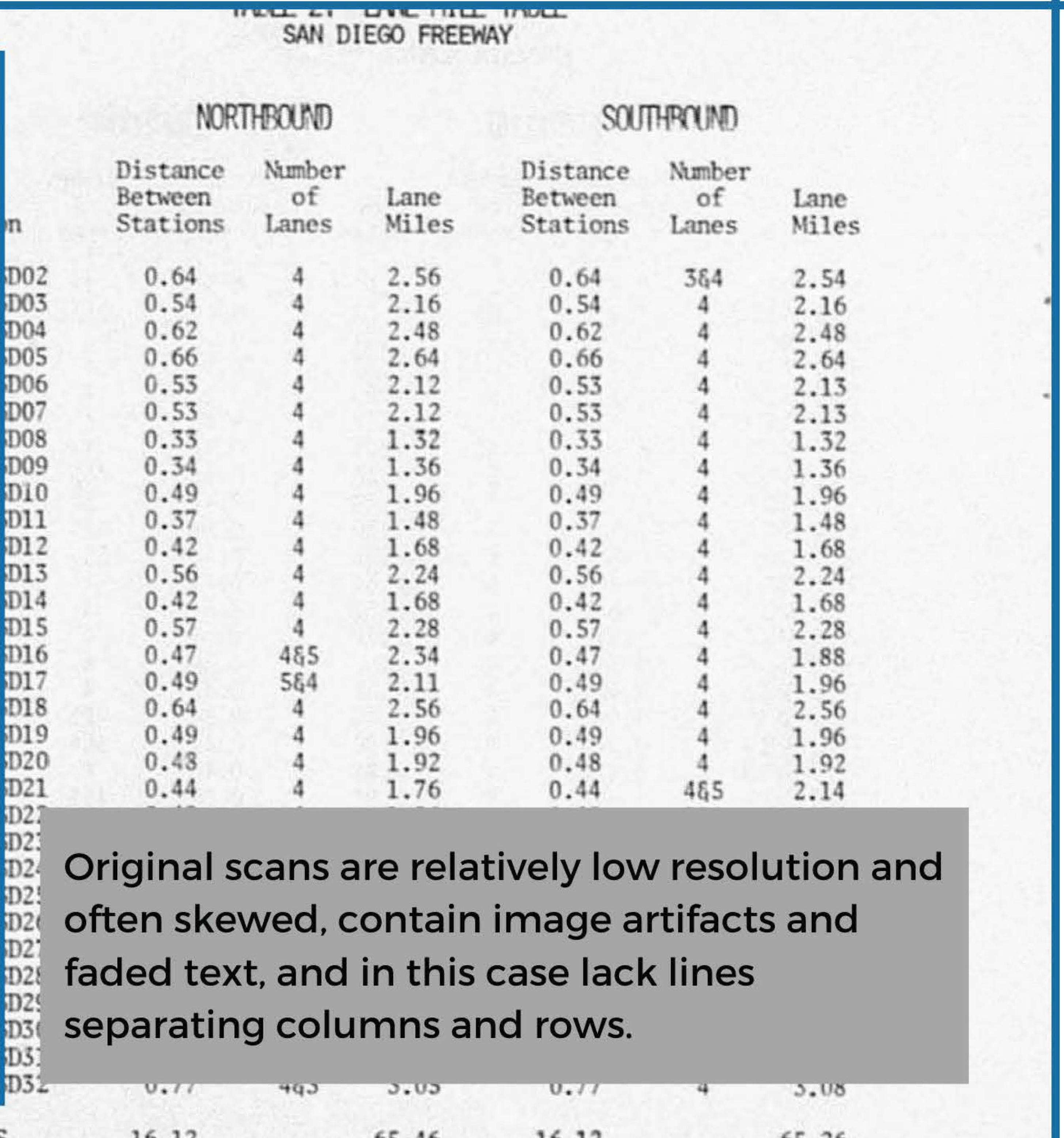
The U.S. Geological Survey uses OmniPage to OCR documents and dabbles in crowdsourcing projects

The National Library of Medicine's PubMed Central converts documents to HTML

The National Institute of Standards and Technology contracts with PubMed Central to convert documents to HTML

