Microtask Digest compared to OCR market leaders

Microtask Digest is a specialized *data extraction service* for processing large amounts of scanned material. Unlike any other OCR solution, Digest utilizes *hybrid intelligence* where state-of-the-art machine intelligence is combined with on-demand human intelligence. This unique combination gives Digest an edge over traditional, machine-only OCR algorithms that struggle with low-resolution scans and artifacts in images.

Digest goes beyond plain OCR by providing customers the data they need in the format they want, regardless of the format of the original material. Automatic structurization, smart data extraction and correlation are all part of the service, ensuring the delivery contains just the right information in the right place.

Adjustable quality assurance process gives customer full control over the quality of the deliverables within the given budget. Quality can be adjusted iteratively while the material is being processed allowing customers to allocate just the right amount of budget for reaching the target quality.

In this whitepaper we compare the performance of Microtask Digest to industry leading OCR solutions using typical low- to medium-quality archive material.

Test material

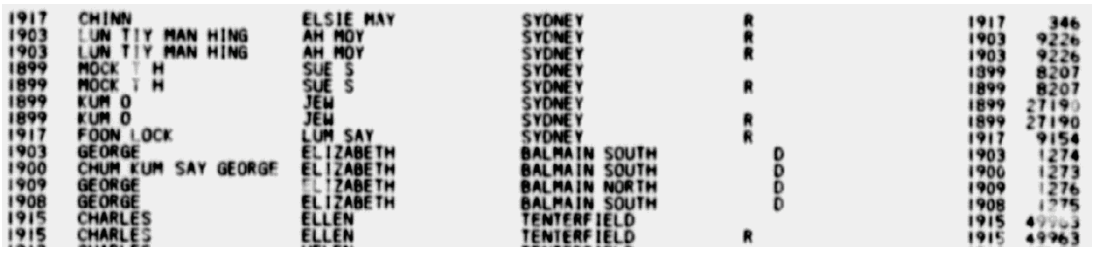
As the test material, we used a project with 130,000 scanned microfilm pages with a relatively low scanning resolution of 2400x1900 pixels per page. By far the largest factor in overall quality was, however, the quality of the original microfilm. Many pages contained scratches and other artifacts. Some pages were originally filmed a bit out of focus.

We took three representative samples of the material, to highlight typical problem areas for OCR engines. In this paper we are showing only crops of the sample images even though all the samples were processed as full page images.

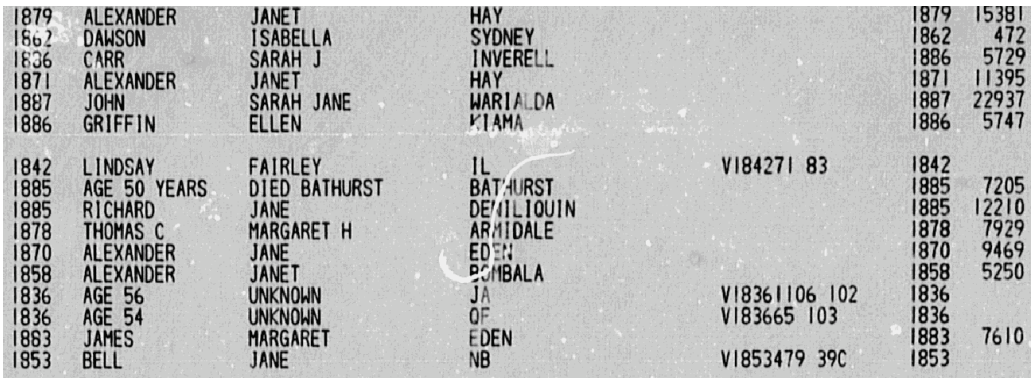
All OCR engines were used “as is” without any additional training. English language model and dictionary was used when available. Microtask Digest built its dictionary automatically from the source material. In addition Digest was configured to use minimal amount of human processing to make results as comparable as possible. Adding more human processing would improve quality in most difficult to read cases (while increasing the cost as well), but even with these settings, we were able to reach 99.6% character level accuracy over the whole material set.