The National Library of Education contracts with the Internet Archive for digitization

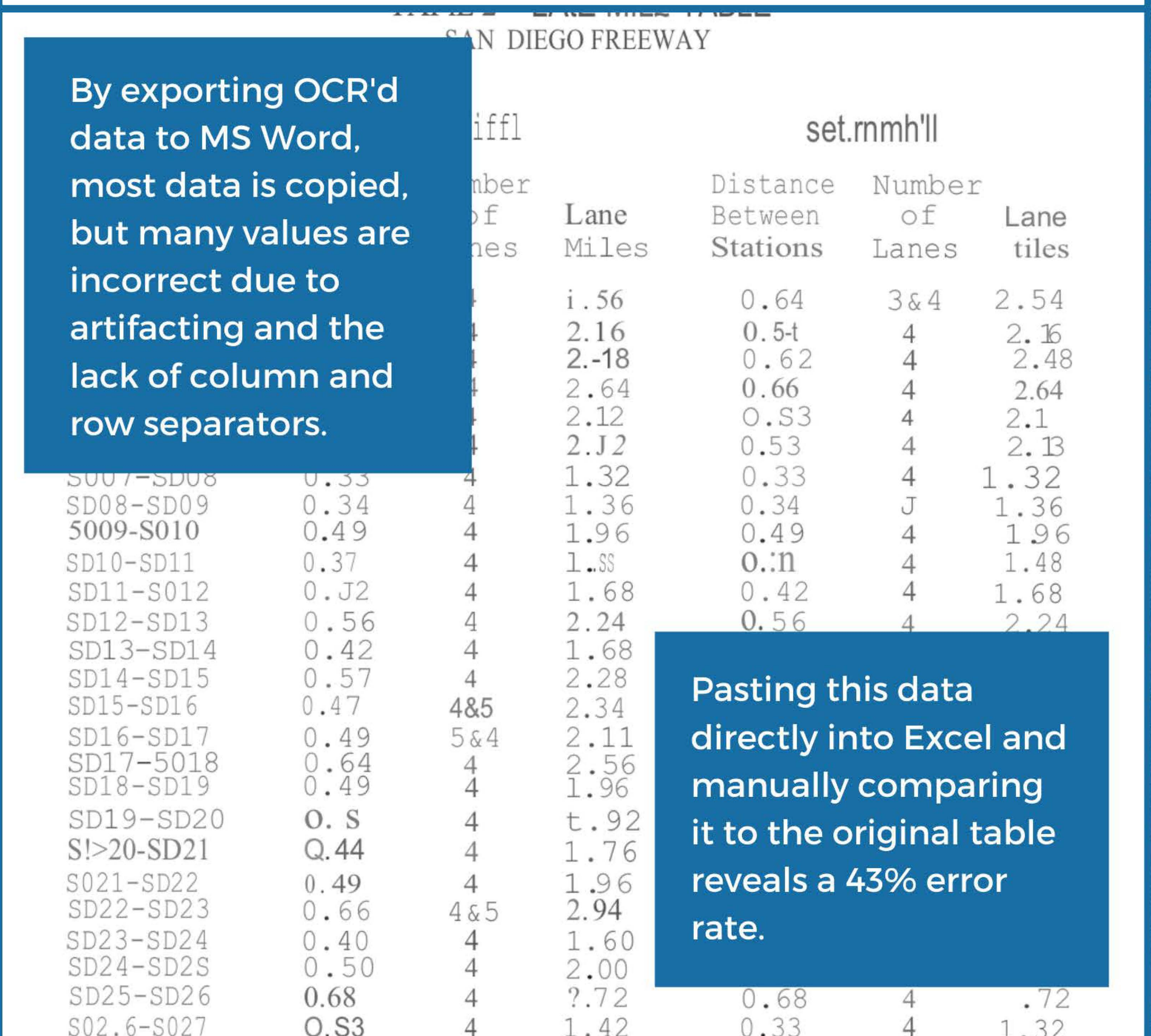
NTL's Current Data Rescue Capability Using Adobe Acrobat Pro



SAN DIEGO FREEWAY

	NORTHBOUND			SOUTHBOUND		
n	Distance Between Stations	Number of Lanes	Lane Miles	Distance Between Stations	Number of Lanes	Lane Miles
D02	0.64	4	2.56	0.64	3&4	2.54
D03	0.54	4	2.16	0.54	4	2.16
D04	0.62	4	2.48	0.62	4	2.48
D05	0.66	4	2.64	0.66	4	2.64
D06	0.53	4	2.12	0.53	4	2.13
D07	0.53	4	2.12	0.53	4	2.13
D08	0.33	4	1.32	0.33	4	1.32
D09	0.34	4	1.36	0.34	4	1.36
D10	0.49	4	1.96	0.49	4	1.96
D11	0.37	4	1.48	0.37	4	1.48
D12	0.42	4	1.68	0.42	4	1.68
D13	0.56	4	2.24	0.56	4	2.24
D14	0.42	4	1.68	0.42	4	1.68
D15	0.57	4	2.28	0.57	4	2.28
D16	0.47	4&5	2.34	0.47	4	1.88
D17	0.49	5&4	2.11	0.49	4	1.96
D18	0.64	4	2.56	0.64	4	2.56
D19	0.49	4	1.96	0.49	4	1.96
D20	0.48	4	1.92	0.48	4	1.92
D21	0.44	4	1.76	0.44	4&5	2.14

Original scans are relatively low resolution and often skewed, contain image artifacts and faded text, and in this case lack lines separating columns and rows.