Sample 1 – high quality, provides a reference

Sample 2 – low quality, small text



Sample 3 – artifacts



Test case 1

Original scan

1936	78 YRS MENTAL HOSP	GLADESVILLE REDFERN	DRUMMOYNE		1936	6868
1921	SMITH	CLARICE	KEMPSEY		1921	7218
1920	LARNACH	LAURA M	LITHGOW		1920	4029
1927	EDWARD C	CLARA L	SYDNEY		1927	17463
1936	POLAK	FLORENCE MAY	WOOLLAHRA		1936	16309
1938	KIRK	MAUREEN THELMA	BULLI		1938	23333
1945	SMITH	LOUISA FLORENCE	MURRUMBURRAH		1945	3855
1937	MILLER	JOAN ANNETTA	LIVERPOOL		1937	22578
1939	0 SHEA	URSULA ANNIE	MURWILLUMBAH		1939	25196
1939	OSHEA	URSUI ٩ ANNIE	MURWILLUMBAH	R	1939	25196

ABBYY FineReader 12 Professional

1936	78 YRS MENTAL HOSP	GLADESVILLE REDFERN	DRUMMOYNE		1936	6868
1921	SMITH	CLARICE	KEMPSEY		1921	7218
1920	LARNACH	LAURA M	LITHGOU		1920	4029
1927	EDUARD C	CLARA L	SYDNEY		1927	17463
1936	POLAK	FLORENCE MAY	UOOLLAHRA		1936	16309
1938	KIRK	MAUREEN THELMA	BULL I		1938	23333
1945	SMITH	louisa Florence	MURRUMBURRAH		1945	3855
1937	MILLER	JOAN ANNETTA	LIVERPOOL		1937	22578
1939	0 SHEA	URSULA ANNIE	MURU1LLUMBAH		1939	25196
1939	OSHEA	URSUI A ANNIE	MURUILLUMBAH	R	1939	?*!? C

FineReader does a pretty good job on the reference image, with some problems in telling U and W apart. The partially visible number in the lower right corner is not captured correctly at all. For some reason, one name in the middle (LOUISA FLORENCE) has been capitalized differently.

ReadIRIS 14

1936	7B YRS P1ENTAL HOSP	GLADESVILLE REDFERN	ORU111tOYNE		1936	686B
1921	SNITH	CLARICE	KENPSEY		1921	7218
1920	LARNACH	LAURA N	LITHGOW		1920	4029
1927	EDWARD C	CLARA L	SYDNEY		1927	17463
1936	POLAK	FLORENCE NAY	WOOLLAHRA		1936	16309
193B	KIRK	NAUREEN THELNA	BULL!		193B	23333
194!,	SNITH	LOUISA FLORENCE	P1URRUI1BURRAH		1945	3855
1937	HILLER	JOAN ANIUTA	LIVERPOOL		1937	22578
1939	OSHEA	UISIA.A MW IE	t1URWILLUI1BAH		1939	25196
1939	OSIEA	UISUI \ ANNIE	NURW I LLUNBAH	R	1939	?SI%

ReadIRIS has clearly the worst performance with a lot of mistakes even in this clear reference image. Number 8 and letter B seem to be difficult for ReadIRIS as well as non-dictionary words such as MURRUMBURRAH. As with FineReader, the last number is not recognized at all.

Omnipage Cloud

936	78 YRS MENTAL HOSP	GLADESVILLE REDFERN	DRUMMOYNE	936	6868
921	SMITH	CLARICE	KEMPSEY	921	7218
920	LARNACH	LAURA M	LITHGOW	920	4029
927	EDWARD C	CLARA L	SYDNEY	927	17463
936	POLAK	FLORENCE MAY	WOOLLAHRA	936	16309
938	KIRK	MAUREEN THELMA	BULLI	938	23333
	SMITH	LOUISA FLORENCE	MURRUMBURRAH	945	3855
937	MILLER	JOAN ANNETTA	LIVERPOOL	937	22578
939	0 SHEA	URSULA ANNIE	MURWILLUMBAH	939	25196
939	OSHEA	URSUI 1 ANNIE	MURWILLUMBAH	939	'Ai%

Omnipage gets the best result of the three commercial OCR engines, but regards the digit 1 in the year as a vertical line, thus removing it altogether. As with others, the last number is not recognized.