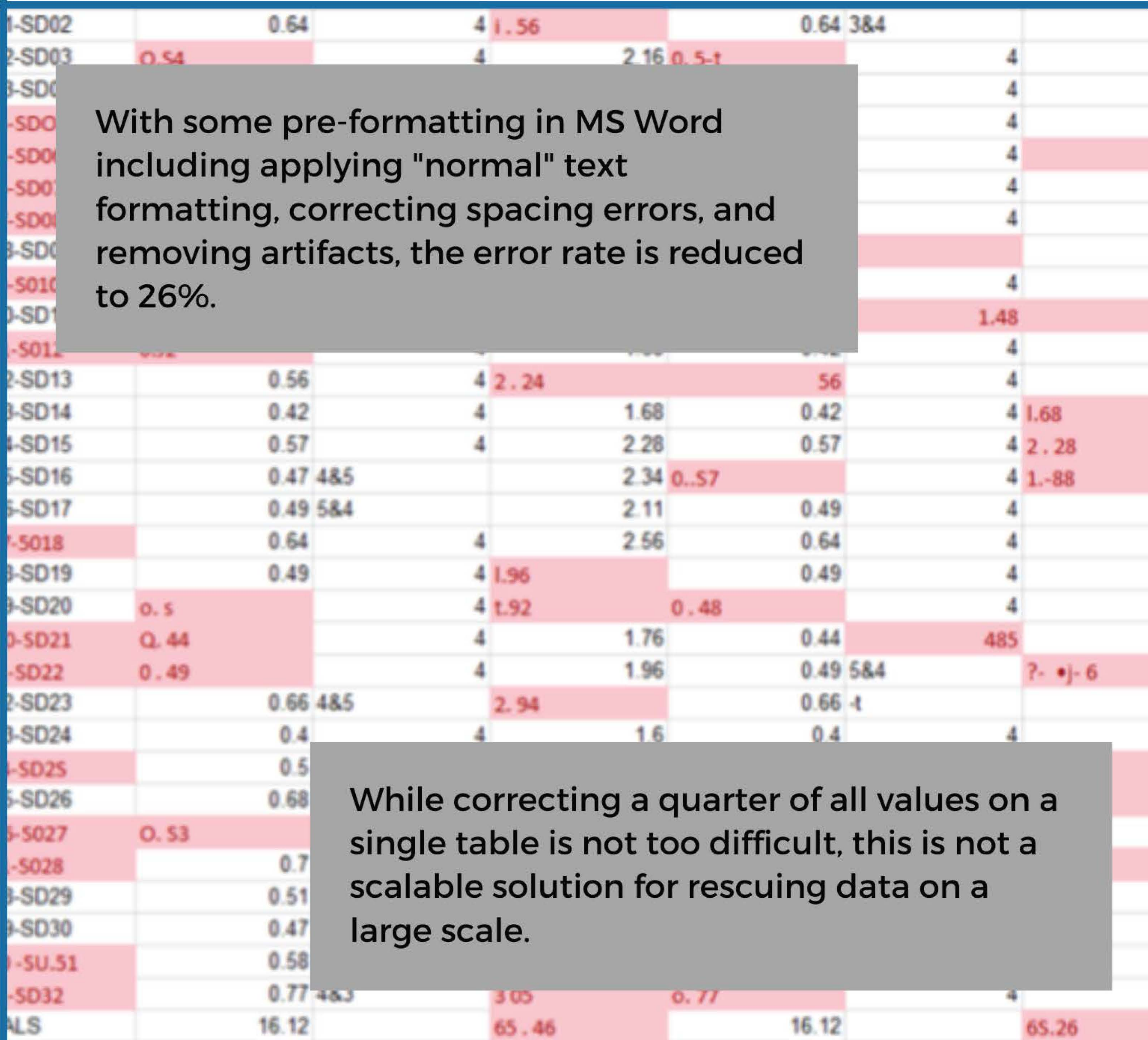


AN DIEGO FREEWAY

		iffl		set.rnmh'll		
		mber f nes	Lane Miles	Distance Between Stations	Number of Lanes	Lane tiles
		4	i.56	0.64	3&4	2.54
		4	2.16	0.5-t	4	2. 16
		4	2.-18	0.62	4	2.48
		4	2.64	0.66	4	2.64
		4	2.12	O.S3	4	2.1
		4	2.J2	0.53	4	2. 13
S007-SD08	0.33	4	1.32	0.33	4	1.32
SD08-SD09	0.34	4	1.36	0.34	J	1.36
5009-S010	0.49	4	1.96	0.49	4	1 96
SD10-SD11	0.37	4	1..SS	0.:n	4	1.48
SD11-S012	0.J2	4	1.68	0.42	4	1.68
SD12-SD13	0.56	4	2.24	0.56	4	2.24
SD13-SD14	0.42	4	1.68			
SD14-SD15	0.57	4	2.28			
SD15-SD16	0.47	4&5	2.34			
SD16-SD17	0.49	5&4	2.11			
SD17-5018	0.64	4	2.56			
SD18-SD19	0.49	4	1.96			
SD19-SD20	O. S	4	t.92			
S!>20-SD21	Q.44	4	1.76			
S021-SD22	0.49	4	1.96			
SD22-SD23	0.66	4&5	2.94			
SD23-SD24	0.40	4	1.60			
SD24-SD2S	0.50	4	2.00			
SD25-SD26	0.68	4	?.72	0.68	4	.72
S02.6-S027	O.S3	4	1.42	0.33	4	1.32

By exporting OCR'd data to MS Word, most data is copied, but many values are incorrect due to artifacting and the lack of column and row separators.

Pasting this data directly into Excel and manually comparing it to the original table reveals a 43% error rate.



1-SD02	0.64	4	i.56	0.64 3&4		
2-SD03	0.54	4	2.16	0.5-t	4	
2-SD13	0.56	4	2.24	56	4	
3-SD14	0.42	4	1.68	0.42	4	1.68
4-SD15	0.57	4	2.28	0.57	4	2.28
5-SD16	0.47 4&5		2.34	0..57	4	1.-88
6-SD17	0.49 5&4		2.11	0.49	4	
7-5018	0.64	4	2.56	0.64	4	
8-SD19	0.49	4	1.96	0.49	4	
9-SD20	o. s	4	t.92	0.48	4	
0-SD21	Q. 44	4	1.76	0.44	485	
-SD22	0.49	4	1.96	0.49 5&4		?- •j- 6
2-SD23	0.66 4&5		2.94	0.66 -t		
3-SD24	0.4	4	1.6	0.4	4	
ALS	16.12		65.46	16.12		65.26

With some pre-formatting in MS Word including applying "normal" text formatting, correcting spacing errors, and removing artifacts, the error rate is reduced to 26%.

While correcting a quarter of all values on a single table is not too difficult, this is not a scalable solution for rescuing data on a large scale.