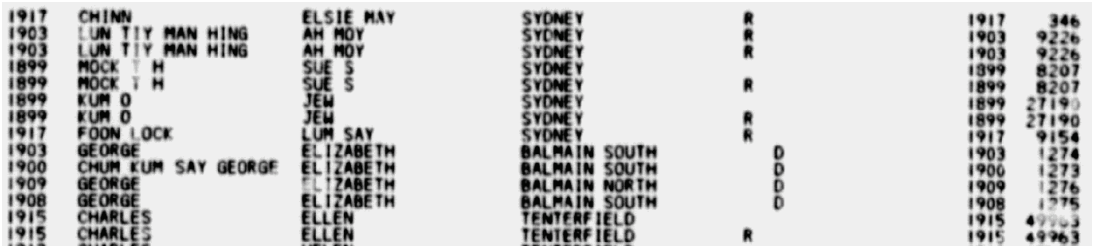
Microtask Digest

1936	78 YRS MENTAL HOSP	GLADESVILLE REDFERN	DRUMMOYNE		1936	6868
1921	SMITH	CLARICE	KEMPSEY		1921	7218
1920	LARNACH	LAURA M	LITHGOW		1920	4029
1927	EDWARD C	CLARA L	SYDNEY		1927	17463
1936	POLAK	FLORENCE MAY	WOOLLAHRA		1936	16309
1938	KIRK	MAUREEN THELMA	BULLI		1938	23333
1945	SMITH	LOUISA FLORENCE	MURRUMBURRAH		1945	3855
1937	MILLER	JOAN ANNETTA	LIVERPOOL		1937	22578
1939	O SHEA	URSULA ANNIE	MURWILLUMBAH		1939	25196
1939	OSHEA	URSHA ANNIE	MURWILLUMBAH	R	1939	79196

Digest delivers best quality, with problems only in areas where the original text has faded out. Even the difficult number in the lower right corner is partially recognized.

Test case 2

Original scan



ABBYY FineReader 12 Professional

If IT	CHINN	ELSIE NAV	SYONE	R	If If	144
190)	I JN ▼ ! Y NAN MIN	AM NOY	SYDNEY	R	IfOJ	922G
IfOI I8ff	LUN T1 Y NAN	AM NOY	SYDNC	R	IfOI	922%
	HIN6 NOCK H	SUE $	SYONE'		• Off	8207
I8ff	MB 1 M	Sue s	SYDNEY	R	1891	•207
I0ff	KUN 0	JEW	SYONE1		• 891	jflw
Ifff	• UN 0	JEW	SV0NI^	R	1891	27190
If 17	ROON LOCI	LUN SAY	SYDNEY	R	• 9I 7	9154
1909	mSm	Elizabeth	BALMAIN SOUTH	0	I90J	1274
1*00	CHUN KUM SAY QI	ELIZABETH	GAL RAIN SOUTH	I	m	IIf 9
If Of	Iif OR Gf		BALMAIN NORTH	0	IfOI	'III
If 08	810R8C	ELIZABETH	BALHAIN SOUTH	0	• 901	• *• ▼5
Ifj5	CHARLES	ELLEN	TENTfRfIf10		jffl	4 » .
1915	«MARIf .	ELLEN	TENURfIELO	R	1915	«f»*J

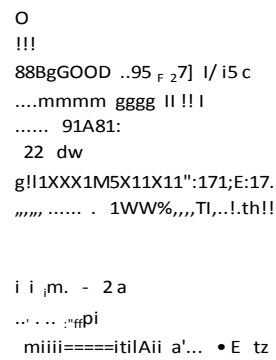
FineReader has trouble reading the blurred small text, although it is able to recognize some words and numbers correctly.

ReadIRIS 14

i,17	CHIHN	(LSI(11.\Y	$YONEY	•		1 •• ,	, ..
1,0)	UJN TI T IIAN H	AH PIOY	$\'()IIfY	•		1,0,	•21-
1,0)	UM TIT IIAN HI	AH PIOY	SYOIIEY	•		ItOJ	•zzb
'"'"	IIOCIC - H	SU($	SYDNEY			Ittt	1201
'"'"	IIOCIC I H	SUf $	SYOIIET	•		•• ••	1201
.. .,	(UII O	JO,	SYOIIEY			••••	111•1
'"'"	JUI! 0	JEW	SYMY	•		""	211•0
1,i1	fOON I DC«	LUii SAT	~YOIIEY	•		1,11	,150
i,01	81:0IIII:	(1.IZAIEIH	IAI.IIIIIN sourH		D	"°'	IZH
ItOO	~ 'IJII SA'r OEO	fLjI •• IH	IAI.1111 IN SOU		0	•• 00	IZU
•• o,	IIfOIIII:	fI ZAtI!IH	IAI.IIAIN II()IIIH		0	1,0,	II7b
,,oe	IIfOIIIIf	fLIZAIEIH	IAI"AIN SOUIH		0	i•ot	t .,.,,
... ~	CHML£$	EII.EN	I!NIUfIfLO			1t1S	• t,-J
1,r5	=m	ELLEN	IENIUFIfLO	•		...~	**•1'-J**

ReadIRIS performs worse than FineReader but at least gets the structure of the page quite right.

Omnipage Cloud


O
!!!
88BgGOOD ..95 F 27] I/ i5 c
....mmmm gggg II !! I
..... 91A81:
22 dw
g!l1XXX1M5X11X11":171;E:17.111n4r221)22"WK1441-XiiS337i
„„„ ..... . 1WW%,,,,Tl,..!.th!!!!:'4141.1.41EWL;:::;1;

i i im. - 2 a
..' . .. :"ffpi
miiii=====itilAii a'... • E tz

Omnipage Cloud service is unable to figure out even the structure of the page, not to mention words or numbers.

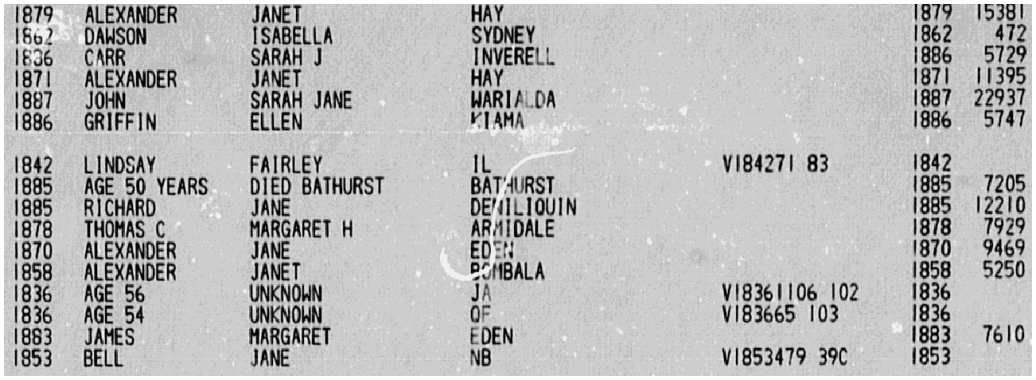
Microtask Digest

1917 CHINN	ELSIKA MAY	SYDNEY	R	1917	346
1903 TAN TIY MAN HING	AH MOY	SYDNEY	R	1903	9226
1903 LUN TIY MAN HING	AH MOY	SYDNEY	R	1903	9226
1899 MOCK H	SUE S	SYDNEY		1899	8207
1899 MOCK I H	SUE S	SYDNEY	R	1899	8207
1899 KUM O	JEW	SYDNEY		1899	27191
1899 KUM O	JEW	SYDNEY	R	1899	27190
1917 KOONTOCK	LUM SAY	SYDNEY	R	1917	9154
1903 GEORGE	ELIZABETH	BALMAIN SOUTH	R	1903	1274
1900 CHUM KUM SAY GEORGE	ELIZABETH	BALMAIN SOUTH	R	1900	1273
1909 GEORGE	ELIZABETH	BALMAIN NORTH	R	1909	1276
1908 GEORGE	ELIZABETH	BALMAIN SOUTH	R	1908	1475
1915 CAKARLES	HELLEN	TENTERFIELD		1915	47794
1915 CHARLES	ELLEN	TENTERFIELD	R	1915	49963

Digest gets the best results by far, even though some of the faded parts are not recognized correctly.

Test case 3

Original scan



1879	ALEXANDER	JANET	HAY		1879	15381
1862	DAWSON	ISABELLA	SYDNEY		1862	472
1836	CARR	SARAH J	INVERELL		1886	5729
1871	ALEXANDER	JANET	HAY		1871	11395
1887	JOHN	SARAH JANE	WARIALDA		1887	22937
1886	GRIFFIN	ELLEN	KIAMA		1886	5747
1842	LINDSAY	FAIRLEY	IL	VI84271 83	1842	
1885	AGE 50 YEARS	DIED BATHURST	BATHURST		1885	7205
1885	RICHARD	JANE	DEWILIOUIN		1885	12210
1878	THOMAS C	MARGARET H	ARMIDALE		1878	7929
1870	ALEXANDER	JANE	EDEN		1870	9469
1858	ALEXANDER	JANET	BOMBALA		1858	5250
1836	AGE 56	UNKNOWN	JA	VI8361106 102	1836	
1836	AGE 54	UNKNOWN	OF	VI83665 103	1836	
1883	JAMES	MARGARET	EDEN		1883	7610
1853	BELL	JANE	NB	VI853479 39C	1853	

ABBYY FineReader 12 Professional

1879	ALEXANDER	JANET	HAY		1879	15381
IBb!	DAWSON	ISABELLA	SYDNEY		1862	472
IF! 36	CARR	SARAH J	INVERELL		1886	5729
1871	ALEXANDER	JANET	HAY		1871	11395
1887	JOHN	SARAH JANE	WAR 1 A.DA		1887	22937
1886	GRIFFIN	ELLEN	KIAMA		1886	5747
1842	LINDSAY	FAIRLEY	IL	VIB427I 83	1842	7205
1885	AGE 50 YEAR	DIED BATHURST	BATHURST		1885	
1885	RICHARD	JANE	DE'IIIIOUIN		1885	12210
I87B	THOMAS C	MARGARET H	ARii I DALE		1878	7929
1870	ALEXANDER	JANE	EO-IJ		1870	9469
1858	ALEXANDER	JANET	BCHBALA	VI836I106 102	1858	5250
1836	AGE 56	UNKNOWN	JA		1836	
1836	AGE 54	UNKNOWN	OF	V183665 103	1836	7610
1883	JAMES	MARGARET	EDEN	VI853479 39C	1883	
1853	BELL	JANE	NB		1853	

FineReader gets most of the text right, but artifacts confuse the algorithm rather badly.

ReadIRIS 14

IB79	ALEXANDER	JANET	HAY		1879	1538 I
1662	DAWSON	I SABELLA	SYDNEY		1862	472
1836	CARR	SARAH J	INVERELL	~	1886	5729
IB71	ALEXANDER	JANET	HAY		1871	11395
IBS7	JOHN	SARAH JANE	UAWLDA		1887	22
IB86	GRIFFIN	ELLEN	1-'.IAHA		1886	5747
IB42	LINDSAY	FAIRLEY	IL .	V 184271 Bl	1842	
IB85	AGE 50 YEARS	DIED BATHURST	BAT\!URST		1885	7205
IB85	RICHARD	JANE	DEWLIOUIN		1885	12210
187B	THOMAS C	MARGARET H	A~ IOALE	a	1878	7929
IB70	ALEXANDER	JANE	EQ^{0}tl		1870	9469
IBSB	ALEXANDER	JANET	%HBALA		1858	5250
1836	AGE 56	UNKNO~N	JA	VI8361106 102	1836	
IB36	AGE 54	UNKNOWN	OF	VIBJ665 103	1836	
1883	JAMES	MARGARET	EDEN		18B3	7610
1853	BELL	JANE	NB	V 1853479 39C	1853	

ReadIRIS has trouble telling numbers 1 and 8 from letters I and B, in addition to the artifact areas.

Omnipage Cloud

	879	ALEXANDER JANET	HAY		879	15381
Rii!		DAWSON ISABELLA	SYDNEY		862	47
E36		CARR SARAH	INVERELL		886	5729
	871	ALEXANDER JANET	HAY		871	11395
887	886	JOHN SARAH JANE	GRIFFIN ELLEN	WARILDA,MA	887	886 22937
	842	LINDSAY FAIRLEY	IL	VI84271 83	842	
	885	AGE 50 YEARS DIED BATHURST	BAT-1011ST		885	7205
	885	RICHARD JANE	DE'IILIOUIN		885	12210
878	870	THOMAS C MARGARET H	ARAIDALE		878	7929
		ALEXANDER JANE	EDT		870	9469
	858	ALEXANDER JANET	[1]3St1BALA		858	5250
	836	AGE 56 UNKNOWN	JA	VI8361106 102	836	
	836	AGE 54 UNKNOWN	OF	V183665 103	836	
	883	JAMES MARGARET	EDEN		883	7610
	953	BELL JANE	NB	V1853479 39C	853	

Omnipage is again regarding the starting 1s as vertical lines, disposing them altogether. Artifacts are confusing the OCR and it is also having trouble figuring out where the column/row breaks are.

Microtask Digest

1879	ALEXANDER	JANET	HAY		1879	15381
1862	DAWSON	ISABELLA	SYDNEY		1862	472
1886	CARR	SARAH J	INVERELL		1886	5729
1871	ALEXANDER	JANET	HAY		1871	11395
1887	JOHN	SARAH JANE	WARIALDA		1887	22937
1886	GRIFFIN	ELLEN	KIAMA		1886	5747
1842	LINDSAY	FAIRLEY	IL	V184271 83	1842	
1885	AGE 50 YEARS	DIED BATHURST	BATHURST		1885	7205
1885	RICHARD	JANE	DENILIOUIN		1885	12210
1878	THOMAS C	MARGARET H	ARMIDALE		1878	7929
1870	ALEXANDER	JANE	EDEN		1870	9469
1858	ALEXANDER	JANET	BOMBALA		1858	5250
1836	AGE 56	UNKNOWN	JA	V18361106 102	1836	
1836	AGE 54	UNKNOWN	OP	V183665 103	1836	
1883	JAMES	MARGARET	EDEN		1883	7610
1853	BELL	JANE	NB	V1853439 39C	1853	

Digest is able to figure out the words behind artifacts with relatively good accuracy. The characters O and Q are extremely similar in this font (DENILIQUIN vs DENILIOUIN) and cannot be told apart without dictionary assistance.

How Digest can beat industry leaders?

Looking at the results from the tests above, it seems incredulous how a new OCR solution can be so much better than industry leaders with dozens of years of R&D behind them.

Unlike general purpose OCR engines such as FineReader and Omnipage, Digest is a very specialized system focusing on very specific material class. Digest algorithms are built for processing massive amounts of very similar pages, using the information gathered across pages to resolve difficult recognition tasks within individual pages. Throwing a single page at Digest to chew on doesn't deliver exceptional results, but feeding another 1000 will increase the quality dramatically.

Using the data from thousands of pages, Digest is able to build dynamic language models and dictionaries based solely on that data. This, in turn, enables it to figure out correct words even when they are partially obscured by artifacts.

OCR engines typically come with hundreds if not thousands of fonts, whereas Digest has none. Instead of trying to match existing fonts to the characters on the scanned image, Digest does human assisted pattern-matching. Having a fully dynamic "font system", Digest learns any font at any size, as long as there is enough material.

Human intelligence is also utilized in other parts of the iterative processing pipeline, to solve hard problems where machine intelligence is having trouble. Relying on human assistance allows the algorithms to be more straightforward and effective, while delivering excellent quality at a competitive price point.

Summary

Microtask Digest is a data extraction service specializing in processing large volumes of similar data. It is offered as a turnkey service instead of licensed software making it very easy and risk-free to use.

Replace expensive keying and unreliable OCR with a hybrid solution that delivers the best of both.

Simply send us your data and the extraction specification and we'll take care of the rest!